

INTRODUCTORY BIOELECTRONICS

INTRODUCTORY BIOELECTRONICS FOR ENGINEERS AND PHYSICAL SCIENTISTS

Ronald Pethig

Stewart Smith

School of Engineering

The University of Edinburgh, UK



A John Wiley & Sons, Ltd., Publication

This edition first published 2013

© 2013, John Wiley & Sons, Ltd

Registered office

John Wiley & Sons, Ltd, The Atrium, Southrn Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Pethig, Ronald.

Introductory bioelectronics : for engineers and physical scientists / Ronald Pethig, Stewart Smith.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-119-97087-3 (cloth)

1. Bioelectronics—Textbooks. I. Smith, Stewart, 1975- II. Title.

QH509.5.P48 2012

572'.437—dc23

2012016834

A catalogue record for this book is available from the British Library.

Print ISBN: 9781119970873

Set in 10/12pt, Times-Roman by Thomson Digital, Noida, India.

Contents

About the Authors	xiii
Foreword	xv
Preface	xvii
Acknowledgements	xix
1 Basic Chemical and Biochemical Concepts	1
1.1 Chapter Overview	1
1.2 Energy and Chemical Reactions	1
1.2.1 Energy	1
1.2.2 Covalent Chemical Bonds	2
1.2.3 Chemical Concentrations	4
1.2.4 Nonpolar, Polar and Ionic Bonds	6
1.2.5 Van der Waals Attractions	7
1.2.6 Chemical Reactions	9
1.2.7 Free-Energy Change ΔG Associated with Chemical Reactions	10
1.3 Water and Hydrogen Bonds	15
1.3.1 Hydrogen Bonds	16
1.4 Acids, Bases and pH	18
1.4.1 The Biological Importance of pH	20
1.4.2 The Henderson-Hasselbalch Equation	21
1.4.3 Buffers	24
1.5 Summary of Key Concepts	25
Problems	26
References	27
Further Readings	27
2 Cells and their Basic Building Blocks	29
2.1 Chapter Overview	29
2.2 Lipids and Biomembranes	29
2.2.1 Fatty Acids	30
2.3 Carbohydrates and Sugars	32
2.4 Amino Acids, Polypeptides and Proteins	34
2.4.1 Amino Acids and Peptide Bonds	35

2.4.2 <i>Polypeptides and Proteins</i>	39
2.5 Nucleotides, Nucleic Acids, DNA, RNA and Genes	43
2.5.1 <i>DNA</i>	43
2.5.2 <i>Ribonucleic Acid (RNA)</i>	47
2.5.3 <i>Chromosomes</i>	50
2.5.4 <i>Central Dogma of Molecular Biology (DNA Makes RNA Makes Protein)</i>	50
2.6 Cells and Pathogenic Bioparticles	51
2.6.1 <i>Prokaryotic and Eukaryotic Cells</i>	52
2.6.2 <i>The Plasma Membrane</i>	53
2.6.3 <i>The Cell Cycle</i>	54
2.6.4 <i>Blood Cells</i>	55
2.6.5 <i>Bacteria</i>	58
2.6.6 <i>Plant, Fungal and Protozoal Cells</i>	60
2.6.7 <i>Viruses</i>	61
2.6.8 <i>Prions</i>	62
2.6.9 <i>Cell Culture</i>	63
2.6.10 <i>Tissue Engineering</i>	64
2.6.11 <i>Cell–Cell Communication</i>	66
2.7 Summary of Key Concepts	70
References	71
Further Readings	71
3 Basic Biophysical Concepts and Methods	73
3.1 Chapter Overview	73
3.2 Electrostatic Interactions	74
3.2.1 <i>Coulomb's Law</i>	74
3.2.2 <i>Ions in Water</i>	78
3.2.3 <i>The Formation of an Ionic Double Layer</i>	79
3.2.4 <i>Ion–Dipole and Dipole–Dipole Interactions</i>	86
3.2.5 <i>Ions in a Membrane or Protein</i>	88
3.3 Hydrophobic and Hydration Forces	90
3.3.1 <i>Hydrophobic Forces</i>	90
3.3.2 <i>Hydration Forces</i>	91
3.4 Osmolarity, Tonicity and Osmotic Pressure	91
3.4.1 <i>Osmoles</i>	91
3.4.2 <i>Calculating Osmolarity for Complex Solutions</i>	92
3.4.3 <i>Osmolarity Versus Tonicity</i>	92
3.5 Transport of Ions and Molecules across Cell Membranes	94
3.5.1 <i>Diffusion</i>	94
3.5.2 <i>Osmosis</i>	95
3.5.3 <i>Facilitated Diffusion</i>	97
3.5.4 <i>Active Transport</i>	97
3.6 Electrochemical Gradients and Ion Distributions Across Membranes	99
3.6.1 <i>Donnan Equilibrium</i>	100
3.7 Osmotic Properties of Cells	103

3.8	Probing the Electrical Properties of Cells	105
3.8.1	<i>Passive Electrical Response</i>	108
3.8.2	<i>Active Electrical Response</i>	108
3.8.3	<i>Membrane Resistance</i>	108
3.8.4	<i>Membrane Capacitance</i>	109
3.8.5	<i>Extent of Ion Transfer Associated with the Membrane Resting Potential</i>	110
3.9	Membrane Equilibrium Potentials	111
3.10	Nernst Potential and Nernst Equation	112
3.11	The Equilibrium (Resting) Membrane Potential	114
3.12	Membrane Action Potential	116
3.12.1	<i>Nerve (Axon) Membrane</i>	117
3.12.2	<i>Heart Muscle Cell Membrane</i>	118
3.13	Channel Conductance	120
3.14	The Voltage Clamp	121
3.15	Patch-Clamp Recording	122
3.15.1	<i>Application to Drug Discovery</i>	123
3.16	Electrokinetic Effects	124
3.16.1	<i>Electrophoresis</i>	124
3.16.2	<i>Electro-Osmosis</i>	129
3.16.3	<i>Capillary Electrophoresis</i>	132
3.16.4	<i>Dielectrophoresis (DEP)</i>	137
3.16.5	<i>Electrowetting on Dielectric (EWOD)</i>	143
	References	145
4	Spectroscopic Techniques	147
4.1	Chapter Overview	147
4.2	Introduction	148
4.2.1	<i>Electronic and Molecular Energy Transitions</i>	148
4.2.2	<i>Luminescence</i>	150
4.2.3	<i>Chemiluminescence</i>	150
4.2.4	<i>Fluorescence and Phosphorescence</i>	150
4.3	Classes of Spectroscopy	151
4.3.1	<i>Electronic Spectroscopy</i>	153
4.3.2	<i>Vibrational Spectroscopy</i>	156
4.3.3	<i>Rotational Spectroscopy</i>	157
4.3.4	<i>Raman Spectroscopy</i>	159
4.3.5	<i>Total Internal Reflection Fluorescence (TIRF)</i>	160
4.3.6	<i>Nuclear Magnetic Resonance (NMR) Spectroscopy</i>	162
4.3.7	<i>Electron Spin Resonance (ESR) Spectroscopy</i>	163
4.3.8	<i>Surface Plasmon Resonance (SPR)</i>	163
4.3.9	<i>Förster Resonance Energy Transfer (FRET)</i>	164
4.4	The Beer-Lambert Law	165
4.4.1	<i>Limitations of the Beer-Lambert Law</i>	168
4.5	Impedance Spectroscopy	170
	Problems	173

References	175
Further Readings	175
5 Electrochemical Principles and Electrode Reactions	177
5.1 Chapter Overview	177
5.2 Introduction	178
5.3 Electrochemical Cells and Electrode Reactions	180
5.3.1 <i>Anodes and Cathodes</i>	181
5.3.2 <i>Electrode Reactions</i>	182
5.3.3 <i>Electrode Potential</i>	184
5.3.4 <i>Standard Reduction Potential and the Standard Hydrogen Electrode</i>	187
5.3.5 <i>The Relative Reactivities of Metal Electrodes</i>	189
5.3.6 <i>The Nernst Equation</i>	192
5.4 Electrical Control of Electron Transfer Reactions	194
5.4.1 <i>Cyclic Voltammetry</i>	197
5.4.2 <i>Amperometry</i>	200
5.4.3 <i>The Ideal Polarised Electrode</i>	201
5.4.4 <i>Three-Electrode System</i>	201
5.5 Reference Electrodes	203
5.5.1 <i>The Silver-Silver Chloride Reference Electrode</i>	204
5.5.2 <i>The Saturated-Calomel Electrode</i>	205
5.5.3 <i>Liquid Junction Potentials</i>	207
5.6 Electrochemical Impedance Spectroscopy (EIS)	208
Problems	212
References	213
Further Readings	213
6 Biosensors	215
6.1 Chapter Overview	215
6.2 Introduction	215
6.3 Immobilisation of the Biosensing Agent	217
6.3.1 <i>Physical Methods</i>	217
6.3.2 <i>Chemical Methods</i>	218
6.4 Biosensor Parameters	218
6.4.1 <i>Format</i>	218
6.4.2 <i>Transfer Function</i>	220
6.4.3 <i>Sensitivity</i>	220
6.4.4 <i>Selectivity</i>	221
6.4.5 <i>Noise</i>	221
6.4.6 <i>Drift</i>	222
6.4.7 <i>Precision and Accuracy</i>	222
6.4.8 <i>Detection Limit and Decision Limit</i>	224
6.4.9 <i>Dynamic Range</i>	226
6.4.10 <i>Response Time</i>	226
6.4.11 <i>Resolution</i>	227

<i>6.4.12 Bandwidth</i>	227
<i>6.4.13 Hysteresis</i>	227
<i>6.4.14 Effects of pH and Temperature</i>	228
<i>6.4.15 Testing of Anti-Interference</i>	228
6.5 Amperometric Biosensors	228
<i>6.5.1 Mediated Amperometric Biosensors</i>	231
6.6 Potentiometric Biosensors	233
<i>6.6.1 Ion Selective Electrodes (ISEs)</i>	235
6.7 Conductometric and Impedimetric Biosensors	237
6.8 Sensors Based on Antibody–Antigen Interaction	240
6.9 Photometric Biosensors	242
6.10 Biomimetic Sensors	245
6.11 Glucose Sensors	247
6.12 Biocompatibility of Implantable Sensors	252
<i>6.12.1 Progression of Wound Healing</i>	252
<i>6.12.2 Impact of Wound Healing on Implanted Sensors</i>	254
<i>6.12.3 Controlling the Tissue Response to Sensor Implantation</i>	254
<i>6.12.4 Regulations for and Testing of Implantable Medical Devices</i>	255
References	255
Further Readings	256
7 Basic Sensor Instrumentation and Electrochemical Sensor Interfaces	259
7.1 Chapter Overview	259
7.2 Transducer Basics	260
<i>7.2.1 Transducers</i>	260
<i>7.2.2 Sensors</i>	260
<i>7.2.3 Actuators</i>	260
<i>7.2.4 Transduction in Biosensors</i>	260
<i>7.2.5 Smart Sensors</i>	261
<i>7.2.6 Passive vs. Active Sensors</i>	262
7.3 Sensor Amplification	262
<i>7.3.1 Equivalent Circuits</i>	262
7.4 The Operational Amplifier	264
<i>7.4.1 Op-Amp Basics</i>	264
<i>7.4.2 Non-inverting Op-Amp Circuit</i>	265
<i>7.4.3 Buffer Amplifier Circuit</i>	266
<i>7.4.4 Inverting Op-Amp Circuit</i>	267
<i>7.4.5 Differential Amplifier Circuit</i>	267
<i>7.4.6 Current Follower Amplifier</i>	268
7.5 Limitations of Operational Amplifiers	269
<i>7.5.1 Resistor Values</i>	269
<i>7.5.2 Input Offset Voltage</i>	269
<i>7.5.3 Input Bias Current</i>	269
<i>7.5.4 Power Supply</i>	270
<i>7.5.5 Op-Amp Noise</i>	270
<i>7.5.6 Frequency Response</i>	270

7.6	Instrumentation for Electrochemical Sensors	271
7.6.1	<i>The Electrochemical Cell (Revision)</i>	271
7.6.2	<i>Equivalent Circuit of an Electrochemical Cell</i>	271
7.6.3	<i>Potentiostat Circuits</i>	272
7.6.4	<i>Instrumentation Amplifier</i>	274
7.6.5	<i>Potentiostat Performance and Design Considerations</i>	275
7.6.6	<i>Microelectrodes</i>	277
7.6.7	<i>Low Current Measurement</i>	277
7.7	Impedance Based Biosensors	278
7.7.1	<i>Conductometric Biosensors</i>	278
7.7.2	<i>Electrochemical Impedance Spectroscopy</i>	280
7.7.3	<i>Complex Impedance Plane Plots and Equivalent Circuits</i>	281
7.7.4	<i>Biosensing Applications of EIS</i>	283
7.8	FET Based Biosensors	284
7.8.1	<i>MOSFET Revision</i>	284
7.8.2	<i>The Ion Sensitive Field Effect Transistor</i>	287
7.8.3	<i>ISFET Fabrication</i>	290
7.8.4	<i>ISFET Instrumentation</i>	291
7.8.5	<i>The REFET</i>	292
7.8.6	<i>ISFET Problems</i>	293
7.8.7	<i>Other FET Based Sensors</i>	293
	Problems	294
	References	296
	Further Readings	296
8	Instrumentation for Other Sensor Technologies	297
8.1	Chapter Overview	297
8.2	Temperature Sensors and Instrumentation	298
8.2.1	<i>Temperature Calibration</i>	298
8.2.2	<i>Resistance Temperature Detectors</i>	298
8.2.3	<i>p-n Junction Diode as a Temperature Sensor</i>	301
8.3	Mechanical Sensor Interfaces	304
8.3.1	<i>Piezoresistive Effect</i>	304
8.3.2	<i>Applications of Piezoresistive Sensing</i>	306
8.3.3	<i>Piezoelectric Effect</i>	311
8.3.4	<i>Quartz Crystal Microbalance</i>	311
8.3.5	<i>Surface Acoustic Wave Devices</i>	315
8.3.6	<i>Capacitive Sensors</i>	317
8.3.7	<i>Capacitance Measurement</i>	319
8.3.8	<i>Capacitive Bridge</i>	320
8.3.9	<i>Switched Capacitor Circuits</i>	322
8.4	Optical Biosensor Technology	325
8.4.1	<i>Fluorescence</i>	325
8.4.2	<i>Optical Fibre Sensors</i>	328
8.4.3	<i>Optical Detectors</i>	329

8.4.4	<i>Case Study: Label Free DNA Detection with an Optical Biosensor</i>	332
8.5	Transducer Technology for Neuroscience and Medicine	335
8.5.1	<i>The Structure of a Neuron</i>	335
8.5.2	<i>Measuring and Actuating Neurons</i>	336
8.5.3	<i>Extracellular Measurements of Neurons</i>	339
	Problems	340
	References	341
	Further Readings	342
9	Microfluidics: Basic Physics and Concepts	343
9.1	Chapter Overview	343
9.2	Liquids and Gases	343
9.2.1	<i>Gases</i>	344
9.2.2	<i>Liquids</i>	346
9.3	Fluids Treated as a Continuum	346
9.3.1	<i>Density</i>	346
9.3.2	<i>Temperature</i>	346
9.3.3	<i>Pressure</i>	347
9.3.4	<i>Maxwell Distribution of Molecular Speeds</i>	349
9.3.5	<i>Viscosity</i>	352
9.4	Basic Fluidics	354
9.4.1	<i>Static Fluid Pressure</i>	354
9.4.2	<i>Pascal's Law</i>	354
9.4.3	<i>Laplace's Law</i>	355
9.5	Fluid Dynamics	356
9.5.1	<i>Conservation of Mass Principle (Continuity Equation)</i>	356
9.5.2	<i>Bernoulli's Equation (Conservation of Energy)</i>	358
9.5.3	<i>Poiseuille's Law (Flow Resistance)</i>	360
9.5.4	<i>Laminar Flow</i>	361
9.5.5	<i>Application of Kirchhoff's Laws (Electrical Analogue of Fluid Flow)</i>	364
9.6	Navier-Stokes Equations	365
9.6.1	<i>Conservation of Mass Equation</i>	366
9.6.2	<i>Conservation of Momentum Equation (Navier-Stokes Equation)</i>	367
9.6.3	<i>Conservation of Energy Equation</i>	369
9.7	Continuum versus Molecular Model	369
9.7.1	<i>Solving Fluid Conservation Equations</i>	370
9.7.2	<i>Molecular Simulations</i>	372
9.7.3	<i>Mesoscale Physics</i>	375
9.8	Diffusion	378
9.9	Surface Tension	383
9.9.1	<i>Surfactants</i>	384
9.9.2	<i>Soap Bubble</i>	384

9.9.3	<i>Contact Wetting Angle</i>	385
9.9.4	<i>Capillary Action</i>	386
9.9.5	<i>Practical Aspects of Surface Tension for Lab-on-Chip Devices</i>	388
	Problems	388
	References	389
	Further Readings	390
10	Microfluidics: Dimensional Analysis and Scaling	391
10.1	Chapter Overview	391
10.2	Dimensional Analysis	391
10.2.1	<i>Base and Derived Physical Quantities</i>	393
10.2.2	<i>Buckingham's π-Theorem</i>	394
10.3	Dimensionless Parameters	400
10.3.1	<i>Hydraulic Diameter</i>	401
10.3.2	<i>The Knudsen Number</i>	403
10.3.3	<i>The Peclet Number: Transport by Advection or Diffusion?</i>	406
10.3.4	<i>The Reynolds Number: Laminar or Turbulent Flow?</i>	406
10.3.5	<i>Reynolds Number as a Ratio of Time Scales</i>	408
10.3.6	<i>The Bond Number: How Critical is Surface Tension?</i>	409
10.3.7	<i>Capillary Number: Relative Importance of Viscous and Surface Tension Forces</i>	410
10.3.8	<i>Weber Number: Relative effects of Inertia and Surface Tension</i>	410
10.3.9	<i>Prandtl Number: Relative Thickness of Thermal and Velocity Boundary Layers</i>	411
10.4	Applying Nondimensional Parameters to Practical Flow Problems	411
10.4.1	<i>Channel Filled with Water Vapour</i>	411
10.4.2	<i>Channel Filled with a Dilute Electrolyte at 293 K</i>	411
10.5	Characteristic Time Scales	412
10.5.1	<i>Convective Time Scale</i>	412
10.5.2	<i>Diffusion Time Scale</i>	412
10.5.3	<i>Capillary Time Scale</i>	413
10.5.4	<i>Rayleigh Time Scale</i>	413
10.6	Applying Micro- and Nano-Physics to the Design of Microdevices	413
	Problems	415
	References	416
	Appendix A: SI Prefixes	417
	Appendix B: Values of Fundamental Physical Constants	419
	Appendix C: Model Answers for Self-study Problems	421
	Index	435

About the Authors

Ronald Pethig is Professor of Bioelectronics in the School of Engineering, University of Edinburgh, and holds PhD degrees in electrical engineering (Southampton) and physical chemistry (Nottingham) with a DSc awarded for work in biomolecular electronics from the University of Southampton. He has enjoyed a long association with the Marine Biological Laboratory, Woods Hole, being elected a Corporation Member in 1982 and an Adjunct Senior Scientist in 2005. Ron is a Fellow of the Institution of Engineering and Technology, and of the Institute of Physics, and is author of *Dielectric and Electronic Properties of Biological Materials* published by John Wiley & Sons, Ltd. He has received several awards, including being the first recipient in 2001 of the Herman P. Schwan Award, and serves as editor-in-chief of the IET journal *Nanobiotechnology* and editor of the Wiley *Microsystem and Nanotechnology* series.

Stewart Smith is an RCUK Academic Fellow with the School of Engineering, University of Edinburgh. He completed his PhD in 2003 at the University of Edinburgh and since then has worked as a researcher at the Scottish Microelectronics Centre working on research ranging from microelectronic test and measurement to Microsystems for biomedical applications. Prior to his appointment to the RCUK fellowship, he was lead researcher on an industrially funded project that developed a prototype implantable drug delivery device for the treatment of ocular disease. A paper resulting from this work later won the 2008 IET Nanobiotechnology Premium Award. Stewart is a member of the technical committee for the IEEE International Conference on Microelectronic Test Structures. His research interests include the design and fabrication of biomedical Microsystems, test structures for MEMS processes and the electrical characterisation of advanced photolithography.

Foreword

There is no doubt that the continued convergence of engineering, science and medicine in the 21st century will drive new treatments, devices, drugs, diagnostics and therapies for healthcare. Worldwide there is a desperate need for effective and economical medical interventions to care for an ageing population that is growing in number and to help lessen the burden on healthcare systems of the frightening rise in chronic diseases and conditions such as diabetes and cardiovascular disease. The rise in chronic illness is to a great extent being driven by lifestyle changes and as countries become more prosperous and industrialised they see the burden of chronic illness rise. The numbers of people affected are notable. For example, the World Health Organisation (WHO) estimates that 346 million people worldwide have diabetes and that diabetes related deaths are set to double between 2005 and 2030. Type II Diabetes is growing because of sedentary lifestyles and obesity. It does not simply bring problems with blood sugar but complications of uncontrolled glucose levels can lead to cardiovascular disease, eyesight problems, renal problems and wound care problems, creating a complex and growing patient load for healthcare providers. Cardiovascular disease is even more prevalent and claimed the lives of 17.6 million in 2008 and the WHO estimates that this will rise to 26.3 million by 2030.

Thus governments and healthcare providers know that changes must be made to reduce chronic disease where possible, and to deliver care effectively and economically to those who are affected by it.

Medical technology and medical devices have a crucial part to play in helping society care for these populations and interventions based on technology and devices are already widespread and growing. The portable glucose meters which diabetics can use to check their blood sugar levels at any time were developed from biosensor technology and have now become a reliable fixture of diabetes treatment. Current research in the field has produced sub- dermal sensors for glucose that can be left in place for up to a week and the future will bring trans-dermal sensors that will use, or modify, the permeability of the skin to extract glucose for analysis. As another example, there is interest in the use of stem cells to grow new tissue or to repair damaged tissue and many of these types of intervention will require tissue scaffolds to guide and nourish the stem cells, thus materials scientists, engineers and life scientists are exchanging information in multidisciplinary research projects for tissue repair.

In terms of healthcare provision, governments, health services and medical companies are embracing the concept of delivering much of the monitoring and therapy for patients within their own homes rather than in hospitals and clinics. Where telehealth systems have been adopted for monitoring they have been well received by patients who can receive daily

reassurance about their conditions by taking and relaying their own measurements to their clinicians. Developing medical situations that cause concern can trigger earlier interventions and treatment through telehealth monitoring and both hospital admissions and mortality are reduced where telehealth is properly implemented. This growing demand for home monitoring requires not only the advanced telecommunications and wireless systems that engineers have developed but more advances in sensor and imaging technology to allow a wide range of conditions to be monitored. This poses a big challenge requiring more bioelectronics based research and development.

It is clear that our current healthcare problems support the need for the training of more engineers and physicists in bioelectronics for medical device and technology development. It is crucial that good training is provided by experienced practitioners in bioelectronics. The fields of medicine, medical technologies and devices are heavily regulated environments and research projects must be based on cognisance of the human body and medical science as well as technology. It is too easy for well meaning teams of engineers and scientists to create research projects that cannot deliver to the clinical interface because key elements of biology, toxicology and the inflammatory response have not been understood. Teams who will make real advances in this sector will include clinicians and engineers and physicists who have knowledge of medical science and bioelectronics.

Beyond medical devices and healthcare needs, the field of bioelectronics has expanded to produce devices with micro and nano scale features that allow the study of individual cells *in vitro* or *in vivo*. Thus, for example, the response of an individual cardiac or neural cell to a pharmacological agent may be studied via a microfabricated biosensor in contact with the cell. The study of individual and group behaviour of cells provides important information for a range of researchers including biologists, materials scientists and pharmacologists. However, this is again a challenging area for researchers and device development and implementation in this field requires an understanding of engineering principles combined with cell biology. Knowledge of bioelectronics is thus a key need for a student entering this field.

Given the wide range of students that can be drawn from the sectors described above and their different needs Professor Pethig and Dr Smith are to be commended for producing an excellent textbook as an introduction to bioelectronics. It is clear from the content and style of the book that in these authors we have real researchers and teachers who perfectly understand the needs of the new student in the subject. All of the key basic elements of cell biology, biophysics and chemistry are clearly set out to ensure that the student understands the basics before the book moves on to introduce the key technologies in the field for sensors, instrumentation and spectroscopy. The book does not shy away from discussing practical problems in systems and the discussion and teaching on the problems of implanting biosensors will shed light on the disappointing results already obtained by many who are already working in this field.

I will be recommending this excellent textbook to my own students and I congratulate Professor Pethig and Dr Smith on their achievement.

Professor Patricia Connolly
FRSE FIET FRSM CEng

Director, Strathclyde Institute of Medical Devices
University of Strathclyde, Glasgow, Scotland

Preface

This book is written for engineering and physical science students studying courses in bioelectronics, biomedical engineering and micro/nano-engineering, at either an undergraduate or postgraduate level, as well as for researchers entering PhD programmes or working on projects in these subject areas. It aims to teach key topics in biology, chemistry, electrochemistry, biophysics, biosensors and microfluidics of relevance to bioelectronics, and also to place this subject into the context of modern biomedical engineering by examining the state of the art in research and commercial applications. Graduates and researchers wishing to bridge the interface between engineering and the life sciences may also find this book helpful.

The book content is derived from selected background material, lecture notes and tutorials provided to postgraduate students studying for the MSc Degree in Bioelectronics at the University of Edinburgh, and to undergraduates studying for the MEng Degree in Electronics with Bioelectronics. PhD students and postdoctoral researchers from different scientific and engineering backgrounds, working on various aspects of biosensors and lab-on-chip devices, also attend some of the lecture courses. Bioelectronics, as introduced to the students and in this textbook, involves the application of electronic engineering and biophysical principles to biology and medicine. An important aspect of this is the development of a communication interface between electronic components and biological materials such as cells, tissue, and organs. The interdisciplinary nature of the subject means that students and researchers will enter bioelectronics courses from different backgrounds, and to accommodate this some of the chapters cover material delivered to the Bioelectronics MSc students as either background revision notes or introductory material. The first two chapters cover basic chemical, biochemical, biological and thermodynamic concepts that are required for an understanding of the content of subsequent chapters. Condensing subjects that normally merit separate textbooks of their own into two chapters certainly risks the content appearing to be too shallow for readers having good background knowledge in chemistry and biology. We have learnt, however, not to underestimate the extent to which engineering graduates appreciate being reminded of such basic concepts as chemical bonds, pH and Avogadro's number, for example, and their background in biological subjects is often not extensive. Some electronic engineers even find it useful to be reminded of how operational amplifiers function, and we do this in Chapter 7, not only as an aid to them but also as introductory background to those having little background in electronics. To provide access to more basic or more extensive treatments of the book content, most chapters contain suggestions for further reading and other reference material.

Bioelectronics is an exciting and growing field of endeavour that will provide important advances for bioengineering and biomedicine. We hope that this textbook will help students and young researchers to become leading lights for such advances.

Acknowledgements

We thank Professor Andrew Mount of the School of Chemistry at the University of Edinburgh for his feedback after careful reading of Chapters 4 and 5. We also thank Laura Bell, Clarissa Lim, Peter Mitchell, Liz Wingett and the late Nicky Skinner in the various editorial and production offices of John Wiley & Sons, for their constant support, help, and above all, patience.

1

Basic Chemical and Biochemical Concepts

1.1 Chapter Overview

This chapter presents the background concepts of chemistry and thermodynamics of relevance to the subject of bioelectronics, and which are discussed further in most chapters of this book. The level of the material covered in this chapter is probably comparable to that covered by most students in pre-university basic chemistry courses. Graduates in engineering and the physical sciences may need to dig deeply into their recollections of such courses, and may also face new concepts. One objective here is to provide an awareness of some basic concepts of the chemical and energetic functioning of biological systems, of which even a modest understanding will go a long way to mastering the interdisciplinary field of bioelectronics.

After reading this chapter readers will gain a refreshed or new understanding of:

- (i) the formation of chemical bonds and how biological systems make use of the change in *Gibbs free-energy* ΔG of chemical reactions to perform the work required to retain their biological viability;
- (ii) chemical concentrations and activity coefficients;
- (iii) the concepts of nonpolar, polar, ionic, and hydrogen bonds;
- (iv) acids, bases and the biological importance of pH and buffers.

1.2 Energy and Chemical Reactions

1.2.1 Energy

A distinguishing characteristic of a living, rather than a nonliving, system is the ability to perform chemical transformations that produce fluxes of matter and energy. This process describes metabolism. Other characteristics that aid the identification of the living state are

molecular organisation into systems of increasing complexity, and the abilities to self-produce and adapt to changes in environmental factors. The minimal level of organisation capable of exhibiting all these characteristics is the cell. The two principal forms of energy are kinetic and potential, associated with motion and stored energy, respectively. Kinetic energy in a molecular system can be interpreted in terms of the motions of its constituent molecules, which we term as heat. This heat can be determined indirectly by measuring the temperature of the molecular system. For heat to perform work (such as by an engine) it must flow from a region of higher to lower temperature. However, living systems are isothermal – they function at constant temperature and cannot utilise heat flow as a source of energy. Instead, living systems utilise the potential energy stored in the chemical bonds of molecules such as glucose or adenosine triphosphate (ATP). Cells continuously degrade such molecules, and the potential energy released when their chemical bonds are broken is used to perform various kinds of work, including the pumping of substances across membranes to produce chemical concentration gradients that in turn serve as sources of stored potential energy. This process, where chemical bond energy is converted into energy stored in the form of a chemical concentration gradient, is an example of the first law of thermodynamics which states that *energy can neither be created nor destroyed*. Other biological examples of this law include photosynthesis where the energy of sunlight absorbed by green leaves is converted into the chemical bond energy of glucose molecules, and in the conversion of chemical bond energy into mechanical and electrical energy by muscle cells and nerve cells, respectively.

All of the metabolic processes that produce the energy fluxes required for maintaining the living state involve the making and breaking of strong, covalent, chemical bonds between atoms in a molecule.

1.2.2 Covalent Chemical Bonds

Most biological molecules contain only six different atoms, namely carbon, hydrogen, oxygen, nitrogen, phosphorus and sulphur. The locations of these atoms in the Periodic Table of Elements are shown in Table 1.1. The electron shells of the atoms are labelled K, L and M. Each shell is composed of one or more subshells that represent the electronic orbitals about the nucleus of the atom. The first shell, K, has one subshell called the 1s shell and can accommodate a maximum of two electrons. The second shell, L, has two subshells (2s, 2p)

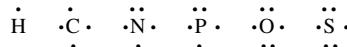
Table 1.1 The locations of hydrogen (H), carbon (C), nitrogen (N), oxygen (O), phosphorus (P) and sulphur (S) in the Periodic Table of Elements

Group								Outer Shell
I H	II	III	IV	V	VI	VII	VIII	K L M
			C	N	O			
				P	S			

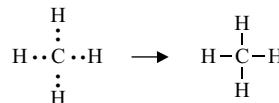
The number of electrons in the outer electron shell of an atom is determined by its group number. Thus carbon (group IV) has four outer electrons and oxygen (group VI) has six outer electrons.

that can accommodate a maximum of eight electrons, with six in the 2p shell. The third shell, M, has three subshells (3s, 3p, 3d) and can accommodate a maximum of 18 electrons, with 10 in the 3d shell.

Electrons in the outer shells have higher average energies than those in the inner shells, and their electron orbitals can extend farther from the nucleus. This contributes to how chemically reactive a particular atom may be in its interaction with other atoms. We can schematically represent the number and arrangement of electrons in the outer electron shells of these atoms as follows [1,2]:

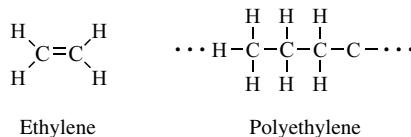


A covalent bond is formed by the sharing of unpaired electrons, one from the outer electron shell of each atom, between the nuclei of two atoms. These shared electrons then enter an electronic orbital that is common to both atoms, acting to reduce the repulsive force between the two positively charged nuclei and to hold them closely together. Thus, the hydrogen atom with one unpaired electron can form only one covalent bond, whilst carbon with four electrons forms four bonds. An example of this is methane (CH_4):

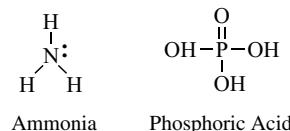


In the methane molecule the carbon atom is covalently bonded to four hydrogens.

In ethylene (C_2H_4) the two carbon atoms are held together by a double bond, and through the polymerisation of ethylene these double bonds are opened up to form the structure of polyethylene:

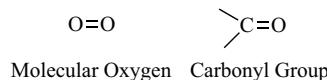


The nitrogen and phosphorus atoms possess five electrons in their outer electronic shells. These atoms can form either three covalent bonds (leaving a lone pair of unbonded electrons) or five covalent bonds. Examples include ammonia (NH_3) and phosphoric acid (H_3PO_4):

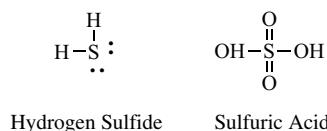


Oxygen contains six electrons in its outer electronic shell (known as the p-shell) and requires just two more electrons to completely fill this shell. It can accomplish this by

forming two covalent bonds with another atom, such as in molecular oxygen (O_2) or in the carbonyl ($C=O$) chemical group:



The sulphur atom can also form two covalent bonds in this manner, as in hydrogen sulfide (H_2S). The outer electronic shell of the oxygen atom has two pairs of electrons that are not involved in covalent bond formation. This, however, does not apply to the sulphur atom, which can form as many as six covalent bonds as in sulphuric acid (H_2SO_4):



1.2.3 Chemical Concentrations

Concentrations of substances dissolved in solutions are often given in terms of weight/volume (e.g. mg/L, or mg/100 mL – a common clinical unit). These units do not depend on knowledge of the molecular structure of the measured substance. For a substance with a known molecular structure, one can define a *mole* of that substance.

Moles and Avogadro's Number: A mole (symbol *mol*) of substance contains as many objects (e.g. atoms, molecules, chemical formula units) as there are atoms in exactly 12 gm of carbon-12. (There are three naturally occurring isotopes of carbon, namely carbon-12, -13 and -14 (^{12}C , ^{13}C , and ^{14}C). ^{12}C is the most abundant and is used as the standard from which atomic masses of all nuclides are measured. The atomic mass of ^{12}C is by definition 12. The radioactive isotope ^{14}C is formed at a constant rate in a chain reaction initiated by cosmic ray protons blasting nuclei in the upper atmosphere to produce neutrons, which in turn bombard nitrogen atoms to form ^{14}C which then combines with oxygen to form carbon dioxide.) The number of atoms in 12 gm of ^{12}C is equivalent to *Avogadro's Number*, of value 6.022×10^{23} , and by convention is given the dimension mol^{-1} . Thus, 1 mol of water molecules contains 6.022×10^{23} water molecules, and 1 mol of *E. coli* comprises 6.022×10^{23} of these microorganisms.

Molar Mass: From the above definition it follows that a mole of any pure substance has mass (in grams) exactly equal to that substance's molecular or atomic mass (in atomic mass units). Molar mass is expressed in units of g/mol. The Dalton (symbol Da) is often employed by biochemists as the unit of molar mass, and is defined as 1 Da = 1 g/mol. Simple chemical compounds have molar masses typically in the range 10–1000 g/mol, and for biopolymers such as proteins and nucleic acids values ranging from 1000 to 5×10^6 are common. The atomic mass values for the most common atoms of biological interest are shown Table 1.2.

Table 1.2 Part of the simplified Periodic Table of Elements to give the mass, in atomic mass units (amu), of some atoms of biological importance

I	II	III	IV	V	VI	VII	VIII
H							
1.008							
		C		N	O	F	
		12.01		14.01	15.99	18.99	
Na	Mg			P	S	Cl	
				30.97	32.07	35.45	
22.99	24.31						
K	Ca						
39.09	40.08						

From Table 1.2 we can deduce that the molar mass of hydrogen gas (H_2) is just over 2 g/mol, and that pure magnesium has a molar mass of 24.31 g/mol. Alternatively, we can say that 1 mol of hydrogen gas is equivalent to 2.016 gm of hydrogen gas, and that 1 mol of pure magnesium is equivalent to 24.31 grams of pure magnesium. Likewise, 58.44 grams of anhydrous (dry) NaCl represents 1 mol of sodium chloride, and 95.21 grams of anhydrous MgCl_2 represents 1 mol of magnesium chloride.

Molar Solution: This is an aqueous solution consisting of one mole of a substance plus enough water to make one litre of solution. Another measure of concentration that we can use is a *Molal Solution*, which is an aqueous solution consisting of one mole of a substance plus 1 kg of water (usually very close to 1 L water). The total volume may thus be more than 1 L. The difference between molar and molal is important for solutions containing a large amount of nonaqueous substance. For example, cream has 20% fat that is homogenised in very small droplets. There will be a 20% difference between the molarity and molality of its salt content, because all the salt will be dissolved in the 80% that is water. Thus, if the *actual* concentration (molality) of the sodium content is 24 mMolal (moles/L water), this could be reported on the product's label as being 20 mMolar (moles/L cream), which for consumers concerned about their daily salt intake can (mistakenly) appear more acceptable.

Concentrations of ions are often given in *Equivalents (or milliequivalents, mEq) per Litre*. The equivalents of an ion are equal to the molarity times the number of charges per molecule. For example, a solvated sodium ion has a single charge, whereas calcium and magnesium ions have two charges. (From Table 1.2 we can see that Ca and Mg are group II atoms, and thus each has two outer valence electrons that are readily donated to two chloride atoms in the formation of calcium chloride and magnesium chloride salts, for example.) Thus, the concept of *Equivalents* is the measure of *Charge concentration*.

Since the molarity of a solution describes the number of individual *particles* dispersed in a given volume of solution, the concept for electrolytes such as sodium chloride is more complicated than for nonelectrolytes because of ionic dissociation. For example, 1 mol of NaCl dissolved in water produces nearly twice as many particles as a mole-equivalent weight of glucose, since the salt dissociates into Na^+ and Cl^- whereas glucose retains its

Table 1.3 Activity coefficient values for some common compounds that dissociate into ions in solution. (Derived from the *CRC Handbook of Chemistry and Physics*, 87th edn, 2006–2007)

Substance	0.01 M	0.05 M	0.1 M	0.5 M	1 M
KCl	0.901	0.816	0.768	0.649	0.604
NaCl	0.903	0.822	0.779	0.681	0.657
MgCl ₂	0.734	0.590	0.535	0.485	0.577
CaCl ₂	0.727	0.577	0.528	0.444	0.495
HCl	0.905	0.832	0.797	0.759	0.811
H ₂ SO ₄	0.542	0.325	0.251	0.146	0.125

single molecule character. Because of electrostatic interaction between the positive ions (cations) and negative ions (anions) there is a statistical probability that at any instant some Na⁺ will be associated with Cl⁻. The electrolyte therefore behaves as if it were not 100% dissociated. Because the electrostatic force between ions decreases with the square of the distance between them, the electrolyte will effectively become more dissociated if the solution is more dilute. Thus, the *activity* (i.e. effective free concentration) of an ion depends on its tendency to dissociate in solution, as well as on its total concentration.

Some activity coefficients (defined as the ratio of the activity divided by the molal concentration) as a function of concentration at 25 °C are given in Table 1.3.

1.2.4 Nonpolar, Polar and Ionic Bonds

In a covalent bond formed between two identical atoms, such as the C–C bond, the bonding electrons are equally shared between the atoms. Such a bond is termed *nonpolar*. Molecules such as Cl₂, H₂ and F₂ are nonpolar, for example. In the description of the concentrations of ions in terms of their equivalents of charge concentration, we have alluded to the concept that different atoms exhibit different tendencies for the sharing of electrons. This tendency can be quantified by their *electronegativity*, using a scale measured from a hypothetical zero to a maximum value of 4.0 (close to that possessed by the most electronegative atom, fluorine). The electronegativity values of some atoms are listed in Table 1.4.

Table 1.4 The electronegativity values for the atoms listed in Table 1.2 based on the Pauling electronegativity scale [4]

I	II	III	IV	V	VI	VII	VIII
H							
2.10							
			C 2.55	N 3.04	O 3.44	F 3.98	
Na	Mg			P	S	Cl	
0.93	1.31			2.19	2.58	3.16	
K	Ca						
0.82	1.00						

We note from Table 1.4 that atoms toward the upper right of the Periodic Table of Elements are more electronegative, and those to the lower left are least electronegative. From this table we can judge that carbon disulphide (CS_2) has almost equal sharing of its electrons when forming its C–S covalent bonds, and so has nonpolar bonds. As a guideline, a maximum difference of $0.4 \sim 0.5$ in electronegativity is often used to define the limit for the formation of a nonpolar bond. For the C–Cl bond there is an unequal sharing of electrons, with electronic charge on average spending more time nearest to the chlorine atom (giving it a slightly negative charge δ^-) and less time near to the carbon atom (making it slightly positively charged δ^+). We say that this bond is a *polar* bond. The H–F bond is particularly polar. Molecules such as NH_3 and H_2O also possess polar bonds, and this lends to them the properties of an electric dipole moment (Figure 3.2, Chapter 3). They will tend to align themselves with an externally applied electric field. Typically, chemical bonds formed between atoms having an electronegativity difference less than 1.6 (but greater than 0.5) are considered to be polar. For larger differences we approach the situation where there is complete transfer of an electron from the least to the most electronegative atom. This type of bond is termed *ionic*. The guideline here is that when the electronegativity difference is greater than 2.0 the bond is considered to be ionic. Common salt (NaCl) is a good example, forming ionic crystals held together by the coulombic forces between the positively charged Na^+ and negatively charged Cl^- atoms. KCl and MgCl_2 are other examples of an ionic solid.

If two highly electronegative atoms are bonded together, the bond between them is usually quite unstable. This occurs in hydrogen peroxide ($\text{H}-\text{O}-\text{O}-\text{H}$), where the strong attractions of bonding electrons towards the two strongly electronegative oxygen atoms make it a highly reactive molecule.

1.2.5 Van der Waals Attractions

Atoms can be defined by a characteristic ‘size’ known as their van der Waals radius. This radius can be determined from investigations of the mechanical properties of gases, from X-ray determinations of the atomic spacing between unbonded atoms in crystals, and from dielectric and optical experiments. The van der Waals radii for some atoms of biological relevance are given in Table 1.5.

If two nonbonding atoms are brought together they initially show a weak bonding interaction, produced by fluctuating electric interactions of the outer electrons of one atom with the positive charge of the other atom’s nucleus, and vice versa. As depicted in Figure 1.1 this can be considered as a fluctuating dipole–dipole interaction between the atoms.

The attraction force between the two atoms increases until their separation distance begins to get less than twice the sum of their van der Waals radii, at which point the two atoms repel

Table 1.5 Van der Waals radii values for some atoms (derived from [5])

Atom	Radius (nm)	Atom	Radius (nm)
Hydrogen	0.12	Oxygen	0.15
Carbon	0.17	Phosphorus	0.18
Chlorine	0.17	Potassium	0.28
Nitrogen	0.16	Sodium	0.23
Magnesium	0.17	Sulphur	0.18

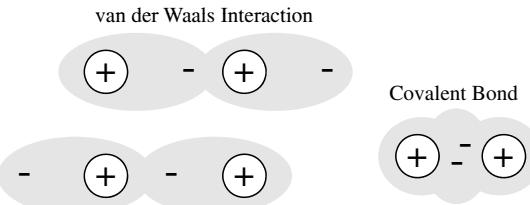


Figure 1.1 Van der Waals attractive interactions arise from dipole-dipole interactions between two nonbonding atoms. This interaction fluctuates in time with how the outer electrons in each atom distribute themselves in their orbitals. In a covalent bond, where a pair of electrons occupies a common (molecular) orbital, the atoms are brought closer together than the sum of their van der Waals radii.

each other very strongly. A mathematically simple model, known as the Lennard-Jones 6–12 potential, can be used to approximate the interaction between a pair of electrically neutral atoms or molecules [3]. The attractive long-range interaction varies as $1/r^6$, where r is the interatomic distance, and the short-range repulsive force is assumed to vary as $1/r^{12}$. The resultant energy is taken as the sum of these two terms and, as shown in Figure 1.2, the equilibrium distance between the two atoms or molecule corresponds to the minimum of the potential energy curve. An insight into the origins of the $1/r^6$ and $1/r^{12}$ dependencies is given in Chapter 3. This model is often used to describe the properties of gases and to model the interatomic interactions in molecular models.

Van der Waals attraction between two atoms is weak, but when many atoms are involved, as occurs for two macromolecular ‘surfaces’ coming into intimate contact, it can become a significant force of attraction. For example, van der Waals interactions make an important contribution to the total force holding together the stable conformations of large molecules such as proteins. When two atoms form a covalent bond, their atomic centers are much closer

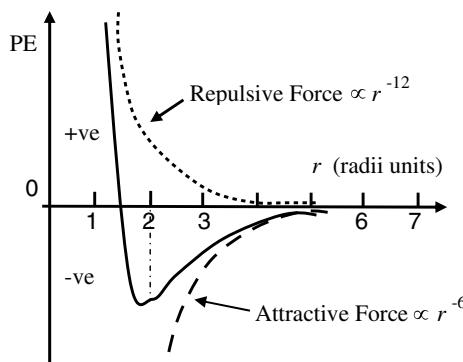
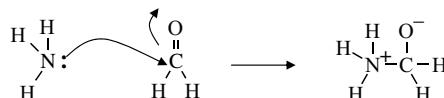


Figure 1.2 The resultant van der Waals force (solid line) can be approximated as the sum of the long-range attractive interaction, assumed to vary as r^{-6} , and the short-range repulsive force which varies as r^{-12} [3]. The equilibrium distance, in radii units r , is located at the minimum of the resulting potential energy (PE) curve.

together than the sum of their van der Waals radii. For example, a single-bonded carbon pair is separated by 0.15 nm, and double-bonded carbons are 0.13 nm apart.

1.2.6 Chemical Reactions

Chemical reactions involve the breaking and forming of covalent bonds, which in turn entails the flow of electrons. When writing down the course of a chemical reaction, curved arrows (known as ‘electron pushing arrows’) are used to symbolise the electron flow. An example is shown below for the reaction between ammonia and formaldehyde, and is a step in the eventual production of hexamine:



In the reaction shown above, one of the lone pair of electrons on the nitrogen atom of the ammonia molecule is donated to form a covalent bond with the carbon atom of formaldehyde. This leaves a positive charge on the nitrogen atom. To maintain the situation that carbon forms four covalent bonds, the double bond is broken down to just a single bond. One of the two electrons so released is used to create the N–C bond, whilst the other is donated to the oxygen atom (leaving it negatively charged).

In accordance with the first law of thermodynamics, the amount of energy released on the formation of a covalent bond is equal to that required for the breaking of it. The strengths of some chemical bonds, in terms of the amount of energy required to break the bond, are given in Table 1.6.

To interpret the concept of chemical bond strength, let us consider the C–C bond. From Table 1.6 we note that one mole of C–C bonds has a bond energy of 346 kJ. The energy required to break a single C–C bond is thus $3.46 \times 10^5 \text{ J}/(6.022 \times 10^{23} \text{ bonds/mol}) = 5.75 \times 10^{-19} \text{ J}$. To break the C–C bond let us assume we need to ‘stretch’ it by 0.2 nm, so that the carbon atoms are spaced 0.35 nm apart, a distance of just over twice the sum of their van der Waals radii. Energy E is expended when a force F moves an object through a distance d ($E=F.d$). The force required to break a C–C bond can thus be estimated as the bond energy $5.75 \times 10^{-19} \text{ J}$ divided by the stretch distance 0.2 nm = 2.87 nN (1 Joule = 1 Newton.meter). To help place this into perspective we can use the concept of tensile strength, which is the force required (per unit area) to break a material. If we think of a C–C

Table 1.6 The strengths of some chemical bonds [6]

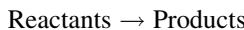
Type of bond	Energy (kJ/mol) ^a	Type of bond	Energy (kJ/mol)
S–S	266	C–N	285
P–O	340	C–C	346
C–O	359	N–H	391
C–H	416	O–H	464
C=C	598	C=O	806

^a 1 kJ/mol = 0.239 kcal/mol.

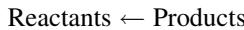
bond as a carbon wire, it will have a cross-sectional area of $\sim 2 \times 10^{-20} \text{ m}^2$. The tensile strength of a C–C bond is thus $\sim 2.87 \text{ nN}/(2 \times 10^{-20} \text{ m}^2) \approx 1.44 \times 10^{11} \text{ N/m}^2 = 144 \text{ GPa}$ (1 Pascal (Pa) = 1 N/m²). By comparison, the tensile strength of materials such as stainless steel and titanium are less than 1 GPa.

Performing a bookkeeping exercise on the number and type of covalent bonds involved in the reaction we have depicted for ammonia and formaldehyde, we see that the C=O bond is replaced with a C–O bond, and that a C–N bond is created. From Table 1.6 we can deduce that if 1 mole of ammonia is reacted with 1 mole of formaldehyde to produce 1 mole of the intermediary product, there is a total bond energy loss of 67 kJ. This loss of bond energy will be given off as heat during the reaction. However, in practice, the amount of heat generated will depend on how much of the original reactants convert into the product. In other words, we need to know the equilibrium state of the reaction.

All chemical reactions reach an equilibrium state. When the reactants are first brought together they react at a rate determined by their initial concentrations. This follows from the *Law of Mass Action*, which states that the rate of a chemical reaction is proportional to the active masses of the reacting substances. We can represent a reaction as:



The concentrations of the reactants decrease as the reaction progresses and their rate of reaction decreases. At the same time, as the amounts of products accumulate they begin to participate in the reverse reaction:



Finally, the stage will be reached where the rates of the forward and reverse reactions become equal and the concentrations of the reactants and products do not change with time. We say that the reaction has reached chemical equilibrium. The equilibrium constant for the reaction is defined as the ratio of the equilibrium concentrations of the products and reactants. For a simple reaction $A + B \leftrightarrow X + Y$

$$K_{eq} = \frac{[X]_{eq} - [Y]_{eq}}{[A]_{eq} - [B]_{eq}}. \quad (1.1)$$

The custom is to use square brackets to designate the equilibrium molar concentrations under standard conditions, namely at 25 °C (298 K) and at a pressure of 1 atmosphere. By convention, when water (H₂O) or hydrogen ions (H⁺) are reactants or products, they are treated in Equation (1.1) as having effective concentrations (activities) of unity (see also Section 1.3). Thus, under standard conditions and starting with 1 M concentrations for all the components of the reaction, when $K_{eq} > 1.0$ the reaction proceeds in the forward direction ($A + B \rightarrow X + Y$), whilst for $K_{eq} < 1.0$ the reaction proceeds in reverse ($A + B \leftarrow X + Y$).

1.2.7 Free-Energy Change ΔG Associated with Chemical Reactions

Because biological systems function at constant temperature and pressure, we can use as a measure of the potential energy released or stored in chemical reactions the concept of Gibbs

free-energy (named after Josiah Willard Gibbs, an early founder of the science of thermodynamics). Gibbs demonstrated that free-energy G is given by the relationship:

$$G = H - TS,$$

where H is the heat energy (also termed enthalpy) of the chemical system, T is the absolute temperature, and S is termed the entropy and provides a measure of the degree of disorder of the system. We are interested in the change of free-energy ΔG that results from a molecule's chemical structure being changed in a chemical reaction. Contributions to the ΔG of a reaction, at constant temperature and pressure, come from the change in heat content between reactants and products, and the change in entropy of the system:

$$\Delta G = \Delta H - T\Delta S.$$

Enthalpy H is released or absorbed in a chemical reaction when bonds are formed or broken. ΔH is thus equal to the overall change in bond energies. We can distinguish between a reaction in which heat is given off (an exothermic reaction) and one in which heat is absorbed (an endothermic reaction). In an exothermic reaction ΔH is negative and the products contain less energy than the original reactants. This is the situation we have deduced for the initial reaction of ammonia with formaldehyde. In an endothermic reaction (heat absorbed) ΔH is positive and the products contain more energy than the reactants.

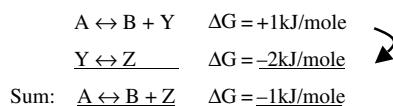
By convention, ΔS is positive when entropy, and thus disorder, increases. The second law of thermodynamics states that the entropy of an isolated system, which is not in equilibrium, will tend to increase over time and to approach a maximum value at equilibrium. The change in Gibbs free-energy ΔG for a spontaneous chemical reaction is always negative, and so a negative value for ΔH (heat given off) and a positive ΔS tend to lead to a spontaneous reaction. Under certain conditions a chemical reaction having a positive ΔG can also occur spontaneously. For this to occur it must be strongly coupled with another reaction having a negative ΔG of larger absolute value than the nonspontaneous reaction. For example, suppose we have the following two reactions:



and



Formation of $B + Y$ in the first reaction will not occur spontaneously, but any Y that is formed is converted spontaneously to Z in the second reaction. This lowers the equilibrium concentration of Y in the first reaction and has the effect of pushing the reaction to produce more $B + Y$. The way this feeds into the sum of the two components of the overall reaction $A \leftrightarrow B + Z$ having a negative ΔG can be represented by:



Energetically unfavourable reactions of the type $A \leftrightarrow B + Y$ are common in cells and are often coupled to reactions having a large negative ΔG . We will see that an important

example of this is the coupling of reactions to the hydrolysis of the molecule adenosine triphosphate (ATP).

The concept of heat being generated or absorbed during a chemical reaction is straightforward – but how do we judge whether there is an increase or decrease in entropy? Determining whether a chemical reaction produces more molecules as products than were originally present in the reactants can provide a simple guide. If the reaction produces an increased number of molecules of the same physical state (i.e. solid, liquid or gas) there will be an increase of entropy, because their constituent atoms will be more randomly dispersed and thus more disordered. Likewise, if a solid reactant converts into either a liquid or gas, its constituent atoms will gain dynamic freedom and thus an increase of entropy.

Chemists define a standard set of conditions to describe the *standard free-energy change* ΔG° of a chemical reaction. The reaction is assumed to occur at 298 K (25 °C) for gases at 1 atmosphere (101.3 kPa) and solute concentrations of 1 M. It can be shown that the relationship between the equilibrium constant K_{eq} and the standard free-energy change ΔG° is given by the expression

$$\Delta G^\circ = -RT \ln K_{\text{eq}} = -2.3RT \log_{10} K_{\text{eq}}, \quad (1.2)$$

where R is the gas constant ($8.314 \text{ J mol}^{-1} \text{ K}^{-1}$) and T is the standard temperature (298 K). Thus, K_{eq} values of 10 and 0.1 litres/mol, for example, correspond to ΔG° values of -5.7 kJ/mol and $+5.7 \text{ kJ/mol}$, respectively. Equation 1.2 can also be written in the form:

$$K_{\text{eq}} = 10^{-\Delta G^\circ / 2.3RT}.$$

When a chemical reaction is not at equilibrium we can interpret the free-energy change for the reaction, ΔG° , as the driving force that tends to drive the reaction towards equilibrium. In this way, we can view ΔG° as an alternative description of the equilibrium constant K_{eq} of a chemical reaction. We should note, however, that this tells us nothing about the rate of a chemical reaction. Many chemical reactions that have a large negative ΔG° value may not proceed at a measurable rate at all! We have seen that chemical reactions involve the breaking and making of covalent bonds. This can only occur if the reacting molecules come close enough together during a collision between them. Some of these collisions may provide the opportunity for the desired sharing of valence electrons between an atom and its ‘target’ reactant, but many other collisions may not. An effective way to increase the rate of a chemical reaction is to add a nonreacting compound, called a catalyst, that facilitates the breaking and making of new bonds between the reactants. Catalysts commonly perform this helpful task by causing a redistribution of the electrons in the reacting atoms, such that the bonds that need to be broken are weakened. This lowers the energy barrier in the reaction ‘pathway’ and increases the reaction rate. After the reaction occurs, the catalyst becomes available to repeat this procedure with another set of reactants. In this way, the catalyst is not consumed in the reaction. The catalysts that take part in biological reactants are large protein molecules called enzymes. As depicted in Figure 1.3 an enzyme catalysed reaction involves the reacting molecule (the substrate) ‘locking’ neatly into an enzyme’s receptor site that is specifically configured for that specific substrate.

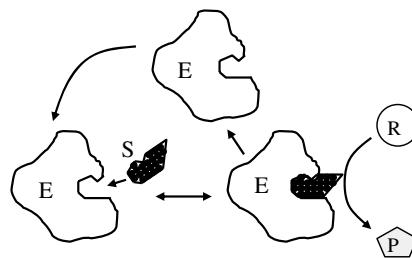
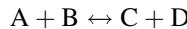


Figure 1.3 In an enzyme catalysed reaction the reactant (substrate S) first binds to a specific site on the enzyme to form an enzyme-substrate complex E-S. The bound substrate then reacts with another reactant to form a product P, leaving the enzyme E available for further reactions.

The overall reaction shown in Figure 1.3 can be written as:



We should also be aware that the standard Gibbs free-energy change ΔG° informs us how far, and in which direction, a reaction will go to reach equilibrium only when the starting concentration of each chemical component is 1 M and at the standard conditions for temperature (298 K) and pressure (101.3 kPa). However, for practically all cases we are only really interested in the actual free-energy change ΔG for reactions where the chemical concentrations are not all the same and not equal to 1 M, and for temperatures other than 298 K. It can be shown for any chemical reaction



that ΔG and ΔG° are related by the expression:

$$\Delta G = \Delta G^\circ + RT \ln \frac{[C][D]}{[A][B]}, \quad (1.3)$$

where the concentrations of the various components and the temperature T are those actually involved. When the reaction reaches the equilibrium state, at which point there is no driving force for the reaction and $\Delta G = 0$, Equation (1.3) reduces to (1.2):

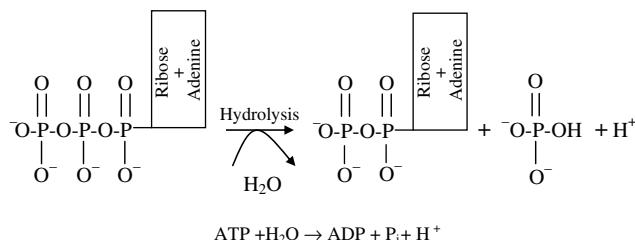
$$0 = \Delta G^\circ + RT \ln \frac{[C]_{eq}[D]_{eq}}{[A]_{eq}[B]_{eq}}$$

or

$$\Delta G^\circ = -RT \ln \frac{[C]_{eq}[D]_{eq}}{[A]_{eq}[B]_{eq}} = -2.3 RT \log_{10} K_{eq}.$$

For almost all types of biological organism, the most important source of free-energy is generated by the breaking of a phosphate bond in adenosine triphosphate (ATP).

ATP consists of adenosine (composed of an adenine group and a ribose sugar) and three phosphate groups (triphosphate), two of which are often referred to as ‘high-energy bonds’. One or both of these bonds can be broken by an enzyme-catalysed hydrolysis reaction. Hydrolysis is a reaction in which a molecule is cleaved into two parts by the addition of a water molecule. The case where one bond is broken, to form adenosine diphosphate (ADP) plus inorganic phosphate (P_i) and a proton (H^+), is depicted below:



The standard Gibbs free-energy change ΔG° for this reaction is -30.5 kJ/mol , which from Equation (1.2) corresponds to an equilibrium constant:

$$K_{eq} = \text{antilog}(-\Delta G^\circ/2.3RT) = \text{antilog}(5.35) = 2.25 \times 105 \text{ (liters/mole)}$$

From Equation (1.1) and noting we can by convention treat the water and hydrogen ion components as having activities equal to 1.0, the equilibrium concentrations of ATP, ADP and P_i satisfy the relationship:

$$K_{eq} = \frac{[\text{ADP}]_{eq} - [P_i]_{eq}}{[\text{ATP}]_{eq}} = 2.25 \times 105 \text{ (liters/mole)}.$$

However, this does not reflect the actual concentrations found in living cells. For example in the cytoplasm of a red blood cell the concentrations of ATP and ADP are typically 2.25 and 0.25 mM, respectively, and inorganic phosphorus P_i has a concentration of 1.65 mM. This gives $[\text{ADP}][P_i]/[\text{ATP}] = 1.83 \times 10^{-4}$, a value some nine orders of magnitude away from the equilibrium situation! The hydrolysis of ATP to ADP is being driven very hard, and this means that the free-energy change available to do useful work is high. We can calculate this free-energy change from Equation (1.3). Assuming a normal physiological temperature of 37°C (310 K) we have:

$$\begin{aligned} \Delta G &= \Delta G^\circ + RT \ln \frac{[\text{ADP}][P_i]}{[\text{ATP}]} \\ &= -30.5 + 2.5 \ln(1.83 \times 10^{-4}) = -30.5 + 2.5(-8.6) = -52 \text{ kJ/mol}. \end{aligned}$$

Thus, the free-energy released by the conversion of ATP into ADP within a cell is significantly larger than the standard free-energy change. Cells use this free-energy to perform many functions, such as: the synthesis of important molecules (DNA, RNA,

proteins, lipids); the pumping across membranes of ions and molecules against their concentration gradients (see Figure 3.11 and discussion in Chapter 3); in the actions of muscle and nerve cells and the firing of neurons, and to maintain a constant body temperature. In plant cells the energy required to convert ADP back into ATP is supplied by trapping the energy of sunlight through photosynthesis. In animal cells and nonphotosynthetic microorganisms it is supplied by the free-energy released in the enzyme-controlled breaking down of the chemical bonds in glucose molecules.

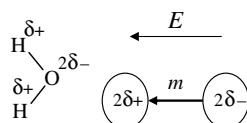
We noted at the beginning of this chapter that a distinguishing characteristic of a living system is its molecular organisation. In a procedure repeated countless times in cell biology laboratories around the world each day a single microorganism, such as *E. coli*, can be isolated from an existing culture and transferred to a new culture dish and placed in an incubator. Apart from this single organism, the dish will typically contain no more than magnesium, sulphate, phosphorus and ammonium ions, water and glucose molecules. Within a day or so, depending on the type of organism, a billion or so new cells will have been generated from the original one. For the case of *E. coli* (a bacterium of cylindrical form about $2\text{ }\mu\text{m}$ long and $1\text{ }\mu\text{m}$ in diameter) its composition includes more than 2 million protein molecules of at least 2000 types, more than 20 million lipid molecules, hundreds of thousands of RNA molecules, and one very large molecule of DNA containing 5 million base pairs. Contained within the volume bounded by its rigid cell wall, formed of interwoven polysaccharides, peptides and lipids, there are also around 300 million ions and metabolites, and some 4×10^{10} water molecules. Thus, the few types of atoms randomly distributed (i.e. of high entropy) in the culture dish have been incorporated into highly organised (low entropy) systems using the free-energy derived from glucose, in an apparently defiant stance against the finality dictated by the second law of thermodynamics! This apparent contradiction of a basic natural law, that all real processes involve the degradation of energy and the dissipation of order, has been clarified by Schrödinger [7], who explained that living organisms extract ‘negentropy’ from their surroundings. What we call the living-state, a state far removed from equilibrium, is maintained by utilising sources of free-energy provided by the continuous flow of energy from the sun to its end point of wasted heat.

1.3 Water and Hydrogen Bonds

We have noted that the *E. coli* bacterium contains around 50 billion water molecules – representing about 90% of its total weight. Water also contributes about 70 ~ 80% of the total weight of animal cells. Most of the biochemical reactions taking part in a cell do so in an aqueous environment. Water is the *mater and matrix* of life. It is the only inorganic liquid that occurs naturally on earth. It is also the only chemical compound that occurs naturally in all three physical states, namely solid, liquid and vapour. According to its molecular size the melting and boiling points of water should be about 100 K lower, and its heat of vaporisation, heat of fusion, and surface tension is higher than that of the comparable hydrides, hydrogen sulphide (H_2S) and ammonia (NH_3) or even than that of most other common liquids. These unique macroscopic properties can only arise from there being strong forces of attraction between the molecules in liquid water – without them water on earth would exist as a gas rather than as a liquid. These strong forces of attraction take the form of hydrogen bonds.

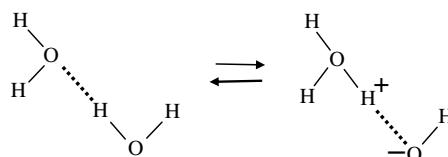
1.3.1 Hydrogen Bonds

A hydrogen atom normally forms just one covalent bond at a time with another atom. However, in Section 1.2.4 we learnt that the water molecule is dipolar. Each hydrogen atom loses a fraction of charge $\delta+$ to the oxygen atom, making the oxygen slightly positively charged by an amount $2\delta+$ to form an electric dipole m , which can align with an externally applied electric field E :



Liquid water is thus a dielectric material possessing a relatively large value of its relative permittivity (dielectric constant) of around 80 at room temperature. The large relative permittivity value for liquid water arises because each water molecule forms transient hydrogen bonds with several other water molecules. The motion of one water molecule dipole moment induced by the electric field is cooperatively coupled to several other water molecules. This cooperative motion enhances the dielectric polarisability of bulk water, because the effective dipole moment per unit volume is larger than that for the situation where each dipole m acted on its own. This makes water a very good solvent for ionic substances (salts) held together by the coulombic force of attraction of the opposite charges of their constituent ions. This force of attraction is inversely proportional to the relative permittivity of the medium in which the ions are embedded. A sodium chloride crystal has a bulk relative permittivity of ~ 6 , and if it is placed in water the force of attraction between the Na^+ and Cl^- ions will be greatly weakened. The crystal will dissociate (dissolve) and the Na^+ and Cl^- ions will be dispersed as free ions in the water. The large electric fields that exist around bare ions such as Na^+ and Cl^- attract water dipoles, which form strongly held hydration sheaths around the ions. Non-ionic substances that possess polar chemical bonds are soluble in water. Sugars, alcohol molecules and many other organic molecules readily dissolve in water because they can form hydrogen bonds with water. Nonionic organic molecules, such as benzene or lipids, that are nonpolar are unable to interact with water in this way, and so are insoluble in water.

Hydrogen bonds form because in liquid water a negatively charged oxygen atom can attract a positively charged hydrogen atom, in a neighbouring water molecule, towards one of its unbonded valence electron pairs. This forms a weak bond of strength around 20 kJ/mol (~ 5 kcal/mol), which is much weaker than the covalent bond between hydrogen and oxygen (~ 460 kJ/mol). The transient nature of hydrogen bonding between molecules of water can be depicted as shown below:



Transient H-bond formations in water

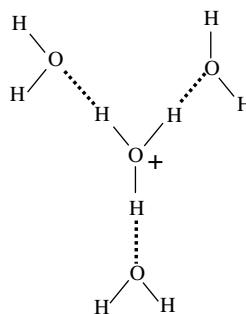
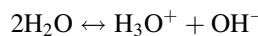


Figure 1.4 Hydronium ion in solution, surrounded by three hydrogen-bonded water molecules.

This diagram shows how the covalent and hydrogen bonds may exchange places from one instant to the next. There is thus a finite (but small) probability that the three hydrogen atoms will associate with one oxygen atom to form a hydronium ion (H_3O^+) leaving another oxygen atom with only one hydrogen to form a hydroxyl ion (OH^-). The dissociation (ionisation) of water can therefore be written as:



The dissociation of water is sometimes written as $\text{H}_2\text{O} \leftrightarrow \text{H}^+ + \text{OH}^-$ to emphasise the production of protons. The positively charged hydrogen atoms (protons) of the hydronium ion attract the electronegative, oxygen, ends of the surrounding water molecules to form a stable hydrated hydronium ion, as shown in Figure 1.4.

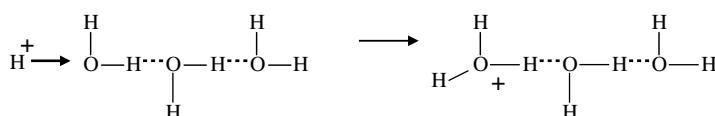
Table 1.7 gives the electrical mobility values for various ions in water, from which we learn that the apparent rate of migration of the H_3O^+ ion in an electrical field is significantly greater than that exhibited by Na^+ and K^+ ions. We have discussed how the electrostatic binding energy of the proton is so large that it has no independent existence in condensed phases such as water. In water it is generally considered to be present as hydronium, H_3O^+ , which gives it an equivalent size between that of the hydrated sodium and potassium ion. The electrical mobility μ of an ion is defined as $\mu = v/E$, where v is the terminal speed acquired under the influence of an electric field E . To a good approximation we can assume that the terminal speed is reached when the accelerating force ($F_a = qE$) is balanced by the Stokes viscous drag force. This viscous force is directly proportional to the size of the ion, and so the electrical mobility of a proton should be of the same magnitude as the Na^+ and K^+ ions.

How do we account for the anomalously high proton mobility? The accepted viewpoint is that a transport process, known as the Grotthuss mechanism, is responsible. This mechanism is named after Theodor Grotthuss [9], who suggested that electrical conduction

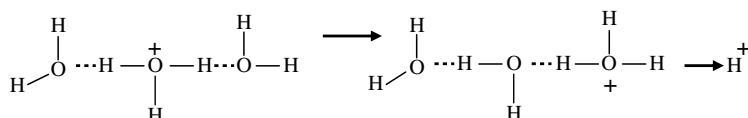
Table 1.7 The electrical mobility of some ions at 25 °C in dilute aqueous solution [8]

Cation	Mobility ($10^{-8} \text{ m}^2/\text{V s}$)	Anion	Mobility ($10^{-8} \text{ m}^2/\text{V s}$)
$\text{H}^+, \text{H}_3\text{O}^+$	36.2	OH^-	20.6
K^+	7.6	Cl^-	7.9
Na^+	5.2	F^-	5.7

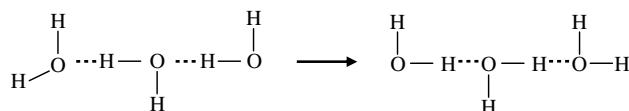
through water resulted from the oxygen atoms simultaneously receiving and passing along a single hydrogen atom. An interesting aspect of this is that at the time of his proposal a water molecule was considered to be OH instead of H₂O, and that an understanding of ions in solution (let alone hydrogen bonds) was at a very primitive level. Nevertheless, his description that throughout the conduction process '*only the water molecules located at the tip of the conducting wires will be decomposed, whereas all those located at intermediate positions will exchange their composing principles reciprocally and alternatively, without changing their nature*' proved to be remarkably insightful. The modern version of the Grotthuss mechanism is depicted in the sequence of events below. The first step involves the injection and binding of a proton into a hydrogen-bond network:



Subsequent steps involve the localised rearrangement of protons and hydrogen bonds, followed by the release of a proton from the hydrogen-bond network:



The final step is the reorientation of the water molecules to re-establish the hydrogen-bonded water structure that existed before the injection of a proton.



A partial analogy for the Grotthuss mechanism is the operation of a bucket brigade, where the buckets move but the people do not. Proton conduction does not involve the diffusion of either the hydronium ions or the protons themselves! A better analogy is electronic conduction along a copper wire. As an electron is injected into the end of a wire at a cathode, another electron is simultaneously ejected from the other end into the anode. This mode of proton transport is of relevance to bioenergetic processes that involve proton diffusion in protein complexes and the pumping of protons across cell membranes, and is the subject of ongoing research [10].

1.4 Acids, Bases and pH

According to the classical Brønsted-Lowry definition of acid and base, H₃O⁺ is *acidic* (it can donate a H⁺ ion) and OH⁻ is *basic* or alkaline (it can accept a H⁺ ion). This definition is named after J.N. Brønsted and T.M. Lowry (Danish and English physical chemists, respectively) who in 1923 independently formulated the protonic definition of acids and bases. Thus, any substance that can form a H⁺ ion is termed an acid, and any substance that

Table 1.8 Acids and Bases

Acids	Bases
Hydrochloric	$\text{HCl} \leftrightarrow \text{H}^+ + \text{Cl}^-$
Carbonic	$\text{H}_2\text{CO}_3 \leftrightarrow \text{H}^+ + \text{HCO}_3^-$
Acetic	$\text{CH}_3\text{COOH} \leftrightarrow \text{CH}_3\text{COO}^- + \text{H}^+$
Water	$\text{H}_2\text{O} \leftrightarrow \text{H}^+ + \text{OH}^-$
	Ammonia $\text{NH}_3 + \text{H}^+ \leftrightarrow \text{NH}_4^+$
	Caustic soda $\text{NaOH} + \text{H}^+ \leftrightarrow \text{Na}^+ + \text{H}_2\text{O}$
	Phosphate $\text{HPO}_4^{2-} + \text{H}^+ \leftrightarrow \text{H}_2\text{PO}_4^-$
	Water $\text{H}_2\text{O} + \text{H}^+ \leftrightarrow \text{H}_3\text{O}^+$

combines with, and thus decreases the concentration of, H^+ ions is termed a base. An acid-base reaction always involves such a so-called conjugate acid-base pair – the proton donor and the proton acceptor (H_3O^+ and OH^- , respectively, for the case of water). Water is said to be *amphoteric* because it acts either as acid or base. Examples of some acids and bases are given in Table 1.8.

The dissociation of water into acid and base is an equilibrium process that follows the *Law of Mass Action*, which states that the rate of a chemical reaction is proportional to the active masses of the reacting substances. For example, the equilibrium constant for the dissociation of water is given by:

$$K_{eq} = \frac{[\text{H}_3\text{O}^+] - [\text{OH}^-]}{[2 \text{H}_2\text{O}]} = \frac{[\text{H}^+] - [\text{OH}^-]}{[\text{H}_2\text{O}]} \quad (1.4)$$

where the brackets denote concentrations in moles per litre. To derive the final right-hand expression of this equation, we have divided the numerator and denominator by $[\text{H}_2\text{O}]$. The concentration of water remains virtually unaltered by its partial dissociation, since (at 25 °C) a litre of pure water contains only 1.0×10^{-7} M of H^+ and an equal number of OH^- ions, whereas the concentration of water in a litre (1000 gm) of pure water is 1000 gm/L divided by the gram molecular weight (18 gm/mol) – namely 55.5 M. Thus, the concentration of water is virtually a constant and it makes no real sense to include it in equation (1.4) as if it were a variable. Equation 1.4 can thus be simplified to:

$$55.5 K_{eq} = [\text{H}^+][\text{OH}^-] \quad (1.5)$$

The constant K_{eq} can be combined with the concentration of water (55.5 M) to give a constant K_w termed the ion product of water. From Equation (1.5) at 25 °C, this is given by:

$$K_w = [\text{H}^+][\text{OH}^-] = 10^{-14}.$$

If $[\text{H}^+]$ for some reason increases, as when an acid substance is dissolved in water, $[\text{OH}^-]$ will decrease so as to keep the product $[\text{H}^+][\text{OH}^-] = 10^{-14}$.

This reaction is the basis for the pH scale, measured as a concentration of H^+ (actually H_3O^+). As shown in the Table 1.9 the pH scale is logarithmic, and typically covers the $[\text{H}^+]$ range from 1 M to 10^{-14} M. The term pH can be thought of as a shorthand term for the *negative log of hydrogen ion concentration*. It is defined as:

$$\text{pH} = \log_{10} \frac{1}{[\text{H}^+]} = -\log_{10} [\text{H}^+]. \quad (1.7)$$

Table 1.9 The pH Scale

	pH	[H ⁺] M	[OH ⁻] M	Example
↑ Increasing Acidic	0	10 ⁰	10 ⁻¹⁴	
	1	10 ⁻¹	10 ⁻¹³	Gastric fluids
	2	10 ⁻²	10 ⁻¹²	Lemon juice
	3	10 ⁻³	10 ⁻¹¹	Vinegar
	4	10 ⁻⁴	10 ⁻¹⁰	Acid soil
	5	10 ⁻⁵	10 ⁻⁹	
	6	10 ⁻⁶	10 ⁻⁸	
Neutral	7	10 ⁻⁷	10 ⁻⁷	Cytoplasm
	8	10 ⁻⁸	10 ⁻⁶	Sea water
	9	10 ⁻⁹	10 ⁻⁵	Alkaline soil
	10	10 ⁻¹⁰	10 ⁻⁴	
↓ Increasing Alkaline	11	10 ⁻¹¹	10 ⁻³	Domestic ammonia
	12	10 ⁻¹²	10 ⁻²	Lime solution
	13	10 ⁻¹³	10 ⁻¹	
	14	10 ⁻¹⁴	10 ⁰	

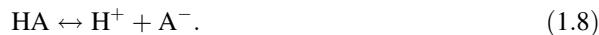
Thus, a 10^{-3} M solution of a strong acid, such as HCl, which dissociates completely in water, has a pH of 3.0, and so on for other concentrations. A solution in which $[\text{H}^+] = [\text{OH}^-] = 10^{-7}$ M has a pH of 7.0, and is said to be neutral – that is, neither acidic nor basic.

1.4.1 The Biological Importance of pH

The hydrogen ion and hydroxyl ion concentrations are important in biological systems because protons are free to move from the H_3O^+ to associate with and thereby neutralise negatively charged groups. Hydroxyl, OH^- , ions are also available to neutralise positively charged groups. This ability to neutralise is especially important in proteins, which as discussed in the next chapter contain both carboxyl ($-\text{COOH}$) and amino ($-\text{NH}_2$) side groups. The acidic carboxyl group can dissociate to give an ionised $-\text{COO}^-$ and H^+ , whereas the basic amino groups can accept a proton to give NH_3^+ . If the acidity of a protein solution is increased (i.e. pH lowered) there will be a relatively large number of H^+ in the solution, so that the probability of a proton neutralising a carboxyl group ($-\text{COO}^- + \text{H}^+ \rightarrow -\text{COOH}$) will be greater than the probability of a hydroxyl ion removing the extra proton from an ionised amino group. Raising the pH of the solution (making it basic) will have the opposite effect. At a certain pH of the solution (the *isoelectric point*), depending on the relative number of carboxyl and amino side-groups, the net charge of a protein molecule will be zero. The electrical properties of proteins (e.g. electrophoretic mobility and electrostatic interactions) are thus determined by the pH of their environment. This is especially evident in the altered properties of an enzyme, because the binding of a substrate molecule to the active site of an enzyme generally involves electrostatic interactions. For example, the pH of the interior (i.e. the cytoplasm) of cells is normally ~ 7.2 . Animal cells contain organelles, such as lysosomes, inside which the pH can be as low as 5. This pH value has been tuned to maximise the functioning of specialised enzymes that degrade proteins, nucleic acids and lipids.

1.4.2 The Henderson-Hasselbalch Equation

The generalised expression for the dissociation of an acid into a proton H⁺ and its anion A⁻ is given by:



Some acids such as HCl dissociate completely and the reaction given above goes to completion:



Other acids such as acetic acid dissociate only partially. In such cases an acid dissociation equilibrium is established with a significant amount of undissociated acid [AH] being present. The acid dissociation constant K_a is derived from the Law of Mass Action:

$$K_a = \frac{[\text{H}^+][\text{A}^-]}{[\text{HA}]} . \quad (1.10)$$

Taking the log of both sides of this equation, we obtain:

$$\log K_a = \log [\text{H}^+] + \log \frac{[\text{A}^-]}{[\text{HA}]} ,$$

or

$$-\log [\text{H}^+] = -\log K_a + \log \frac{[\text{A}^-]}{[\text{HA}]} .$$

Substituting pH for $-\log [\text{H}^+]$ and pK_a for $-\log K_a$ we obtain:

$$pH = pK_a + \log \frac{[\text{A}^-]}{[\text{HA}]} ,$$

or

$$pH = pK_a + \log \frac{\text{proton acceptor}}{\text{proton donor}} . \quad (1.11)$$

This is the Henderson-Hasselbalch equation, which permits the calculation of the degree of dissociation of an acid, given the pH of the solution and the pK_a of the acid. Exercises in the use of this equation are given in the problems section at the end of this chapter. Acids, such as hydrochloric, nitric and sulphuric, which ionise completely when added to water are classified as *strong* acids. The pH of their aqueous solutions is determined simply from their concentration in water using the relationship pH = $-\log[\text{acid}]$ as given by Equation (1.7). Likewise, the pH of a strong base is given by pH = 14 + log[base]. Acids that do not ionise completely are classified as *weak* acids, and as a rough guide have pK_a values greater than around 1.0. It is possible to have a concentrated solution of a weak acid, or a dilute solution of a strong acid. Most acids formed of organic molecules are weak acids. A good example is

ethanoic acid, commonly known as acetic acid, having a pK_a value of 4.76 and which reacts with water to produce ethanoate and hydronium ions:



However, the back reaction is more dominant than the forward one – the ions react to reform the acetic acid and water. Only a small fraction of the acetic acid molecules are converted into ions, and the rest remain undissociated. Just because an acid may be classified as a weak one, this does not imply that it is relatively safer to handle. For example, hydrofluoric acid with a pK_a value of 3.15 is a weak acid that ionises in water according to the standard manner:



However, hydrofluoric acid is extremely toxic. It readily penetrates through the skin and then reacts with calcium and magnesium ions in the blood and tissues, resulting in a serious imbalance of these important physiological ions that can lead to damage, sometimes fatal, to nerves and the heart. Furthermore, when hydrofluoric acid approaches a concentration of 100% its acidity dramatically increases as a result of the following reaction:



From Table 1.4 we see that fluorine has a large (the largest) electronegativity value, which means that the toxic FHF^- anion is stabilised by a very strong H–F hydrogen bond. Bioelectronics engineers are quite likely to come across hydrofluoric acid during the microfabrication of sensors and lab-on-chip devices, because it is used to dissolve the oxide on silicon wafers, and to etch channels in glass substrates, for example. So be warned!

The form of Equation (1.11) can be represented as a titration curve, to show the fraction of undissociated acid as a function of pH. Examples for acetic acid and lactic acid are shown in Figure 1.5.

The pK_a of any acid is equal to the pH at which half of the acid molecules are dissociated and half are not. This can be verified by noting that for $\text{pH} = pK_a$, then from Equation (1.11)

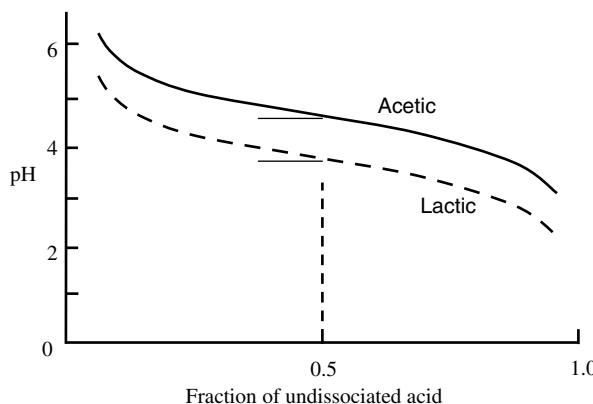
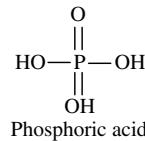


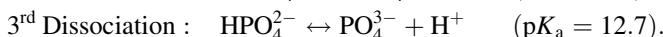
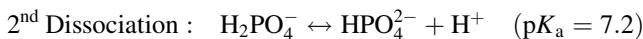
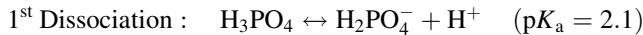
Figure 1.5 Titration curves for acetic acid ($pK_a = 4.76$) and lactic acid ($pK_a = 3.86$).

we have $\log([A^-]/[HA]) = 0$, corresponding to $[A^-] = [HA]$. At one pH unit below its pK_a value, 91% of an acid's molecules will not be dissociated and will be present as $[HA]$. At one pH unit above its pK_a , 91% of an acid will be in the $[A^-]$ form.

Some acids have more than one chemical group that can act as a proton donor. An example of a biologically important acid is phosphoric acid:



The three OH groups in the phosphoric acid molecule do not dissociate simultaneously to act as proton donors, but do so with discrete pK_a constants according to the following three reactions taken in order:



The variation of the pH of an aqueous solution of phosphoric acid (H_3PO_4) at 298 K, as a function of an added acid, is shown in Figure 1.6. The effect of the separate dissociations of its three hydrogen atoms can clearly be seen in this titration curve. On adding another acid to the phosphoric acid solution, the pH of the solution decreases. At a pK_a value, however, the increase of pH is less than if the phosphoric acid had been absent, because protons released into the solution by the additional acid are taken up by the ionised (A^-) form of the phosphoric acid. Likewise, if we have an acid of pH value near to its pK_a value and add to it an alkali (base) such as sodium hydroxide (NaOH), the pH will increase but at a slower rate than

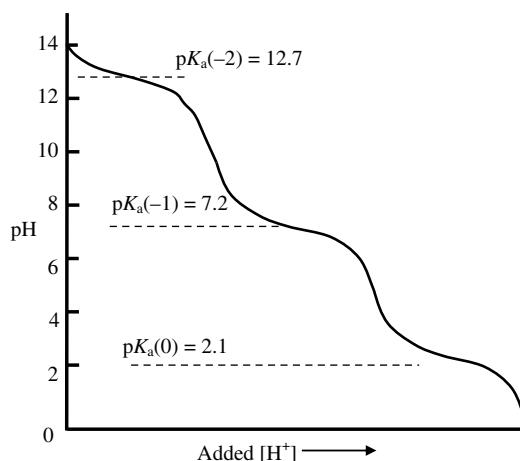
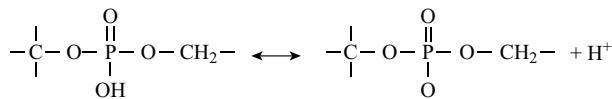


Figure 1.6 The titration curve for phosphoric acid (H_3PO_4) at 298 K, as a function of an added acid. As discussed in the text, the greatest buffering capacity of a conjugate acid-base system is obtained when $\text{pH} = \text{pK}$.

if the phosphoric acid was absent. Hydroxyl ions generated by the added base are neutralised by protons released by the phosphoric acid. This property of an acid or base is known as *buffering*, and as shown in Figures 1.5 and 1.6, the ability to act in this way decreases rapidly for pH values one unit above or below the pK_a .

Phosphate forms part of the structure of DNA and RNA, and is linked to two carbon atoms of adjacent ribose sugars. The OH group can dissociate to give a free proton – hence the reason why DNA and RNA molecules are classified as acids. The dissociation process is depicted below:



The pK_a for the dissociation of the OH proton in DNA and RNA is ~ 3.0 , and so at physiological pH (~ 7.2) each phosphate residue in DNA or RNA is negatively charged.

1.4.3 Buffers

Because of the effect of pH on the ionisation of basic and acidic groups in proteins and other biological molecules, the pH of intra- and extracellular fluids must be held within the narrow limits in which the enzyme systems have evolved. Deviations of one pH unit or more generally disrupt the orderly functioning of a living system. This sensitivity to the acidity of the aqueous intracellular milieu exists in part because reaction rates of different enzyme systems become mismatched and uncoordinated. For example, the pH of human blood is maintained at pH 7.4 by means of natural pH buffers. As discussed above, a buffered system is one that undergoes little change in pH over a certain pH range upon addition of relatively large amounts of an acid or a base.

A buffer must contain an acid $[HA]$ to neutralise added bases, and a base $[A^-]$ to neutralise added acids. The greatest buffering capacity of such a conjugate acid-base pair occurs when the acid $[HA]$ and base $[A^-]$ concentrations are both large and equal. Referring to the Henderson-Hasselbalch equation (1.11) we see that this situation exists when the pH is equal to the pK_a of the acid component of the buffering system (since $\log 1 = 0$). This is demonstrated in Figures 1.3 and 1.4 for that portion of a titration curve along which there is relatively little change in pH on addition of hydroxide.

The most effective buffer systems are combinations of weak acids and their salts. The former dissociate only slightly, thus ensuring a large reservoir of HA , whereas the latter dissociate completely, providing a large reservoir of A^- . Added H^+ therefore combines with A^- to form HA , and added OH^- combines with H^+ to form HOH . As H^+ is thereby removed, it is replaced by dissociation of HA . The most important inorganic buffer systems in the body fluids are the bicarbonates and phosphates. Amino acids, peptides and proteins, because of their weak-acid side groups, form an important class of organic buffers within and outside biological cells. An acid-base indicator is often added to the medium used to culture cells in the laboratory. These indicators are often large soluble organic molecules whose colour in its acid form (when combined with a hydrogen) differs from that when it is in its anionic, conjugate base, form. The indicator molecule is chosen to have a pK_a value close to the pH (e.g. pH 7.4) at which the medium is to be maintained.

Table 1.10 Some common buffering systems and their useful buffering pH ranges

Buffering system	Useful buffering range
Hydrochloric acid and potassium chloride	pH 1.0 – pH 2.2
Hydrochloric acid and glycine	pH 2.2 – pH 3.6
Citric acid and sodium citrate	pH 3.0 – pH 6.2
Acetic acid and sodium acetate	pH 3.7 – pH 5.6
Sodium hydroxide and potassium dihydrogen phosphate	pH 5.8 – pH 8.0
Sodium tetraborate and hydrochloric acid	pH 8.1 – pH 9.2
Sodium hydroxide and glycine	pH 8.6 – pH 10.6
Sodium hydroxide and sodium bicarbonate	pH 9.6 – pH 11.0
Sodium hydroxide and sodium hydrogen phosphate	pH 11.0 – pH 11.9
Sodium hydroxide and potassium chloride	pH 12.0 – pH 13.0

A list of common buffering systems and their corresponding useful buffering pH ranges are given in Table 1.10.

Phosphate buffers are also commonly used, and these are prepared by adding set amounts of 0.1 M disodium hydrogen phosphate to either defined amounts of 0.1 M hydrochloric acid to obtain buffers covering the range pH 7 to pH 9, or to defined amounts of sodium hydroxide to cover the range pH 10 to pH 11.

1.5 Summary of Key Concepts

Biological systems are able to perform chemical transformations that produce fluxes of matter and energy to maintain a highly organised molecular system that is far from being in thermodynamic equilibrium with its environment. Because biological systems function at constant temperature, they cannot utilise the *flow* of heat as a source of energy to achieve this state. Instead, they utilise the *potential energy* stored in chemical covalent bonds. The energy of most value to biological systems is the change in *Gibbs free-energy* ΔG of a chemical reaction. Biological systems utilise the free-energy released from the chemical bonds of molecules such as adenosine triphosphate (ATP) or glucose, and couple these reactions with those that synthesise proteins and other required substances, and to reactions that perform work such as muscle contraction or the pumping of ions against their chemical gradients across membranes.

Covalent chemical bonds are formed when electronic orbitals are shared between two atoms. The extent to which electronic charge is shared between the bonded atoms depends on their relative values of electronegativity. In polar bonds more electronic charge is located near the more electronegative atom which assumes a small negative charge, and the atom of lower electronegativity takes on an equal but slightly positive charge. Water molecules are formed of two polar O–H bonds, and this leads to mutual attractions of its molecules through a network of *hydrogen bonds*. Without these hydrogen bonds water would not exist as a liquid on earth, but as a gas. The polar form of a water molecule, and the hydrogen-bonded network, lend to liquid water a large value of relative permittivity (dielectric constant). This gives liquid water excellent properties as a *solvent* for ionic salts (weakening ionic bonds to the point where the ionic crystal lattice dissociates into free ions) and the

ability to dissolve nonionic substances that possess polar chemical bonds (e.g. sugars and alcohol molecules) by forming hydrogen bonds with them. Organic molecules that are non-polar and nonionic are not soluble in water.

A water molecule can dissociate by breaking one of its H–O bonds into a positively charged hydrogen ion (proton) and a negatively charged hydroxyl ion. Although hydrogen ions rapidly combine with water molecules to form hydronium ions (H_3O^+), it is useful to consider them in their nascent form and to refer to the concentration $[\text{H}^+]$ of hydrogen ions in a solution – expressed as the pH of the solution ($\text{pH} = -\log[\text{H}^+]$). In pure water at 298 K, $[\text{H}^+] = 10^{-7} \text{ M}$, to give a pH value of 7.0. The pH of a biological fluid, such as the cytoplasm inside a cell or fluid within intracellular organelles, is an important parameter controlling the biochemical reactions that maintain the viability of a living system. Biological systems employ different types of organic and inorganic molecules as pH buffers to maintain biological fluids at their optimal pH.

Certain combinations of atoms, such as the hydroxyl ($-\text{OH}$), carboxyl ($-\text{COOH}$), carbonyl ($-\text{C}=\text{O}$), and amino ($-\text{NH}_2$) groups, are common components of biologically important molecules. Each one of these groups has distinct chemical properties associated with their tendency to gain or lose protons. An appreciation of this greatly simplifies an understanding of many aspects of the chemistry of living systems.

Problems

- 1.1. What is meant by equilibrium in a chemical reaction?
- 1.2. Explain how you would prepare the following aqueous solutions:
 - (a) 100 mL of 0.1 M NaCl
 - (b) 10 mL of 50 mM KCl
 - (c) 100 mL of 1 M MgCl_2 (from a bottle of the anhydrous salt)
 - (d) 100 mL of 1 M MgCl_2 (from a bottle of the hydrate $\text{MgCl}_2(\text{H}_2\text{O})_3$).
 - (e) 2 μM KCl solution.
- 1.3. What are the properties of water that make it a good solvent for ions and molecules in a cell?
- 1.4. Describe the formation of hydrogen bonds in water. If water did not contain hydrogen bonds would it exist as a solid, liquid or gas at room temperature?
- 1.5. Hydrochloric acid is an example of a strong acid, and sodium hydroxide is classed as a strong base. They can be assumed to fully dissociate (ionise) when dissolved in water. On this basis, calculate the pH of the following solutions:
 - (a) 1 M HCl
 - (b) 1 μM HCl
 - (c) 10 mM NaOH.
- 1.6. A 1 M solution of acetic acid ($\text{p}K_a = 4.75$) has a pH of 2.4. What percentage of the acetic acid molecules is dissociated? Is acetic acid a strong or weak acid? (Domestic vinegar comprises ~ 1 M acetic acid.)
- 1.7. Lactic acid ($\text{C}_3\text{H}_6\text{O}_3$) in solution can lose a proton from its acidic group to produce the lactate ion $\text{CH}_3\text{CH}(\text{OH})\text{COO}^-$. Calculate the $\text{p}K_a$ of lactic acid, given that when the concentration of lactic acid is 0.01 M and the concentration of lactate is 0.087 M, the pH is 4.8.

1.8. Using the Henderson-Hasselbalch equation, and by defining a base concentration factor α as

$$\alpha = \frac{[A^-]}{[A^-] + [HA]}$$

show that the greatest buffering capacity of a conjugate acid-base pair occurs for $[A^-] = [HA]$. (Hint: Find the value for α where $d(\text{pH})/d\alpha$ has its lowest value.)

References

- [1] Lewis, G.N. (1916) The atom and the molecule. *Journal of the American Chemical Society*, **38**, 762–785.
- [2] Pauling, L. (1960) Chapter 2, in *The Nature of the Chemical Bond*, 3rd edn, Cornell University Press.
- [3] Lennard-Jones, J.E. (1931) Cohesion. *Proceedings of the Physical Society*, **43**, 461–482.
- [4] Pauling, L. (1932) The nature of the chemical bond, IV: the energy of single bonds and the relative electronegativity of atoms. *Journal of the American Chemical Society*, **54**, 3570–3582.
- [5] Bondi, A. (1964) Van der Waals volumes and radii. *Journal of Physical Chemistry*, **68**, 441–451.
- [6] Housecroft, C.E. and Constable, E.C. (2010) Chapter 4, in *Chemistry*, 4th edn, Prentice Hall.
- [7] Schrödinger, E. (1944) *What is Life?* Cambridge University Press.
- [8] Atkins, P.W. and De Paula, J. (2002) Chapter 27, in *Physical Chemistry*, 7th edn, W.H. Freeman.
- [9] de Grotthuss, C.J.T. (1806) Mémoire sur la décomposition de l'eau et des corps qu'elle tient en dissolution à l'aide de l'électricité galvanique. *Annales de Chimie (Paris)*, **58**, 54–74. An English translation, by Régis Pomès, of this paper is published in: *Biochimica et Biophysica Acta* 1757: 871–875, 2006.
- [10] Wright, C.A. (2006) Chance and design: Proton transfer in water, channels and bioenergetic proteins. *Biochimica et Biophysica Acta*, **1757**, 886–912.

Further Readings

- Alberts, B., Johnson, A., Lewis, J. et al. (2007) Chapter 2, in *Molecular Biology of the Cell*, 5th edn, Garland Science.
- Eisenberg, D. and Kauzmann, W. (1969) *The Structure and Properties of Water*, Oxford University Press.
- Harold, F.M. (2001) *The Way of the Cell*, Oxford University Press.
- Nelson, D.L. and Cox, M.M. (2009) Chapters 1, 2 & 13, in *Lehninger Principles of Biochemistry*, 5th edn, W.H. Freeman.
- Schrödinger, E. (1944) *What is Life?* Cambridge University Press, (reprinted 1962, and as a Canto edition, 2003).

2

Cells and their Basic Building Blocks

2.1 Chapter Overview

The chemical composition of a typical bacterium and animal (mammalian) cell is shown in Table 2.1.

Leaving aside the water content of a cell, macromolecules such as proteins, nucleic acids (DNA, RNA), and polysaccharides make up a large percentage of a cell's mass. The building blocks for these macromolecules are small organic molecules, namely fatty acids, sugars, amino acids and nucleotides. This chapter describes the chemical structures and functions of these molecular building blocks, and the biological importance of the macromolecules and macrostructures they combine to form. A summary description is then given of how these macromolecules and microstructures interact and function in different types of cell.

After reading this chapter a basic understanding should be obtained of:

- (i) basic biochemical and biophysical properties of fatty acids, carbohydrates, amino acids and nucleotides;
- (ii) basic biochemical and biophysical properties of biomembranes, proteins, DNA and RNA, and their biological roles;
- (iii) the genetic code and the central dogma of molecular biology (DNA makes RNA makes protein);
- (iv) the differences between prokaryotic and eukaryotic cells;
- (v) mammalian blood cells and the immune system;
- (vi) bacteria, viruses and prions;
- (vii) cell life-cycle, cell culture, tissue engineering and cell-cell communication.

2.2 Lipids and Biomembranes

Cells of higher organisms are separated, but not isolated, from their surroundings by their cytoplasmic membrane, which also serves to act as anchors for proteins that transport or

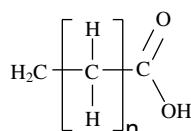
Table 2.1 Approximate chemical composition of a typical bacterium and mammalian cell. (Adapted from Alberts *et al.* [1])

Chemical component	Percentage of total cell weight	
	Bacterium	Animal cell
Water	70	70
Inorganic ions (e.g. Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Cl^{2-})	1	1
Amino acids, nucleotides, and other small molecules	1	1
Metabolites (e.g. glucose, fatty acids)	2	2
Macromolecules (proteins, nucleic acids, polysaccharides)	24	21
Lipids	2	5

pump specific chemicals into or out of a cell. Membranes also define the boundaries of intracellular organelles and the nucleus in eukaryotic cells. The main structural components of biological membranes are lipids, which exist as derivatives of fatty acids. The term ‘lipid’ covers a wide range of molecules, including oils, waxes, sterols, certain (fat-soluble) vitamins and fats. The one property they all share in common is that they are hydrophobic. When placed in water individual lipid molecules will adopt a configuration that leads to minimum contact with water molecules, and will cluster into a group with other lipid molecules. This is exemplified by the formation of oil droplets in water, and how lipids in an aqueous medium segregate into a separate nonaqueous phase.

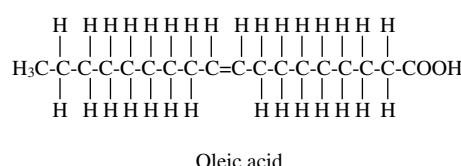
2.2.1 Fatty Acids

Fatty acid molecules contain a hydrocarbon chain, commonly consisting of 16 or 18 carbon atoms. An acidic carboxyl group (COOH) is attached to one end of this chain. Stearic acid $\text{CH}_3(\text{CH}_2)_{16}\text{COOH}$ and arachidic acid $\text{CH}_3(\text{CH}_2)_{18}\text{COOH}$ are examples, whose general chemical structure is shown below:



Chemical structure of saturated fatty acids

Stearic ($n=16$) and arachidic ($n=18$) acid are examples of fatty acids with no double ($\text{C}=\text{C}$) bonds in their hydrocarbon chain, and are termed as being *saturated*. If the hydrocarbon chain contains one or more double $\text{C}=\text{C}$ bonds the fatty acid is termed *unsaturated* – an example of which is oleic acid $\text{CH}_3(\text{CH}_2)_7\text{CH}(\text{CH}_2)_7\text{COOH}$:



The two hydrogen atoms attached to the carbons in the $\text{C}=\text{C}$ double bond of oleic acid lie on the same side of the bond, and this configuration is known as the *cis* form. This *cis*

configuration introduces a bend in the hydrocarbon chain. The other possible configuration, known as the *trans* form in which the two hydrogen atoms are situated on opposite sides of the C=C double bond, does not result in a bent hydrocarbon chain. The reason why butter and lard are solid at room temperature is because they are composed of saturated fatty acids whose straight hydrocarbon chains can pack closely together. Easily spreadable butter substitutes (e.g. margarine) contain unsaturated fatty acids that are unable to pack closely together because of the ‘kinks’ in their hydrocarbon chains, and have a softer form than butter at room temperature. Plant oils contain polyunsaturated fatty acids (with multiple C=C double bonds) and are liquid at room temperature.

A fatty acid molecule thus has two chemically distinct parts – a long hydrophobic chain that is not very reactive chemically, and a carboxyl group (COOH) which when ionised as COO^- is chemically active and hydrophilic. Molecules such as these, which contain both hydrophobic and hydrophilic regions, are termed *amphipathic*. Fatty acids by themselves will not form a membrane that is capable of acting as a boundary between an aqueous medium and the aqueous cytoplasm of a cell. In aqueous media fatty acids will tend to form clusters, with the hydrocarbon chains packed together inside and the carboxylic acid groups directed outwards towards the surrounding water molecules. To form biomembranes fatty acids need to be converted into a structure that readily form sheets of lipid bilayers. The most common ones adopted in nature are phospholipids composed of two fatty acid side chains attached to a negatively charged (and hence hydrophilic) phosphate group via a glycerol molecule. The two fatty acid ‘tails’ may both be saturated, unsaturated, or adopt one of each form. As shown in Figure 2.1, in some phospholipids the ‘head’ of the molecule may be increased in size with the addition of an amine which can ionise to the hydrophilic form NH_3^+ .

Phospholipids can spontaneously form sheets of bilayers, two molecules thick, in an aqueous environment. As depicted in Figure 2.2, the hydrocarbon tails keep away from the water by aligning themselves in the middle of the bilayer structure. The close packing of the hydrocarbon tails is stabilised by van der Waals interactions, and the fluidity of the bilayer interior is influenced by the number of C=C double bonds in the hydrocarbon structures of the tails. The polar head groups are stabilised through hydrogen bonding to water molecules, as well as by electrostatic interactions between the phosphate and amine groups. As shown

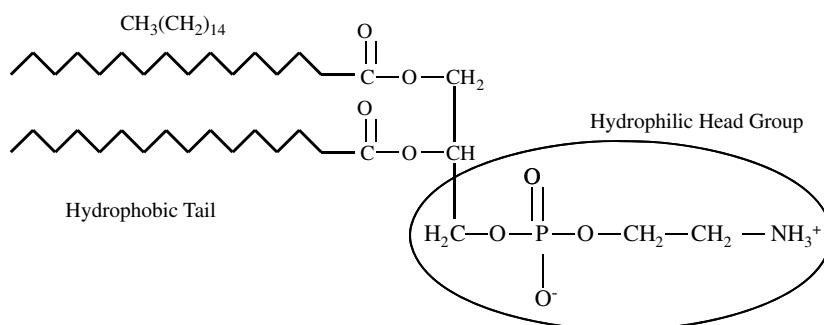


Figure 2.1 The chemical structure of a typical phospholipid (in this case phosphatidylethanolamine) to show its hydrophobic tail and hydrophilic head group.

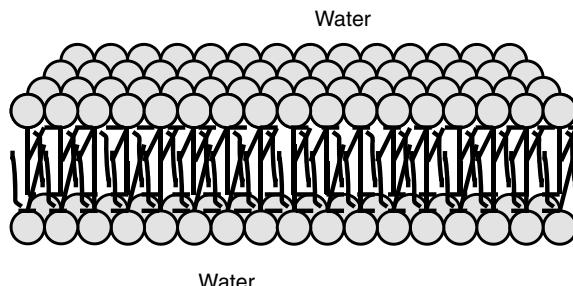


Figure 2.2 Schematic representation of a phospholipid bilayer. The small spheres represent the hydrophilic heads groups, and the lines are the hydrophobic hydrocarbon tails of individual phospholipid molecules.

schematically in Figure 2.3 for a fat cell, the outer membrane of a cell is formed by a spherical lipid bilayer structure that encloses the cytoplasm and internal cell structures.

Apart from their importance as precursors to phospholipids, fatty acids are used as a source of energy by tissues. Fat cells, known as adipocytes, contain one large droplet of lipid (see Figure 2.3). When triggered by hormones such as adrenaline these cells release fatty acids into their surrounding environment (normally blood), which are then broken down into smaller molecules identical to those derived from the breakdown of glucose.

2.3 Carbohydrates and Sugars

Carbohydrates are composed of carbon atoms and the atoms that form water molecules, namely hydrogen and oxygen. Simple carbohydrates, called mono-saccharides, have the chemical structure $(\text{CH}_2\text{O})_n$ and are often referred to as simple sugars. The number ‘n’ of carbon atoms ranges from 3 to 7 and the corresponding sugar molecules are called trioses, tetroses, pentoses, hexoses and heptoses. We will learn later in this chapter that two pentose sugars (ribose and deoxyribose) are essential components of DNA and RNA. An important

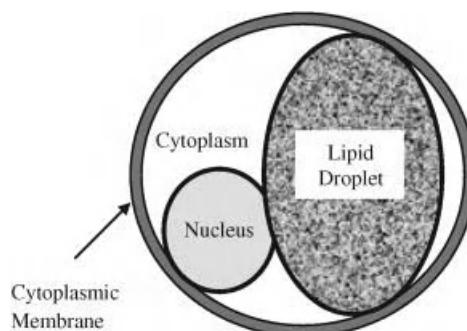


Figure 2.3 Schematic representation of a fat cell (adipocyte).

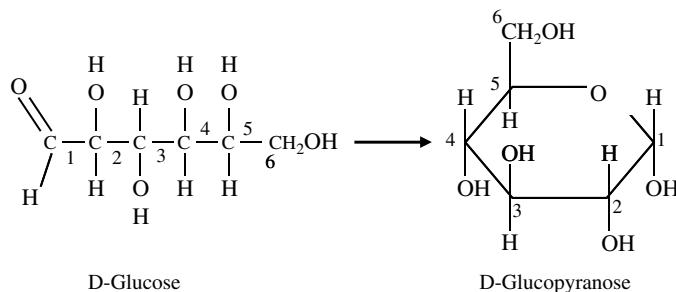


Figure 2.4 The linear and ring form of D-Glucose.

hexose is glucose ($C_6H_{12}O_6$) because when it is broken down in cells of higher organisms it releases free energy. As shown in Figure 2.4 the linear structure of glucose can form a ring structure arising from the reaction of the aldehyde at the 1 carbon with the hydroxyl group on the 5 carbon, to form glucopyranose. A less common ring structure (glucofuranose) is formed by the reaction of the 1 carbon aldehyde with the hydroxyl on the 4 carbon. The chemical formula of a monosaccharide does not therefore fully describe the molecule. For example, a different sugar is formed if the hydrogen and hydroxyl groups attached to the 2 carbon of the D-glucose molecule switch places. This sugar (mannose) cannot be converted to glucose without breaking and making the relevant covalent bonds. Each of the sugars can also exist in either of two forms that are mirror images of each other, called the D-form and the L-form. The D- or L-form of a molecule signifies the direction, *dextro* (right) or *levro* (left), in which the plane of polarisation of light rotates when passing through a solution of the molecules. The most common form of sugars found in, and metabolised by, biological systems is the D-form.

Two monosaccharides can be linked by a covalent bond to form a disaccharide, a third can be added to form a trisaccharide, and so on to form a polymeric chain known as an *oligosaccharide*. If more than around 50 sugar subunits (mers) are joined together the resultant structure is called a polysaccharide. Each covalent bond, known as a glycosidic bond, is formed between an $-OH$ group on one sugar and an $-OH$ group on another by a *condensation reaction* that results in a loss of a molecule of water. This reaction is depicted in Figure 2.5 for the formation of one molecule of the disaccharide known as sucrose composed of glucose and a fructose molecule. Polysaccharides, containing hundreds or even thousands of sugar subunits, can act as energy stores in cells – an example being glycogen stored in liver cells. They also form part of the structures of connective tissues, are part of the composition of mucus and slime, and serve to lubricate bone joints. A polysaccharide of glucose, namely cellulose, is the main structural component of plant cell walls and as such is probably the most abundant organic molecule on Earth.

Proteins and lipids, known as glycoproteins and glycolipids, have oligosaccharide chain attachments and are important building blocks of cell membranes. The attached oligosaccharide chains increase the water solubility of the proteins and lipids and act to orientate them at the interface between the membrane and the surrounding aqueous medium. Oligosaccharides are also responsible for the grouping of human blood cells according to the

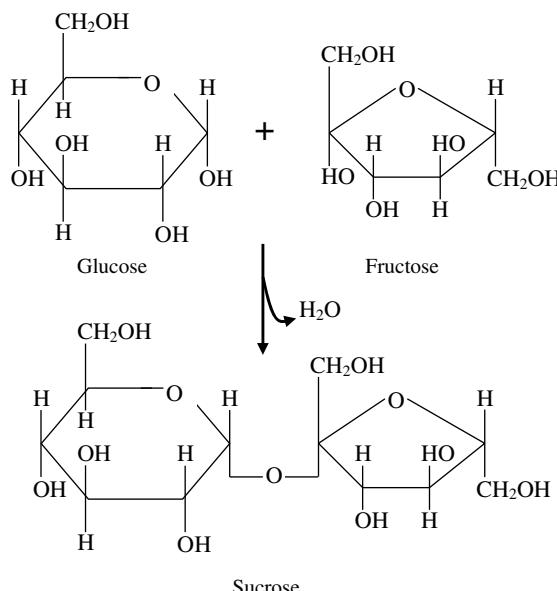


Figure 2.5 Formation of sucrose from the condensation reaction of glucose with fructose.

ABO system, and for the rules that dictate allowed transfusion of blood between donors and recipients.

For example, a group O person can only receive blood from group O individuals, but can donate blood to persons of any blood group. On the other hand, a person with the AB group can receive blood from any donor, but can only be a donor to another AB individual. Such rules are determined by the type of oligosaccharide linked to a protein or lipid in the outer membrane of a person's red blood cells, and how this molecular structure can perform as an antigen. If blood of one group is injected into someone having a different blood group, white blood cells (known as B cells) in the recipient's blood may 'recognise' the foreign antigen and then generate antibodies that selectively attach themselves to it. These antibodies, which may already have existed in the recipient's blood serum, are large protein complexes known as immunoglobulins. Very large numbers of different antibodies are normally created in a person's blood from an early age as a result of exposure to antigens present on bacteria or plants, for example. The attachment of antibodies to their specific antigens in turn triggers an immune response that involves specialised cells, known as phagocytes, locating and then ingesting the antibody-coated foreign blood cells. The reason why blood group O individuals can be a donor to all ABO groups is because they do not have either the A- or B-antigen on their red blood cells. However, their blood serum naturally contains both anti-A and anti-B antibodies against the A and B blood group antigens, which does not permit their receiving blood from A, B or AB donors.

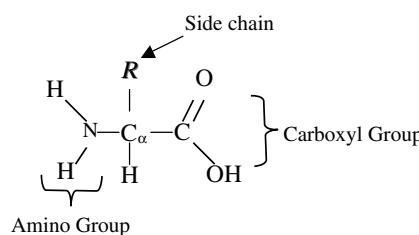
2.4 Amino Acids, Polypeptides and Proteins

Proteins are the working molecules of cells. They provide a cell with structural rigidity; pumps to drive ions and metabolites across membranes; catalysts for a vast range of

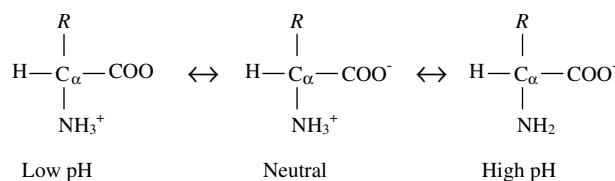
biochemical reactions including the functioning of genes; and ‘motors’ to provide motility of objects within cells and of some whole cells. Our present understanding of how proteins are able to perform this amazing range of functions has taken more than 200 years to achieve. Although the importance to biological processes of substances we now call proteins was appreciated before 1800, the word ‘protein’ was first used in the scientific literature by the Dutch chemist G.J. Mulder in 1838 – a term suggested to him by J.J. Berzelius [1,2]. Using refinements of the chemical analysis procedures of Lavoisier, Liebig, Gay-Lussac, Dalton and others [3], Mulder concluded that proteins were composed of a very large number of carbon, hydrogen, nitrogen and oxygen atoms. Based on the principle that no molecule can contain a nonintegral number of atoms, he obtained the chemical formula $C_{400}H_{620}N_{100}O_{120}P_1S_1$ for egg albumin, and exactly the same formula for serum albumin, but with two sulphur atoms instead of one. It is now known that phosphate groups can be linked to the protein structure, so that phosphorus, unlike sulphur, is not considered to be an intrinsic atomic component of a protein molecule. Proteins are thus macromolecules of large molecular weight. At that time it was also found that when proteins are subjected to the hydrolytic action of boiling acid, they decompose into relatively simple crystalline substances – now known as amino acids. Amino acids are the monomers that make up the polymeric chains of proteins. By 1903, 18 of the common 20 amino acids had been isolated and characterised, with the last two, methionine and threonine, being found in 1922 and 1936, respectively [4,5].

2.4.1 Amino Acids and Peptide Bonds

The 20 common amino acids, apart from proline, contain an amino group ($-NH_2$) and an acid carboxyl group ($-COOH$). Proline possesses an *imino* group ($-NH-$) instead of an amino group. In accordance with the description of acids and bases in Chapter 1, at normal physiological pH the acidic carboxyl group is ionised as $-COO^-$, and the basic groups are ionised as $-NH_3^+$ (or $=NH_2^+$ for the case of proline). The basic chemical structure of an amino acid is shown below:

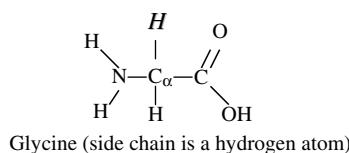


The predominant forms of an amino acid across the pH range are as follows:



In neutral solutions ($\text{pH} \approx 7$) amino acids exist predominantly in the dipolar, doubly ionised form, called a zwitterion.

The central carbon atom, called the alpha-carbon (C_α), is bonded to an amino (or imino) group, a carboxyl group and a hydrogen atom. A variable chemical group R , termed the *side chain*, is also bonded to the C_α carbon, and is what gives an amino acid its special characteristic. Amino acids with a side chain bonded to the C_α carbon are referred to as *alpha amino acids*, and are by far the most common form found in nature. Glycine has the simplest side chain, namely a single hydrogen atom. This lends to glycine, with its two hydrogen atoms about the C_α carbon atom, the property of symmetry. The remaining amino acids do not possess such symmetry – and so have two mirror-image (stereoisomeric) structures, called the D and L forms, as described for sugars in Chapter 1. Only the L forms of amino acids are found in protein molecules, but D-amino acids form part of bacterial cell walls and occur in some antibiotics.



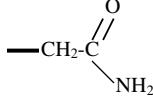
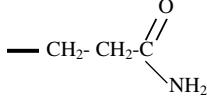
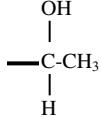
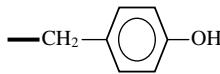
The chemical structures of the 20 common amino acids are given in Tables 2.2–2.4, and are classified according to whether their side chain is hydrophilic or hydrophobic.

Five of the side chains in the 20 common amino acids are ionisable, and their pK values are given in Table 2.4. The main factors to consider when determining whether a side chain

Table 2.2 Amino acids with hydrophobic (nonpolar) side chains R

Amino acid	Side chain structure R	Amino Acid	Side Chain Structure R
Alanine (Ala)	—CH ₃	Isoleucine (Ile)	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{C}-\text{CH}_2-\text{CH}_3 \\ \\ \text{H} \end{array}$
Leucine (Leu)	$\begin{array}{c} \text{CH}_3 \\ \diagdown \\ -\text{CH}_2-\text{CH} \\ \diagup \\ \text{CH}_3 \end{array}$	Methionine (Met)	—CH ₂ —CH ₂ —S—CH ₃
Phenylalanine (Phe)	—CH ₂ — $\begin{array}{c} \text{CH}_3 \\ \\ \text{C}_6\text{H}_4 \end{array}$	Proline (Pro)	$\begin{array}{c} \text{H}_2 \quad \text{H}_2 \\ \diagdown \quad \diagup \\ \text{C}_\alpha \quad \text{C} \\ \diagup \quad \diagdown \\ \text{N} \quad \text{C} \\ \diagdown \quad \diagup \\ \text{H}_2 \quad \text{H}_2 \end{array}$
Tryptophan (Trp)	$\begin{array}{c} \text{H}_2 \quad \text{H} \\ \diagdown \quad \diagup \\ \text{C} \quad \text{N} \\ \diagup \quad \diagdown \\ \text{H}_2 \quad \text{C}_6\text{H}_4 \end{array}$	Valine (Val)	$\begin{array}{c} \text{CH}_3 \\ \diagdown \\ -\text{CH} \\ \diagup \\ \text{CH}_3 \end{array}$

Table 2.3 Amino acids with hydrophilic (uncharged, polar) side chains R

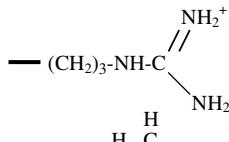
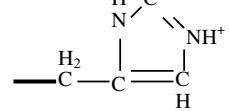
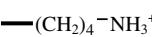
Amino acid	Side chain structure R	Amino acid	Side chain structure R
Asparagine (Asn)		Cysteine (Cys)	—CH ₂ -SH
Glutamine (Gln)		Glycine (Gly)	—H
Threonine (Thr)		Serine (Ser)	—CH ₂ OH
		Tyrosine (Tyr)	

is hydrophobic or hydrophilic are:

- carbon and nonpolar groups do not readily hydrogen-bond to water, and are thus hydrophobic;
- oxygen and nitrogen can hydrogen-bond to water, and are thus hydrophilic;
- ionisable groups (e.g. --COO^- , --NH_3^+ or $=\text{NH}_2^+$) are hydrophilic;
- polar groups are hydrophilic.

A covalent chemical bond can be formed between two amino acids to form a dipeptide, involving the amino group of one amino acid and the carboxyl group of the other. This bond,

Table 2.4 Amino acids with hydrophilic (charged) side chains R

Amino acid	Side chain structure R	Amino acid	Side chain structure R
	Positively charged (pH < pK)		Negatively charged (pH > pK)
Arginine (Arg) pK ~ 12		Aspartic acid (Asp) pK ~ 4.7	—CH ₂ —COO ⁻
Histidine (His) pK ~ 6.5		Glutamic acid (Glu) pK ~ 4.7	—CH ₂ —CH ₂ —COO ⁻
Lysine (Lys) pK ~ 10.2			

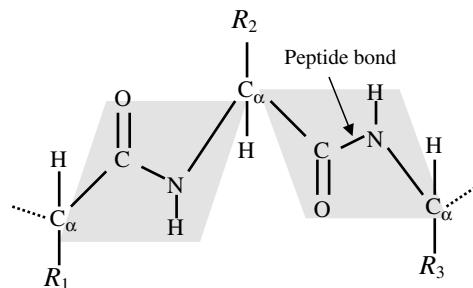
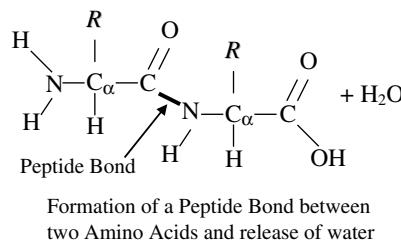
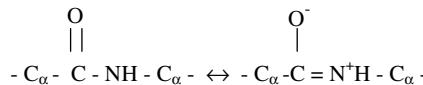


Figure 2.6 The peptide bond has a ‘resonant’ partial double-bonding of the carbon and nitrogen atoms. This results in the peptide group of atoms to lie in the same plane [7].

known as a peptide bond, results from the elimination of a molecule of water in a so-called condensation reaction:



The carboxyl ($\text{C}=\text{O}$) and nitrogen atom forming the peptide bond between the two amino acid *residues* exhibit a resonating partial double-bond character, as depicted below:



Because of this resonance bonding, the six atoms of the peptide group (the two alpha-carbons, the carbon, oxygen and nitrogen plus hydrogen) all lie in the same plane, as shown in Figure 2.6. Independent rotation of two planar peptide groups about their connecting alpha-carbon is possible. As shown in Figure 2.7, the relative conformation of a pair of planar peptide groups can be defined by two dihedral angles ϕ and ψ , where ϕ is the angle of rotation about the $\text{C}_\alpha-\text{C}$ bond, and ψ is the angle of rotation about the $\text{N}-\text{C}_\alpha$ bond.

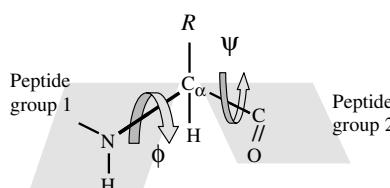


Figure 2.7 The relative conformation of two adjacent planar peptide groups in a dipeptide is defined by the angles of rotation ϕ and ψ about the connecting C_α atom.

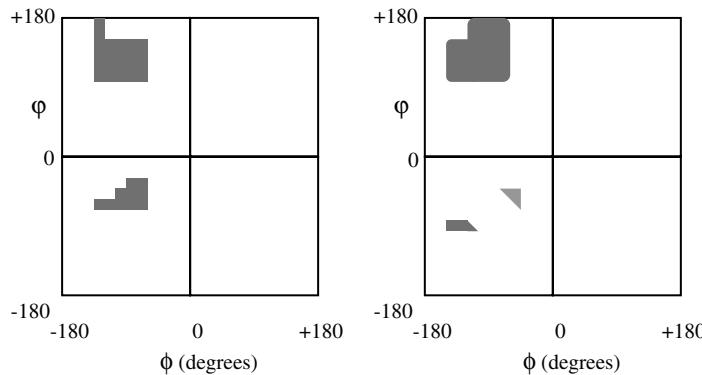


Figure 2.8 Ramachandran plots to show (left) the permissible conformations of valine and isoleucine side groups and (right) for a perfect helix of poly-L-alanine, in terms of the angles ϕ and ψ shown in Figure 2.7. (Adapted from Ramachandran and Sasickharan [7, p. 337].)

Although the angles ϕ and ψ shown in Figure 2.7 can theoretically both assume all the values from 0 to 360° , physically realisable conformations are in fact limited by restrictions on the allowed van der Waals contact distances between different atoms in the dipeptide, and especially for those atoms that are located in the side-chain R . Although $\phi=\psi=0^\circ$ is used as a reference point, it is in fact sterically prohibited because this would require the hydrogen atom in the NH group of peptide unit 1 and the oxygen atom in peptide unit 2 attempting to occupy the same space. The situation $\phi=\psi=180^\circ$ corresponds to where the two peptide units are fully extended and lie in the same plane. The allowed values for ϕ and ψ can be graphically presented in the form of a Ramachandran plot [7], examples of which are shown in Figure 2.8. Glycine, with a side chain R comprising a single hydrogen atom, is far less sterically hindered than the examples shown in Figure 2.8 for valine and isoleucine, and the allowed values for ϕ and ψ for glycine occupy all four quadrants of the Ramachandron plot and permit a broad range of conformations. With increasing length and complexity of the side chain R , the degree of steric freedom becomes increasingly smaller.

2.4.2 Polypeptides and Proteins

Three amino acids can covalently bond together to form a tripeptide, and when four do so we have a tetrapeptide, and so on to form resulting structures known as oligopeptides. When many amino acid residues bond together to form a long structure, the result is known as a polypeptide chain, as depicted in Figure 2.9. Proteins are formed from one or more polypeptide chains. To be able to perform their biological function (e.g. as an enzyme or a structural element such as a microfilament) proteins fold into one or more specific spatial conformations dictated by the sequence of residues in their polypeptide chains and the corresponding permitted values for the rotational angles ϕ and ψ . Protein sizes range from a lower limit of around 50 to several thousand amino acid residues. An average protein contains around 300 residues. Very large aggregates can be formed from protein subunits, for example many thousand actin molecules assemble into a microfilament.

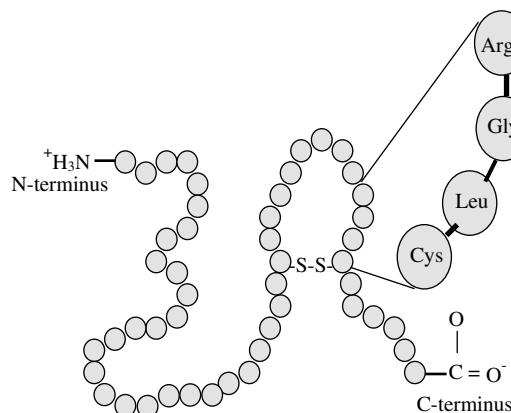


Figure 2.9 A polypeptide is formed of a chain of amino acid residues. It has a defined chemical orientation, with an N-terminus (bearing a free amino group) and a C-terminus (bearing a free carboxyl group). A disulphide bond linking two cysteine residues is shown.

There are four distinct categories of a protein's structure:

- **Primary structure:** This is defined by the amino acid sequence of the polypeptide chains. A specific gene in a cell determines the primary structure of a protein. As will be described later in this chapter, a specific sequence of nucleotides in DNA is transcribed into mRNA, which is then read by structures called ribosomes, in a process called translation. The sequence of a protein is unique to that protein, and defines its structure and function. The primary structure is held together by the covalent peptide bonds made during the process of protein biosynthesis or translation by ribosomes. These peptide bonds provide rigidity to the protein. The primary structure can also be defined by the covalent bonding of sulphur atoms between two cysteine residues in the same or different polypeptide chains. These bonds are termed disulphide bridges, and an example is given in Figure 2.9.
- **Secondary structure:** This refers to the arrangement of parts of a polypeptide chain into highly regular substructures, the most prominent of which are the alpha helix and the beta-pleated sheet structures shown in Figure 2.10. Hydrogen bonds are responsible for

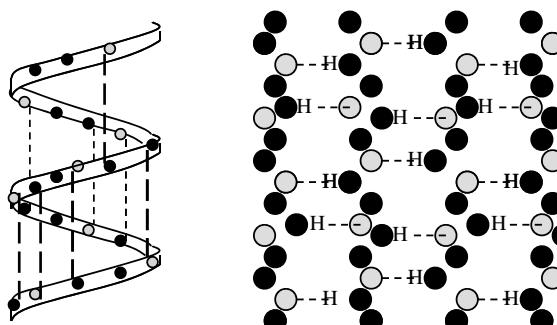


Figure 2.10 The α -helix (left) structure is held together by hydrogen bonds ($\text{NH} \cdots \text{O}$) that are formed nearly parallel to the long axis of the helix. The β -sheet (right) is held together by hydrogen bonds between adjacent sections of a polypeptide chain.

stabilising these two structures. The conformations of the amino acid residues in the alpha helix correspond to values for ϕ of -45° to -50° and $\psi = -60^\circ$, and each turn of the helix includes 3.6 residues. Each residue participates in a hydrogen bond, so that each successive helix turn is held in place to an adjacent helix turn by three to four hydrogen bonds. The residues in a beta-pleated sheet structure have conformations with $\phi = -135^\circ$ and $\psi = +135^\circ$, and is also held together by hydrogen bonds. However, because water-amide hydrogen bonds are generally stronger than amide-amide hydrogen bonds, these secondary structures are stable only when the local concentration of water is sufficiently low, as for example in the fully folded protein state.

- **Tertiary structure:** This is the 3D structure of a single protein molecule, involving the spatial arrangement of the secondary structures, including the folding of parts of the polypeptide chain between α -helices and β -sheets. As depicted in Figure 2.11, it describes the completely folded and compacted polypeptide chain. Several polypeptide chains can be combined into a single protein molecule through ionic interactions (salt bridges) between oppositely charged ionised side-chains, hydrogen bonds, hydrophobic ‘bonding’ interactions, disulphide bridges and intermolecular van der Waals forces between nonpolar groups. As a general rule, the hydrophilic (charged and polar) amino acid residues are located on the outside of a folded protein, with the hydrophobic residues buried inside the polypeptide structure.
- **Quaternary structure:** The forming of a complex of several protein molecules, or protein subunits, that function as part of a larger assembly or protein complex is referred to as a quaternary structure, an example of which is depicted in Figure 2.11. A protein may shift between several, reversible, similar structures in performing its biological function, either as an enzyme controlling chemical reactions or as a structural element.

Hair is primarily composed of keratin – a fibrous structural protein that is also a key structural component of skin and finger nails. Disulphide bonds between cysteine residues in the keratin structure give hair its elasticity. A strand of straight hair can be transformed into curly hair (or curly transformed to straight) by wrapping it around curved (or straight) rods and breaking the disulphide bonds using a reducing solution of sodium or ammonium thioglycolate at a pH of $8 \sim 10$, together with applied heat (the earliest methods used a mixture of cow urine and water!). After $15 \sim 30$ minutes an oxidising lotion (e.g. hydrogen peroxide) and an

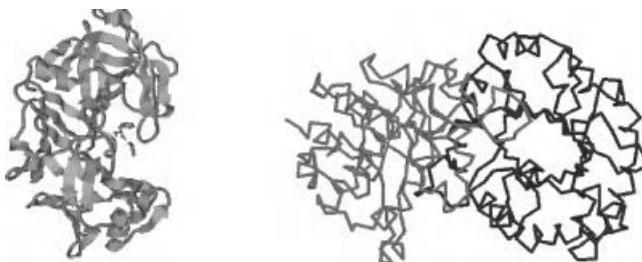


Figure 2.11 The tertiary structure (left) describes the links between α -helices and β -sheets and all of the noncovalent interactions that stabilise the correct folding of a single polypeptide chain in a protein. The quaternary structure (right) refers to the noncovalent interactions that hold together several polypeptide chains into a single protein molecule, as for example chymotrypsin, haemoglobin and RNA polymerase.

alkali neutraliser is applied to bring down the pH and to close the disulphide bonds again to reform the hair into the shape of the rod.

The numbers of constituent amino acid residues and polypeptide chains for several proteins are presented in Table 2.5.

Some proteins, known as conjugated proteins, contain chemical components that are not amino acid residues. These components, known as prosthetic groups, are held in place in the protein structure through interactions with peptide unit side chains. Metalloproteins contain complexes that incorporate metals such as zinc, calcium and copper, and haemoglobin contains an iron complex known as a haem or porphyrin group. Casein found in milk contains a phosphate group, and is an example of a phosphoprotein. The lipoproteins found in blood contain lipids, and glycoproteins contain carbohydrates, of which the immunoglobulin antibody is a good example.

For specified conditions and physiological state of a cell, its complete set of proteins is known as its *proteome*, whose study and application is given the name *proteomics* (rather like genomics for the study of genes). Proteomics includes determinations of: the dynamics and structure of proteins; their abundance (expression) and localisation in a cell; chemical modifications of proteins after their translation from mRNA, and the extent and types of protein-protein interactions. An important post-translational modification is the phosphorylation (addition of a phosphate) of many enzymes and structural proteins involved in cell signalling processes. Cell signalling and cell-cell communication, to be described further in *Section 2.6.11*, governs basic cellular activities and coordinates cell actions. The ability of cells to perceive and correctly respond to their microenvironment is the basis of their development, tissue repair, immunity, and the regulation of their internal environment to maintain a stable and constant physiological state (homeostasis). Errors in cell signalling can result in various diseases, including cancer and diabetes, and by understanding this diseases may be treated effectively. Proteomics also offers the promise to identify targets for new drugs. For example, if a certain enzyme is known to be implicated in a disease, its three-dimensional structure can provide the information to design a drug molecule that fits into the active site of the enzyme and blocks its activity. As knowledge of the genetic differences between individuals is gained, the development of personalised drugs will become possible.

A protein's function is determined by its three-dimensional structure, which in turn is determined by the linear sequence(s) of the amino acids in the polypeptide chain(s) of which

Table 2.5 The composition and molecular weight of some proteins

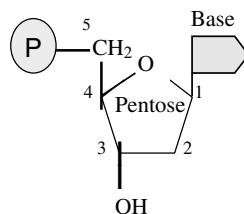
Protein	Number of		Molecular weight
	Residues	Polypeptide chains	
Cytochrome -c	105	1	12 330
Lysozyme	129	1	13 930
Myoglobin	153	1	16 890
Haemoglobin	574	4	64 500
Serum Albumin	609	1	68 500
RNA Polymerase	4158	5	450 000

it is composed. Instructions for the assembly of the amino acid sequence are coded by the linear sequence of nucleotides of the nucleic acid DNA.

2.5 Nucleotides, Nucleic Acids, DNA, RNA and Genes

Nucleic acids are responsible for storing the information and instructing the cell about the proteins it should synthesise. The chemical nature of this genetic process was discovered in 1944, and the description of the structure of the DNA double helix was given by Watson and Crick in 1953 [8]. The DNA double helix has assumed the symbol for the discipline of molecular biology, whose primary function is elucidation of the nature and methods of replication and expression of genetic information. The exciting way in which molecular biology evolved has been described by Judson in his colourful book *The Eighth Day of Creation*, [9]. What follows here is a basic description of the molecular actors and actions of this story.

The two information-storing molecules in cells are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Proteins are polymers constructed from 20 different monomers (the amino acids) but DNA and RNA are comprised of just four monomers – called nucleotides. A nucleotide is composed of a phosphate group (P) and a ‘base’ linked together by a five-carbon sugar molecule (pentose):



The bases found in DNA are adenine, guanine, cytosine and thymine, often abbreviated as A, G, C and T. In RNA the thymine base (T) is replaced by uracil (U).

In DNA the pentose sugar molecule is *deoxyribose*, whereas in RNA it is *ribose*.

2.5.1 DNA

Nucleic acids consist of polymer chains of nucleotides, formed in a condensation reaction to create a *phosphodiester* bond, in which, as shown in Figure 2.12, a water molecule is released (as with the production of a glycosidic bond between sugars or a peptide bond between amino acids). Two nucleotides joined by such a bond forms a dinucleotide, and a trinucleotide represents a single strand of DNA containing three nucleotides. Additional nucleotides can be added to produce a long DNA single strand having a defined chemical orientation. One end (the so-called 3' end) of a DNA strand has a free hydroxyl group (attached to carbon 3 of the sugar), whilst the other end (the 5' end) has a phosphate group. This orientation has important implications regarding the properties of DNA.

At the beginning of Chapter 1, the ability to self-produce was given as a defining characteristic of a living system. The metaphor has even been proposed that living things are ‘chemical machines, whose object is to make two where there was one before’ [10]. The entity in a cell that is responsible for this ‘secret of life’ is DNA. The biologically native state of DNA is a double helix composed of two intertwined single strands of DNA, as schematically shown in Figure 2.13.

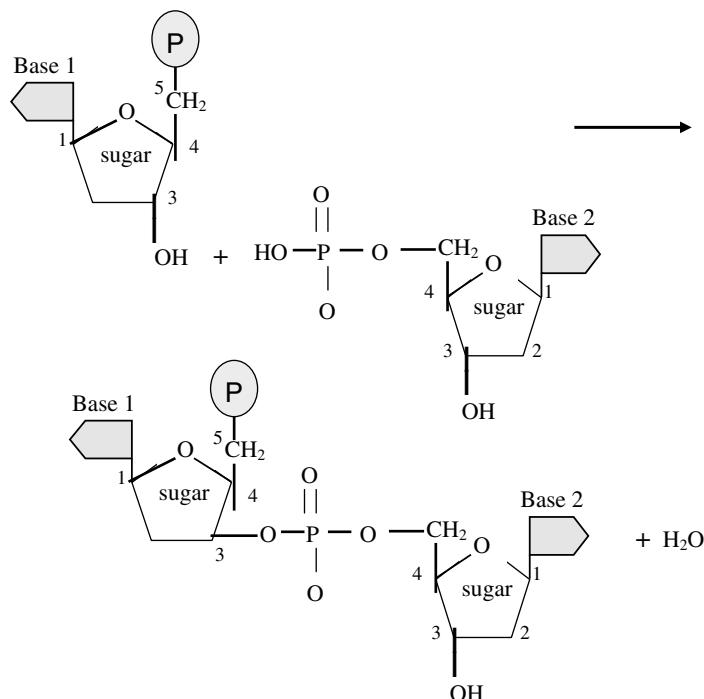


Figure 2.12 The condensation reaction that links two nucleotides with a phosphodiester bond.

As indicated in Figure 2.13, the two single strands of DNA in the double helix structure proceed from carbon 5' to 3', but are directed in opposing directions. The two DNA strands are held together by hydrogen bonds linking their bases, and only if the two strands of the helix are antiparallel can the members of each base pair fit together within the double helix. The two DNA strands can, in theory, form either a right-handed or left-handed helix, but the structure of the sugar-phosphate backbone is such that the right-handed helix is the more favourable geometry. As first proposed by Watson and Crick [8] (with their famous understatement: '*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*') the size, shape and

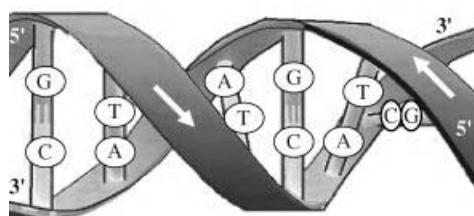


Figure 2.13 A schematic of the DNA double helix.

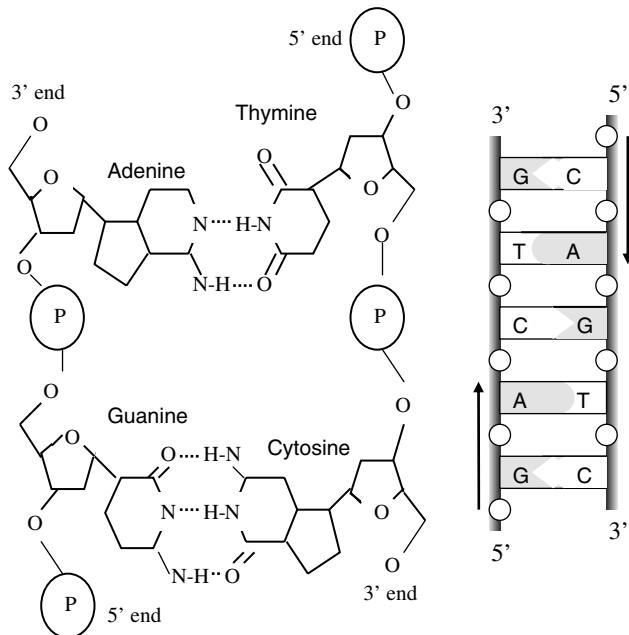


Figure 2.14 A schematic of the double-stranded DNA, showing the base-pair complementarity of the A-T and G-C pairings.

chemical composition of the bases dictates that base A is always paired with T, and G is paired with C. As shown in Figure 2.14 the A-T pair is held together by two hydrogen bonds and the G-C pair by three. To break the G-C pair thus requires more energy (87.9 kJ/mol) than that required to break the A-T pair (50.6 kJ/mol) [11]. This difference is reflected in the finer details of how the DNA polymer is copied. X-ray studies have determined that the stacked bases are regularly spaced 0.34 nm apart along the helix, and that the length of one complete helix turn is 3.4 nm (to give ~10 pairs of bases per turn). The hydrogen bonds between the bases gives the double helix considerable stability and rigidity, but also allows the double helix a good degree of flexibility, enabling long DNA chains to coil up to form supercoils or condensed structures of very large molecular weight. The polypeptide alpha-helix shown in Figure 2.10 is far less flexible, because the hydrogen bonds hold together adjacent parts of the helix.

From Table 2.5 we note that the number of amino acid residues in proteins ranges from hundreds to several thousands – whereas DNA molecules are typically very much larger. For example, the DNA molecule in the single chromosome of an *E. coli* bacterium comprises just under five million base pairs (this number of base-pairs defines the genome size of *E. coli*). Thus, if fully stretched out its DNA would have a length of ~1.5 mm – some three orders of magnitude longer than the *E. coli* bacterium itself! So how does the chromosome package itself inside this bacterium? The solution lies in the flexibility of a DNA double helix that allows it to coil and fold into a superhelix. This can be simulated by continuously twisting an elastic band and slowly bringing the ends together, so that the twisted band first forms small coils that then proceed to curl into a tight knot. Human cells contain 46 chromosomes,

containing a total of 3.2×10^9 base-pairs. If the DNA from all 46 chromosomes of a single human cell were to be connected and straightened out, its total length would be ~ 2 m!

To assist in the packaging of this DNA into the nucleus of a human cell it is wrapped around protein molecules, called histones, to form structural units called nucleosomes that are spaced at regular intervals along the main DNA chain, rather like beads on a string. Arrays of nucleosomes form chromatin fibres that are then further packaged into chromosomes. This form of DNA packaging occurs in cells with a nucleus (eukaryotic cells) but not in prokaryotic cells such as bacteria, where typically the total DNA forms a large circular molecule. A gene corresponds to a stretch of DNA that contains the sequential information for the production of proteins or RNA chains that have functional roles in the cell. Some stretches of DNA do not encode for proteins or RNA, and at present a quite large percentage of this so-called ‘junk’ DNA has no known biological function. The entirety of the genes and noncoding sequences of DNA in a cell is called its genome. Many types of virus do not possess DNA, and instead their genome consists of the coding information contained in another polynucleic acid called RNA.

At each division of a cell in an organism an exact copy must be made of its genome. How is this accomplished? In Figure 2.14 we see that one strand, say S1, of the DNA double helix is composed of a sequence of nucleotides that is exactly mirrored by its complimentary bases in the other strand S2. Thus, strand S1 can serve as a template for making a new S2 strand, and strand S2 can serve the same purpose for S1. This replication process can be observed, using an electron microscope, in the form of Y-shaped structural forks moving along a DNA double helix (Figure 2.15).

The forks shown in Figure 2.15 are produced by initiator proteins that bind to the DNA with the result that strand S1 separates away from strand S2. This requires the breaking of the hydrogen bonds that hold together the base pairs, and tends to happen at an A-T base pair because the energy required is about half of that required to break the G-C base pair. A group of proteins are then attracted to the exposed fork, an important member of which is the enzyme DNA polymerase that places a new complimentary DNA strand onto each separated S1 and S2 to produce two new DNA double-helices. DNA polymerase can synthesise a new DNA strand continuously in only one direction, because it adds new nucleotides to the 3' end of a strand – not to the 5' end, as depicted in Figure 2.16. This limitation is overcome through a discontinuous process whereby the polymerase ‘backstitches’ short sections of DNA onto the 5' end strand with the assistance of another enzyme called DNA ligase. The new nucleotides enter these processes as nucleoside triphosphates, the breaking of whose bonds

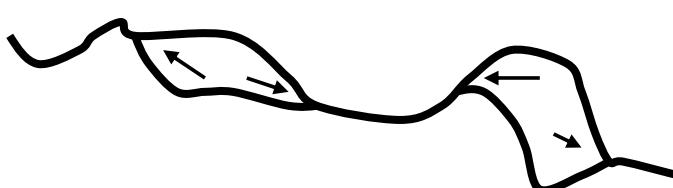


Figure 2.15 A schematic representation (as viewed with an electron microscope) of a portion of a DNA double helix in the process of being replicated. The arrows indicate the locations of four Y-shaped forks and their directions of propagation (Derived from Kornberg & Baker [12].)

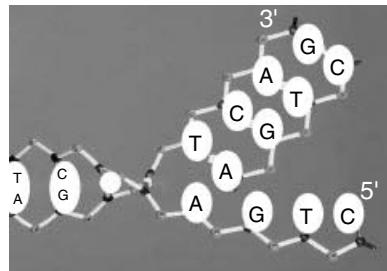


Figure 2.16 At a replication fork DNA polymerase continuously synthesises a new DNA strand by adding new nucleotides to the separated strand with the 3' end. This produces one copy of the original DNA double helix. Complimentary bases are discontinuously ‘backstitched’ onto the other strand, to form another copy of the original DNA molecule.

provides the free energy required for the polymerisation reaction. DNA polymerase also serves as its own ‘proofreader’ to ensure that the correct complimentary base has been inserted. In this way less than one error is typically made for every ten million new base pairs formed. A variety of DNA repair enzymes also continuously scan for and correct replication mistakes, or replace damaged nucleotides, using the uncorrupted DNA strand as the template.

The research that led to the understanding of DNA replication, and to the discovery of DNA polymerase, is described in a book by Arthur Kornberg who in 1959 was awarded the Nobel Prize in Physiology and Medicine for this work [13].

2.5.2 Ribonucleic Acid (RNA)

RNA is very similar to DNA but differs in a few important structural details. In a cell, whereas DNA takes the form of a double-stranded helix, RNA is single-stranded. This means that a RNA chain is much more flexible than DNA and can fold up into a variety of three-dimensional shapes containing sections of single strand loops and double helices wherever parallel strands are able to form complimentary nucleotide base pairs. An example of this can be found in the transfer RNA molecule shown in Figure 2.17. Some of the shapes that RNA molecules can adopt enable them to perform catalytic functions. Another difference is that the RNA nucleotides contain ribose (DNA contains deoxyribose – a type of ribose that lacks one oxygen atom) and has the base uracil rather than thymine present in DNA. Examples of complimentary A-U pairing in a RNA molecule can be seen in Figure 2.17.

Different types of RNA are central to the synthesis of proteins and are transcribed from DNA by enzymes called RNA polymerases. These enzymes bind to the DNA in the nucleus of eukaryotic cells, separate the two strands of the nuclear DNA, and pair ribonucleotide bases to the template DNA strand according to the Watson-Crick base-pairing interactions shown in Figure 2.14 (with uracil replacing thymine). Thus, referring to the replication of a DNA molecule depicted in Figure 2.16, the action of RNA polymerase is to produce a strand of RNA with a nucleotide sequence CUGA (rather than the sequence CTGA of a DNA strand if DNA polymerase had been in action). Many RNA polymerases can act on a single strand of DNA at the same time to speed up the transcription process. Roger Kornberg, the son of Arthur Kornberg mentioned for his work leading to the discovery of DNA polymerase, was

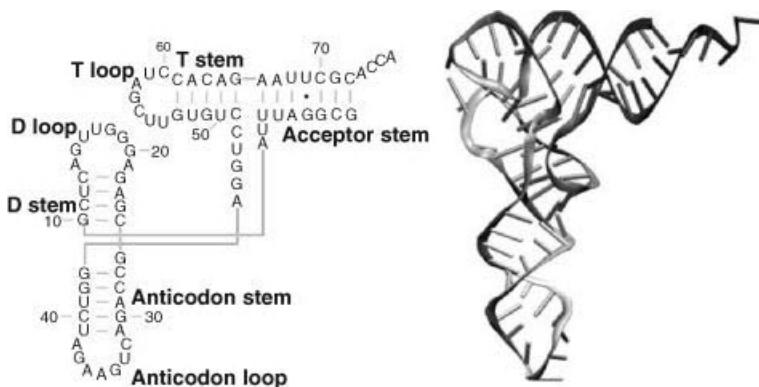


Figure 2.17 The nucleotide sequence and 3-Dimensional structure of a transfer RNA (tRNA) molecule. (Reproduced with permission from Flores, S.C., and Altman, R.B., RNA, 16: 1769–1778, 2010.)

awarded a Nobel Prize in 2006 for his detailed molecular images of RNA polymerase during various stages of the transcription process.

A type of RNA called messenger RNA (mRNA) carries coding information, obtained from the DNA template, in the form of tri-nucleotide units called codons that each code for a single amino acid. Strands of mRNA then interact in the cytoplasm with protein structures called ribosomes (in recent years ribosomes have become important targets in the search for new antibiotics to fight the emergence of drug resistant bacteria). In eukaryotic cells the mRNA is formed inside the nucleus, and has to pass through pores in the nuclear membrane to locate organelles known as ribosomes in the cytoplasm. Ribosomes consist of proteins and ribosomal RNA polymers, which together act as a molecular ‘machine’ to read mRNA and to translate the information it carries into the production of amino acid chains that form proteins. Different types of transfer RNA (tRNA) molecules mediate this process by transferring a specific amino acid to the growing peptide chain. The different tRNA molecules can be attached to only one type of amino acid, and each one contains a three base anticodon that can base pair to the corresponding codon on the mRNA chain. The ‘anticodon’ arm whose loop contains the anticodon is shown in Figure 2.17, together with the 7 base-pair stem that attaches to an amino acid. This process is shown schematically in Figure 2.18.

There are $4^3 = 64$ different codon combinations possible with a triplet codon of three nucleotides, and all 64 codons are assigned for either amino acids or start and stop signals during translation of the mRNA code into a polypeptide sequence. Because there are only 20 common amino acids, there is some redundancy in the assignment of the mRNA triplet codons, as shown in Table 2.6. The codons UAA, UGA and UAG are used as instructions to stop the translation process.

The translation example shown in Figure 2.18 shows a strand of mRNA with the initial codon AUG that serves as an initiation site, where translation into a polypeptide chain begins, and also as the code to produce methionine. The mRNA sequence, CAGGUUCGUG-GAUGC, that follows is translated into a chain of five amino acids comprising glutamine-valine-arginine-glycine-cystine. In Figure 2.18 the tRNA molecule that added arginine to the

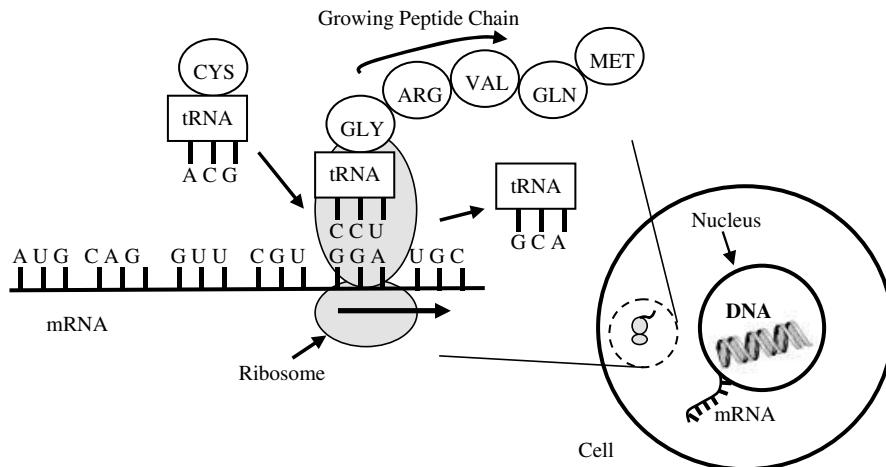


Figure 2.18 Translation of mRNA (from right to left) into a peptide chain. The ribosome begins at the start triplet codon (AUG) at the 3' end of the mRNA, which also codes for methionine. The triplet codons (CAGGUUCGUGGA) that follow produce glutamine, valine, arginine and glycine in the growing peptide chain. A transfer RNA molecule, with its anticodon ACG, brings cystine towards the codon UGC site on the mRNA.

peptide chain is shown leaving the ribosome, whilst a tRNA with its anticodon ACG carries a cystine molecule to the site on the ribosome where the translation process takes place.

The fact that RNA is able to both store information and catalyse chemical reactions has led some biologists to the conclusion that RNA predates DNA in the evolution of living systems [14]. The reign of the RNA world on Earth may have lasted between 3.6 and 4.2 billion years

Table 2.6 The synthesis (translation) of the common 20 amino acids by ribosomes employs the triplet codons (three nucleotide sequences) carried by mRNA listed in this table

Amino acid	Triplet codon	Amino acid	Triplet codon
Alanine	GCU, GCC, GCA, GCG	Leucine	UUA, UUG, CUU, CUC, CUA, CUG
Arginine	CGU, CGC, CGA, CGG, AGA, AGG	Lysine	AAA, AAG
Asparagine	AAU, AAC	Methionine	AUG
Aspartic acid	GAU, GAC	Phenylalanine	UUU, UUC
Cysteine	UGU, UGC	Proline	CCU, CCC, CCA, CCG
Glutamine	CAA, CAG	Serine	UCU, UCC, UCA, UCG, AGU, AGC
Glutamic acid	GAA, GAG	Threonine	ACU, ACC, ACA, ACG
Glycine	GGU, GGC, GGA, GGG	Tryptophan	UGC
Histidine	CAU, CAC	Tyrosine	UAU, UAC
Isoleucine	AUU, AUC, AUA	Valine	GUU, GUC, GUA, GUG

ago. This would have been based on RNA genomes that are copied and maintained through the catalytic function of RNA. Remnants of this may still exist in some microenvironments that have survived to this day, and the construction of artificial RNA-based life from synthetic oligonucleotides may be possible. Until relatively recently, a difficulty with this concept has been the fact that the single-stranded RNA normally expands into a chain one nucleotide at a time, and that in the primordial world RNA did not have enzymes to catalyse this reaction. In 2009, however, it was demonstrated that under favourable conditions of temperature and pH in water small fragments of RNA can fuse into larger lengths of 120 nucleotides and more [15]. This enzyme- and template-independent synthesis of long oligomers of RNA in water, from chemicals that would have existed in prebiotic times, certainly approaches the concept of spontaneous generation of (pre)genetic information.

2.5.3 *Chromosomes*

Earlier we deduced that, if stretched out and joined end to end, the total amount of DNA in each human cell (apart from red blood cells which do not have a nucleus) would have a length of about 2 m. This total DNA, the human genome, contains approximately 3×10^9 nucleotides and is distributed as long lengths of DNA in chromosomes. Apart from the germ cells (eggs and sperm) a typical human cell contains two copies of 22 of these chromosomes, numbered from 1 to 22 in order of diminishing physical size. Females have two X chromosomes and males one X and one Y chromosome to give a total of 46 chromosomes. The X chromosome is inherited from the mother and the Y chromosome from the father. The 22 chromosomes, plus the X and Y chromosome, can be distinguished from one another by staining with dyes that distinguish between DNA that is rich in either A-T or G-C nucleotide base pairs. Each chromosome type can be identified by the distinctive patterns of coloured bands along them, and chromosomal abnormalities can also be detected.

2.5.4 *Central Dogma of Molecular Biology (DNA Makes RNA Makes Protein)*

This *dogma* was first enunciated by Francis Crick [16] and states that the sequential structural information stored in a protein cannot be transferred to another protein or to a nucleic acid. (Crick used the word ‘dogma’ by way of a catch phrase without realising its implied interpretation – in fact he wished his concept to be considered as an hypothesis [Ref. 9, Chapter 6]) In living systems there are three major classes of linear biopolymer, namely DNA, RNA and proteins, whose monomer sequences encode information. There are 9 conceivable direct transfers of information possible between these three classes, as depicted in Figure 2.19a. The transfer of information is assumed to be an error-free transfer in which the molecular sequence of one biopolymer is used as a template to construct another biopolymer with a molecular sequence that depends entirely on that template. Transfers that can occur in all cells, known as general transfers, are the three cases of DNA → DNA (DNA replication), DNA → RNA (transcription), and RNA → protein (translation). Special transfers are ones which do not occur in most cells but may occur in special circumstances, such as in virus-infected cells, and are the three cases of RNA → RNA, RNA → DNA, and DNA → Protein. A known example of the RNA-DNA transfer takes place in retroviruses, where DNA is synthesised using RNA as a

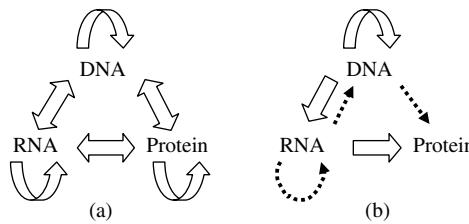


Figure 2.19 (a) The 9 conceivably possible direct transfers of information between DNA, RNA and proteins. (b) The central dogma of molecular biology states that only the transfers represented by the block arrows are possible. The dotted arrows indicate special transfers under specific conditions, such as those involving retroviruses or artificially in a test-tube.

template. An enzyme known as reverse transcriptase carries out this process. The human immunodeficiency virus (HIV) is a retrovirus and is the cause of AIDS.

After HIV has bound to a target cell, normally one of the vital blood cells of the immune system, the RNA content of the virus and various enzymes including reverse transcriptase and ribonuclease, and protease, are injected into the cell. The single strand of viral RNA genome is then transcribed into double strand DNA and integrated into a chromosome of the host cell, which can lead to possible reproduction of the virus [17].

The three special transfer possibilities are included in Figure 2.19b as dotted arrows. Finally, the three remaining transfer possibilities, known as unknown transfers, are considered never to occur and are not included in Figure 2.19b.

The block arrows in Figure 2.19b therefore summarise the normal flow of biological information, namely that DNA information can be transcribed into messenger RNA, and then translated into the synthesis of proteins using the information in mRNA as a template. The master template stored in DNA can itself be replicated, so that the cycle DNA-makes-RNA-makes-protein can be repeated in a new generation of cells and biological organisms. The information stored in a protein cannot be used to make new DNA, RNA or another protein.

2.6 Cells and Pathogenic Bioparticles

We complete this chapter with a summary description of different types of cells, including blood cells and bacteria, along with viruses and prions as examples of pathogenic bioparticles.

The cell is the structural and functional unit of all known living organisms. It is the smallest unit of an organism that is classified as living, and is often called the building block of life. Organisms, such as most bacteria, consist of a single cell but other organisms such as animals are composed many cells, they are multicellular. Humans, for example, comprise around 10^{14} cells of typical diameter $10\text{ }\mu\text{m}$, and each of mass around 1 nanogram. Cells are limited in size by the ratio between their outer surface area and their volume. A small cell has more surface area through which to exchange nutrients, gases and other chemicals between the external and internal cell media than a large cell for a given volume of nucleus. There is also a limit to the amount of biochemical processes that a nucleus can control in a cytoplasm.

2.6.1 Prokaryotic and Eukaryotic Cells

The names for these two basic cell types are derived from the Greek word *Karyose*, meaning ‘kernel’, which in biology is used to refer to the nucleus of a cell. *Pro* means *before*, and *eukaryotic* means *true or good*. Thus, *Prokaryotic* means *before a nucleus*, and *eukaryotic* means *possessing a true nucleus*.

These two cell types perform many similar biological functions. Both are enclosed by plasma membranes, filled with cytoplasm, and loaded with ribosomes. Both have DNA which carries the archived instructions for operating the cell. The DNA in the two cell types has the same structural form, and the genetic code for a prokaryotic cell is exactly the same genetic code used in eukaryotic cells. Eukaryotic animal cells are generally thought to have descended from prokaryotes that lost their cell walls. The cell wall has pores that allow materials to enter and leave the cell, but they are not very selective about what passes through. The plasma membrane, which lines the inner cell wall surface, provides the final filter between the cell interior and the environment. With only the flexible plasma membrane left to enclose them, the primordial prokaryotes would have been able to expand in size and complexity. **Eukaryotic cells are generally ten times larger than prokaryotic cells and have membranes enclosing interior components, the organelles.** Like the exterior plasma membrane, these membranes also regulate the flow of materials, allowing the cell to segregate its chemical functions into discrete internal compartments.

Other important differences between prokaryotic and eukaryotic cells include:

- **Eukaryotic cells have a true nucleus, representing the largest organelle in a cell, bound by a double membrane. Prokaryotic cells have no nucleus.** The purpose of the nucleus is to sequester the DNA-related functions of the big eukaryotic cell into a smaller chamber, for the purpose of increased efficiency. This function is unnecessary for the prokaryotic cell, because it’s much smaller size means that all materials within the cell are relatively close together.
- **Eukaryotic DNA is linear; prokaryotic DNA is circular and has no ends.**
- **Eukaryotic DNA is complexed with proteins called *histones*, and is organised into chromosomes. Prokaryotic DNA is *naked*, meaning that it has no histones associated with it, and it is not formed into chromosomes.** A eukaryotic cell contains a number of chromosomes; a prokaryotic cell contains only one circular DNA molecule and a varied assortment of much smaller circlets of DNA called *plasmids*. The smaller and simpler prokaryotic cell requires far fewer genes to operate than the eukaryotic cell.
- Both cell types have a large number of ribosomes, but the ribosomes of the eukaryotic cells are larger and more complex than those of the prokaryotic cell. Ribosomes are composed of a special class of RNA molecules (ribosomal RNA, or rRNA) and a specific collection of different proteins. A eukaryotic ribosome is composed of five kinds of rRNA and about eighty kinds of proteins. Prokaryotic ribosomes are composed of only three kinds of rRNA and about 50 kinds of protein.
- **The cytoplasm of eukaryotic cells takes the form of a gel-like material filled with a complex collection of organelles, many of them enclosed in their own membranes. A prokaryotic cell contains no membrane-bound organelles.** This is a significant difference and the source of the vast majority of the greater complexity of the eukaryotic cell. There is much more space within a eukaryotic cell than within a prokaryotic cell, and many of the

Table 2.7 The characteristic differences between prokaryotic and eukaryotic cells

Feature	Prokaryote	Eukaryote
Size	Small: $0.5 \sim 5 \mu\text{m}$	$5 \geq 50 \mu\text{m}$
Genetic material	Circular DNA (in cytoplasm)	DNA in form of linear chromosomes (in nucleus)
Organelles	Few present	Many organelles
Cell walls and other structures	Rigid, formed from glycoproteins. (Bacteria also contain flagellum, plasmid and capsule)	Fungi: Rigid, formed from polysaccharides (chitin). Plant: Rigid, formed from polysaccharides (e.g. cellulose). Animals: No cell wall

organelles structures, like the nucleus, increase the efficiency of bioreactions by confining them within small volumes. If the organelles are removed, the soluble part of the cytoplasm that remains is called the cytosol, consisting mainly of water and dissolved substances such as mineral salts and amino acids.

A summary of the main characteristics that distinguish prokaryotic from eukaryotic cells is given in Table 2.7.

2.6.2 The Plasma Membrane

All living cells have a plasma membrane that encloses their contents and serves as a semi-porous barrier to the outside environment. The membrane acts as a boundary, holding the cell constituents together and keeping other substances from entering. The plasma membrane is permeable to specific molecules, however, and allows nutrients and other essential elements to enter the cell and waste materials to leave the cell. Small molecules, such as oxygen, carbon dioxide, and water, are able to pass freely across the membrane, but the passage of larger molecules (e.g. amino acids and sugars) is carefully regulated. The biophysical and electrical properties of membranes are discussed further in Chapter 3.

According to the accepted current model, known as the *fluid mosaic model*, the plasma membrane is composed of a phospholipid bilayer (see Figure 2.2). Individual lipids and proteins can move freely within the bilayer as if it was a fluid, and the membrane-bound proteins form a mosaic pattern when looking at the membrane surfaces. Within the phospholipid bilayer of the plasma membrane, many diverse proteins are embedded, while other proteins simply adhere to the surfaces of the bilayer. Some have carbohydrates attached to their outer surfaces and are referred to as *glycoproteins*. The positioning of proteins on the plasma membrane is related in part to the organisation of the filaments that comprise the cytoskeleton, which help anchor them in place. The cytoskeleton forms the framework of a cell. It consists of protein microfilaments and larger microtubules that support the cell, to give it its shape and help with the movement of its internal organelles. The arrangement of proteins also involves the hydrophobic and hydrophilic regions found on the surfaces of the proteins. The hydrophobic regions of the protein associate with the hydrophobic interior of the plasma

membrane, whereas hydrophilic regions extend past the surface of the membrane into either the cytosol of the cell or the outer environment. Many of the transmembrane protein structures form channels and pumps.

An important membrane ion pump is the $\text{Na}^+ \text{-K}^+$ pump. This pump actively transports Na^+ ions out of a cell, and K^+ ions into a cell, against their electrochemical gradients. This is described in more detail in Chapter 3. The $\text{Na}^+ \text{-K}^+$ pump plays a direct role in regulating the osmolarity of the cytosol, along with the action of water channel proteins (aquaporins) that allows water to flow down its activity gradient into or out of a cell. To avoid influencing the ion gradients across the membrane, an aquaporin permits the rapid passage of water molecules but blocks the passage of ions. This is achieved through the special structure of the aquaporin channel, which consists of a narrow pore lined on one side by hydrophilic amino acids and on the other side by hydrophobic amino acids. The water molecules follow the path, one by one in single file, presented by the hydrophilic groups, to which they make transient hydrogen bonds with carbonyl oxygens [18]. The pore diameter is too small to permit the passage of hydrated ions and as explained in the next chapter the energy required to remove the hydration shell around an ion, plus the electrostatic interaction of a bare ion with a hydrophobic surface, presents an insurmountable energy barrier for the passage of an ion. We also learnt in Chapter 1 that protons in solution exist as the hydronium ion $(\text{H}_2\text{O})_3\text{H}_3\text{O}^+$, shown in Figure 1.4, so that along with Na^+ , K^+ , Ca^{2+} and Cl^- ions we can also understand why aquaporins do not allow the passage of H^+ ions (protons).

2.6.3 The Cell Cycle

A necessary ability of all living cells is to duplicate their genomic DNA and to pass identical copies of it to their daughter cells. All growing cells perform this function in their cell cycle, which consists of two periods, namely the period of cell division and an interphase period of cell growth. The ways in which prokaryotic and eukaryotic cells coordinate their DNA synthesis and cell division are quite different.

As indicated above, the genome of a prokaryotic cell is a single circular molecule of DNA. In rapidly growing prokaryotes, such as bacteria, its DNA is replicated throughout much of the cell cycle process. The circular chromosome of the mother cell is attached to the internal plasma membrane surface to facilitate the DNA-replication process. When this replication is complete, the new chromosome is attached at another site on the membrane. A new membrane and cell form in the region between the points of attachment of the two chromosomes, parts of which invaginate to produce a septum that divides the cell. The two daughter cells separate, each one with its chromosome attached to its inner membrane surface. The cell cycle is complete.

The life-cycle scheme for all growing and dividing eukaryotic cells is shown in Figure 2.20. Nerve cells and striated muscle cells do not divide. These cells do continue to synthesise RNA, proteins and membrane material, but do not replicate their DNA. The red blood cells of mammals also do not divide – they do not possess a nucleus.

The major part of the cell cycle shown in Figure 2.20 comprises the G_1 (first gap), the S (synthetic) phase, and the G_2 (second gap) period. These make up the interphase when new DNA, membrane material and other macromolecules are synthesised. The remaining, relatively short, part of the cell cycle is the M (mitosis) period during which time the cell divides (cytokinesis). The replication of DNA and the synthesis of histone proteins take place only

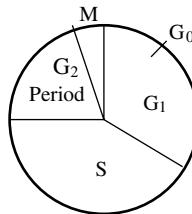


Figure 2.20 The cell cycle for dividing mammalian cells. During the interphase period consisting of G₁, S and G₂, DNA and other cellular materials such as RNA, protein and membrane are synthesised. Cell division occurs during the mitosis (M) period. Terminally differentiated cells, or cells in a culture with depleted nutrients, stop at the G₀ state.

during the S phase, during which each DNA double-helix is replicated and quickly combined with histones and other chromosomal proteins. During the two gap periods, G₁ and G₂, no net synthesis of DNA occurs but damaged DNA can be repaired. During the G₂ period a cell should contain two copies of each of the DNA molecules existing in the cell at G₁. Throughout the interphase (G₁, S and G₂) there is continuous cell growth and synthesis of macromolecules such as RNA, proteins and membranes. Finally, during the relatively brief mitotic (M) period the cell divides, and identical copies of the DNA are distributed to each of the two daughter cells. Nonreplicating cells such as nerve or striated muscle cells are in the so-called terminally differentiated state and are usually halted in the cell cycle at the G₀ stage shown in Figure 2.20. During cell culture growth of mammalian cells the S, G₂ and M periods are roughly constant, whereas the G₁ period can vary greatly depending on the culture conditions. If the culture medium becomes depleted of the required nutrients or hormones, for example, they can remain at the G₀ stage for many hours or days, until stimulated back into G₁ by the addition of the missing growth medium components.

2.6.4 Blood Cells

An adult human body contains 5 ~ 6 litres of blood, 55% of which is a liquid called plasma or serum (when clotting factors are removed). The remaining 45% comprises the blood cells. The blood performs important functions. These functions include: carrying oxygen to the tissues and collecting carbon dioxide; conveying nutritive substances (e.g. amino acids, sugars, mineral salts) and collecting waste material to be excreted; transporting hormones, enzymes and vitamins. A very important function is to protect an organism against disease agents, using the immune response provided by lymphocytes, the phagocytic activity of leukocytes, and the bactericidal power of the plasma. These topics are of relevance to bioelectronics for situations where a sensor or other type of electronic device is implanted in the body. Tissue damage will occur on implanting such a device, and its performance can be degraded as a result of the inflammatory and immune responses involved in the healing of this wound. This is described further in Chapter 6, Section 6.12.

The cellular composition of normal human blood is given in Table 2.8. 1 μ L of whole blood (equivalent to about one-fortieth of a drop of blood) contains up to six million red blood cells and a much smaller number (~300) of B cells. An increase in the number of lymphocytes is normally an indication of a viral infection, whilst a lower than normal concentration can be associated with an increased rate of infection after surgery or traumatic

Table 2.8 The composition of 1 μL of human blood (equivalent to $\sim 1/40$ th of a drop of blood)

Erythrocytes (red cells)	5 ~ 6 million
Platelets	$\sim \frac{1}{4}$ million
Leukocytes (white cells)	\sim 7 thousand
	Comprising:
	4400 Granulocytes
	400 Monocytes
	2200 Lymphocytes
	Comprising:
	1500 T-cells
	400 NK-cells
	300 B-cells

injury. A reduction of T cells occurs when the human immuno-deficiency virus (HIV) infects and destroys T cells.

2.6.4.1 The Plasma

If a sample of blood is centrifuged, the cells sediment to the bottom of the tube to leave $\sim 55\%$ of the sample at the top in the form of a slightly alkaline ($\text{pH } 7.4$) and pale yellow fluid, comprising 90% water and 10% solid matter (9 parts organic and 1 part mineral). The organics include amino acids, glucose, hormones, lipids, proteins and vitamins. The minerals take the form of ions such as Na^+ , K^+ , Ca^{2+} , Mg^{2+} and Cl^- .

2.6.4.2 Platelets (Thrombocytes)

The main function of platelets is to prevent the loss of blood in injured tissues, by aggregating and releasing chemicals to promote blood coagulation. Released substances include serotonin which reduces the diameter of damaged blood vessels, and fibrin to trap cells and form a clot. They have a diameter of $2 \sim 3$ microns and are not considered to be real cells.

2.6.4.3 Erythrocytes (Red Cells)

These cells are rich in haemoglobin (~ 250 million per cell), which is a protein able to bind oxygen and thus responsible for providing oxygen to tissues. Where there is a high concentration of oxygen in the body, such as in the alveoli of the lungs, each haemoglobin molecule binds four oxygen molecules to form oxyhaemoglobin. When an erythrocyte reaches tissue with low oxygen concentration the haemoglobin releases these oxygens. Erythrocytes are also partly responsible for recovering carbon dioxide produced as waste, but most CO_2 is carried by plasma in the form of soluble carbonates. The mean lifetime of erythrocytes is about 120 days, at which time they are retained by the spleen and then phagocytized (eaten) by macrophages.

In man and in all mammals, erythrocytes are devoid of a nucleus and have the shape of a biconcave lens, which allows more room for haemoglobin and raises the cell surface and

cytoplasmic volume ratio. These characteristics maximise the efficiency of oxygen diffusion by these cells. In fishes, amphibians, reptilians and birds, erythrocytes do have a nucleus.

2.6.4.4 Leukocytes (White Cells)

These cells are responsible for the defence of the organism, and are of two types, namely *granulocytes* and *lymphoid cells*. The number ratio of white to red blood cells in normal human blood is $\sim 1:700$.

Granulocytes contain granules in their cytoplasm, which have different properties including a different affinity towards neutral, acid or basic stains. Granulocytes can thus be distinguished as neutrophils, eosinophil (or acidophils) and basophils;

- *Neutrophils*: Act to phagocytose bacteria and are present in large numbers in the pus of wounds. They are unable to renew the lysosomes used in digesting the bacteria and die after having phagocytosed a few of them.
- *Eosinophils*: Attack parasites and phagocytose antigen-antibody complexes.
- *Basophils*: Possess a phagocytory capability, but also secrete anticoagulant and vasodilatory substances such as histamines and serotonin.

Lymphoid Cells consist of two types, namely lymphocytes and monocytes:

- Lymphocytes are cells, which besides being present in blood populate the lymphoid tissues and organs (e.g. thymus, bone marrow, spleen). They are slightly larger than erythrocytes, and have a nucleus that occupies nearly all of the internal cellular volume. They are also the main constituents of the immune system, which is the defence against the attack of pathogenic microorganisms such as viruses, bacteria, fungi and protista (e.g. unicellular organisms). Lymphocytes produce *antibodies*, which appear on their outer plasma membrane. An antibody is a molecule able to ‘recognise’ and bind itself to molecules called *antigens*. As for all proteins, these antibodies are coded by genes. On the basis of a recombination mechanism of some of these genes, every lymphocyte produces antibodies of a specific molecular shape. The number of lymphocytes circulating in the blood is so large that they are able to recognise practically all the chemicals existing in the organism, both its own natural and foreign ones. They recognise hundreds of millions of different molecules!

The cells of the immune system, chiefly lymphocytes, cooperate amongst themselves to activate, boost or make more precise the immune response. To attain this scope, there exist different types of lymphocytes, with different functions, namely B and T lymphocytes. When the B cells are activated, they quickly multiply and secrete hosts of antibodies, which on meeting microorganisms with complementary shape (*epitopes*) bind to them and form complexes to immobilise the microorganisms. Other cells, which are not specific but able to recognise antibodies, phagocytose these complexes. In their turn, the T cells are divided into three categories: Tc (cytotoxic) cells kill infected cells directly by inducing them to undergo apoptosis (programmed cell death); Th (helpers) assist in activating B cells to make antibody responses; Ts (suppressors) suppress the activity of other T cells and are crucial for self tolerance. The immune system also produces *memory cells*, which are deactivated lymphocytes ready to be reactivated on further encounters with the same antigen.

Another population of lymphocytes in the peripheral blood and lymphoid organs do not have receptors for antigens. These lymphocytes have a nonspecific defence function that is not activated by Th lymphocytes. These cells represent the more ancient component of the immune system and they are characterised by their cytotoxic activity. They are called *Natural Killer(NK)* cells. Apart from killing viruses, bacteria, infected and neoplastic (abnormal) cells, these lymphocytes also regulate the production of other haematic cells such as erythrocytes and granulocytes.

- Monocytes are the precursors of *macrophages*. They are larger blood cells, which after attaining maturity in the bone marrow, enter the blood circulation where they stay for 24–36 hours. Then they migrate into the connective tissue, where they become macrophages and move within the tissues. In the presence of an inflammation site, monocytes quickly migrate from blood vessels and start an intense phagocytory activity. Macrophages also cooperate in the immune defence by exposing molecules of digested bodies on their membrane, presenting them to more specialised cells such as B and T lymphocytes.

2.6.5 *Bacteria*

Bacteria are prokaryotic cells and are about one-tenth the size of eukaryotic cells, typically $0.5 \sim 5.0 \mu\text{m}$ in length. They display a wide range of morphologies. Most are either spherical (cocci), rod-shaped (bacilli) or spiral-shaped (spirilla). Many bacterial species exist simply as single cells, but others associate in characteristic patterns. For example, *Streptococcus* form chains, and *Staphylococcus* group together into clusters. Only a small number of bacterial species cause disease in humans. Some of those that do so can only replicate inside the body and are called *obligate pathogens*. Other bacteria, called *facultative pathogens*, replicate in environments such as water or soil and only cause disease on encountering a susceptible host. *Opportunistic pathogens* are bacteria that are normally harmless but have a latent ability to cause disease in an injured or immuno-compromised host.

A plasma membrane encloses the contents of a bacterial cell and acts as a barrier to hold nutrients, proteins and other essential components of the cytoplasm within the cell. Many important biochemical reactions are driven by concentration gradients of ions and molecules across the membrane. Bacteria lack a nucleus, mitochondria, and the other organelles present in eukaryotic cells. Their genetic material is typically a single circular chromosome located in the cytoplasm in an irregularly shaped body called the nucleoid, which contains the chromosome with associated proteins and RNA. Around the outside of the plasma membrane is the bacterial cell wall, composed of peptidoglycan and made from polysaccharide chains cross-linked by peptides. There are broadly speaking two different types of cell wall in bacteria – Gram-positive and Gram-negative, according to their reaction with the Gram stain. Gram staining involves applying crystal violet to a heat-fixed smear of a bacterial culture, followed by the addition of iodine, rapid decolourisation with alcohol or acetone, and counterstaining. After decolourisation the Gram-negative cell has lost its outer membrane and loses its purple colour, whereas the Gram-positive cell remains purple. The counter-stain, safranin or basic fuchsin, is applied last to give decolourised Gram-negative bacteria a pink or red colour. This staining procedure works because Gram-positive bacteria possess a thick cell wall containing many layers of peptidoglycan and teichoic acids. The crystal violet-iodine complex becomes trapped within the peptidoglycan multilayers. Gram-negative bacteria, on the other hand, have

a relatively thin cell wall consisting of a few layers of peptidoglycan surrounded by a second lipid membrane containing lipopolysaccharides and lipoproteins. These differences in structure can produce differences in antibiotic susceptibility. As a general rule Gram-negative bacteria are more pathogenic, because the lipopolysaccharide in their outer membrane breaks down into an endotoxin which increases the severity of inflammation. The human body does not contain peptidoglycan, and humans produce an enzyme called lysozyme that attacks the open peptidoglycan layer of Gram-positive bacteria. Gram-positive bacteria are also more susceptible to antibiotics such as penicillin, which inhibit a step in the synthesis of peptidoglycan.

Bacteria are often grown in solid media, such as agar plates, to isolate and identify pure cultures of a bacterial strain. However, liquid growth media are used when measurement of growth or large volumes of cells are required. Growth in stirred liquid media occurs as an even cell suspension, making the cultures easy to divide and transfer, although isolating single bacteria from liquid media is difficult. The use of selective media (media with specific nutrients added or deficient, or with antibiotics added) can help identify specific organisms. Bacterial growth follows three phases. When bacteria first enter a high-nutrient environment that allows growth, the cells need to adapt to their new environment. The first phase of growth is the *lag phase*, a period of slow growth when the cells are adapting to the high-nutrient environment and preparing for fast growth. The lag phase has high biosynthesis rates, as proteins necessary for rapid growth are produced. The second phase of growth is the logarithmic phase, marked by rapid exponential growth. The rate at which cells grow during this phase is known as the *growth rate*, and the time it takes the cells to double is known as the *doubling time* or *generation time*. During log phase, nutrients are metabolised at maximum speed until one of the nutrients is depleted and starts limiting growth. The final phases are the *stationary phase*, followed by the *death phase*, caused by depleted nutrients. The cells reduce their metabolic activity, consume nonessential cellular proteins, and then die.

A practical way to monitor bacterial growth and to determine the doubling time is to periodically measure the optical absorbance of a sample of a bacterial suspension. The size of a typical bacterium is such that it will scatter light of wavelengths around $450 \sim 600$ nm. An example of an absorbance plot taken over the lag phase, the logarithmic phase and into the stationary phase of a growing culture of *Micrococcus luteus* is given in Figure 2.21.

Exponential growth of the number of bacteria n with time t , commencing with an initial number n_0 , can be written as:

$$n(t) = n_0 \exp(kt) = n_0 2^{t/T}$$

where k is the growth constant defined as the number of times per unit time of growing by a factor exponential e . T is the time it takes for the number to double, namely the doubling time. If the absorbance plot is converted to a \log_2 plot, as indicated in Figure 2.20, a change of 1 absorbance units means the absorbance has doubled. The slope of this \log_2 plot thus gives the time for the culture to double in number density. The derived doubling time for the *M. luteus* sample is close to 71 minutes. The ‘textbook’ doubling times for bacteria such as *M. luteus* and *E. coli* are often given as 30 and 20 minutes, respectively. However, these values are valid only for optimum conditions of temperature, nutrient concentration and cell density and when no growth suppressing substances are present. Many bacteria produce such substances if their cell density becomes too high. Examples of other typical doubling times are 2 hours and 24 hours, respectively, exhibited by *B. subtilis* and *M. tuberculosis*, respectively.

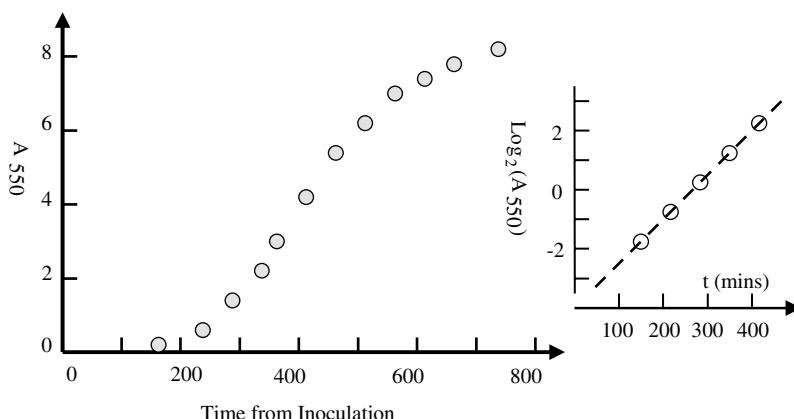


Figure 2.21 Growth curve obtained for *Micrococcus Luteus* obtained by measuring the absorbance at 550 nm of the culture at periodic times after inoculation at time zero (unpublished work). The reciprocal of the $\log_2(A)$ plot at the logarithmic growth phase equals the bacteria doubling time (~ 71 mins).

2.6.6 Plant, Fungal and Protozoal Cells

Plant cells deposit an extracellular matrix around themselves, composed of a crosslinked network of a polysaccharide called cellulose, to form a thick cell wall. Secondary cell walls contain additional molecules to add rigidity and further protection. The cellulose gives the primary cell wall high tensile strength, which allows the plant cell to build up a high internal hydrostatic pressure, known as turgor pressure. This turgor pressure may reach 10 or more atmospheres, in order to attain osmotic equilibrium with the external environment. At equilibrium there is no net influx of water into the cell despite the large osmotic difference caused by the higher concentration of solutes inside the cell wall compared to that outside it. The turgor pressure provides rigidity and the main driving force for plant growth.

Fungi include both unicellular yeast cells, such as *Saccharomyces cerevisiae*, and filamentous, multicellular, moulds such as those found on mouldy fruit or bread. Most pathogenic fungi exhibit *dimorphism*, which is the ability to grow in either yeast or mould form. The yeast-to-mould or mould-to-yeast transition is frequently associated with infection. For example, some fungi grow as a mould at low temperatures in soil, but then change to a harmful yeast form when inhaled into the lungs.

Protozoan parasites exist as single cells and frequently require more than one host in a complex life-cycle. The most common protozoal disease is malaria, transmitted to humans by the bite of the female of any of 60 species of *Anopheles* mosquito. The most intensively studied of the malaria-causing parasites, *Plasmodium falciparum*, exists in eight distinct forms, and requires both human and mosquito hosts to complete its life-cycle. Because fungi and protozoan parasites are eukaryotes, their pathogenic varieties are difficult to kill with drugs without harming the host. The tendency of fungal and parasitic infecting organisms to switch amongst several different forms during their life-cycles also makes them more difficult to treat. A drug that is effective at killing one form is often ineffective at killing another.

form, which therefore survives the treatment. As a result, antifungal and antiparasitic drugs are often less effective and more toxic than antibiotics.

2.6.7 Viruses

Viruses are not cells. They are bioparticles that can vary from simple helical or icosahedral shapes, to more complex structures. They can reproduce only inside a host cell. They are about 1/100th the size of bacteria, with diameters ranging from around 10 to 300 nm. And so, unlike cells and bacteria which can be viewed using a conventional light microscope, most viruses can only be seen using scanning and transmission electron microscopes. A complete virus particle, known as a virion, consists of nucleic acid (DNA or RNA) surrounded by a protective coat of protein called a capsid, made from proteins encoded by the viral genome. The capsid shape serves as the basis for morphological distinction (helical, icosahedral, envelope, complex).

Some species of virus surround themselves with a modified form of one of the host cell membranes, either the outer membrane of the infected host cell, or internal membranes such as nuclear membrane or endoplasmic reticulum. The virus thus gains an outer lipid bilayer, known as a viral envelope. This membrane is studded with proteins coded for by the viral genome and host genome; the lipid membrane itself and any carbohydrates present originate entirely from the host cell. The influenza virus and HIV use this strategy. Most membrane enveloped viruses are dependent on the envelope for their infectivity. The *complex* viruses possess a capsid that is neither purely helical, nor purely icosahedral, and may possess extra structures such as protein tails or a complex outer wall. For example, the T4 bacteriophage has a complex structure consisting of an icosahedral head bound to a helical tail with protruding protein tail fibres. This tail structure acts like a molecular syringe, attaching to the bacterial host and then injecting the viral genome into the cell.

All viral genomes encode three types of proteins, namely proteins for replicating the genome, proteins for packaging the genome and delivering it to more host cells, and proteins that modify the structure or function of the host cell to enhance the replication of the virions. Many viral genomes also encode a protein that subverts the host's normal immune defence mechanisms. The design of effective antiviral drugs is difficult because viruses use the host cell's ribosomes to make their proteins and enable viral replication. Antibiotics such as tetracycline specifically poison bacterial ribosomes, but it is not possible to find an equivalent drug to target viral ribosomes without destroying the host cells. In most cases viral infections in animals cause an immune response that eliminates the infecting virus. These immune responses can also be produced by vaccines that give immunity to a viral infection. The best strategy for containing viral diseases is thus to prevent them by vaccinating the potential hosts. Vaccination programs have effectively eliminated the smallpox virus, and the eradication of poliomyelitis is well advanced.

Viral vaccines take the form of being either live or inactivated. Live vaccines are prepared from attenuated strains that are devoid of pathogenicity but are capable of inducing a protective immune response. They multiply in the human host and provide continuous antigenic stimulation over a period of time. For inactivated whole virus vaccines, the original virus is grown by normal virus culture methods, as for example in tissue culture (the polio Salk vaccine), in eggs (influenza) or mouse brain (rabies Semple vaccine). The virus is then inactivated by formalin or B-propiolactone so that the replication function of the virus is destroyed.

Each year the World Health Organization (WHO) recommends specific vaccine viruses for vaccine production, but then individual countries make their own decision for licensing of vaccines in their country. For vaccination against influenza, three or more vaccine viruses are chosen to maximise the likelihood that the main circulating viruses during the upcoming flu season will be well covered by the vaccine. For example, WHO recommended that the Northern Hemisphere's 2010–11 seasonal influenza vaccine should contain the following three vaccine viruses:

- an A/California/7/2009 (H1N1)-like virus,
- an A/Perth/16/2009 (H3N2)-like virus,
- a B/Brisbane/60/2008-like virus.

This recommended composition of the 2010–11 seasonal vaccine for the Northern Hemisphere, including the EU and the United States, was the same composition that was recommended for the Southern Hemisphere's 2010 influenza vaccines. In the nomenclature used, the letter *A* or *B* corresponds to the type of human influenza which tend to become pandemic, the numbers correspond to virus strains, and the letters *H* and *N* refer to the two viral proteins *hemagglutinin* and *neuraminidase*. Neuraminidase promotes influenza virus release from infected cells and facilitates virus spread within the respiratory tract.

2.6.8 Prions

Prions are infectious agents in the form of misfolded proteins that replicate and propagate in the host cell. They cause various neurodegenerative diseases in mammals, a well-known example being *bovine spongiform encephalopathy* (BSE), otherwise known as mad cow disease. This fatal disease (Creutzfeld-Jacob disease) can be transmitted to humans who eat infected beef, and can also be transmitted from human to human via blood transfusions. The brain tissue develops holes and takes on a spongelike appearance. The DNA code for making prion protein is in a gene that all mammals possess and is mainly active in nerve cells.

The prion has the identical amino acid sequence to the normal form of the protein. The only difference between them is in their folded three-dimensional structure. The misfolded protein can cause a normal folded form to unfold and to aggregate with other prions to produce regular helical structures called *amyloid* fibres. The prion is thus able to cause the normal protein form to adopt its misfolded prion conformation, causing it to become infectious. This is equivalent to prions being able to replicate themselves in the host cell. If an amyloid fibre is broken into smaller pieces, each piece can initiate the prion polymerisation process in a new cell. The prion can therefore propagate as well as replicate. Furthermore, if consumed by another host organism, the newly formed misfolded prions may transmit the infection to that organism. How polypeptide chains explore their conformation-energy space and fall into a global free energy minimum state corresponding to their correct three-dimensional folded form is poorly understood. The linear order of the peptides is reliably given by the pertinent gene's DNA sequence, but referring to Figures 2.7 and 2.8 we can appreciate that the number of possible polypeptide folding possibilities must be enormous (one of the authors of this book once heard Sydney Brenner remark that there are possibly as many folding possibilities as there are proteins!). The chances of misfolded protein production must be high, so what

special circumstances are required to generate the relatively uncommon prion disease? Finding the answers to such questions are currently the objectives of active research.

2.6.9 Cell Culture

This is the process by which cells, mainly mammalian, are grown under controlled *ex vivo* conditions. Applications of cell culture include the manufacture of viral vaccines and many products of biotechnology. Biological products produced by recombinant DNA (rDNA) technology in animal cell cultures include enzymes, synthetic hormones, antibodies and anti-cancer agents.

Cells to be cultured can be obtained in several ways. White cells purified from blood are capable of growth in culture (human red blood cells do not possess a nucleus containing the DNA required for cell replication). Mononuclear cells can be released from soft tissues by enzymatic digestion with enzymes, such as collagenase, trypsin, or pronase, which break down the extracellular matrix. Alternatively, in a method known as explant culture, pieces of tissue can be placed in growth media and the cells that grow out are available for culture. Cells that are cultured directly from an organism are known as primary cells. With the exception of some derived from tumors, most primary cell cultures have limited lifespan. After a certain number of population doublings, cells undergo the process of senescence and stop dividing, while generally retaining viability. An established or immortalised cell line has acquired the ability to proliferate indefinitely, either through random mutation or by deliberate genetic modification.

Cells are grown and maintained at an appropriate temperature and gas mixture (typically, 37 °C, 5% CO₂ for mammalian cells) in a cell incubator. Culture conditions vary widely for each cell type, and variation of conditions for a particular cell type can result in different phenotypes being expressed. Aside from temperature and gas mixture, the most commonly varied factor in culture systems is the growth medium. Recipes for growth media can vary in pH, glucose concentration, growth factors, and the presence of other nutrients. The growth factors used to supplement media are often derived from animal blood, such as calf serum. The culture medium is also generally supplemented with antibiotics, fungicides, or both to inhibit contamination.

Cells can be grown in *suspension* or as *adherent* cultures. Some cells naturally live in suspension, without being attached to a surface, such as cells that exist in the bloodstream. There are also cell lines that have been modified to be able to survive in suspension cultures so that they can be grown to a higher density than adherent conditions would allow. Adherent cells require a surface, such as tissue culture plastic, which may be coated with extracellular matrix components to increase adhesion properties and provide other signals needed for growth and differentiation. Most cells derived from solid tissues are adherent. Another type of adherent culture is *organotypic culture* which involves growing cells in a three-dimensional environment as opposed to two-dimensional culture dishes. This 3-D culture system is biochemically and physiologically more similar to *in vivo* tissue, but is technically challenging to maintain because of factors such as limited diffusion of chemicals to and from the innermost cells.

As cells reach confluence, they must be subcultured or passaged. Failure to subculture confluent cells results in reduced mitotic index and eventually cell death. The first step in

subculturing monolayers is to detach cells from the surface of the primary culture vessel by trypsinisation or mechanical means. The resultant cell suspension is then subdivided or reseeded into fresh cultures. Secondary cultures are checked for growth, fed periodically, and may be subsequently subcultured to produce tertiary cultures. The time between passaging cells depends on the growth rate and varies with the cell line.

As cells generally continue to divide in culture, they generally grow to fill the available area or volume. This can generate the following issues:

- nutrient depletion in the growth media;
- accumulation of dead (apoptotic/necrotic) cells;
- cell-to-cell contact can stimulate cell cycle arrest, causing cells to stop dividing, and is known as contact inhibition or senescence;
- cell-to-cell contact can stimulate cellular differentiation.

Amongst the common manipulations carried out on cells in culture are media changes, passaging cells, and transfecting cells. These are generally performed using tissue culture methods that rely on sterile techniques, which aim to avoid contamination with bacteria, yeast, or other cell lines. Manipulations are typically carried out in a biosafety hood or laminar flow cabinet to exclude contaminating microorganisms. Antibiotics (e.g. penicillin and streptomycin) and antifungals can also be added to the growth media. As cells undergo metabolic processes, acid is produced and the pH decreases. Often, a pH indicator is added to the medium in order to measure nutrient depletion. Passaging (also known as subculture or splitting cells) involves transferring a small number of cells into a new vessel. Cells can be cultured for a longer time if they are split regularly, as it avoids the senescence associated with prolonged high cell density. Suspension cultures are easily passaged with a small amount of culture containing a few cells diluted in a larger volume of fresh media. For adherent cultures the cells first need to be detached, using a mixture of trypsin and EDTA, for example. A small number of detached cells can then be used to seed a new culture. Another common method for manipulating cells involves the introduction of foreign DNA by transfection. This is often performed to cause cells to express a protein of interest. DNA can also be inserted into cells using viruses.

2.6.10 Tissue Engineering

This is the interdisciplinary field that employs cell biology, engineering, and materials science to repair or replace portions of diseased tissues (e.g. bone, cartilage, blood vessels, bladder, etc.). The term *regenerative medicine* is also used in this context, but in this case, emphasis is placed on the use of *stem* cells to produce tissues. There are three main strategies:

1. Design and grow human tissues *in vitro* for later implantation. For example, skin graft replacements, for treatment of burns, are grown in tissue culture.
2. Implantation of devices, which may or may not contain cells, that induce the regeneration of functional human tissues. For example, growth factors may be used to assist in biomaterial-guided tissue regeneration, or polymers can be assembled into three-dimensional configurations, to which cells attach and grow to reconstitute tissues.

3. Development of external devices containing human tissues designed to replace the function of diseased internal tissues. This involves establishing primary cell-lines, placing them on or within structural matrices and implanting them inside the body.

Examples include artificial:

- bladders
- bone
- bone marrow
- cartilage (using chondrocytes for knee repair)
- liver (using hepatocytes)
- pancreas (using islet of Langerhans cells (beta-cells) to produce and regulate insulin for control of diabetes)
- skin (using fibroblasts).

Cells are often categorised by their source:

- *Autologous* cells are obtained from the same individual to whom they will be reimplanted. This reduces problems associated with rejection and pathogen transmission, but is not possible for patients suffering from a genetic disease.
- *Allogenic* cells come from the body of a donor of the same species.
- *Xenogenic* cells are those isolated from individuals of another species. In particular, animal cells have been used quite extensively in experiments aimed at the construction of cardiovascular implants.
- *Syngenic* or *isogenic* cells are isolated from genetically identical organisms, such as twins, clones, or highly inbred research animal models.
- *Primary* cells are from an organism.
- *Secondary* cells are from a cell bank.

Cells are often implanted or *seeded* into an artificial structure (scaffold) capable of supporting three-dimensional tissue formation. Scaffolds usually serve at least one of the following purposes:

- allow cell attachment and migration;
- deliver and retain cells and biochemical factors;
- enable diffusion of vital cell nutrients and expressed products;
- exert certain mechanical and biological influences to modify the behaviour of the attached cells.

To achieve the goal of tissue reconstruction, scaffolds must meet some specific requirements. A high porosity and an adequate pore size are necessary to facilitate cell seeding and diffusion throughout the whole structure of both cells and nutrients. Biodegradability is often an essential factor since scaffolds should preferably be absorbed by the surrounding tissues without the necessity of surgical removal. The rate at which degradation occurs has to coincide as much as possible with the rate of tissue formation. This means that while cells are fabricating their own natural matrix structure around themselves, the scaffold is able to

provide structural integrity within the body and eventually it will break down leaving the newly formed tissue to take over the mechanical load. The ability to be able to inject scaffold material into a tissue is also important for some clinical uses.

Many different materials (natural and synthetic, biodegradable and permanent) have been investigated. Examples of these materials are collagen and some polyesters. A commonly used synthetic material is polylactic acid (PLA) – a polyester which degrades within the human body to form lactic acid, a naturally occurring chemical which is easily released from the body. Similar materials are polyglycolic acid (PGA) and polycaprolactone (PCL). The degradation mechanism for these materials is similar to that of PLA, but PGA exhibits a faster rate and PCL a slower rate of degradation compared to PLA.

Scaffolds may also be constructed from natural materials. Examples include collagen or fibrin, and polysaccharidic materials such as chitosan. Glycosaminoglycans (GAGs) have also proved suitable in terms of cell compatibility, but some issues with potential immunogenicity still remain. Amongst GAGs, hyaluronic acid in combination with cross-linking agents (e.g. glutaraldehyde) has also been used as scaffold material. Chemically functionalised groups incorporated into the scaffolds may also be useful in the delivery of small molecules (drugs) to specific tissues.

2.6.10.1 Stem Cells

Stem cells are undifferentiated cells with the ability to divide in culture and give rise to different forms of specialised cells. According to their source, stem cells are divided into *adult* and *embryonic* stem cells – the first class being *multipotent* and the latter mostly *pluripotent*. Multipotent progenitor cells have the potential to give rise to a limited number of different cell lineages. An example of a multipotent stem cell is a haematopoietic cell, a blood stem cell that can develop into several types of blood cells but cannot develop into muscle or brain cells, for example. Pluripotency is the ability of the human embryonic stem cell to differentiate or become almost any cell type in the body. Some cells are *totipotent* in the earliest stages of the embryo. Totipotency is the ability of a single cell to divide and produce all the differentiated cells in an organism, including extra-embryonic tissues such as the placenta. While there is still a large ethical debate related with the use of embryonic stem cells, there is active research and clinical studies directed towards using stem cells to repair diseased or damaged tissues, or to grow new organs.

2.6.11 Cell–Cell Communication

Most cells in multicellular organisms emit and receive chemical signals. They do this to organise themselves into a coordinated system. **A cell usually receives a chemical signal by the binding of a signalling molecule onto a receptor protein on the cell membrane surface.** This binding activates one or more sequences of *intracellular signalling proteins* that are directed towards influencing the activity of target *effector proteins* inside the cell. The types of effector protein are diverse. They can be a component of the cytoskeleton that influences the shape or motility of a cell, or an enzyme involved in a metabolic pathway, the regulation of gene expression, or operation of an ion channel, for example. The different ways an extracellular signal protein can activate signalling pathways within a cell are depicted in Figure 2.22.

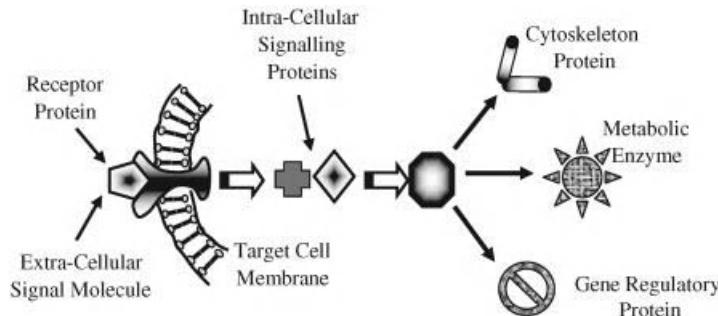


Figure 2.22 This binding of a signalling protein to a membrane receptor protein can activate a sequence of intracellular signalling proteins that are designed to influence the activity of a target effector protein inside the cell. Depending on the type of effector protein, this can result in a change of the shape or motility of a cell, alter cell metabolism or the regulation of gene expression, for example.

The chemical signals secreted from a signalling cell and communicated to another cell are also diverse, and include peptides, amino acids, proteins, nucleotides, proteins, steroids, or hormones, as well as dissolved gases. Some signalling molecules, such as proteins and carbohydrates, are synthesised in the endoplasmic reticulum of the signalling cell and then processed further by its Golgi complex to enable its release by exocytosis into the extracellular medium. Other signalling molecules are small enough to diffuse through the membrane of the signalling cell, to be either released into the extracellular medium or to remain attached to the outer surface of the signalling cell. Signalling molecules that remain attached to the outer surface of the signalling cell can only perform their function through contact with the receiving cell. Most of the signal molecules released into the extracellular medium are hydrophilic and, as shown in Figure 2.23, bind to cell-surface receptor proteins that span across the membrane of the receiving cell. As is also shown in Figure 2.23, small hydrophobic signal molecules have to be transported by proteins to the target cell, where they can then diffuse across the cell membrane to interact directly with an internal effector protein.

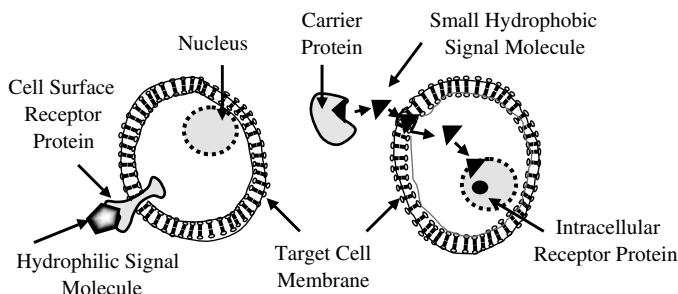


Figure 2.23 Hydrophilic signal molecules released into the extracellular medium can bind directly to cell-surface receptor proteins, but small hydrophobic signal molecules are transported by proteins to the target cell, where they can then diffuse across the cell membrane to interact directly with an internal effector protein.

The concentration of molecular signals in the extracellular medium is very low, typically around 0.1 nM or smaller. To compensate for this low concentration the binding of a signal molecule to its receptor on the target cell is very specific, and is usually characterised by a high affinity constant value exceeding 10^8 litres/mole. (The affinity constant is also known as the association or equilibrium constant K_{eq} described in Chapter 1).

There are four main types of signalling processes between cells:

1. Contact-Dependent

As its name implies, contact-dependent signalling between cells can only occur if, as shown in Figure 2.24, there is direct membrane–membrane contact between the cells. This type of signalling is important in the early stages of an organism's development, and in immune responses. For example, a type of phagocyte known as a dendritic cell possesses a large number of receptors on its membrane surface that can recognise invading microbes or their products in the body and then ingest them. This activates the dendritic cell to produce proteins on its surface (we say it *expresses* the microbial antigen on its surface) and to migrate to a nearby lymph node. Through contact with the small subset of T cells that express a receptor for the microbial antigen expressed on the dendritic cell surface, the T cells are activated to proliferate and then migrate to the infection site to help eliminate the infection, either by killing infected cells or by activating macrophages that then phagocytose the microbes. Some of the activated T cells remain in the lymph node, and can activate B cells to produce antibodies against the microbial antigens. Whereas the influence of activated B cells can extend over long distances, through release of their manufactured antibodies into the bloodstream, activated T cells have to make direct contact to other cells.

2. Synaptic

Nerve cells and neurons are the most sophisticated of those able to perform long-range communication with other cells. They typically possess long axons that terminate at specialised signal transmission sites known as synapses. Neurons can be activated by stimuli from other neurons or by an externally generated event, and when this occurs they send action

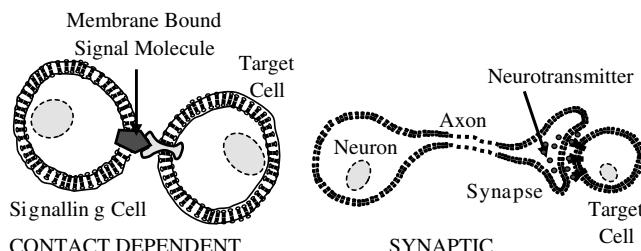


Figure 2.24 Left: Contact-dependent signalling between cells can only occur if there is direct membrane–membrane contact between the cells. This type of signalling is important in the early stages of an organism's development, and in immune responses. Right: When an action potential reaches the synapse at the end of a neuron's axon, chemical signals called neurotransmitters are secreted and diffuse quickly to their target cell across a narrow synaptic gap. Synaptic signalling is efficient in terms of the speed and repetition rates at which signals can be transmitted between cells.

potentials at speeds up to 100 m/s along their axons. When an action potential reaches the synapse at the end of an axon, chemical signals called neurotransmitters are secreted and diffuse quickly and specifically to their target cell across a synaptic gap of width less than 100 nm. This is shown schematically in Figure 2.24. The concentration of a secreted neurotransmitter (e.g. acetylcholine) is quite high at around 0.5 mM, and so the receptors on the postsynaptic target cell need possess only a relatively low affinity for their target signal chemical. This aids the quick dissociation of the signalling chemical from its target cell, after which it can either be destroyed by a specific enzyme or recycled after capture by a membrane transport protein. Synaptic signalling is therefore very efficient in terms of the speed and repetition rates at which signals can be transmitted between cells. The electrical activity of neurons is described further in Chapter 3.

3. Paracrine

In this type of cell signalling, often employed as developmental cues made by embryonic cells, secreted signal molecules only affect target cells in close proximity to the signalling cell (Figure 2.25). To ensure that paracrine signals only act locally, their long-range diffusion is prevented through capture by polysaccharide chains present in the extracellular matrix or attached to neighbouring cells. Molecules known as *antagonists* are also secreted in order to block the signal's activity, and they do this by either binding directly to the signalling molecule or to receptor sites on neighbouring cells. The signalling action may in fact be so local as to only affect the signalling cell itself. This is known as *autocrine signalling*, and is a strategy used by cancer cells, for example, to defend their own survival and to stimulate cell division.

4. Endocrine

In this case the signalling cells secrete signal molecules into the extracellular fluid, where they can diffuse over long distances to their target cells. As shown in Figure 2.25, a good example of this is the secretion of hormones into the bloodstream, which can carry them to target cells over the entire body.

2.6.11.1 Gap Junctions

Gap junctions are water-filled channels (diameter ≤ 1.5 nm) formed by proteins that directly connect the cytoplasms of adjacent cells. They commonly occur in connective tissues,

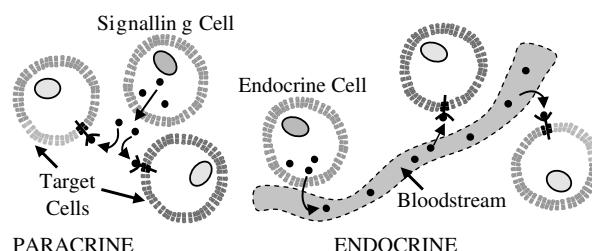


Figure 2.25 In paracrine signalling the secreted signal molecules only affect target cells in close proximity to the signalling cell. This is often used as developmental cues made by embryonic cells to each other. Right: An example of endocrine signalling is the secretion of hormones from an endocrine cell into the bloodstream, which carries them to target cells over the entire body.

epithelial (skin) and liver cells, as well as for some types of neurons, for example. Gap junctions allow cell neighbours to communicate both chemically and electrically with each other. This is required in the liver, to coordinate the response to nerve signals that connect directly to only a subpopulation of the liver cells. Small molecules (≤ 1000 daltons) such as inorganic ions, sugars, amino acids, nucleotides, vitamins (but not proteins and nucleic acids) can pass readily through gap junctions. Microelectrodes inserted into cells connected by gap junctions reveal that they are electrically connected. Defective gap junctions can lead to such health problems as congenital deafness and peripheral nerve dysfunction.

2.7 Summary of Key Concepts

Fatty acids in the form of phospholipids can spontaneously form bilayer structures that are used to construct biomembranes, such as the cytoplasmic membrane that separates a cell from its external environment or the membranes of organelles inside a cell. Carbohydrates in the form of sugars are also used as a source of energy by cells and tissues. The importance of amino acids to the cell comes from their role in making proteins, which are polymers of amino acids covalently linked together by peptide bonds into a long polypeptide chain. Polypeptide chains fold into a three-dimensional structure to form proteins. A protein's function is determined by its three-dimensional structure, which in turn is determined by the linear sequence of the amino acids in its polypeptide chain. Instructions for the assembly of the amino acid sequence are coded by the linear sequence of nucleotides of the nucleic acid DNA. DNA in cells of higher organisms takes the form a double helix, in which each polynucleotide strand can act as a template for the synthesis of the other strand. The process of DNA replication results in a copying of the genetic information from cell to daughter cell and from parents to their offspring.

Selected portions of the DNA are transcribed into several types of RNA by enzymes called RNA polymerases. This transcription process produces single strands of RNA complementary to one strand of DNA – with uracil (U) replacing the thymine (T) of DNA. Some types of RNA can fold up into structures that can act as catalysts. Single strands of messenger RNA (mRNA) contain coding information, obtained from the DNA template, in the form of a sequence of tri-nucleotide units called codons. Each codon specifies a single amino acid to be translated into a polypeptide chain by the interaction of the mRNA strand with a ribosome in the cytoplasm. Transfer RNA (tRNA) mediates this process and provides the corresponding amino acid. The central dogma of molecular biology, that *DNA makes RNA makes Protein*, implies that the sequential structural information stored in a protein by its amino acid sequence cannot be transferred to another protein or to a nucleic acid. The normal flow of biological information is that DNA information can itself be replicated, and then transcribed into messenger RNA from whence it is translated into amino acids to form proteins.

The cell is the structural and functional unit of all known living organisms. There are two basic types of cell, those that do not possess a nucleus (prokaryotes) and those that do have a nucleus (eukaryotes). The nucleus serves to concentrate the DNA-related functions so as to increase their efficiency, which is not necessary for the much smaller prokaryotic cell because the internal reacting materials are relatively close together and many of its biochemical functions are less complicated. Prokaryotic and eukaryotic cells differ significantly in their cell division schemes. For example, the condensation and decondensation of the DNA that occurs in eukaryotic cells during mitosis does not occur at all in prokaryotes. Both

cell types possess a plasma membrane that encloses their contents and serves as a semi-porous barrier to the outside environment. Small molecules, such as oxygen, carbon dioxide, and water, are able to pass freely across this membrane, but the passage of larger molecules (e.g. amino acids and sugars) is carefully regulated. Unlike most prokaryotes, eukaryotic cells also contain extensive internal membranes, which enclose subcellular structures called organelles.

Blood and its cellular components perform important functions, such as the transport of oxygen and nutrients to tissues and providing protection against disease agents, through the immune response provided by lymphocytes and the phagocytic activity of leukocytes, for example. Most cells in multicellular organisms emit and receive chemical signals, so as to organise themselves into a coordinated system. The binding of a signalling molecule by the target cell can activate intracellular signalling proteins, that in turn can influence target effector proteins inside the cell, to influence its shape, motility, metabolism or gene expression, for example.

References

- [1] Alberts, B., Johnson, A. Lewis, J. et al. (2007) Chapter 2 (Tables 2.3 & 2.4), in *Molecular Biology of the Cell*, 5th edn, Garland Science.
- [2] Mulder, G.J. (1838) Zusammensetzung von Fibrin, Albumin, Leimzucker, Leucin u.s.w. *Annalen der Pharmacie*, **28**, 73–82.
- [3] Vickery, H.B. (1950) The origin of the word protein. *Yale Journal of Biology and Medicine*, **22**, 387–393.
- [4] Holmes, F.L. (1963) Elementary analysis and the origins of physiological chemistry. *ISIS*, **54**, 50–81.
- [5] Vickery, H.B. and Schmidt, C.L.A. (1931) The history of the discovery of the amino acids. *Chemical Reviews*, **9**, 169–318.
- [6] Vickery, H.B. (1972) The history of the discovery of the amino acids II. A review of amino acids described since 1931 as components of native proteins. *Advances in Protein Chemistry*, **26**, 81–171.
- [7] Ramachandran, G.N. and Sasickharan, V. (1968) Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, **23**, 283–437.
- [8] Watson, J.D. and Crick, F.H.C. (1953) A structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
- [9] Judson, H.F. (1979) *The Eighth Day of Creation*, Simon & Schuster, New York.
- [10] Harold, F.M. (2001) *The Way of the Cell*, Oxford University Press.
- [11] Guerra, C.F., Bickelhaupt, F.M., Snijders, J.G. and Baerends, E.J. (2000) Hydrogen bonding in DNA base Pairs: reconciliation of theory and experiment. *Journal of the American Chemical Society*, **122**, 4117–4128.
- [12] Kornberg, A. and Baker, T.A. (2005) *DNA Replication*, 2nd edn, University Science Books, Sausalito, Ca.
- [13] Kornberg, A. (1989) *For the Love of Enzymes: The Odyssey of a Biochemist*, Harvard University Press.
- [14] Joyce, G.F. (2002) The antiquity of RNA-based evolution. *Nature*, **418**, 214–221.
- [15] Costanzo, G., Pino, S., Ciciriello, F. and Di Mauro, E. (2009) Generation of long RNA chains in water. *Journal of Biological Chemistry*, **284** (48), 33206–33216.
- [16] Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
- [17] Greene, W.C. (1993) AIDS and the immune system. *Scientific American*, **269** (3), 99–105.
- [18] Lee, K.L., Kozono, D., Remis, J. et al. (2005) Structural basis for conductance by the archaeal aquaporin AqpM at 1.68 Å. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (52), 18932–18937.

Further Readings

- Alberts, B., Johnson, A. Lewis, J. et al. (2007) Chapter 2, in *Molecular Biology of the Cell*, 5th edn, Garland Science.
Bauer, W.R., Crick, F.H.C. and White, J.H. (1980) Supercoiled DNA. *Scientific American*, **243** (1), 118–133.
Becker, W.M., Reece, J.M. and Peonie, M.F. (1995) *The World of the Cell*, 3rd edn, Benjamin/Cummings Publ. Co., Redwood City, CA.
Capra, J.D. and Edmonson, A.B. (1977) The antibody combining site. *Scientific American*, **236** (1), 50–59.

- Crick, F.H.C. (1966) The genetic code III. *Scientific American*, **215** (4), 55–62.
- Dickerson, R.E. (1983) The DNA helix and how it is read. *Scientific American*, **249** (6), 94–111.
- Crick, F. (1988) *What Mad Pursuit: A Personal View of Scientific Discovery*, Basic Books Inc., New York.
- Doolittle, R. (1985) Proteins. *Scientific American*, **253** (4), 88–99.
- Judson, H.F. (1979) *The Eighth Day of Creation*, Simon & Schuster, New York (reprinted, Cold Spring Harbor Laboratory Press, New York, 1996).
- Lake, J.A. (1981) The ribosome. *Scientific American*, **245** (2), 84–97.
- lodish, H.F. and Rothman, J.E. (1978) The assembly of cell membranes. *Scientific American*, **240** (1), 48–63.
- Nelson, D.L. and Cox, M.M. (2009) Chapters 2, 5 and 6, in *Lehninger Principles of Biochemistry*, 5th edn, W.H. Freeman.
- Perutz, M.F. (1964) The hemoglobin molecule. *Scientific American*, **211** (5), 64–76.
- Perutz, M.F. (1978) Hemoglobin structure and respiratory transport. *Scientific American*, **239** (6), 92–125.
- Rich, A. and Kim, S.H. (1977) The three-dimensional structure of transfer RNA. *Scientific American*, **238** (1), 51–62.
- Sharon, N. (1980) Carbohydrates. *Scientific American*, **243** (5), 90–116.
- Tanford, C. and Reynolds, J. (2001) *Nature's Robots: A History of Proteins*, Oxford University Press.
- Watson, J.D. (1968) *The Double Helix*, Simon & Schuster, New York, (Touchstone Edition, 2001).

3

Basic Biophysical Concepts and Methods

3.1 Chapter Overview

This chapter describes some of the concepts that have been developed to describe and examine the biophysical properties of biological systems at the molecular and cellular level. A comprehensive coverage would include the techniques that have been developed to describe the molecular folding and conformational transitions involved in the functioning of enzymes and nucleic acids; the forces that stabilise the structures of proteins in membranes; cable theory analyses of axons and muscle fibres, for example. Some of the material covered in this chapter will go some way to provide an introduction to such topics, but the main emphasis is to provide an understanding of the main biophysical concepts of relevance to engineers and scientists working in areas such as the development and applications of biosensors, microfluidics and lab-on-chip technologies.

After reading this chapter readers will gain a basic understanding of:

- (i) electrostatic forces acting on ions in solution, in membranes and at charged surfaces, and the concept of the Debye screening length;
- (ii) the principal modes of transport (diffusion, osmosis, active) of ions and molecules across membranes;
- (iii) hydrophobic and hydration forces acting on bioparticles;
- (iv) the concepts of osmolarity, tonicity and osmotic pressure of relevance to cells;
- (v) electrochemical gradients and ion distributions across membranes;
- (vi) osmotic properties of cells;
- (vii) the passive and active electrical properties of membranes and cells;
- (viii) membrane equilibrium potentials, action potentials and ion channel conduction;
- (ix) voltage and patch clamp techniques for investigating membrane ion channels;
- (x) biological applications of electrokinetic effects (electrophoresis, electro-osmosis, dielectrophoresis) and electrowetting on dielectric.

3.2 Electrostatic Interactions

3.2.1 Coulomb's Law

Coulomb's law, also known as Coulomb's inverse square law, describes the electrostatic interaction between electrically charged particles, and is named after the French scientist Charles Augustin de Coulomb. The two most significant charged particles are the proton and electron, which carry positive and negative charges of 1.6×10^{-19} coulombs, respectively. Charged atoms or molecules in solution are called ions. Positively charged ions will migrate in an electric field to the negative electrode (the cathode) and are called cations, whilst negatively charged ions are called anions and will migrate to the positively charged anode.

Coulomb's law states that:

The magnitude of the Electrostatics force of interaction between two point charges is directly proportional to the scalar multiplication of the magnitudes of charges and inversely proportional to the square of the distances between them.

The idealised point charges are considered to be small in size compared to their separation distance r , and the force F (units of newton) acting simultaneously on such point charges q_1 and q_2 is given by

$$F = K \frac{q_1 q_2}{r^2}. \quad (3.1)$$

A positive value for F implies that the force is repulsive, while a negative force implies it is attractive. In Equation (3.1) the constant K is the coulomb force constant, which in SI units is given by

$$K = \frac{1}{4\pi\epsilon_0}. \quad (3.2)$$

and is related to the defined electromagnetic properties of free space (vacuum) by the equation derived by James Clerk Maxwell:

$$c^2 = \frac{1}{\mu_0 \epsilon_0}, \quad (3.3)$$

where c is the speed of electromagnetic radiation (light) in vacuum ($2.99\ 792\ 458 \times 10^8 \text{ m s}^{-1}$) and μ_0 is the magnetic permeability of free space, defined as $4\pi \times 10^{-7} \text{ H m}^{-1}$. From Equations (3.2) and (3.3) the permittivity of free space ϵ_0 is derived to be $8.85\ 418\ 782 \times 10^{-12} \text{ F m}^{-1}$ and the Coulomb force constant as $8.9\ 875\ 517\ 871\ 764 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$. In electrostatic units (esu) and gaussian units (statcoulomb) the unit charge is defined in such a way that the Coulomb constant K in Equation (3.1) is a dimensionless quantity equal to 1. The inverse square law described by Equation (3.1) results from the force field due to an isolated point charge being uniform in all directions and attenuating with distance r at the same rate as the expansion of the surface area ($4\pi r^2$) of a sphere centred on the point charge.

Equation 3.1 represents the scalar form of Coulomb's law. To obtain both the magnitude and the direction of the force we require the vector-form:

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2 (r_1 - r_2)}{|r_1 - r_2|^3} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2} \hat{r}_{21}. \quad (3.4)$$

The direction of the force F is given by the unit vector \hat{r}_{21} , lying parallel with the line directed from charge q_2 to q_1 . If the two charges are alike (either both positive or both negative) the product $q_1 q_2$ in Equation (3.4) is positive and charge q_1 will experience a repulsive force of direction given by \hat{r}_{21} . If the two charges have opposite signs, the product $q_1 q_2$ is negative and an attractive force will be felt by charge q_1 directed along $-\hat{r}_{21}$. These vector relationships are shown in Figure 3.1.

Coulomb's law can be applied to describe the forces that hold together the atomic constituents in a molecule, as well as the forces that bind atoms and molecules together to form solids and liquids. It can also be used to describe the force between the positively charged nucleus and each of the negatively charged electrons in an atom. Equations 3.1 and (3.4) assume that the two charges are stationary or moving slowly. This is known as the electrostatic approximation. Rapidly moving charges create magnetic fields that alter the force on the two charges.

Two charges are shown in Figure 3.1, but Coulomb's law can be extended to derive the force acting on a point charge as a result of its electrostatic interaction with any number of other point charges. This is achieved using the law of superposition by the vector addition of the individual electrostatic forces that act alone on that point charge. The force vector F acting on a small point charge q by a system of N discrete charges is thus given by:

$$F = \frac{q}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i(r - r_i)}{|r - r_i|^3} = \frac{q}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i}{R_i^2} \hat{R}_i, \quad (3.5)$$

where q_i and r_i are the magnitude and position respectively of the i^{th} charge, \hat{R}_i is a unit vector pointing in the direction from charge q_i to charge q , and R_i is the magnitude of the distance of separation between charges q_i and q . The resultant force vector F lies parallel to the electric field vector at the position of the point charge q , with that point charge removed.

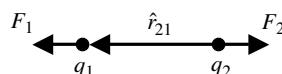


Figure 3.1 Two point charges, distance r apart, experience a force F given by Coulomb's law. If the two charges are of the same polarity, charge q_1 will experience a repulsive force directed along the unit vector \hat{r}_{21} . If the two charges are of opposite polarity the direction of the attractive force acting on charge q_1 is given by $-\hat{r}_{21}$.

The work W required (at constant temperature and pressure) to bring a charge q_1 from infinity to a distance r from charge q_2 is given by:

$$W = - \int_{\infty}^r F dr = - \frac{q_1 q_2}{4\pi\epsilon_0} \int_{\infty}^r \frac{1}{r^2} dr = \frac{q_1 q_2}{4\pi\epsilon_0} \frac{1}{r}. \quad (3.6)$$

Performing work of this amount increases by the same amount the potential energy U of the system of two charges. The potential energy is therefore given by:

$$U = \frac{q_1 q_2}{4\pi\epsilon_0 r} = q_1 \phi \quad (3.7)$$

with

$$\phi = \frac{q_2}{4\pi\epsilon_0 r} \frac{1}{r}. \quad (3.8)$$

The parameter ϕ represents the Coulomb *potential* arising from the charge q_2 . The difference between potential energy U (units of Joule) and potential ϕ (units of Volt) should be noted and understood. If the product $q_1 q_2$ in Equation (3.7) is positive then work is required to bring the two charges closer together and the potential energy U is positive. On the other hand, if the product $q_1 q_2$ is negative then work is required to keep the two charges apart and the potential energy assumes a negative value.

The magnitude of the electric field E created by the single point charge q_2 at a distance r from that charge is given by the negative gradient of the potential at that point:

$$E = -\nabla\phi = \frac{1}{4\pi\epsilon_0 r^2} \frac{q_2}{r} \hat{r}. \quad (3.9)$$

Symbol ∇ is the gradient operator, $\hat{i}(\partial/\partial x) + \hat{j}(\partial/\partial y) + \hat{k}(\partial/\partial z)$, with $\hat{i}, \hat{j}, \hat{k}$ unit vectors along axes x, y, z. For a positive point charge ($+q$) the electric field is directed radially away from the location of the charge, and for a negative charge ($-q$) it is directed in the opposite sense towards the charge. The magnitude of the force acting on charge q_1 arising from the electric field created by q_2 is given by:

$$F = q_2 E. \quad (3.10)$$

The values obtained for the potential energy U and potential ϕ using Equations (3.7) and (3.8) assume that a charge, for example an isolated ion, is located in a vacuum. For an ion located in a solution, account must be taken of how easily the molecules surrounding the ion can be electrically polarised. This polarisation will counteract the local electric field E and reduce the magnitudes of U and ϕ . The polarisability of a material is determined by its dielectric permittivity $\epsilon = \epsilon_0 \epsilon_r$, where ϵ_r is known as the relative permittivity and is sometimes also referred to as the dielectric constant (ϵ_r varies with temperature and the ac frequency of an applied electric field, and so is not a constant). For ions located in a medium

other than vacuum, the equations for the potential energy, potential and electric field are modified to the forms:

$$U = \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r} \frac{1}{r} = q_1 \phi \quad (3.11)$$

$$\phi = \frac{q_2}{4\pi\epsilon_0\epsilon_r} \frac{1}{r} \quad (3.12)$$

$$E = \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_2}{r^2} \hat{r}. \quad (3.13)$$

Vacuum is completely unresponsive in terms of its electrical polarisability and so will have $\epsilon_r = 1$. At 298 K and low ac frequencies, water, ethanol and benzene exhibit ϵ_r values of around 78.5, 24, and 2.3, respectively. The relatively high ϵ_r value for water is related to the H₂O molecule possessing a large dipole moment (see Figure 3.2) enhanced by the formation of hydrogen bonds to neighbouring water molecules, as discussed in Chapter 1 (Sections 1.2.4 and 1.3.1). An isolated water molecule consists of two polar O–H bonds, each having a dipole moment of 5×10^{-30} C m (1.51 debye units). The O–H bond length (l) in a water molecule is 95.8 p.m., so that its dipole moment ($= l\delta q$) implies that the magnitudes of $\delta+$ and $\delta-$ defined in Figure 3.2 are each 5×10^{-20} C, corresponding to about one-third of a full electronic charge q . The vector sum of these two dipolar bonds gives a total dipole moment for a water molecule of 6.2×10^{-30} C m (1.855 D). As shown in Figure 3.2 the convention is to direct the dipole moment vector from the negative charge to the positive charge. A dipole will experience a torque when subjected to an external electric field, tending to align the dipole vector with the field to give an orientation of lower potential energy. The orientational polarisability of a water molecule in the bulk water phase is increased as a result of the cooperative reorientations required of its neighbouring hydrogen-bonded water molecules. This increases the effective dipole moment of a water molecule in the condensed phase to around 8.3×10^{-30} C m (~2.5 D) and results in the high relative permittivity exhibited by water. Ethanol possesses a smaller dipole moment (1.69 D) and so exhibits a smaller ϵ_r value. The benzene molecule is symmetrical (see Figure 3.2) and so has a zero dipole moment and

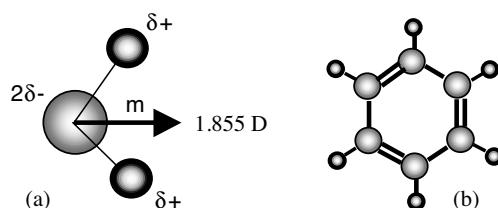


Figure 3.2 (a) The dipole moment m (1.855 debye units) of an isolated water molecule (H₂O) arises from the vector sum of the moments (each 1.51 D) of its two polar OH bonds that form a bond angle of 104.5°. The magnitudes of $\delta+$ and $\delta-$ are each about 0.32 of a full electronic charge. (b) The benzene molecule (C₆H₆) is symmetrical and so has a zero dipole moment.

does not exhibit orientational polarisability. Its relative permittivity value of 2.3 arises solely from electronic polarisations.

Because we are manipulating the charges under the conditions of constant temperature and pressure, it follows from the discussion in Section 1.2.7 of Chapter 1 that in thermodynamic terms the potential energy U given by Equation (3.11) takes the form of Gibbs free-energy. From the second law of thermodynamics, contributions to the free-energy U of a system (including a system of charges) in thermal equilibrium with its surroundings, at constant temperature and pressure, will come from its heat energy (enthalpy H) and entropy S according to the equation (see also Section 1.2.7 of Chapter 1):

$$U = H - TS. \quad (3.14)$$

3.2.2 Ions in Water

In an ionic crystal such as NaCl we can estimate, from the Van der Waals radii values given in Table 1.5, that the closest separation distance between a Na^+ and Cl^- ion is 0.4 nm. From Equation (3.10), and taking the relative permittivity of a sodium chloride crystal as $\epsilon_r \sim 6$, the Coulomb energy of attraction of the ion pair is about 9.6×10^{-20} J (i.e. $23 kT$ at room temperature). When a crystal of NaCl is placed in water the ions at the crystal surface will interface with a medium of relative permittivity $\epsilon_r \sim 80$, and when fully immersed their Coulomb energy will approach a value of -4×10^{-21} J ($1.8 kT$), which is only a little more than the energy of $3 kT/2$ associated with thermal fluctuations and insufficient to result in stabilisation of the ion pair. This is why salt crystals dissolve and dissociate in water.

To understand this dissociation process more clearly we can find the relative contributions of the enthalpy and entropy changes from Equation (3.14). The entropy S term can be approximated as $-\partial U / \partial T$, which from Equation (3.10) gives:

$$S = -\frac{\partial}{\partial T} \left(\frac{q_1 q_2}{4\pi \epsilon_0 \epsilon_r r} \right) = \frac{q_1 q_2}{4\pi \epsilon_0 \epsilon_r^2 r} \frac{\partial \epsilon_r}{\partial T} = U \frac{1}{\epsilon_r} \frac{\partial \epsilon_r}{\partial T}. \quad (3.15)$$

The value of ϵ_r for water at 273 K is 87.9 and to a good approximation between 273 and 315 K exhibits a percentage fall of -0.4% per Kelvin. From Equation (3.15) this gives $S = -4 \times 10^{-3} U$. The value of TS in Equation (3.14) at 298 K is thus $-1.19U$, in other words larger than the total free energy U ! This implies that the attraction of sodium and chloride ions is mainly driven by entropy and not enthalpy. A dominating negative value for TS implies that work has been expanded on the system of ions and water so as to create a more ordered system. The ions themselves are only weakly associated and so this increase in order must be associated with how the water molecules interact with the ions. This interaction involves the torques induced on the water molecules by the interactions of their dipole moments with the electric fields around the charged ions. This restricts the orientational rotations of the water molecules and creates a ‘hydration shell’ of orientated water molecules around an ion, as schematically depicted in Figure 3.3. This reduction of orientational mobility and dipole alignments creates the increase of order responsible for the large negative value we have deduced for ST in Equation (3.14). The reduction of orientational mobility of the water molecules in this ‘hydration shell’ will result in a reduction of the local value of ϵ_r .

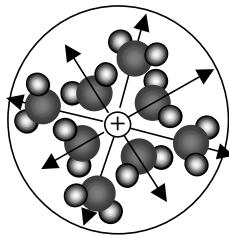


Figure 3.3 The electrostatic interaction between an ion and its surrounding water molecules forces them to align their dipole moments along the field lines created by the ion. This local ordering and rotational mobility of the water molecules represents a significant reduction of entropy of the ion-water system.

At room temperature the effective ϵ_r value for the water structure closest to the ion could be as low as 10, instead of the value ~ 80 for normal bulk water. The actual value for U will therefore be larger than that predicted by Equation (3.11) if the local structuring of water molecules is ignored.

A positively charged ion will orientate a neighbouring water molecule such that the negative component of its dipole is directed towards the ion. This will serve to screen the ion's positive charge and to reduce its Coulomb potential ϕ . Water molecules will orientate in the opposite sense around a negatively charged ion and so also act to screen this charge. This screening is also enhanced by the tendency on average of neighbouring ions to overcome thermal vibrations and to be attracted to a countercharged ion. So, although the interaction between counter ions is sufficiently weak for their salt to dissociate and dissolve in water, the balance between electrostatic forces and thermal agitation is such that on average ions with the same charge will tend to avoid each other and those of opposite charge to spend more time near each other. This weak association of counter ions will increase as their average separation distance decreases (i.e. the salt concentration increases) and this is reflected in the activity coefficient values given in Table 1.3. For example, from Table 1.3 we can deduce that a 10 and 100 mM solution of KCl will exhibit an osmolarity of 180 and 154 mOsm, respectively, rather than the value of 200 mOsm if it were to act as an ideal ionic solution.

3.2.3 The Formation of an Ionic Double Layer

The distribution of ions around a charged particle is determined by the balance between electrostatic forces and thermal agitation. A quantitative description of the distribution of ions in thermal equilibrium in an electrostatic field can be obtained by combining the Poisson equation of electrostatics with the Boltzmann distribution. The Poisson equation generalises the electrostatic Coulomb potential $\phi(r)$ to a volume distribution of charge density $\rho(r)$:

$$\nabla^2 \phi(r) = -\frac{\rho(r)}{\epsilon_0 \epsilon_r}. \quad (3.15)$$

Equation 3.15 reduces to the Coulomb potential described by Equation (3.8) when $\rho(r)$ is a point charge in a uniform dielectric medium. For an arbitrary collection of ions of number

densities (m^{-3}) c_i and valences z_i the charge distribution is given by

$$\rho(r) = q \sum_i z_i c_i(r), \quad (3.16)$$

where q is the charge on an electron. The earliest model, known as the Helmholtz model, describes the distribution of ions at the interface between a charged surface and an electrolyte as a parallel-plate capacitor. One plate of this capacitor contains the charge on the electrified surface and the other plate (known as the Helmholtz plane) contains the ions, of opposite charge polarity to that on the surface, electrostatically attracted to it from the electrolyte. The thickness of this electrical double layer is taken to be the diameter of the ions attracted to the charged surface. If we define σ to be the surface charge density, it is balanced by an equivalent amount of ionic charge of opposite polarity in the solution. This balance can be expressed by the relationship

$$\sigma = - \int_0^\infty \rho(r) dr. \quad (3.17)$$

In the Helmholtz model the counter ionic charge density $\rho(r)$ takes the form of a layer of charges at the charged surface. An obvious oversimplification of the Helmholtz model is that thermal motions of the ions in the electrolyte are ignored. These thermal motions will cause $\rho(r)$ to form a diffused distribution rather than a layer.

The concentration c_i of ions in thermodynamic equilibrium with the electrolyte solution, as a function of their distance r from a charged surface, is related to the electrostatic potential $\phi(r)$ using the Boltzmann distribution as follows:

$$c_i(r) = c_{i\infty} \exp\left(\frac{-qz_i\phi(r)}{kT}\right). \quad (3.18)$$

The parameter $c_{i\infty}$ is the concentration of ions in the bulk solution, far enough away from the charged object that the value of $\phi(r)$ is zero. If the electrolyte is an aqueous sodium chloride solution, for example, this corresponds to equal concentrations $[\text{Na}^+]$ and $[\text{Cl}^-]$ of the sodium and chloride ions. As we move from the bulk solution towards a negatively charged surface, for example, we would find that $[\text{Na}^+]$ increases and $[\text{Cl}^-]$ decreases. Substituting Equation (3.18) into Equation (3.16) gives

$$\rho(r) = q \sum_i z_i c_{i\infty} \exp\left(\frac{-qz_i\phi(r)}{kT}\right). \quad (3.19)$$

Using this result to eliminate $\rho(r)$ from Equation (3.15) we obtain the Poisson-Boltzmann equation:

$$\nabla^2 \phi(r) = - \frac{q}{\epsilon_o \epsilon_r} \sum_i z_i c_{i\infty} \exp\left(\frac{-qz_i\phi(r)}{kT}\right). \quad (3.20)$$

This equation describes the electrical potential $\phi(r)$ at the interface between a charged object and an electrolyte solution, taking into account the screening of this potential by counterions. For example, it can in principle describe the spatial composition of the ionic ‘atmosphere’ around an ion or a charged particle. For such situations, where r is the only

relevant coordinate, the appropriate form of the vector operator ∇^2 involves spherical coordinates. However, Equation (3.20) cannot be solved analytically using spherical coordinates. Depending on the geometry of the system and the boundary conditions, solving Equation (3.20) may require the use of approximations. One such approximation is to assume that the electrostatic interactions of the ions in the solution are weak ones, so that $qz_i\phi(r)/kT \ll 1$. This allows the linear form ($e^x = 1 + x + x^2/2! + \dots$) of the exponential function to be used in Equation (3.20) to give:

$$\nabla^2\phi(r) = -\frac{q}{\epsilon_0\epsilon_r}\sum_i z_i c_{i\infty} \left(1 - \frac{qz_i\phi(r)}{kT}\right). \quad (3.21)$$

A further simplification can be made by noting that for a sufficiently large distance r the electrical potential $\phi(r)$ is zero, and $d\phi(r)/dr$ is also zero. This corresponds to electrical neutrality of the solution, so that

$$q\sum_i z_i c_{i\infty} = 0.$$

Adopting this boundary condition as r tends to infinity, Equation (3.21) can then be written as:

$$\nabla^2\phi(r) = \frac{q^2}{\epsilon_0\epsilon_r kT} \sum_i z_i^2 c_{i\infty} \phi(r). \quad (3.22)$$

Equation 3.22 is referred to as the linear form of the Poisson-Boltzmann equation, and can be written as

$$\nabla^2\phi(r) = \kappa^2\phi(r). \quad (3.23)$$

3.2.3.1 The Debye Screening Length

If the variable $\phi(r)$ in Equation (3.23) is transformed to a variable with units of $1/\kappa$, this equation will contain no parameters. This means that $1/\kappa$ must represent a fundamental unit when considering electrostatic interactions in ionic solutions. From Equations (3.22) and (3.23)

$$\kappa^2 = \frac{q^2}{\epsilon_0\epsilon_r kT} \sum_i z_i^2 c_{i\infty}$$

From which we can determine that $1/\kappa$ has units of length, and is an important parameter known as the Debye length. We can interpret its significance by stating that for distances r shorter than the Debye length the electrostatic interactions will be strong, but for much larger distances the interactions will be weak because of ionic screening. The factor $1/\kappa$ can thus be taken to be the ionic screening distance. Its value is given by

$$1/\kappa = \sqrt{\frac{\epsilon_0\epsilon_r kT}{2q^2 I N_A 10^3}}. \quad (3.24)$$

In this equation N_A is the Avogadro constant ($6.022 \times 10^{23} \text{ mol}^{-1}$); we have converted the ionic density $c (\text{m}^{-3})$ to the ionic strength $I (\text{mol l}^{-1})$ of the solution by assuming a simple monovalent salt ($z_i = 1$) such as NaCl, so that

$$\sum_i z_i^2 c_{i\infty} = 2IN_A 10^3.$$

For an aqueous 10 mM solution of NaCl at 298 K we can calculate $1/\kappa$ to be 3.07 nm (assuming $\epsilon_r = 80$). From Equation (3.24) we note that the Debye length is inversely proportional to the square root of the solution's ionic strength, so that for a 1 M solution it decreases to 0.31 nm. We would expect the ionic screening to increase as the number of ions per unit volume increases. The ionic strength also increases as z_i^2 , so that solutions containing multi-valent salts (e.g. CaCl₂) will be more effective at screening electrostatic interactions.

3.2.3.2 The Gouy-Chapman Equation

If we wish to consider the case of a charged membrane surface, or to approximate the curved surface of a particle as a planar surface, the only important dimension is the distance normal to the surface, which we will take to be the x-direction. In one dimension Equation (3.15) is written as

$$\frac{d^2\phi(x)}{dx^2} = -\frac{\rho(x)}{\epsilon_0 \epsilon_r}, \quad (3.25)$$

so that Equation (3.20) takes the form:

$$\frac{d^2\phi(x)}{dx^2} = -\frac{q}{\epsilon_0 \epsilon_r} \sum_i z_i c_{i\infty} \exp\left(\frac{-qz_i\phi(x)}{kT}\right). \quad (3.26)$$

This form of the Poisson-Boltzmann equation can be solved analytically without converting it to the linear form. For a monovalent salt solution such as NaCl (i.e. $z = \pm 1$) of number density c , and noting that as the distance x tends to infinity $d\phi/dx$ tends to zero, integration of Equation (3.26) gives:

$$\frac{d\phi(x)}{dx} = \sqrt{\frac{2kTc}{\epsilon_0 \epsilon_r}} \cdot (\exp(-q\phi(x)/2kT) - \exp(q\phi(x)/2kT)). \quad (3.27)$$

Using the relationship between the surface charge density σ and the counter charge density $\rho(r)$ given by Equation (3.17), then from Equation (3.25):

$$\sigma = \epsilon_0 \epsilon_r \int_0^\infty \frac{d^2\phi(x)}{dx^2} dx = -\epsilon_0 \epsilon_r \frac{d\phi(0)}{dx}. \quad (3.28)$$

This equation relates the electric field (i.e. the gradient of the electric potential) at the surface to the surface charge density, where we can define $\phi(0)$ to be the surface potential. Substituting Equation (3.27) into Equation (3.28) we obtain the important Gouy-Chapman equation:

$$\sigma = -\sqrt{2\epsilon_0\epsilon_r kTc} \cdot [\exp(-q\phi(0)/2kT) - \exp(q\phi(0)/2kT)]. \quad (3.29)$$

If we adopt the linear approximation for the exponential terms and insert the Debye screening length $1/\kappa$ introduced for Equation (3.24), the Gouy-Chapman equation reduces to the simple form:

$$\sigma = \epsilon_0\epsilon_r\kappa\phi(0). \quad (3.30)$$

Equation 3.30 describes a proportionality between the surface charge σ and the surface potential $\phi(0)$ analogous to the relationship $q=VC$ between charge and voltage for a capacitance C . From Equation (3.30) we can consider the term $\epsilon_0\epsilon_r\kappa$ to be an effective capacitance per unit area, with the Debye length $1/\kappa$ representing the distance between the two charge-carrying plates. This further supports the concept of the charge distribution of ions at the interface between a charged surface and an electrolyte taking the form of an electrical double layer.

Integration of Equation (3.27) and adopting the assumption that $qz_i\phi(r)/kT \ll 1$, together with the linear form of the exponential, we obtain the relationship:

$$\phi(x) = \phi(0) \exp(-x\kappa) \quad (3.31)$$

Equation 3.31 indicates that the electrostatic potential falls exponentially with distance into the electrolyte, reaching a value of $0.37\phi(0)$ at a distance equal to the Debye screening length $1/\kappa$.

3.2.3.3 Stern's Modification of the Gouy-Chapman Equation

The assumptions used in deriving Equation (3.29) do not hold for high surface charge densities and high potential gradients (fields). The predicted concentrations of ions attracted to the charged surface can be unrealistically high, sometimes above the saturation level for a salt. Multivalent salt ions can also be attracted so strongly to the surface as to bind to it. Otto Stern introduced two modifications to the Gouy-Chapman theory, described in detail in the book by Aveyard and Haydon ([1], p. 231). The first modification simply takes account of the fact that an ion cannot get closer to the charged surface than its own radius, and the second modification is to allow for specific binding of ions to the charged surface in what is called the Stern layer. Within a distance of the Debye length other ions form a diffuse layer, and the electrostatic potential falls exponentially as described by Equation (3.31), with the potential $\phi(0)$ being replaced with the value $\phi(a)$ at the interface of the Stern and diffuse layers. These two ion distributions and the corresponding profile of the electrostatic potential are shown in Figure 3.4.

Bedzyk *et al.* [2] have determined the ion distribution in an electrolyte solution in contact with a charged polymerised phospholipid membrane using x-ray standing waves, and found it to qualitatively agree with the Gouy-Chapman-Stern model.

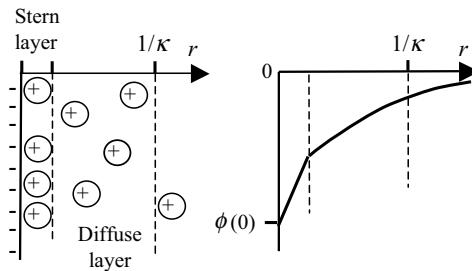


Figure 3.4 The Gouy-Chapman-Stern description of the electrical double layer at a charged surface. On the left, cations are shown strongly bound in the Stern layer at a negatively charged surface. Other cations form a diffuse layer within a distance corresponding to the Debye length $1/\kappa$. As shown on the right the electrostatic potential falls linearly within the Stern layer, and then follows an exponential fall. For distances beyond the Debye length the charged surface is ionically screened from the bulk electrolyte solution.

The charge densities σ_{St} and σ_{Dl} in the Stern layer and the diffuse layer add up to the total charge σ on the surface, and together these charged layers act as two capacitors in series. The total capacitance C is given by:

$$C = \frac{C_{St}C_{Dl}}{C_{St} + C_{Dl}}. \quad (3.32)$$

The contribution to the total capacitance of the Stern layer tends to be unaffected by changes in the ionic strength of the solution. At low ionic strengths the Debye screening length $1/\kappa$ is relatively large so that the effective capacitance $\epsilon_0\epsilon_r\kappa$ of the diffuse layer is low and the total capacitance C given by Equation (3.32) tends to be dominated by the diffuse layer. At high ionic strengths the Debye length is small, C_{Dl} is large and now the total capacitance tends to be dominated by C_{St} .

3.2.3.4 Activity Coefficient of Ions in Solution

We discussed in Chapter 1 (Table 1.3) and Section 3.2.2 of this chapter that the activity coefficient of ions in solution decreases as the ionic concentration increases. We are now in a position to quantify this effect. The distribution of ions around a central ion in solution will have spherical symmetry, and this can be obtained by writing the linearised form of the Poisson-Boltzmann Equation (3.23) in spherical coordinates:

$$\frac{1}{r} \frac{d^2(r\phi(r))}{dr^2} = \kappa^2 \phi(r). \quad (3.33)$$

Integration of this differential equation, after multiplying through by r , gives the result

$$r\phi(r) = A \exp(-r\kappa) + B \exp(r\kappa), \quad (3.34)$$

in which A and B are constants of integration. To satisfy the boundary condition that as r tends to infinity, the potential $\phi(r)$ tends to zero, constant B must equal zero. Thus, from Equation (3.34)

$$\phi(r) = \frac{A \exp(-r\kappa)}{r} \quad (3.35)$$

and the electric field $E(r)$ is given by

$$E(r) = \frac{d\phi(r)}{dr} = \frac{A \exp(-r\kappa)}{r^2} \left(\kappa + \frac{1}{r} \right). \quad (3.36)$$

We can now use the boundary condition that at the surface of an ion of radius a there is no ionic screening of the potential, so that the electric field at the surface of the ion is given by Equation (3.13). For $r = a$, and setting $E(r)$ in Equation (3.36) to the value given by Equation (3.13), we obtain the following value for A :

$$A = \frac{q}{4\pi\epsilon_0\epsilon_r} \left[\frac{\exp(a\kappa)}{1 + a\kappa} \right].$$

Substituting this result into Equation (3.35) gives

$$\phi(r) = \frac{q}{4\pi\epsilon_0\epsilon_r r} \left[\frac{\exp(-\kappa(r - a))}{1 + a\kappa} \right]. \quad (3.37)$$

This result gives the electrostatic potential energy around an ion in solution. Substituting this into Poisson's Equation (3.15) in the form of Equation (3.33) the corresponding counterion charge density around a single ion is given as

$$\rho(r) = \frac{q\kappa^2}{r} \left[\frac{\exp(-\kappa(r - a))}{1 + a\kappa} \right]. \quad (3.38)$$

The charge contained in a cylindrical shell of thickness dr at a distance r from the ion centre is $4\pi r^2 \rho(r) dr$. Plots of this function for two different ionic strengths of NaCl are given in Figure 3.5.

From Figure 3.5 the maximum number density of counterions around an ion occurs at a distance corresponding to the Debye length $1/\kappa$. Beyond this distance the central ion is effectively screened from further electrostatic interactions. Equation 3.37 can thus be separated into two terms – the Coulombic term given by Equation (3.12) and the term that describes the ionic screening by the counterion distribution shown in Figure 3.5:

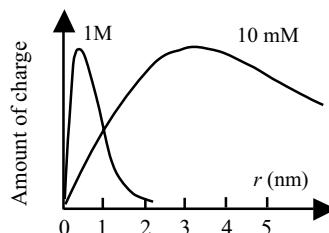


Figure 3.5 Plots of the counterion charge density around a single ion in water, as a function of radial distance r , for two NaCl solutions. The 10 mM solution at 298 K has a Debye length of 3.07 nm, which reduces down to 0.31 nm for the 1 M solution.

$$\phi(r) = \frac{q}{4\pi\epsilon_0\epsilon_r r} + \frac{q}{4\pi\epsilon_0\epsilon_r r} \left[\frac{\exp(-\kappa(r-a))}{1+a\kappa} - 1 \right].$$

Following the logic described for Equation (3.6) we can calculate the work W done on the ionic cloud that screens the central ion (the right-hand term of the above equation) on incrementally charging the central ion (the left-hand term) by integrating the right-hand term to give:

$$W = \frac{1}{4\pi\epsilon_0\epsilon_r a} \left(\frac{1}{1+a\kappa} - 1 \right) \int_0^q q dq = -\frac{q^2}{8\pi\epsilon_0\epsilon_r} \left(\frac{1}{a+1/\kappa} \right). \quad (3.39)$$

To correct for ionic screening, this work W is added to the Gibbs free energy G :

$$G = G^\circ + RT \log_e c$$

of an ideal solution of ionic concentration c . The activity coefficient γ for a single ion is defined as:

$$kT \log_e \gamma = W.$$

At relatively low ionic concentration the Debye screening length $1/\kappa$ is much larger than the ion radius a , so that from Equation (3.39) we can approximate the activity coefficient to be

$$\log_e \gamma = -\frac{q^2 \kappa}{8\pi\epsilon_0\epsilon_r}. \quad (3.40)$$

Equation 3.40 is known as the Debye-Hückel limiting law for the activity coefficient of an ionic solution. From Equation (3.24) we see that κ is proportional to the square root of the ionic strength of the solution. Equation 3.40 thus predicts the expected result that the activity coefficient approaches zero as the ionic strength approaches zero. All electrolyte solutions are ideal at infinite dilution of their ions.

3.2.4 Ion–Dipole and Dipole–Dipole Interactions

The interaction between an ion of charge Q and a neighbouring water molecule is shown in Figure 3.6. The dipole moment of the water molecule is represented as two equal and opposite point charges, $+δq$ and $-δq$, separated by a distance d of the order 0.1 nm.

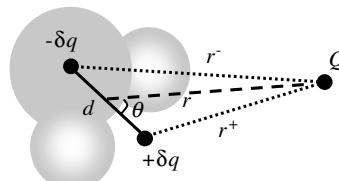


Figure 3.6 The interaction between a charge Q and a water dipole. The dipole is represented as two equal and opposite charges separated by a distance d .

The electrostatic interaction between charge Q and the dipole will be the sum of the Coulomb interaction between Q and $+\delta q$ and that between Q and $-\delta q$:

$$U = \frac{\delta q \cdot Q}{4\pi\epsilon_0\epsilon_r r^+} - \frac{\delta q \cdot Q}{4\pi\epsilon_0\epsilon_r r^-},$$

where the distances r^+ and r^- are as defined in Figure 3.6. For the case where d (~ 0.1 nm) is small compared to r , this equation can be approximated through the following steps:

$$\begin{aligned} U &= \frac{1}{4\pi\epsilon_0\epsilon_r} \left(\frac{\delta q \cdot Q}{(r + 1/2(d \cos \theta))} - \frac{\delta q \cdot Q}{(r - 1/2(d \cos \theta))} \right) \\ &\approx -\frac{1}{4\pi\epsilon_0\epsilon_r} \frac{\delta q \cdot Q d \cos \theta}{(r^2 - (d^2/4) \cos^2 \theta)} \approx -\frac{m \cdot Q \cos \theta}{4\pi\epsilon_0\epsilon_r r^2} \end{aligned} \quad (3.41)$$

The final step in Equation (3.41) includes defining the dipole moment as $m = d\delta q$.

The charge-dipole potential energy varies as $1/r^2$ and thus falls off more rapidly than the Coulomb potential energy with its $1/r$ dependence. The interaction range is thus shorter. Repeating the calculation above for the case of the interaction of two molecular dipoles, we find that the interaction depends on the orientation of both dipoles with respect to the line drawn between their centers and decreases as a function of $1/r^3$. The range of interaction is even smaller than that for the electrostatic interaction between a charge and a dipole. These findings justify the diagram drawn in Figure 3.3 to show how the ‘structured’ hydration shell around the ion is confined to a relatively small spherical volume defined by around two to three water molecule diameters.

In Chapter 1 the weak attractive force between neighbouring atoms, known as the Van der Waals force, was described in terms of an induced dipole–dipole interaction. In Figure 1.2 this force is shown to vary as $1/r^6$. This represents a very short-range force, and we are now able to approach an understanding of why this is so. Molecules that carry no charge or lack a permanent dipole moment can have an induced moment given to them as a result of an interaction with an electric field. The magnitude of this induced moment (m_i) is determined by magnitude of the field and the polarisability α of the molecule according to the relationship:

$$m_i = \alpha E. \quad (3.42)$$

The potential energy of the interaction between the field created by a charge Q and an induced dipole moment is then given by inserting the above expression for the induced moment into Equation (3.41), using Equation (3.9) as the field, and assuming that the induced moment is aligned with the field:

$$U = -\frac{\alpha Q^2}{(4\pi\epsilon_0\epsilon_r)^2 r^4}. \quad (3.43)$$

From Equations (3.11) and (3.41) the potential ϕ_m and field E_m ($-\nabla\phi_m$) of a dipole of moment m is given by:

$$\begin{aligned}\phi_m &= -\frac{m}{4\pi\epsilon_0\epsilon_r r^2} \\ E_m &= \frac{3m}{4\pi\epsilon_0\epsilon_r r^3}\end{aligned}\quad (3.44)$$

Using Equation (3.44) for the dipole fields, and following the procedure described to obtain Equation (3.41), the sum of the four Coulomb terms in the interaction between a dipole and an induced dipole leads to the following form for the potential energy:

$$U \propto \frac{\alpha m^2}{r^6}.$$

This provides an insight into the attractive dipole-dipole force represented by the $1/r^6$ term in the Lennard-Jones 6–12 potential described in Chapter 1 and shown in Figure 1.2. The steric repulsion term ($1/r^{12}$) simply describes the fact that atoms and their electron shells cannot occupy the same space, and so describes a steep repulsion at a short range that balances out the longer range attraction force. Assigning to the steric repulsion term a $1/r^{12}$ dependence is not based on any real physical theory, but is mathematically convenient and describes quite well the experimental data on intermolecular spacing in crystals [3].

3.2.5 Ions in a Membrane or Protein

A modified form of Equation (3.11) can be used to calculate the self free-energy (also known as the electrostatic self-energy or Born energy) of a single charge in a medium. The potential energy change when a small increment of charge δq is brought to the surface of a sphere that already carries a charge q can be described by:

$$\partial U = \frac{q\delta q}{4\pi\epsilon_0\epsilon_r r} \frac{1}{r}. \quad (3.45)$$

The total work required, and hence the total free-energy, to carry out this charging process from an initial zero charge to a final charge Q , is given by:

$$U = \frac{1}{4\pi\epsilon_0\epsilon_r r} \int_0^Q q\delta q = \frac{Q^2}{8\pi\epsilon_0\epsilon_r r} s. \quad (3.46)$$

As an approximation we can assume the value for the radius r to be used in Equation (3.46) to be the ionic radius, namely 0.095 nm and 0.133 nm for the sodium and potassium ion, respectively. If the medium surrounding the ion is water we should be careful in choosing the appropriate value for ϵ_r . The value of 78.5 at 298 K results from unhindered rotational freedom of a water dipole and the cooperative motions of neighbouring hydrogen bonded water molecule. The concept of permittivity is thus a macroscopic one, and so assigning an effective ϵ_r for the structured water immediately surrounding an ion is not straightforward. The relative permittivity of 5 obtained for water at infrared frequencies, where dipolar orientations contribute nothing to the polarisability, can be taken as the lower value for the first

layer of water molecules around an ion. Normal tetrahedral hydrogen-bond associations will be restricted for the second layer of water, and by the fourth layer the normal bulk dielectric property of water can reasonably be assumed to have been attained. The $1/r$ dependency given by Equation (3.43) indicates that the coulomb potential is a relatively long range one. Taking a value for ϵ_r of $30 \sim 40$ in Equation (3.46) for the case of an ion in water is probably appropriate.

We will now employ Equation (3.46) to estimate the free energy transfer required to take a sodium ion from an aqueous environment into the hydrocarbon interior of a cell membrane. The relative permittivity of the membrane interior can be taken as $\epsilon_r \sim 2.35$ based on values of 2.31 for stearic acid ($\text{CH}_3(\text{CH}_2)_{16}\text{COOH}$) and 2.42 for palmitic acid ($\text{CH}_3(\text{CH}_2)_{14}\text{COOH}$) given in the *CRC Handbook of Chemistry & Physics* [4]. From Equation (3.21), with $r = 0.095 \text{ nm}$ and assigning an effective value for ϵ_r of 35, the potential energy of a Na^+ ion in water is $3.4 \times 10^{-20} \text{ J}$, and when located in a medium of the same permittivity as a hydrocarbon membrane ($\epsilon_r = 2.35$) is $2.45 \times 10^{-19} \text{ J}$. The difference between these potential energies is $2.1 \times 10^{-19} \text{ J}$, which at 298 K corresponds to an energy difference of $51.3 \text{ } kT$. The factor kT corresponds to the mean thermodynamic energy available to an ion when it is in equilibrium with its environment at temperature T . A value of $51.3 \text{ } kT$ (127 kJ/mole) therefore represents a considerable energy barrier to be surmounted, and is depicted schematically in Figure 3.7. Equation 3.43 does not take account of the image force that will be present at the interface between two materials of different permittivity, but provided that the membrane thickness is of the order of 4 nm or more (which it is for a cell membrane) this correction can be ignored [5].

A thermally activated process will be required to surmount the energy barrier shown in Figure 3.5, and so the rate of ion transfer will be proportional to $\exp(-\Delta U/kT)$. For ΔU equal to $51.3 \text{ } kT$ this exponential factor has a value of 5×10^{-23} . A pure lipid bilayer therefore presents an impermeable barrier to biologically important ions such as Na^+ , K^+ and Cl^- , and even more so for a Ca^{2+} ion with its double charge. This demonstrates why the presence of ion channels, formed of water-filled proteinaceous pores, is of such significance in cell membranes. The water molecules in these pores maintain the relatively high ϵ_r environment normally experienced by an ion as it crosses the lipid membrane. The fact that some

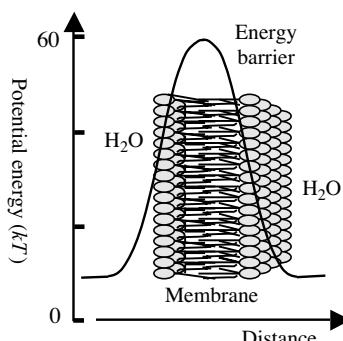


Figure 3.7 A large energy barrier ($\sim 51 \text{ kT}$ units in height) confronts the passage of a sodium ion across a pure bilipid membrane structure. The membrane structure effectively acts as an impermeable barrier to electrically charged particles.

membrane ion channels exhibit selectivity differences of more than 1000 towards different ions also indicates that variations in the molecular structure of the proteins forming the walls are of great importance. Differences in ionic radii and the physical diameter of a particular ion channels type are also considerations of importance.

The effective permittivity of the hydrophobic interior of a protein molecule can be estimated to have the same value as the inner region of a cell membrane [6]. The energy barrier presented to an ion wishing to penetrate into the protein from an external aqueous environment will thus be of the same order as that required of an ion entering a cell membrane. So, unless the protein contains a small pocket of water molecules that are accessible to the surrounding aqueous medium, it is very difficult indeed to place a charge inside a protein. A bare ion lacking an extensive hydration shell would represent a highly disruptive influence if located within a protein. This is why the ionisable amino-acid residues in a protein structure are located on the outside of the molecule and make contact with the surrounding aqueous medium.

3.3 Hydrophobic and Hydration Forces

3.3.1 *Hydrophobic Forces*

The familiar saying that ‘oil and water do not mix’ is basically a description of the hydrophobic force. This is commonly observed as oily liquids aggregating to form a separate phase from water, or the formation of beads of water on a waxy leaf surface, for example. The aversion (phobia) that nonpolar substances have to water explains the creation of the term ‘hydrophobic’.

At the molecular level the hydrophobic effect is an important driving force in the folding of protein structures. Water soluble proteins have structures in which the side groups of hydrophobic amino acids such as alanine, tryptophan and valine, are brought together to form a hydrophobic core situated as far as possible from the surrounding solvent. Hydrophilic (water-loving) amino acids such as arginine, aspartic acid and lysine, are situated on the protein surface such that their polar or charged side groups interact with surrounding water molecules. The principal driving force behind the protein folding process is the minimising of the number of hydrophobic side groups exposed to water. Hydrophobic forces also drive the formation of cell membranes and the insertion of membrane proteins into their nonpolar lipid interiors. They are also important energetic factors in the tertiary structures of DNA through stacking interactions between hydrocarbon bases.

The hydrophobic effect is entropy driven. It involves neither a repulsive force between nonpolar molecules and water, nor an attractive force between nonpolar molecules. A decrease in entropy results from a hydrocarbon molecule causing a disruption of the normal hydrogen-bond network between water molecules. Each hydrogen-bond has a strength of around 20 kJ/mol. As pictured in Figure 3.8 the nonpolar hydrocarbon is unable to form hydrogen bonds, and so the hydrogen bonding of water molecules at its surface is disrupted and partially reconstructs to form a solvation shell. The water molecules in this shell have restricted rotational and translational mobilities, and this represents an unfavourable free energy of the system. The overall disruptive effect to the system is reduced by hydrophobic molecules aggregating together so as to reduce their surface area exposed to water.

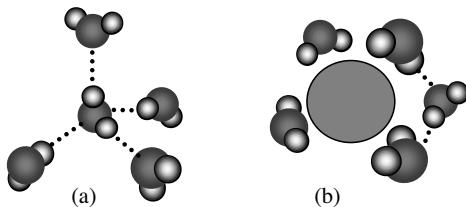


Figure 3.8 (a) Water molecules forming their normal tetrahedral arrangement of hydrogen bonds (dotted lines) in bulk water. (b) Water molecules at the surface of a hydrophobic body are restricted in orientation as they attempt to form hydrogen bonds with other water molecules.

3.3.2 Hydration Forces

Surfaces that consist of polar molecules or ionisable acidic and basic groups attract water dipoles or can form hydrogen bonds with them. They are hydrophilic surfaces. An example is the surface of the phospholipid bilayer structure shown in Figure 2.2, where the polar phosphate head groups also carry a negative charge. As two hydrated lipid bilayers come closer together they experience a repulsive force. Depending on the chemical composition of the phospholipids head group, this force increases exponentially with decreasing separation distance, and prevents them from approaching closer than around 2–4 nm. Measurements of this effect are not performed by physically pushing membrane surfaces together, but by osmotically taking away water and thus reducing the chemical potential of the water [7]. Measurement of the membrane separation as a function of the water potential leads to a value of the hydration force, which is related to the work required to remove the hydration layer on each membrane surface. The water molecules are strongly attracted to the polar and charged head groups and rearrange themselves around these groups as the membrane surfaces approach each other. At very close separation the chemical potential is similar to that of the water molecules shown nearest to an ion in Figure 3.3.

A hydrophilic surface, and especially a cell membrane surface, will carry a net surface charge. Electrostatic interactions should therefore also be involved as membrane surfaces approach each other. The evidence is, however, that such interactions are overcome by the hydration forces for separation distances below around 2.5 nm. The hydration force can be overcome by protein-protein interactions between adjacent membranes, or by divalent cations (e.g. Ca^{2+} , Mg^{2+}) that can link between two adjacent negative phosphate head charges and also disrupt water molecule binding. The coming together and subsequent fusing of membranes is an important process in cell biology. Hydration repulsion is therefore a key factor in ensuring that membrane fusion is a highly controlled and not a random process.

3.4 Osmolarity, Tonicity and Osmotic Pressure

3.4.1 Osmoles

This refers to the number of impermeable particles dissolved in a solution, regardless of charge. This is important for determining the diffusional movement of water, as for example

across a cell membrane. For substances that maintain their molecular structure when they dissolve (e.g. a sugar molecule such as glucose, proteins, DNA), the *osmolarity* and the *molarity* are essentially the same. For substances that dissociate (e.g. an ionic salt such as NaCl) when they dissolve, the osmolarity is the number of free particles times the molarity. Thus for a pure NaCl solution, a 1 mM solution would be 2 mOsmolar (1 mOsm each for Na and Cl). A 1 mM MgCl₂ solution would have an osmolarity of 3 mOsm (for simplicity we are assuming unity activity coefficients). When measured as osmoles per litre, one obtains the *osmolarity*. For osmoles per kg water, one obtains *osmolality*.

3.4.2 Calculating Osmolarity for Complex Solutions

As described above, the osmolarity of a simple solution is equal to the molarity times the number of particles per molecule. However, real solutions can be much more complex. For example:

- proteins with many equivalents/L may only contribute a small amount to the osmolarity, since they consist of a few very large ‘particles’;
- not all the solution volume is aqueous; for example, blood plasma has 7% dissolved proteins and lipids;
- not all ions are free in a solution; cations may be bound to other anions or to proteins.

For complete accuracy, all constituents should be included in the calculation. However, such aspects as those given above can leave many uncertainties when calculating the osmolarity of a solution like blood plasma. Therefore, we sometimes have to take shortcuts that provide us with good approximations.

For example, we can obtain a good estimate for plasma osmolarity by taking the reported Na concentration (mEq/Litre plasma) and doubling this value. This obviously erroneous calculation (given all of the above) gives a result close to the correct one, since the errors tend to cancel each other! In some clinical settings, one must also account for the effects of elevated plasma glucose or urea.

3.4.3 Osmolarity Versus Tonicity

Consider the situation shown in Figure 3.9, where a protein solution is separated from a protein-free buffer solution by a semi-permeable membrane. If buffer components (e.g. salts, small sugars, amino-acids) and solvent may pass through the membrane, but the protein cannot (i.e. is impermeable), then we have a nonequilibrium situation. The chemical activity (effective concentration) of water in the protein solution will be lower than that in the solution on the other side of the membrane; consequently, water will tend to flow into the protein solution. The strength of this tendency is termed osmotic pressure, and can be measured by the amount of excess pressure required to prevent water flow across the membrane. (This device is called an osmometer.)

Osmotic pressure P_{osm} is proportional not only to the concentration C of the solute but also to the absolute temperature T :

$$P_{osm} = RTC = \frac{nRT}{V}, \quad (3.47)$$

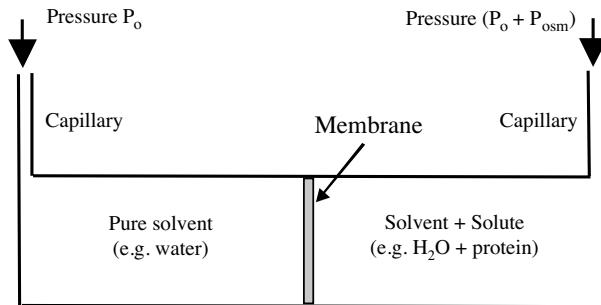


Figure 3.9 A protein solution separated from a protein-free solution by a semi-permeable membrane. The excess pressure P_{osm} required to prevent water flowing into the protein solution down its activity gradient is termed the osmotic pressure.

where n is the number of mole equivalents of solute, R is the molar gas constant and V is the volume in litres (solute molecules in solution behave thermodynamically like gas molecules). The molar gas constant has the value $0.082 \text{ L atm K}^{-1} \text{ mol}^{-1}$ (or $8.314 \text{ J K}^{-1} \text{ mol}^{-1}$). Like the gas laws, however, this expression for osmotic pressure holds true only for dilute solutions, and so corrections must be made for concentrated solutions and the activity coefficients of electrolytes. For example, to calculate the osmotic pressure of a 100 mM aqueous solution of NaCl we note from Table 1.3 that at 298 K the activity coefficient is 0.78. Thus, the mole equivalent/litre of a 100 mM NaCl salt solution = $2 \times 0.1 \times 0.78 = 0.156 \text{ equiv/L}$. According to Equation (3.47) the osmotic pressure is:

$$P_{osm} = \frac{(0.156 \text{ mol})(0.082 \text{ Latm/K mol})(298 \text{ K})}{1 \text{ L}} = 3.81 \text{ atm.}$$

Osmolarity measures the effective gradient of water, assuming that all the osmotic solute is completely impermeant. It is simply a count of the number of dissolved particles. Therefore a 300 millimolar solution of glucose, a 300 millimolar solution of urea, and a 150 millimolar solution of NaCl each have the same osmolarity.

However, a cell placed in each of these solutions would behave very differently. In a 150 mM NaCl solution there would be equal osmotic strengths on both sides of the cell membrane, so that the cell should maintain the same volume. NaCl in its ionic dissociated form cannot cross the membrane, and the cytoplasm is mostly equivalent to a 150 mM NaCl solution. On the other hand urea is very permeable through most cell membranes and so it exerts little osmotic force against a real cell and its membrane. A cell placed in 300 mM urea would rapidly swell because urea would enter the cell down its concentration gradient, followed by water down its activity gradient.

Tonicity is a functional term that describes the tendency of a solution to resist expansion of the intracellular volume. Two solutions are isosmotic when they have the same number of dissolved particles, regardless of how much water would flow across a given membrane barrier. In contrast, two solutions are isotonic when they would cause no water movement across a membrane barrier, regardless of how many particles are dissolved. In the example given

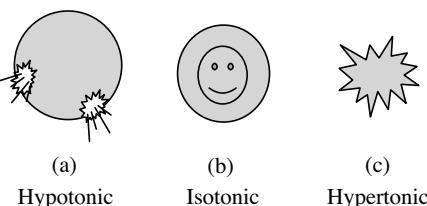


Figure 3.10 Cells suspended in various solutions. (a) Cells swell and burst in a *hypotonic* solution such as pure water. (b) Cells are ‘happy’ in an *isotonic* solution such as physiological strength saline. (c) Cells shrink and shrivel in a *hypertonic* solution such as a concentrated salt solution.

above, a 150 mM NaCl solution would be isosmotic to the inside of a cell, and it would also be isotonic – the cell would not swell or shrink when placed in this solution. On the other hand, a 300 mM urea solution, while still isosmotic would cause the cell to swell and burst (due to its permeability). This isosmotic urea solution is not isotonic. Instead, it has a lower tonicity and is termed as being *hypotonic*. (A solution of higher tonicity is called *hypertonic*.) These various situations are cartooned in Figure 3.10.

3.5 Transport of Ions and Molecules across Cell Membranes

The main function of the plasma membrane is to protect the interior of the cell from the outside world. It controls the passage of incoming and outgoing substances, and maintains the ionic concentrations of various substances. It is also selectively permeable, allowing some molecules into the cell and keeps others out. A good example is the blood–brain barrier, which allows the passage of some substances into the brain, but screens out toxins and bacteria (although HIV and bacterial meningitis can cross this barrier). Substances allowed to cross this barrier include water, CO₂, O₂, glucose, amino acids and antihistamines.

Cells obtain the ions and molecules they need from their surrounding fluid. This involves their transport across membranes – the plasma membrane as well as those membranes that bound the nucleus, endoplasmic reticulum, and mitochondria. The principle modes of transport are diffusion, osmosis, facilitated diffusion and through active transport.

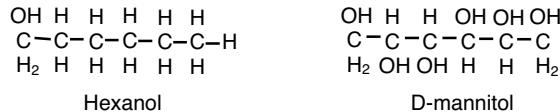
3.5.1 Diffusion

Molecules move spontaneously down their concentration gradient (i.e. from a region of higher to a region of lower concentration) by diffusion, a process that is facilitated by their Brownian motion. As a result of diffusion molecules reach an equilibrium where they are evenly distributed and no net movement of molecules occurs across the membrane. The membrane is permeable to water molecules and other small ones such as oxygen and carbon dioxide. Oxygen is nonpolar and so diffuses very quickly. Carbon dioxide and water molecules are polar, but are also very small and so diffuse freely in and out of the cell. Other substances do not.

If a solute molecule comes into contact with the lipid layer of the membrane, it may enter the lipid phase by virtue of its thermal energy and cross the lipid bilayer, to emerge into the aqueous phase on the other side of the membrane. To leave the aqueous phase and enter the

lipid phase, a solute must first break all its hydrogen bonds with water. This activity requires kinetic energy of ~ 5 kcal per hydrogen bond. Moreover, the solute molecule crossing the lipid phase of the membrane must dissolve in the lipid bilayer. Its lipid solubility will therefore play a role in determining whether or not it will cross the membrane by simple diffusion.

It is therefore evident that those molecules with a minimum of hydrogen bonding with water will most readily enter the lipid bilayer, whereas the probability is low that polar molecules such as water and inorganic ions will dissolve in the bilayer. Consider for example the structures of two 6-carbon molecules – hexanol and D-mannitol:



Note the difference in the number of hydroxyl (OH) groups. Hexanol is poorly soluble in water and highly soluble in lipids, whereas mannitol is highly soluble in water and poorly soluble in lipids owing to its hydrogen-bonding capacity. Thus, even though they are of the same size, hexanol diffuses across membranes much more readily than mannitol.

The probability P with which a molecule will cross a membrane can be expressed as:

$$P = \frac{D_m K}{d},$$

in which D_m is the diffusion coefficient of the molecule within the membrane (the more viscous the membrane or the larger the molecule the lower this value), K is the partition coefficient of the molecule, and d is the thickness of the membrane. A simple way to determine the partition coefficient is to shake the test substance in a closed tube containing equal amounts of water and olive oil. The partition coefficient K is determined from the relative solubilities in water and oil at equilibrium by the equation:

$$K = \frac{\text{solute concentration in oil}}{\text{solute concentration in water}}.$$

The diffusion of water through the plasma membrane is of such importance to the cell that it is given a special name: *osmosis*.

3.5.2 Osmosis

Osmosis is a special term used for the diffusion of water through cell membranes. Although water is a polar molecule, it is able to pass through the lipid bilayer of the plasma membrane. Selective transport of water molecules, in single file, also takes place through pores formed by transmembrane proteins called aquaporins (Peter Agre received the 2003 Nobel Prize in Chemistry for their discovery).

Water passes by diffusion from a region of higher to a region of lower water concentration. Osmosis through a selectively permeable membrane is illustrated in Figure 3.11. Water is never transported actively – it never moves against its concentration gradient. However, the concentration of water can be altered by the active transport of solutes, and thus its movement in and out of the cell can be controlled.

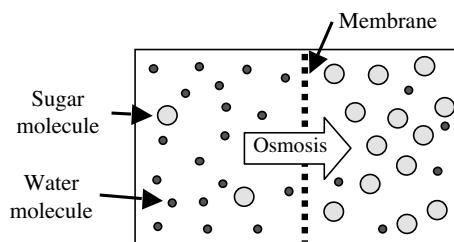


Figure 3.11 Osmosis is the term given to the diffusion of water from a region of high concentration (high potential) of water molecules to a region of low concentration (low potential) across a partially permeable membrane. Water is shown here passing from a dilute to a high concentration of impermeable sugar molecules.

3.5.2.1 Hypotonic Solutions

If the concentration of water in the medium surrounding a cell is greater than that of the cytosol, the medium is said to be hypotonic. Water enters the cell by osmosis. A simplified depiction of this osmotic process is shown in Figure 3.12. As shown in Figure 3.10 a red blood cell placed in a hypotonic solution (e.g. 0.1% salt solution) will burst (haemolysis) as a result of the influx of water. White blood cells with their nucleus and more extensive cytoskeleton will expand but are less likely to burst. Bacteria and plant cells avoid bursting in hypotonic solutions because of strong cell walls. These allow the buildup of *turgor* pressure within the cell, until it equals the osmotic pressure and osmosis ceases.

As depicted in Figure 3.10 when red blood cells are placed in a 0.9% salt solution, they neither gain nor lose water by osmosis. Such a solution is said to be isotonic. The extracellular fluid of mammalian cells is isotonic to their cytoplasm. This balance must be actively maintained because of the large number of organic molecules dissolved in the cytosol but not present in the external fluid. These organic molecules exert an osmotic effect that, if not compensated for, would cause the cell to take in so much water that it would swell and might even burst. This fate is avoided by pumping sodium ions out of the cell with the sodium-potassium pump.

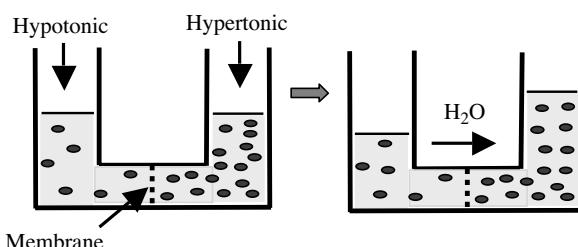


Figure 3.12 Water passing by osmosis across a selectively permeable membrane from a hypotonic to a hypertonic solution.

If red blood cells are placed in sea water ($\sim 3\%$ salt) they lose water by osmosis and the cells shrink and shrivel up. Sea water is hypertonic to the cytosol of the red cells. Water will diffuse from the cytoplasm down its activity gradient in an attempt to dilute out the salt solution, following the same basic process shown in Figure 3.12. Sea water is also hypertonic to the external fluid of most marine vertebrates. To avoid fatal dehydration these animals must continuously drink sea water and then desalt it by pumping ions out of their gills by active transport. Marine reptiles (turtles and snakes) use special salt glands for the same purpose. If a plant tissue is placed in sea water, the cell contents shrink away from the rigid cell wall in a process called plasmolysis.

However, lipid bilayers are impermeable to most essential ions and molecules, such as: K^+ , Na^+ , Ca^{2+} (cations); Cl^- , HCO_3^- (anions); small hydrophilic molecules like glucose and mannitol; macromolecules like proteins and RNA. Cells solve this problem by means of facilitated diffusion and active transport.

3.5.3 Facilitated Diffusion

Large polar molecules, such as glucose and amino acids, cannot diffuse across the plasma membrane of a cell. Ions such as Na^+ and Cl^- also cannot pass through by simple diffusion. These molecules and ions pass through transmembrane protein channels instead by a process known as facilitated diffusion. These proteins, or assemblies of proteins, are embedded in the plasma membrane to form a water-filled channel through which an ion or molecule can pass down its concentration gradient into or out of the cell. This is depicted in Figure 3.13. Molecules will randomly move through the channel or pore by diffusion. This requires no energy. It is a passive process, just as for osmosis. The transmembrane channels that permit facilitated diffusion can be opened or closed. They are said to be *gated*. Many ion channels open or close in response to binding a small signalling molecule or *ligand*. Apart from ligand-gated ion channels, there are also mechanically-gated, voltage-gated, and light-gated channels.

3.5.4 Active Transport

Cells must maintain ion concentration gradients across their plasma membrane (see Table 3.2). Passive transport cannot achieve this. Active transport is the pumping of molecules or

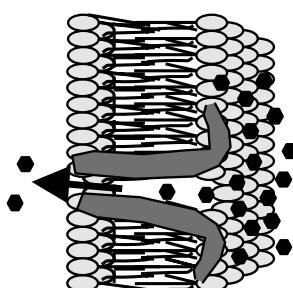


Figure 3.13 Ions and large polar molecules can pass through a cell membrane by facilitated diffusion along water channels formed by transmembrane proteins.

ions through a membrane *against* their concentration gradient. It requires a transmembrane protein (usually a complex of them) called a transporter, as well as energy in the form of ATP. The energy of ATP may be used directly or indirectly. Some transporters bind ATP directly and use the energy of its hydrolysis as described in Chapter 1 to drive active transport. Other transporters use the energy already stored in the gradient of a directly-pumped ion. Direct active transport of the ion establishes a concentration gradient. When this is relieved by facilitated diffusion, the energy released can be harnessed to the pumping of some other ion or molecule.

3.5.4.1 Sodium-Potassium Pump

The cytosol of animal cells contains a concentration of potassium ions as much as 20-times higher than that in the extracellular fluid. Conversely, the extracellular fluid contains a concentration of sodium ions as much as 10-times greater than that within the cell. These concentration gradients are established by the active transport of both ions, using the same transporter called the sodium-potassium pump (or Na^+/K^+ ATPase). This pump uses the energy from the hydrolysis of ATP described in Chapter 1 to actively transport 3 Na^+ ions out of the cell for each 2 K^+ ions pumped into the cell. This result is shown schematically in Figure 3.14. Almost one third of all the energy (ATP) generated by the mitochondria in animal cells is used solely to run this pump!

As will be discussed later in this chapter, the action of the sodium-potassium pump accomplishes several vital functions:

- Accumulation of sodium ions outside of the cell draws water out of it and thus enables the cell to maintain osmotic balance.
- The gradient of sodium ions is harnessed to provide the energy to run several types of indirect pumps.
- It helps establish a net charge across the plasma membrane – producing a resting membrane potential that prepares nerve and muscle cells for the propagation of action potentials leading to nerve impulses and muscle contraction.

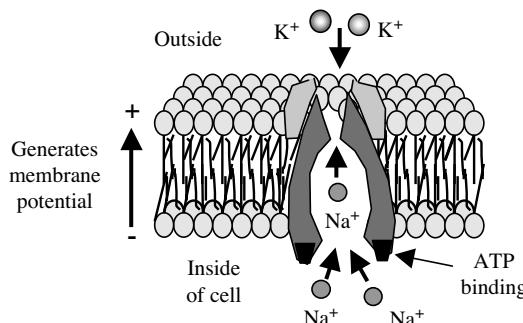


Figure 3.14 The sodium-potassium pump uses the free energy released by the hydrolysis of ATP to pump out 3 sodium ions for 2 potassium ions pumped into the cell. This generates a membrane potential of around -70 mV (negative with respect to the reference zero potential of the extracellular medium).

Table 3.1 Summary of the methods of transport of molecules and ions across the plasma membrane of cells

	Active or passive	Molecules that move	Direction	Requires energy?	Protein required?
Diffusion	Passive	Small hydrophobic	Down gradient	No	No
Osmosis	Passive	Water	Towards high conc of solutes	No	No
Facilitated Diffusion	Passive	Selected by specific transporter	Down gradient	No	Yes
Active Transport	Active	Selected by specific transporter	Pumped in or out (depends on transporter)	Yes (ATP)	Yes

A summary of the methods of transport of ions and molecules across the cell membrane is given in Table 3.1.

3.6 Electrochemical Gradients and Ion Distributions Across Membranes

If a molecule bears an electric charge its net ionic flux across a membrane will be determined not only by the permeability of the membrane to that ion and the concentration gradient of the ion, but also by the electric potential difference between the two sides of the membrane. The following are the main factors of importance:

For charged molecules (e.g. Na^+ , K^+ , Cl^- , Ca^{2+} , amino acids) two forces act to produce net passive diffusion of the species across a membrane:

- (a) The chemical gradient arising from differences in the concentration of the substance on the two sides of the membrane, and
- (b) The electrical field (i.e. difference in potential across the membrane divided by membrane thickness) experienced by the ion as it enters the membrane.

A positively charged ion will tend to move in the direction of increasing negative potential. The sum of the combined forces of concentration gradient and electrical gradient acting on an ion determine the net *electrochemical gradient* acting on the ion. It follows that there must be a potential difference just sufficient to balance and counteract the chemical gradient acting on an ion, so as to prevent a net trans-membrane flux of the ion in question. The potential at which an ion is in electrochemical equilibrium is termed the *equilibrium potential*. The value of this potential depends on several factors, the most prominent being the ratio of concentrations of the ion in question. We will cover this in more detail in Section 3.10, but for the present we can state that for a monovalent ion at 298 K the equilibrium potential is equal to:

$$0.059 \times \log_{10} (\text{ratio of extracellular to intracellular concentration}) \text{ Volts.}$$

Thus, a 59 mV potential difference across the membrane has the same effect on the net diffusion of that ion as a transmembrane concentration ratio of 10:1 for that ion.

Passive diffusion of an ion species will therefore take place against its chemical concentration gradient if the electrical gradient (i.e. potential difference) across the membrane opposes, and exceeds, the concentration gradient. For example, if the interior of a cell is more negative than the equilibrium potential for K^+ , potassium ions will diffuse into the cell even though the intracellular concentration of K^+ is much higher than the extracellular concentration. Electrical forces, of course, do not act directly on uncharged molecules such as sugars.

3.6.1 Donnan Equilibrium

In 1911, the physical chemist Frederick Donnan examined the distribution of diffusible solutes separated by a membrane that is freely permeable to water and electrolytes, but totally impermeable to one species of ion confined to one of two compartments. In this situation, as Donnan discovered, the diffusible solutes become unequally distributed amongst the two compartments. We will consider the experiment outlined in Figures 3.15 and 3.16.

The experiment involves four stages, the first two of which are shown in Figure 3.15.

- Pure water is placed in the two chambers separated by a semipermeable membrane. KCl is dissolved into chamber 1.
 - The dissolved salt (K^+ and Cl^-) will diffuse through the membrane until the concentration of K^+ and Cl^- become equal on both sides.
- The third and fourth stages are shown in Figure 3.16.
- The potassium salt of a nondiffusible anion (a macromolecule such as a protein P^- (which may have multiple negative charges) is added to the solution in chamber 1.
 - The K^+ and Cl^- ions redistribute until a new equilibrium is established by movement of some K^+ from chamber 2 into chamber 1 to compensate for the nondiffusible P^- ions, as well as a net transfer of Cl^- into chamber 2.

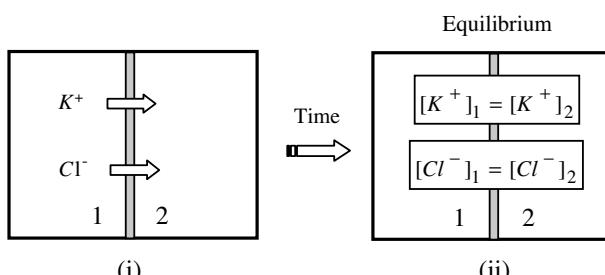


Figure 3.15 (i) KCl is dissolved in chamber 1, the dissociated ions diffuse through the semi-permeable membrane into chamber 2. (ii) At equilibrium the concentrations of the ions are equal on both sides of the membrane.

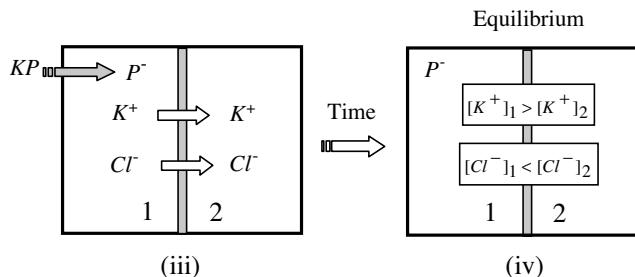


Figure 3.16 (iii) The potassium salt of a nonpermeable anion (e.g. a protein molecule P^-) is added at stage (ii) of Figure 3.13 into chamber 1. The K^+ and Cl^- ions will distribute until (iv) new equilibrium concentrations of the ions is established in chambers 1 and 2.

The *Donnan equilibrium* is characterised by a reciprocal distribution of the anion and cation such that:

$$\frac{[K^+]_1}{[K^+]_2} = \frac{[Cl^-]_2}{[Cl^-]_1} \quad (\text{square brackets indicate concentrations}). \quad (3.48)$$

At equilibrium, the diffusible K^+ is more concentrated in the compartment in which the nondiffusible anion P^- is confined than in the other, whereas the diffusible anion Cl^- becomes less concentrated in that compartment than in the other.

This equilibrium situation arises from the following physical requirements:

- (a) There must be electroneutrality in both compartments. Within each compartment the total number of positive charges must equal the total number of negative charges. In this example, $[K^+] = [Cl^-]$ in compartment 2.
- (b) The diffusible ions K^+ and Cl^- must, statistically, cross the membrane in pairs to maintain electrical neutrality. The probability that they will cross together is proportional to the product $[K^+][Cl^-]$.
- (c) At equilibrium the rate of diffusion of KCl in one direction through the membrane must equal the rate of KCl diffusion in the opposite direction. Thus, at equilibrium the product $[K^+][Cl^-]$ in one compartment must be equal to the same product in the other compartment. This is the relationship given by Equation (3.48).

An algebraic expression for the equilibrium condition can be derived by assigning to the chamber 2 the concentrations $[K^+]_2 = [Cl^-]_2 = x$. We also let $[Cl^-]_1 = y$, and for the added protein salt we assign $[KP] = z$. The protein anion P^- will remain in chamber 1, and so at final equilibrium $[K^+]_1 = (y + z)$. The equality of the product $[K^+][Cl^-]$ in the two compartments at final equilibrium can thus be expressed as:

$$y(y + z) = x^2. \quad (3.49)$$

This equation holds if P^- is not present, for in that case K^+ and Cl^- are equally distributed and $z = 0$ and $x = y$.

Rearrangement of Equation (3.49) gives, at equilibrium:

$$\frac{x}{y} = \frac{y+z}{x}, \quad (3.50)$$

to show that as the concentration z of the nondiffusible anion P^- is increased, the concentrations of the diffusible ions will become increasingly divergent.

Example 3.1

Figure 3.17 shows the initial (before equilibrium) situation of a dialysis bag, containing 300 mM of a sodium-protein salt in 1L of pure water, immersed into a vessel containing 200 mM NaCl in 1 litre of pure water. The dialysis bag membrane allows small ions to pass through it, but not protein molecules. Assuming that one sodium ion completely dissociates from the protein, determine the equilibrium concentrations of sodium and chloride ions within and outside the dialysis bag.

Solution:

Consider the two conditions that determine the final equilibrium:

- (i) Chemical equilibrium for the mobile salt:

Activity of NaCl outside the bag = Activity of NaCl inside the bag:

$$[\text{Na}^+]_{\text{out}} \cdot [\text{Cl}^-]_{\text{out}} = [\text{Na}^+]_{\text{in}} \cdot [\text{Cl}^-]_{\text{in}} \quad (\text{i})$$

- (ii) Macroscopic Electroneutrality:

$$\text{Outside : } [\text{Na}^+]_{\text{out}} = [\text{Cl}^-]_{\text{out}} \quad (\text{ii})$$

$$\text{Inside : } [\text{Protein}^-] + [\text{Cl}^-]_{\text{in}} = [\text{Na}^+]_{\text{in}} \quad (\text{iii})$$

We know the following:

$$[\text{Cl}^-]_{\text{out}} + [\text{Cl}^-]_{\text{in}} = 200 \text{ mM} \quad (\text{iv})$$

$$[\text{Na}^+]_{\text{out}} + [\text{Na}^+]_{\text{in}} = 500 \text{ mM} \quad (\text{v})$$

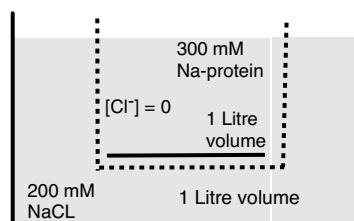


Figure 3.17 Initial concentrations of protein and NaCl across a dialysis membrane.

$$[\text{Protein}^-] = 300 \text{ mM} \quad (\text{vi})$$

Thus, we have 5 independent equations and 4 unknowns – implying that there is more than one way to solve this problem.

Step 3: Steps to a solution (here is one route):

In (i) substitute for $[\text{Cl}^-]_{\text{out}}$ using (ii):

$$\{[\text{Na}^+]_{\text{out}}\}^2 = [\text{Na}^+]_{\text{in}} [\text{Cl}^-]_{\text{in}} \quad (\text{vii})$$

In (vii) substitute for $[\text{Cl}^-]_{\text{in}}$ using (iii) and (vi):

$$\{[\text{Na}^+]_{\text{out}}\}^2 = [\text{Na}^+]_{\text{in}} \{[\text{Na}^+]_{\text{in}} - 300\}$$

Finally, on substituting for $[\text{Na}^+]_{\text{in}}$ using (v) we obtain:

$$\{[\text{Na}^+]_{\text{out}}\}^2 = (500 - [\text{Na}^+]_{\text{out}})(200 - [\text{Na}^+]_{\text{out}}).$$

$$\text{Leading to : } 700[\text{Na}^+]_{\text{out}} = 10,000$$

$$\text{To give } [\text{Na}^+]_{\text{out}} = 143 \text{ mM.}$$

$$\text{From (v) : } [\text{Na}^+]_{\text{in}} = 357 \text{ mM;}$$

$$\text{From (ii) \& (iv) : } [\text{Cl}^-]_{\text{out}} = 143 \text{ mM}, [\text{Cl}^-]_{\text{in}} = 57 \text{ mM.}$$

3.7 Osmotic Properties of Cells

We will now begin to consider the properties of the cell membrane that are responsible for the different concentrations of ions that are maintained inside and outside the cell – and for the regulation of cell volume.

Ionic Steady State:

Although the intracellular concentrations of inorganic solutes at ionic steady state conditions differ somewhat amongst different cell types and different organisms, certain generalisations can be made:

- The most concentrated inorganic ion in the cytosol is K^+ , which is typically 10–30 times as concentrated in the cytosol as in the extracellular fluid. Conversely, the internal concentrations of free Na^+ and Cl^- are typically less (\sim one-tenth or less) than the external concentrations.
- The intracellular concentration of Ca^{2+} is maintained several orders of magnitude below the extracellular concentration. This situation is due in part to active transport of Ca^{2+} out of the cell, across the membrane, and in part to the sequestering of this ion within such organelles as the mitochondria and cytoplasmic reticulum. As a result, the activity of Ca^{2+} in the cytosol is generally well below 1 μM .

Table 3.2 A typical mammalian muscle cell has the following steady state internal (cytosol) and extracellular concentrations of sodium, potassium, calcium, chloride and macromolecular anions A⁻

Internal ion concentration (mM)	Extracellular ion concentration (mM)
Na ⁺ : 10; K ⁺ : 140; Ca ²⁺ : <10 ⁻⁶ ; Cl ⁻ : 3~4; A ⁻ : 140	Na ⁺ : 120; K ⁺ : 2.5; Ca ²⁺ : 2.0; Cl ⁻ : 120

- Cell membranes are typically far more permeable (~ 30 times) to K⁺ than to Na⁺. Membrane permeability to chloride varies. In some cells it is similar to potassium, whilst in others it is lower. The permeability to Na⁺ is low – but not low enough to prevent sodium from leaking steadily into the cell.

The steady state internal and extracellular concentrations of ions for a typical mammalian muscle cell are given in Table 3.2.

In view of the general leakiness of the cell membrane, the question arises as to what degree the Donnan equilibrium described in Section 3.6 contributes to the steady-state ionic distributions between the cell interior and cell exterior. Three related factors are involved:

1. A preponderance of net negative charge resides in the form of anionic sites such as carboxyl on peptide and protein molecules that are nonpermeant and thus trapped within the cell. These charges must be balanced by positively charged counterions such as Na⁺, K⁺, Mg²⁺, and Ca²⁺.
2. Because such ‘immobile’ anionic sites are trapped within the cell by the inability of the parent peptides and proteins to cross the outer cell membrane, we have a natural situation similar in some respects to the artificial situation we considered in Section 3.6 to illustrate Donnan Equilibrium. If K⁺ and Cl⁻ were the only diffusible ions, an equilibrium situation in the cell would indeed develop similar to that shown in Figure 3.16. However, the cell membrane is leaky to Na⁺ and other inorganic ions. With time the cell would load up with these ions if they were simply allowed to accumulate. This in turn would cause osmotic movement of water into the cell, causing it to swell.
3. Such osmotic disasters are avoided by the ability of the cell membrane to pump out Na⁺, Ca²⁺, and some other ions at the same rate they leak in, keeping the intracellular Na⁺ concentration about an order of magnitude lower than the extracellular concentration. This active pumping confers on the membrane an effective impermeability to Na⁺ and Ca²⁺. As a result, the concentrations of these ions are not allowed to come into equilibrium, and the cell in fact behaves very much on the surface *as if* it were in a state of Donnan Equilibrium. In spite of this resemblance, the unequal distribution of ions represents a *steady state* requiring the continual expenditure of energy (to pump ions) rather than a true equilibrium.

Since K⁺ and Cl⁻ are by far the most concentrated and most permeant ions in the tissue, they distribute themselves in a way similar to that in an ideal Donnan Equilibrium – namely that the KCl concentration product ([K⁺] \times [Cl⁻]) of the cell interior will approximately equal the KCl concentration product of the extracellular solution, providing the membrane permeabilities of chloride and potassium are both high relative to those of other ions present. This is shown in Figure 3.18.

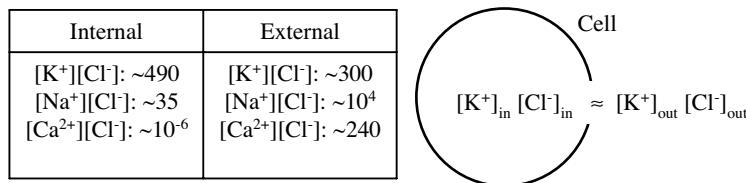


Figure 3.18 The potassium and chloride ions are the most concentrated and membrane permeable ions in tissue. They distribute themselves within and outside cells similar to that expected for ideal Donnan equilibrium.

The combination of an asymmetric distribution of ions between the intra- and extracellular fluids, together with the selective ion permeability of the cell membrane, give rise to the membrane equilibrium potentials listed in Table 3.3. This will be discussed in more detail later in this chapter.

3.8 Probing the Electrical Properties of Cells

Electrical phenomena in living tissues can be detected by placing two electrodes in the tissue to measure the field set up by electric currents flowing through the extracellular fluids. Since these currents originate across cell membranes, a more direct and quantitative approach is to measure electrical events across the membrane of a single cell. This measurement is done by comparing the electric potential of one side of the membrane with that of the other side. One sensing electrode is placed in electrical continuity with the outside of the cell, and another is inserted inside the cell. The difference between these two potentials is the membrane potential V_m and is always given as the intracellular potential relative to the extracellular potential, which is arbitrarily defined as zero. A simple electrical stimulating and recording arrangement is shown in Figure 3.19.

Table 3.3 Ratios of the external and internal cell concentrations of important ions for a typical mammalian muscle cell and human red blood cell

Cell Type	Ion	Conc. ratio (out/in)	Equil. potential mV (calc.)	Measured potential (mV)
Muscle	Na^+	12	+67	−90 mV
	K^+	0.026	−98	
	Ca^{2+}	15 000	+123	
	Cl^-	30	−90	
Red blood cell	Na^+	18	+74	−10 to −14
	K^+	0.05	−77	
	Ca^{2+}	52 000	+139	
	Cl^-	1.6	−12	

The calculated equilibrium and measured membrane potential is also given. (Derived from Wilfred D. Stein, *Channels, Carriers, and Pumps*, Academic Press, p. 37, 1990.)

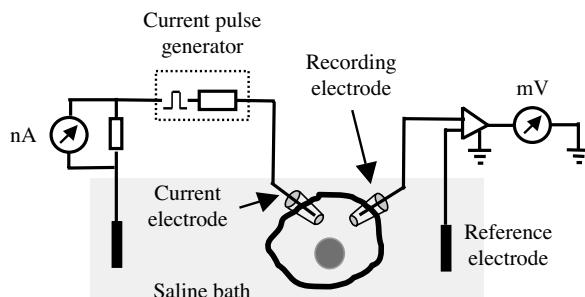


Figure 3.19 A basic system for the stimulation and recording of the electrical properties of a cell membrane. Ingoing or outgoing current pulses are applied to the cell. The difference between the potential of an electrode inserted into the cytosol and an external electrode gives the membrane potential V_m .

As shown in Figure 3.19 the cell is immersed in a physiological saline solution containing a reference electrode. Glass capillary microelectrodes, with tip diameters less than 0.1 micron and filled with an electrolyte such as 3 M KCl, can be inserted into cells with negligible damage to their membranes. The tip resistance of such microcapillary electrodes can approach values of $20 \sim 50 \text{ M}\Omega$ and so will not act as a short-circuit across the membrane and can serve as a voltage probe. The first step is to insert the tip of such a recording electrode through the membrane of the cell. Before the tip of this microelectrode enters the cell, it and the reference electrode are at the same potential (taken to be reference zero). When the fine capillary tip penetrates the membrane, the cytoplasm is in continuity with the electrical connection to a voltage amplifier via a fine column of electrolyte that fills the inside of the capillary electrode (e.g. a 3 M solution of KCl). As the tip of the recording microelectrode is advanced, penetration of the plasma membrane is indicated by the sudden appearance of a negative potential shift of the voltage trace (see Figure 3.20). The steady negative potential recorded by the electrode tip in the cytoplasm is the resting potential V_{rest} (mV). All cells that have been investigated have a negative resting potential, which can be as high as

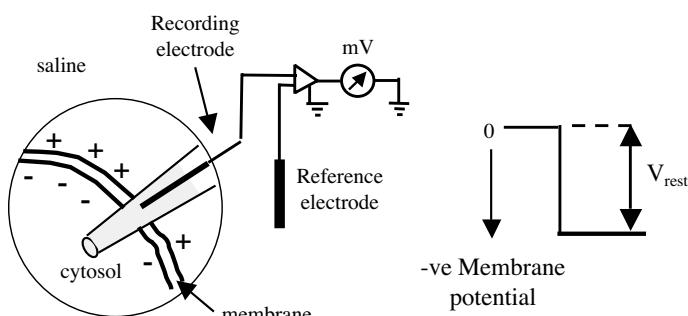


Figure 3.20 When a high resistance glass capillary voltage probe is inserted into a cell it records a negative potential with respect to the outside of the cell. This is the resting membrane potential (V_{rest}).

-100 mV . The potential sensed by the intracellular electrode does not change as the tip is advanced further into the cell. Thus, the entire potential difference between the cell interior and cell exterior exists across the surface membrane and in the regions immediately adjacent to the inner and outer membrane surfaces.

The electrical properties of the cell membrane can be examined by causing a pulse of current to pass through the membrane so as to produce a perturbation in the membrane potential. A second microelectrode, the current electrode shown in Figure 3.19, can deliver such a current. The current from this electrode, in the form of a current pulse generated by applying a step voltage in series with a high value resistance ($>1\text{ G}\Omega$), flows across the membrane in either the inward (bath to cytoplasm) or the outward direction depending on the polarity of the step voltage. When current pulses are passed so that positive charge is removed from inside the cell via the current electrode, the potential difference across the membrane increases (hyperpolarises). The intracellular negative potential is increased (e.g. from -60 to -70 mV). With hyperpolarisation, the membrane (with some exceptions) produces no response other than a positive potential change due to the applied current. If a current pulse is passed from the electrode into the cell, positive charge will be added to the inner surface of the cell membrane. This charge causes the potential difference across the membrane to decrease, and the cell is then said to become *depolarised* (e.g. from -60 to -50 mV). These two types of response are shown in Figure 3.21.

As the strength of the outward pulse is intensified, depolarisation will increase, as shown in Figure 3.21. *Excitable cells*, such as nerve, muscle, and many receptor cells, exhibit a *threshold potential* at which the membrane will produce a strong active response. This is known as the *action potential* shown in Figure 3.22. The action potential is caused by the activation of membrane channels permeable to sodium, which themselves are activated by the reduction in voltage difference between the two sides of the cell membrane. The opening of the sodium channels in response to depolarisation and the resulting flow of sodium ions into the cell provide an example of *membrane excitation*. The mechanisms underlying the action potential and other instances of membrane excitation will be considered later in this chapter.

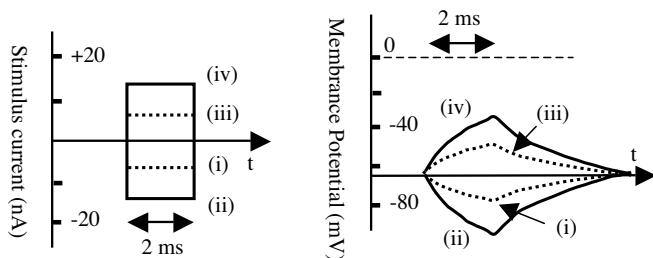


Figure 3.21 When a current pulse is applied that removes positive charge from inside a cell, hyperpolarisation of the membrane occurs. The intracellular negative potential is increased (e.g. from -60 to -70 mV). A current pulse of opposite polarity will add positive charge to the inner surface of the cell membrane, causing depolarisation of the membrane (e.g. from -60 to -50 mV).

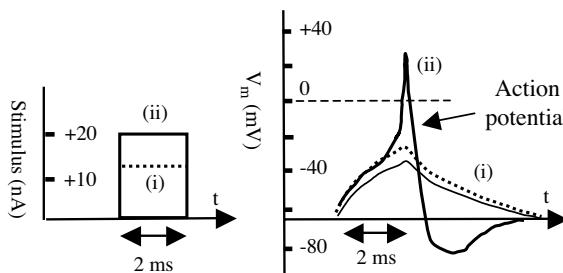


Figure 3.22 Excitable cells, such as nerve, muscle, and many receptor cells, exhibit a threshold depolarisation potential at which the membrane will produce a strong active response. This is known as the action potential.

We can now appreciate that cell membranes respond to stimuli with two quite different classes of electrical behaviour – namely, *passive* and *active* behaviour:

3.8.1 Passive Electrical Response

This is always produced when an electric current is forced across a biological membrane, because of the electrical capacitance and conductance properties of the membrane. Passive responses occur independently of any molecular changes that open or close gated ion channels in the membrane. The resistance (reciprocal of conductance) of a cell membrane is associated with leakage pathways that allow inorganic ions to cross the membrane. The capacitance of a membrane is a measure of the extent to which the ion impermeability of the membrane leads to separation of electrical charges across the membrane.

3.8.2 Active Electrical Response

Such responses, known as membrane excitations, are found in excitable tissue such as nerve, muscle, and sensory receptors. They depend on the opening and/or closing of numerous *ion channels* (also called *membrane channels*) in response to a stimulus. Some ion channels are *gated* (i.e. opened and shut) by changes in voltage across the membrane, while others are opened by the binding of transmitter or messenger molecules. Other channels, primarily in sensory receptor cells, are activated by specific stimulus energies such as light (photoreceptors) or mechanical strain (mechanoreceptors). When a certain group of channels selectively permeable to a certain species of ion is opened, a current may be carried across the membrane. As in the case of sodium channels, such a current normally produces a voltage signal across the membrane. As we will later in this chapter, the gating of ion channels is the immediate cause for nearly all electrical activity in living tissue.

3.8.3 Membrane Resistance

The passive resistance of a membrane is a measure of its permeability to ions. In saline solutions the resistivity of pure phospholipids is as high as $10^{13} \Omega \text{ m}$. This can

be compared to 298 K values that range from 0.6 to 0.8 $\Omega \text{ m}$ for prepared physiological solutions (buffers). The value for sea water is $\sim 0.2 \Omega \text{ m}$. A 4 nm thick lipid bilayer can be estimated to have a specific resistance of $40 \text{ k}\Omega \text{ m}^2$. The significantly lower resistivities of biological membranes (typically $0.01 \sim 1 \Omega \text{ m}^2$) therefore can be assumed to arise from structures other than the lipid bilayer itself. These structures are protein-bounded aqueous pores (aquaporins) and various ion channels embedded in the lipid. The density of different channels typically range from 50 to 500 per μm^2 , with conductances of $1 \sim 100 \text{ pS}$. However, many of these channels may not be ‘open’ at any given time.

If a step pulse of steady current is applied across the membrane, the membrane potential shifts by ΔV_m from the resting value. ΔV_m depends on the magnitude of the applied current ΔI and the membrane resistance R , which can be determined from Ohm’s Law:

$$\Delta V_m = R \Delta I$$

Consider two spherical cells, one small and the other large and both with membranes having the same *specific resistance* R_m to electric current (i.e. the same resistance per unit square area of membrane). For a given increment of current ΔI inserted into the cells, the large cell will show a smaller increment of voltage ΔV_m because the same current will flow through a larger area of membrane. Because the input resistance of a cell (i.e. the total resistance encountered by current flowing into or out of a cell) is a function of both membrane area A and specific resistance R_m of a cell, it is useful when comparing membranes of different cells to correct for the effect of membrane area on the current density. Thus, the specific membrane resistance is calculated as:

$$R_m = RA = \frac{\Delta V_m}{\Delta I} A \quad (\text{ohms m}^2).$$

3.8.4 Membrane Capacitance

Because they are very thin ($\sim 4 \text{ nm}$) and virtually impermeable to ions over most of their surface area, cell membranes can violate the principle of electroneutrality at the *microscopic* scale. Negative charges accumulated at or near one surface of a membrane will interact electrostatically over the short distance of the membrane thickness, with positive charges on the other side of the membrane. The ability of the cell membrane to accumulate and separate electric charge is called its membrane capacitance. Electronic engineers can view this situation as a very thin dielectric (the lipid bilayer) sandwiched between two conductors (electrolytes) representing the basic form of a capacitor. Cell membranes contain a lipid bilayer of about 3 nm in thickness (verified by electron microscopy) with proteins protruding on each side. Assuming a total thickness of 4 nm for the insulating region of the membrane, and that the lipid content has a relative permittivity of 2.35 (as estimated in Section 3.2.4), the membrane capacitance for a smooth cell surface can be calculated to be $\sim 5.2 \text{ mF/m}^2$.

The equivalent circuit for a cell membrane to describe the charging and discharging of the membrane on application and then removal of a current pulse is shown in Figure 3.23. The

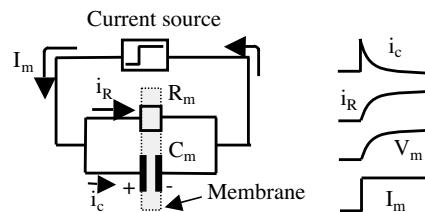


Figure 3.23 The equivalent circuit for a cell membrane can be represented as a parallel combination of the membrane resistance R_m and capacitance C_m . Time courses are shown for the resistive current i_R , capacitive current i_C , membrane potential V_m (across the membrane resistance and capacitance) on induction of a membrane current (I_m) pulse.

relationship between potential V and time during the charging of the membrane capacitance is given by:

$$V(t) = V_o e^{-\tau/R_m C_m},$$

where the time to fall to $1/e$ of its initial value is the time constant given by $\tau = R_m C_m$. Having determined the membrane specific resistance R_m , the membrane capacitance can be determined from measurement of the membrane time constant (typically $0.1 \sim 10$ ms). Experimental values obtained for C_m are normally larger than the theoretical value of ~ 5 mF/m² determined above for a smooth membrane structure. The experimental value (typically $10 \sim 30$ mF/m²) obtained correlates closely with the extent to which the area of an otherwise smooth membrane surface is increased as a result of the presence of membrane folds and protuberances, such as blebs and microvilli, for example.

3.8.5 Extent of Ion Transfer Associated with the Membrane Resting Potential

We have learnt that cells have a negative resting membrane potential of about -70 mV. How many ions must be transferred across the membrane to produce this potential? We can derive an estimate of this by considering a cell of radius $10 \mu\text{m}$ having a membrane capacitance of 10 mF/m^2 . The total surface area of this cell is $4\pi 10^{-10} \text{ m}^2$, to give a total membrane capacitance of $4\pi 10^{-12} \text{ F}$. To set up a potential of 70 mV will require a charge Q given by $Q = VC = 70 \times 10^{-3} \times 4\pi 10^{-12} = 8.8 \times 10^{-13} \text{ C}$. Dividing by the Faraday constant ($9.65 \times 10^4 \text{ C/mol}$) leads to the result that we require the equivalent of $9.1 \times 10^{-18} \text{ mol}$ of monovalent ions to be transferred across the membrane to generate a membrane potential of -70 mV. The cell volume is $(4\pi 10^{-15})/3 \text{ m}^3 = (4\pi 10^{-12})/3 \text{ litres}$. The potassium content of a cell of this volume, when the potassium is present at 150 mM , is roughly $(4\pi 10^{-12})/3 \times 0.15 = 6.3 \times 10^{-13} \text{ moles}$ (we will assume an activity coefficient of unity). Thus, to charge the membrane to -70 mV requires as little as $1.4 \times 10^{-3} \%$ of the cell's total potassium to be transferred across the membrane. The rule of electroneutrality – that positive charges must equal negative charges remains essentially unviolated at the *macroscopic* scale. The imbalance of charges exists only at the *microscopic* scale across the membrane thickness.

3.9 Membrane Equilibrium Potentials

The electrical energies of the transmembrane potentials of cells are responsible for nearly all the electrical phenomena that occur in the animal body. These potentials originate from two features of biological membranes:

- assymetrical distribution of ions between the intracellular and extracellular compartments;
- selective permeability of the membrane.

Consider the chamber shown in Figure 3.24. It is divided into two compartments by a membrane that is selectively permeable to potassium ions. We will assume that both compartments initially contain a 1 mM KCl solution. Electrodes inserted into the compartments would record that no potential difference occurs across the membrane. The hypothetical membrane is permeable to K^+ but *not* to Cl^- , and so it is possible for the K^+ ions to diffuse across the membrane on their own. On average for every potassium ion that passes in one direction through the membrane another potassium ion will pass through in the opposite direction. As long as the two compartments contain the same concentration of KCl the net flux of K^+ ions is zero and the potential difference across the membrane remains zero.

If the concentration of KCl in compartment 1 is suddenly increased to 10 mM, as shown in Figure 3.24b, a net diffusion of K^+ ions will take place through the potassium-selective membrane from compartment 1 to compartment 2. This net transfer of positive ions will create a potential difference across the membrane such that compartment 1 is more negative than compartment 2. The K^+ concentration gradient across the membrane represents a chemical potential difference that initially drives diffusion of K^+ from compartment 1 to 2. Each additional K^+ that diffuses from 1 to 2 adds its positive charge to that side, and Cl^- is left behind since it cannot cross this hypothetical membrane. This generates an electrical potential difference (a back emf). Thus, each K^+ ion now entering the membrane has two forces acting on it: a *chemical* potential difference. favouring net K^+ flux from 1 to 2, and an *electrical* potential difference. favouring net K^+ flux from 2 to 1. These two opposing forces

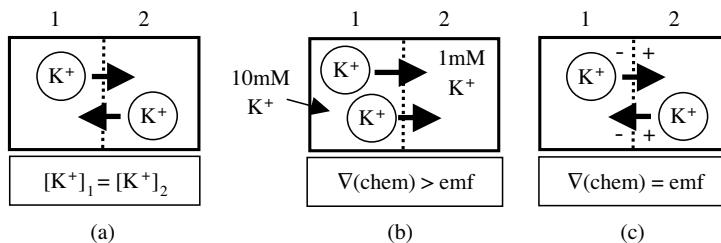


Figure 3.24 (a) A chamber divided into two compartments by a membrane permeable only to K^+ ions. Each compartment contains 1 mM KCl. The net flux of K^+ ions across the membrane is zero and no potential difference occurs across the membrane. (b) The concentration of KCl in compartment 1 is increased to 10 mM. K^+ ions diffuse across the membrane down their chemical gradient $\nabla(\text{chem.})$ and initiate a back emf across the membrane. (c) At electrochemical equilibrium the electrical potential difference across the membrane that exactly opposes diffusion down the chemical gradient is termed the equilibrium potential for the K^+ ion.

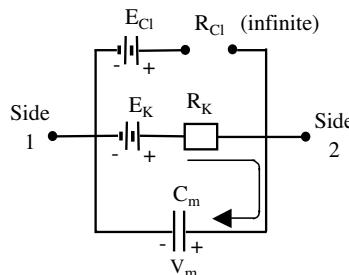


Figure 3.25 Equivalent circuit for the process leading to the electrochemical equilibrium process shown in Figure 3.22. E_K supplies the emf for K^+ to carry current through the membrane's potassium channel, which then creates a positive charge on side 2 of the membrane capacitance C_m .

come into equilibrium and remain balanced, as depicted in Figure 3.24c. The potassium ion is then said to be in *electrochemical equilibrium*. The potential difference that is established across a membrane in this way is termed the *equilibrium potential* for the ion in question (in this case, the potassium equilibrium potential, E_K). In our hypothetical situation here this potential difference E_K will be maintained indefinitely provided there is no leakage of Cl^- ions across the membrane. As discussed in Section 3.8.5, when considering the development of a potential across a cell membrane, very few ions actually diffuse across a unit area of the membrane before the equilibrium potential E_K is established. The concentrations of potassium ions in compartments 1 and 2, 10 mM and 1 mM respectively, therefore remain virtually unchanged during the overall process shown in Figure 3.24.

If for any reason the membrane potential V_m is not at the equilibrium potential E_X for an ion X, there will exist an emf acting on that ion, emf_X , equal to the difference between V_m and E_X :

$$\text{emf}_X = V_m - E_X.$$

Clearly, when $V_m = E_X$, ion X will experience no emf and will be in electrochemical equilibrium across the membrane.

An equivalent electrical circuit is shown in Figure 3.25 for the process shown in Figure 3.22 of the development of the membrane potential. Positive charge (in the form of potassium ions) driven by the emf acting on potassium (i.e. $V_m - E_K$) leaks through the potassium conductance (i.e. R_K) of the membrane so as to accumulate on the other side of the membrane. When the voltage across the capacitance of the membrane equals the potassium equilibrium potential (i.e. when $V_m - E_K = 0$) net diffusion of K^+ ions ceases and the system is at equilibrium – side 2 positive with respect to side 1. Although the electrochemical gradient for the chloride ion is in the opposite direction, it has no effect because our hypothetical membrane is impermeable to chloride ions (i.e. R_{Cl} in Figure 3.25 is effectively an open circuit).

3.10 Nernst Potential and Nernst Equation

The Nernst equation is one of the most widely used mathematical relationships in studies of bioelectric phenomena. Its derivation is based on the concept of a thermodynamic

equilibrium between the osmotic work that is required to move a given number of ions across a membrane in one direction, and the electrical work required to move the same number of charges back across the membrane in the opposite direction. The potential across the cell membrane that exactly opposes net diffusion of a *particular* ion through the membrane is called the Nernst potential for *that* ion. The magnitude of this potential is determined by the ratio of the concentrations of that specific ion on the two sides of the membrane. The greater this ratio the greater is the tendency for that ion to diffuse in one direction, and thus the greater the potential V required to prevent its diffusion.

The free energy G of a system is only really of interest when that system undergoes some kind of change that results in a change ΔG of the free energy. Free energy has been defined such that ΔG directly measures the amount of disorder created. According to the 2nd law of thermodynamics a physico-chemical reaction can proceed spontaneously only if this results in a net increase of disorder (entropy) of the total system. Energetically favourable processes are therefore those that decrease free energy and have a negative ΔG , and the relevant example of this here is the diffusion of an ion from a region of high to one of a low ionic concentration.

Movement of an ion down its concentration gradient across the plasma membrane and into a cell is accompanied by a favourable free-energy change per mole:

$$\Delta G_{\text{conc}} = -RT \ln \frac{[\text{Conc outside}]}{[\text{Conc inside}]}, \quad (3.51)$$

where R is the universal gas constant: ($R = 8.31 \text{ J K}^{-1} \text{ mol}^{-1}$) and T is the absolute temperature (Kelvin). (As noted earlier in this chapter, solute molecules in solution behave thermodynamically like gas molecules. We can therefore use the ideal gas law $PV = nRT$, with n the number of moles of the solute molecule.) Moving the ion into a cell across a membrane whose inside is at a potential V relative to the outside will cause an additional free-energy change (per mole of ion moved) given by:

$$\Delta G_V = zFV, \quad (3.52)$$

where F is the Faraday Constant ($F = 9.648 \times 10^4 \text{ C mol}^{-1}$) and z is the number of charges on the ion. At the point where the concentration and potential gradients just balance $\Delta G_{\text{conc}} + \Delta G_V = 0$. From Equations (3.51) and (3.52) this leads to:

$$zFV - RT \ln \frac{[\text{Conc outside}]}{[\text{Conc inside}]} = 0$$

and from this we obtain the Nernst Equation:

$$V = \frac{RT}{zF} \ln \frac{[\text{Conc outside}]}{[\text{Conc inside}]} = 2.3 \frac{RT}{zF} \log_{10} \frac{[\text{Conc outside}]}{[\text{Conc inside}]} \quad (3.53)$$

For a univalent cation, $z = +1$, and $2.3 \frac{RT}{zF} = 59.1 \text{ mV}$ at 298 K . At 298 K , we obtain a value for the Nernst Potential V of -59.1 mV for the case where $C_{\text{in}}/C_{\text{out}} = 10$. If a univalent cation diffuses down this concentration gradient to the outside of the cell, V will become less negative. Thus, for the situation $C_{\text{in}}/C_{\text{out}} = 1$, $V = 0$.

Table 3.4 The intra- and extracellular ion concentrations, together with the equilibrium potentials and ion channel conductance values, for frog muscle membranes at 298 K. (Derived from Ove Sten-Knudsen, Biological Membranes, Cambridge University Press, 2002, pp. 389 and 391)

Ion species	Intracellular conc. (mM)	Extracellular conc. (mM)	Equilibrium potential (mV)	Ion channel conductance G (S/m ²)
Na ⁺	13	110	+55	0.8×10^{-2}
K ⁺	138	2.5	-103	85×10^{-2}
Cl ⁻	3	112.5	-93	170×10^{-2}

The intra- and extracellular ion concentrations, together with the equilibrium potentials and ion channel conductance values, are given for frog muscle membranes in Table 3.4.

From Table 3.4 we obtain values for [K⁺]_{out} and [K⁺]_{in} of 2.5 mM and 138 mM, respectively, for a frog muscle cell. From the Nernst Equation we calculate that the K⁺ equilibrium potential (V_K) at 298 K is -103 mV (agreeing with the value given in Table 3.4). At this potential difference of -103 mV across the membrane there is no net flow of K⁺ across the membrane. For any particular membrane potential V_m the net force tending to drive a particular type of ion out of the cell through that ion-specific channel is proportional to the difference between V_m and the equilibrium potential for the ion. For K⁺ it is $V_m - V_K$. The ion current through the potassium channel will be $(V_m - V_K)G_K$ where G_K is the conductance represented by the number of potassium channels per unit area of membrane.

In the situation where the membrane potential V_m is -103 mV, from Table 3.4 we can see that, although the potassium ion current through the membrane would be zero, there would be a sodium ion current of $(V_m - V_{Na})G_{Na} = -158 \times 0.8 \times 10^{-2} = -1.3 \text{ mA/m}^2$, as well as a chloride ion current = -17 mA/m². This gives a net membrane current of -18.3 mA/m² (i.e. current flowing out of the cell across the membrane).

How then, in the presence of these various ion concentrations, and ion channels with their different equilibrium potentials and conductance values, does the membrane potential ever attain an equilibrium value? Can we calculate such an equilibrium membrane potential? The answers to these questions form our next subject matter.

3.11 The Equilibrium (Resting) Membrane Potential

It is clear from Section 3.10 that the relative trans-membrane concentrations, together with the ease with which different ions can cross the membrane, determine their relative contributions to the potential they produce in diffusing across the membrane. On this basis, and by making the assumption that there is a uniform gradient of potential in going from one side of the membrane to the other side, Goldman [8] derived the following equation for the steady state resting potential V_m across a cell's membrane (taking into account the dominant ions that can permeate through that membrane):

$$V_m = \frac{RT}{F} \ln \left(\frac{P_{Na}[Na^+]_{out} + P_K[K^+]_{out} + P_{Cl^-}[Cl^-]_{in}}{P_{Na}[Na^+]_{in} + P_K[K^+]_{in} + P_{Cl^-}[Cl^-]_{out}} \right), \quad (3.54)$$

where P_{ion} is the permeability for that ion (m/sec). This is commonly referred to as the Goldman equation, but is also known as the Goldman-Hodgkin-Katz or GHK equation. Values for

Table 3.5 Values for the intra- and extracellular ion concentrations, together with the membrane permeabilities, for the frog muscle membrane (derived from Ove Sten-Knudsen, *Biological Membranes*, Cambridge University Press, 2002, pp. 389 and 391)

Ion species	Intracellular conc. (mM)	Extracellular conc. (mM)	Permeability P (m/sec)
Na^+	13	110	0.02×10^{-8}
K^+	138	2.5	2.0×10^{-8}
Cl^-	3	112.5	4.0×10^{-8}

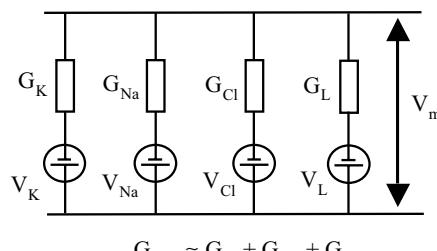
the intra- and extracellular ion concentrations, together with the membrane permeabilities, are given in Table 3.5 for the frog muscle membrane.

Inserting the values from Table 3.5 into Equation (3.54):

$$V_m = \frac{RT}{F} \ln \frac{0.02 \times 10^{-8}.110 + 2.0 \times 10^{-8}.2.5 + 4.0 \times 10^{-8}.3.0}{0.02 \times 10^{-8}.13 + 2.0 \times 10^{-8}.138 + 4.0 \times 10^{-8}.112.5} = -93.4 \text{ mV.}$$

Another way to evaluate the total current and membrane potential for a membrane is to consider the membrane as a heterogeneous structure consisting of a variety of distinct elements that are arranged in parallel in the membrane. This so-called *mosaic* membrane model consists of a matrix into which are embedded numerous, closely packed, minute cylinders that all penetrate the membrane in its entirety. These cylinders (*ion channels*) have the unique property of displaying *ion-specific permeability*.

Consider the mosaic membrane, shown in Figure 3.26, equipped with three different types of ion channel specifically permeable to Na^+ ions, K^+ ions, and Cl^- ions. On each side of the membrane are different concentrations of Na^+ , K^+ and Cl^- ions, which will strive by the diffusional forces to attain equal concentrations on both sides for all three kinds of ion. However, before this situation can arise this diffusion process will have established a *diffusion potential* across the membrane and also across all three types of ion channel. We will now derive the magnitude of the equilibrium *membrane potential* V_m .



$$G_{\text{total}} \approx G_K + G_{\text{Na}} + G_{\text{Cl}}$$

Figure 3.26 Equivalent circuit for a mosaic membrane possessing three different ion channels. The conductances G are represented by the number of ion channels per unit area for each ionic species. The battery V_L represents the contribution to the membrane current from other ionic sources. These other sources are relatively insignificant.

As discussed in Section 3.10, the currents that each type of ion carry are given by $(V_m - V_{ion})G_{ion}$ where G_{ion} is the membrane conductance for that ion channel type:

$$I_{Na} = G_{Na}(V_m - V_{Na})$$

$$I_K = G_K(V_m - V_K) \quad ,$$

$$I_{Cl} = G_{Cl}(V_m - V_{Cl})$$

where V_m is the equilibrium membrane potential.

When the equilibrium membrane potential V_m is established, the sum of these ion channel currents will be zero (Kirchoff's 2nd law):

$$G_{Na}(V_m - V_{Na}) + G_K(V_m - V_K) + G_{Cl}(V_m - V_{Cl}) = 0.$$

Solving for V_m leads to the result:

$$V_m = \frac{G_{Na}V_{Na} + G_KV_K + G_{Cl}V_{Cl}}{G_{total}} = \frac{G_{Na}}{G_{total}}V_{Na} + \frac{G_K}{G_{total}}V_K + \frac{G_{Cl}}{G_{total}}V_{Cl}. \quad (3.55)$$

The steady state membrane potential can thus be described as the sum of the products of the equilibrium potential and membrane conductance for each participating ion divided by the total conductance of the membrane for the ions that cross the membrane. We can define the *Transferance Number* or *Transport Number* T_j of an ion as $T_j = G_j/G_{total}$. These quantities will obey the relation:

$$\sum \frac{G_j}{G_{total}} = \sum T_j = 1. \quad (3.56)$$

Thus, although the equilibrium potential V_j for a particular ion may be very large, its contribution to the membrane potential V_m may be insignificant if the permeability to that particular ion is much less than those of the other ions ($T_j \ll 1$). Although the procedures for formulating Equations (3.54) and (3.55) are fundamentally different, they both express the same result, namely that the magnitude of the membrane potential is determined by that ion whose flux through the membrane is dominant.

From Table 3.4 for frog muscle, $G_{total} = G_{Na} + G_K + G_{Cl} = 2.56 \text{ S/m}^2$. From Equation (3.55):

$$V_m = 0.003 \times 55 + 0.332(-103) + 0.665 \times (-93) = -95.9 \text{ mV}.$$

This result is close to the value -93.4 mV obtained using Equation (3.54) and to the experimental result obtained by Hodgkin and Horowicz [9] of -95 mV .

3.12 Membrane Action Potential

In Section 3.8 we learnt that in response to an appropriate (electrical) stimulus the membranes of excitable cells, such as nerve and muscle, exhibit a strong active response known as the action potential.

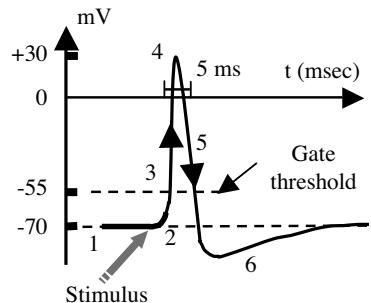


Figure 3.27 The action potential initiated by the electrical stimulation of an axon membrane is shaped by a rapidly activating inward flow of Na^+ ions into the cell followed by a more slowly activating outward K^+ current. The six main processes involved are described in the main text.

3.12.1 Nerve (Axon) Membrane

A schematic of an action potential generated by a nerve axon is shown in Figure 3.27.

The action potential takes the form of a depolarisation from its resting state ($\sim -70 \text{ mV}$) followed by repolarisation to that resting state. In other words the potential of the membrane's inner surface briefly goes from a negative to a positive potential with respect to the outer medium, before reestablishing the rest potential. Hodgkin and Huxley [10] in work that led to the award of the Nobel Prize in 1963 demonstrated that the action potential involves an early inward current of Na^+ ions (down its electrochemical gradient) into the cell followed by a later outward current of Na^+ ions (also down its electrochemical gradient).

The process involves several steps:

1. The condition at the resting membrane potential of around -70 mV . (Na^+ is more concentrated on the outside and K^+ on the inside of the axon membrane.)
2. An electrical stimulus causes 'fast' voltage-gated Na^+ channels to open. If the opening is sufficient to drive the interior potential from -70 to -55 mV the process of activation continues. In other words there is an action threshold potential corresponding to where the membrane is depolarised by a stimulus of $\sim 15 \text{ mV}$.
3. Having reached the action threshold more Na^+ channels open. The Na^+ influx drives the interior of the cell membrane up to about $+30 \text{ mV}$. The process to this point is called depolarisation.
4. After $\sim 1 \text{ ms}$, the Na^+ channels close and 'slow' K^+ channels open. Because the K^+ channels are much slower to open the depolarisation has time to be completed. (Having both Na^+ and K^+ channels open at the same time would drive the membrane potential towards its resting value and prevent the creation of the action potential.)
5. With the voltage-gated K^+ channels open the membrane begins to repolarise back towards its resting potential.
6. The repolarisation typically overshoots the resting potential to about -90 mV . This is called hyperpolarisation and is important for the transmission of signals (at $\sim 100 \text{ m/sec}$) along the neuron's axon. Hyperpolarisation prevents the neuron from receiving another stimulus during this time, or at least raises the threshold for any new stimulus. Part of the importance of hyperpolarisation is in preventing any stimulus already sent along an axon

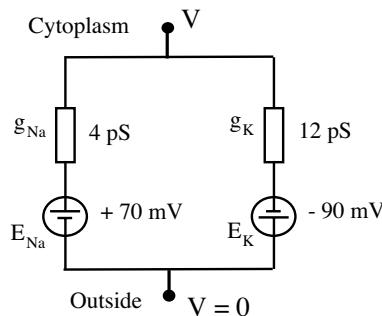


Figure 3.28 The equivalent circuit to describe the contributions of the sodium and potassium pumps in producing an action potential in an axon membrane is shown in Figure 3.26. Included in this figure are the approximate values for the Na and K conductances (g_{Na} and g_{K}) derived from Hille ([11], pp. 329–30) and the equilibrium potentials (E_{Na} and E_{K}) for these two ions.

from triggering another action potential in the opposite direction. In other words, hyperpolarisation assures that the signal is proceeding in one direction.

7. After hyperpolarisation the sodium-potassium pump eventually brings the membrane back to its resting state of -70 mV .

The equivalent circuit to describe the contributions of the sodium and potassium pumps in producing an action potential in an axon membrane is shown in Figure 3.28. Included in this figure are the approximate values for the Na and K conductances (g_{Na} and g_{K}) and the equilibrium potentials (E_{Na} and E_{K}) for these two ions.

During any time interval the electrical charging of the membrane capacitance is determined by the magnitude and polarity of resultant current generated by the equivalent circuit. The voltage V appearing across the membrane will change at a rate dictated by this current and the membrane capacitance according to the relationship:

$$C_m dV/dt = -g_{\text{Na}}(V - E_{\text{Na}}) - g_{\text{K}}(V - E_{\text{K}}).$$

The way in which action potentials are propagated along the axon of a neuron will be described in Section 3.13. For now it is sufficient to state that the action potential does not decrease in strength with distance propagated along an axon. It is an *All-or None* phenomenon – action potentials either happen completely or not at all.

3.12.2 Heart Muscle Cell Membrane

The action potentials in an axon membrane are produced by voltage-gated sodium and potassium channels. For heart muscle membranes voltage-gated calcium (Ca^{2+}) channels, instead of sodium channels, play important roles. For example, in ventricular myocytes, high-voltage activated Ca^{2+} channels contribute towards keeping the cell depolarised for several hundred milliseconds. Cardiac pacemaker cells possess no functional Na^+ channels, and so their entire action potentials are generated by voltage-gated opening of Ca^{2+} channels. Electrical stimulations arising from action potentials of a heart muscle cell are conducted from one cell

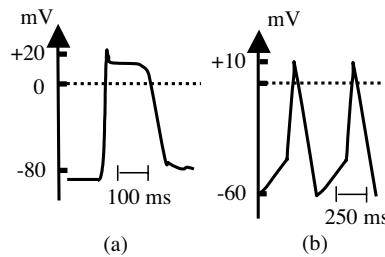


Figure 3.29 (a) Cardiac action potential. (b) Pacemaker cell action potential. (Derived from [11].)

to all the cells that are adjacent to it, and thence to all the cells of the heart. A typical action potential exhibited by a cardiac muscle cell and a pacemaker cell are shown in Figure 3.29.

The resting membrane potential phase (~ -85 mV) of the cardiac action potential is associated with diastole of the chamber of the heart, the period of time when the heart fills with blood after systole (heart contraction). If the resting membrane potential becomes too positive, the cell may not be excitable, and conduction through the heart may be delayed, increasing the risk for arrhythmias (irregular heart beats). The ‘plateau’ phase of the cardiac action potential is sustained by a balance between inward movement of Ca^{2+} through calcium channels and outward movement of K^+ through the K^+ channels. During the rapid repolarisation phase the Ca^{2+} channels close, while the K^+ channels remain open, ensuring a net outward current and a return to the resting potential.

Two voltage-dependent calcium channels play critical roles in the physiology of cardiac muscle, namely the L-type ('L' for Long-lasting) and T-type ('T' for Transient) voltage-gated calcium channels. These channels respond differently to voltage changes across the membrane. L-type channels respond to higher membrane potentials, open more slowly, and remain open longer than T-type channels. Because of these properties, L-type channels are important in *sustaining* an action potential, while T-type channels are important in *initiating* them. Because of their rapid kinetics, T-type channels are commonly found in cells undergoing rhythmic electrical behaviour. T-type calcium channels are also found in the so-called pacemaker cells of the heart, which control the heart beat. (T-type channels are also commonly found in some neuron cell bodies involved in rhythmic activity such as walking and breathing.) L-type channels are the targets of a class of drugs called dihydropyridines, which block the currents produced by these channels.

In addition to stimulus from adjacent cells, certain cells (pacemaker cells) of the heart have the ability to undergo *spontaneous depolarisation*, in which an action potential is generated without any electrical stimulation from nearby cells. This is called cardiac muscle automaticity. The cells that can undergo spontaneous depolarisation the fastest are the primary pacemaker cells of the heart, and set the heart rate. This spontaneous depolarisation is due to the plasma membranes within the heart that have reduced permeability to potassium ions but still allow passive transfer of calcium ions, allowing a net charge to build. The normal activity of the pacemaker cells of the heart is to spontaneously depolarise at a regular rhythm, whilst abnormal automaticity involves the abnormal spontaneous depolarisation of cells of the heart. This abnormality typically causes arrhythmias (irregular rhythms) in the heart.

3.13 Channel Conductance

Conductance values are given in Figure 3.28 for the sodium and potassium selective pumps. These pumps are envisaged to take the form of aqueous pores or channels through the cytoplasmic membrane. Supporting evidence for this concept is the high permeability and high ionic throughput rates measured for single channels. A simple model for a membrane channel is shown in Figure 3.30. It takes the form of a cylinder that spans across the membrane, and is open to the solutions on either side of the membrane. The overall effective electrical resistance of the channel consists of three components, namely the resistance of the cylindrical pore itself together with the resistance of the regions adjacent to the open ends of the channel ([11] Chapter 11).

A well-studied example of the physical and electrical properties of a channel is that of the mechanically gated MscL channel located in the cell envelope of bacteria. This channel enables fast adjustments of turgor pressure in response to a sudden reduction of osmotic pressure. When the tension forces acting on the membrane approach the point where lysis of the membrane can occur, the MscL channel forms a large nonselective pore and acts as a safety valve by releasing osmolytes from the cell interior. The MscL channel has no selectivity towards anions and cations, and its effective conductance is directly proportional to the conductivity of the surrounding bulk electrolyte. This channel appears therefore to act as a classical ohmic resistor. The conductance G of a cylindrical pore of the form shown in Figure 3.30 can therefore be calculated as:

$$G = \frac{\sigma A}{l} = \frac{\sigma \pi r^2}{l}, \quad (3.56)$$

where σ is the conductivity of the solution filling the pore, and l and r are the length and radius of the pore, respectively. For the MscL channel from *E. coli* values for r and l are approximately 1.6 nm and 4 nm, respectively [12]. The conductivity of a typical electrolyte for mammalian cells is 1.7 S m^{-1} . Using this conductivity value in Equation (3.56) gives a channel conductance of 3.4 nS . This conductivity is three-orders greater than the value of 4 pS given for the selective sodium channel in Figure 3.28 and leads to the question as to the relevance of Equation (3.56) for other ion channels. In Chapter 10 the dimensionless Knudsen number is introduced as the means for deciding whether or not a microfluidic system can be analysed using macroscopic concepts. Equation (3.56) represents a macroscopic

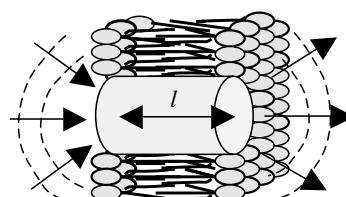


Figure 3.30 A simple model for a membrane channel consists of an open cylinder that spans the membrane. Its overall electrical resistance comprises the resistance of the cylindrical pore when filled with electrolyte together with the resistances of the regions leading up to the open ends of the channel.

description of channel conductance. The Knudsen number (Kn) is the ratio of the inter-molecular spacing of the fluid (i.e. the diameter 0.28 nm of a water molecule) to the characteristic dimension of the fluidic system, which for a cylindrical channel is its diameter. For the MscL channel $\text{Kn} = 0.28/3.2 = 0.09$. In Chapter 10 we find that this places the MscL at close to the limit where a continuum, macroscopic, model is appropriate. This is consistent with the finding that the conductance of the MscL channel is proportional to a wide range of electrolyte conductivity values (see [12]).

Two other well studied channels are the acetylcholine receptor (AChR) and the gramicidin A channel. The AChR channel has a radius of 0.3 nm and a length of 0.6 nm, and its structure contains rings of negatively charged amino acids at the extracellular and the cytoplasmic portions of the channel [13]. Equation 3.56 predicts an AChR channel conductance value of 0.6 nS in a 100 mM KCl solution (1.29 S m^{-1}). The actual measured conductance is 0.08 nS [13]. This near eightfold discrepancy will arise from the fact that the corresponding Kn value of 0.47 places the problem of calculating the conductance of the AChR channel into the meso-scale region between the continuum approximation and a model involving discontinuous, dynamic, molecular physics (see Chapter 10). The presence of the rings of negative charges at the openings of the AChR channel also indicate that its effective conductance will also be determined by the rates at which ions can diffuse towards (or be repelled from) the channel openings. The gramicidin A channel has an effective radius of 0.4 nm, a length of 2.5 nm and a measured conductance for K^+ ions in 0.1 M KCl of 21.6 pS ([11], Chapter 11). The theoretical conductance based on Equation (3.56) is 0.26 nS, a value some 12-times larger than the measured one. The Kn value for the gramicidin channel is 0.7, which like the AChR channel places it in the meso scale and explains why a macroscopic model is not appropriate for calculating its conductance. Some of the molecular models that have been explored to understand the conductance of the smallest diameter ion channels are described by Hille ([11], Chapter 11).

3.14 The Voltage Clamp

To explain the ‘overshoot’ shown in Figure 3.27 of the membrane potential to $+30 \sim +40 \text{ mV}$ observed at the peak of the action potential, Hodgkin *et al.* [14] formulated the so-called *sodium hypothesis*. Basically, this assumes that the initial change in the membrane potential only consists of a selective increase in the permeability to sodium that is large enough to dominate the diffusion regime for a short time. In the extreme case one might expect an overshoot of $\sim 60 \text{ mV}$ (the equilibrium potential V_{Na} for the sodium ions) but never a substantially higher value. They demonstrated that replacement of the extracellular NaCl by choline chloride, glucose or sucrose, molecules which do not penetrate the membrane, resulted in a reduction of the action potential in proportion to the reduction of the extracellular Na^+ concentration, whereas the resting membrane potential remained unchanged. Replacement of the normal extracellular fluid by a hypertonic solution having an excess of sodium resulted in an increase in the overshoot of a magnitude that fitted with that predicted by the Nernst equation.

The amounts of Na^+ and K^+ entering or leaving the axoplasm during the activity of the membrane were determined using of radioactive tracers. Measurements were performed on the squid axon because its diameter of $500 \sim 1000 \mu\text{m}$ allows capillary electrodes to be readily inserted into the axoplasm (mammalian nerve axons have much smaller diameters of

less than $10\ \mu\text{m}$). It was found that in the course of an action potential there was a net Na^+ entry of about $4\ \text{pmol}/\text{cm}^2$ ($\sim 20\,000$ ions across $1\ \mu\text{m}^2$) and a K^+ loss from the cytoplasm of the same amount. However, these experiments did not provide any information on the temporal course of the inward and outward flows of sodium and potassium. If these two oppositely directed ionic movements do generate a change in the membrane potential of the form of the action potential, the first event must be a charging of the membrane's inside by an inwards-directed current to a positive value (e.g. $+40\ \text{mV}$) that is followed by an outward-directed current that leads to the repolarisation of the membrane back down to around $-70\ \text{mV}$. If these currents are carried by Na^+ and K^+ ions there must be a time lag between the Na^+ entry and the K^+ outflow.

Any attempt to demonstrate the separate contributions of these two currents on the propagating action potential would face the problem that the membrane potential changes with time along the length of the axon (in the manner of a wave packet). Consequently, the membrane current is composed partly of ionic currents crossing the membrane and partly of a component used to charge the membrane capacitance C_m to a changing membrane potential ($i = dq/dt = C \cdot dV/dt$). Therefore, to measure the ionic currents a method was required to eliminate the complications arising from the charging of the membrane capacitance. This method is called the Voltage Clamp technique [14].

By means of the voltage clamp technique (described in more experimental detail in Chapter 8) the membrane potential can be maintained (clamped) at an arbitrary prechosen value and then changed and clamped almost instantaneously to a new chosen value. This can be achieved irrespective of the changes in the ionic currents that might follow as a result of changes in the driving force on the ions in the membrane, and changes in the membrane's permeability to one or several ionic species. Because of the instantaneous potential displacement from one level to another, the membrane capacitance changes charge only at the instant of the membrane potential change. Therefore, the currents that may be measured during the voltage clamp (where $dV/dt = 0$) are exclusively ionic currents that flow through the membrane. This is equivalent to connecting the axon's inside and outside with a controllable constant voltage generator. As described in Chapter 8, this is achieved by inserting into the axoplasm an extra electrode (a so-called current electrode) that is connected to an electronic feedback circuit that, despite changes in membrane permeabilities, supplies a current of just the strength and direction to ensure that the membrane potential remains at a given predetermined level. Changes in the membrane current with time at this clamped membrane potential will provide information about the changes in the membrane permeability to the surrounding ions.

3.15 Patch-Clamp Recording

The direct way to investigate the functioning of membrane ion channels is to record the current which flows through an open channel, or to measure the changes in membrane potential produced by an imposed current. As depicted in Figure 3.20, the potential across a cell membrane can be measured by inserting a glass microelectrode into the cell and measuring the difference between its recorded potential and one registered by a reference electrode located in the extracellular medium. The voltage-clamp technique allows the membrane potential to be held at a constant value so that the current that flows through the membrane at any particular potential can be measured. However, some cells are too small to allow their penetration

by a microelectrode. Also, the plasma membranes of neurons and muscle cells contain a very large number of voltage-gated ion channels, and the membranes of cells that do not exhibit electrical excitations contain a variety of other types of gated channel. The total current crossing a cell membrane is the algebraic sum of the currents flowing through all of these channels, which means that the functioning of a single ion channel is not possible using the conventional voltage-clamp technique. These issues can be overcome using the patch-clamp technique developed by Neher and Sakmann [15,16].

This technique (described further in Chapter 8) employs the fact that a clean, fire-polished, glass micropipette pressed against a cell can fuse to its membrane to form a very high resistance seal ($\geq 10^9 \Omega$) of good mechanical stability. This isolates a small patch of the membrane on the cell and the ion channels it contains can then be investigated through either electrical or chemical manipulation. The high resistance seal means that current can only enter or leave the micropipette through open channels in the isolated patch of membrane. Neher and Sakmann [15] were thus able to report the first recording of the activity of a single-channel (an acetylcholine-activated channel). By applying mild suction, a patch of membrane can be removed from a cell, so that the current through a single channel can be recorded as a function of different compounds exposed to the inside (cytoplasmic) membrane surface of a cell. By increasing the suction, an excised patch can also be prepared having its external membrane surface (outside-out) exposed to an outside solution. Measurements can be made of the millisecond kinetics for membrane ion currents as low as 10^{-12} A ($\sim 10^7$ ions/sec), with voltage clamping to give precise control over channel voltage-gating.

3.15.1 Application to Drug Discovery

The specific and regulated functioning of membrane ion channels plays important roles in many physiological processes. These include electrical signalling in the brain and heart, the secretion of hormones into the bloodstream, the transduction of sensory signals, the regulation of blood pressure and immune responses. Defects in ion-channel function can therefore result in profound physiological effects, and more than 55 different inherited ion channel diseases, termed as ‘channelopathies’, have in fact been identified [17]. Ongoing patch-clamp studies of ion-channel function and modulation, coupled with the identification of specific genetic defects that lead to ion-channel related diseases, are providing insights into the relationship between ion-channel structure and function. A well studied example is the potassium-ATP (K_{ATP}) channel, where mutations of its pore-forming proteins lead to an impaired ability of ATP to bind to the channel and thus to inhibit the channel’s transport of potassium ions. This can increase K_{ATP} channel currents sufficiently enough to reduce electrical activity in nerves and muscles, leading to such diseases as diabetes, epilepsy and muscle weakness. This understanding has led to the development of drugs that specifically block (K_{ATP}) channels. The therapeutic action of many existing drugs (e.g. local anaesthetics, sedatives, antianxiety and antidiabetic) is through their interaction with membrane ion channels. Table 3.6 summarises some of diseases that are related to ion channel dysfunction.

Pharmaceutical companies have in the past faced problems in developing high-throughput assays able to randomly screen their large libraries (typically $> 25\,000$) of potential ion-channel drugs. Many existing ion-channel drugs were developed without knowledge of the precise drug target and mode of action at the molecular level. Although molecular-induced modulation of ion channels can be observed using ion-sensitive or voltage-sensitive

Table 3.6 Some diseases related to ion channel dysfunction (derived from [17])

Ion Channel	Diseases
K^+	Diabetes; epilepsy
Na^+	Epilepsy; heart; hypertension
Ca^{2+}	Angina; cardiac arrhythmia; epilepsy; hypertension; migraine; muscle weakness; chronic pain
Cl^-	Constipation; cystic fibrosis; deafness; epilepsy; kidney

fluorescent dyes, for example, this lacks the precision, temporal resolution, and voltage control that can be obtained using patch-clamp measurements. However, conventional patch-clamp studies are too technically demanding and laborious for the primary screening of potential ion-channel drugs. Recently, automated and medium-throughput techniques have been developed and are beginning to have an impact on the drug discovery ‘pipeline’ by providing high quality, information-rich, and biologically-relevant assays [17]).

3.16 Electrokinetic Effects

3.16.1 Electrophoresis

Electrically charged particles are induced to move in an electrical field – an effect called electrophoresis. In most practical situations the field is DC and spatially uniform. In this situation the force exerted on a particle carrying a net charge Q by an applied electric field E is given by:

$$F_e = QE. \quad (3.57)$$

The particle will accelerate until there is a balance between the electrophoretic force and the frictional force that opposes the particle’s motion through the fluid. To a first approximation we can consider the main resistive force to motion to be the Stokes viscous force:

$$F_S = 6\pi\eta av \quad (3.58)$$

where a and v are the particle’s radius and steady-state velocity, respectively, and η is the dynamic viscosity of the fluid medium. The steady-state velocity attained when the forces balance is:

$$v = \frac{Q}{6\pi\eta a}. \quad (3.59)$$

The electrophoretic mobility μ_e can then be defined as:

$$\mu_e = \frac{v}{E}. \quad (3.60)$$

However, this analysis does not take into account the fact that the moving particle will carry a thin layer of fluid along with it. We will assume that this fluid flow behaves in the same way

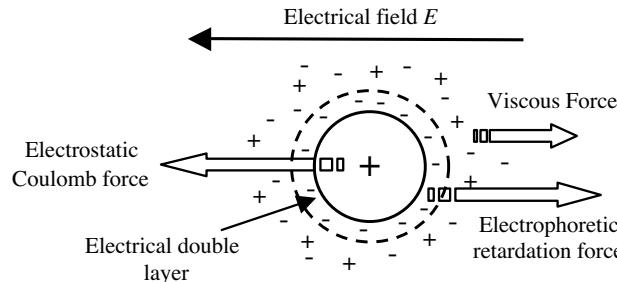


Figure 3.31 The steady-state velocity of a charged particle in an electric field is determined by the balancing of the Coulomb electrostatic force that accelerates the particle, with the Stokes viscous drag force and the retardation resulting from the interaction of the field on the counterions in the electrical double layer.

as the Couette flow described in Section 9.3.5 and Figure 9.6 of Chapter 9. The fluid layer immediately in contact with the particle's surface will match the velocity of that surface – we will have the condition known as zero slip at the interface between the particle surface and the fluid. This fluid layer will contain a distribution of counterions $\rho(x)$ of the form shown in Figure 3.4. Because, by definition, these counterions carry a charge of opposite polarity to that at the particle surface, they will experience a force given by Equation (3.57) but in the opposite sense to that experienced by the particle. This effect will produce a retarding force on the particle, so that the effective electrophoretic mobility will be less than that predicted by Equation (3.60). We have the overall scheme of forces acting on the charged particle as shown in Figure 3.31.

We will approximate the case shown in Figure 9.6, by assuming that the spatial gradient dv/dx of velocity of the fluid associated with the particle decreases linearly as a function of distance x from this interface. The shear stress τ exerted on each fluid layer is thus given by Equation (9.14) of Chapter 9:

$$\tau = \eta dv/dx.s, \quad (3.61)$$

where η is the dynamic viscosity of the fluid. At the steady-state condition of electrophoresis, the velocity of each fluid layer will be constant. The shear force and electric force acting on each volume element $A dx$ of the layer must therefore be equal and opposite:

$$\eta A \frac{dv}{dx} = E \rho(x) A dx$$

or

$$E \rho(x) = \eta \frac{d^2 v}{dx^2}.$$

Replacing $\rho(x)$ using the Poisson Equation (3.15) we obtain

$$-\epsilon_0 \epsilon_r E \frac{d^2 \phi(x)}{dx^2} = \eta \frac{dv}{dx}.$$

As boundary conditions for the integration of this equation, we can assume that as x tends to infinity the charged particle is effectively screened by the counterions (i.e. $\phi(x)$ tends to zero, and that the velocity of a fluid layer is zero (the charged particle is moving relative to the bulk fluid). We will also depart from the Gouy-Chapman model of the electrical double layer by defining the potential ζ as the potential at the boundary between the surface layer of fluid moving at the same velocity as the particle, rather than the potential exactly at the surface of the charged particle. On integration we thus obtain the result

$$-\varepsilon_0 \varepsilon_r E \zeta = \eta v(0).$$

Thus, by adopting the definition of electrophoretic mobility given in Equation (3.60), from the above equation we obtain the following relationship:

$$\mu_e = \frac{v}{E} = \frac{\varepsilon_0 \varepsilon_r \zeta}{\eta}. \quad (3.62)$$

This is known as the Helmholtz-Smoluchowski equation, and represents the improvement made by Smoluchowski in 1903 to Helmholtz's original theory. The potential ζ is known as the *zeta potential*. No adequate theory appears to definitely relate the zeta potential at the fluid slip plane to the electrostatic potential $\phi(0)$ used in Equation (3.31) for the Gouy-Chapman model, where it defines the potential right at a charged surface. Equation 3.62 works well in solutions of high ionic strength, where the double-layer thickness κ^{-1} , defined as the Debye screening length in Equation (3.24), is much less than the particle radius a (i.e. $a\kappa \gg 1$). This is the case for most aqueous electrolytes because the effective Debye screening length is at most a few nanometers (see Figure 3.5). However, for nano-sized colloidal particles and very low ionic strength solutions, as depicted in Figure 3.32, the Debye length can exceed the particle radius. The retardation force caused by the atmosphere of counterions thus acts much further from the particle surface and is reduced. For the case $a\kappa < 1$, an improvement to the theory introduced by Hückel in 1924 predicts the following relationship for the electrophoretic mobility of charged nanoparticles in very dilute electrolytes:

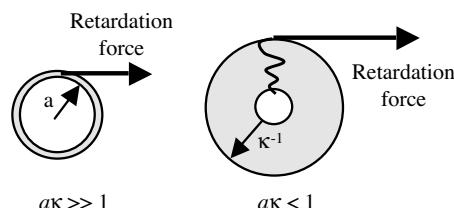


Figure 3.32 Equation 3.62 provides a good description of the electrophoretic mobility of a charged particle in an electrolyte of high ionic strength, where the thickness (Debye length $1/\kappa$) of the electrical double layer is much smaller than the particle radius a . Nano-sized particles in an electrolyte of low ionic strength can have a Debye length much larger than the particle radius. In this situation the retarding force transmitted to the particle is weakened and Equation (3.63) more accurately describes the electrophoretic mobility.

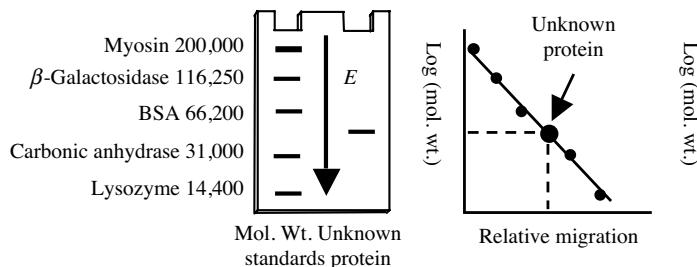


Figure 3.33 The molecular weight of an unknown protein can be determined by comparing its electrophoretic migration through a pH buffered gel against the behaviour of a set of known proteins. The stained protein samples are inserted at the top of gel chamber, and a high electric field E is applied using electrodes built into the gel chamber.

$$\mu_e = \frac{2\epsilon_r\epsilon_0\zeta}{3\eta}. \quad (3.63)$$

An important biological application of electrophoresis is in the separation of a mixture of proteins according to their size (molecular weight) and electrophoretic mobility. If an electric field is applied to an aqueous solution containing free proteins, convection streams caused by local heating effects at the electrodes or within the solution can occur. These convection fluid currents will disturb the electrophoretic mobility of the protein molecules. This is avoided by mixing the proteins into a pH buffered stabilising medium such as agar or gels composed of silica, agarose or acrylamide, for example. The electrophoretic separation of protein molecules through a gel will then be based on both molecular sieving, related to the pore size of the gel, as well as the electrophoretic mobility of the molecules. The smaller proteins (those of lower molecular weight) will migrate more rapidly through the gel than those of larger molecular weight. Sections of the gel containing the isolated bands of a stained protein mixture can be removed for further analysis. The molecular weight of an unknown protein can be determined by comparing its migration through the gel against the migrations of protein molecular weight standards. This is shown schematically in Figure 3.33.

3.16.1.1 Isoelectric Focusing of Proteins

As described in Chapter 2, protein molecules carry charged amino groups on their surface. Table 2.4 of Chapter 2 lists the pK values for these functional groups. The pK corresponds to the pH at which half of the members of that group are protonated. As the pH changes, the net charge on a protein's surface will change. At high pH, most proteins will have many deprotonated surface groups, and will carry a net negative charge. At low pH, with many protons added to the surface, most proteins have a net positive charge. At some intermediate pH, different for every protein type, the *net charge* on the protein will be zero. The protein is not without charged groups, it carries equal numbers of positive and negative charges. The pH at which a protein has a net charge of zero is designated its isoelectric point (pI).

A protein dissolved in buffer at its pI has no net charge and thus no net electrophoretic mobility. In isoelectric focusing (IEF) a pH gradient is established along the length of a gel,

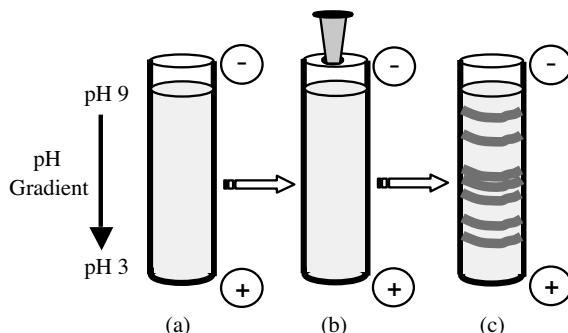


Figure 3.34 Isoelectric focusing is used to determine the pI of a protein, or to separate proteins based on their pI values. (a) An amphoteric fluid is mixed into a gel and a stable pH gradient established by applying an electric field. (b) A protein solution is added and the electric field reapplied. (c) Stained proteins distributed along pH gradient according to their pI values.

as shown in Figure 3.34. Proteins migrate through this gradient until they reach their pI. As shown in Figure 3.34, the gradient is set up so that negatively charged molecules migrate into a decreasing pH region. If a protein is in a region where the pH is above its pI, it has a negative charge and moves to a lower pH. If it is in a pH below its pI, it has a positive charge which moves it into higher pH regions. This gives rise to the self-focusing aspect of IEF, as proteins are continually swept back into tight bands centred on the appropriate pI. IEF is thus an equilibrium electrophoresis system, run until protein movement ceases.

3.16.1.2 Ampholytes are Used to Set Up the pH Gradient

The pH gradient in the isoelectric focusing gel shown in Figure 3.34 is generated by the inclusion of ampholytes, which are low molecular weight amphoteric molecules. Amphoteric molecules can react as either an acid or a base. A mixture of ampholytes is used, each having a different pI. Like protein molecules, the ampholytes migrate through the gel until they reach a region where the pH is equal to their pI. Unlike the proteins, the ampholytes are present in high enough concentration to change their local pH. The gel is set up with a uniform mixture of ampholytes throughout, and its anodic and cathodic ends are immersed in dilute acid and base respectively. Ampholytes near the ends of the gels will be positively charged near the positive electrode, and negatively charged near the negative electrode. They therefore begin to migrate into the gel, with the most charged (i.e. the ones furthest from their pI) moving the fastest. Over time they separate into zones of defined pH. If the ampholyte system is well designed, a smooth gradient of pH is created.

Various mixtures of amphoteric substances have been used as ampholytes, namely amino acids, proteins, and synthetic poly acidic, poly basic, molecules. Proteins can be good ampholytes, but they interfere with analyses of protein samples by introducing new proteins into the mixture. Polycarboxylic acid polyamines are the most commonly used ampholytes. These molecules have excellent buffering capacity across a broad pH range, and are usually provided in a molecular weight range of 300–500 daltons. Their sole disadvantage is that they may bind tightly to the sample proteins, due to ionic interactions, and can be very difficult to remove.

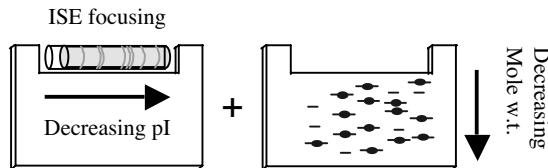


Figure 3.35 Two-dimensional polyacrylamide electrophoresis (2D-PAGE) combines the isoelectric focusing technique shown in Figure 3.34 and the gel electrophoresis method shown in Figure 3.33.

By combining the gel electrophoresis and isoelectric focusing techniques shown in Figures 3.33 and 3.34, two-dimensional gel electrophoresis can be performed. This technique, given the acronym 2D-PAGE, is depicted in Figure 3.35. Thousands of different proteins can be resolved from a single sample using 2-D gel electrophoresis. By combining mass spectrometric surface analysis of the gels, over 2500 proteins that comprise the *E.coli* proteome is capable of being detected from a single *E.coli* sample [18].

3.16.2 Electro-Osmosis

We have seen that electrophoresis is the motion of a charged molecule or particle in a fluid that is induced by an electric field. Electro-osmosis, on the other hand, is a phenomenon where liquid is induced to flow through a narrow channel or capillary by an applied electric field. For this effect to occur, immobilised electric charges must be present on the inner surface of a channel wall in contact with the liquid. This surface charge can arise from the adsorption of charged species in the liquid, or to have ionisable groups as part of the wall structure. A way to ensure that the walls are charged is to fabricate them from glass or fused silica, where deprotonation of Si-OH silanol groups occurs above pH 3 to form negatively charged silanoate ($\text{Si}-\text{O}^-$) groups. This surface charge induces the formation of an electric double layer by attracting ions of opposite charge from the buffer solution, as shown in Figure 3.36.

If an electrical field is applied along the axis of the channel, a volume Coulombic force ($\rho.E$) will be exerted on the buffer solution. However, the net charge density ρ in the buffer is significantly different from zero only in a thin annular region, within the Debye length

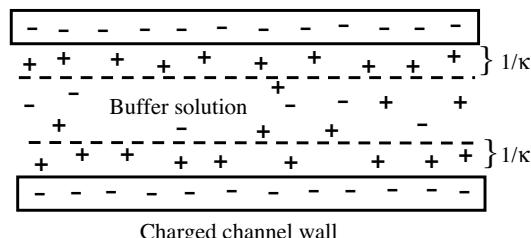


Figure 3.36 The formation of an electrical double layer at the surface of internal charged channel or capillary wall. The net charge density ρ in the buffer solution is significantly different from zero only in a thin annular region, within the Debye length of thickness $1/\kappa$.

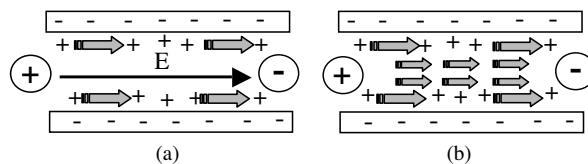


Figure 3.37 (a) On application of an electric field the counterions in the electrical double layers at the channel wall are accelerated and induce local fluid flow. (b) Shear forces accelerate neighbouring fluid lamina, until a steady-state is reached where all fluid lamina move at a uniform velocity given by Equation (3.64).

region close to the channel wall, as shown in Figure 3.36. Therefore, only the counterions in the fluid close to the wall will experience a Coulombic accelerating force and induce fluid movement – in a direction that depends on the field direction and polarity of the counterions. Due to the radial velocity gradient that is formed, the adjacent fluid annuli will be accelerated by the momentum transfer caused by viscous forces until the velocity gradient approaches zero across the whole radius of the capillary. This evolution of the electro-osmosis flow profile is shown in Figure 3.37.

As depicted in Figure 3.37, the charged fluid layer ‘drags’ the adjacent fluid layer along, until finally the entire channel moves at a uniform velocity. The ‘stationary-plate/moving-plate’ scheme outlined in Figure 9.6 of Chapter 9, to describe the viscosity of a Newtonian fluid, has in effect been created. Numerical simulations by Dose and Guiochon [19] demonstrated that this process develops on a timescale between $100\ \mu\text{s}$ and 1 ms. After that time, the whole fluid inside the channel moves at a constant velocity, with the resulting flow profile across the capillary being of a rectangular ‘plug’ shape as shown in Figure 3.38 [20]. This uniform velocity profile occurs if the channel characteristic length (diameter) is at least 7-times that of the electric double layer thickness (Debye length), and if other sources of fluid acceleration such as convection due to Joule heating are absent. This velocity profile is very different from that of pressure-driven flow, which has the parabolic profile shown in Figure 3.38, and is described further in Chapter 9. As a special characteristic of electro-osmotically pumped systems, fluid zones can be transported without significant hydrodynamic dispersion. This is of particular interest in capillary electrophoresis and other elements of microfluidic devices.

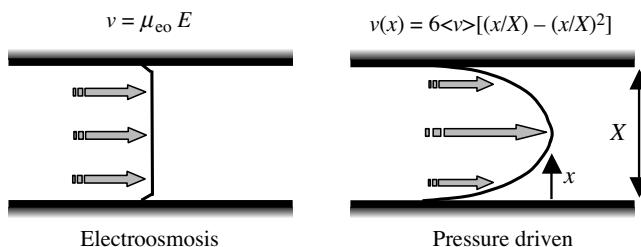


Figure 3.38 The rectangular ‘plug’ flow profile of electro-osmosis differs from the laminar (parabolic) flow profile induced by pressure-driven flow (see Figure 9.15 and Equation (9.22) in Chapter 9).

Under the assumption that the buffer viscosity η and the permittivity $\epsilon_0\epsilon_r$ are the same in the double layer as in free solution, the Smoluchowski theory described in Section 3.16.1 is valid, and the velocity v_{eo} of a fluid undergoing electro-osmosis with an applied electric field E is given by:

$$v_{eo} = \frac{\epsilon_0\epsilon_r\zeta E}{\eta} \quad (3.64)$$

As for electrophoresis of a charged particle, the zeta potential ζ in this equation is the potential difference at the interface between the tightly held counterions close to the charged surface (the channel wall) and the bulk solution. This interface is called the *slip plane* or the *surface of shear*. The electro-osmotic velocity therefore depends on the characteristics (viscosity η and permittivity $\epsilon_0\epsilon_r$) of the fluid undergoing transport, as well as the channel wall material that determines the surface charge and the value of ζ . Fluid properties such as pH, ionic strength or composition, can have an enhanced or opposing effect. For example, in a bare glass capillary, the surface charge increases with pH. If the pH is increased through addition of high concentrations of a metallic salt, the ionic strength will increase and reduce the double layer thickness, which in turn will reduce electro-osmosis.

Microfluidic devices made from glass tend to have a well-characterised surface charge that varies predictably as a function of fluid pH and composition. Under physiological conditions, glass and silica have a negative zeta potential. With reduction of the local pH, the silanol groups on the glass surface become protonated and the zeta potential falls in magnitude. Many different surface modification techniques have been developed for glass, which allow the user to change the surface charge or to alter its biocompatibility (e.g. cell adhesion or nonfouling coatings). However, with an increasing use of polymeric components (silicone, Mylar, Teflon) electro-osmotic behaviour is less predictable. Biological fluids, in particular, can lead to problems such as protein adsorption on polymeric surfaces.

Some of the advantages and disadvantages of employing electro-osmosis to drive fluid flow in microchannels or capillaries can be summarised as follows:

Advantages:

- Uniform flow profile.

This results in uniform retention times for all particles in a given section of a device, which can greatly simplify calculations and analysis. Because fluids move as a bolus, the leading and trailing edges of materials are minimised. This reduces the time and material required to change solutions in a device.

- No moving-part pumps are required.
- A simple fluidic interface.

The interface between the source of pumping (electrodes) can be as simple as two wires placed into holes in the device. Unlike pressure-driven flow, we do not require a leak-tight interface between the source of the hydraulic force and the fluid being driven.

Disadvantages:

- Strong dependence on the electrochemical properties of channel wall and fluid.

If a device is expected to process a variety of fluids or a fluid of unknown pH and ionic strength, the electro-osmotic velocity will be unpredictable.

Table 3.7 Summary of variable parameters that can be adjusted to influence electro-osmotic mobility (μ_{eo}) measurements

Parameter	Effect	Comments
Electric field	μ_{eo} changes proportionately	Joule heating may result if too high. Efficiency and resolution may decrease if lowered too much
Solution pH	μ_{eo} decreased at low pH and increased at high pH	Simple and practical. May change charge or structure of a solute such as a protein
Ionic strength of solution	If increased, the zeta potential and μ_{eo} decrease	High ionic strength can generate high current and Joule heating
Temperature	Changes fluid viscosity (2–3% per °C)	Can be controlled automatically
Surfactant	Adsorbs to capillary wall via hydrophobic and/or ionic interactions	Anionic surfactants can increase μ_{eo} , whilst cationic ones can decrease or reverse μ_{eo}
Organic modifier	Usually decreases μ_{eo} by changing zeta potential and viscosity	Often requires experimentation to determine complex changes. Can significantly alter selectivity
Covalent coating	Bonding of chemicals to capillary wall	Can alter hydrophilicity and surface charge of wall. May not be stable
Neutral hydrophilic polymer	Adsorbs to capillary wall via hydrophobic interactions	Decreases μ_{eo} by shielding surface charge and increasing viscous drag

- Often requires high voltages (typically in the kV to MV range). This requires isolation of the electrodes from the sample fluid to avoid the products of electrolysis (bubbles, acid or base production) from entering the sample fluid, whilst at the same time retaining electrical connectivity.
- Heat produced by the electric field may have to be dissipated.

A summary of the various ways to control electro-osmotic induced fluid flow in microfluidic devices is given in Table 3.7.

3.16.3 Capillary Electrophoresis

Capillary electrophoresis, also known as capillary zone electrophoresis, can be used to separate ionic species by their charge and frictional forces and mass. As discussed in Section 3.16.1, in conventional electrophoresis electrically charged molecules or particles move in a liquid under the influence of an electric field. The technique of capillary electrophoresis was designed in the 1960s to separate species based on their size to charge ratio in the interior of a small capillary filled with an electrolyte. This technique provides an alternative approach to gel electrophoresis to alleviate thermal convection problems in free solution electrophoresis.

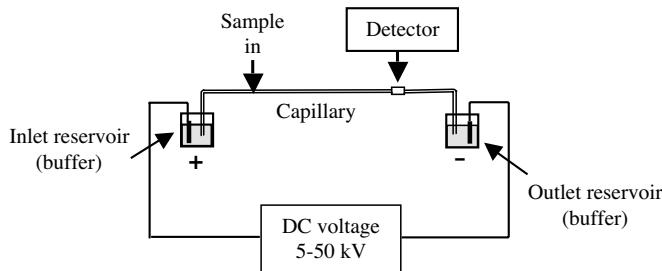


Figure 3.39 The components of capillary electrophoresis equipment consist of a narrow-bore capillary filled with an electrolyte buffer. The migration of an injected sample along the capillary is induced by applying a high-voltage across electrodes at the inlet and outlet reservoirs.

The essentials of capillary electrophoresis are shown in Figure 3.39. A narrow-bore capillary, typically of length $25 \sim 75$ cm, outside diameter $300 \sim 400$ μm , inside diameter $25 \sim 75$ μm , is filled with an electrolyte buffer to maintain a fixed pH of the solution. Due to its narrow bore, heat is dissipated efficiently by the capillary and allows high electric fields to be used. A liquid sample is introduced into the capillary via capillary action, pressure, or siphoning. The migration of the sample along the capillary is then initiated by an electric field, applied between the inlet and outlet reservoirs using electrodes energised by a high-voltage. All molecular components of the sample, whether or not they are positively or negatively charged, will be pulled through the capillary in the same direction (towards the cathode) by electro-osmosis. This is because the electro-osmotic flow of the buffer solution will be greater than the electrophoretic movement of the analytes. Even small, triply charged, anions will be directed to the cathode by the relatively powerful electro-osmosis of the buffer solution. Negatively charged components of the sample are thus retained longer in the capillary due to their conflicting electrophoretic mobilities. Experimentally, the electro-osmotic mobility can be determined by measuring the retention time of an electrically uncharged analyte. The velocity (u) of an analyte in an electric field can then be defined in terms of the sum of its electrophoretic and electro-osmotic mobility:

$$u = u_e + u_{eo} = (\mu_e + \mu_{eo})E.$$

The order of migration of species along the capillary is shown in Figure 3.40. Neutral species travel at the same velocity as the fluid towards the cathode (assuming that the capillary walls are negatively charged). Small multiply charged cations migrate quickly, whereas small multiply charged anions are strongly retarded. Thus, the chemical components of an injected sample separate as they migrate along the capillary, according to their characteristic electrophoretic mobilities. They are detected near the outlet reservoir in the form of an electropherogram, which displays the detected ‘peaks’ as a function of time, with the various ‘peaks’ spaced apart according to the different retention times of the analyte components. An example is given in Figure 3.41 of the separation at pH 4.5 of a set of proteins having pI values of 7.0 or higher (see Table 3.8). The proteins shown in Figure 3.41 each have a pI higher than 4.5 and so will have protonated amino acid side chains (see Table 2.4 of

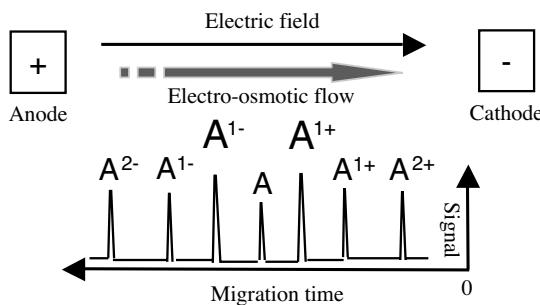


Figure 3.40 A schematic to show the order of migration of charged species of different size and charge, carried by electro-osmotic flow, during capillary electrophoresis. The chemical components of an injected sample will separate as they migrate along the capillary to an extent determined by their characteristic electrophoretic mobilities.

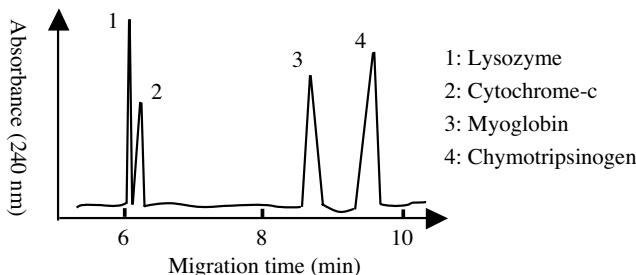


Figure 3.41 The migration times of four basic proteins ($pI \geq 7$) obtained at pH 4.5 by capillary electrophoresis (based on data presented by [20]).

Chapter 2) and carry a net positive charge. From Tables 2.5 and 3.8 we find that cytochrome-c and lysozyme have similar molecular weights and pI values, and so it is not surprising to see that their electro-osmotic migration times through a capillary are quite similar. Chymotrypsinogen, with a molecular weight of 25.7 kDa, is about twice the size of cytochrome-c, and this is reflected in its much longer migration time.

A common form of detection is by optical absorbance at visible and UV wavelengths. A section of the capillary is used as an optical cell, and this ‘on-tube detection’ enables

Table 3.8 Isoelectric points (pI) of some proteins

Protein	pI	Protein	pI
Pepsin	<1.0	Haemoglobin	6.8
Egg albumin	4.6	Myoglobin	7.0
Serum albumin	4.9	Chymotrypsinogen	9.5
Urease	5.0	Cytochrome-c	10.7
β -lactoglobulin	5.2	Lysozyme	11.0

detection of separated analytes with no loss of resolution. In general, capillaries used in capillary electrophoresis are coated with a polymer for increased stability, but the portion of the capillary used for optical detection must be optically transparent. Bare capillaries can break relatively easily and, as a result, capillaries with transparent coatings are available to increase the stability of the cell window. The path length of the detection cell in capillary electrophoresis ($\sim 50 \mu\text{m}$) is far less than that ($\sim 1 \text{ cm}$) used in normal spectrometers. According to the Beer Lambert Law, described in Section 4.4 of Chapter 4, the sensitivity of the detector is proportional to the path length of the cell. To improve the sensitivity the path length can be increased. The capillary tube itself can be expanded at the detection point, creating a ‘bubble cell’ with a longer path length or additional tubing can be added at the detection point. Both of these methods, however, will decrease resolution of the separated analytes.

The resolution R_s of capillary electrophoresis is often given [22] as:

$$R_s = 0.177(\mu_1 - \mu_2) \sqrt{\frac{V}{D_{av}(\mu_{av} + \mu_{eo})}},$$

where μ_1 and μ_2 are the effective mobilities of the more mobile and less mobile analyte, respectively, μ_{av} is their average mobility, D_{av} is their average diffusion coefficient, μ_{eo} is the electro-osmotic mobility, and V is the applied voltage. This expression has been greatly refined in subsequent work by Rawjee and Vigh [23]. Maximum resolution is attained when the electrophoretic and electro-osmotic mobilities are similar in magnitude and opposite in sign. In addition, high resolution requires low diffusion coefficients, and a low electro-osmotic velocity with a correspondingly increased analysis time.

Fluorescence detection can also be used in capillary electrophoresis for samples that naturally fluoresce or are chemically modified to contain fluorescent tags (see Section 4.2.4 in Chapter 4). Ethidium bromide, fluorescein and green fluorescent protein are common fluorophore tags. This mode of detection offers high sensitivity and improved selectivity for these samples, but cannot be utilised for samples that do not fluoresce. Laser-induced fluorescence has been used with a detection limit as low as 10^{-18} to 10^{-21} mol . This sensitivity arises from the high intensity of the incident light and the ability to accurately focus the laser on the capillary. In order to obtain the identity of sample components, capillary electrophoresis can be directly coupled with mass spectrometers or the surface enhanced Raman spectroscopy (SERS) technique described in Chapter 4. In most systems, the capillary outlet is introduced into an ion source that utilises electrospray ionisation. The resulting ions are then analysed by the mass spectrometer. This requires volatile buffer solutions, which will affect the range of separation modes that can be employed and the degree of resolution that can be achieved. For SERS, capillary electrophoresis eluants can be deposited onto a SERS-active substrate. Analyte retention times can be translated into spatial distance by moving the SERS-active substrate at a constant rate during capillary electrophoresis. This allows the subsequent spectroscopic technique to be applied to specific eluants for identification with high sensitivity. As shown by Lin *et al.* [24] SERS-active substrates can be chosen that do not interfere with the spectrum of the analytes.

Capillary electrophoresis can be incorporated into lab-on-chip devices (see Figure 3.42) and the electrical control of fluid flow in networks of channels and capillaries, without the

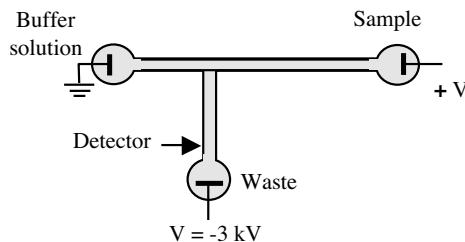


Figure 3.42 Electro-osmosis can be used with capillary electrophoresis in lab-on-chip devices to both control fluid flow, fluidic mixing, and in this simple example to analyse the components of a sample using capillary electrophoresis. The direction and rate of fluid flow in the various arms of this fluidic T-junction can be controlled by the magnitude and polarity of the applied voltages, without the use of valves.

use of valves, is also possible. Early demonstrations of such lab-on-chip techniques were described by Seiler *et al.* [25].

The various advantages of capillary electrophoresis can be summarised as follows:

- small sample sizes ($1\text{--}10 \mu\text{l}$) can be used, and high sensitivity obtained (see Figure 3.43). Femtomole (10^{-15} M) and zeptomole (10^{-21} M) detection levels can be achieved using UV and fluorescence detectors, respectively;
- relatively fast separation times ($1 \sim 45 \text{ min}$) can be achieved;
- easy technique to use with predictable selectivity;
- automation is possible;
- extremely high separation efficiencies are attainable;
- reproducible results can be obtained;
- Can be coupled to a wide range of detectors, such as UV/Visible absorption spectroscopy, mass spectrometry, electrochemical, conductivity, with laser-induced fluorescence being the most sensitive option.

Examples of the various ways in which capillary electrophoresis can be employed for the study and characterisation of biological molecules are given in Table 3.9.

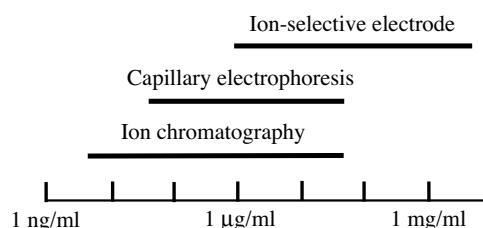


Figure 3.43 A comparison of the sensitivities achieved by different detection techniques.

Table 3.9 Summary of main applications of capillary electrophoresis for the study and characterisation of biological molecules

■ Proteins, peptides	■ Small molecules
– Purity, heterogeneity	– Pharmaceuticals
– Physical properties	Assay
(Mol Wt., pI, peptide mapping)	Purity
– Binding studies	Chiral separation
– Identification	– Forensics
– Quantitation	– Environment
– Immunoassays	– Foods, additives
■ Nucleobases, oligonucleotides	– Ions (organic and inorganic)
– Polymerase chain reaction	
– DNA fingerprinting	
– Clinical applications	

3.16.4 Dielectrophoresis (DEP)

Dielectrophoresis (DEP) is defined as the translational motion of electrically neutral matter in a nonuniform electric field. The field can be either AC or DC. DEP is capable of selectively isolating, concentrating, or purifying target bioparticles when present in complex mixtures. Examples include the isolation of stem cells, cancer cells and bacteria from blood for therapy or further analysis. DEP also lends itself readily to miniaturisation and automation, either as standalone microdevices or as the means for rapid and efficient sample collection and preparation [26].

The DEP collection of particles at electrode edges is shown in Figure 3.44. An effect not possible by magnetophoretic manipulation of magnetic particles, namely the attraction to *as well as the repulsion* from a magnetic pole of the same particles under essentially the same conditions, is possible by DEP and is also shown in Figure 3.44. The ability to attract or repel particles from electrodes is an important aspect of DEP. The translational forces producing

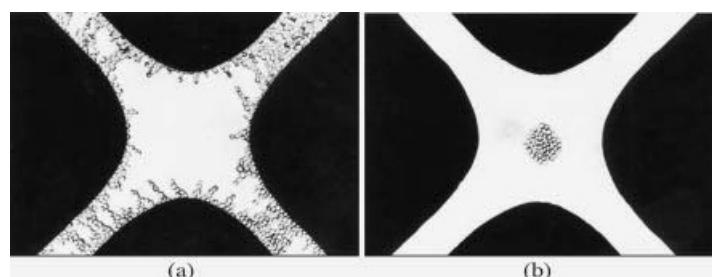


Figure 3.44 (a) Cells collecting at electrode edges under the influence of positive dielectrophoresis. (b) Cells repelled from electrode edges into a field ‘cage’ under the action of negative dielectrophoresis. The only difference in experimental conditions is the frequency of electrical excitation of the electrodes.

these effects arise from the interaction of the particle's dipole moment m with the non-uniform electric field.

3.16.4.1 Force imposed on an Infinitesimal Dipole in an Electric Field Gradient

A useful starting point in formulating the DEP behaviour of particles is to estimate the net force exerted on a small physical dipole when placed in a nonuniform electric field. In Section 3.2.4 a dipole is described as consisting of equal and opposite charges $+\delta q$ and $-\delta q$ located a vector distance d apart. We define the dipole moment as $m = \delta qd$. From Equation (3.41) the electric potential ϕ at a distant point (but not at the dipole origin) from the dipole is given by:

$$\phi = \frac{|m| \cos \theta}{4\pi\epsilon_0\epsilon_r r^2}, \quad (3.65)$$

where θ and r are, respectively, the polar angle and radial position (measured from the centre of the dipole) in spherical coordinates. This result is obtained on condensing the dipole to the limit of a double-point singularity in a manner, where as d tends to zero, q is increased in magnitude so that the dipole moment m remains constant. As in our earlier treatment of electrostatic interactions, the quantity $\epsilon_0\epsilon_r$ is the absolute permittivity of the medium surrounding the dipole. A plot of the equipotentials around a dipole is shown in Figure 3.45. In this figure the potential at any location is given by the sum of the individual potentials generated by the two point charges at that location. This follows from the *Principle of Superposition*, and is shown more clearly in the 3-D plot shown in Figure 3.45b. The electric potential ϕ (Volts) is a scalar quantity. The electric field E generated around the dipole at any point is a vector quantity $E = -\nabla\phi$, where ∇ is the *del* (gradient) vector operator. The total electric field at any point is the *vector* sum of the electric fields due to each charge. The field of a dipole is cylindrically symmetrical about the dipole axis, so that from Equation (3.65) the radial and transverse components of the field intensity in any meridian plane are given by:

$$E_r = -\frac{\partial\phi}{\partial r} = \frac{1}{2\pi\epsilon_0\epsilon_r} \frac{m \cos \theta}{r^3}; \quad E_\theta = -\frac{1}{r} \frac{\partial\phi}{\partial\theta} = \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{m \sin \theta}{r^3}. \quad (3.66)$$

In Figure 3.46 the dipole is located in a nonuniform electric field E , with the negative charge $-q$ having a vector position r . The dipole contribution to the total electric field cannot exert a force on itself, and so is not included in the field E of Figure 3.46. (If this were to be

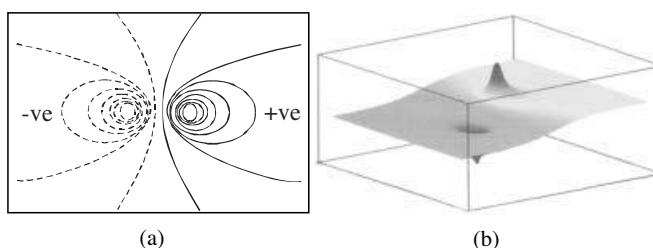


Figure 3.45 (a) A two-dimensional plot of the equipotentials around an electric dipole. (b) A three-dimensional plot of the equipotentials shown in (a).

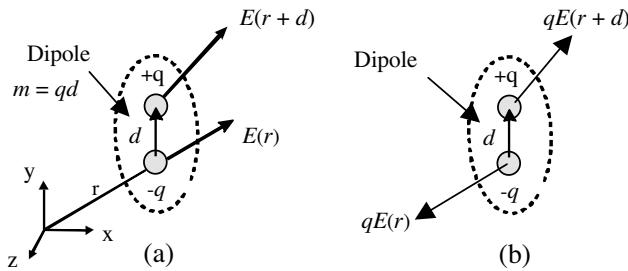


Figure 3.46 (a) A dipole located in an electric field gradient, with the negative charge $-q$ having a vector position r . (b) Because the two dipole charges each experience different values of the electric field, they experience different Coulombic forces. This gives rise to a dielectrophoretic force acting on the dipole.

the case, it would be rather like the dipole ‘pulling itself up by its own bootlaces’!) The total Coulombic force \mathbf{F} acting on the dipole is given by:

$$\mathbf{F} = q\mathbf{E}(r+d) - q\mathbf{E}(r). \quad (3.67)$$

We can simplify this equation using a vector Taylor series expansion,¹ and expand the electric field about position r as:

$$\mathbf{E}(r+d) = \mathbf{E}(r) + d \cdot \nabla \mathbf{E}(r) + \frac{d^2}{2!} \cdot \nabla^2 \mathbf{E}(r) + \text{higher order terms} \quad (3.68)$$

so that $\nabla \mathbf{E}$ (pronounced grad E) is the field gradient.² If the dipole particle size is small compared to the length scale of the field nonuniformity, we can ignore the terms containing d^2 and higher in Equation (3.68). Substituting Equation (3.67) into (3.68) and condensing the dipole (i.e. taking the limit $d \rightarrow 0$) in such a way that the dipole moment m (given by $m = qd$) remains finite, we obtain the force on an infinitesimal dipole as:

$$\mathbf{F} = q(\mathbf{E} + d \cdot \nabla \mathbf{E}) - q\mathbf{E} = qd \cdot \nabla \mathbf{E} = m \cdot \nabla \mathbf{E}. \quad (3.69)$$

This force is known as the *Dielectrophoresis* (DEP) force and can be interpreted as the energy which must be expended to withdraw the dipole particle from the local field E into a region where there is no field.

3.16.4.2 Dielectrophoretic Forces on Particles

We can make the conceptual transition from dealing with hypothetical point dipoles to real particles using the fact that materials are polarised when subjected to an electrical field. This polarisation produces surface charges on the material, creating in effect a macroscopic

¹ Taylor’s series: $f(x+h) = f(x) + h f'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \dots$ (where $f', f'', f''' \dots$ are the 1st, 2nd, 3rd differentials, and so on).

² $\nabla = i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z}$ and $E = -\nabla V$ (written as $E = -\nabla V$) where V is the electric potential. E thus belongs to a special class of vector fields that can be expressed as the gradient of a scalar field (or scalar point function) and is said to be an *irrotational* field.

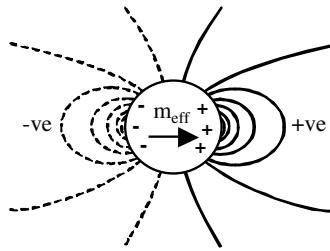


Figure 3.47 The induced surface charges on a spherical particle are assumed to form an effective dipole moment m_{eff} that generates a similar potential distribution as the point-charge dipole shown in Figure 3.45.

dipole. The induced surface charge density $\Delta\rho$ gives a measure of the *polarisability* of the dielectric material. We can relate this to a molecular property of the dielectric by defining a polarisability parameter p as the *average induced dipole moment per unit volume*.

This relationship between induced surface charge on a polarised particle and an effective dipole moment is shown in Figure 3.47. Our task is to formulate an expression for this effective dipole moment and to make sure it generates the same potential contours around itself as the point dipole moment shown in Figure 3.45.

The effective dipole moment m_{eff} of a particle of radius R , suspended in a medium of permittivity ϵ_m and polarised in a field E , to replace the dipole moment m in Equation (3.69) is:

$$m_{\text{eff}} = pvE = 4\pi\epsilon_0\epsilon_m R^3 f(\epsilon)E, \quad (3.70)$$

where $f(\epsilon)$ is the effective polarisability (per unit volume) of the particle – generally referred to as the Clausius-Mossotti factor (named after a German and Italian scientist who made early contributions in the 1800's to an understanding of dielectric properties). From Equations 3.69 and (3.70):

$$F_{\text{DEP}} = 4\pi\epsilon_0\epsilon_m R^3 f(\epsilon)(E \cdot \nabla)E = 2\pi\epsilon_0\epsilon_m R^3 f(\epsilon)\nabla E^2. \quad (3.71)$$

Equation 3.71 teaches several important facts, namely that the DEP force F_{DEP} is zero if the field is uniform (i.e. $\nabla E = 0$), and that the force depends on:

- the *square* of the applied electric field magnitude;
- the *polarisability* p of the particle;
- the *effective volume* v of the particle;
- the *geometry* of the electrodes producing the nonuniform field.

The product $(E \cdot \nabla)E$ in Equation (3.71) has dimensions of V^2/m^3 . The voltage-squared dependence indicates that the direction of the DEP force is insensitive to the polarity of the applied field – either DC or AC voltages can be used to energise the electrodes. The same DEP force can also be produced using a smaller applied voltage if the electrode dimensions are scaled down accordingly. For example, with all other factors remaining fixed, a one hundred-times smaller voltage can produce the same DEP force if the electrode dimensions are

scaled down 1000-fold. To produce a significant DEP force on a biological cell, for example, requires a value for $(E \cdot \nabla)E$ of at least $10^{12} \text{ V}^2/\text{m}^3$. With suitably scaled electrodes this can be achieved using an applied voltages of the order 1 V.

Biological particles such as bacteria and cells exhibit a conductivity associated with mobile ions in their structures, and the medium in which they are suspended is usually a conducting electrolyte. When A.C. fields are applied, these conduction losses can be described in the form of either a complex permittivity ϵ_p^* :

$$\epsilon_p^* = \epsilon_0 \epsilon_p - \frac{j\sigma_p}{\omega}$$

or a complex conductivity σ_p^* :

$$\sigma_p^* = \sigma_p + j\omega \epsilon_0 \epsilon_p,$$

where j is the imaginary vector ($j = \sqrt{-1}$) and ω is the angular frequency ($\omega = 2\pi f$) of the applied A.C. field. The Clausius-Mossotti factor $f(\epsilon)$ in Equation (3.70) can be expressed as a complex function using either of the two equivalent forms:

$$f(\epsilon) = \left(\frac{\epsilon_p^* - \epsilon_m^*}{\epsilon_p^* + 2\epsilon_m^*} \right) \quad \text{or} \quad f(\epsilon) = \left(\frac{\sigma_p^* - \sigma_m^*}{\sigma_p^* + 2\sigma_m^*} \right). \quad (3.72)$$

The total current in the particle can be considered to comprise two elements – one associated with field-induced movement of free charges such as ions, and the other arising from the field-induced perturbation of bound charges known as a displacement current. At low frequencies, as $\omega \rightarrow 0$, the current is dominated by the conduction of free charges and there is essentially no phase difference between the field and the current. At high frequencies, $\omega \rightarrow \infty$, the dielectric displacement current dominates and the particle acts like a capacitor with the current leading the applied field by a phase angle close to $\pi/2$ radians. Depending on the relative values of the permittivity ϵ_p of the particle and that of the surrounding medium ϵ_m , $f(\epsilon)$ will have a value bounded by the limits $-0.5 \leq f(\epsilon) \leq 1.0$. A positive value for $f(\epsilon)$ corresponds to the effective moment m_{eff} being colinear with the applied field E , whilst m_{eff} acts against E when $f(\epsilon)$ is negative, which in turn corresponds to either a positive or negative DEP force, respectively. Examples of these two types of DEP response are shown in Figure 3.44. From Equation (3.72) it is clear that $f(\epsilon)$ is a complex variable (having real and imaginary components). The translational DEP force results from the electric field interacting with the in-phase component of the induced dipole moment, and so the real component of $f(\epsilon^*)$ should appear in Equation (3.71) and is this is commonly signified as $\text{Re}[f(\epsilon^*)]$.

3.16.4.3 Dielectrophoresis of Cells

Bioparticles, such as cells and bacteria, have complicated structures and certainly cannot be modelled as homogeneous spheres. A simple method to describe many bioparticles is by means of the so-called *multishell* model. A simple example of this is shown in Figure 3.48a in the form of the single-shell model used to describe a red blood cell (or a liposome) of radius R . The membrane (thickness d) is assigned a specific conductance g_m and specific capacitance C_m , with $g_m = \sigma_m/d$, and $C_m = \epsilon_m/d$. C_m values for cells are large (reflecting the

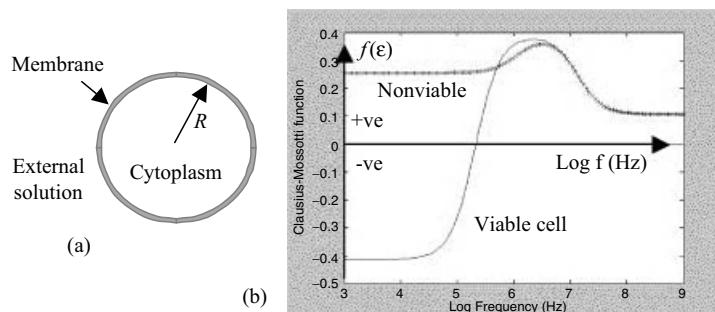


Figure 3.48 (a) The single-shell model of a cell can be used cells (e.g. a mammalian red blood cell) that has no nucleus, or a liposome, of radius R . (b) Frequency variations of the Clausius-Mossotti functions $f(\epsilon)$ for a viable and nonviable red blood cell.

ultra-thin nature of the membrane) – typically of the order $10 \sim 20 \text{ mF/m}^2$. For viable cells the membrane acts as a barrier to passive ion flow across it, and so g_m has a low value – typically $\sim 1 \text{ mS/m}^2$ or lower.

Red blood cells (erythrocytes) are normally discoid in shape, but when suspended in an electrolyte they can take the form of a spheroid ($\sim 7 \mu\text{m}$ in diameter). Human erythrocytes, unlike white blood cells (leukocytes), do not possess a nucleus and can be represented as a thin membrane (the single shell) surrounding the cytoplasm. A 3-shell model is required to represent a cell with a nucleus, comprising four separate compartments: a 1st shell to represent the cytoplasmic membrane; a 2nd shell to represent the cytoplasm; and a 3rd shell to represent the membrane surrounding the nucleus. The Clausius-Mossotti function for a multishell particle is obtained by evaluating *effective* values for the relative complex permittivity ϵ_p^* or conductivity σ_p^* of the particle. The term *effective* is used to signify that a heterogeneous (multishell) particle may be replaced conceptually with one having homogeneous *smeared-out* bulk properties, such that substitution of this homogeneous particle with the original heterogeneous one would not alter the electric field in the surrounding medium. A numerical method for achieving this has been described by Huang *et al.* [27]. The corresponding Clausius-Mossotti function is then obtained by substituting this effective value for ϵ_p^* into Equations (3.71) and (3.72). The frequency variation of this function is shown in Figure 3.48b for a model of a red blood cell with an intact membrane, and for one whose membrane has been damaged and no longer acts as a high resistance to passive ionic conduction. For frequencies below $\sim 100 \text{ kHz}$ nonviable, damaged, cells experience positive DEP and the viable ones negative DEP. This effect can be used to separate viable from nonviable cells that are suspended in a fluid that flows through a DEP chamber.

In DEP studies of mammalian cells (e.g. blood cells, cancer cells, stem cells) the suspending medium commonly takes the form of a low conductivity ($10 \sim 180 \text{ mS/m}$) electrolyte containing sufficient concentrations of sugars (e.g. mannitol, sucrose, dextrose) to raise the osmolarity to the normal physiological level of around 280 mOs/kg . For viable cells the plasma membrane acts as an electrical insulator to passive ion conduction, and thus at low frequencies the cell will appear as an insulating object suspended in a conducting medium. This corresponds to a negative polarisability factor $f(\epsilon)$ so that viable cells will exhibit negative DEP at low frequencies, as shown for the example of a viable red blood cells in

Figure 3.48b. With increasing frequency the electrical field begins to penetrate into the conductive cytoplasm. Electronic engineers will recognise this as capacitive coupling between the suspending medium and cytoplasm, where the effective capacitance of the plasma membrane shorts out the membrane resistance. The effective conductivity of the cytoplasm will be less than that ($\sim 1.4 \text{ S/m}$) of a pure physiological strength electrolyte because of the presence of insulating bodies and structures (e.g. protein cytoskeleton, lipid membranes). Values for σ_p in the range $0.1 \sim 0.5 \text{ S/m}$ are commonly deduced for viable cells above 1 MHz, so that depending on the choice of suspending medium conductivity we can achieve the condition $\sigma_p > \sigma_m$. From this we deduce that, with increasing frequency, a transition from negative to positive DEP may occur. The frequency value, commonly referred to as the DEP crossover frequency f_{xo} , of such a transition occurs when $\text{Re}[f(\epsilon^*)]$ is zero. For the usual experimental conditions used in DEP experiments on cells, this crossover frequency occurs at a frequency of the order 100 kHz. Theoretical modelling of the Clausius-Mossotti factor $f(\epsilon)$ of the type shown in Figure 3.48b predicts that a second crossover of the DEP response (from positive to negative DEP) can occur at a frequency above 100 MHz for some cells when suspended in suitable media. This effect corresponds to where displacement currents begin to dominate over ionic conduction effects as the frequency increases, and where the effective permittivity of the cell interior is less than that of the surrounding medium. The first systematic experiments to measure this so-called 2nd DEP crossover frequency were those of Chung *et al.* [28] for myeloma cells.

A cell's plasmic membrane acts as a capacitor because it is constructed like one – namely a thin dielectric situated between two conductors (the outer and inner electrolytes). For a cell of radius R suspended in an electrolyte of conductivity σ_m the membrane capacitance C_{cm} can be determined from a measurement of the first (lower frequency) DEP crossover frequency f_{xo1} , using the following relationship:

$$C_{cm} = \frac{\sqrt{2}}{2\pi R f_{xo1}} \sigma_m.$$

This equation assumes that the high resistance value of the cell membrane has not been impaired due to damage or the onset of cell death, for example. For a fixed cell radius, the effective membrane capacitance of a smooth cell will be less than that for a cell having a complex cell surface topography associated with the presence of microvilli, blebs, membrane folds or ruffles, for example. This will influence the value observed for f_{xo1} , which has important implications for applying DEP to characterise and selectively isolate target cells from other cells [26].

3.16.5 Electrowetting on Dielectric (EWOD)

In Chapter 9, surface tension is described as an inherently dominant force in the microscale. This force can be modified and controlled electrically. An early demonstration of this of relevance to lab-on-chip technologies was the demonstration by Pollack *et al.* [29] of rapid manipulation of discrete microdroplets along a linear array of electrodes. The direct manipulation of discrete droplets offers the means to integrate microfluidic systems without the need for conventional pumps, valves or channels. Such systems can be flexible, power efficient and capable of performing complex and highly parallel microfluidic processing tasks. To achieve this effect a voltage is applied, as shown in Figure 3.49, between a conducting liquid

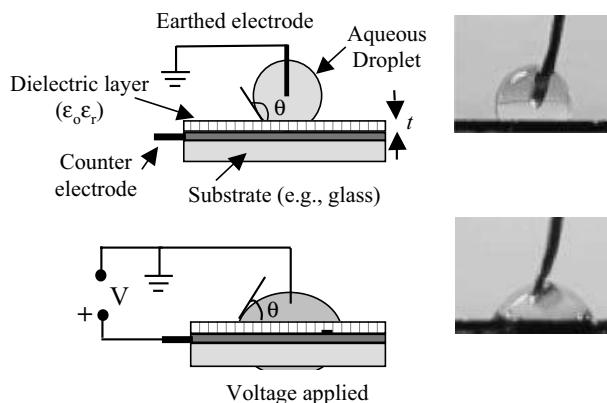


Figure 3.49 The EWOD effect is achieved by applying a voltage between a conducting liquid droplet at rest on a thin dielectric layer of poor wettability, situated above a counter electrode. The resulting charge that accumulates at the solid–liquid interface results in the contact angle θ falling below 90° , which is equivalent to a transition from a hydrophobic to a hydrophilic state if the droplet is aqueous.

droplet (e.g. an electrolyte) at rest on a dielectric layer and a counter electrode positioned below this dielectric. For maximum effect the dielectric surface should be of poor wettability – for example, a hydrophobic surface if the droplet is aqueous. The resulting charge that accumulates at the solid–liquid interface leads to a change in contact angle from $>90^\circ$ to $<90^\circ$, as shown in Figure 3.49. This is equivalent to a transition from a nonwetting to a wetting state. This effect is known as electrowetting on dielectrics (EWOD).

The liquid drop shown in Figure 3.49 is in contact with a solid insulator film of thickness t and permittivity $\epsilon_0\epsilon_r$. A voltage V is applied between the conducting liquid and a counter electrode situated beneath the dielectric film. Before the voltage is applied we assume that the solid–liquid interface is not electrically charged. When a voltage is applied the conducting liquid drop and the counter electrode form a capacitor C of value proportional to the area A_{S-L} formed by the solid–liquid interface at the base of the droplet. The surface capacitance C (per unit area) is:

$$C = \frac{\epsilon_0\epsilon_r}{t}$$

The wetted dielectric surface will attain a charge of magnitude $Q = VC$, and this will decrease the surface potential energy of the water molecules at the solid–liquid surface. The water molecules will gain cohesive energy arising from their dipole attraction to these surface charges, and so reduce their potential energy. To a first order approximation the electrostatic energy stored in the capacitor ($\frac{1}{2}CV^2$) can be incorporated into an expression for the voltage-dependent solid–liquid interfacial energy $\gamma_{S-L}(V)$ to give:

$$\gamma_{S-L}(V) = \gamma_{S-L}(0) - \frac{\epsilon_0\epsilon_r}{2t} V^2, \quad (3.73)$$

where $\gamma_{S-L}(0)$ is the interfacial energy with no voltage applied. The V^2 dependence indicates that either a direct current voltage, of positive or negative polarity, or an alternating current

voltage can be applied across the dielectric layer. For an ac rather than a dc voltage, V^2 is replaced by V_{peak}^2 in Equation (3.73).

The contact angle θ will be modified according to Young's equation (see Chapter 9):

$$\gamma_{S-A} = \gamma_{S-L}(0) + \gamma_{L-A} \cos \theta(0) = \gamma_{S-L}(V) + \gamma_{L-A} \cos \theta(V), \quad (3.74)$$

where γ_{S-A} and γ_{L-A} are the solid-air and liquid-air surface energies, respectively.

From Equations (3.73) and (3.74) we obtain:

$$\cos \theta(V) = \cos \theta(0) + \frac{\varepsilon_0 - \varepsilon_r}{2t\varepsilon_r} V^2. \quad (3.75)$$

From Equation (3.75) we can estimate that a voltage of ~ 160 V is required to lower θ from 110° to 70° for an aqueous droplet ($\gamma_{L-A} \sim 70 \times 10^{-3}$ N m $^{-1}$) if PTFE (Teflon) or PET ($\varepsilon_r \sim 2$) of thickness ~ 5 μm is used as the dielectric. Equation 3.75 also indicates that the electro-wetting effect is enhanced if the dielectric thickness t is reduced and ε_r is increased. The field across the dielectric is V/t and this should be maintained at a value well below that where charge injection and dielectric breakdown occurs. In the example considered above of 160 V applied across a dielectric thickness of 5 μm , the corresponding field is 32 MV m $^{-1}$, which is below the dielectric strength value of 60 MV m $^{-1}$ for PTFE, for example.

Some polymers can be vapour-deposited as thin hydrophobic dielectric films – a common example being various forms of poly(p-xylene) known as parylene that have values for ε_r of around 3 and a dielectric strength ~ 7 MV m $^{-1}$. Other dielectrics being investigated for EWOD applications include the high- κ (high dielectric constant) oxides that have replaced silicon dioxide as the gate material in the latest CMOS devices. Silicon oxynitride, for example, has a value for ε_r of around 8, a dielectric strength greater than 1000 MV m $^{-1}$, and can be formed as a submicron layer on conducting silicon.

Through the suitable physical arrangement and electrical switching of electrodes, EWOD can be used to control the motion and delivery of fluid droplets in microfluidic devices. A droplet situated mainly on an electrode element, but also overlapping an adjacent electrode area, can be induced to relocate onto this neighbouring electrode by bringing the first electrode to earth potential and applying a voltage of sufficient magnitude to the neighbouring electrode (so as to change the leading dielectric-liquid contact angle from above to below 90°). A droplet can also be split into two separate portions by applying voltages to both of its adjacent electrodes. Refinements of the dielectric layers in EWOD devices, in terms of their dielectric strength, high permittivity, and surface wetting characteristics, is an active research area (e.g. [30]). EWOD can also be integrated with other technologies such as dielectrophoresis and optoelectronic tweezers to provide the means for single cell sample preparation and analysis on lab-on-chip devices (e.g. [31]).

References

- [1] Aveyard, R. and Haydon, D.A. (1973) *An Introduction to the Principles of Surface Chemistry*, Cambridge University Press.
- [2] Bedzyk, M.J., Bommarito, G.M., Caffrey, M. and Penner, T.L. (1990) Diffuse-double layer at a membrane-aqueous interface measured with X-ray standing waves. *Science*, **248** (4951), 52–56.
- [3] Lennard-Jones, J.E. (1924) On the determination of molecular fields. *Proceedings Royal Society London, A106*, 463–477.

- [4] Lide, D.R. (ed.) (2001–2002) *CRC Handbook of Chemistry & Physics*, 82nd edn.
- [5] Parsegian, A. (1969) Energy of an ion crossing a low dielectric membrane: solutions to four relevant electrostatic problems. *Nature*, **221**, 844–846.
- [6] Pethig, R. and Kell, D.B. (1987) The passive electrical properties of biological systems: their significance in physiology, biophysics and biotechnology. *Physics in Medicine & Biology*, **32**, 933–970.
- [7] Rand, R.P. (1981) Interacting phospholipids bilayers: measured forces and induced structural changes. *Annual Reviews of Biophysics and Bioengineering*, **10**, 288–314.
- [8] Goldman, D.E. (1943) Potential, impedance, and rectification in membranes. *The Journal of General Physiology*, **27**, 37–60.
- [9] Hodgkin, A.L. and Horowicz, P. (1959) The influence of potassium and chloride ions on the membrane potential of a single muscle fibre. *Journal of Physiology*, **148**, 127.
- [10] Hodgkin, A.L. and Huxley, A.F. (1952) Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo. *Journal of Physiology*, **116**, 449–472.
- [11] Hille, B. (1992) *Ionic Channels of Excitable Membranes*, 2nd edn, Sinauer Associates, Inc., Sunderland, Massachusetts.
- [12] Sukharev, S., Durell, S.R. and Guy, H.R. (2001) Structural models of the MscL gating mechanism. *Biophysical Journal*, **81**, 917–936.
- [13] Imoto, K., Busch, C., Sakmann, B. *et al.* (1988) Rings of negatively charged amino acids determine the acetylcholine receptor channel conductance. *Nature*, **335**, 645–649.
- [14] Hodgkin, A.L., Huxley, A.F. and Katz, B. (1949) Ionic currents underlying activity in the giant axon of the squid. *Archives Des Sciences Physiologiques*, **3**, 129.
- [15] Neher, E. and Sakmann, B. (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, **260**, 799–802.
- [16] Neher, E. and Sakmann, B. (1992) The patch clamp technique. *Scientific American*, **266**, 44–51.
- [17] Ashcroft, F.M. (2006) From molecule to malady. *Nature*, **440**, 440–447.
- [18] Loo, R.R.O., Cavalconi, J.D. Van Bogelen, R.A. *et al.* (2001) Virtual 2-D electrophoresis: Visualization and analysis of the *E.coli* proteome by mass spectroscopy. *Analytical Chemistry*, **73**, 4063–4070.
- [19] Dose, E.V. and Guiochon, G. (1993) Timescales of transient processes in capillary electrophoresis. *Journal of Chromatography A*, **652** (1), 263–275.
- [20] Rice, C.L. and Whitehead, R. (1993) Electrokinetic flow in a narrow cylindrical capillary. *Journal of Physical Chemistry*, **69** (11), 4017–4024.
- [21] Schmalzing, D., Piggee, C.A. Foret, F. *et al.* (1993) Characterization and performance of a neutral coating for the capillary elecrophoretic separation of biopolymers. *Journal of Chromatography A*, **652** (1), 149–159.
- [22] Giddings, J.C. (1969) Generation of variance, theoretical plates, resolution and peak capacity in electrophoresis and sedimentation. *Separation Science*, **4** (3), 181–189.
- [23] Rawjee, Y.Y. and Vigh, G. (1994) A peak resolution method for the capillary electrophoretic separation of the enantiomers of weak acids. *Analytical Chemistry*, **66** (5), 619–627.
- [24] Lin, H., Natan, M.J. and Keating, C.D. (2000) Surface enhanced Raman scattering: A structure-specific detection method for capillary electrophoresis. *Analytical Chemistry*, **72** (21), 5348–5355.
- [25] Seiler, K., Fan, H.Z., Fluri, K. and Harrison, D.J. (1994) Electrosomotic pumping and valveless control of fluid flow within a manifold of capillaries on a glass chip. *Analytical Chemistry*, **66**, 3485–3491.
- [26] Pethig, R. (2010) Dielectrophoresis: Status of the theory, technology and applications. *Biomicrofluidics*, **4**, 028811.
- [27] Huang, Y., Hölzle, R., Pethig, R. and Wang, X.B. (1992) Differences in the AC electrodynamics of viable and non-viable yeast cells determined through combined dielectrophoresis and electrorotation studies. *Physics in Medicine & Biology*, **37**, 1499–1517.
- [28] Chung, C., Waterfall, M. Pells, S. *et al.* (2011) Dielectrophoretic characterisation of mammalian cells above 100 MHz. *Journal of Electrical Bioimpedance*, **2**, 64–71.
- [29] Pollack, M.G., Fair, R.B. and Shenderov, A.D. (2000) Electrowetting-based actuation of liquid droplets for microfluidic applications. *Applied Physics Letters*, **77** (1), 1725–1726.
- [30] Cahill, B.P., Giannitsis, A.T. Land, R. *et al.* (2010) Reversible electrowetting on silanized nitride. *Sensors & Actuators*, **B144**, 380–386.
- [31] Shah, G.J., Ohta, A.T. Chiou, E.P.Y. *et al.* (2009) EWOD-driven droplet microfluidic device integrated with optoelectronic tweezers as an automated platform for cellular isolation and analysis. *Lab on Chip*, **9**, 1732–1739.

4

Spectroscopic Techniques

4.1 Chapter Overview

Spectroscopic techniques are widely employed in biosensors and biomedical tests. The majority of these are photometric, simple examples of which include monitoring changes of the colour of a pH indicator, or of an analyte as it reacts with an immobilised reagent. More complicated examples include monitoring the intensity of reflected radiation in a surface plasmon resonance device, or ratio metric measurement of Förster energy resonance transfer between two chromophores. The electromagnetic waves used in photometric sensors range from frequencies as high as 10^{15} Hz down to 10^9 Hz. Most optical-based sensing techniques offer the significant advantage over electronic ones in that the transmission of photometric information through optical cables or space is not susceptible to electrical interference. A form of spectroscopy increasingly employed in electrochemical-based sensors is known as impedance spectroscopy. In this technique the effective charge-transfer resistance at the sensing electrode of a device can be monitored and quantified by measuring its electrical impedance for frequencies at the opposite end of the electromagnetic scale (typically 100 kHz to sub-hertz frequencies). In this chapter the theoretical and practical fundamentals of the various forms of spectroscopy will be described and serve as the background for the photometric biosensors described in Chapters 6 and 7.

After reading this chapter readers will gain a basic understanding of:

- (i) photometry based on the absorption, emission, and scattering of light;
- (ii) electronic, vibrational, rotational and Raman spectroscopy;
- (iii) nuclear magnetic resonance (NMR) and electron spin resonance (ESR);
- (iv) techniques that employ: Total internal reflection fluorescence (TIRF); surface plasmon resonance (SPR); Förster resonance energy transfer (FRET);
- (v) the concepts of molar absorption coefficient, absorbance and transmittance;
- (vi) the Beer-Lambert law and its limitations;
- (vii) Impedance Spectroscopy.

4.2 Introduction

Photometric spectroscopy is the analysis of the electromagnetic (EM) radiation absorbed, emitted or scattered by atoms or molecules when they undergo transitions between discrete energy states. EM radiation can be considered to consist of individual photons of energy E given by:

$$E = h\nu = hc/\lambda,$$

where h is Planck's constant ($h = 6.62 \times 10^{-34}$ J s) and c is the velocity of light in a vacuum ($c = 3 \times 10^8$ m/s). The radiation takes the form of orthogonally associated electric and magnetic waves, which as shown in Figure 4.1 extend from low energy radio waves up to and beyond high energy X-rays. Propagated EM waves have a frequency of oscillation ν and wavelength λ related through the equation $c = \nu\lambda$. As detailed in Table 4.1, the region of the EM spectrum visible to the human eye is restricted to a narrow range of energies, corresponding to wavelengths from ~ 400 nm (violet) to ~ 700 nm (red). Spectroscopists often present their data as plots of absorption, or emission, or light scattering intensity as a function of either the radiation wavelength or the wavenumber (given by the reciprocal of wavelength, often expressed in units of cm^{-1}).

4.2.1 Electronic and Molecular Energy Transitions

In Chapter 1 the chemical bonds in a molecule are described as the sharing of electrons between the nuclei of adjacent atoms. Fluctuating motions of electrons in molecular orbitals can give rise to attractive dipole-dipole (van der Waals) interactions between adjacent

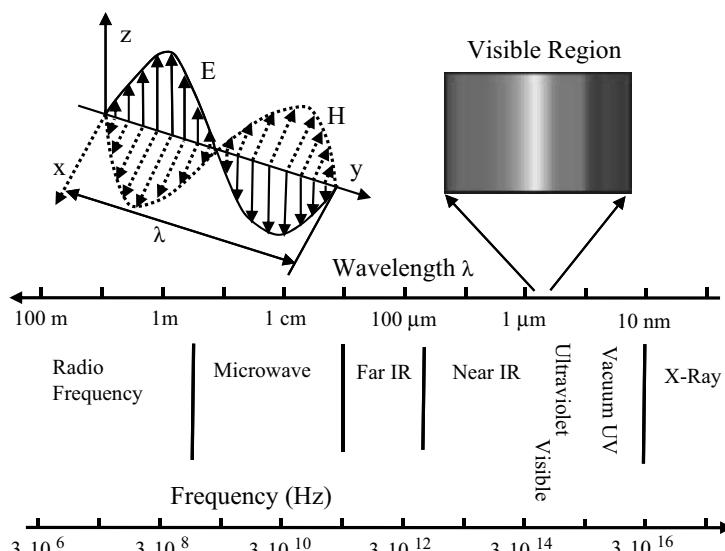


Figure 4.1 Electromagnetic (EM) radiation consists of orthogonal electric (E) and magnetic (H) fields that can propagate as sine waves over a wide range of frequencies. The visible region ($\lambda = 400\text{--}700$ nm) occupies a narrow part of this range.

Table 4.1 The colours, wavelengths and frequencies of light in the visible region of the EM spectrum

Colour	Wavelength (λ) (nm)	Frequency (ν) Hz/ 10^{14}
Ultraviolet	<300	>10
Violet	390	7.7
Blue	450	6.7
Green	520	5.8
Yellow	580	5.2
Red	700	4.3
Infrared	>1000	<3.0

molecules, whereas the positive charges carried by the nuclei give rise to repulsive forces. The combination of these attractive and repulsive Coulombic forces leads to an equilibrium distance between nuclei and creates a stable molecular structure (see Figure 1.2 in Chapter 1).

According to quantum mechanical theory molecules possess discrete states, or quanta, of energy. These states are called the energy levels of the atom or molecule. Changes in energy can arise from the redistribution of electrons within molecular orbitals and also, as depicted in Figure 4.2, as a result of perturbations (vibrations) of their chemical bond lengths or changes of molecular rotation. Each of these energy transitions occur with characteristic time scales across a wide range of energies. The energies involved in changing electron distributions are of the order of several electron volts with time scales in the range 10^{-14} to 10^{-17} seconds. Vibrational energy levels are separated by around 0.04 to 0.4 eV (timescales of $\sim 10^{-13}$ s to 10^{-14} s) and rotational transitions typically involve energies at most one-tenth of this, with timescales from around 10^{-12} s to 10^{-10} s. Electronic redistributions between orbitals, and changes of the vibrational and rotational motions of molecules, can be induced by light (electromagnetic radiation) and studied by spectroscopy. Examples of such transitions are shown in Figure 4.3. In Chapter 3 we learnt that the factor kT corresponds to the mean thermodynamic energy available to a particle when it is in equilibrium with its environment at temperature T (k is the Boltzmann constant). At room temperature kT has a value of 0.025 eV. From this we can deduce that electronic transitions can only occur from the ground electronic state, whereas many rotational states are already ‘occupied’ at room

- Electronic transitions in atoms and molecules
 $10^{14} \sim 10^{17}$ Hz (Visible, UV, X-ray)
 - Bonded nuclei vibrate with respect to each other
 $10^{13} \sim 10^{14}$ Hz (Infrared)
 - Molecules rotate
 $10^{10} \sim 10^{12}$ Hz (Microwaves)
-

Figure 4.2 Electronic transitions in atoms and molecules, and bonded nuclei exhibit vibrational and rotational motions across a wide range of time scales.

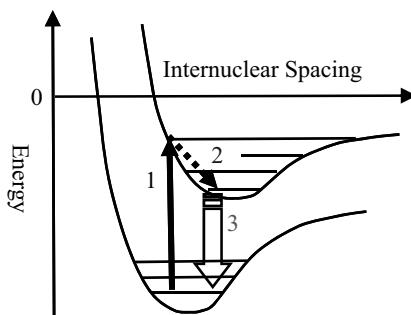


Figure 4.3 Electronic transitions between molecular energy levels can include: (1) absorption of light leading to the excitation of an electron to a higher energy level; (2) relaxation to a lower energy state as a result of energy lost to molecular vibrations; (3) radiative decay (fluorescence) back down to the ground state.

temperature and so rotational transitions can occur from a wide range of energy levels, not only from the ground state. Analysis of the electromagnetic radiation absorbed, emitted or scattered by atoms or molecules as they undergo transitions between two discrete energy states can provide information about the electronic structure of atoms, as well as the bond lengths, bond angles and bond strengths of molecules.

4.2.2 Luminescence

This is a general term for the emission of radiation from a cool object (*incandescence* is the emission of light from hot metals, as for example from the filament of a light bulb). The two main categories of luminescence that can be excited by optical irradiation are fluorescence and phosphorescence, whilst luminescence arising from a chemical reaction is known as chemiluminescence.

4.2.3 Chemiluminescence

This is the emission of light driven by a chemical reaction, of which the most common natural form is *bioluminescence* employed by some aquatic animals to hunt for food (e.g. angler fish), by some insects (e.g. glow worms, fireflies) to attract mates, and by some marine microorganisms to signal distress.

4.2.4 Fluorescence and Phosphorescence

These are the two modes of radiative decay of electronically excited states. Electronically excited states have a finite lifetime, and in most cases the energy of excitation is dissipated into random thermal motions of surrounding atoms with no luminescence observed. However, the excited state may also lose energy through radiative decay accompanied by the emission of a photon as the electron falls back to its lowest orbital energy state (see Figure 4.3). If photon emission occurs directly from the excited state following absorption of the excitation radiation it is known as *fluorescence*. Fluorescence is therefore a very fast response and disappears rapidly after the radiation being absorbed is switched off. If there is

a time delay following absorption of the excitation radiation, the photon emission is known as *phosphorescence*. The delay associated with phosphorescence is a consequence of an electron in its excited state reversing its spin to form what is termed an excited triplet state (two electrons having parallel spins instead of antiparallel spins in a singlet state). This reversal of spin can occur as a result of spin-orbital coupling with an adjacent large atom. This excited triplet state can lose energy through thermal exchange with the surrounding solvent until it reaches the lowest triplet energy. The direct transfer of a triplet state down to its ground singlet state is forbidden and so the energy associated with the excited triplet state remains stored until a spin reversal again takes place through spin-orbital coupling with the large atom. At this point the energy released on getting back to the ground state is released as radiation (phosphorescence).

Naturally fluorescent biomolecules include those containing haem (iron containing) and redox proteins (e.g. haemoglobin, myoglobin, cytochrome-c, ferredoxins, rhodopsin) and pigments such as fluorescein, flavins, coumarin and cyanine. The chemical group in these molecules responsible for the fluorescence is called a fluorophore. Such groups will absorb radiation of a specific wavelength and re-emit radiation at a longer wavelength, whose value is dependent on the amount of energy lost in the transition according to the nature of the fluorophore and to some extent its chemical environment. An important example of a protein fluorophore was first identified and isolated from the jellyfish *Aequorea victoria* – namely the green fluorescent protein (GFP) which exhibits bright green fluorescence when exposed to blue light. The GFP gene can be inserted into a cell or organism, and the subsequent appearance of green fluorescence indicates that the gene of interest has been successfully taken up and expressed by the cell or organism under study [1]. This has been achieved for a wide range of bacteria, cell types, fungi and plants, flies, fish and animals, and has been applied to many important applications. For example, proteins are too small to be observed even under an electron microscope, but their fluorescence can readily be observed. Thus, by infecting a T-cell with HIV containing GFP labelled proteins researchers have found how HIV proteins pass from an infected cell to a noninfected cell. Furthermore, the fusion of GFP and GFP-like proteins to another protein does not alter the function or location of the protein in a cell. This has been used to follow the differentiation and proliferation of cells during embryonic development. A new scientific field known as optogenetics has evolved from GFP technology and is now routinely used by neuroscientists to study and control the switching on and off of individual neurons in a brain. The 2008 Nobel Prize in Chemistry was awarded to those (Martin Chalfie, Osamu Shimomura and Roger Tsien) who pioneered this GFP technology.

4.3 Classes of Spectroscopy

The various classes of spectroscopy involve different kinds of interaction of incident EM radiation with a test sample. The basic modes of practical operation and output spectra for the three main classes of optical spectroscopy (absorption, emission, scattering) are summarised in Figure 4.4.

The absorption, emission and scattering of EM radiation can only result if there is an interaction between the electric or magnetic field component of the EM radiation with the atomic or molecular structure of the test sample. The conditions that allow for interactions with the

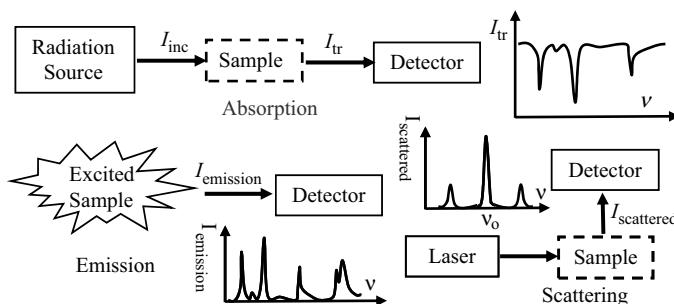


Figure 4.4 The three main classes of spectroscopy (absorption, emission, scattering) differ with respect to their mode of operation and output spectra.

electric field component of the EM radiation can be summarised as follows:

- Atoms or molecules can absorb EM radiation if its energy exactly matches that required to excite an electron from its ground state up to a higher energy state. This corresponds to *electronic spectroscopy* and is illustrated in Figure 4.5. Radiation will be re-emitted at the same frequency as the incident EM radiation, and if $\sim 180^\circ$ out of phase with it will result in an attenuation of the incident radiation.
- Molecules can absorb EM radiation and produce vibrational transitions if this absorption results in a change in the dipole moment of a molecule during its vibration. This corresponds to *vibrational spectroscopy*.
- Molecules in the gaseous state can absorb EM radiation and lead to rotational transitions only if they possess either a permanent electric or magnetic dipole moment. This gives rise to *rotational spectroscopy*.

The corresponding conditions for interactions with the magnetic field component are:

- Atoms or molecules can absorb EM radiation if they possess permanent magnetic dipole moments or magnetic moments associated with the spin states of electrons or nuclei.

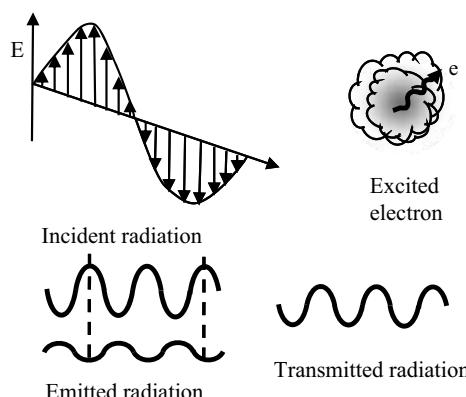


Figure 4.5 Incident EM radiation of energy equal to a transition between two electron energy levels in an atom will be absorbed and then re-emitted. The re-emitted radiation will occur at the same frequency, and if $\sim 180^\circ$ out of phase with the incident radiation will be attenuated.

Raman spectroscopy, named after the Indian scientist C.V. Raman, is used to study vibrational and rotational transitions and relies on inelastic scattering (Raman scattering) of monochromatic light, usually produced by a laser of wavelengths fixed at a value that can extend from the near ultraviolet through the visible and into the near infrared. Interaction of the laser light with vibrational or rotational modes of a molecule excites the molecule from the ground state to a virtual energy state. When the molecule relaxes it emits a photon and returns to a different rotational or vibrational state. The difference between the original state and the final state results in the energy of the emitted radiation being shifted up or down from that of the incident radiation. If the final vibrational state energy of the molecule is greater than the initial ground state, then from the principle of conservation of energy the emitted photon will be shifted to a lower frequency. This shift in frequency is termed as a Stokes shift. If the final vibrational state is less energetic than the initial state, the emitted photon is shifted to a higher frequency in what is termed as an anti-Stokes shift. The condition required for rotational Raman spectroscopy is that the molecule must have an anisotropic polarisability. The polarisability of a molecule is a measure of the extent to which it is distorted in an electric field. For a single atom this polarisability is isotropic, the distortion of the atom is the same irrespective of the direction of the applied field. Most molecules, however, exhibit anisotropic polarisabilities, their polarisability does depend on the field direction. This allows Raman rotational spectroscopy to be performed on molecules that do not possess a permanent electric or magnetic dipole moment, as required for normal rotational (microwave) spectroscopy. For vibrational Raman spectroscopy to occur the polarisability should change as the molecule vibrates. This condition exists for most molecules, including homonuclear diatomic molecules that do not possess a permanent dipole moment and are therefore not able to be examined by normal vibrational (infrared) spectroscopy.

Spectroscopic instruments can take on various forms of complexity in their design. Designs that can readily be miniaturised are of particular interest to designers and users of biosensors, and examples that fall into this category are shown in Figures 4.6 and 4.7

4.3.1 Electronic Spectroscopy

This is the measurement of the absorption or emission of EM radiation associated with *transitions* of electrons between energy states in an atom or molecule. Transitions between the ground, the lowest most stable, state and higher states of an atom involve energies in the X-ray region of the EM spectrum. Those of relevance to biosensors mostly correspond to energies in the UV and visible regions, and are associated with the excitation of electrons between energy levels characteristic of the molecular structure. Examples of this include the excitations that give rise to the colours of pigments, dyes, plants and flowers. The absorption spectra for two forms of chlorophyll are shown in Figure 4.8.

Leaves containing chlorophyll absorb wavelengths in the red and blue regions of the visible spectrum, and so reflect colours in the green region. Chlorophyll is not a very stable compound and is continually synthesised, a process requiring sunlight and warmth. Leaves therefore lose their chlorophyll molecules in Autumn. Another pigment often found in leaves is carotene, which absorbs blue-green and blue light and so reflects in the yellow part of the spectrum. Carotene, which can protect chlorophyll from oxidative damage and whose energy of light absorbed is transferred to chlorophyll, is more stable and persists in leaves when the

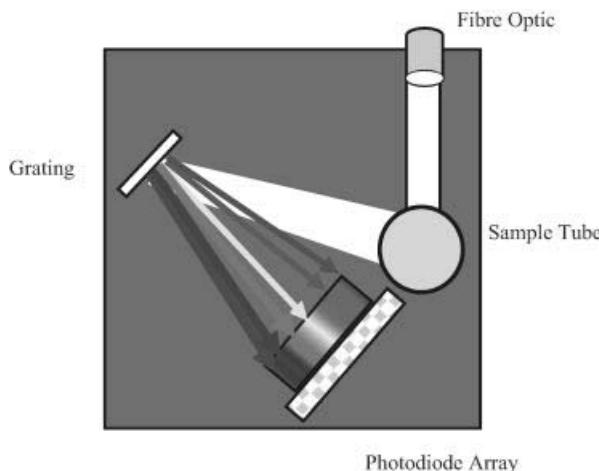


Figure 4.6 In this schematic of an absorbance spectrometer, a light beam passing through the test sample is reflected off an elliptical mirror at the rear of the sample container onto a grating polychromator and analysed by a linear array of photodiodes.

chlorophyll disappears. Anthocyanins, which absorb blue, blue-green and green light and so reflect in the orange and red region, are produced when sugars in the leaves increase sufficiently to react with certain proteins. The changing colours of leaves during Autumn result from the disappearance of chlorophyll and the retention of the carotene and anthocyanin pigments. The rule to remember is that if a substance absorbs in the visible range, it will be seen as the complementary colour to that absorbed. A material absorbing in the red will be seen as green since red and green are complementary colours. A complementary colour corresponds to the colour of white light if that colour is removed. Thus, if a solution appears yellow or blue, respectively, it has a chromophore absorbing in the purple or orange, respectively.

The group of atoms in a molecule most strongly involved in absorption of radiation is called a chromophore. An important example is the C=C double bond whose action as a chromophore arises from optical excitation of one of the electrons forming the double bond

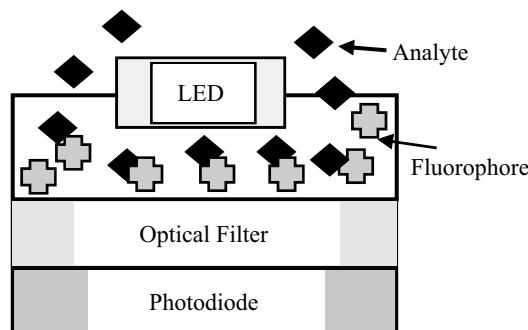


Figure 4.7 A solid-state fluorometer for detecting the change of fluorescence of a fluorophore on combination with a target analyte molecule.

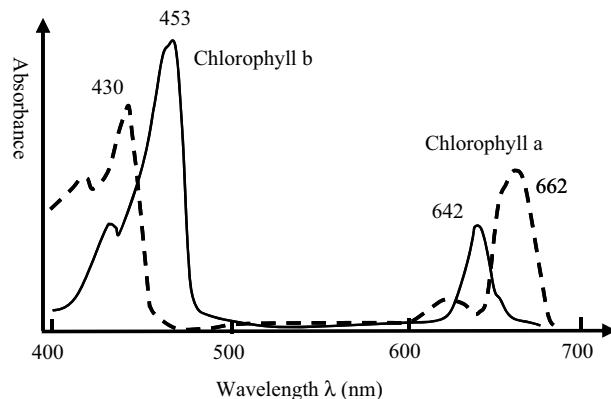


Figure 4.8 Absorption peaks for chlorophyll lie in the red and blue regions of the visible spectrum, and arise from electronic transitions. Leaves that contain chlorophyll therefore reflect green light, which is the colour of white light with red and blue light removed.

(called a π -electron) into an antibonding π^* orbital (chemists refer to this as a π to π^* -star transition). The absorption required for this electronic transition is in the UV. When the C=C double bond is part of a linear chain of conjugated carbon atoms ($-\text{C}=\text{C}-\text{C}=\text{C}-$) the π,π^* absorption shifts towards longer wavelengths and into the visible part of the spectrum. The retina of the eye contains a chromophore known as ‘visual purple’, which consists of a conjugated compound (retinol) attached to a protein that absorbs over the entire visible region.

Chromophores with increasingly more complex chemical structures exhibit absorption peaks that increasingly move towards longer wavelengths and exhibit more intense absorption peaks. This last effect is mainly associated with the increase in molecular size, and thus of the effective cross-sectional area presented as resonant absorbing sites for incident photons. A simple representation of this is shown in Figure 4.9.

An important concept in understanding spectra obtained by electronic spectroscopy is known as the Frank-Condon Principle. This is based on the fact that nuclei are far more

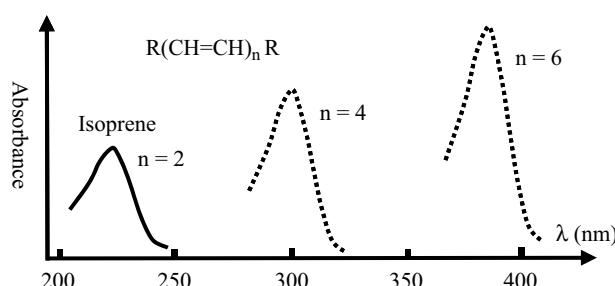


Figure 4.9 A schematic to show how the absorption spectrum of a linear conjugated hydrocarbon compound increases in absorbance magnitude and wavelength as the number of double bonds (n) increases. Isoprene with two double bonds has an absorption peak at 220 nm in the UV. β -carotene (from carrots) with $n = 9$ absorbs most strongly between 400–500 nm in the green-blue part of the spectrum.

massive than electrons, so that the assumption can be made that during the time taken by an electron to transition between energy states the nuclei do not change their internuclear positions. This is the basis for representing the initial excitation and final radiative transition back to the ground state as vertical lines in Figure 4.3. When an electron is excited into a higher molecular energy level, this can alter the electronic spatial distribution and hence the coulomb forces that determine the equilibrium length of a chemical bond. The situation shown in Figure 4.3 implies that the internuclear distance increases during the time period where the excited electron lost energy in interactions with vibrations of the molecule.

4.3.2 Vibrational Spectroscopy

Vibrational transitions, involving the stretching, bending or other forms of deformation of chemical bonds only produce an absorption or emission spectrum if the vibration interacts with EM radiation. Apart from the special case of Raman spectroscopy, this dictates that there must be a *change* in the dipole moment of the molecule. Energy levels between vibrational states are closer together than molecular orbital levels, so that transitions between vibrational energy states result in emission or absorption of energies in the infrared, rather than the visible, region of the EM spectrum. Diatomic molecules composed of dissimilar atoms (e.g. CO, HCl, NO, HF) produce vibrational (infrared) spectra because the magnitude of the permanent electric dipole moment changes as the bond length changes during vibration. However, homonuclear diatomic molecules (H_2 , O_2 , N_2 , Cl_2 , Br_2 , I_2 , F_2) do *not* produce infrared spectra because they do *not* possess a permanent dipole moment, and one is *not* produced during vibration. A molecule does not have to possess a permanent dipole moment to be infrared active. For example, because of its linear symmetry the carbon dioxide molecule does not possess a permanent dipole moment, but as shown in Figure 4.10 an asymmetrical stretching or bending of this molecule results in the appearance of a dipole moment and the presence of infrared absorptions.

If the two atoms of a chemical bond are displaced slightly from their equilibrium separation, the force acting to restore their original positions is to a first approximation proportional to the change in the bond length. The two atoms can thus be treated as two point masses connected by a massless spring, corresponding to the classic example of a harmonic

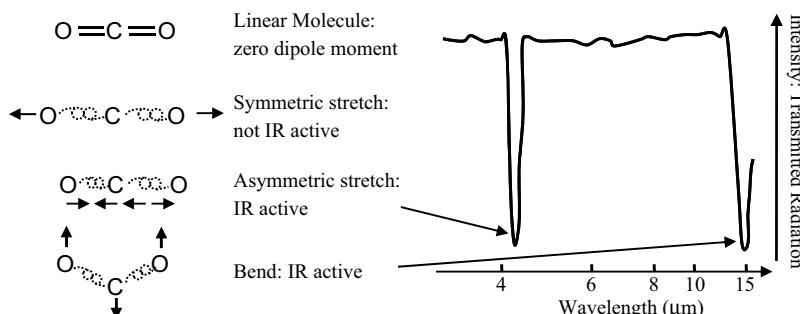


Figure 4.10 A molecule that does not possess a permanent dipole moment can be infrared active. Although the carbon dioxide molecule does not possess a permanent dipole moment, asymmetrical stretching or bending results in a dipole moment and an associated infrared absorption.

oscillator. For two atom masses m_1 and m_2 , and a spring force constant k (a measure of the chemical bond strength), the frequency of harmonic oscillation is given by:

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k(m_1 + m_2)}{m_1 m_2}}. \quad (4.1)$$

As for all atomic and molecular energies, the vibrational energy levels are quantised. The allowed vibrational energy levels are given by:

$$E_{\text{vibr}} = (n + 1/2)\hbar\nu \quad (n = 0, 1, 2, \dots),$$

where \hbar is Planck's constant. The selection rule for allowed transitions between vibrational energy levels is $\Delta n = \pm 1$ (a positive transition corresponds to absorption of energy from a lower to higher energy level, a negative one to emission). The transitional energy between vibrational states with quantum numbers $n + 1$ and n is therefore:

$$\Delta E = \left(n + \frac{3}{2}\right)\hbar\nu - \left(n + \frac{1}{2}\right)\hbar\nu = \hbar\nu.$$

The transition energy is thus independent of quantum number, so that all transitions associated with a particular molecular vibration occur at a single frequency given by Equation (4.1). Molecules with strong bonds (large k) between low mass atoms have high vibrational frequencies. Vibrations involving the bending of bonds tend to occur at lower frequencies (higher wavelengths) than stretching vibrations because they are generally less stiff. This is exemplified in Figure 4.10 for the carbon dioxide molecule.

At high vibrational excitations the separation of the atoms in a bond may extend so far beyond their normal equilibrium locations that the assumption they occupy a parabolic energy well, and hence exhibit harmonic motion, is no longer valid. This behaviour is termed anharmonicity and results in the allowed energy levels becoming closer together than described here.

4.3.3 Rotational Spectroscopy

This class of spectroscopy is also known as microwave spectroscopy. As shown in Figure 4.11, if a molecule possesses a permanent dipole moment (i.e. it is a polar molecule)

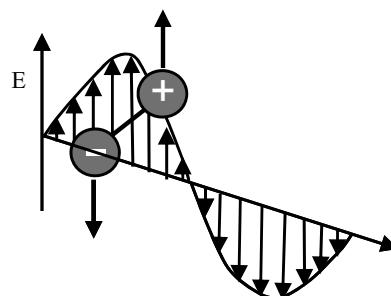


Figure 4.11 A molecule possessing a permanent dipole moment can interact with the electric field component of EM radiation, inducing a rotational torque which can cause its rotation rate to increase or decrease.

it can experience a rotational torque when exposed to an electric field, causing it to rotate more quickly (in excitation) or more slowly (in de-excitation). If a molecule, such as the oxygen molecule, does not have a permanent electric dipole moment but does possess a magnetic dipole moment, it can couple to the magnetic field component of EM radiation and exhibit rotational absorption.

The classical rotational kinetic energy of a molecule is expressed in terms of angular momentum J and rotational inertia I :

$$E_{\text{rot}} = (J^2/2I). \quad (4.2)$$

The quantised angular momentum number J ($J = 0, 1, 2, \dots$) when inserted into Equation (4.2) gives the energy of rotation confined to the values:

$$E_J = \frac{\hbar^2}{8\pi^2 I} J(J + 1) \quad \text{with } J = 0, 1, 2, \dots$$

The absorption and emission of EM radiation associated with changes in molecular rotation typically occurs at microwave frequencies. Rotational spectroscopy is practical *only in the gas phase*, where changes in rotation correspond to transitions between the molecule's rotational quantum numbers. In gases at high pressure, or in liquids and solids, the rotational motion is usually quenched due to molecular collisions.

Molecules can undergo a change in rotational energy at the same time as a change in vibrational energy, and this can be studied spectroscopically. The combination of the vibrational and rotational energy is given by:

$$E = E_{\text{vibr}} + E_{\text{rot}} = \left(n + \frac{1}{2}\right)hv + \frac{\hbar^2}{8\pi^2} \frac{J(J + 1)}{I}.$$

The allowed transitions between vibrational and rotational energy levels, and their combination, are shown in Figure 4.12.

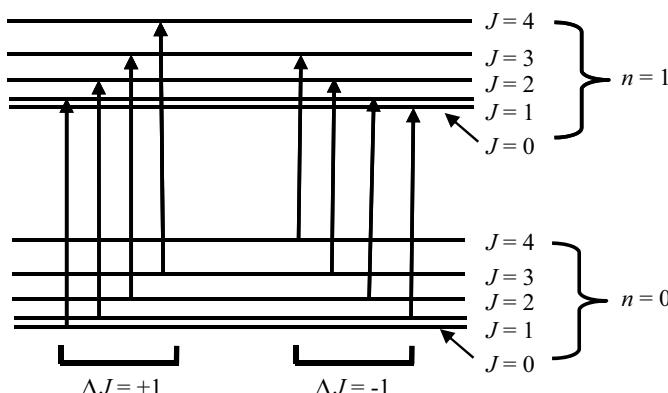


Figure 4.12 The allowed transitions between energy states for molecules experiencing a combination of vibrational and rotational energy changes. The $\Delta J = 0$ transition is forbidden.

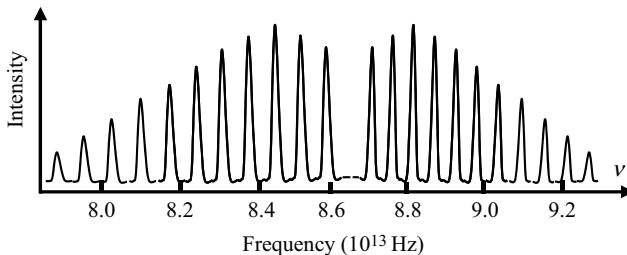


Figure 4.13 A schematic of the combined vibrational-rotational absorption spectrum for HCl, corresponding to the allowed transitions shown in Figure 4.11. The absence of the central frequency peak results from the $\Delta l = 0$ transition being forbidden (based on [2,3]).

The classic example of a combined vibrational and rotational absorption spectrum is that exhibited by the HCl molecule, a schematic of which is shown in Figure 4.13. The absorption peaks correspond to the allowed energy transitions shown in Figure 4.12. The spacing between the absorption peaks shown in Figure 4.13 can be used to compute the rotational inertia I of Equation (4.2). It should be noted that the peak corresponding to the central frequency, given by Equation (4.1), is absent from the absorption spectrum. This is the direct result of the selection rule for rotational spectroscopy that the molecule must possess a permanent dipole moment. During a rotational energy transition the transition dipole moment must not vanish and there must also be conservation of angular momentum when a photon (which possesses angular momentum) is emitted or absorbed by a molecule. These conditions are only met for rotational transitions $\Delta J = \pm 1$. The transition $\Delta J = 0$ is forbidden.

4.3.4 Raman Spectroscopy

We have noted that the selection rule for rotational Raman spectra is that the molecule must possess an anisotropic polarisability. It must experience nonsymmetrical distortion in an electric field. A distortion of a molecule in an electric field will return to that form after a rotation of 180° in the field, in other words twice every cycle. This leads to a quantised rotational Raman selection rule of $\Delta J = \pm 2$. When the molecule makes a transition with $\Delta J = +2$, the emitted photon has a lower frequency than the incident photon that excited the molecule, producing what is known as a Stokes shift. When the molecule makes a transition with $\Delta J = -2$, the emitted photon leaves the molecule with higher energy than the incident photon. There is an effective transfer of rotational energy from the molecule to the incident light beam, and the scattered light of higher frequency produces what are known as anti-Stokes lines.

If the molecules to be examined by Raman spectroscopy are adsorbed onto a roughened metal surface, or a surface coated with metallic nanoparticles, the intensity of the Raman signal is increased. This is considered to primarily arise from an excitation in the metal surface of collective electronic oscillations known as plasmons, which in turn magnifies the electric field component E of the incident radiation, and is known as *surface enhanced Raman spectroscopy*. It is particularly effective if the surface features have dimensions smaller than the wavelength of the incident radiation, and when the plasmon frequency is in resonance with the incident radiation. The energy of the incident radiation that excites the

Raman modes, and that of the emitted Raman signal, is each enhanced by a factor that is proportional to E^2 . The total enhancement is thus proportional to E^4 , and if both the incident and emitted radiation frequencies are close to the plasmon resonant frequency the total enhancement can be as great as 10^{10} . This permits the detection of single molecules by Raman spectroscopy.

4.3.5 Total Internal Reflection Fluorescence (TIRF)

TIRF has been used in various forms of biosensor. The basics of this are outlined in Figure 4.14, to show a laser light beam propagating along a glass waveguide, which can take the form of a microscope slide, for example. Total internal reflection of this light beam is possible if the refractive index n of the glass is larger than that of the medium making contact with the glass. Thus, should a light beam propagating through glass ($n = 1.5$) strike an interface with an aqueous medium ($n = 1.35$) at a sufficiently high angle of incidence it is refracted along a direction that is parallel to the interface. This refractive behaviour is governed by Snell's law:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2,$$

where θ_1 is the angle of the incident beam in the medium of higher refractive index n_1 and θ_2 is the refracted beam angle in the lower-index n_2 medium. At the critical angle of incidence the refraction direction becomes parallel to the interface between the two media (90° relative to the normal), so that the critical angle is given by:

$$\sin \theta_c = n_2/n_1$$

For our case of the light passing from glass ($n_1 = 1.5$) to an aqueous medium ($n_2 = 1.35$) this gives $\theta_c = 64.15^\circ$. At a larger angle of incidence than this the incident light beam is totally reflected back into the glass waveguide.

If the reflected wave has the same amplitude as the incident one, the incident and reflected components of the magnetic field vector parallel to the boundary of light incidence, and that

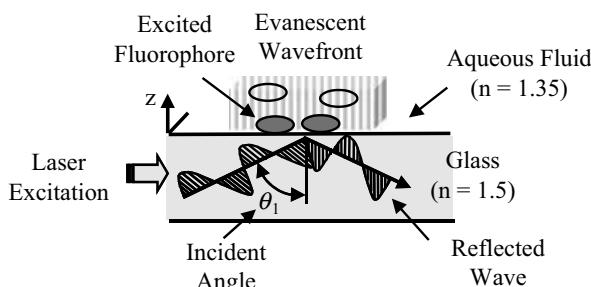


Figure 4.14 Total internal reflection fluorescence (TIRF) makes use of the EM evanescent wave generated by the interference between an incident and reflected light wave. Total internal reflection occurs above a critical angle of incidence, provided that the refractive index n of the light guide is greater than that of the adjacent medium. Excitation of fluorophores by the evanescent wave is confined to a region within ~ 100 nm from the waveguide-fluid interface. In this case two of the four fluorophores shown are excited.

of the magnetic flux density perpendicular to this boundary, superimpose destructively. The electric field components parallel to the boundary plane (x-direction) will however superimpose constructively. There can therefore be no solution of Maxwell's equations at the boundary plane without a nonvanishing transmitted wave. However, the incident and reflected waves have equal energy and so the principle of energy conservation demands that the transmitted wave cannot be a sinusoidal travelling wave. The only possible solution is that the transmitted wave is a standing wave that decays exponentially – in other words an evanescent wave. The solution for the electric field E beyond the interface (neglecting its time dependence) is given by [4]:

$$E(x, z) = A \exp(jkx) \exp(-z/d),$$

where $k = (2\pi/\lambda)n_2 \sin \theta$ is the component of the wave vector in the z-direction parallel to the interface, and λ is the vacuum wavelength of the light beam. The vector amplitude of the transmitted field A depends on the amplitude and the polarisation of the incident field, and the angle of incidence. The characteristic attenuation distance d depends on the angle of incidence, and is given by:

$$d(\theta_1) = \frac{\lambda}{2\pi} (n_1^2 \sin^2 \theta_1 - n_2^2)^{-1/2}.$$

At distance d (often referred to as the penetration depth) the magnitude of the evanescent electric field vector drops to 1/e (i.e. 0.37) of its initial value at the interface. The intensity of the evanescent wave is:

$$I(z) = C |A|^2 \exp(-2z/d),$$

where C is a constant for a given polarisation of the incident light beam. For incident radiation of wavelength $\lambda = 450$ nm, for example, the value for d is 304 nm (for $n_1 = 1.5$, $n_2 = 1.35$, $\theta_1 = 66^\circ$). At distances $z = 10$ and 300 nm from the interface, respectively, the evanescent wave intensity I will have fallen by 6% and 86%, respectively. The effective penetration depth of the evanescent wave decreases rapidly as the angle of incidence increases, and this can be adjusted to obtain the desired degree of penetration. For example, with $\theta_1 = 70^\circ$ in our present example, $d = 176.8$ nm and the corresponding evanescent wave intensity at a distance of 300 nm will have fallen by 97%. Thus, the evanescent wave does not effectively extend beyond a distance of several hundreds of nanometers into the adjacent medium, and typically less than a distance equal to the wavelength of the incident light.

In a typical biosensor application, the biosensing agent (e.g. an antibody or specific RNA sequence) is immobilised onto a glass slide or polymer film that acts as the waveguide for the incident light beam (e.g. [5–7]). If the glass slide is thin enough, multiple internal reflections can be generated along the slide. The zigzag path of the guided beam produces a periodic evanescent wave intensity distribution, each area of which can be ‘spotted’ with a different antibody to provide multiple analyses of different analytes. If a divergent light beam is used, with a small spread of incident angles, a homogeneous intensity distribution can be obtained after a long enough pathlength. The binding of analytes to the surfaces of nanoparticles can also be studied, where incident light is totally reflected within the particles (e.g. [8]). Fluorophores, potentially able to be excited into fluorescence by a wavelength within the bandwidth

of the incident laser beam, are introduced into an aqueous medium in contact with the functionalised surface. These fluorophores have also been designed to bind specifically to the target analyte, and because the evanescent field intensity falls off exponentially this provides a very sensitive method for detecting the binding of the analyte with the surface immobilised sensing agent. Excitation of fluorophores in the bulk of the aqueous medium is avoided, thus confining the fluorescence emission to a very thin region and giving a much higher signal-to-noise ratio than is achieved compared to conventional fluorescence techniques. The fluorescence can be collected and detected using the optics of a microscope focussed onto the surface of the glass slide.

4.3.6 Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR measures the resonant absorption of radiofrequency radiation when nuclei (with non-zero spin angular momentum) undergo a transition between spin states whose energies have been separated by an externally applied magnetic field. A nucleus with nonzero spin behaves like a magnet. In the presence of a magnetic field B the states acquire different values of energy:

$$E = -Bm_I g_I \mu_N,$$

where g is a numerical g-factor that is a measured characteristic of the nuclei. The parameter μ_N is known as the nuclear magneton, and its value is inversely proportional to the mass of the nucleus. For example, the two spin states ($m_I = \frac{1}{2}$ and $m_I = -\frac{1}{2}$) of a hydrogen ^1H nucleus in a magnetic field are separated by energy:

$$\Delta E = \frac{1}{2}Bg_I\mu_N - \left(-\frac{1}{2}Bg_I\mu_N \right) = Bg_I\mu_N.$$

More nuclei will be in the lower of the two energies (according to the Boltzmann distribution law). EM radiation of energy resonant with the energy separation induces transitions between the spin states and is strongly absorbed. The resonance frequency, $\nu = \frac{\Delta E}{\hbar} = \frac{Bg_I\mu_N}{\hbar}$, is proportional to the strength of the applied magnetic field and is in the radiofrequency region of the spectrum. Inductive search coils are used to provide and detect the radio-frequency signals.

NMR provides information on the different chemical environments of nuclei in a molecule. The nuclear magnetic moments of a nucleus will interact with the local magnetic field. This local field differs from the actual applied field because electronic orbital angular momentum induced in neighbouring nuclei by the applied field produces currents that create small additional magnetic fields. The resonance condition for a particular nucleus therefore depends on the magnitude of this additional induced field, and the shift away from the resonance condition defined by the applied field is known as the *chemical shift* of the nuclei. The magnitude of the resonance absorption is proportional to the number of equivalent nuclei in the same environment. Coupling between the magnetic moments of neighbouring nuclei produces fine structure in the absorbance signals and can be deciphered to give information of the magnetic and hence chemical environment of the molecule.

NMR in the form of Magnetic Resonance Imaging (MRI) or Magnetic Resonance Tomography (MRT) is used for medical imaging. This utilises the spin magnetic moment possessed by protons in water, as well as the fact that protons in different tissues relax back to their equilibrium states at different rates. The density of proton nuclear spins; the relaxation times (T_1 and T_2) that characterise the way that the spins relax back to their thermal equilibrium populations; spectral chemical shifts arising from interactions of proton magnetic moments with the moments of other types of nuclei; are used to construct images. MRI is particularly useful for imaging soft tissue (for which X-ray imaging is not effective) of high water content and little density contrast, such as the brain, muscle, connective tissue and tumours.

4.3.7 *Electron Spin Resonance (ESR) Spectroscopy*

ESR, or electron paramagnetic resonance (EPR) as it is also called, is analogous to NMR spectroscopy. The magnetic fields at which molecules containing unpaired electrons come into resonance with monochromatic radiation is studied. Molecules possessing unpaired electrons are termed paramagnetic, there magnetic moments align with an externally applied magnetic field. An electron possesses half-integral ($\frac{1}{2}$) spin angular momentum, which gives rise to two possible spin orientations distinguished by the quantum numbers $m_s = +\frac{1}{2}$ and $m_s = -\frac{1}{2}$. In the absence of an externally applied electric field the energies of these two orientations of the electron magnetic moment are equal, and as for NMR a magnetic field is used to cause separation of their energies. Transitions between the two states are induced by the absorption of EM radiation resonant with the energy separation.

The energy separation of the two electron spin orientations is proportional to the Bohr magneton, which takes the same form as the nuclear magneton but instead is inversely proportional to the mass of an electron. The mass of an electron is 1837 times smaller than that of a proton, and so the energy separation of the two electron spin orientations is greater for a given applied magnetic field than for nuclear spin orientations. Therefore, ESR resonance absorptions occur in the microwave rather than the radiofrequency region of the spectrum. The microwave generation and detection is performed using waveguides.

ESR differs from NMR in an important respect. The Pauli exclusion principle requires that whenever two electrons occupy one orbital their spins must be paired. It is therefore only possible to reorientate the spin of an electron if the electron is unpaired, and so ESR is restricted to chemical species with an odd number of electrons (radicals, so-called triplet states, and d-metal complexes). The intensity of an ESR absorption obtained for a sample will be proportional to the concentration of unpaired electrons present. ESR signals can show hyperfine structure that depend on the degree of coupling of the unpaired electron with magnetic nuclei, and this can therefore provide information on the electronic structure of radicals. If a reagent does not possess an unpaired electron, it can be ‘spin-labelled’ by incorporating into it a known paramagnetic molecule.

4.3.8 *Surface Plasmon Resonance (SPR)*

The excitation of surface plasmons by light to produce a surface plasmon-polariton is termed as surface plasmon resonance. Plasmons are quasi-particles arising from the quantisation of the harmonic oscillations of free electrons in a metal about the fixed positive charges of its crystal lattice. Surface plasmons are collective electronic oscillations that are confined to a

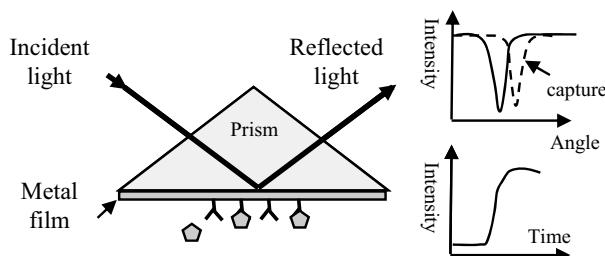


Figure 4.15 At a critical angle of light incidence the photons interact with surface plasmons in a metal film to create polaritons. This critical angle is sensitive to changes of the refractive index of the adjacent medium probed by the created evanescent wave, and changes of this (e.g. caused by analyte capture to an immobilised probe) can be monitored as a change in reflected light intensity.

metal surface and can interact strongly with incident polarised light to create polaritons. Coupling of light (photons) into surface plasmons can be achieved using a prism or optical grating to match the photon and surface plasmon wave vectors. A schematic of this is shown in Figure 4.15.

Basically, a laser beam is focused onto a thin ($\leq 50\text{ nm}$) metal film deposited onto a glass prism. At a critical angle of incidence of the light beam, energy is absorbed by the metal to excite the surface plasmons into resonance, and the intensity of the reflected light beam is significantly reduced. This strong coupling of the electric field component of the photon beam with the plasmons produces polaritons that are able to conduct along the metal surface. An evanescent EM wave is also created that decays exponentially from the metal surface into the adjacent medium, such as air or an aqueous fluid. EM radiation (light) of frequency below a plasmon resonance is reflected because the electrons screen the electric field component of the radiation, but at higher frequencies they are unable to respond fast enough and the light is transmitted. In most metals plasmon resonances occur in the ultraviolet, giving them a shiny appearance in visible light. For metals such as gold and copper, the resonances occur in the visible range and the metal takes on its characteristic colour.

The angle of incidence of the light beam to achieve the plasmon resonance is very sensitive to any changes of the effective dielectric properties (i.e. the refractive index) of the medium adjacent to the metal surface and within the influence of the evanescent wave. Changes of surface refractive index can result from the adsorption of molecules to the metal's surface, or by the capture of an analyte to a ligand attached to this surface. These effects can be monitored by observing changes in the critical angle of incidence, or more sensitively by detecting the sharp increase of the reflected light intensity as it shifts away from its original minimum value.

4.3.9 Förster Resonance Energy Transfer (FRET)

Named after Theodor Förster, the German scientist who formulated the original theory, this is the process whereby an electronically excited chromophore may transfer energy (without the emission of radiation) to a neighbouring chromophore as a result of the alignment and coupling of their dipoles. When the two chromophores are fluorescent, the term *fluorescence resonance energy transfer* is sometimes used. This resonance interaction does not involve a collision of the two chromophores, and also occurs without

conversion to thermal energy by way of induced molecular vibrations. The donor is the chromophore that absorbs the incident radiation energy, and the acceptor is the one that eventually receives this energy. For this resonance transfer to occur efficiently there must be an overlap of the emission spectrum of the donor chromophore with the adsorption spectrum of the acceptor. The separation distance between the two chromophores should also typically be no more than a critical distance (of the order 10 nm) and the FRET efficiency also depends on the degree of parallel alignment of the donor emission dipole moment and the acceptor adsorption dipole moment.

FRET leads to a decrease in fluorescence intensity of the donor chromophore and to the lifetime of its excited state. The fluorescence intensity of the acceptor is also increased. This means that FRET can be quantified as a ratio-metric determination of the two fluorescent signals. If the acceptor does not fluoresce, the resonant energy transfer can be monitored as a quenching of the donor fluorescence. Both of these methods can be employed in biosensor designs. For example, the acceptor molecule can be immobilised and fluorescence monitored to detect the presence of a donor analyte. Genetically encoded GFP and related fluorescent dye proteins have been employed in FRET measurements in biological cells, to monitor protein–protein interactions as well as the annealing of RNA oligonucleotides, for example. FRET activity can also be used to monitor the action potentials of neurons and heart muscle cells. The donor chromophore is a phospholipid-based molecule that binds only to the extracellular surface of the cytoplasmic cell membrane, whilst the acceptor molecule is a negatively charged hydrophobic molecule that can bind to either the external or internal membrane surface. Excitable cells such as neurons in their resting state have a membrane potential of ~ 70 mV with respect to the extracellular medium, and so the negatively charged acceptor molecules prefer to be located on the outer membrane surface, along with the anchored donor molecules. If the cells are radiated with light absorbed by the donor molecules, FRET activity can be detected as the characteristic fluorescence emission of the acceptor. An induced action potential will cause the membrane potential to reverse polarity (depolarises) to a value of $\sim +30$ mV. The acceptor molecules will now be repelled from the outer membrane surface, and attracted across the membrane to the positive internal surface. This transition can occur in less than half a second. At a donor-acceptor separation distance equal to that of the membrane thickness, FRET can no longer take place. This sequence of events is depicted in Figure 4.16, for the example where the donor can be excited to emit a red fluorescence, and the acceptor blue fluorescence. In the resting membrane state the donor and acceptor molecules can be close enough for FRET to take place, and blue fluorescence is emitted by the acceptor. When the cell undergoes an action potential event, the donor and acceptor molecules are separated and only the red fluorescence of the donor is evident.

The FRET procedure (also known as voltage sensitive dye imaging) depicted in Figure 4.16 has been used to identify neuron circuits that underlie rhythmic patterns of electrical output in nerve fibres [9]; the optical recording of individual neuron action potentials and synaptic potentials [10]; and the electrical activity of heart muscle cells [11]. Further examples of the application of FRET are given in Chapter 8.

4.4 The Beer-Lambert Law

The Beer-Lambert law, also known as Beer's law, relates the optical absorbance A of a chemical analyte to its concentration $[C]$. This law is usually given as the linear relationship:

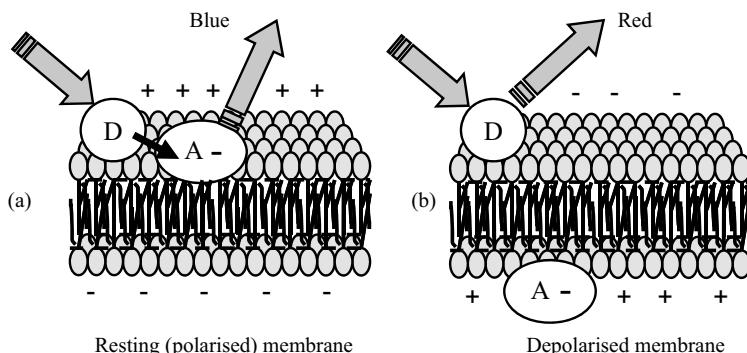


Figure 4.16 The behaviour of electrically excitable cells can be studied by monitoring how the blue fluorescence of a negatively charged acceptor molecule ceases when the membrane potential reverses polarity during an action potential event. The resting state is characterised (in this example) by emitted blue fluorescence arising from FRET from the excited donor molecule, which when energetically uncoupled from the acceptor through their separation being greater than the critical distance fluoresces in the red.

$$A = a(\lambda)l [C],$$

where $a(\lambda)$ is a wavelength-dependent absorption coefficient, and l is the optical path length through the analyte sample. If the concentration $[C]$ is given in terms of molarity, this equation becomes:

$$A = \epsilon l [C], \quad (4.3)$$

where ϵ is the molar absorption coefficient with units of $M^{-1} \text{ cm}^{-1}$. Ultraviolet, visible and infrared spectrometers usually display the absorption data in terms of transmittance T or %-transmittance. Transmittance T is defined as:

$$T = \frac{I_{tr}}{I_{inc}},$$

where I_{inc} is the incident light intensity, and I_{tr} is the light intensity after it has been transmitted through the sample. The relation between absorbance A and transmittance T can be derived using the scheme described in Figure 4.17.

In Figure 4.17 a sample of length l is shown that contains N absorbing species per cm^3 . Incident light of intensity I_{inc} enters the sample at $x=0$, and leaves it at $x=l$ as transmitted light of intensity I_{tr} . We will assume that each individual absorbing species at resonance presents a cross-sectional area σ to photons passing through the sample. (Typically $\sigma \sim 10^{-16} \text{ cm}^2$ for molecules.) At the resonant frequency, photons striking this molecular area are totally absorbed (they operate as black objects), but otherwise the photons continue passing through the sample. It is also assumed that the solution containing the analyte molecules does not absorb radiation over the range of wavelengths employed in the spectroscopic analysis. Assigning the light intensity incident on an infinitesimally thin slice dx at $x=x$ to be I_x , and dI to be the intensity absorbed in this slice,

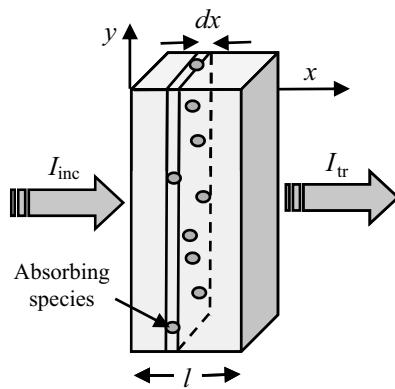


Figure 4.17 A sample of length l , containing resonating absorbing species, subjected to incident light of intensity I_{inc} , exits the sample as transmitted light of intensity I_{tr} . An infinitesimally thin slice dx absorbs light of intensity dI .

the fraction of light absorbed per unit area of the slice is equal to $\sigma N dx$. The fraction of light transmitted per unit area is thus:

$$\frac{dI}{I_x} = -\sigma N dx.$$

For the whole sample length we need to sum up (integrate) this result:

$$\int_0^l \frac{dI}{I_x} = -\sigma N \int_0^l dx$$

to give:

$$\ln(I_{\text{tr}}) - \ln(I_{\text{inc}}) = -\sigma N l$$

or

$$-\log_{10}\left(\frac{I_{\text{tr}}}{I_{\text{inc}}}\right) = \frac{1}{2.303} \sigma N l.$$

The relationship between the absorbance A and transmittance T is defined as:

$$A = -\log_{10} T = -\log_{10}\left(\frac{I_{\text{tr}}}{I_{\text{inc}}}\right) = \frac{1}{2.303} \sigma N l. \quad (4.4)$$

The concentration [C] of the absorbing species is given by:

$$[C] = N(1/N_{\text{av}})10^3 = N(6.023 \times 10^{20})^{-1} \text{ M (moles/litre)}$$

($N_{\text{av}} = \text{Avogadro's Number} = 6.022 \times 10^{23}$). Equation 4.4 can therefore be written as:

$$A = -2.61 \times 10^{20} \sigma [C]l. \quad (4.5)$$

Defining the molar absorption coefficient $\varepsilon = 2.61 \times 10^{20} \sigma (\text{M}^{-1} \text{cm}^{-1})$ then from Equation (4.5) we obtain the Beer-Lambert law given by Equation (4.1). Absorption peaks in the visible and ultraviolet range of wavelengths typically correspond to values for ε of $10^4 \sim 10^5 \text{ M}^{-1} \text{cm}^{-1}$. The molar absorption coefficient has units expressed in terms of the reciprocal of the product of concentration and optical path length. This can also be given in units of $\text{cm}^2 \text{mmol}^{-1}$ ($1 \text{M}^{-1} \text{cm}^{-1} = 1 \text{cm}^2 \text{mmol}^{-1}$) to more clearly indicate that ε is a molar cross-section for absorption. The larger the effective cross-sectional area of a molecule at resonance, the larger will be the reduction of the intensity of an incident light beam.

Example 4.1

The intensity of light of wavelength 254 nm is attenuated to 16% of its incident value after passing through an alcohol solution of 50 mM benzene contained in a 1.0 mm thin quartz cuvette. Calculate the absorbance A and the molar absorption coefficient ε . What would be the transmittance through a 2.0 mm thick cuvette?

Solutions:

The absorbance A is given by Equation (4.4):

$$A = -\log(T) = -\log(0.16) = 0.8.$$

The molar absorption coefficient ε is obtained from the Beer-Lambert Law (Equation 4.3)

$$\varepsilon = A/(l[C]) = 0.8/[(0.1)(50 \times 10^{-3})] = 160 \text{ M}^{-1} \text{cm}^{-1}.$$

From Equation (4.5) the absorbance is directly proportional to the light path distance through a sample. Doubling the pathlength (from 1 to 2 mm) will therefore lead to an absorbance of 1.6. The transmittance T from Equation (4.4) is given by:

$$T = \text{antilog}(-1.6) = 0.025.$$

This is equivalent to 97.5% of the incident light being absorbed by the sample.

4.4.1 Limitations of the Beer-Lambert Law

For a fixed optical pathlength through a sample, the linear relationship between absorbance and concentration given by Equation (4.1) no longer holds at very high values of the absorbance. The critical level for this will depend on the concentration of the analyte and its molar absorption coefficient. A value of [C] larger than around 10 mM often merits caution in this respect. At high concentrations analyte molecules in close proximity with

each other can experience electrostatic interactions, which can alter the absorbance coefficient ϵ as well as the refractive index of the sample. At higher concentrations and longer pathlengths there is also an increasing probability that two or more absorbing species will lie in the same optical path. This will lead to an underestimate of the molar absorption coefficient. The chemical equilibrium of some samples (e.g. dissociation of molecular salts) can also shift with increasing concentration. Active fluorescence or phosphorescence of the sample, as well as the scattering of light by particulate matter can also limit the linearity of the Beer-Lambert law. Practical limitations can also occur as a result of stray light entering the sample or the optical system. Because the absorbance coefficient ϵ is a function of wavelength, the use of nonmonochromatic radiation can lead to errors. This can be minimised if measurements can be made in a wavelength region close to a maximum absorption band having a relatively flat profile. The incident radiation should also take the form of parallel light rays, all of which travel the same distance through the sample.

Example 4.2

A sensor operates on the principle that an immobilised reagent R of known molar absorption coefficient will change colour when it reacts with a target analyte A to form a chemical complex according to the reaction $A + R \leftrightarrow AR$. The chemical equilibrium constant K is known for this reaction. Can the analyte concentration be determined by measuring the height of the distinct absorption peak of the AR complex?

Solutions:

The chemical equilibrium constant K for the reaction is given by:

$$K = \frac{[AR]}{[A][R]}. \quad (4.6)$$

The total concentration $\langle R \rangle$ of the immobilised reagent is given by the sum of the free reagent available for the reaction and the amount already present in the complex:

$$\langle R \rangle = [R] + [AR]. \quad (4.7)$$

Substituting $[AR]$ from Equation (4.6) into Equation (4.7):

$$\langle R \rangle = [R] + K[A][R] = [R](1 + K[A])$$

$$\text{to give the free reagent concentration } [R] = \frac{\langle R \rangle}{1 + K[A]}. \quad (4.8)$$

Substituting this result into Equation (4.7) we obtain:

$$[AR] = \frac{K[A]\langle R \rangle}{1 + K[A]}. \quad (4.9)$$

This equation does not provide a simple linear relationship between the concentration of the analyte $[A]$ and the absorption peak measured for the complex $[AR]$ and the

concentration of the analyte $[A]$. At sufficiently low analyte concentrations (where $[A] \ll 1/K$), Equation (4.9) predicts that the absorbance will be proportional to $[A]$. With increasing analyte concentration the absorbance will increase and approach a constant level for $[A] \gg 1/K$.

However, by combining Equations (4.8) and (4.9) the ratio of the absorbances exhibited by $[AR]$ and $[R]$ is given by:

$$\frac{[AR]}{[R]} = K[A]. \quad (4.10)$$

This equation gives a linear relationship proportional to the concentration $[A]$ of the analyte and independent of the reagent concentration. Thus, unless only low concentrations of an analyte are expected, spectrometric sensors that incorporate an immobilised reagent will require absorbance measurements to be made at *two* different wavelengths. Measurement of the absorbance of the product $[AR]$ is not in general sufficient.

4.5 Impedance Spectroscopy

This form of spectroscopy does not involve the interaction of EM radiation with a test sample. The principle method of impedance spectroscopy is measurement of the alternating current (*ac*) resistance of a system. This is usually accomplished by applying a small amplitude *ac* voltage perturbation and detecting the *ac* current response. Small perturbations are studied to ensure that the system is exhibiting as near to a linear response as possible (see Figure 4.18). The term *impedance spectroscopy* is used because the current response is determined for a range of *ac* frequencies, rather than for only one set frequency. The impedance Z of the system is defined as the ratio of the voltage-time signal $V(t)$ to the induced current-time signal $I(t)$:

$$Z = \frac{V(t)}{I(t)} = \frac{V_0 \cos(\omega t)}{I_0 \cos(\omega t - \phi)} = Z_o \frac{\cos(\omega t)}{\cos(\omega t - \phi)}, \quad (4.11)$$

where V_o and I_o are the peak values of the *ac* voltage and current signals, ω is the angular frequency ($2\pi f$) and ϕ is the phase shift between the current-time and voltage-time signals. These relationships are shown in Figure 4.18.

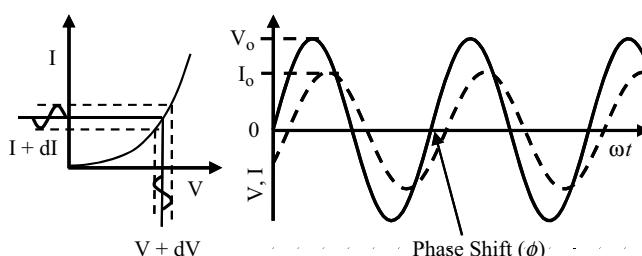


Figure 4.18 Sinusoidal voltage and current waveforms, with the current (I) lagging the voltage (V) by ϕ radians.

In complex notation, and making use of Euler's relationship, Equation (4.11) can be written as:

$$Z = Z_o \frac{\exp(j\omega t)}{\exp(j\omega t + \phi)} = Z_o \exp(j\phi) = Z_o(\cos \phi + j \sin \phi),$$

where $j = \sqrt{-1}$. The impedance is thus a complex quantity having real (Re) and imaginary (Im) components, commonly expressed as follows:

$$Z(\omega) = \text{Re}(Z) + j \text{Im}(Z) = Z' + iZ''. \quad (4.12)$$

For electrochemical and biosensor applications of impedance spectroscopy, the real component is associated with a resistance element R , and the imaginary component with a capacitance C having reactance $X (=1/(j\omega C))$. A system comprising of resistance and reactance will exhibit an impedance of the form:

$$Z(\omega) = \frac{R_o + j\omega\tau R_\infty}{1 + j\omega\tau} = R_\infty + \frac{R_o - R_\infty}{1 + j\omega\tau}, \quad (4.13)$$

where R_o and R_∞ are the limiting low- and high-frequency values of the equivalent series resistance, and τ is a characteristic time constant ($\tau = RC$). Representing the complex impedance of a RC system in the form $Z = R - jX$, these series components are

$$R = R_\infty + \frac{R_o - R_\infty}{1 + \omega^2\tau^2}; \quad X = \frac{(R_o - R_\infty)\omega\tau}{1 + \omega^2\tau^2}. \quad (4.14)$$

Equation 4.13 can be written in the form:

$$u + v = (R_o - R_\infty)$$

with

$$u = Z(\omega) - R_\infty \quad \text{and} \quad v = j\omega\tau(Z(\omega) - R_\infty).$$

In the complex plane u and v are orientated 90° to each other, with a vector sum equal to the real and constant quantity $(R_o - R_\infty)$. The right angle included by u and v is thus inscribed in a semicircle of radius $(R_o - R_\infty)/2$, which represents the locus of all values of the effective resistance R and reactance X as ω varies from a very low to a high frequency. The locus is a semicircle, and not a circle, because the capacitive reactance is only a negative quantity. A semicircle of this form, centred at x_1 , is described by the equation:

$$(x - x_1)^2 + y^2 = [(R_o - R_\infty)/2]^2. \quad (4.15)$$

By eliminating $(\omega\tau)^2$ from the two equations of (4.14) we can derive the following relationship:

$$(R - R_\infty)(R - R_\infty) + X^2 = 0. \quad (4.16)$$

On adding $[(R_o - R_\infty)/2]^2$ to both sides of Equation (4.16) a relationship of the same form as Equation (4.15) can also be derived:

$$[(R - R_\infty) - (R_o - R_\infty)/2]^2 + X^2 = [(R_o - R_\infty)/2]^2. \quad (4.17)$$

For a simple series RC system the locus of all values of the effective resistance R and reactance X , as we proceed from low to high frequencies, will thus be a semicircle of radius $(R_o - R_\infty)/2$ with its centre at $(R_o + R_\infty)/2$. The maximum value for X is $(-[R_o - R_\infty]/2)$, with a corresponding value of $R = (([R_o + R_\infty]/2))$, and occurs at the frequency where $\omega\tau = 1$.

Impedance spectroscopy can be applied to the study and analysis of a wide range of materials such as dielectric or ionic solids and liquids, as well as electrochemical reactions whose fundamental properties can be modelled as a series of combined electrical circuit analogues. A simple dielectric material, containing one type of dipole species and where electrode polarisation effects can be neglected, can be represented as a parallel combination of a resistance R and capacitance C . The frequency dependent complex impedance of such a system can be presented as a complex plane impedance plot (frequently misnamed as a Nyquist diagram in the electrochemistry literature) of Z' versus Z'' as shown in Figure 4.19. The ideal shape of the arc shown in Figure 4.19 is a semicircle with its centre on the real axis of the complex impedance plane. At very low frequencies the capacitive reactance ($1/\omega C$) is very large, equivalent to an open circuit, and the absolute impedance $|Z|$ is equal to the resistance R , with the phase angle $\phi = 0^\circ$. At very high frequencies the capacitive reactance becomes negligibly small, acting as a short circuit across the resistor, and $|Z| = 0$, with $\phi = 90^\circ$. These data can also be presented as a Bode plot, where unlike the complex impedance plot the frequency information is shown explicitly as either the logarithm of absolute impedance $|Z|$ or phase angle ϕ plotted as a function log frequency.

For materials containing mobile ions and where electrode reactions can occur, the equivalent electrical circuit becomes more complicated so as to take into account not only the bulk RC properties, but also electrode reactions, diffusion of ions and their adsorption at electrodes, and possible ion-pairing or other charge recombination effects. The major task is then to identify the most suitable physico-chemical model to adopt and if possible assign this to a pertinent equivalent circuit. This last task is often not straightforward – for example it can be shown by topological analysis [12] that 11 nonidentical circuits can give any single impedance versus frequency characteristic! Physicochemical insight is therefore required to identify which circuit is applicable.

The parallel RC circuit shown in Figure 4.19 can commonly represent the geometrical capacitance and bulk resistance of a sample, leading to the concept of a dielectric relaxation time $\tau = RC$ of the material. Parallel RC responses are also representative of an electrode reaction, where we can define a reaction resistance and a capacitance associated with an electrical double-layer at the electrode surface. An equivalent circuit of the form shown in

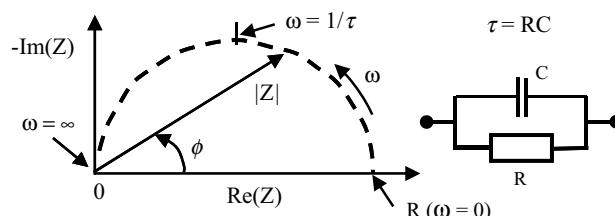


Figure 4.19 The complex impedance Z of a simple RC network, as a function of frequency, can be represented as a plot of the real and imaginary components of Z .

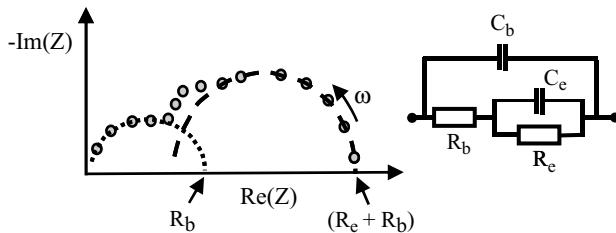


Figure 4.20 A plot of the complex impedance of an equivalent circuit representing a system consisting of bulk resistance R_b , a geometrical capacitance C_b , and a parallel combination of R_e and C_e , to represent an electrode reaction characterised by a reaction resistance and an electrical double-layer capacitance.

Figure 4.20 can be used to represent the combination of a geometrical capacitance and a bulk resistance in series with an electrode reaction. The bulk resistance is shown in series with the parallel electrode reaction because the charging and discharging of the electrical double layer involves charge transfer with the bulk material. If the associated RC time constants for the bulk dielectric relaxation time and the electrode reaction are well separated, then the Nyquist plot will take the form of two separated semicircles. In the more general case where there is an overlap of the time constants, the complex impedance plane plot takes the form shown in Figure 4.20.

The arcs shown in Figures 4.19 and 4.20 rarely take the form of semicircles with centres located on the real axis of the complex impedance plane. One common cause of this is the presence of diffusion controlled processes, for example, in the system under study. These processes can appear as distributed resistance and capacitance elements having a spread of relaxation times, or as frequency-dependent resistance elements. A procedure to analyse inclined or depressed semicircular arcs is described by Lemaitre *et al.* [13].

Problems

- 4.1. Figure 4.21 shows the absorption spectrum for a $10 \mu\text{M}$ solution of a dye dissolved in water, obtained using a 1 cm quartz cuvette. Peak absorptions are observed at 290 and 670 nm.
 - (a) Based on the data given in Table 4.1, explain the logic you would use to derive the colour of the dye. What is this colour?
 - (b) Estimate the percentage of incident radiation absorbed at wavelengths of (i) 290 nm; (ii) 420 nm; (iii) 670 nm.
 - (c) Estimate a value for the effective cross-sectional area of this dye molecule at 670 nm.
- 4.2. The evanescent wave in TIRF spectroscopy decays exponentially in energy with penetration depth, with a characteristic distance d given by:

$$d = \frac{\lambda}{4\pi} (n_1^2 \sin^2 \theta_i - n_2^2)^{-1/2}.$$

- (a) Calculate the value of d for the case of a laser beam (operating at 455 nm) directed at a 75° angle of incidence at an interface between quartz ($n = 1.55$) and a liquid ($n = 1.33$).

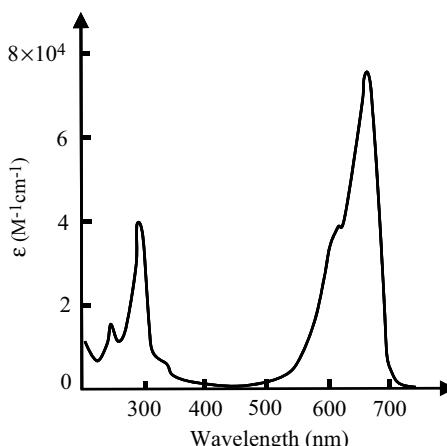


Figure 4.21 The absorption spectrum exhibited by a dye molecule. What is the observed colour of this dye?

- (b) Calculate the fractional intensity of the evanescent wave at a penetration distance equal to d , and also at 300 nm.
- (c) How might TIRF be applied in a biosensor?
- 4.3. The Cu^{2+} ion is the chromophore responsible for the visible absorption of copper sulphate (CuSO_4) in aqueous solution. Using a 5 mm cuvette, the transmittance of 0.1 M CuSO_4 at a wavelength of 600 nm was determined to be 0.3. Calculate the molar absorption coefficient of solvated Cu^{2+} and its effective cross-sectional area for photon capture.
- 4.4. A sensor operates by detecting changes in the intensity of light transmitted through a transparent substrate. A reagent R , immobilised on this substrate, selectively complexes with the target analyte A according to the reaction:

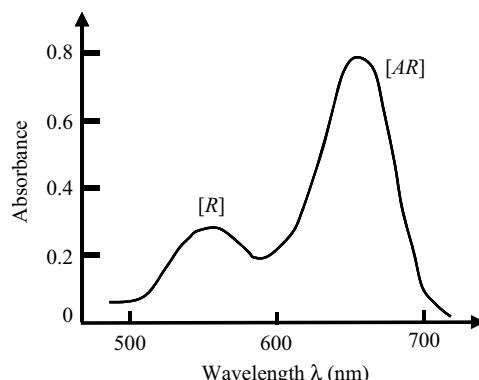
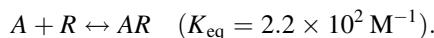


Figure 4.22 Absorption spectrum of light passing through a transparent substrate on which an immobilised a reagent (R) reacts with a target analyte (A) to form a chemical complex AR .

Based on the absorption spectrum shown in Figure 4.22, estimate the concentration of the analyte in the test sample. Assume that the effective cross-sectional areas of the *R* and *AR* chromophores are the same at their peak absorption wavelengths.

References

- [1] Tsien, R.Y. (1998) The green fluorescent protein. *Annual Review of Biochemistry*, **67**, 509–545.
- [2] Clermontel, D., Michel, J.P., Khatibi, P. and Vu, H. (1978) Vibration-rotation absorption spectra of HCl and HBr at very high foreign gas densities. *Infrared Physics*, **18** (3), 229–232.
- [3] Tipler, P.A. and Llewellyn, R.A. (1999) Chapter 9, in *Modern Physics*, 3rd edn, W.H. Freeman.
- [4] Klein, M.V. and Furtak, T.E. (1986) Chapter 2, in *Optics*, 2nd edn, John Wiley & Sons, Inc.
- [5] Engstrom, H.A., Andersson, P.O. and Ohlson, S. (2006) A label-free continuous total-internal-reflection-fluorescence-based immunoassay. *Analytical Biochemistry*, **357** (2), 159–166.
- [6] Tang, Y.J., Chen, Y., Yao, M., Zou, Z.X. *et al.* (2008) Total internal reflection fluorescence spectroscopy for investigating the adsorption of a porphyrin at the glass/water interface in the presence of a cationic surfactant below the critical micelle concentration. *Journal of Fluorescence*, **18** (2), 261–267.
- [7] Brandenburg, A., Curdt, F., Sulz, G., Ebling, F. *et al.* (2009) Biochip readout system for point-of-care applications. *Sensors & Actuators*, **B139**, 245–251.
- [8] Charlton, C., Gubala, V., Gandhaman, R.P., Prasad, R. *et al.* (2011) TIRF microscopy as a screening method for non-specific binding on surfaces. *Journal of Colloid and Interface Science*, **354** (1), 405–409.
- [9] Caciato, T.W., Brodfuehrer, P.D., Gonzalez, J.E., Jiang, T. *et al.* (1999) Identification of neural circuits by imaging coherent electrical activity with FRET-based dyes. *Neuron*, **23** (3), 449–459.
- [10] Stein, W., Stadele, C. and Andras, P. (2011) Single-sweep voltage-sensitive dye imaging of interacting identified neurons. *Journal of Neuroscience Methods*, **194** (2), 224–234.
- [11] Ella, S.R., Yang, Y., Clifford, P.S., Gulia, J. *et al.* (2010) Development of an image-based system for measurement of membrane potential, intracellular Ca(2+) and contraction in arteriolar smooth muscle cells. *Microcirculation (New York, NY: 1994)*, **17** (8), 629–640.
- [12] Foster, R.M. (1932) Geometrical circuits of electrical networks. *Transactions of the American Institute of Electrical Engineers*, **51**, 309–317.
- [13] Lemaitre, L., Moors, M. and Van Peteghem, A.P. (1983) The estimation of the charge transfer resistance by graphical analysis of inclined semicircular complex impedance diagrams. *Journal of Applied Electrochemistry*, **13**, 803–806.

Further Readings

- Abbas, A., Linman, M.J. and Cheng, Q.A. (2011) New trends in instrumental design for surface plasmon resonance-based biosensors. *Biosensors & Bioelectronics*, **26** (5), 1815–1824.
- Axelrod, D. (1989) Total internal-reflection fluorescence microscopy. *Methods in Cell Biology*, **30**, 245–270.
- Bally, M., Halter, M., Voros, J. and Grandin, H.M. (2006) Optical microarray biosensing techniques. *Surface and Interface Analysis*, **38** (11), 1442–1458.
- Daghestani, H.N. and Day, B.W. (2010) Theory and applications of surface plasmon resonance, resonant mirror, resonant waveguide grating, and dual polarization interferometry biosensors. *Sensors*, **10** (11), 9630–9646.
- Kastrup, L. and Hell, S.W. (2004) Absolute optical cross section of individual fluorescent molecules. *Angewandte Chemie-International Edition*, **43**, 2–5.
- Macdonald, J.R. (2005) *Impedance Spectroscopy: Theory, Experiment and Applications*, 2nd edn, John Wiley & Sons, New York.
- Wu, J.S., Liu, W.M., Ge, J.C. *et al.* (2011) New sensing mechanisms for design of fluorescent chemosensors emerging in recent years. *Chemical Society Reviews*, **40** (7), 3483–3495.

5

Electrochemical Principles and Electrode Reactions

5.1 Chapter Overview

Many biosensors operate by detecting or controlling electrochemical reactions involving the target analyte itself, or an electroactive chemical product of the analyte. The reactions occur at an electrode (variously termed the working, sensing or indicator electrode) whose main function is to electrochemically interact with chemical species close to the electrode surface by way of an electron transfer reaction. The reaction rate and type can be controlled by changing the potential of the electrode, with respect to a reference electrode, and is often quantitatively analysed by applying a constant or time-varying potential change and measuring the electrode current response, or alternatively imposing a current and measuring the response of the electrode potential. The basic principles of electrochemistry and electron transfer reactions at electrode surfaces are presented in this chapter, and will serve to introduce concepts covered in the following chapters on biosensors and instrumentation.

After reading this chapter readers will gain a basic understanding of:

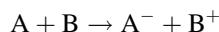
- (i) electron transfer and redox reactions at electrode-solution interfaces;
- (ii) anodic and cathodic reactions in electrochemical and electrolytic cells;
- (iii) the concept of the standard reduction potential;
- (iv) the application of the Nernst equation for determining the relationship between an electrode potential and the relative concentrations of the reactants in a redox couple;
- (v) the theory and application of reference electrodes;
- (vi) the basics of amperometry and cyclic voltammetry;
- (vii) Electrochemical Impedance Spectroscopy.

5.2 Introduction

Electrochemistry is the study of electron transfer processes that normally occur at electrode-solution interfaces, in what are termed *redox* reactions. The term *redox* reaction is shorthand for *reduction-oxidation* reaction. A large number of chemical reactions can be regarded as the outcome of a reduction and an oxidation process:

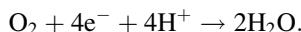
- Reduction: $A + e^- \rightarrow A^-$; (molecule A *receives* an electron)
- Oxidation: $B \rightarrow B^+ + e^-$. (molecule B *loses* an electron).

The overall result (sum) of these two reactions gives:



Such reactions can be created by mixing molecules A and B, where the electron released by B in the oxidation step is transferred to a nearby molecule A, which undergoes reduction. This is shown schematically in Figure 5.1.

In virtually all cells that contain a nucleus, an electron transfer scheme of the form depicted in Figure 5.1 occurs along the mitochondrial electron transport chain in a process that generates stored energy in the form of ATP molecules. A series of metallo-protein molecules facilitate the controlled transfer of electrons from strongly reducing agents (molecules containing hydrogen created by nutrient metabolism in the citric acid cycle) to highly oxidising agents (oxygen molecules). The stepwise energy released in this process is used to create a proton concentration gradient to drive the synthesis of ATP. We can deduce the rate of electron flow by examining the overall chemical reaction:



Adult humans during normal physical activity consume ~ 0.4 litres of oxygen per minute in this reaction. At standard temperature and pressure the volume a mole of gas occupies is 24.5 litres, and so this consumption is equivalent to ~ 16 mM per min of oxygen. The reaction involves the transfer of four electrons for every oxygen molecule consumed, and this equates to the transfer of ~ 64 mM, or 3.85×10^{22} , electrons per minute. If this occurred on a metal surface it would create a current flow of 103 Amps!

An electron transfer step in Complex I of the mitochondrial electron transport chain involves the iron (Fe) group in a protein containing an iron-sulfur active site being reduced as a result of receiving a high energy electron (initially donated by NADH) from a flavin

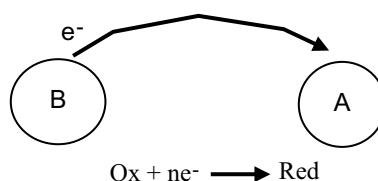


Figure 5.1 A redox reaction in which molecule A gains an electron donated by the oxidation of molecule B. The overall reaction is $A + B \rightarrow A^- + B^+$.

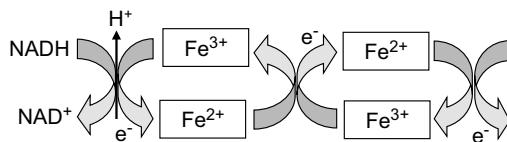
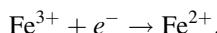
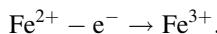


Figure 5.2 Electron transfer reactions down the electron energy gradient in complex I of the mitochondrial electron transport chain. Two electrons accepted initially from NADH pass one at a time through a chain of iron-sulfur groups embedded in proteins. The free energy released drives protons across the inner membrane of the mitochondrion to generate an electrochemical proton gradient that then drives ATP synthesis.

mononucleotide (FMN) molecule. This reaction, in which the 3+ oxidation state of Fe is *reduced* to the 2+ state, is written as:



The Fe^{2+} and Fe^{3+} states are normally referred to as the Fe(II) and Fe(III) forms, respectively. The next step in the electron transport chain in Complex I involves the Fe^{2+} group handing on an electron to a lower energy level in a neighbouring iron-sulfur site. The Fe^{2+} group thus becomes oxidised to Fe^{3+} as a result of passing on this electron:



The large amount of energy that would normally be released in one ‘blast’ from the energetically favourable reaction, that effectively combines hydrogen and oxygen to form water, is thus instead released in 15 or more small steps. This controlled release of free energy drives a series of three proton pumps to create a pH gradient across the inner membrane of the mitochondrion, that in turn drives the synthesis of ATP. These proton pumping reactions can be represented schematically as in Figure 5.2.

The reduction and oxidation reactions shown in Figure 5.2 can be arranged to each take place at electrode surfaces, as depicted in Figure 5.3. Each of these electrode reactions then forms what is known as an *electrochemical half-cell*. If the two electrodes were to be connected by a conducting wire, so that electrons given up in the oxidation of Fe^{2+} at one electrode are carried forward for the reduction reaction at the other electrode, we would then have a complete *electrochemical cell*.

Other examples of electron transfer reactions occurring at electrodes, namely the electroplating of copper onto a metal and the corrosion of iron, are shown in Figure 5.4.

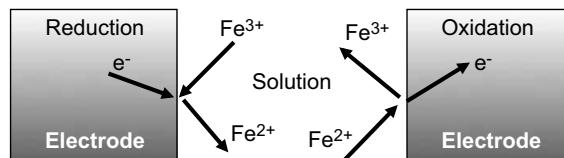


Figure 5.3 The reduction (left) and oxidation (right) reactions shown in Figure 5.2 can each take place at an electrode surface.

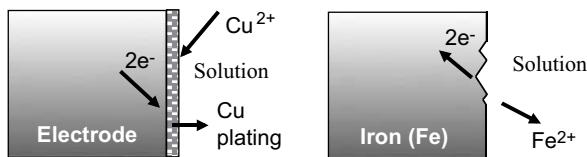


Figure 5.4 Examples of electron transfer reactions at electrodes. The electroplating of copper involves the reduction of solvated copper II (cupric) ions, and the corrosion (oxidation) of iron releases iron II (ferrous) ions into solution.

5.3 Electrochemical Cells and Electrode Reactions

Redox reactions taking place in a complete electrochemical cell, comprising two isolated half-cells and two solutions of molecules A and B, are shown in Figure 5.5. The solutions are contained within separate compartments, separated by an ion porous membrane, so that direct interaction between molecules A and B is not possible. However, allowing ion transfer between the two compartments ensures that no build up of net charge occurs in each half-cell to affect the electron transfer. An electrode is placed into each compartment. One of the electrodes acts as a source of electrons and the other as a sink, so that solution B gives up electrons to one electrode and solution A collects electrons from the other. The two electrodes are connected by a conducting wire that acts as a pathway for electrons given up by B to be carried to A. This forms a complete electrochemical cell.

A discharging electrochemical cell (also called a Galvanic or Voltaic cell) therefore produces electrical energy from spontaneous chemical reactions that occur within it. The driving force for the electron flow and reaction is the potential difference (voltage) generated between the two electrodes. A battery consists of two or more electrochemical cells connected in series.

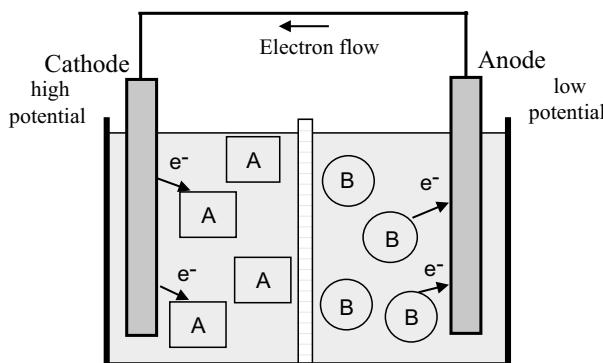


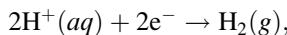
Figure 5.5 In an electrochemical cell the *reduction* reaction occurs at the *cathode*, inducing a *positive potential* relative to the solution. A *negative potential* relative to the solution is induced at the *anode* as a result of its *oxidation* reaction. An ion porous membrane allows the flow of ions between the two halves of this cell.

5.3.1 Anodes and Cathodes

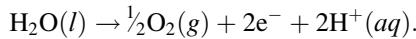
By convention, as shown in Figure 5.5, the electrode where oxidation occurs is called the anode, and the electrode where reduction occurs is called the cathode. In an electrochemical cell, which acts as a source of electricity, the cathode is at a higher potential than the anode. This arises because, when chemical species *A* undergoes reduction, electrons are withdrawn from the electrode (the cathode) to leave a positive charge on it. In the compartment containing the anode, oxidation corresponds to *B* transferring electrons to the electrode, so giving it a negative charge – corresponding to a lower, more negative, potential.

For the case of an *electrolytic cell* in which *electrolysis* occurs, an otherwise nonspontaneous reaction in the cell is driven by an externally applied direct electric current. As the Greek word *lysis* (meaning to break up) implies, we can say that an electrochemical cell decomposes chemical compounds using electrical energy. The amount of electrical energy required is equal to the Gibbs free energy of the reaction plus any losses, such as heat. The anode (by definition) is still the site for oxidation, but now electrons must be withdrawn from the chemical species. At the cathode there must also be a supply of electrons available for reduction. Therefore, the anode must now be made positive with respect to the cathode. A simple example is the electrolysis of water, shown schematically in Figure 5.6, where water is broken down into hydrogen gas and oxygen gas. Hydrogen is evolved at the cathode and oxygen at the anode.

The overall electrolysis reaction shown in Figure 5.6 involves two half-reactions. At the cathode, where reduction takes place, the reduction of hydrogen cations produces hydrogen gas, leading to a removal of electrons and a build up of positive charge:



whilst at the anode an oxidation reaction generates oxygen gas, with electrons accumulating on the anode:



It is this charge difference between the cathode and anode that produces the external current. We have written the oxidation reaction equation in the direction shown because the

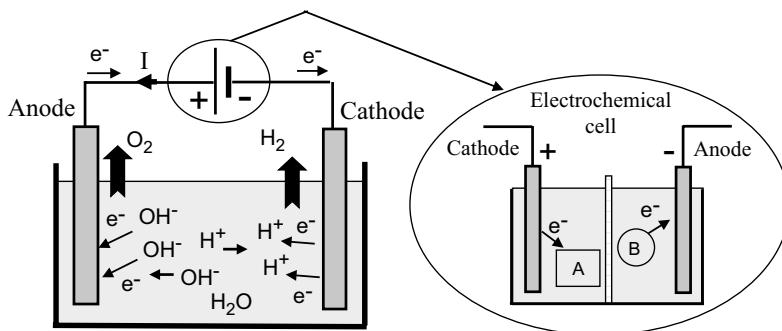
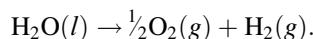


Figure 5.6 The electrolysis of water in an electrolytic cell. The current *I* required for the electrolytic reaction is supplied by an electrochemical (Galvanic) cell (or dc power supply). The direction of electron flow e^- is opposite to that of the conventional current *I*.

Table 5.1 The definition and electron transfer characteristics of the anode and cathode for an electrochemical cell (battery) and a cell supporting electrolysis

Location	Electrochemical cell	Electrolytic cell
Anode	<ul style="list-style-type: none"> • Site of oxidation • The negative terminal • Releases electrons to external circuit 	<ul style="list-style-type: none"> • Site of oxidation • The positive terminal • Releases electrons to external circuit
Cathode	<ul style="list-style-type: none"> • Site of reduction • The positive terminal • Accepts electrons from external circuit 	<ul style="list-style-type: none"> • Site of reduction • The negative terminal • Accepts electrons from external circuit

number of electrons gained in the oxidation half-reaction must equal the number of electrons lost in the reduction half-reaction. The nomenclatures (*aq*), (*g*), (*l*) and (*s*) are used to denote the physical states (aqueous solvation, gaseous, liquid or solid, respectively) of the chemicals. Adding together these two half-reactions gives the overall reaction as:



The two electrons ($2e^-$) appearing either side of the overall reaction cancel each other. Also, in accordance with the chemical formula for water the volume ratio of hydrogen to oxygen evolved is exactly 2 : 1. The number of electrons that are effectively conducted through the external wire is twice the number of hydrogen molecules that are generated, and four times the number of generated oxygen molecules. As given in Chapter 1 (Section 1.4) a litre of pure water at room temperature contains 1.0×10^{-7} M of H^+ and an equal number of OH^- ions. The electrolysis of water is therefore a slow process, but the electrolysis can be increased on adding dilute sulphuric acid, for example, to the water. This increases the concentration of H^+ in the solution and thus an increased rate at which current can be transported by H^+ conductivity.

Most students of physics and engineering are introduced to the terms *anode* and *cathode* as the terminals of an electrochemical cell (a battery) that drives conventional electric current through an external circuit. Confusion can arise with these terms when considering the case, as shown in Figure 5.6, where an electrochemical cell drives current through an electrolytic cell. Such confusion can be avoided by noting that in both cases we have the cathode defined as the site of chemical reduction, and the anode as the site of chemical oxidation. As summarised further in Table 5.1, in both cases the cathode accepts electrons from the external circuit, and the anode releases electrons to the external circuit. The only difference lies in the polarity of the electric potential of the two terminals.

5.3.2 Electrode Reactions

An electrode reaction is a chemical process involving the transfer of electrons into or out of the surface of a metal or semiconductor. This may be a reduction process whereby a species is reduced by the gain of electrons from the electrode, as in the following three examples:

- (i) $\text{Cu}^{2+} + 2e^- \rightarrow \text{Cu}$
- (ii) $\text{Fe}^{3+} + e^- \rightarrow \text{Fe}^{2+}$
- (iii) $2\text{H}_2\text{O} + 2e^- \rightarrow \text{H}_2 + 2\text{OH}^-$.

By convention, the current flowing for a *reduction* reaction is a *negative* quantity. Alternatively, the charge transfer may be an anodic reaction in which a species is oxidised by the loss of electrons to the electrode, as in the following three examples:

- (i) $2\text{H}_2\text{O} - 4\text{e}^- \rightarrow \text{O}_2 + 4\text{H}^+$
- (ii) $2\text{Cl}^- - 2\text{e}^- \rightarrow \text{Cl}_2$
- (iii) $\text{Pb} + \text{SO}_4^{2-} - 2\text{e}^- \rightarrow \text{PbSO}_4$.

By convention, the current flowing for an *anodic* process is a *positive* quantity.

The above examples indicate the possible diversity of electrode reactions. Furthermore, the electroactive species may be organic or inorganic; electrically neutral or charged; a species dissolved in solution; the solvent itself; a film on the electrode surface; or even the electrode material itself. Moreover, the product may be dissolved in solution, a gas, or a new phase on the electrode surface (e.g. growth of aluminium oxide on an aluminium electrode).

The reduction of a chemical species requires the transfer of an electron from occupied electron energy levels close to the Fermi level of the electrode to an unfilled molecular orbital (MO) of the chemical. Likewise, oxidation requires the transfer of an electron from an occupied MO to an unoccupied level near the Fermi level. These electron transfers, which involve the quantum mechanical tunnelling of electrons between the Fermi level and a molecular orbital, are depicted in Figure 5.7.

The Fermi energy level corresponds to the mean free energy of the most energetic electrons in the metal. At normal temperatures the electrons in a metal occupy energy levels below (more negative) than the reference level of zero electron volts, shown in Figure 5.7, which corresponds to the energy of an electron at rest some distance away from the electrode surface. An electron at rest has zero kinetic energy, and if removed far enough away from the influence of any kind of electric charge it will have zero potential energy. The energy transition required of an electron to relocate from the Fermi level to the reference zero level is known as the work function. The work function values for some metals are given in Table 5.2. The relative work function (hence Fermi energy) values for a gold and platinum

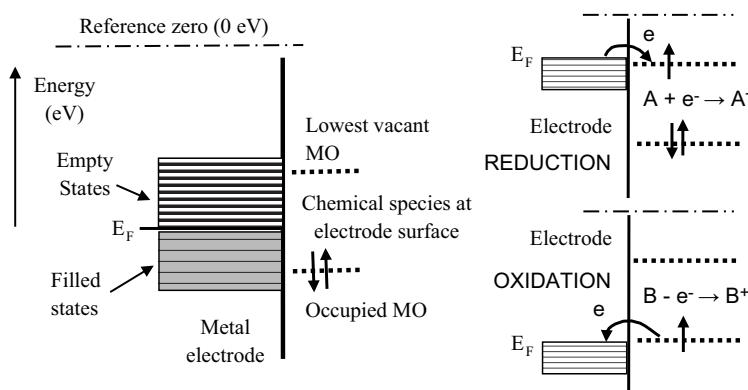


Figure 5.7 Reduction involves the transfer of an electron from the electrode's Fermi level E_F to an unfilled molecular orbital (MO) of a chemical species situated at the electrode surface. Oxidation involves the transfer of an electron from an occupied MO to an unoccupied level near the Fermi level.

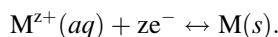
Table 5.2 Work function values for some common metals

Metal	Work function (eV)
Silver	4.26
Mercury	4.49
Copper	4.65
Gold	5.10
Platinum	5.65

electrode are shown in Figure 5.8, in relation to the highest electron occupied and lowest unoccupied molecular orbital level of a chemical species assumed to be at the electrode surface. Based on purely thermodynamic reasoning, the chemical may more readily be oxidised by a platinum electrode than by a gold one.

5.3.3 Electrode Potential

If a metal electrode M is dipped into its own metallic salt solution (i.e. a solution containing the corresponding metal ions M^{z+}) some of the atoms in the solid may dissolve into the solution as M^{z+} ions. Each atom that does this leaves z electrons behind, which results in a negative charge on the electrode and a positive charge (cation) in the solution. An equilibrium state is reached when the rates of atom escape and capture become equal and the following *redox equilibrium* is set up:



This is shown for the case of monovalent ions in Figure 5.9. The amount and polarity of the net charge on an electrode will depend on where the equilibrium lies for this reaction.

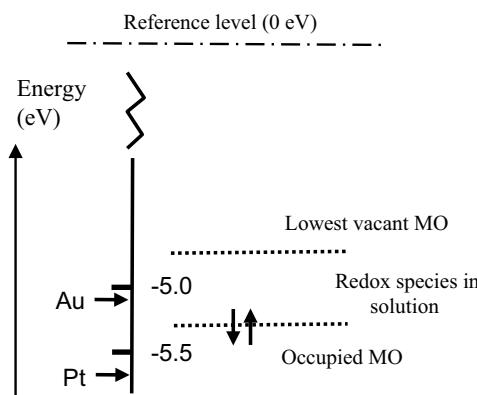


Figure 5.8 The relative Fermi energies for a gold and platinum electrode are shown with respect to the molecular orbital energy levels for a chemical species at the electrode surface (based on Table 5.2 and figure 1.1.3 of [1]). At their equilibrium (zero-current) potentials the chemical will more readily be oxidised by the platinum electrode than by the gold electrode. If their potentials are moved towards more positive values the gold electrode will, on purely thermodynamic grounds, more readily reduce the chemical than the platinum electrode.

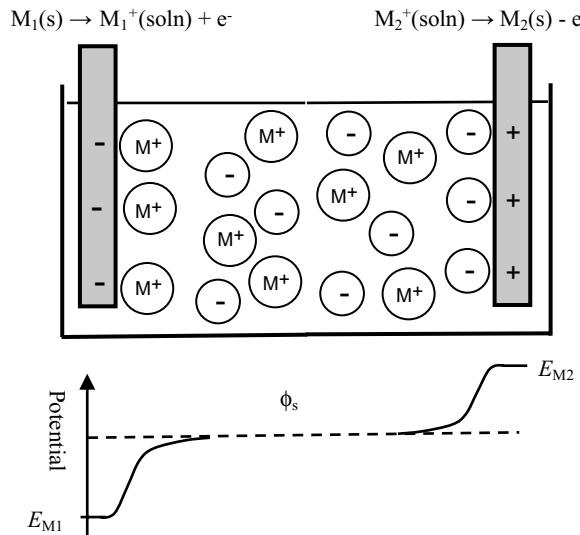


Figure 5.9 The amount and polarity of the net charge on an electrode immersed into a solution containing its metal salt will depend on where the equilibrium lies for the two electrode reactions shown. The potential difference (voltage) appearing between the two electrodes is given by $[(E_{M1} - \phi_s) - (E_{M2} - \phi_s)] = (E_{M1} - E_{M2})$. If E_{M2} is defined as the reference potential, then E_{M1} is the electrode potential of electrode M1 with respect to electrode M2.

If the charge is positive at equilibrium, it will appear to an observer as a positive electric potential, because it is found that more work is needed to bring up a positive test charge from infinity than to an uncharged electrode. If the charge on the electrode is negative, the observer reports that the electrode has a less positive (more negative) potential because less work is needed to bring up the positive test charge.

At equilibrium a potential difference is established across the electrical double layer at the metal-solution interface. The potential difference (voltage) appearing between the two electrodes in Figure 5.9 is given by:

$$[(E_{M1} - \phi_s) - (E_{M2} - \phi_s)] = (E_{M1} - E_{M2}),$$

where ϕ_s is the potential of the bulk solution. It is important to distinguish between the terms ‘potential’ and ‘voltage’. (Electronic engineers often use the term ‘applying a voltage’ to a circuit location, when what they are doing is applying a potential difference to this location with respect to a reference ground plane.) If E_{M2} is assigned to be the potential of a reference electrode, then E_{M1} is the potential with respect to that of the reference electrode M₂. The absolute value of an electrode potential cannot be determined – it can only be given with respect to another potential. If the metallic salt solution into which the electrode is immersed is of unit chemical activity, then at the standard condition of atmospheric pressure (101.3 kPa) at 25 °C (298 K), the equilibrium electrode potential is known as the *standard electrode potential*.

The establishment of an electrode potential is produced at the expense of the loss of free energy of the system. If E is the electrode potential of the electrode, and the electrode

reaction involves the transfer of n electrons, the electrical work available from the electrode is nFE (Joules), where F is the Faraday constant of value 9.65×10^4 C/mol. For a reversible electrode reaction the free energy change is given by:

$$\Delta G = \Delta G^\circ + RT \ln\left(\frac{[\text{reduced form}]}{[\text{oxidized form}]}\right) = \Delta G^\circ + RT \ln\left(\frac{[M]}{[M^{Z+}]}\right), \quad (5.1)$$

where ΔG° is the standard free energy. In Equation (5.1) the term $[M]$ represents the effective concentration (chemical activity) of the metal electrode. This can be considered to remain constant and so the normal convention is to assign $[M]$ as having a constant activity of 1. The energy per mol of any uncharged component of a solution is characterised by its chemical potential π_i . This energy is the Gibbs energy, and is a measure of the component's ability to do useful work (e.g. electrical or osmotic work). In a dilute solution the chemical potential π_i of a given component i of concentration C_i is given by

$$\pi_i = \pi_i^\circ + RT \ln C_i,$$

where π_i° is the standard chemical potential at C_i equal to 1 molar. The work done to add a charged species to a solution of electric potential ϕ differs from that required when the potential is zero. If the species ' i ' is an ion of charge $z_i e$, the extra work per ion is $z_i e \phi$, where z_i is the charge number. The electrochemical potential Π_i is related to the chemical potential π_i of an uncharged species by:

$$\Pi_i = \pi_i + N_A z_i e \phi = \pi_i + z_i F \phi.$$

When a positive ion (z_i positive) is in a region of positive potential, the electrochemical potential is greater than the chemical potential. This corresponds to the ion having a greater tendency to escape from the region or to undergo chemical change. The opposite is true for a negative ion in the same region – it is less reactive. An important consequence of this is that at the equilibrium state the *electrochemical potential of each chemical species is the same in every phase*. This is an important concept for understanding the functioning of a reference electrode, such as the silver-silver chloride or calomel electrode, for example.

A good example of a practical electrochemical cell (or more precisely a combination of two half-cells) is shown in Figure 5.10 in the form of the Daniell Cell. This cell consists of a zinc electrode immersed into a zinc sulphate solution as one half-cell, together with a copper electrode immersed in a copper sulphate solution. These two half-cells are ionically connected via a glass tube containing a gel saturated with a potassium chloride solution. This so-called 'salt bridge' prevents charge building up in each half-cell during the electron transfer reactions at each electrode. A porous ceramic can be used for this purpose instead of the salt bridge. Although the frit will produce small junction potentials, it is experimentally more convenient than preparing a conventional salt bridge. The Gibbs free energy (chemical potential) of the system is related to the emf E (1.1 V) of this cell by:

$$\Delta G = -nFE, \quad (5.2)$$

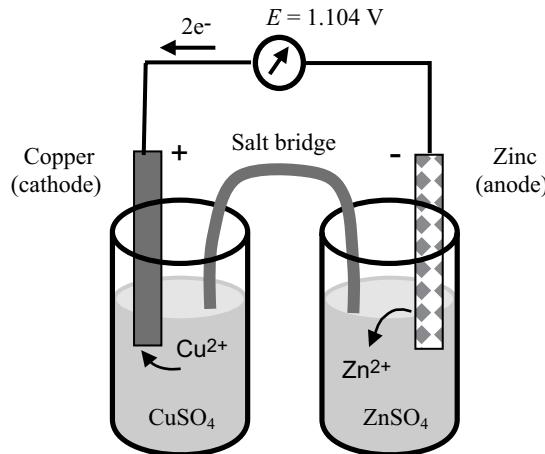


Figure 5.10 The Daniell Cell. The electrochemistry of this cell is discussed in detail in Example 5.1.

where n is the number of electrons transferred ($n = 2$ in this case). From Equations 5.1 and 5.2 the electrode potential is given as:

$$E = -\frac{\Delta G^\circ}{nF} - \frac{RT}{nF} \ln \left(\frac{[\text{reduced form}]}{[\text{oxidised form}]} \right) = -\frac{\Delta G^\circ}{nF} + \frac{RT}{nF} \ln \left(\frac{[M^{Z+}]}{[M]} \right). \quad (5.3)$$

The factor $(-\Delta G^\circ/nF)$ in Equation 5.2 is termed the *standard reduction potential* E° .

5.3.4 Standard Reduction Potential and the Standard Hydrogen Electrode

The reduction potential (also known as the *redox potential* or *oxidation-reduction potential*) is a measure of the tendency of a chemical species to acquire electrons and hence to be reduced. The reduction potential is measured in Volts. Each chemical species has its own intrinsic reduction potential. The more positive the potential, the greater is the species' affinity for electrons, and the greater its tendency to be reduced. Reduction potentials of chemicals in aqueous solutions are determined by measuring the potential difference between an inert sensing electrode (e.g. platinum, gold, graphite) in contact with the solution and a stable reference electrode connected to the solution by a salt bridge. The inert metal acts as a source or sink of electrons but takes no further part in the redox reaction. So far in this chapter we have considered the case of electrodes immersed in a solution (e.g. Figures 5.9 and 5.10). It is also possible to establish an equilibrium potential at a gas-metal electrode. The gas is passed over the surface of the electrode, which is immersed in a solution of ions related to the gas. If chlorine is the gas then the solution must contain chloride ions. Hydrogen can also be used as the gas, and so the solution should contain H^+ ions (protons). The *Standard Hydrogen Electrode* (SHE) is the reference from which all standard reduction potentials are determined (see Figure 5.11).

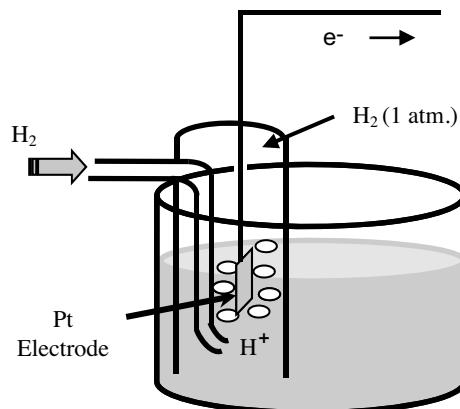
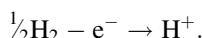


Figure 5.11 The standard hydrogen electrode (SHE) defines the zero reference level for the determination of the standard reduction potential of another half-cell system (with a shared electrolyte of hydrochloric acid). The temperature is 25 °C and hydrogen gas is passed at a pressure of one atmosphere over a pure platinum electrode.

Hydrogen gas can readily be oxidised to form protons:



If this reaction is performed under *standard* conditions – the Gibbs free energy is *defined* to be *zero*. Standard conditions are defined to be a temperature of 25 °C (298 K) with pure hydrogen gas being supplied at a pressure of one atmosphere (101 kPa) and passed over a pure platinum electrode. This half-cell reaction, in turn, defines the zero reference level for the determination of the standard reduction potential of another half-cell system coupled to a half-cell containing a standard hydrogen electrode (without the presence of junction potentials). The standard reduction potential (E°) is measured under standard conditions (298 K, a chemical activity of 1 for each ion participating in the reaction at a pressure of 1 atmosphere for each gas that is part of the reaction, and metals in their pure standard state). Standard reduction potentials for some reactions are given in Table 5.3.

The reactions given in Table 5.3 are spontaneous in the direction as presented if the standard potential E° is greater than zero, and are spontaneous in the reverse direction to that presented if the standard potential is less than zero. If an equation is reversed, so that the reactants become the products, the sign of E° must also be reversed. Thus, referring to the chemical reaction shown in Figures 5.2 and 5.3, from Table 5.3 we can define the reduction of the ferric ion as follows:



This informs us that the standard reduction potential (E_0) for this reaction is +0.771 Volts, as referenced against the standard hydrogen electrode (SHE). For the oxidation reaction (hydrolysis) of water, leading to the production of oxygen gas, we have to reverse the relevant reaction given in Table 5.3:



Table 5.3 Standard reduction potentials for some common half-cell reactions. (P. Vanysek, *CRC Handbook of Chemistry and Physics*, 87th edn, Boca Raton, 2007).

$\frac{1}{2}$ -Cell reaction	Standard potential E° (Volts)
$\text{F}_2 + 2\text{H}^+ + 2\text{e}^- \leftrightarrow 2\text{HF}$	+3.053
$\text{Au}^{3+} + 3\text{e}^- \leftrightarrow \text{Au}$	+1.498
$\text{O}_2 + 4\text{H}^+ + 4\text{e}^- \leftrightarrow 2\text{H}_2\text{O}$	+1.229
$\text{Br}_2 + 2\text{e}^- \leftrightarrow 2\text{Br}^-$	+1.066
$\text{Ag}^+ + \text{e}^- \leftrightarrow \text{Ag}$	+0.7996
$\text{Fe}^{3+} + \text{e}^- \leftrightarrow \text{Fe}^{2+}$	+0.771
$\text{Cu}^+ + \text{e}^- \leftrightarrow \text{Cu}$	+0.521
$\text{Cu}^{2+} + 2\text{e}^- \leftrightarrow \text{Cu}$	+0.3419
$\text{Hg}_2\text{Cl}_2 + 2\text{e}^- \leftrightarrow 2\text{Hg} + 2\text{Cl}^-$	+0.26808
$\text{AgCl} + \text{e}^- \leftrightarrow \text{Ag} + \text{Cl}^-$	+0.22233
$2\text{H}^+ + 2\text{e}^- \leftrightarrow \text{H}_2$	0.0000
$\text{CO}_2 + 2\text{H}^+ + 2\text{e}^- \leftrightarrow \text{HCOOH}$	-0.199
$\text{PbSO}_4 + 2\text{e}^- \leftrightarrow \text{Pb} + \text{SO}_4^{2-}$	-0.3588
$\text{Fe}^{2+} + 2\text{e}^- \leftrightarrow \text{Fe}$	-0.447
$\text{Cr}^{3+} + 3\text{e}^- \leftrightarrow \text{Cr}$	-0.744
$\text{Zn}^{2+} + 2\text{e}^- \leftrightarrow \text{Zn}$	-0.7618
$2\text{H}_2\text{O} + 2\text{e}^- \leftrightarrow \text{H}_2 + 2\text{OH}^-$	-0.8277
$\text{Al}^{3+} + 3\text{e}^- \leftrightarrow \text{Al}$	-1.662
$\text{K}^+ + \text{e}^- \leftrightarrow \text{K}$	-2.931
$\text{Ca}^+ + \text{e}^- \leftrightarrow \text{Ca}$	-3.80

This informs us that the electrolysis of water does not occur spontaneously at pH 0, but has to be driven by a voltage of at least 1.23 Volts, applied externally across the electrodes of the electrolysis cell. Multiplying up or dividing down the various quantities throughout a reaction equation does not change the E° value, because the ratio of the reactant to product concentrations is not changed. For example, doubling up the quantities in the water electrolysis reaction is written as:

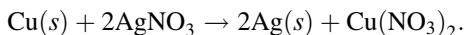


5.3.5 The Relative Reactivities of Metal Electrodes

The E° values given in Table 5.3 for metal reactions provides the means to judge the relative tendency for a reduction reaction to occur at an electrode made of that metal, compared to that of the reduction of an H^+ ion under standard conditions (i.e. at a hydrogen electrode). All of the metals appearing at the top of Table 5.3 (those having the largest positive E° values) have high reduction potential – they can be easily reduced and so act as strong oxidising agents. From Table 5.3 we can see that silver and copper are better oxidising agents than Zn^{2+} or Al^{3+} , for example. On the other hand, the large negative reduction potential (-3.8 V) of a calcium electrode makes it very difficult to reduce Ca^+ ions to Ca atoms. However, Ca^+ readily loses electrons to act as a reducing agent. In summary, as the reduction potential increases (i.e. its negative value decreases) the tendency of the electrode to behave

as a reducing agent decreases. Thus, metals such as calcium (Ca) and potassium (K) act as good reducing agents, whereas metals such as silver (Ag) and gold (Au) are very poor reducing agents.

Metals having the lower reduction potentials are not readily reduced but are easily oxidised to their ionic state by losing electrons. These displaced electrons can reduce a metal that possesses a higher reduction potential, and this can lead to an oxidised metal displacing a metal of larger reduction potential from its salt solution. For example, we can see from Table 5.3 that copper has a lower reduction potential than silver. If copper metal is added to a silver nitrate (AgNO_3) salt solution, silver atoms will precipitate from the solution as they are replaced by copper atoms to form copper nitrate salt:



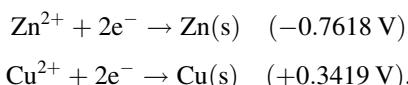
In general a metal at a lower position (standard potential) in Table 5.3 can displace the metals lying above it in the table from the solutions of their salts. Such metals are more reactive in displacing the other metals. Thus, of the elements listed in Table 5.3, calcium is the most electropositive element in solutions and fluorine the most electronegative.

Example 5.1

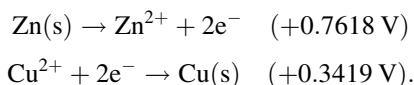
With the aid of Table 5.3 calculate the emf produced by the Daniell cell shown in Figure 5.10.

Solution:

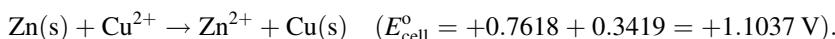
Write down the two half-reactions as given in Table 5.3:



One of these reaction equations (along with the sign of its E° value) must be reversed, because the number of electrons gained in one half-reaction must equal the number of electrons lost in the other half-reaction. Also, the sum of these two half-reactions gives the value of the cell emf, which must have a positive value for an electrochemical (Galvanic) cell because both reactions are spontaneous. The two half-reactions we require are therefore:



Adding these two half-reactions, and noting the sum of the potentials of the two half-cells:



The emf produced by the Daniell cell shown in Figure 5.10 is therefore $\sim +1.1 \text{ V}$. The positive value for the cell emf indicates that the change in Gibbs free energy ΔG for the

overall reaction is negative, and so proceeds spontaneously in the direction as shown above (from left to right) with the zinc electrode acting as the negative terminal.

An alternative way to obtain this result is to calculate the difference between the reduction potentials of the two half-reactions, to give the voltage appearing across the zinc and copper electrodes. The absolute value of this difference is 1.1 V. The polarity of each metal terminal is ascertained from the fact that electrons will flow to the half-cell having the more positive standard reduction potential. In other words, electrons will flow into the copper electrode, and so in terms of conventional current will act as the positive terminal.

In practice, the cell emf will depend on temperature and the relative concentrations of reactants and products. If the concentrations of the reactants increase relative to those of the products, the cell reaction becomes more spontaneous and the emf will increase. If the cell is used as a voltage source to drive an external electric current, the reactants will be consumed to form more products, and the emf will fall.

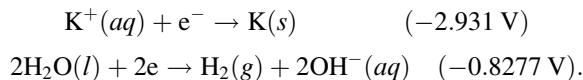
Example 5.2

Derive the value of the voltage that must be applied in order to electrolyse an aqueous solution of potassium bromide (KBr).

Solution:

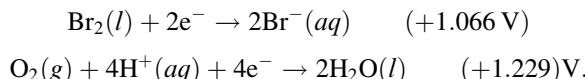
From our consideration of the electrolysis of water (Figure 5.6) it follows that if an electrolytic cell is operating with an aqueous solution, then the water can also be reduced at the cathode or created at the anode. These reaction equations must also be included in our calculations.

The possible reduction (negative E°) reactions at the cathode are therefore:



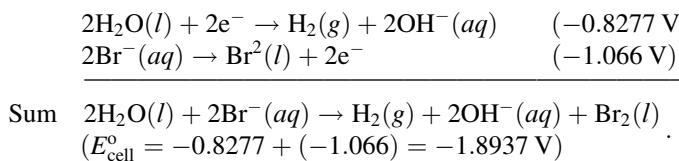
Of these two reactions, the one involving water (H_2O) has the largest (least negative) E° value, and is thus the most energetically favourable reduction reaction to take place at an electrode. Hydrogen gas, $\text{H}_2(g)$, will thus be produced at the cathode.

The possible oxidation (positive E°) reactions at the anode are:



The reaction involving bromine has the lowest E° value, and will be the most favourable oxidation reaction to occur at an electrode. Liquid bromine $\text{Br}_2(l)$ will thus be produced at the anode (at pH 0). The minimum voltage required to electrolyse potassium bromide is the sum of the E° values for these two most energetically favourable (and thus dominant) reaction equations. This sum must also have a negative value (electrolysis

being a nonspontaneous reaction) and so the bromine reaction equation must be reversed before performing the summation:



A minimum voltage of ~ 1.9 V must therefore be applied to the cell to achieve the required electrolysis of an aqueous solution of potassium bromide.

5.3.6 The Nernst Equation

In Example 5.1 it was noted that the emf of a cell is sensitive to changes in temperature and the relative concentrations of the reactants and products. The Nernst equation provides a quantitative way to determine the shift of an equilibrium potential E away from the standard reduction potential E° as a result of changes in the temperature and activities (equal to concentrations if dilute solutions) a_O and α_R of the oxidised and reduced species, respectively. The Nernst equation can be obtained directly from Equation 5.3 derived from the calculation of the free energy change of a reversible electrode reaction (and using chemical activity α rather than concentration):

$$E = E^\circ + \frac{RT}{nF} \ln \left(\frac{a_O}{a_R} \right)$$

or

$$E = E^\circ + 2.3 \frac{RT}{nF} \log_{10} \left(\frac{a_O}{a_R} \right) \quad (5.4)$$

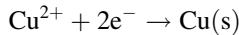
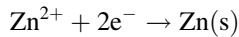
in which R is the universal gas constant ($8.31 \text{ J K}^{-1} \text{ mol}^{-1}$), n is the number of electrons involved in the electron transfer process, and F is the Faraday constant ($9.648 \times 10^4 \text{ C mol}^{-1}$). Thus, at $T = 25^\circ\text{C}$ (298 K) for a redox reaction involving a *single* ($n = 1$) electron transfer process:

$$E = E^\circ + 2.3 \frac{8.31 \text{ J.K}^{-1} \text{ mol}^{-1} \times 298.15 \text{ K}}{9.648 \times 10^4 \text{ C.mol}^{-1}} \log_{10} \left(\frac{a_O}{a_R} \right).$$

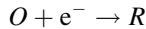
For reactions at a solid electrode surface we can take the electrode's activity to remain constant at 1. It is rarely, if ever, the case that a metal electrode is in equilibrium with its anion at the metal-solution interface. For example, all of the half-cell reactions given in Table 5.3 involve a metal cation. Therefore, at $T = 25^\circ\text{C}$ for a *single* ($n = 1$) electron transfer reaction where, as is normally the case the metal is the reduced species and $\alpha_R = 1$, the Nernst equation takes the form:

$$E = E^\circ + 0.059 \log_{10} \left(\frac{a_O}{1} \right) = E^\circ + 0.059 \log_{10}[O]. \quad (5.5)$$

The final right-hand term of this equation assumes that the concentration $[O]$ is sufficiently low to equate chemical activity to concentration. If we review the two half-cell reactions in the Daniel cell:



and inspect the convention used to denote a redox reaction (see Figure 5.1)



we find that the solid electrodes Zn(s) and Cu(s) equate to the reduced species R . For the unusual case, as might conceivably occur with a semiconductor electrode, the solid electrode material represents the oxidised species, the Nernst equation becomes:

$$E = E^\circ + 0.059 \log_{10} \left(\frac{1}{a_R} \right) = E^\circ - 0.059 \log_{10} [R]. \quad (5.6)$$

The direction of electron flow for a cell composed of two different half-cells can be predicted by comparing their redox potentials. A half-cell that accepts electrons from a standard hydrogen electrode is defined as having a positive redox potential, and a half-cell that donates electrons to a standard hydrogen electrode is defined as having a negative redox potential. Electrons will flow from the half-cell having the more negative E° value to the half-cell having the less negative (or positive) E° potential.

Example 5.3

A half-cell consists of an inert electrode immersed in a solution containing the $\text{Fe}^{3+}/\text{Fe}^{2+}$ couple. Determine the electrode potential, with respect to that of a standard hydrogen electrode, corresponding to the situation where the concentration of the Fe^{3+} ions is 100-times less than that of the Fe^{2+} ions.

Solution:

From Table 5.3, $E^\circ = +0.771$ for the $\text{Fe}^{3+}/\text{Fe}^{2+}$ couple, and so from equation (5.4) the electrode potential is (with $n = 1$):

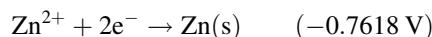
$$E = 0.771 + 0.059 \log_{10} (1/100) = 0.771 + [0.059 \times (-2)] = 0.889 \text{ V.}$$

Example 5.4

A zinc electrode in an electrochemical cell is immersed into a solution containing its own metallic salt. If the local concentration of Zn^{2+} ions at the electrode surface is 0.5 M, calculate the electrode potential with respect to the standard hydrogen electrode.

Solution:

The relevant electrochemical reaction involves the reduction by the zinc electrode of the oxidised species Zn^{2+} :



Equation 5.4 therefore takes the form (with $n = 2$):

$$E = -0.7618 + 0.118 \log_{10}(100/1) = -0.7618 + (0.118 \times 2) = -0.5258 \text{ V.}$$

5.4 Electrical Control of Electron Transfer Reactions

Electron transfer reactions at an electrode surface can be controlled by changing the electrical potential of the electrode. We can approach an understanding of this by considering how an electric current can be induced in a metallic conductor. The energy band model of a metal depicts the free electrons partially occupying a band of delocalised energy levels up to the Fermi energy level. If a potential difference (i.e. a voltage) is applied to the ends of a metal wire in an electrical circuit, electrons will percolate down the induced gradient of energy levels in the metal to produce a current, as shown in Figure 5.12. The *free energies* of the filled electron energy states are increased in the metal end connected to the negative battery terminal, and are lowered at the positively biased end.

It is important to distinguish between potential energy and electrical potential. The potential, with respect to a reference level, of a metal connected to a negative battery terminal will be lowered because less work will be required to bring a positive charge up to it. On the other hand, the potential energy of its electron energy states will be increased because they are negatively charged, and can facilitate a reduction reaction by raising the energies of electrons at the Fermi energy of the metal to where they can make a transition (usually by tunnelling) into an unoccupied molecular orbital of a chemical species adsorbed or near the metal's surface. This achieves a reduction reaction. For a positively biased metal, its electronic energy states are lowered, making a reduction process less likely for a given chemical species but improving the chance that an oxidising reaction can occur. These two possibilities are demonstrated in Figure 5.13.

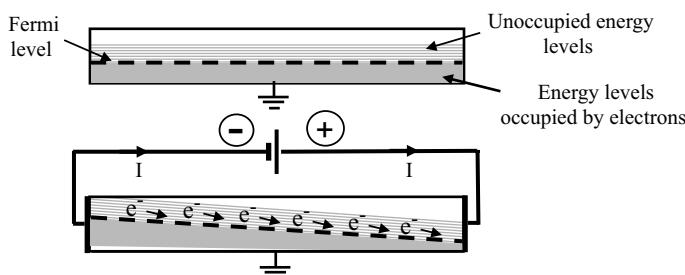


Figure 5.12 Electrons in the partially filled conduction band of a metal wire can be induced to flow down an externally induced potential energy gradient.

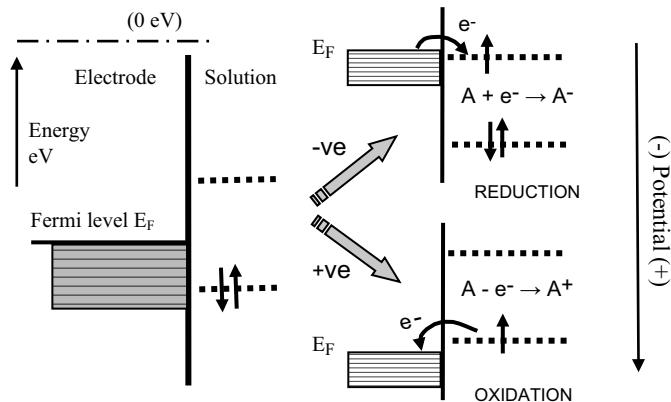
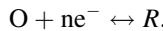


Figure 5.13 Electron transfer reactions at an electrode surface can be controlled by making the electrode potential more positive or negative. The situations shown here correspond to a more positive potential giving rise to the oxidation, and a more negative potential leading to reduction, of a chemical species.

We will consider a half-cell composed of an electrode immersed in a solution containing a chemical species that exhibits a reversible reaction:



where n is the number of electrons transferred in the reaction. The electrode is assumed to be an inert metal (e.g. platinum or gold) so that no electron transfer occurs across its surface when immersed in an electrolyte. In other words it performs as an *ideal polarised electrode*. To simplify our discussion we will also assume that the counter electrode completing the electrochemical cell is of sufficiently large surface area that its current density is very small. This will serve to make this counter electrode nonpolarisable, being able to conduct the cell current without changing its potential. This counter electrode could, for example, be platinum foil acting as a standard hydrogen electrode, separated from the first half-cell by a porous glass frit or a membrane. The voltage-current characteristics of the complete electrochemical cell are therefore determined solely by the performance of the first electrode, which we will call the *working electrode*. Also, the chemical concentrations of O and R are low enough that concentrations, rather than chemical activities, can be used in the Nernst equation (5.4).

If the working electrode potential is adjusted to enable an oxidation reaction where R is oxidised to O ($R \rightarrow O + ne^-$), the associated anodic (oxidation) current density is by convention assigned a positive value. To an external observer a conventional electric current is directed into the electrode. A negative value is given to the current associated with a reduction (cathodic) reaction where O is reduced to R ($O + ne^- \rightarrow R$), and conventional current flows away from the electrode and along the wire. At the condition of *dynamic equilibrium* the rates of oxidation and reduction are equal. Thus, the oxidation exchange current density I_O and the reduction exchange current density I_R density are equal and opposite to give a zero net exchange current density I_o :

$$I_O + (-I_R) = I_o = 0 \quad (5.7)$$

The magnitude of the exchange currents will each depend on the surface concentrations $[O]_s$ and $[R]_s$ of the electroactive species O and R, respectively, and on the electron transfer rate constants k_O and k_R :

$$I_O = nF[O]_s k_O; \quad I_R = -nF[R]_s k_R; \quad (5.8)$$

At the equilibrium state the concentrations of O and R at the electrode surface remain constant with time. There are no concentration gradients of the electroactive reactants O and R at the surface of the working electrode, and so the potential of the working electrode will remain steady at the standard potential value E given by the Nernst equation.

If a voltage is now applied across the cell so as to raise the working electrode potential to a value 0.24 V more positive than E° , for example, a steady state condition can only arise if an anodic (oxidation) current flows so as to change the concentration ratio $[O]_s:[R]_s$ from 1:1 to a situation close to 1000 : 1. The Nernst equation in fact demands that this be the case. Under steady state conditions, with an electrode potential $(E - E^\circ) = +0.24$ V, the oxidised form (O) of the redox couple is by far the most dominant species. As first observed by Tafel in [2], cell currents are often related exponentially to the value of $(E - E_o)$. This implies that the rate constants in Equation (5.8) depend on the applied potential, and the modern interpretation of this effect is embodied in the Butler-Volmer equation (named after two physical chemists, John Butler of England and Max Volmer of Germany):

$$I = I_o \left[\exp\left(\frac{\alpha_A nF(E - E^\circ)}{RT}\right) - \exp\left(-\frac{\alpha_C nF(E - E^\circ)}{RT}\right) \right]. \quad (5.9)$$

The factor α_A is the transfer coefficient for the electron tunnelling process involved in the anodic transfer of an electron from a molecular orbital in an oxidised species to the electrode, and α_C is the corresponding transfer coefficient for the cathodic reaction. For simple transfer processes $\alpha_A + \alpha_C = 1$, and it is commonly assumed that $\alpha_C \approx \alpha_A \approx 0.5$. The quantity $(E - E^\circ)$ is termed the *over-potential* to define the deviation from the equilibrium potential. Raising the working electrode potential to a value 0.24 V more positive than E° represents a high positive value for the over-potential. In this case the second term of Equation (5.9) can be neglected to give the anodic current density as:

$$\log(I_{OX}) = \log I_o + \frac{\alpha_A nF}{2.3RT} (E - E^\circ). \quad (5.10)$$

If the electrode potential is set to a value 0.24 V more negative than E° , steady state conditions will arise through a cathodic current leading to $[O]_s/R_s = 0.001$. The reduced form (R) then becomes the dominant species at the electrode surface. The second term in Equation (5.9) is now the dominant one and the cathodic current density is given by:

$$\log(I_R) = \log I_o - \frac{\alpha_C nF}{2.3RT} (E - E^\circ). \quad (5.11)$$

The total current given by Equation (5.9) is shown in Figure 5.14, as the sum of the two exponential components of the anodic and cathodic currents given by Equations (5.10) and (5.11). It is important to note that the current-potential response shown in Figure 5.14, based on the Butler-Volmer equation, is valid only for the situation where the electrode reaction is

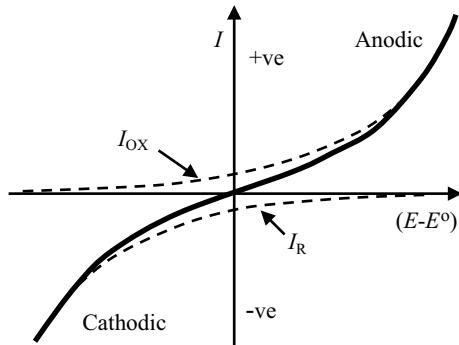


Figure 5.14 The current-potential response of an electrode reaction according to the Butler-Volmer equation (5.9). The total current is the sum of the anodic and cathodic currents given by equations (5.10) and (5.11). Any rate limiting steps associated with the *mass-transfer* of electroactive species between the electrode surface and the bulk electrolyte are not included.

controlled by the charge-transfer kinetics at the electrode surface. No account is made of the rate limiting step which may result from the relatively slow diffusion of an oxidised or reduced electroactive species away from the electrode surface into the bulk electrolyte, or the diffusion of these species to the electrode surface from the bulk electrolyte. Such *mass transfer* is required in order to maintain the concentration ratio $[O]_s:[R]_s$ dictated by the Nernst equation.

Based on the treatment presented by Bard and Faulkner [1, ch. 3], Equation (5.9) can be modified to take into account the influence of diffusion-controlled mass-transfer as follows:

$$I = I_o \left[\frac{[O]_s(t)}{[O]_{bulk}} \exp\left(\frac{\alpha_A nF(E - E^o)}{RT}\right) - \frac{[R]_s(t)}{[R]_{bulk}} \exp\left(-\frac{\alpha_C nF(E - E^o)}{RT}\right) \right]. \quad (5.12)$$

When the electrode reaction is controlled by the diffusion of the electroactive species (mass-transfer controlled) the current has a limiting value I_{lim} given by:

$$I_{lim} = \frac{nAFD}{\delta} [C]_{bulk}, \quad (5.13)$$

where A is the electrode surface area, D and $[C]_{bulk}$ are the diffusion coefficient and bulk concentration of the limiting electroactive species, respectively, and δ is the distance from the electrode surface into the bulk electrolyte over which the diffusion process is effective (the diffusion layer thickness). Diffusion processes are described in Chapter 10.

5.4.1 Cyclic Voltammetry

This is the name given to the experimental procedure whereby the potential (relative to a reference electrode) of a working electrode immersed in a solution containing an electroactive species is cycled at a steady rate either side of the equilibrium potential value E^o . The resulting current flowing through the counter electrode is monitored in a quiescent solution. The potential-time waveform has a symmetrical ‘saw-tooth’ profile, with the same positive

and negative sweep rates, that can range from a few millivolts up to 100 Vs^{-1} . Thus, for example, the potential at any time t for a negative-going voltage sweep is given by:

$$E(t) = E_i - vt,$$

where E_i is the initial potential and v is the linear sweep rate (Vs^{-1}). When the potential of the working electrode is more positive than E° , the electroactive species may become oxidised and produce an anodic current (i.e. electrons passing from the solution to the working electrode). On the return voltage scan, as the potential of the working electrode becomes more negative than E° , reduction may occur and give rise to a cathodic current. A schematic of such a cyclic voltammogram is shown in Figure 5.15, and reflects the International Union of Pure and Applied Chemistry (IUPAC) convention that the anodic current is plotted in the upper (positive) half of the potential-current plot, with the cathodic current given in the lower (negative) half. However, in many textbooks (e.g. [1]) and scientific publications (mostly from laboratories in the USA) the IUPAC convention is not adopted, so that the cathodic (negative) and anodic (positive) currents are placed in the upper and lower halves of the plot, respectively!

We can understand the basic shape of a voltammogram by rearranging the Nernst equation to form a time-dependent relationship:

$$(E_i - vt - E^\circ) = \frac{RT}{nF} \ln \left(\frac{[O]_s(t)}{[R]_s(t)} \right).$$

This equation can also be written:

$$\frac{[O]_s(t)}{[R]_s(t)} = \exp \left[\frac{nF}{RT} (E_i - vt - E^\circ) \right] \quad (5.14)$$

This form of the Nernst equation emphasises that the relative concentrations of the controlling electroactive species at the electrode surface are time-dependent in cyclic voltammetry. The example shown in Figure 5.15 corresponds to the situation where the electrode interfaces with a solution containing only the oxidised form of an electroactive species. The reduced form (R) is not present in the solution. The electrode potential is initially held at a

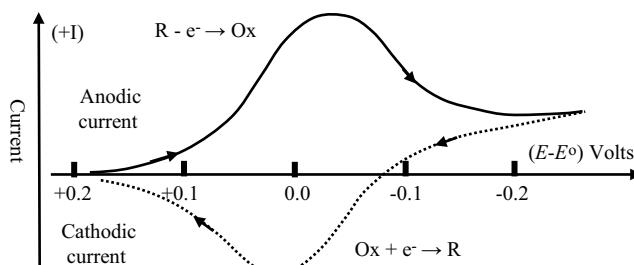


Figure 5.15 A cyclic voltammogram for a reversible redox reaction. Starting at a value above the standard reduction potential E° , with only the oxidised form of an electroactive species present, the electrode potential E is ramped to a value below E° , and then back up again. This generates the reduction current peak (solid line) followed by the oxidation current peak (dotted line).

potential sufficiently more positive than E° , so that no charge-transfer at the electrode occurs under steady state conditions. As the electrode potential approaches the E° value, reduction of the oxidised species commences, a reduction current flows and the concentration $[O]_s$ gets smaller. The reduced species (R) will diffuse away from the electrode and a concentration gradient of (O) is also created. Diffusion of (O) from the bulk electrolyte to the electrode increases and this leads to an increase of the reduction current. This process continues as the electrode potential gets more negative, until the potential falls below E° . At this point the surface concentration of (O) approaches zero according to Equation (5.14), and mass transport of O to the electrode surface attains a maximum rate to produce a peak of the reduction current. Beyond this stage of the negative potential sweep the concentration of (O) is depleted and the reduction current approaches the limiting value given by Equation (5.13). On initiation of the reverse potential sweep, the large concentration of the reduced species $[R]_s$ at the electrode surface provides favourable conditions for their reoxidation. As the electrode potential approaches and then rises above E° the generated concentration gradient of (R) at the electrode causes an increase of the oxidation current, which then passes through a maximum value before falling as the concentration of oxidisable species (R) is depleted.

As shown in Figure 5.15, the peaks of the reduction and oxidation currents occur either side of the standard reduction potential E° . The height and width of these current peaks depend on the rate at which the potential is cycled, as well as the kinetics of the charge-transfer processes at the electrode surface. Other factors include the rates of desorption of (O) and (R) from the electrode surface and their rates of diffusion (mass transfer) between the bulk solution and the electrode. These various contributions are shown schematically in Figure 5.16. Cyclic voltammograms contain a significant amount of information on the control of electrode reactions!

The geometry of the electrode can also influence the shape of a voltammogram. For a large area, flat, electrode, the diffusion (mass-transfer) processes depicted in Figure 5.16 are restricted to a planar surface over the electrode. A microelectrode protruding from a substrate, on the other hand, will have access to a much larger, hemi-spherical, diffusion surface. Cyclic voltammograms obtained using micro- or nano-scale electrodes will not exhibit the

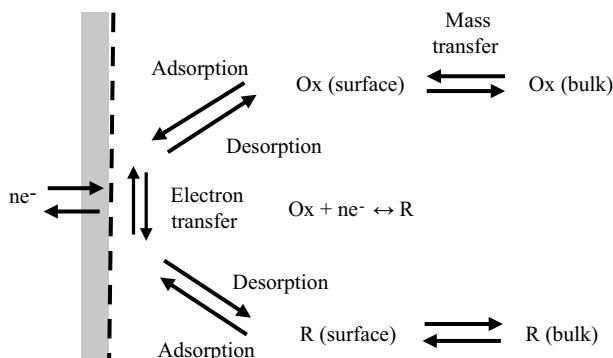


Figure 5.16 The flux of electrons ($-I/nF$) across an electrode surface for a reversible redox reaction is controlled by the kinetics of the electrochemical electron transfer and the mass transport of reduced (R) and oxidised (O) species to and away from the metal surface. The mass transfer involves the diffusion and/or migration of R and O down concentration gradients and electric fields.

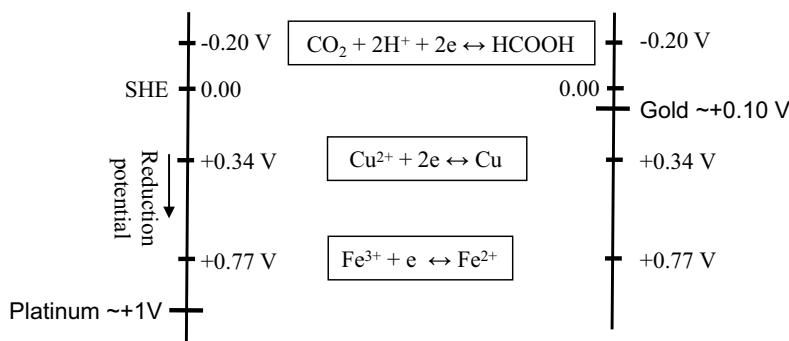


Figure 5.17 The standard reduction potentials for some electron transfer reactions, referenced against the standard hydrogen electrode (SHE). The approximate locations of the Fermi levels (vs. SHE) are shown for a platinum and gold electrode.

same level of mass-transfer controlled characteristics as those obtained using large area flat electrodes.

5.4.2 Amperometry

The most common class of biosensor operates as an amperometric device, where measurement is made of the current arising from an electrode reaction involving an electroactive analyte. The choice of electrode material is an important consideration for such devices. In general, as we have seen, when the potential of an electrode is moved from an equilibrium state towards more negative potentials, the chemical species that will be reduced first is the oxidant in the redox couple with the least negative (or more positive) standard reduction potential E° . Figure 5.17 depicts the relative situations for inert platinum and gold electrodes immersed in an aqueous solution containing iron and copper ions with dissolved carbon dioxide. The approximate locations of the Fermi levels for platinum and gold are based on the work function values given in Figure 5.8. As the potential of a platinum electrode is lowered to less positive potentials with respect to the hydrogen electrode reference level, the first species reduced will be Fe^{3+} , since the E° of the $\text{Fe}^{3+}/\text{Fe}^{2+}$ couple is the least negative (i.e. most positive), followed by Cu^{2+} and CO_2 . When the potential of an electrode is made progressively more positive, the chemical species that will be oxidised first is the reductant in the redox couple of least positive (or more negative) E° . Thus, for a gold electrode in an aqueous solution containing Cu , Fe^{2+} and CO_2 , copper will be the first to be oxidised as its potential is made more positive, followed by Fe^{2+} . The associated series of electrode current peaks produced by these electrochemical reactions are shown schematically in Figure 5.18. The dotted curves in Figure 5.18 represent the currents that would be observed if the potential had been increased slowly in small incremental steps, rather than as a relatively fast potential ramp.

In the form of amperometry known as *one-electrode amperometry*, the potential of a working (or indicator) electrode is maintained at a constant value with respect to a reference electrode. The current is measured after the introduction of an analyte to the electrochemical cell. The electrode potential is chosen to be close to the known value of E° for an electrochemical reaction involving the target analyte. For the case of *two-electrode amperometry*, a

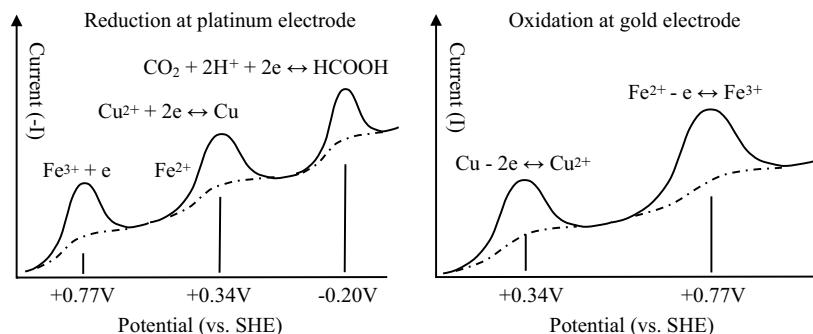


Figure 5.18 Reduction and oxidation currents produced as a function of the potential applied to platinum and gold electrodes, respectively, for the series of redox reactions shown in Figure 5.17. The current peaks are obtained using an applied linear voltage ramp, whilst the dotted curves show the steady-state current.

small constant potential difference is maintained between two working electrodes – one of which will sustain an anodic current and the other a cathodic current at the commencement of an electrochemical reaction. At steady state conditions these two currents will be of equal magnitude, but opposite polarity, and the electrode potentials will adjust to accommodate this.

5.4.3 The Ideal Polarised Electrode

An electrode at which no electron transfer can occur across the interface between the metal and a pure solution, regardless of the potential imposed by an outside source of voltage, is called an *Ideal Polarised Electrode*. While no real electrode can behave in this way over the whole potential range that can be applied to an electrode-solution system, ideal polarisability can be approached over certain limited potential ranges. For example, consider Figure 5.19 which shows the situation for a mercury electrode in contact with a degassed and clean potassium chloride (KCl) solution. At sufficiently positive potentials the mercury can oxidise, and at a very negative potential of -2.1 V the potassium ion (K^+) can be reduced. However, in the potential range between these processes, no electron transfer reactions occur. The reduction of water is thermodynamically possible in this region between $+0.25\text{ V}$ and -2.1 V , but occurs at a very slow rate at a mercury surface unless quite negative potentials are reached. The only other faradaic currents that could occur in this potential range would arise from electron transfer reactions involving trace impurities, such as metal ions, oxygen, and organic chemical species. With a clean and degassed KCl solution, such currents would be quite small.

5.4.4 Three-Electrode System

The overall chemical reaction taking place in an electrochemical cell consists of the net effect of two half-cell reactions. Most electrochemical experiments or devices (e.g. sensors) are concerned with electron transfer reactions that occur at only one of the electrode – namely the working electrode. An example of this is cyclic voltammetry described in

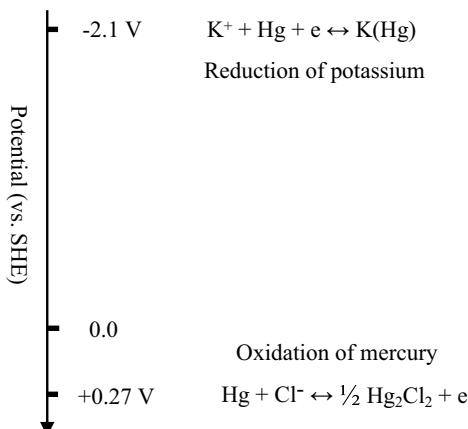


Figure 5.19 A mercury electrode in contact with a degassed potassium chloride solution containing no trace impurities behaves as an ideal polarised electrode. Over the 2.35 V range, between the oxidation of mercury and the reduction of potassium, no electron transfer reactions can take place.

Section 5.4.1. An experimental cell could therefore consist of the working electrode (also termed the indicator or sensing electrode) coupled with a counter electrode that also functions as the reference electrode. The working electrode's potential would be monitored or controlled with respect to this reference electrode, and the current response monitored. However, it is preferable to use a three-electrode system, the basic form of which is shown in Figure 5.20.

In Figure 5.20 the working electrode (WE) defines the electrode-solution interface under study. It should behave as a chemically inert and ideal polarised electrode. The reference electrode (RE) maintains a constant reference potential by operating as a nonpolarised electrode, a situation obtained by ensuring that no, or minimal, current flows through it. The purpose of the counter electrode (CE) is to supply the current required by the working electrode without in any way limiting or influencing the measured response of the electrode reaction. To achieve this it should have a much greater surface area than the working electrode. The feedback circuit with the operational amplifier drives the current between the

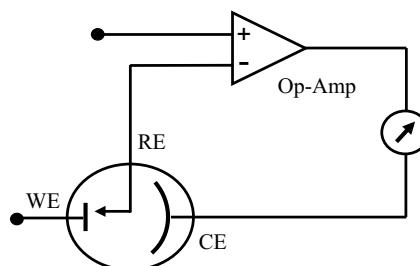


Figure 5.20 A three-electrode electrochemical cell, consisting of the working electrode (WE), a reference electrode (RE) and the counter electrode (CE). The operational amplifier drives the current between the working and counter electrode, but negligible current passes through the reference electrode.

working and counter electrode, while ensuring that none passes through the reference electrode circuit. This circuit maintains stability of the reference potential, and is referred to as a potentiostat. Further discussion of the three-electrode system and instrumentation for electrochemistry can be found in Chapter 7.

5.5 Reference Electrodes

The role of a reference electrode is to provide a fixed potential, which does not vary, during a potentiometric experiment or whilst an ion selective electrode is being used as a sensor. In all experiments it will be necessary to relate the potential of the reference electrode to other voltage scales, for example to the standard hydrogen electrode – the agreed standard for thermodynamic calculations. However, a hydrogen electrode is not a particularly convenient electrode to make or operate for routine use – and is potentially hazardous because it uses flowing hydrogen. Therefore, in practice, other *secondary* reference electrodes are used.

The concept of a reference electrode can be understood in terms of a complete electrochemical cell being composed of two $\frac{1}{2}$ -cells, as described in Section 5.3. The net cell potential is the difference of the electrode potentials of these two $\frac{1}{2}$ -cells. If one of the electrode potentials can be arranged to be of a fixed and unchanging value, then the net potential of the cell will only depend on the electrode reaction occurring at the other electrode (often referred to as the working or indicator electrode).

An essential feature, therefore, of a reference electrode is that it should provide a stable and reproducible electrode potential, and be relatively insensitive to changes in temperature. Compared to the hydrogen electrode it must also be easy to make and safe to use. In amperometric experiments the potential between the indicator (working) electrode and the reference electrode is controlled by a potentiostat, and as the reference half cells maintains a fixed potential, any change in applied potential to the cell appears directly across the working electrode-solution interface. The reference electrode serves the dual purpose of providing a thermodynamic reference and also isolates the working electrode to be the electrode-solution interface under electrochemical examination. In practice, however, any measuring device must draw current to perform the measurement. A good reference electrode should therefore be able to maintain a constant potential even if a few microamperes are passed through it. We say that the reference electrode should not be substantially polarised during the experiment or sensing operation. Ideally, it should be nonpolarisable.

A satisfactory reference electrode must also exhibit reversibility. This means that there should be negligible departure from a dynamic equilibrium state, with minimal susceptibility to outside electrochemical disturbances. Dynamic equilibrium will consist of a continuous flow of forward and reverse electron transfer reactions (i.e. cathodic and anodic) at the electrode surface. These exchange currents, which overall are self-cancelling, should be as high as possible. When the electrode is in use, either in measurement or as a sensor, the net current associated with this use should be much smaller than the exchange current, otherwise there will be a displacement of the electrode potential from its equilibrium value.

Most reference electrodes take the form of a three-phase electrode that is reversible to anions in solution. Their potentials are determined by the activity of these anions according to the Nernst equation. The three phases are in contact and in equilibrium with each other, and consist of a metal, a sparingly soluble solid salt formed by the cation of the metal with the anion to which the electrode is reversible, and a solution containing this

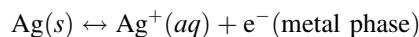
same anion. The most common and simplest example of this is the silver-silver chloride reference electrode.

5.5.1 The Silver-Silver Chloride Reference Electrode

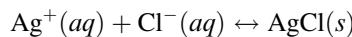
This electrode is normally represented as:



and consists of metallic silver, solid silver chloride, and an electrolyte solution containing a soluble chloride anion. The potential of this electrode is determined by the reactions shown schematically in Figure 5.21 and given in equation form below:



and



with a standard reduction potential of +0.22233 V (vs. SHE).

The potential of the electrode is determined at equilibrium by the Nernst equation (5.4):

$$E = E^\circ + \frac{RT}{nF} \ln a(\text{Ag}^+).$$

This corresponds to the equilibrium potential of a silver, silver ion electrode. If the activity solubility product K_s ($= a(\text{Ag}^+) \cdot a(\text{Cl}^-)$) in the solid silver chloride phase is assumed to be constant, then from Equation (5.4) the potential for the silver-silver chloride electrode is given as:

$$E = E_{\text{Ag}, \text{Ag}^+}^\circ + \frac{RT}{nF} \ln \left(\frac{K_s}{a(\text{Cl}^-)} \right)$$

or

$$E = E_{\text{Ag}, \text{AgCl}}^\circ - \frac{RT}{nF} \ln a(\text{Cl}^-).$$

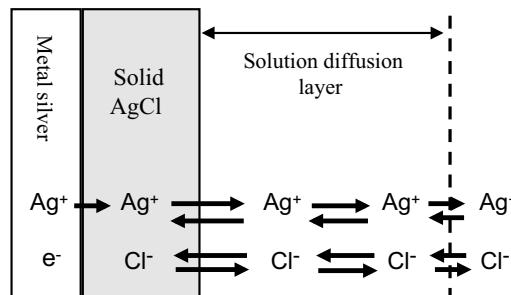


Figure 5.21 A schematic of the reactions that determine the equilibrium potential of the silver-silver chloride electrode.

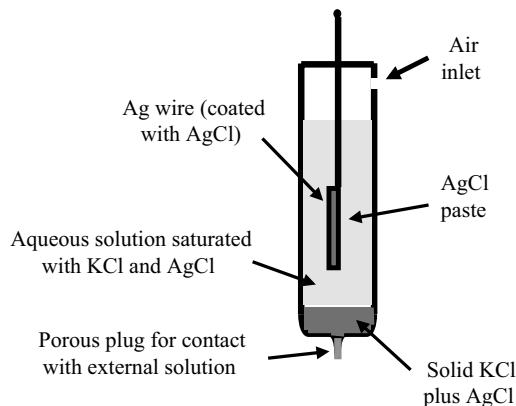


Figure 5.22 Schematic of the silver-silver chloride reference electrode.

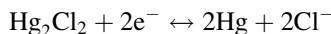
This expression shows how the potential of the silver-silver chloride electrode varies with the activity of chloride ions in the surrounding solution. The standard reduction potential value of +0.22223 V given in Table 5.3 corresponds to the recommendation by Bates and MacAskill [3] that silver-silver chloride electrodes should be standardised with 0.01 M Cl⁻ ion concentration (corresponding to an activity *a*(Cl⁻) of 0.00904 at 25 °C).

Commercially available silver-silver chloride electrodes commonly consist of a cylindrical glass tube containing a 4 M solution of KCl saturated with AgCl. The lower end of the tube is sealed with a porous ceramic frit. A ‘chloridised’ silver wire coated with a layer of silver chloride is immersed into the filling solution, which connects to the measuring system via a low-noise cable. A schematic of this system is shown in Figure 5.22.

A solid deposit of sodium chloride, with a smaller proportion of silver chloride, is included to ensure that the concentration of the chloride ions remains constant in the aqueous solution.

5.5.2 The Saturated-Calomel Electrode

Calomel is an old name for mercurous chloride (Hg₂Cl₂). As shown in Figure 5.23, this electrode consists of a mercury pool in contact with a paste made by mixing mercurous chloride powder and saturated potassium chloride solution. A constant chloride ion concentration is maintained in the paste through contact with a saturated potassium chloride solution. The half-cell reaction equation is:



and as given in Table 5.3 has a standard reduction potential of +0.26808 V (vs. SHE). The reduction potentials of the calomel electrode for various concentrations of the chemical components are given in Table 5.4.

A reference scale to convert between the standard hydrogen electrode, the silver-silver chloride and the calomel electrode is given in Figure 5.24.

Table 5.4 Standard reduction potential values of the saturated-calomel electrode for various concentrations of its chemical components. (P. Vanysek, *CRC Handbook of Chemistry and Physics*, 87th edn, pp. 8–21, Boca Raton, 2007).

Composition	Standard potential E° (Volts)
Saturated NaCl (SSCE)	+0.2360
Saturated KCl	+0.2412
1 Molal KCl	+0.2800
1 Molar KCl (NCE)	+0.2801
0.1 Molar KCl	+0.3337

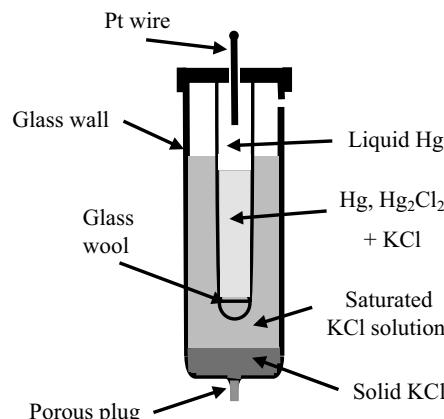


Figure 5.23 Schematic of the saturated-calomel reference electrode.

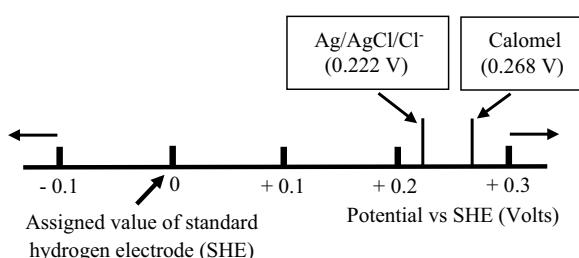


Figure 5.24 A scale indicating the relative potentials of the silver-silver chloride and the saturated-calomel electrode with respect to the standard hydrogen electrode.

Example 5.5

In Table 5.3 the standard reduction potential E° for the Zn^{2+}/Zn couple is given as -0.7618 Volts. What would be the value of E° if measured against a silver-silver chloride reference electrode?

Solution:

From Figure 5.24 (and Table 5.3) we find that the standard potential of the silver-silver chloride electrode with respect to the standard hydrogen electrode is $+0.2223$ Volts. The E° value for the Zn^{2+}/Zn couple vs. the Ag/AgCl electrode is thus equal to $[-0.7618 - (+0.2223)] = -0.9841$ Volts.

5.5.3 Liquid Junction Potentials

The standard voltage given by a reference electrode is only correct if there is no additional voltage supplied by a liquid junction potential formed at the porous plug between the filling solution and the external test solution. Liquid junction potentials can appear whenever two dissimilar electrolytes come into contact. At this junction, a potential difference will develop as a result of the tendency of the smaller and faster ions to move across the boundary more quickly than those of lower mobility. From Table 10.4 of Chapter 10 we can see that the chloride ion has a larger diffusion coefficient in water than the sodium ion, whereas potassium and chloride ions have roughly similar diffusion coefficients. Formation of a junction potential between two NaCl solutions of different concentrations is shown in Figure 5.25.

These potentials are difficult to reproduce, tend to be unstable, and are seldom known with any accuracy. Steps should therefore be taken to minimise them. Using 4 Molar KCl as the inner filling solution in the silver-silver chloride electrode, for example, has the advantage that the K^+ and Cl^- ions have nearly equal mobilities and form what is known as an equi-transferrant solution. Also, the electrolyte concentration is much higher than that of the test sample solution – thus ensuring that the major portion of the current is carried by these ions. A third factor in minimising the junction potential is the fact that there is a small but constant flow of electrolyte out from the reference electrode thus inhibiting any back-diffusion of sample ions – although this is less important with modern gel electrolytes.

Representative examples of liquid junction potentials are given in Table 5.5. Although of relatively small magnitude, liquid junction potentials have to be algebraically added to the E° factor in the Nernst Equation 5.3. Any variations in their values during analyses can be a major source of potential drift and error in measurements.

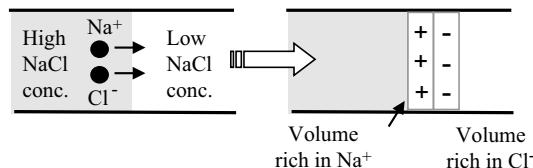


Figure 5.25 An example of a junction potential forming at the interface between two electrolytes, as a result of a charge separation due to the difference in diffusion mobility of sodium and chloride ions.

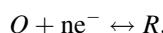
Table 5.5 Liquid junction potentials for some electrolyte interfaces. The polarity given is with respect to the side of the junction containing the solution on the left of the two cited solutions

Junction	Potential (mV)
0.1 M NaCl 0.1 M KCl	-6.4
0.1 M NaCl 3.5 M KCl	-0.2
1 M NaCl 3.5 M KCl	-1.9
0.1 M HCl 0.1 M KCl	+27.0
0.1 M HCl 3.5 M KCl	+3.1

5.6 Electrochemical Impedance Spectroscopy (EIS)

As described in Chapter 4, impedance spectroscopy can be used to analyse electrical processes occurring in a system. The technique is particularly sensitive to changes in both surface and bulk effects, and as such is a valuable technique to apply for electrochemical research and sensor applications.

A good example to consider is a simple reversible reaction at a working electrode of the form



where both the oxidised and reduced species, O and R , are soluble in the aqueous solution at the electrode. A complex impedance plot can be obtained by applying an AC voltage to the electrochemical cell. The resulting current-potential response is given by Equation (5.9). Expansion of the exponential functions in this equation into their series forms gives

$$I = I_o \left[1 + \frac{\alpha_A nF(E - E^o)}{RT} + \frac{1}{2!} \left(\frac{\alpha_A nF(E - E^o)}{RT} \right)^2 + \dots \right] - I_o \left[1 - \frac{\alpha_C nF(E - E^o)}{RT} + \frac{1}{2!} \left(\frac{\alpha_C nF(E - E^o)}{RT} \right)^2 - \dots \right].$$

For small perturbations of the working electrode's potential ($|E - E^o| \ll [RT/\alpha nF]$) the quadratic and higher terms can be ignored, and for $\alpha_C = \alpha_A = 0.5$, we have

$$I = I_o \frac{nF}{RT} (E - E^o) \quad (5.11)$$

with

$$I_o = I_O = nF[\text{O}]_s k_O,$$

or

$$I_o = I_R = -nF[\text{R}]_s k_R,$$

depending on whether the potential is perturbed to a value above or below E_o , respectively. Thus, if the potential of the working electrode is perturbed by just a few millivolts (≤ 5 mV) about the value of the equilibrium potential E° , we can assume that the current-potential response is approximately linear. The ratio $(E - E^\circ)/I$ has dimensions of resistance, and the concept of a *charge-transfer resistance* R_{ct} can be introduced and defined as

$$R_{ct} = \frac{RT}{nF I_o}. \quad (5.12)$$

This resistance component is small when the exchange current I_o is large. As shown in Figure 5.16, apart from the impedance to current flow related to charge transfer at the electrode-solution interface, there is also effective impedance related to diffusion-controlled mass transfer. For example, the situation where O is reduced to R can be considered as three steps:

- (i) mass transport (by diffusion) of O from the bulk electrolyte to the electrode surface;
- (ii) charge transfer reaction (kinetic control) that converts O to its reduced form R ;
- (iii) mass transport of R from the electrode surface into the bulk electrolyte.

The overall impedances related to these processes can be represented as a linear combination of a resistor R_s and capacitor C_s . The charge transfer resistance R_{ct} can be separated from the mass transport processes to give [1, ch. 9] the angular frequency ($\omega = 2\pi f$) dependencies for R_s and C_s as:

$$R_s = R_{ct} + \frac{\sigma}{\omega^{1/2}}; \quad C_s = \frac{1}{\sigma \omega^{1/2}}.$$

For a planar diffusion front over an electrode of surface area A , σ is given by:

$$\sigma = \frac{RT}{n^2 F^2 A \sqrt{2}} \left[\frac{1}{D_{OX}^{1/2} [O]_b} + \frac{1}{D_R^{1/2} [R]_b} \right],$$

where D_O and D_R are the diffusion coefficients for the electroactive species of bulk fluid concentration $[O]_b$ and $[R]_b$, respectively. The impedance Z of the linear combination of R_s and C_s is

$$Z = R_s + \frac{1}{j\omega C_s} = R_{ct} + \frac{\sigma}{\omega^{1/2}} + \frac{\sigma}{i\omega^{1/2}},$$

which takes the form of a conventional resistance element R_s in series with a frequency-dependent element known as the Warburg impedance Z_w :

$$Z_w = \text{Re}(Z_w) + \text{Im}(Z_w) = \frac{\sigma}{\omega^{1/2}} - j \frac{\sigma}{\omega^{1/2}}.$$

The Warburg impedance possesses equal real and imaginary components, and so takes the form of a constant phase element, characterised by a constant phase angle of 45° . The magnitude of Z_w is:

$$|Z_w| = \sqrt{\left(\frac{\sigma}{\omega^{1/2}}\right)^2 + \left(\frac{\sigma}{\omega^{1/2}}\right)^2} = \sigma\sqrt{\frac{2}{\omega}}. \quad (5.13)$$

The relative values of Z_w and R_{ct} provide an indication of the balance between mass transport control and charge transfer kinetics of an electrode reaction. Unless the charge transfer process exhibits very slow kinetics, at high frequencies the Warburg impedance will be negligible compared with the charge transfer resistance, whereas at very low frequencies it will have the dominating influence on an electrode reaction.

The total impedance of an electrochemical cell will include contributions from the counter electrode as well as the working electrode. In order to be able to focus on the reaction occurring at the working electrode, the impedance of the counter electrode is reduced to insignificance by making its surface area as large as possible. The impedance of the working electrode should also include the effective capacitance C_{dl} of the electrical double layer at its surface, which will appear as an element in parallel with the charge transfer resistance and Warburg impedance. The bulk electrolyte between the two electrodes will appear as a series resistance R_b , and this should also be considered. The geometrical capacitance shown in Figure 4.2 of Chapter 4 will be negligibly small, and the reactance it presents will be short-circuited by the bulk electrolyte resistance. The overall equivalent circuit for a simple reversible reaction at a working electrode system can therefore take the form shown in Figure 5.26.

At low frequencies, the real and imaginary impedance components of this equivalent circuit approach the limiting values:

$$\text{Re}(Z) = R_b + R_{ct} + \frac{\sigma}{\omega^{1/2}}$$

$$-\text{Im}(Z) = \frac{\sigma}{\omega^{1/2}} + 2\sigma^2 C_{dl}.$$

Eliminating the factor $\sigma/\omega^{1/2}$ gives the relationship:

$$-\text{Im}(Z) = \text{Re}(Z) - (R_b + R_{ct} - 2\sigma^2 C_{dl}) \quad (5.14)$$

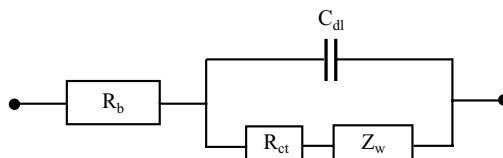


Figure 5.26 Equivalent circuit that includes a charge transfer resistance R_{ct} that controls the kinetics of a simple reversible electrode reaction, together with the Warburg impedance Z_w that controls the mass transport. The resistance R_b of the bulk electrolyte and the capacitance C_{dl} of the electrical double layer at the electrode are also included.

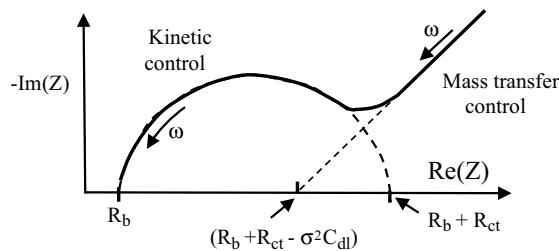


Figure 5.27 Complex impedance plot for the equivalent circuit of an electrode reaction shown in Figure 5.26. Mass transfer control operates at low frequencies, and kinetic control occurs at high frequencies.

A plot of $-\text{Im}(Z)$ versus $\text{Re}(Z)$ should thus take the form of a straight line of unit slope, intercepting the $\text{Re}(Z)$ axis at $(R_b + R_{ct} - \sigma^2 C_{dl})$. At very high frequencies, where the Warburg impedance becomes negligible, the real and imaginary impedance components of the equivalent circuit shown in Figure 5.25 approach the limiting values:

$$\begin{aligned}\text{Re}(Z) &= R_b + \frac{R_{ct}}{1 + \omega^2 C_{dl}^2 R_{ct}^2} \\ -\text{Im}(Z) &= \frac{\omega C_{dl} R_{ct}^2}{1 + \omega^2 C_{dl}^2 R_{ct}^2}.\end{aligned}$$

Eliminating the frequency ω from these two equations, and proceeding in the same way as described in Chapter 4 (Equations 4.13–4.17) gives the result:

$$\left(\text{Re}(Z) - R_b - \frac{R_{ct}}{2} \right)^2 + (-\text{Im}(Z))^2 = \left(\frac{R_{ct}}{2} \right)^2 \quad (5.15)$$

In the high-frequency range, a plot of $-\text{Im}(Z)$ versus $\text{Re}(Z)$ should thus take the form of a semicircle of radius $R_{ct}/2$, centred on the $\text{Re}(Z)$ axis at $(R_b + R_{ct}/2)$. A complex impedance plane plot that incorporates equations (5.14) and (5.15) to describe the frequency variation of the impedance of the equivalent circuit of Figure 5.26 is shown in Figure 5.26.

In Figure 5.27 the frequency range of measurement of $\text{Re}(Z)$ and $\text{Im}(Z)$ would typically extend from 10^{-3} Hz to 1 MHz, to show the kinetically controlled region of the electrode reaction (the semicircle) and the mass transport (diffusion) controlled region represented by the straight line of unit slope. It should be noted, however, that not all electrode reactions involve simple, homogeneous and reversible, reactions of the form $\text{O} + n\text{e}^- \leftrightarrow \text{R}$. For example, the chemical species formed by electron transfer at the electrode surface may not be stable in the bulk electrolyte. It may simply be an intermediate which undergoes another chemical change to form a final product. A heterogeneous reaction might also occur, whilst the oxidised or reduced species is absorbed on the electrode, involving further chemical reactions such as the formation of dimers that might involve further charge transfer, for example. The equivalent circuit shown in Figure 5.26 would not be sufficient to describe such deviations from a homogeneous and reversible reaction. Extra components to accommodate more than one charge transfer resistance or mass transfer impedance would have to be added. The corresponding impedance plots would be more complicated than the ideal form depicted in Figure 5.27.

Problems

- 5.1. Some sensors operate by monitoring changes of an electrochemical reaction at an electrode surface. An important relationship between the electrode potential E and the equilibrium concentrations of the redox couple involved in the electrochemical reaction is provided by the Nernst Equation:

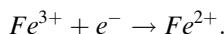
$$E = E^o + 2.303 \frac{RT}{nF} \log \frac{[C_{ox}]}{[C_{red}]}.$$

- (a) Define the symbols E^o , $[C_{ox}]$ and $[C_{red}]$ in this equation.
 (b) At 25°C the factor $2.303 \frac{RT}{nF}$ has a numerical value of 0.0592 (for $n = 1$).

What are the units of this factor?

What is the significance in designating $n = 1$?

- 5.2. Identify which of the following equations correctly interpret the Nernst Equation for the reaction:



(a) $E = E^o + 0.059 \log_{10} \left(\frac{[Fe^{2+}]}{[Fe^{3+}]} \right)$

(b) $E = E^o - 0.059 \log_{10} \left(\frac{[Fe^{3+}]}{[Fe^{2+}]} \right)$

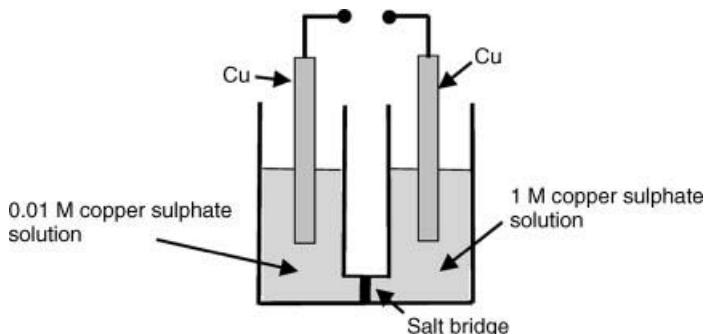
(c) $E = E^o + 0.059 \log_{10} \left(\frac{[Fe^{3+}]}{[Fe^{2+}]} \right)$

(d) $E = E^o - 0.059 \log_{10} \left(\frac{[Fe^{2+}]}{[Fe^{3+}]} \right).$

- 5.3. Write the Nernst Equation to describe the potentials and concentrations that pertain to the following reaction:



- 5.4. (a) Calculate the magnitude and polarity of the potential appearing between the two copper electrodes (at 25°C) of the electrochemical cell shown in the figure below.



- (b) Identify the electrode where copper ions will deposit on the electrode, if the electrodes are connected with a metal wire.

5.5. A reference electrode is often incorporated into an electrochemical sensor. What are the characteristics required of a reference electrode, and what function does it perform?

5.6. An electrochemical process at an electrode can be characterised by a charge transfer resistance of value given by the following expression:

$$R_{ct} = \frac{RT}{nF I_o}$$

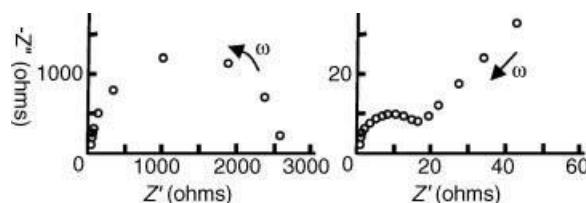
- (a) Derive the units for R_{ct} .

(b) Would a high value of R_{ct} correspond to a low or a high charge transfer rate at the electrode surface? Explain your reasoning.

(c) How can R_{ct} be determined experimentally?

(d) The value for I_o is determined to be 0.2 mA at 20 °C for a one-electron transfer process. Calculate the corresponding value for R_{ct} .

5.7. The complex impedance plane plots obtained for two different electrode reactions are shown below, for the frequency range 10^{-4} – 10^4 Hz.



- (a) Estimate the value of the charge-transfer resistance for both electrode reactions.

(b) What can you deduce in terms of the different processes that control the impedance to current flow of these two electrode reactions?

References

- [1] Bard, A.J. and Faulkner, L.F. (2001) *Electrochemical Methods: Fundamentals and Applications*, 2nd edn, John Wiley & Sons, New York.
- [2] Tafel, Julius (1905) Über die Polarisation bei Kathodischer Wasserstoffentwicklung. *Zeitschrift Fur Physikalische Chemie-Frankfurt*, **50**, 641–712.
- [3] Bates, R.G. and MacAskill, J.B. (1978) Standard potential of the silver-silver chloride electrode. *Pure and Applied Chemistry*, **50**, 1701–1706.

Further Readings

- Janz, G.J. and Ives, D.J.G. (1968) Silver, silver chloride electrodes. *Annals of the New York Academy of Sciences*, **148** (1), 210–221.
- Thomas, F.G. and Henze, G. (2001) *Introduction to Voltammetric Analysis*, CSIRO Publishing, Collingwood, Australia.
- Wang, J. (2006) *Analytical Electrochemistry*, 3rd edn, John Wiley & Sons, Hoboken, New Jersey.

6

Biosensors

6.1 Chapter Overview

Biosensors are analytical devices for detecting chemical analytes. They consist of a sensitive biological element, which reacts selectively with the target analyte, and a transducer with its associated electronic signal processing and output display. The transducer or detector element often depend on the spectroscopic and electrochemical techniques described in Chapters 4 and 5. In this chapter attention will be focused on the biologically sensitive element of a biosensor and the physicochemical transduction processes. The electronic measurements and instrumentation aspects are covered in Chapters 7 and 8.

After reading this chapter the reader will gain a basic understanding of:

- (i) the characteristics required of a biosensor and the main considerations to be taken into account in its design;
- (ii) the physical and chemical methods used to immobilise the biosensing agent;
- (iii) the format and important operating parameters of a biosensor (e.g. transfer function, precision, accuracy, detection limit, effects of pH and temperature);
- (iv) amperometric biosensors;
- (v) potentiometric biosensors and ion-selective electrodes;
- (vi) conductometric biosensors;
- (vii) potentiometric and impedimetric biosensors;
- (viii) immunosensors;
- (ix) photometric biosensors;
- (x) biomimetic sensors;
- (xi) glucose sensors and their continuing development;
- (xii) controlling and testing the biocompatibility of implantable sensors.

6.2 Introduction

Chemical sensors are defined as measurement devices which utilise chemical or biological reactions to detect and quantify a specific analyte or reaction event. The terms *sensor*, *transducer*, *detector*, are often used to mean the same thing. They are devices that convert one form of energy into another and produce a usable energy output in response to a specific

Table 6.1 Required characteristics of a biosensor
(often to be defined by the user)

• High selectivity	• Good resolution
• Suitable sensitivity	• Utility
• Good dynamic range	• Field portability
• Attractive price	• Safe to use
• Low running costs	• Ruggedness
• Simplicity of use	• Reproducibility
• Reliability	• Ease of calibration
• Speed of response	• Stability
• Accuracy	• Precision

measurable input. For chemical sensors, a transducer plus a chemically active surface is termed a sensor. To distinguish biosensors from chemical sensors we define a biosensor as one which uses as its active detection component a biomolecule (enzyme, antibody, nucleic acid or cell membrane receptor), organelles, microorganisms, biological tissue, or a biomimetic polymer. Broadly speaking, biosensors can be categorised as falling either into a catalytic or affinity mode of operation. Catalytic biosensors employ an active biocomponent, usually an immobilised enzyme or catalytically active polynucleotides (DNAzymes) that reacts specifically with the analyte, but also cell membrane patches or whole-cells as the transducing element. Such sensors can also be used to detect toxic chemicals or as drug discovery tools by detecting the rate of inhibition of a biocatalytic reaction. Affinity biosensors exploit the specific interaction of a ligand (the analyte) and a biological receptor, and include immunosensors that exploit antibody-antigen interactions together with nucleic acid biosensors that involve the complementary hybridising of oligonucleotides.

A key advantage of a biosensor is its high selectivity for detection of the target analyte in a complex test sample. Purification and other processing of the sample before its analysis are thus reduced to a minimum. Applications of biosensors include: medical care in clinics and hospitals; diagnostic tests in a laboratory or doctor's surgery; determination of food quality; detection of environmental pollutants; industrial process control; law enforcement and military defense systems.

The characteristics required of a biosensor are summarised in Table 6.1.

The main considerations to be taken into account when designing a biosensor are:

- the molecule (analyte) to be detected and how it is to be delivered to the sensor;
- the choice of biological sensing agent (see Figure 6.1) to ensure specific detection of the target analyte;

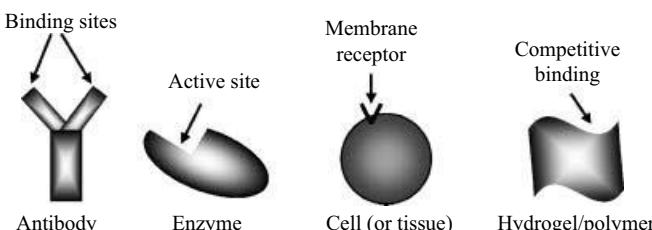


Figure 6.1 The bioactive sensing element in a biosensor can be of various forms. In all cases the desired result is a highly specific detection of the target analyte.

Table 6.2 Classes of biosensor operation principles

Electrochemical:	
• Potentiometric	Optical
• Amperometric	• Spectroscopy
• Impedimetric	• Fluorescence
Conductometric	• Polarimetry
Immunosensing	Thermal
• Antibody-antigen	• Thermistor
	Piezoelectric
	• Quartz crystal

- the most suitable method to immobilise the sensing agent;
- the dynamic range of analyte measurement;
- the detection method (e.g. optical, electrochemical, conductometric);
- the mode of electronic readout;
- ways to avoid false signals (e.g. from interfering sources) and to improve sensitivity;
- whether for single use (disposable) or multiple use;
- sensor fabrication.

The main classes of biosensor, in terms of their operating principles, are listed in Table 6.2. These various types of biosensor will be described in this chapter, together with examples of their application.

6.3 Immobilisation of the Biosensing Agent

A critical step in creating a biosensor is the preparation and preservation of the biological detection agent. For the case of enzymes they need to be extracted from their normal biological matrix and purified. In some cases isolated and purified enzymes are chemically unstable or have impaired activity through loss, for example, of their required cofactors. In other cases it can be prohibitively expensive to isolate the required enzyme in a purified form. Considerable efforts have therefore been directed towards the use of microbial or tissue-based sensors. The methods that can be used to entrap or immobilise the various types of biosensing agent can roughly be divided into physical and chemical methods.

6.3.1 Physical Methods

Adsorption and entrapment are the main physical procedures that can be used. Materials such as activated carbon, alumina, graphite powder, ion-exchange resins, polystyrene and silica gel can be used as inert matrices in which to immobilise enzymes, and microorganisms can be absorbed onto membrane supports made from alginate gel, carbon paste, paper and polypropylene, for example. The immobilised biodetection agents can then be confined to the working electrode by means of a nylon or polyamide net, or very commonly by means of a cellophane dialysis membrane. Enzymes and extracts from plant tissue can also be mixed with carbon paste to form thick-film electrodes that can be screen-printed onto a polyester material.

The electrochemical deposition of platinum onto a platinum electrode produces a surface layer, known as platinum black consisting of porous microplatinum particulates, that readily absorbs protein material. Enzymes, and even single cells, can be entrapped in the pores of polymer films made from materials such as polyacrylamide or polyester sulfonate.

6.3.2 Chemical Methods

Direct chemical crosslinking or covalent bond attachment can be used to link an enzyme to a high molecular weight passive protein. Glutaraldehyde has commonly been used as a cross-linking reagent of enzymes to bovine serum albumin, collagen, egg albumin and gelatin, for example. The two aldehyde (COH) groups of the glutaraldehyde molecule can react with the amino groups of proteins to form a thick gel composed of protein molecules linked together via strong imine ($C=N$) bonds. This gel solidifies into a solid when the solvent evaporates. Covalent bonding of an enzyme to a solid support can also occur if the solid possesses exposed carbonyl groups that can form imine bonds to amino groups of the enzyme.

Direct immobilisation of an enzyme in a matrix that permits direct electronic interaction with an electrode can be achieved by electrochemical polymerisation. A solution containing the enzyme and a monomer such as aniline or pyrrole is subjected to electrochemical oxidation using potentiometry or cyclic voltammetry, resulting in the enzyme being immobilised in an electronically conducting polymer, such as polyaniline or poly-pyrrole (Uchiyama *et al.*, [1]). Enzymes that have been crosslinked with glutaraldehyde can also be immobilised in electrochemically polymerised nonconducting polymers [2].

A technique commonly called the *avidin-biotin system* takes advantage of the high affinity constant ($\sim 10^{-15} \text{ mol}^{-1} \text{ L}$) between the two proteins avidin and biotin. Avidin is first coated onto the sensing surface, the biomolecule to be immobilised is then attached to a biotin molecule and allowed to interact with the avidin layer. Although the affinity between avidin and biotin is very high, covalent bonding is not involved and so multiple washing and reuse of the same sensing surface can be used. Robust self-assembled monolayers (SAMs) can be formed by immersing a gold substrate into a high purity solvent containing a surfactant, such as ethanol containing alkanethiols possessing free thiol (SH) groups. These groups are then used to link the sensing biomolecule to the monolayer.

6.4 Biosensor Parameters

6.4.1 Format

A generic form of a biosensor is shown in Figure 6.2. An immobilised bioactive element is exposed to a sample to be tested for its content of a specific analyte. Chemical or physical changes that occur when the bioactive element interacts with the analyte are transduced into an optical or electrical signal, whose amplitude depends on the analyte concentration. The transducer can take on many forms, such as acoustic, amperometric, colourimetric, conductometric, impedimetric, optic or potentiometric.

Examples of biosensor designs are shown in Figures 6.3 and 6.4. In the optic design of Figure 6.3a, a solution containing the analyte is admitted into a chamber whose inner surface is coated with a protein having receptor sites that bind specifically to the analyte. A candidate protein for this purpose would be concanavalin A (Con A), which has receptor sites that

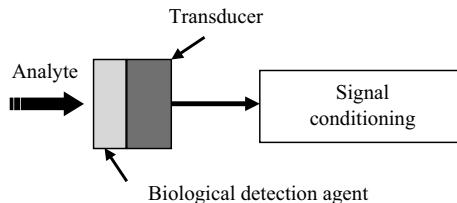


Figure 6.2 A biosensor consists of an immobilised bioactive material that interacts with the target analyte. This interaction is transduced into an output signal, most usually optical or electrical.

bind specifically to various sugars, glycoproteins and glycolipids. If the analyte to be detected is glucose, fluorescent molecules that specifically bind to glucose are added to the test sample and either injected directly, or diffused through a porous membrane, into the chamber. An optical fibre interfaced with the chamber can be used to irradiate the sample so as to excite the fluorescent molecules, and to then collect the emitted fluorescence. The emitted fluorescence can be detected as the current generated by a photo-detector connected to the other end of the optical fibre. Thus, as a general principle, competitive binding of the tagged analyte to the protein coated on the chamber surface causes a reduction of the bulk fluorescence. The resulting reduction of the measured photocurrent is used to determine the initial concentration of the analyte. The sensitivity of detection can be improved by subtracting or comparing the output of a similar chamber coated with a protein that does not bind to the analyte.

In the sensor layout shown in Figure 6.3b, an incident light beam excites surface plasmon resonance and an associated evanescent wave at the surface of a metal film. The critical angle of incidence to achieve this effect will change if the refractive index of the adjacent medium is altered as a result of the binding of analyte molecules to molecular receptors coated on the metal surface. This analyte capture can be monitored as a change of the reflected light intensity.

The amperometric sensor design shown in Figure 6.4a monitors the rate at which an immobilised enzyme selectively oxidises the analyte, corresponding to the enzyme's natural substrate. The basic design of a conductometric biosensor is shown in Figure 6.4b, and operates on the principle that many enzymatic reactions involve either the consumption or

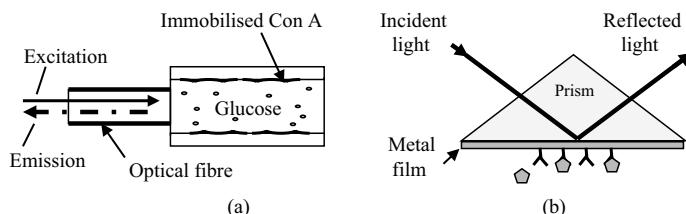


Figure 6.3 Examples of optometric biosensor designs: (a) A glucose sensor that operates by detecting the reduction of fluorescence emitted by tagged glucose molecules as they bind to a protein (ConA) coating. (b) The binding of analyte molecules to immobilised receptor molecules is detected as a change in either the critical angle of incidence, or reflected intensity, of a light beam that induces surface plasmon resonance in a metal film.

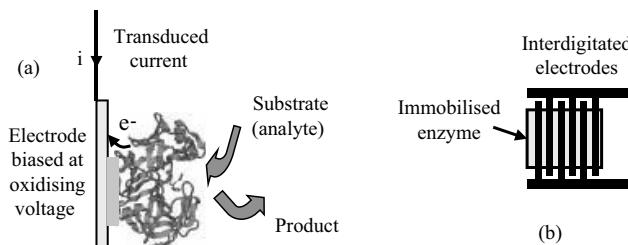


Figure 6.4 Examples of electrical biosensor designs: (a) An amperometric sensor that operates by detecting the change in the redox state of an immobilised enzyme when it reacts with its natural substrate (the analyte). (b) A conductometric sensor that detects changes in the concentration of ionic species when the analyte reacts with an immobilised enzyme coated onto a microelectrode array.

production of electrically charged chemical species. Thus, a change in the conductivity of the test sample can be expected if the analyte reacts with an immobilised enzyme or cell membrane that contains specific receptors. This change in conductivity can be detected by immobilising a biosensing matrix onto an interdigitated array of gold or platinum electrodes, fabricated by photolithography or vacuum spraying onto a thin plastic film, glass or ceramic substrate. A small-amplitude AC voltage (e.g. 25 mV peak-to-peak with no dc bias voltage) is commonly applied to the electrodes at a frequency of 10 kHz~100 kHz. A reference signal can be incorporated into the design by coating a second electrode with only the immobilising matrix (e.g. a crosslinking albumin-glutaraldehyde gel). The differential output signal between the biologically active and a reference electrode array can then be monitored.

6.4.2 Transfer Function

A sensor detects a chemical input S_{in} and transduces or converts this to a more useful form S_{out} , that is a function of S_{in} , such that:

$$S_{\text{out}} = f(S_{\text{in}})$$

The function $f(x)$ is the transfer function of the sensor, an expression of the relationship between the input and the output. This could take a number of different forms, including an equation, a graph of the response or a qualified calibration curve. It is usually preferable for this to be a linear function, or linear to a simple mathematical function (e.g. logarithmic), of S_{in} .

6.4.3 Sensitivity

The sensitivity of the sensor can be defined as the slope of the output characteristic or transfer function curve ($\delta S_{\text{out}}/\delta S_{\text{in}}$) over a specified linear range. The sensitivity of a biosensor can be determined by calibrating its response against reference samples of known analyte concentration. For the calibration example shown in Figure 6.5 we can state for specified operating conditions that the sensor exhibits a sensitivity of 24 nA/mM over a linear range up to 8.0 mM.

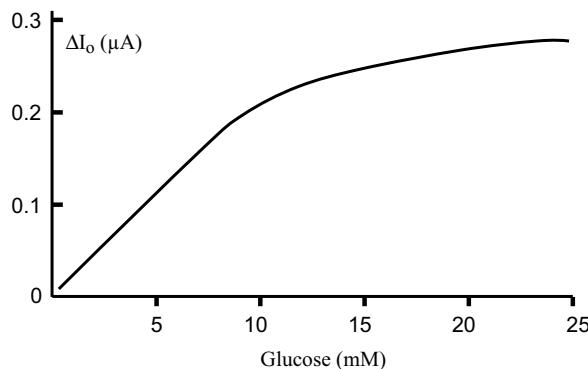


Figure 6.5 Calibration curve for a conductometric glucose sensor.

6.4.4 Selectivity

Many sensors will be sensitive to parameters other than the one being measured, such as the ambient temperature, pH or interfering chemicals. The selectivity is defined as the sensitivity to the desired parameter divided by the sensitivity to the interfering parameter. A good biosensor should therefore be as sensitive as possible to the target analyte, and insensitive to any other physico-chemical inputs it is likely to encounter.

6.4.5 Noise

Noise is a random fluctuation of the output of a sensor which is essentially unrelated to the input parameter. Common sources of noise include temperature fluctuations, electromagnetic interference, instability of the bioactive sensitive element or of the electronic circuitry, mechanical vibrations and fluid flow artifacts such as bubbles. Noise can be quantified as the root mean square x_{rms} of a sample (x_1, x_2, \dots, x_n) of output signals for a given time period:

$$x_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Random errors arising from noise can be reduced by signal processing, such as filtering, but this is often accompanied by a reduction of the dynamic response of the sensor. The sensitivity can be improved by using a reference sensor $R2$, coated with an inert material that does not detect the analyte, but is exposed to the same random disturbances as the active sensor $R1$. The random disturbances are cancelled out if the output signal is taken as either the difference ($R1-R2$) or ratio ($R1/R2$) of the two sensor readings (see Figure 6.6).

In addition to the sensor noise, the instrumentation system will also introduce different forms of noise and these will often dominate. The distribution of the noise will vary across the frequency spectrum, for example white noise is constant with frequency (across some defined bandwidth) while pink noise is proportional to $1/f$. The Signal to Noise Ratio (SNR) of a sensor output is an important parameter that will often determine the sensitivity or minimum resolution, as well as the detection limits of a sensor system.

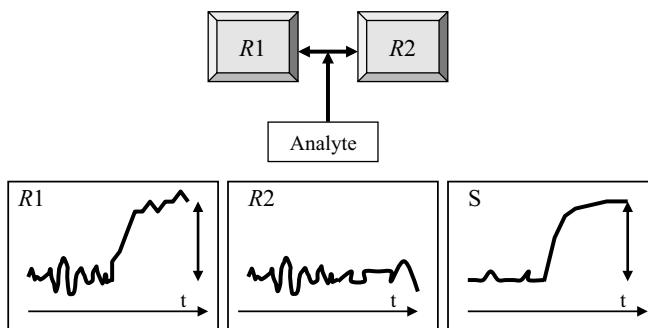


Figure 6.6 Sensitivity can be improved using a reference sensor R_2 , with the bioactive sensor R_1 , and taking the output signal S as either the difference or ratio of the R_1 and R_2 outputs.

6.4.6 Drift

A slow degradation or physico-chemical change of the bioactive sensing agent can result in a baseline drift of the output signal (see Figure 6.6). The baseline should be stable when there is no interaction between the bioactive sensing agent and the analyte. This baseline (rms value) drift can be quantified in terms of response units per time interval, as indicated in Figure 6.7.

The long-term stability of a sensor, as well as short-term baseline drifts, should also be investigated. This can be achieved by comparing the sensor output, for a fixed analyte concentration, over a period of one or more months. Long-term drift can indicate that the sensor is slowly degrading due to adverse effects from its environment and this may be a particular problem for implanted medical sensors.

6.4.7 Precision and Accuracy

The accuracy of a measurement defines how close the measured value is to the true value, while the precision describes how the measured value changes with repeated

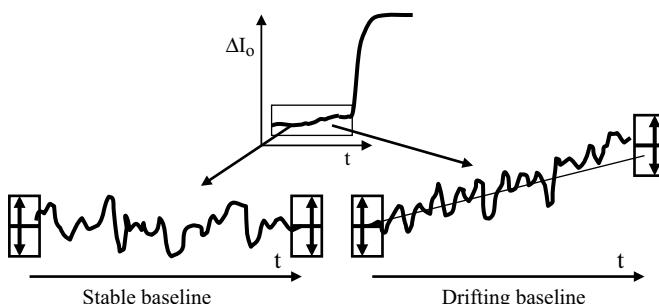


Figure 6.7 A stable baseline output signal with no analyte present is desirable, but can be accommodated by defining the baseline drift if present.

measurements. With a large dataset the accuracy can be thought of as the offset between the mean measured value and the true value of the parameter, while the precision is the variability of the data. Measurements can be precise but inaccurate, accurate but imprecise, both or neither. Obviously we would prefer it if the sensor output is both accurate and precise but it is generally better to be approximately right than precisely wrong. In any case, no statement of a measured value is complete without knowledge of the errors involved. A statistical measure of the precision of a set of measurements can be obtained by determining the standard deviation σ :

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} s$$

where N is the number of measured data values, and \bar{x} is the mean value of all the individual measurements x_i . A low value for the standard deviation indicates that the measured values tend to lie in a narrow range about the mean or average value, whereas a high standard deviation implies that the measured data values are spread over a wide range and thus lacks precision. Another way of viewing this is to define *imprecision* as the standard error of a measured value. A better indicator of precision is given by the coefficient of variation (also known as the relative standard deviation) defined as the ratio of the standard deviation to the mean, and which is often quoted as a percentage. Examples of such statistical procedures and results are given in Table 6.3 for the glucose sensor.

The relative standard deviation value of 0.02 derived in Table 6.3 informs us that glucose concentrations can be determined to within a precision of $\pm 2\%$, but this does not inform us to how close a glucose reading will be to the ‘true’ concentration. The accuracy can be determined by correlating the sensor readings with measurements obtained using an accepted, very accurate, method. Infrared spectroscopy is one well tested method. The example shown in Figure 6.8 gives the correlation and regression line for 50 measurements on various test samples using the glucose sensor (characterised by Figure 6.5 and Table 6.3) and infrared spectroscopy. The correlation coefficient r , also known as the cross-correlation coefficient, is a number between 0 and 1 that gives the quality of the least squares fitting of the two sets of experimental data. If there is complete correlation, then $r = 1.0$. Values of $r \geq 0.9$ are considered to represent strong correlation, and so the correlation coefficient of 0.996 given in Figure 6.8 indicates that the glucose sensor provides readings of high accuracy (with a precision of $\pm 2\%$).

Table 6.3 Statistical analyses of readings taken by the glucose sensor for a test sample containing $\sim 10 \text{ mM}$ glucose

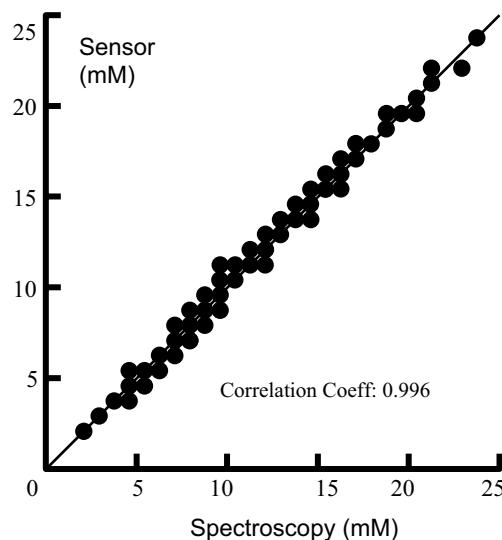


Figure 6.8 The accuracy of a sensor can be determined by correlation with a standard reference measurement. This example shows the correlation and regression line for 50 measurements of the glucose content of a clinical samples using standard spectrophotometry and the test glucose sensor.

6.4.8 Detection Limit and Decision Limit

Although a calibration plot of the form shown in Figure 6.5 suggests that the biosensor has an analytical range that extends from almost zero analyte concentration to an upper limit of 25 mM, it is not possible to make accurate measurements down to analyte concentrations approaching that of a so-called *blank* sample that contains no analyte. Sufficient analyte must be present in a test sample to produce an analytical signal that can be distinguished from *analytical noise* – the term used to describe the signal produced in the absence of the analyte. This has led to the concept of *Detection Limit*, also known as the *Limit of Detection*. The IUPAC *Compendium of Chemical Terminology* (2nd edn, 1997) defines *Detection Limit* (in analysis) as:

‘The minimum single result which, with a stated probability, can be distinguished from a suitable blank value’. Furthermore: ‘The limit defines the point at which the analysis becomes possible and this may be different from the lower limit of the determinable analytical range.’

A commonly accepted formula for calculating the Limit of Detection (LoD) is:

$$\text{LoD} = \text{Blank Value} + k\sigma_{\text{blank}} \quad (6.1)$$

where the *Blank Value* is the mean result obtained for a set of repeated measurements on blank samples, σ_{blank} is the standard deviation of the blank results, and k is a numerical factor chosen according to the required level of confidence. (The IUPAC recommend values for k of either 2 or 3, but provide no statistical guidance.) As depicted in Figure 6.9, the level of confidence can be interpreted as the probability that a blank sample will be falsely

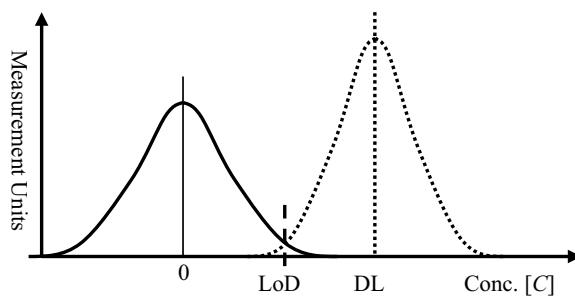


Figure 6.9 The solid curve shows the cumulative standard normal distribution $\Phi(z)$ of errors obtained for blank samples. The limit of detection (LoD) is shown for $z = 1.645$, corresponding to $\Phi(z) = 0.95$ ($\Phi(z) = 0.5$ at $z = 0$). The dotted curve is the cumulative distribution of errors obtained for an analyte concentration just above the LoD. Setting the sensor decision limit (DL) as shown will give the probability of a false negative reading as $\sim 5\%$.

determined to contain an analyte (i.e. a false positive). Commonly accepted values for this probability are 5 and 1%, respectively, which for a normal (Gaussian) distribution of results corresponds to values for k of 1.645 and 2.326, respectively. The IUPAC recommended values for k of either 2 or 3, respectively, correspond to probabilities for a false positive result from a blank sample as 2.3 or 0.14%, respectively.

The conversion from the units of measurement (e.g. microamps, millivolts, optical absorbance) to concentration C of the test analyte is made through the sensitivity determined from the slope S of the calibration curve at low analyte concentrations. From Equation (6.1) we can define the *Limit of Concentration* (LoC) as:

$$\text{LoC} = \text{Blank Conc} + kS\sigma_{\text{blank}}$$

Because the analyte concentration of the blank is zero, this is simplified to:

$$\text{LoC} = kS\sigma_{\text{blank}}$$

From Figure 6.9 we can appreciate that the LoC should not to be interpreted as the smallest concentration that can be measured, but rather the concentration at which a decision can be made as to whether or not an analyte is present in a test sample. It corresponds to where the signal can, to a specified level of confidence, be distinguished from the background noise. If we assume a symmetrical cumulative normal distribution for the analytical errors in a measurement result, then for an analyte concentration equal to the limit of detection there will be a 50% probability of obtaining a false negative result. In terms of the operating procedure for a sensor, a distinction should thus also be made between the limit of detection and what can be termed as the *Decision Limit* for interpreting measurement results. A detection decision can be set to reduce the probability of a false negative to an acceptable level of 5 or 1%, for example. This is demonstrated in Figure 6.9, to show the cumulative distribution of analytical error for an analyte concentration a little above the limit of detection calculated using Equation (6.1).

Example 6.1

The following data was obtained by running 20 calibrations of a biosensor using standard solutions of the designated analyte, together with a blank sample containing no analyte. Which of the standards lie below the limit of detection (calculated using Equation (6.1) with $k = 2.33$)? What would be a suitable value to set for the decision limit of the sensor?

Analyte conc. (mM)	Mean signal (nA)	σ (nA)
Blank (0)	5.2	1.5
5	7.0	1.4
10	9.7	1.7
15	12.4	1.5
20	15.1	1.4
25	17.8	1.3
50	29.5	1.2
75	36.4	1.3

Solution:

From Equation (6.1) with $k = 2.33$ and the results obtained for the blank sample, the $\text{LoD} = 5.2 + 3.5 = 8.7 \text{ nA}$. The 5 mM standard is below this level of detection. The mean result for the 15 mM standard is more than 2σ above the LoD and would be an appropriate level to set for the decision limit. Choosing the decision limit as 10 mM would give a significant chance of obtaining a false negative because there is an overlap of the analytical error distributions of the blank and 10 mM sample readings.

6.4.9 Dynamic Range

The range of a sensor is the stated maximum and minimum values of the target analyte that can be measured. This could be the linear part of the transfer function curve or be defined by the limit of detection at the lower end of the scale and some upper limit where the output saturates. A related parameter is the *linearity* of the sensor, which defines the level of deviation from a purely linear response over the specified sensor range.

For the calibration curve shown in Figure 6.5 the minimum and maximum values are 1 and 23 mM, respectively, with a nonlinear response above 7 mM. A distinction between the limit of detection and the decision limit can be important in environmental monitoring, food safety and drug tests, for example, where the incidence of false positive and false negative readouts should be kept to a minimum. However, when using a glucose or oxygen sensor in a clinical setting, the limits of detection are normally well below the expected ranges of physiological concentration. In such cases the application of a decision limit is not relevant.

6.4.10 Response Time

Sensors do not change output state immediately when an input parameter change occurs. Rather, it will change to the new state over a period of time called the response time. The

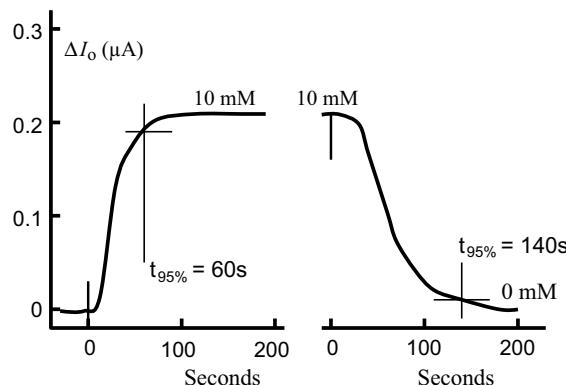


Figure 6.10 Response times typically found for biosensors of the designs shown in Figure 6.3a.

response time can be defined as the time required for a sensor output to change from its previous state to a final settled value within a tolerance band of the correct new value. Typical responses for the glucose sensor designs of Figure 6.3 are given in Figure 6.10. The response time, taken in this case to be the time required for the sensor output to attain 95% of the final steady reading, is 60 seconds.

Figure 6.10 also shows how the response settles back to the initial reading after the test sample has been replaced with a standard solution containing no glucose. The corresponding response time is 140 seconds.

6.4.11 Resolution

This can be thought of as the minimum fluctuation in the input parameter which produces a detectable change in the output. It will not only be related to the amplitude of the signal but also the timescale of the fluctuation, such that small but slow changes in the input may cause a measurable output fluctuation where small, fast changes do not. Therefore the resolution is dependent on the response time of the sensor.

6.4.12 Bandwidth

The bandwidth is the useful frequency range of a sensor. The high frequency (low pass) limit will be determined by how quickly the sensor responds to an instantaneous change in the input. Other sensors may also have a high-pass characteristic such that they only produce a useful output when the input changes. The output of these sensors will decay to some nominal value after a step change in the input and this will determine the low frequency cutoff.

6.4.13 Hysteresis

Ideally the output of a sensor will depend entirely on the input parameter and not on the previous history of the input. However in many cases this is not what happens and the output for a given input will vary depending on whether the input is increasing from a low level or decreasing from a higher level.

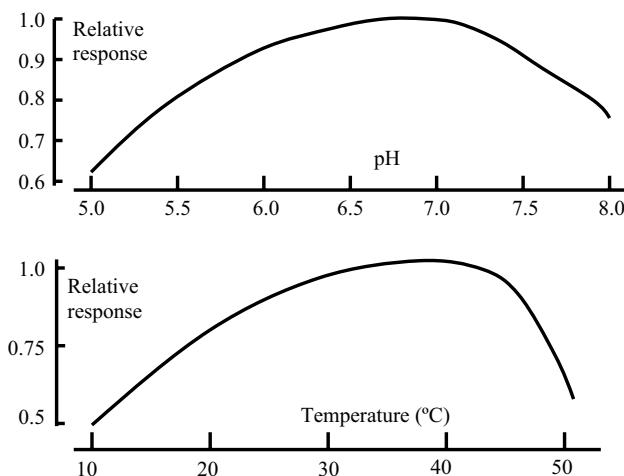


Figure 6.11 Typical responses expected for a biosensor as a function of pH and temperature.

6.4.14 Effects of pH and Temperature

From Chapter 1 (Section 1.4.1) we learnt that the binding of a substrate molecule to the active site of an enzyme generally involves electrostatic interactions, and that these interactions are sensitive to the local acidity of their environment and tuned to physiological pH. We can therefore expect the performance of a biosensor to be optimum at around pH 7. The rate of a biological reaction is also temperature dependent and for a typical activated process will double for a temperature rise of around 10 °C, until at a sufficiently high temperature the structure of the enzyme will degrade and its activity diminish. The typical responses of a biosensor to changes in pH and temperature are shown in Figure 6.11, and these responses must be taken into account when selecting the operational pH and temperature of a biosensor.

6.4.15 Testing of Anti-Interference

Biological samples will often contain substances that can interfere with the desired selectivity of a sensor. For example, various sugars, glycoproteins and glycolipids can bind strongly to the protein concanavalin A used in the design of Figure 6.3a, and electroactive chemicals such as ascorbic and uric acid could produce conductivity changes in using the design of Figure 6.4b. The testing of a prototype sensor should therefore include studies of possible interfering agents. An example of this is shown in Figure 6.12, where addition of ascorbic acid and uric acid to a test sample is found to have negligible effect on the readings from the conductometric glucose sensor.

6.5 Amperometric Biosensors

These are the most commonly reported class of biosensor, and take the form of an enzyme-coupled electrochemical sensor. Referring to Figure 6.2, an electrode serves as the signal

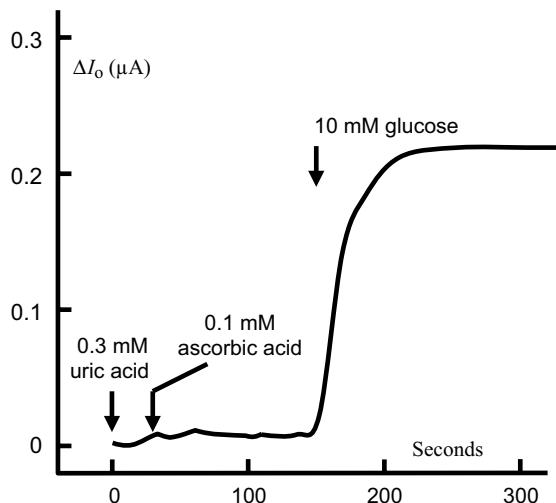
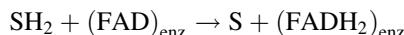


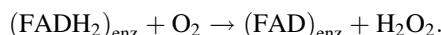
Figure 6.12 An anti-interference test to demonstrate that a conductometric glucose sensor is not influenced by the presence of uric acid or ascorbic acid at concentrations expected for a clinical sample.

transducer, where the measurable response is an electrical current arising from a redox reaction. The electrode is maintained at a specific potential with respect to a reference electrode, and the current produced is proportional to the concentration of the analyte that acts as the enzyme's substrate. Enzymes that have been used in amperometric biosensors are the oxidoreductases, also known as redox enzymes. They consist of two main classes, namely flavoenzymes and NAD(P)⁺/NAD(P)H-dependent dehydrogenases.

Flavoenzymes contain as their cofactor a nucleic acid derivative of riboflavin in the form of either flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN). These cofactors act as the redox active site and exhibit highly reversible electrochemistry. The FMN is noncovalently bound to the enzyme structure, but in some cases FAD may be covalently bound to an amino acid residue such as cystine, histidine or tyrosine. Each catalytic cycle consists of two distinct half-reactions – first the reduction of the oxidised form of the flavoenzyme by the reduced form of its substrate, followed next by the reoxidation of the flavoenzyme through the transference of electrons to an oxidised acceptor. In the following example:

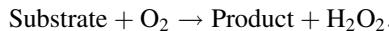


the ligand SH_2 is the reduced form of the substrate and $(\text{FAD})_{\text{enz}}$ is the oxidised form of the flavoenzyme. Oxygen can readily diffuse to the FAD group and act as an electron acceptor and be reduced to hydrogen peroxide in the following natural reaction:



The hydrogen peroxide produced can readily diffuse out and away from the reaction site in the enzyme. The overall reaction performed by a flavo-oxidoreductase, in the presence of

oxygen, is thus of the form:



The concentration of an analyte, acting as the substrate, can then be determined by amperometric detection of the oxygen consumed or of the hydrogen peroxide (H_2O_2) produced. Two of the most commonly used enzymes in clinical biosensors are glucose oxidase and cholesterol oxidase for the determination of the blood levels of glucose and cholesterol, respectively. During measurement the working electrode may act as an anode or a cathode. For example, a glucose sensor that uses glucose oxidase could measure the oxidation of the H_2O_2 produced by polarising the working electrode at +0.6V (vs. standard calomel electrode) to monitor the reaction:



Alternatively, the oxygen consumed by the enzymatic reaction could be measured by the fall in the oxygen reduction current by setting the working electrode to a negative potential of -0.92V (vs. Ag-AgCl) to monitor the reaction:



If an electrode is to be used as the transducing element in a flavoenzyme-based sensor, it will have to replace oxygen as the natural electron acceptor for the reduced (FADH_2)_{enz} form. This can be accomplished by performing the reaction in an oxygen deprived environment, which clearly precludes a clinical application. If kinetically efficient electron transfer between an immobilised redox enzyme and the working electrode can be achieved then, as depicted in Figure 6.13, the direct electrical coupling of a peroxidase enzyme (e.g. horse radish peroxidase) to an electrode can be used to detect hydrogen peroxide produced by a co-immobilised oxidase enzyme. The peroxidase catalyses the reaction shown in Equation (6.3). Direct electron communication between the peroxidase and an electrode can be facilitated by changing the surface morphology of the electrodes at the nanoscale, by coating the surface with metal nanoparticles or carbon nanotubes, for example.

A problem often encountered in this type of biosensor is that impurities or other chemical species in the sample material may also be electroactive at a potential close to that being applied to the working electrode. For example, ascorbic acid and uric acid are commonly

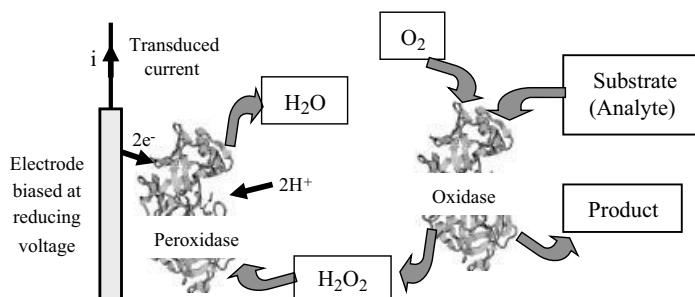


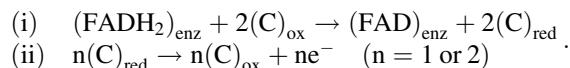
Figure 6.13 Direct, nonmediated, electron transfer between an electrode and two enzymes (an oxidase and a peroxidase).

found together in many biological samples and are oxidised at the same anodic potential of 0.35 volts (cf. Ag/AgCl). Uric acid is a chemical created when the body breaks down substances called purines (found in some foods and drinks, such as liver, anchovies, mackerel, dried beans, peas, beer, and wine). High levels of uric acid in the body can be dangerous, and this can be checked in a urine sample. However, ascorbic acid could also be present, and so cause a faulty analysis of the uric acid if an electrochemical analysis is made.

The other class of oxidoreductases commonly used in amperometric biosensors are NAD(P)⁺/NAD(P)H-dependent dehydrogenases. These enzymes catalyse the removal of hydrogen from a substrate in an oxidation-reduction reaction, followed by the transfer of the hydrogen to an acceptor molecule. An example is alcohol dehydrogenase that in humans and animals breaks down alcohols to aldehydes or ketones, with the reduction of nicotinamide adenine dinucleotide (NAD⁺) to NADH. In yeast and some bacteria, alcohol dehydrogenases can also catalyse the opposite reaction as part of a fermentation process. Sensors that incorporate this kind of enzymatic reaction typically rely on changes in optical absorbance at specific wavelengths to monitor the redox state of the NAD⁺/NADH cofactor, or of an oxidoreduction mediator that catalyzes the electron transfer between an electrode and the cofactor.

6.5.1 Mediated Amperometric Biosensors

All of the redox enzymes have a cofactor, such as FAD, located deeply within their structure. Therefore, direct electron transfer via electron tunnelling to the surface of a conventional metal electrode is kinetically not favourable. This can be overcome using synthetic or biologically active charge carriers as intermediates between the redox active site in the enzyme and the electrode. The mechanism for this mediated electron transfer to the electrode can be represented by the following two steps:



The second step occurs as a reaction at the electrode surface, where $n(\text{C})_{\text{red}}/n(\text{C})_{\text{ox}}$ is assumed to be a one- or two-electron redox couple. The redox potential of this mediating couple should be such as to permit an energetically favourable electron transfer from the reduced enzyme to the electrode. An example of such a scheme is shown schematically in Figure 6.14 using potassium ferrocyanide $[\text{K}_4\text{Fe}(\text{CN})_6]$ whose anion $[\text{Fe}(\text{CN})_6]^{4-}$ (known as ferrocyanide) readily oxidises to ferricyanide:



The redox potential of this couple is $\sim +0.45 \text{ V}$ (vs. SHE), whilst that for the FAD/FADH₂ or FMN/FMNH₂ couple is $\sim -0.23 \text{ V}$ (vs. SHE). From Chapter 5 (Section 5.3.4) it follows that, with its more positive redox potential, the ferri/ferrocyanide couple provides an energetically favourable potential gradient for electron transfer from a flavin redox site of an enzyme. Ferri/ferrocyanide has been used in glucose sensors, for example, as an intermediate between the redox centre of glucose oxidase and the working electrode. Ferrocyanide and ferricyanide are also impermeable to the plasma membranes of cells, and this makes them useful as an extracellular electron receptor probe in the study of redox reactions in cells.

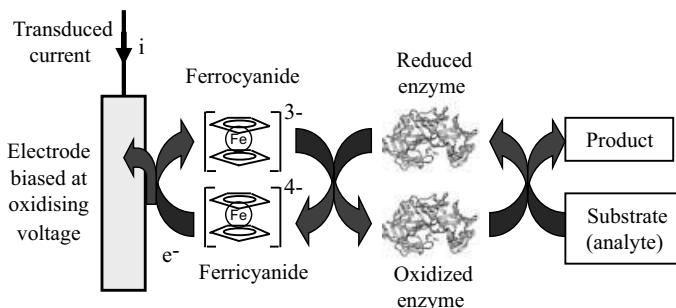


Figure 6.14 The ferri/ferrocyanide couple possesses a more positive redox potential than the flavin redox centre of oxidoreductase enzymes and can thus provide an energetically favourable pathway for mediated electron transfer from the enzyme to an electrode biased at a suitable potential.

Any increase in ferrocyanide can be attributed to secretions of reductants or to electron transport events occurring across the membrane, and can be monitored spectroscopically at a wavelength of 535 nm (the ferro-form is yellow and the ferri-form is red).

Other mediators commonly used are phenylene-diamine, diamino-benzidine, hydroquinone, osmium or ruthenium complexes and dyes. A class of planar organometallic complexes, known as conducting organic salts, can also serve as mediators. They consist of stacks of alternating electron donor molecules (e.g. tetracyanoquino-dimethane) and electron acceptors (e.g. tetrathiafulvalene) and can act as metallic conducting pathways between redox couples and an electrode surface.

Electron transfer mediators should possess the following properties:

- They must react rapidly with the reduced form of the enzyme.
- They must be sufficiently soluble, in both their oxidised and reduced forms, to be able to diffuse rapidly between the active site of the enzyme and the electrode surface.
- The reduced form of the mediator should not readily react with oxygen.
- They should be stable in both their reduced and oxidised forms and should not react with other chemicals in the sensor.
- They must be able to compete with the enzyme's natural substrate (e.g. molecular oxygen in the case of oxidases).
- Their redox potential should provide an appropriate potential gradient for electron transfer between an enzyme's active redox site and an electrode (see Figure 6.14).
- They should exhibit reversible electrochemistry and have a large rate constant for the interfacial electron transfer at the electrode surface.

It has become common practice to refer to biosensors that employ electron transfer mediators as *second generation* biosensors. Historically, in the development of biosensors they came after the *First generation* ones in which the normal product of an enzyme reaction diffuses to the electrode and produce an electrical response. The term *third generation* is given to those biosensors where the enzyme reaction itself causes the measured response, without direct involvement of either a reaction product or mediated electron transfer. This is discussed further in Section 6.8 that describes the development of the glucose sensor.

6.6 Potentiometric Biosensors

These are less common than amperometric biosensors. In a potentiometric biosensor the potential difference between the indicator (working) electrode and a reference electrode is measured without polarising the complete electrochemical cell. In other words, minimal electric current is permitted. The working electrode assumes a potential difference between the reference electrode whose magnitude depends on the activity $[A]$ (effective concentration) of a specific analyte in the test solution. Changes in this potential are given by the Nernst equation (see Chapter 5, Section 5.3.5), which at room temperature is:

$$E = E^o + \frac{RT}{nF} \ln[A] = E^o + \frac{0.059}{n} \log_{10}[A].$$

The most common form of potentiometric biosensor incorporates an ion selective membrane in an arrangement shown in Figure 6.15. Two chambers, one containing the test solution and the other a fixed concentration of the ion to be measured, are separated by a membrane. This membrane is selectively permeable to, or selectively binds, the ion to be measured. An electrode (e.g. Ag/AgCl) is located in each chamber and connected through a high-input impedance millivoltmeter. As an alternative to the membrane, the indicator electrode can be coated with a coating that selectively permeable to the ion to be detected.

The voltmeter shown in Figure 6.15 measures the difference ΔE between the Nernst potentials E_{test} and E_{ref} at the electrodes placed in the test and the reference (fixed) solution, respectively:

$$\Delta E = \left(E^o + \frac{0.059}{n} \log_{10}[A_{test}] \right) - \left(E^o + \frac{0.059}{n} \log_{10}[A_{ref}] \right)$$

to give a linear relationship

$$\Delta E = K + \frac{0.059}{n} \log_{10}[A_{test}]s \quad (6.4)$$

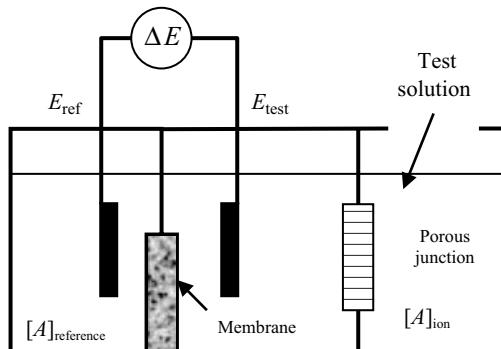


Figure 6.15 Schematic of a potentiometric biosensor that incorporates an ion selective membrane to separate two reference electrodes immersed in solutions of different concentrations of the ion to be measured.

in which K ($= 0.059/n \log[A_{ref}]$) is a constant for a particular reference arrangement and is equal to the intercept of the straight line extrapolated back to zero test analyte concentration. Sensors can be made sensitive to various ions (e.g. hydrogen, fluorine, iodine, chlorine ions) in addition to gases such as carbon dioxide and ammonia. Enzymes that through their natural reactions can change the concentration of any of these ions or gases can be immobilised in the sensor as the means to measure their substrate concentrations, or to detect inhibitors (e.g. heavy metal ions, insecticides) or modulators of the enzyme.

Ideally, the potential difference between the indicator and reference electrode is proportional to the logarithm of the test ion activity or gas fugacity. However, this is only the case when:

- the membrane or indicator electrode surface layer is 100% selective for the test analyte; or
- there is a constant or low enough concentration of interfering ions; and
- potential differences at various phase boundaries (junction potentials) are either negligible or constant, except at the membrane-sample solution interface.

In many cases the indicator electrode takes the form of a pH electrode to measure the activities of enzymes (and hence the concentration of the specific substrate for that enzyme) which produce or consume protons as a result of catalysis. Examples of enzymes that can be used in this way are urease, glucose oxidase, penicillinase and acetylcholinesterase – to monitor the concentrations of urea, glucose, penicillin, and the neurotransmitter acetylcholine (or some pesticides that inhibit acetylcholine-esterase), respectively.

Suitably modified ion selective field effect transistors (ISFETs) can also be used as potentiometric biosensors (Schöning and Poghossian [3]). As described in more detail in Chapter 7, Section 7.8, ISFETs consist of a p-type silicon substrate with two n-doped regions, known as the source and drain. The gate dielectric, typically SiO_2 , is covered by an ion selective membrane which is selectively permeable to specific ions, such as K^+ , Ca^{2+} and F^- . BioISFETs have been constructed using enzymes (an EnFET), where local pH changes resulting from the enzyme activity can be detected. Devices that incorporate antibodies (ImmunFET) and DNA probes (GenFET) face significant problems in converting bio-recognition events into measurable signals. CellFETs take the form of an ISFET in which cells are constrained to the area of the gate insulator. The cell state or viability can be monitored after the administration of chemical agents to the cell suspension, and such devices have the potential as valuable tools for drug discovery environmental toxicity. However, in general, the practical development of BioFETs has been slow because of problems associated with incompatibility of most biomolecule immobilisation methods with ISFET microfabrication, packaging and encapsulation. The long term stability of bioISFETs has also been an issue.

Related devices, known as light addressable potentiometric sensors (LAPS), involve the coupling of a transient photocurrent to an insulated n- or p-doped silicon thin layer in contact with an electrolyte. The transient photocurrent is induced using an intensity modulated light source such as a light emitting diode (LED), and its magnitude depends on the potential at the silicon-electrolyte interface. This technology has been applied to bacterial detection (Gehring *et al.* [4]).

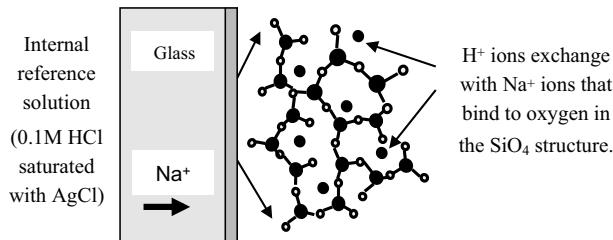


Figure 6.16 The glass membrane of a pH electrode consists of an irregular silicate lattice structure, composed of 22% Na_2O , 6% CaO , 72% SiO_4 .

6.6.1 Ion Selective Electrodes (ISEs)

Ion-selective electrodes are designed to respond selectively to *one* particular ion, hopefully to the exclusion of other ion types. They contain a thin membrane capable of only allowing the desired ion to diffuse through it, or to bind to it. The pH electrode is the most well known and simplest example, and can be used to illustrate the basic principles of ISEs.

6.6.1.1 The pH Electrode

The essential component of a pH electrode is a thin glass membrane (~0.1 mm thick) having an irregular silicate lattice structure (see Figure 6.16) of chemical composition 22% Na_2O , 6% CaO , 72% SiO_4 . When this glass membrane is immersed in a test solution, a hydrated gel of thickness around 10 nm is formed at its surface. Hydrogen ions in the external solution can diffuse into this hydrated gel layer and displace sodium ions, which then diffuse across the membrane to produce an equilibrium potential. In the vast bulk of the glass membrane, the structure remains dry and all the exchange sites are occupied by sodium ions. An ion-exchange equilibrium is established between the sodium and hydrogen ions. Hydrogen ions are the only ones to bind significantly to the hydrated gel layer and to act in this way.

From Equation (6.4), noting that $n = +1$ for hydrogen ions and that pH is defined as the negative logarithm of the hydrogen ion concentration, the potential difference created across the glass membrane is given by:

$$\Delta E = K - 0.059 \text{ pH} \text{ (Volts).} \quad (6.5)$$

Regular calibration of a pH electrode should be made against known standards. Measurement of pH cannot be more accurate than that of a standard, which is typically ± 0.01 .

The following sources of error are common in the measurement of pH:

- Junction potential: If ionic strengths differ between analyte and the standard buffer, the junction potential will differ and result in an error of ± 0.01 .
- Junction Potential Drift: Caused by slow changes in $[KCl]$ and $[AgCl]$ – *re-calibrate!*
- Sodium Error: At very low $[H^+]$, corresponding to levels above $\sim pH 12$, the ion-exchange equilibrium favours Na^+ and the apparent pH is lower than the true pH.
- Acid Error: At high $[H^+]$, corresponding to levels below $\sim pH 1$, the measured pH is higher than actual pH. The hydrated gel layer is saturated with hydrogen ions.

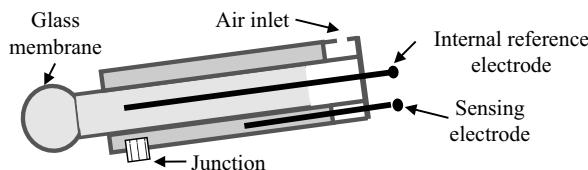


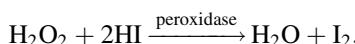
Figure 6.17 A pH sensor commonly takes the form of concentric tubes, with the inner one containing the reference electrode and a bulbous glass membrane. The sensor is immersed vertically into the test solution, to cover the membrane and ceramic junction that serves as a salt bridge to the sensing electrode.

- Equilibration Time: Takes ~30s to minutes for electrode to equilibrate with solution.
- Hydration of Glass Membrane: A dry membrane will not respond to H^+ correctly.
- Temperature: Calibration needs to be done at same temperature of measurement.
- Cleaning: Contaminants on probe will cause reading to drift until properly cleaned or equilibrated with analyte solution.

A pH electrode commonly consists of the concentric format shown in Figure 6.17.

6.6.1.2 Other Types of Ion-Selective Electrode

We have seen that the important sensing element of a pH electrode is a silicate system based on a network of SiO_2 molecules. By incorporating different metal oxides into this molecular network, such as those of aluminium, boron, calcium, lithium, potassium and sodium, specialised glass membranes can be formulated to favour different types of ion-exchange equilibria. Sensors that respond specifically to ions such as Na^+ and Ag^+ can be made in this way. Chalcogenide glasses, based on molecular networks of AsS , AsSe and AsTe can also be formed to act as sensors for divalent ions such as Ca^{2+} , Cd^{2+} and Pb^{2+} . Iodide ion selective electrodes can also be made using a glass membrane formed by the co-precipitation of silver iodide and silver sulphide. The monitoring of the hydrogen peroxide formed by oxidoreductase enzymes provides an example of how an ion selective electrode can be used in a biosensor. Rather than exploiting the electro-oxidation reaction given by Equation (6.2) the hydrogen peroxide can be used to oxidise hydrogen iodide to iodine in the presence of a peroxidase enzyme:



The decrease in iodide ion concentration, as measured by an iodide-selective electrode, can be used to determine the amount of hydrogen peroxide produced.

Instead of a glass membrane, ion-specific sensors can be made using ionic conductors, such as the one for fluorine ions shown in Figure 6.18. Lanthanum fluoride (LaF_3) forms a regular crystal lattice structure, but when doped with europium lattice defects are produced in the form of vacant sites. The F^- ions can readily migrate through this lattice structure by hopping from one vacancy site to another. If the membrane shown in Figure 6.15 takes the form of such a fluoride ion conductor, the sensor becomes mainly selective to fluoride ions. Ion selective polymeric membranes can also be formed by mixing an ionophore (an ion

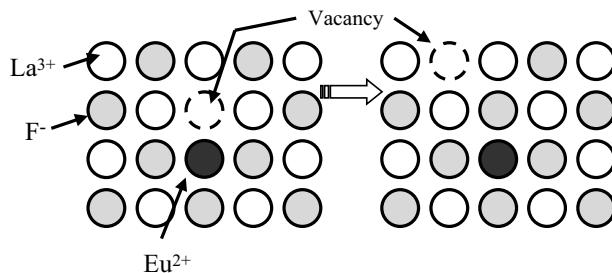


Figure 6.18 Solid state membrane for the electrode of a fluorine sensor. The membrane consists of a lanthanum fluoride (LaF_3) crystal doped with europium. Fluorine ions migrate through the crystal by hopping between crystal vacancies caused by the presence of the europium ions in the LaF_3 crystal lattice. This produces a potential difference across the membrane.

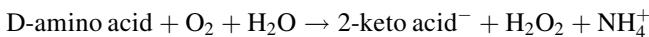
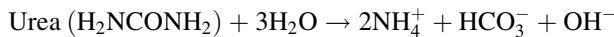
carrier) with a hydrophobic polymer such as polyvinylchloride (PVC) and a plasticiser to make the rigid plastic more ‘fluidic’. Some of these ionophores can be synthetic, based on crown ethers, for example, whilst others occur as natural biological molecules, such as enniatin (NH^{3+}), gramicidin (H^+ , Na^+ , K^+), ionomycin (Ca^{2+}) and valinomycin (K^+). A limitation for some of these sensors, especially those based on natural biological ionophores, is that they will only work effectively over a narrow pH range.

In contrast to the pH glass membrane, other ion-selective membranes are less ion-specific and can permit the passage of some of the other ions which may be present in the test solution, thus causing the problem of ionic interference. Also, most other ISEs have a much lower linear range and higher detection limit than the pH electrode. Many exhibit a curved calibration line in the region 10^{-5} to 10^{-7} M and very few can be used to determine concentrations below 1×10^{-7} M. Thus, for test solutions containing a low concentration of the ion to be measured, it may be necessary to construct a calibration graph with several points in order to define the slope more precisely in the nonlinear range. Also, the calculation of ionic concentration is far more dependent on a precise measurement of the potential difference than is the pH, because the pH depends on the order of magnitude of the concentration rather than the precise value. For example, it would take an error of nearly 6 millivolts to cause a change of 0.1 pH units, but only a 1 millivolt error will cause a 4% error in the calculated concentration of a monovalent ion (and 8% for a divalent ion). This arises because the theoretical value for the slope at 298 K is 59.2 mV for monovalent ions and 29.6 mV for divalent ions (e.g. $\text{antilog}(1/59) = 1.04$). It should also be remembered that pH is defined as the negative log of the *activity* of the ion (which is measured directly by any ISE) but most measurements of other ions require the actual concentration, which can differ significantly from its activity in a sample with complex matrices and high ionic strength.

6.7 Conductometric and Impedimetric Biosensors

The simplest form of a conductometric sensor uses interdigitated electrodes of the form shown in Figure 6.4b coated with an immobilised enzyme whose reaction with its natural substrate creates, or changes the nature of, electrically charged species.

Candidate enzyme reactions include the following urea/urease and amino acid oxidase reactions:



in which the initially uncharged substrates are hydrolysed to yield charge-bearing species. Other possibilities for conductometric monitoring of enzyme activity include: the generation of ionic groups by amidases; the separation of unlike charges by decarboxylases; protonic conduction induced by the action of esterases; the action of kinases in changing the degree of association of ionic groups; changes in the size (hence mobility) of charged species by phosphatases. The enzymes can be immobilised in a gel formed by crosslinking glutaraldehyde with a passive protein such as albumin or covalently bonding it to collagen. A conductometric bi-enzymatic biosensor using immobilised *Chlorella vulgaris* microalgae as bioreceptors has been described by Chouteaua *et al.* [5], in which algae were immobilised on interdigitated electrodes with glutaraldehyde-treated albumin. Local conductivity variations caused by algae alkaline phosphatase and acetylcholinesterase activities were detected. These two enzymes are known to be inhibited by distinct families of toxic compounds, namely heavy metals for alkaline phosphatase, carbamates and organophosphorous pesticides for acetylcholinesterase. The bi-enzymatic biosensors were tested to study the influence of heavy metal ions and pesticides. The limit of detection (LOD) for Cd^{2+} and Zn^{2+} ions was determined to be 10 parts/billion (ppb) after 30 minutes exposure, whereas Pb^{2+} produced no significant inhibition as this ion appeared to adsorb preferentially on the albumin matrix. Paraoxon-methyl was found to inhibit *C. vulgaris*. The biosensors were exposed to different mixtures of the ions and pesticide, and no synergistic or antagonist effects were observed.

The close separation of electrode pairs in the interdigitated geometry means that a measurable ac current response can be obtained using small-amplitude ac voltages (e.g. 10 mV pk-pk, 1 kHz). Because the charged groups are detected altogether, unless the electrodes are first coated with an ion selective layer, the method is relatively nonselective. Also, the sensitivity achieved will depend on the relative change of conductance produced by the enzymatic reaction to the initial conductance of the medium in contact with the electrodes. This limits *in vivo* applications of conductometric biosensors. The sensitivity can be enhanced by including in the sensor design a reference dummy electrode that incorporates an inactivated enzyme or is isolated from the enzyme reaction. In the designs of enzyme-based conductometric sensors described by Mikkelsen and Rechnitz [6] detection limits of 1 μM and linear ranges of two orders of magnitude were obtained.

The transduction element of a conductometric biosensor described by Muhammad and Alocija [7] for the detection of foodborne pathogens is shown in Figure 6.19. This transducer consists of polyclonal antibodies immobilised onto a nitrocellulose membrane between two silver electrodes. The polyclonal antibodies can be chosen to be specific for a wide range of *Salmonella* species and *E. coli* serotypes, for example. A liquid sample containing a target antigen is applied to a sample pad and drawn by capillary action over a reaction pad, past the immobilised antibodies and into an absorbing collection pad. The reaction pad consists of a fibre-glass membrane onto which is absorbed an antibody complexed to conducting polyani-line. Before the sample is applied, the gap between the electrodes in the capture pad represents an electrical open circuit (see Figure 6.19).

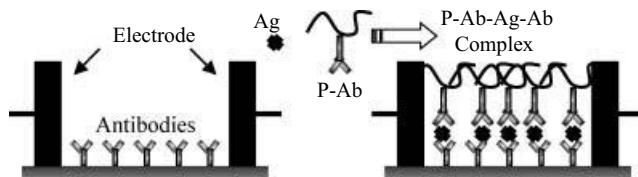


Figure 6.19 The transducing element of a conductometric biosensor for the detection of foodborne pathogens [4]. The pathogenic antigen (Ag) binds to an antibody (Ab) labelled with conducting polyaniline (P). This forms a sandwich complex (P-Ab-Ag-Ab) with immobilised antibodies on the capture pad to create a conducting bridge between the two electrodes.

After applying the sample the solution carrying the antigen flows to the reaction pad, dissolves the adsorbed antibody-polyaniline layer, and any complexes formed as a result of antibody-antigen binding are carried into the capture pad containing the immobilised antibody. At this point a second antibody-antigen reaction occurs and produces a sandwich structure, as shown in Figure 6.19. The polyaniline component of the sandwich creates molecular conducting links that bridge across the two silver electrodes, and is detected as a decrease in the resistance between the electrodes. Any unbound nontarget pathogenic organisms are carried by capillary flow to the absorption membrane. The detection limits determined by Muhammad and Alocija [7] for *Salmonella* and *E. coli* O157:H7 were determined to be 8.3 ± 0.6 and 7.9 ± 0.3 colony forming units (CFU) per ml, respectively, which can provide an excellent sensitivity for yes/no (qualitative) detection.

A development of this form of detection method has been described by Weizmann *et al.* [8] for DNA detection, involving carbon nanotube-DNA nanowire devices and oligonucleotide-functionalised enzyme probes. The key aspect of this sensor design is a DNA-linked-CNT wire motif, which forms a network of interrupted carbon nanotube wires connecting two electrodes. Sensing occurs at the DNA junctions linking CNTs, followed by amplification using enzymatic metalisation leading to a conductometric response. The DNA analyte detection limit is 10 fM with the ability to discriminate single, double, and triple base pair mismatches.

In the search for new methods to miniaturise biosensing elements, investigations are in progress to use self-assembled monolayers (SAMs) as immobilising matrices for enzymes and DNA. SAMs are nanoscale surface deposits formed through a spontaneous adsorption process of organic molecules onto various substrates such as gold, silver, iron, copper, and glass. A binary SAM can be produced in which each monolayer consists of two different molecules and functional groups. Once a binary-SAM is produced, it is possible by electrochemical means, to selectively remove parts of the outer monolayer to create defects in the binary-SAM, or to completely remove the outer layer to form a uni-SAM. This procedure can be exploited to give a desired spatial distribution of uni- and binary-SAMs on an electrode surface, to obtain structures similar to those obtained using conventional photolithography. Park *et al.* [9] employed electrochemical impedance spectroscopy to monitor the growth and electrochemical parameters of SAMs, and a typical result is shown schematically in Figure 6.20.

The results shown in Figure 6.20 take the form of a complex plane impedance plot as depicted in Figures 4.20 and 5.27 of Chapters 4 and 5. Park *et al.* [9] used the equivalent

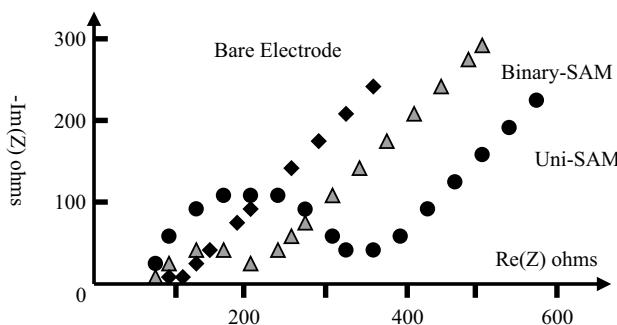


Figure 6.20 Complex impedance plots, based on results described by Park *et al.* [6] obtained with a bare electrode and one coated with a self-assembled monolayer (SAM). Comparison with Figure 5.27 indicates that the SAM increases the charge transfer resistance between the electrode surface and a redox probe.

electrical circuit shown in Figure 5.26 to fit all of their experimental results. The SAMs were grown on planar gold electrodes whose effective surface area had been increased by the electrochemical deposition of gold nanoparticles, and ferro/ferricyanide redox probes were used in the impedance spectroscopy measurements. A comparison of Figures 6.20 and 5.27 indicates that a SAM coating on the electrodes increased the charge transfer resistance across the electrode-solution interfaces. Results such as those schematically depicted in Figure 6.20 were used by Park *et al.* [9] to characterise the differences in electrochemical parameters (electron transfer rate and resistance, Warburg impedance and electrical double layer capacitance) of different types of uni-SAMs and binary-SAMs. An optimum SAM structure for a conductometric biosensing application would ideally exhibit a high electron transfer rate and low double layer capacitance. Electrochemical impedance spectroscopy can serve as a tool to aid the optimum design of SAMs in terms of their molecular constituents and patterning. It could also serve as the detection platform in a biosensor that employs SAMs or other membrane structures to immobilise electroactive transduction elements at an electrode surface.

6.8 Sensors Based on Antibody–Antigen Interaction

The sensor shown in Figure 6.19 relies on interactions between immobilised antibodies and antigens. Antibodies are immune system-related proteins called immuno-globulins. Each antibody consists of four polypeptide chains – two heavy chains and two light chains joined to form a Y-shaped molecule, as depicted in Figure 6.21a. The amino acid sequence in the tips of the 'Y' varies greatly amongst different antibodies. This variable region, composed of 110~130 amino acids, gives the antibody its specificity for binding antigen. The variable region includes the ends of the light and heavy chains, and is further subdivided into hyper-variable (HV) and framework (FR) regions. Hypervariable regions have a high ratio of different amino acids in a given position relative to the most common amino acid in that position. The HV regions directly contact a portion of the antigen's surface. The FR regions, which have more stable amino acids sequences, form a beta-sheet structure which serves as a scaffold to hold the HV regions in position to contact the antigen.

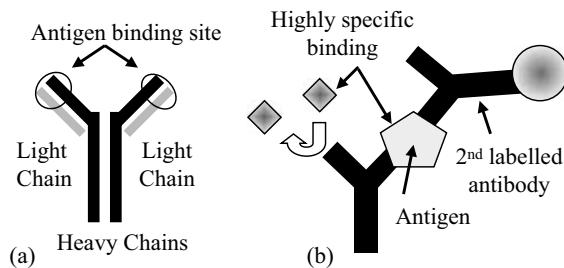


Figure 6.21 (a) An antibody consists of four polypeptides – two heavy chains and two light chains forming a ‘Y’ shaped molecule. The amino acid sequence in the tips of the ‘Y’ varies greatly amongst different antibodies, and this gives the antibody its specificity for binding antigen. (b) Signal amplification can be obtained using a second antibody labelled with a fluorescent dye, magnetic bead, conducting particle or enzyme, for example.

Antibodies are produced by immunising animals (rabbits, mice or rats) with the antigen. B-cells in their blood initiate an immune response by producing antibodies. After checking for the presence of these antibodies, the B-cells are extracted and grown in cultures to increase their number and the amount of antibodies produced. There are many types of B-cells in blood and they produce a variety of polyclonal antibodies having different affinities. Monoclonal antibodies, all having identical affinities, are produced by taking B-cells from the immunised animal and fusing them with myeloma cells. These hybridoma cells are then cultured (expanded) and the antibodies they produce harvested.

An antigenic determinant, a site on the antigen that the immune system responds to by making antibody against it, is frequently a unique structure on the antigen. For example, in hen egg white lysozyme a glutamine at position 121 (Gln 121) protrudes away from the antigen surface. The antibody’s HV region forms an opening to surround the antigen’s protruding Gln 121. Hydrogen bonds stabilise the antibody–antigen interaction. In addition to hydrogen bonds other weak interactions, such as van der Waals forces, hydrophobic and electrostatic forces, improve the binding specificity between antibody and antigen. These interactions occur over large and sometimes discontinuous regions of the molecules, thus improving the binding affinity. Water molecules in spaces between the antigen and the antibody also contribute significantly to the binding energy by creating additional hydrogen bonds.

When incorporated into a sensor, the antibody is usually immobilised on a solid surface that can be passivated to avoid adventitious binding of the antigen to the surface, rather than to the antigen binding site of the antibody. The sensor shown in Figure 6.19 makes use of a second antibody–antigen reaction to amplify the transduction signal, using a conducting molecular tag. Another example of this technique is shown in Figure 6.21b where a common technique known as competitive binding (see Figure 6.22) is employed. In this example a fluorescent tag on the antibody is replaced by the target antigen because it exhibits a much stronger binding affinity than the fluorophore. When used in a TIRF device, the fluorescence excited by the evanescent wave is reduced after the fluorophore is replaced by the antigen. The second antibody could also have an enzyme or magnetic bead attached to it, which extends the methods of detection to include conductometric, optical, electrochemical and magnetic sensing.

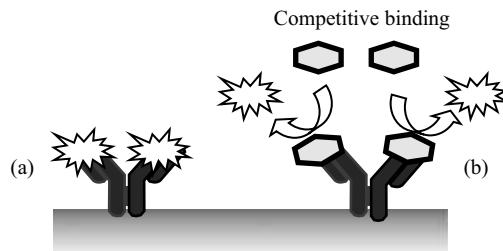


Figure 6.22 (a) Immobilised antibody, with fluorescent tags bound to its antigen binding sites, in a total internal reflection fluorescence (TIRF) sensor. (b) TIRF signal decreases when an antigen displaces the fluorescent tags as a result of competitive binding.

6.9 Photometric Biosensors

Photometric biosensors (also known as opto-biosensors) operate by directing a light beam at a biochemically active surface and detecting changes in absorption, scattering, refractive index, fluorescence or chemiluminescence using a photodiode or photomultiplier. Direct label-free detection relies on a change of the intrinsic optical property of the biosensing element as a result of its interaction with the target analyte. An example is the spectral change that occurs in the visible range for the redox active flavin group of an oxidoreductase enzyme on reacting with its substrate. However, it can be difficult to separate out background absorbance or fluorescence and such sensors can suffer from nonspecific binding events and low analytical sensitivity to low molecular mass analytes. These problems can often be addressed by employing a reference sensing region next to the bioactive site. However, many optical biosensors use a fluorophore that is covalently bound to a chemical label or probe, in which case fluorescence intensity or lifetime becomes the analytical parameter. The choice of fluorophore and covalent bonding can be chosen to give an analytical wavelength to suit the available light source and optical detection system, and to provide for readily measurable fluorescence lifetimes in the microsecond to millisecond range.

Fluorescence intensity (F) obeys a linear relationship with concentration of the form described by the Beer Lambert law (Equation (4.3) of Chapter 4):

$$F = I_o \varepsilon [C] k l \phi \quad (6.6)$$

where I_o is the intensity of the incident (laser) light beam, ε is the molar absorption coefficient of the chromophore, $[C]$ its concentration, l the optical path length through the labelled substrate, k is a geometry factor related to such factors as the arrangement of optical filters and prisms, and ϕ is the quantum yield of fluorescence. Measurement of the intensity can suffer from interference associated with the background fluorescence and scattering of light by the biosensing element and the test sample, as well as an unsteady light source intensity and baseline drifts of the photodetector output. These can be compensated for by ratio-metric measurements at two wavelengths of fluorescence using an inert second fluorophore, or with a donor and acceptor dye and exploitation of the Förster Resonance Energy Transfer (FRET) technique described in Section 4.3.9 of Chapter 4.

One of the first examples of an opto-biosensor was described by Wolfbeis *et al.* [10], who exploited the fact that the phosphorescence and fluorescence of many dyes (e.g. flavins,

chlorophyll and porphyrins) can be quenched by oxygen. A ruthenium-based dye was bound to the surface of ion exchange beads to form an aqueous emulsion, which was dried and mixed with silicone to form a 50 mm thick film. This film was then hydrated by submerging in warm water. The membrane embedded dye was excited into fluorescence at a wavelength of 470 nm to produce a fluorescence emission at 610 nm. Tests were performed with alcohol oxidase and catalase coembedded into the dye-bead emulsion before mixing with silicone, and this provided a sensor capable of responding to ethanol in the 0.1–1.0 M range. The method relied on the fact that alcohol oxidase oxidises ethanol, with consumption of oxygen and production of hydrogen peroxide, whilst catalase decomposes hydrogen peroxide to water and oxygen. This innovative work clearly demonstrated the promise of combining the sensitivity of fluorescence and the selectivity of enzymatic reactions.

Oxidised luminol, in the presence of hydrogen peroxide, exhibits chemiluminescence in the form of emitted blue light. This can be used to detect the production of hydrogen peroxide by oxidases, and in its most elaborate form has been developed into a multi-analyte bio-sensor chip. In the original form of this described by Marquette *et al.* [11] six different oxidases, specific for choline, glucose, glutamate, lactate, lysine and uric acid, were immobilised onto microbeads. Each bead type was then mixed with luminol coated beads in a photopolymer and deposited, as shown in Figure 6.23, as small dots of 800 μm diameter along six microchannels formed on top of a glassy carbon electrode. The luminol was poised at an oxidised state by applying a +0.85 V potential between the glassy carbon electrode and a platinum pseudo-reference electrode. Chemiluminescence emitted by any of the spots of immobilised enzyme and luminol, after injection of test samples into the microchannels, was detected using a charge coupled device (CCD) camera. A schematic of the luminescence reaction involving a choline coated and a luminol coated bead is included in Figure 6.23. The detection ranges obtained were 1–25 μM (uric acid), 1 μM -0.5 mM (glutamate and lysine), 20 μM -2 mM (glucose) and 2 μM -0.2 mM (choline and lactate), and the multi-functional sensor chip was successfully used to detect glucose, lactate and uric acid in human serum, without the need of internal calibration of the sample.

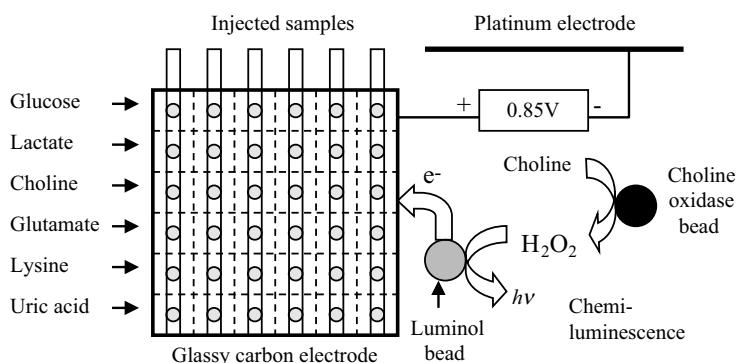


Figure 6.23 A schematic of the multifunctional sensor chip described by Marquette *et al.* [11]. Six oxido-reductase enzymes, specific for glucose, lactate, choline, glutamate, lactate, lysine and uric acid, are coated onto microbeads and immobilised with luminol coated beads in small spots of a photopolymer. The generated hydrogen peroxide and oxidised luminol react to produce chemiluminescence.

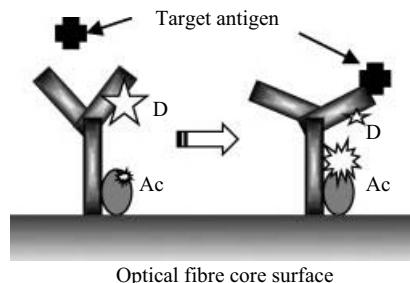


Figure 6.24 Principle of operation of the FRET immunosensor described by Ko and Grant [9]. Evanescent radiation from the fibre core excites the donor (D) fluorophore. When the target antigen binds to the antibody, a conformational change occurs and results in energy transfer from the donor to the acceptor (Ac) fluorophore.

FRET systems have often been employed in biosensors, and an immunosensing example described by Ko and Grant [12] for the detection of *Salmonella* antigens is shown in Figure 6.24. The evanescent wave emitted from the core of an optical fibre is used to excite donor fluorophores attached to IgG antibodies that are immobilised on the surface of the fibre core. A protein, fluorescently labelled to act as the acceptor in the FRET process, is bound to the Fc portion of the antibodies, enhancing their correct orientation on the core surface for binding to the target pathogens in the sample solution. In the absence of the pathogen, the fluorescence occurs at the emission wavelength of the donor fluorophore, with little or no fluorescence from the acceptor. On binding to the target pathogen, a conformational change occurs in the 3D structure of the IgG antibody leading to a decrease in distance between the donor and acceptor fluorophores. The FRET process can then occur, leading to an emission from the acceptor. Ratiometric detection is then possible and the ratio of donor to acceptor emission provides a measure of the binding state between the antibody and targeted pathogen on the surface of the optical fibre. Because this sensing principle integrates the basic principles of FRET with the inherent conformational changes of an antibody as it binds to an immobilised antigen, the likelihood of a false positive signal is reduced. The sensor was tested using buffered saline and homogenised pork sample solutions doped with *Salmonella typhimurium*. The limit of detection obtained for these two tests were 10^3 cells/ml and 10^5 colony forming units (CFU) per gram within a 5 minutes' response time, respectively.

Figure 6.25 shows the method described by Ueberfeld and Walt [13] in the development of a reversible fluorescent DNA probe that can be used to determine the concentration of single-stranded DNA in solution. The probe consists of a single-stranded DNA molecule that is labelled with a FRET donor and acceptor fluorophore. The fluorescent dyes Cy3 and Cy5 were chosen by Ueberfeld and Walt as the donor and acceptor fluorophores, respectively. The excitation wavelength of the donor was 530 nm, and the Cy3 (donor) and Cy5 (acceptor) emissions were detected at 565 and 666 nm, respectively. In its nonhybridised state the DNA molecule adopts a stem-loop conformation. Upon hybridisation with the target DNA strand, the loop conformation cannot coexist with the resulting double-stranded section of DNA, and so the probe opens up to force the donor and acceptor fluorophores apart. The FRET process weakens, resulting in a fluorescence intensity increase of the donor and a fluorescence intensity decrease of the acceptor. The ratio of the acceptor-to-donor fluorescence intensities is

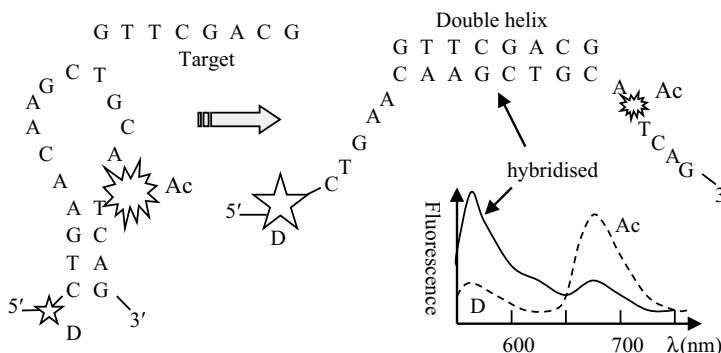


Figure 6.25 Principle of operation of the DNA sensor described by Ueberfeld and Walt [10]. The loop formed by the nucleotide probe sequence CAAGCTGC is straightened out when it hybridises to the complimentary target oligonucleotide sequence. This increases the distance between the FRET donor (D) and acceptor (Ac) labels and results in an increase and decrease, respectively, of the donor and acceptor fluorescence intensities.

independent of the amount of probe and so provides a quantitative measure of the free DNA target concentration. For the probe to exhibit reversible binding to the target, the free energy of the hybridised probe-target DNA duplex should be of the same magnitude as the free energy of the stem formation. The stem length and the length of the loop region complementary to the target were chosen to allow for such reversible binding.

6.10 Biomimetic Sensors

These sensors attempt to imitate Nature's elegant biosensing schemes whereby the initial recognition and subsequent binding of a target biomolecule is coupled to a cascade of signalling and amplification events. For example, the sense of smell involves the specific binding of an odorant molecule to a protein and the selective binding of this odorant-protein complex to an olfactory receptor protein, followed by signal transduction and amplification involving the opening of a large number of membrane ion channels and stimulation of neuron action potentials to the brain. An early example of a biomimetic sensor that uses whole cells is that of Peter *et al.* [14] for detecting halogenated hydrocarbons, based on the liberation of halogen ions by the action of the enzyme alkyl-halidohydrolase present in *Rhodococcus* bacteria. The design of this potentiometric sensor involves immobilising the cells close to the membrane of an ion-selective (e.g. chloride or bromide) electrode. Using whole cells as the sensing element avoids the often complicated and expensive process involved to isolate the required enzyme from its native cell membrane or cytoplasm. A major disadvantage, however, can be the sensor's lack of selectivity to a specific target molecule due to the presence of many other enzymes in a cell. For this reason efforts have been made to apply much simpler systems, such as functionalised lipid bilayers or vesicles, which to some extent mimic the structure of the cell membrane and can act as both the recognition and transduction surface. A noteworthy example is described by Cornell *et al.* [15] in which the conductance of a population of gated ion channels is switched by a molecular recognition event. The sensor represents an impedance device that can be integrated into a microelectronic circuit and used in a wide range of biological media, including blood. The active element consists of a lipid

bilayer, each layer containing gramicidin ion channels, anchored to a gold electrode. The ion channels in the outer layer are mobile and can form conducting dimers with the ion channels located in the inner membrane leaflet. When an electric potential is applied between the gold substrate and the external solution, ions can flow through each conducting dimer between the external solution and the gold surface. When an ion channel in the outer leaflet of the membrane captures a target protein molecule, it is no longer able to form a conducting dimer across the membrane, and so this ionic conducting element is switched off. Because each conducting dimer can conduct up to a million ions per second, this switching off represents a high gain event. This gated-ion sensor can be used with most types of receptor molecule, including antibodies and nucleotides, and envisaged applications include cell typing, the detection of large proteins, viruses, antibodies, DNA, electrolytes, drugs, pesticides and other low-molecular-weight compounds.

Song and Swanson [16] have described an optical analogue to the gated ion-channel bio-sensor described by Cornell *et al.* [15], for the detection of biological toxins such as the cholera, ricin and shiga toxin. These toxins take the form of a protein made of two subunits – A and B. Subunit A is the catalytic subunit, responsible for the cytotoxic activity, while subunit B recognises the target cell through multiple binding sites for receptors (known as gangliosides) on the cell surface. The chemical structure of a particular ganglioside determines its specificity towards the binding of a target protein molecule. Song and Swanson's sensor mimics this action using optically tagged ganglioside receptors whose mobility in the upper leaflet of a bilayer membrane triggers an optical fluorescence change upon protein binding. The bilayer membranes are composed of natural phosphatidylcholine containing fluorophore-labelled gangliosides, and are formed into either vesicles or as a coating on silica microspheres. On capture of a target protein by this artificial cell membrane, signal-transduction can occur through a fluorescence quenching process or via the Förster resonance energy transfer (FRET) process described in Chapter 4 (Section 4.3.9). In the first of these two signal-transduction schemes (see Figure 6.26a the fluorophore-tagged gangliosides are uniformly incorporated into the outer membrane leaflet surface at a concentration low enough to ensure that interaction between optically excited fluorophores is unlikely. A strong fluorescence signal should therefore be observed because quenching of the fluorescence through mutual interactions should not occur. However, the binding of subunit B of a toxin protein to two or more of the mobile fluorophore-tagged gangliosides will bring the fluorophores into close proximity and trigger a decrease of fluorescence intensity through a fluorescence self-quenching mechanism. Changes of other fluorescence properties such as polarisation and lifetime should also occur. For the second signal-transduction scheme (see Figure 6.26b the ganglioside receptors are covalently attached to either a FRET donor or acceptor fluorophore. They can fluoresce independently to give a strong fluorescence of the donor and weak fluorescence of the acceptor if only the donor is optically excited. The binding-induced aggregation of a FRET donor/acceptor pair triggers an efficient energy transfer to boost the acceptor fluorescence at the expense of the donor fluorescence. The biomimetic membrane surface therefore participates directly in the signal transduction, because a simultaneous two-colour change is directly induced by the multiple binding of the target toxin protein to the membrane incorporated ganglioside receptors.

Song and Swanson [16] also describe a third possible signal-transduction scheme, which involves a combination of the distance-dependent fluorescence quenching and a FRET processes described in Figure 6.26b. This makes it possible for one ganglioside acceptor to

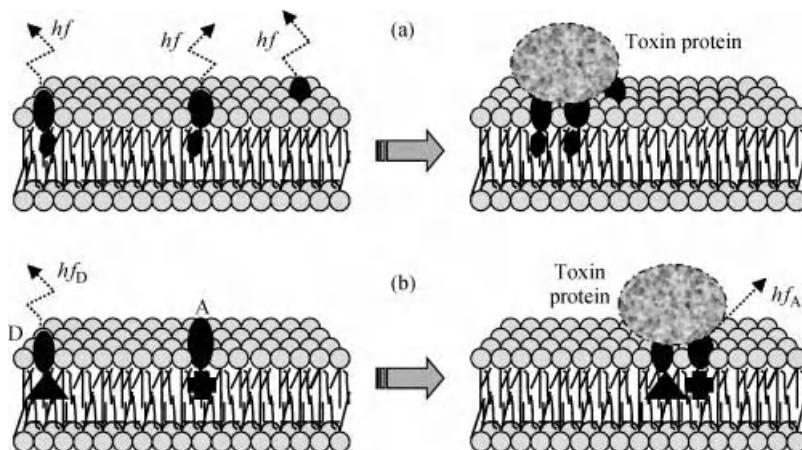


Figure 6.26 Signal-transduction schemes for a biomimetic sensor that detects toxin proteins binding to an artificial cell membrane. (a) The binding of a toxin protein to two or more mobile fluorophore-tagged gangliosides triggers a decrease of fluorescence intensity as a result of a distance-dependent fluorescence self-quenching mechanism. (b) The binding-induced aggregation of ganglioside receptors labelled with FRET donor/acceptor pairs results in reduction of the donor (D) fluorescence and an increase of the acceptor (A) fluorescence. (Based on [16].)

quench the fluorescence of more than 100 FRET donor molecules, resulting in significant signal amplification and a more efficient usage of the mimetic membrane surface area.

6.11 Glucose Sensors

In this section we present an overview of the development of the glucose sensor. A comprehensive coverage is not attempted, but references to key papers and reviews are included so that interested reader can gain a wider appreciation of this important subject. Of all the biosensors, the glucose sensor has the longest history of development and is by far the most commonly used, estimated [17] to account for approximately 85% of the total biosensor market. This status reflects the large and continuing increase of the number of people suffering from diabetes and their need for continuous monitoring of blood glucose level. Normal glucose levels in human blood range from 4 to 5.9 mM and a diagnosis for diabetes is given if this rises above 6.9 mM (8 hours after fasting). Hypoglycaemic episodes (low glucose) cause blackouts and can be life-threatening, whereas periods of hyperglycaemia (high glucose) can cause circulatory disease, strokes, blindness, kidney failure and nerve degeneration. Reliability and accuracy are therefore particularly important for a glucose sensor. Smaller markets for glucose sensors are in the bioprocessing, food and drink industries, and in laboratories developing renewable fuel cells, for example. The following summary of the development of the various generations of enzymatic glucose sensors illustrates some practical principles that are common to many other enzyme-based biosensors.

The principal enzymic component of a glucose biosensor has been glucose oxidase, commonly abbreviated to either GOx or GOD, which is a dimeric protein containing two FAD

flavin groups. Each flavin group acts as a redox centre, located deep within the protein's roughly spheroidal structure of dimensions around $12\text{ nm} \times 10\text{ nm}$, and is reduced in the catalytic reaction with glucose to produce gluconolactone according to the following reaction:



The gluconolactone hydrolyses to gluconic acid. In the presence of oxygen the flavin group is reoxidised as follows:



The first glucose sensor was described by Clark and Lyons [18], and operated by monitoring the oxygen consumed by glucose oxidase immobilised onto the platinum cathode of a conventional potentiometric 'Clark' oxygen sensor. A problem with this form of glucose sensor is that it is sensitive to changes in the background level of oxygen as well as to changes in glucose concentration. This problem was overcome by Updike and Hicks [19] who designed a glucose sensor that incorporated two platinum cathodes (-0.65 V), both coated with gels of immobilised glucose oxidase (see Figure 6.27). The enzyme in one of the gels had been inactivated by heat treatment, and so was unresponsive to glucose. Its platinum cathode, however, remained sensitive to changes in oxygen tension. The other platinum electrode, with its coating of an active enzyme gel, remained responsive to both oxygen and glucose. By monitoring the difference between the outputs of the two platinum cathodes, each referenced against the same Ag/AgCl electrode, changes in glucose concentration could be measured with relatively little sensitivity towards changes in oxygen tension.

The reactions given in Equations (6.7) and (6.8) suggest that oxidation of the generated hydrogen peroxide, as given by Equation (6.3) can also be used to determine glucose concentrations. This was first demonstrated by Guilbault and Lubrano [20] and as such represented the very first amperometric biosensor. The anodic sensing of hydrogen peroxide gives a current directly proportional to the glucose concentration, and the oxygen so generated is able to mediate the regeneration of the catalytic FAD centre to maintain the reaction of Equation (6.7). These early developments represent the *first generation* of biosensors.

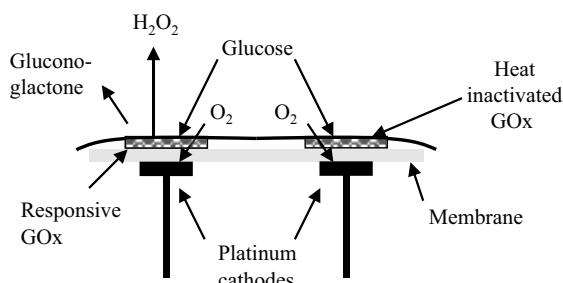


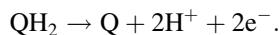
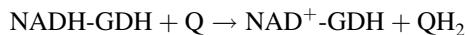
Figure 6.27 Dual cathode design of a glucose sensor that compensates for variations in background oxygen tension (based on [19]). The platinum cathode coated with active glucose oxidase responds to both glucose and oxygen, whilst the other cathode coated with heat inactivated GOx responds only to oxygen. The difference in cathode current outputs is referenced against the same Ag/AgCl electrode.

A significant problem for the first generation glucose sensors is their dependence on oxygen to regenerate the oxidised form of the FAD group in glucose oxidase. The oxygen available in a blood sample, for example, is insufficient to sustain the oxidation of glucose, and this can result in an underestimate of the glucose level. Also, the potential range at which hydrogen peroxide is oxidised coincides with the oxidation potentials of compounds such as uric acid and ascorbic acid (vitamin C), found in blood at concentrations of around 0.05 and 0.3 mM, respectively. Amperometric currents that overestimate the blood glucose level can therefore arise from such interferences, a result that could mask hypoglycaemia or induce this condition through inappropriate dosing of insulin, and possibly leading to a hypoglycaemic coma. Other sugars consumed in food, such as maltose and galactose, can also act as interfering electrochemical responses to glucose sensors. Another problem for amperometric sensing of blood glucose sensing, especially for implanted electrodes, is electrode fouling by blood proteins, platelets and even bacteria. This problem can be reduced by using polymeric coatings, such as polyethylene glycols, that resist protein absorption, or to co-immobilise the glucose oxidase with an anticoagulant such as heparin. An interesting innovation has been the introduction of coatings that give a controlled release of nitric oxide (NO) that inhibits the adhesion of platelets and bacteria to the sensor [21].

The major issues associated with oxygen dependence are addressed by the second generation of enzyme-based glucose sensors, in which nonphysiological electron transfer mediators replace oxygen as the substrate for the glucose oxidase enzyme. The reoxidation of these mediators at the electrode results in an amperometric current. The most commonly used mediator in commercial glucose sensors is ferricyanide, and others such as ferrocene derivatives, quinones and transition-metal complexes have also been used. These mediators each possess the following desired properties: low toxicity; low molecular weight and poor solubility to provide high diffusion constants without forming molecular complexes; reversible redox behaviour and low enough redox potential to minimise the oxidation of interfering compounds such as ascorbic and uric acid; good chemical stability and a low tendency to form other compounds. During long-term use of the sensor, however, the mediators may leach away and no longer remain close enough to the enzyme and electrode surface. The presence of dissolved oxygen may also compete with the mediator as a substrate for the enzyme, resulting not only in reduced sensor efficiency but also in a build-up of hydrogen peroxide. Innovations have been the development of electron-conducting redox hydrogels that electronically ‘wire’ the enzyme to an electrode [22] and the use of glucose dehydrogenase, instead of glucose oxidase, as the active enzyme (Heller *et al.*, US Patent 6,120,676; [23,24]). Glucose dehydrogenase (GDH) is a class of oxidoreductase having various cofactors, including NAD and FAD, catalysing the following reaction:



The two substrates of GDH are thus glucose and an electron accepting molecule, such as a quinone complex (Q). The reaction steps involved are:



In the glucose sensor described by Heller *et al.* [24] a redox hydrogel, consisting of a polymer bound osmium $\text{Os}^{2+}/\text{Os}^{3+}$ redox couple is used as the electron transfer mediator. The redox centres of the enzymes are tethered directly to the water-swollen crosslinked polymer network of the gel, and so there are no leachable components. Also, the negative formal potential of around -0.2 V (vs Ag/AgCl) for the $\text{Os}^{2+}/\text{Os}^{3+}$ couple means that the oxidised Os^{3+} is not reduced by the positively biased Ag/AgCl counter electrode. The current to be measured is therefore not internally short-circuited between the electrodes. These electrodes comprise screen-printed Ag/AgCl and carbon strips placed $50\ \mu\text{m}$ apart, and operate with a very small blood sample of just $300\ \text{nL}$. The latest version of this commercially available glucose sensor is subcutaneously implanted, and transmits the amperometric data every minute [23]. These sensors are fabricated using mass production techniques, such as screenprinting of electrodes, and have brought down costs to allow for single-use and disposability.

Advances are being made in the development of *third generation* glucose sensors. The aim is to achieve direct electron transfer between the enzyme and the electrode, without the need for natural or synthetic mediators. The major problem in achieving direct electron transfer from the FADH_2 of the GOx enzyme to an electrode is the fact that the FADH_2 group is buried too deeply within the protein structure for there to be efficient direct electron tunnelling to a planar electrode. The development of mesoporous electrode materials, such as nanostructured TiO_2 synthesised using a carbon nanotube template [25], offers an electrode surface that entraps the enzyme and permits direct electron transfer from the enzyme to the electrode. Glass carbon electrodes, modified by the addition of a mixture of carbon nanotubes and cadmium-telluride quantum dots, have also been found to promote the direct electrochemistry of GOx. Cyclic voltammograms, of the form shown in Figure 5.15 of Chapter 5, are observed with anodic and cathodic peak potentials (vs. Ag/AgCl) of -0.359 V and -0.385 V, respectively, with the small peak potential separation of $26\ \text{mV}$ indicating that fast electron transfer to the electrode was occurring [26]. Of significance was the finding that on adding glucose to the electrode buffer solution the reduction peak current decreased linearly with the concentration of glucose up to $0.7\ \text{mM}$, deviating from the linear relationship with higher glucose concentration in accordance with the expected kinetics of the GOx reaction [26].

In summary, first generation enzyme-based glucose sensors are disadvantaged by their oxygen dependency and their sensitivity to interference from other electroactive compounds found in blood. The second generation, now dominating the glucose sensor market, achieves oxygen independence using synthetic electron transfer mediators and minimise electroactive interferences by operating at a lower amperometric potential. The development of nano-scale, mesoporous, electrode surfaces may signal the emergence of third generation glucose sensors. However, all of these enzymatic glucose sensors have the common fault that they suffer from chemical instability. They are also constrained to the pH range 2–8, and a maximum operating temperature of $\sim 44\ ^\circ\text{C}$. The efforts and associated costs involved to ensure stability and long shelf-life of the immobilised enzyme in a glucose sensor are not trivial. Over many years there have therefore been efforts to identify a suitable nonenzymatic system in which glucose can be oxidised directly at a ‘bare’ electrode. The electro-oxidation processes for glucose at a number of electrodes, namely platinum, gold, nickel, copper and carbon, have been investigated using various voltammetric and amperometric techniques. Each electrode material presents its own problems associated with lack of selectivity and chemical fouling of the electrode, electroactive interference, the slow kinetics of glucose oxidation, and in some cases an inability to operate at physiological pH. This work, directed

Table 6.4 SWOT analysis of amperometric glucose sensors

Strengths	Weaknesses
<ul style="list-style-type: none"> • Dominates commercial market • Good dynamic response • Accurate results free from interference • Cheap disposable electrode • Compact, cheap, portable instrument 	<ul style="list-style-type: none"> • Invasive (finger-prick or subcutaneous) • Complicated enzyme immobilisation • Chemical instability of enzyme • Small dynamic range (saturation kinetics) • Limited pH and temperature range
Opportunities	Threats
<ul style="list-style-type: none"> • Increasing world markets resulting from increasing number of diabetics • Maintain future market share through development of fourth generation, non-enzymatic, glucose sensors 	<ul style="list-style-type: none"> • Development of noninvasive sensors that will take over its major market share.

towards the development of a *fourth generation* (nonenzymatic) class of glucose sensors has been reviewed by Toghill and Compton [27]. A SWOT analysis of amperometric glucose sensors is presented in Table 6.4.

Finally, increasing efforts are directed towards the sensing of glucose based on fluorescence intensity and lifetime. Perceived advantages include an improved sensitivity and the potential for non-invasive measurement using near-infrared radiation. Non-invasive glucose monitoring can be accomplished by measurement of cell auto-fluorescence due to NAD(P)H, and fluorescent markers of mitochondrial metabolism can signal changes in extracellular glucose concentration. Several receptors have been employed to detect glucose in fluorescence sensors, such as concanavalin A (Con A), enzymes such as hexokinase/glucokinase, bacterial glucose-binding protein, and boronic acid derivatives which bind the diol groups of sugars. Techniques being explored include fluorescence resonance energy transfer (FRET) between a fluorescent donor and an acceptor, either within a protein which undergoes glucose-induced changes in conformation or through competitive displacement; measurement of glucose-induced changes in the intrinsic fluorescence of enzymes. An introduction to these promising approaches has been given by Pickup *et al.* [28]. A SWOT analysis of fluorescence-based glucose sensors is presented in Table 6.5.

Table 6.5 SWOT analysis of fluorescent-based glucose sensors

Strengths	Weaknesses
<ul style="list-style-type: none"> • Noninvasive (skin penetrating infrared radiation) • Improved sensitivity 	<ul style="list-style-type: none"> • Purification and extraction of FRET donor-acceptor pair non-trivial and costly • Glucose receptors tend to precipitate over time
Opportunities	Threats
<ul style="list-style-type: none"> • Promising technology that will attract major investments for development • Dominate world market of glucose sensors 	<ul style="list-style-type: none"> • May not proceed beyond research and development stage • Complexity and cost of technology may preclude competitive entry into the market

6.12 Biocompatibility of Implantable Sensors

Biosensors are increasingly being developed for use as implantable medical devices, examples of which are the glucose sensor described in this chapter, or the interfacing of electronic probes with nerves and brain tissue described in Chapter 8. Some damage to tissue will result from the implantation of a sensor, and a reduction of the sensor's performance can result from biofouling and fibrous encapsulation associated with the inflammatory and immune responses involved in the healing of this wound. A basic understanding of the different stages of the wound healing process can assist efforts to manage or control the adverse effects of this to a sensor's performance. The regulations and methods for testing the safety of implanted sensors should also be understood.

6.12.1 Progression of Wound Healing

The steps involved in the progression of wound healing commence with the initial tissue damage:

- **Injury:** The level of tissue damage will vary. Some implanted sensors have a geometrical construction that allows them to be implanted through a needle, while larger devices such as pacemakers will require more invasive surgery. Generally any implant will lead to damage to vascular tissue and blood vessels. Immediately after injury the blood vessels will constrict to limit the blood flow and to allow a clot to form, in a process called haemostasis. The blood vessels will then dilate so as to allow the cells that control inflammation to migrate through the walls of the now porous blood vessel. The mixture of fluid, blood, inflammatory response cells and immune factors that will pool at the injury site is known as exudate.
- **Blood/Material Interactions:** The next step in the healing process is the formation of clots resulting from an interaction between blood platelet cells and proteins such as fibrin and fibronectin which interact to form a plug to prevent further bleeding. The proteins will also coat the implanted device and this can have an immediate impact on the sensitivity of the sensor by preventing the transport of analyte to the sensitive area. Platelets will also release chemicals that encourage the migration of inflammatory-response cells to the wound and keep nearby blood vessels porous so that these cells can reach the wound site.
- **Provisional matrix formation:** The material of the clot is usually known as the provisional matrix, because it forms a substrate for other cells to interact with and heal the wound. This matrix will eventually be replaced with granulation tissue. This clot formation can affect the performance of the implanted sensor in a process known as protein biofouling, and especially so if the pores of the sensor membrane becomes blocked with material which impedes analyte transport.
- **Immune cell recruitment:** Chemical signals released during wounding and clot formation attract immune- and inflammatory-response cells to the site of the tissue injury. Neutrophils, known as polymorphonuclear leukocytes (PNMs) or neutrophil granulocytes are the most abundant type of inflammatory-response cells found in the blood. They are short-lived, existing for no more than one or two days in the blood and act as phagocytes, taking the form of mobile cells that are attracted by chemotaxis to and then engulf bacteria and other pathogens. Neutrophils also release other chemicals involved in the inflammatory response to attract and activate other cells that are required for healing.

Macrophages and lymphocytes are involved in the secondary, chronic, stage of inflammation which is a normal part of healing so long as this process does not persist for too long. Lymphocytes are active in immune-response reactions as a result of antibody production. If the implant is sterile and fabricated using materials that do not cause adverse immune reactions, then lymphocytes will not be significantly active at the implant site. Macrophages will be active, however, because they act to remove dead cells, old neutrophils and other pathogens. These cells do not migrate to the site but are created by maturing monocytes, and other mobile white blood cells that have been stored in the lymph glands ready to respond to tissue injuries in the body.

- **Chronic Inflammation:** A certain amount of inflammation is a standard part of the healing process, but if the chronic stage persists for too long it can lead to on-going damage of healthy tissue. Generally this is only found in the elderly, or those with a compromised immune system, and can result in ulceration and other chronic infections. Foreign bodies can also cause chronic inflammation through a number of different mechanisms, which can also vary from person to person. An important consideration in this respect is the nature of the physical properties of the interface between the implanted device and the tissue. A hard, unyielding interface or a dense implant that can move within the wound site can lead to constant recurring injury and excitation of inflammatory responses.
- **Granulation Tissue Formation:** Within a few days of injury, as the inflammatory response hopefully recedes, new cells begin to appear and proliferate at the implant site. These include fibroblasts, which are responsible for the production of the extracellular protein material such as collagen, and endothelial cells which form the vascular system. These cells form granulation tissue to replace the provisional matrix, which is disaggregated and consumed by macrophages. The granulation tissue forms a fibrous structure that is a precursor to the extracellular matrix and new capillary blood vessels.
- **Angiogenesis:** The cells involved in granulation tissue formation require a supply of blood and so new capillaries are formed by vascular endothelial cells in a process known as angiogenesis. This blood supply gives the granulation tissue its characteristic red colour.
- **Foreign Body Reaction:** Phagocytes such as neutrophils and macrophages are too small to be able to engulf the average implanted medical device and so they may undergo a reaction known as frustrated phagocytosis, where instead they produce enzymes and other chemicals designed to degrade and destroy a foreign body. Macrophages will also associate to form large conglomerations known as *foreign body giant cells* in order to achieve the same objective.
- **Fibrous Capsule Formation:** Macrophages and foreign body giant cells can exist at the surface of an implant for the whole duration of its time there. Whether or not they remain biologically active during this time is not known with certainty, but this probably depends on the properties of the implant surface. To shield the remaining tissues in the body from the implant and limit the extent of the inflammatory response, fibroblasts will build up a capsule of fibrous collagen around the device. This material will typically lack capillaries and be avascular, and so the sensor can effectively be isolated from the very environment it is designed to monitor. Fibrous capsule formation appears to be minimal on porous surfaces, compared to flat, smooth and nonporous surfaces, although the pore size is an important factor. If too small, the surface may appear too smooth and hinder cell adherence but lead to increased growth of a fibrous capsule. If too large this can encourage the ingrowth of inflammatory-response cells, such as macrophages, which can damage the

implant. Pores sizes of around 1 μm diameter appear to inhibit fibrous tissue growth and attachment of inflammatory-response cells, but encourage advantageous vascularisation around the sensor.

6.12.2 Impact of Wound Healing on Implanted Sensors

The initial impact is that of protein biofouling, which might not totally destroy the sensor's functioning but can inhibit diffusion of the analyte to the sensing element and also reduce sensitivity. Inflammatory cells which are frustrated in attempts to phagocytose an implant may produce chemicals that can degrade parts of the implant, particularly fragile sensor membranes. These chemicals might also act as sensing interferants. The most common endpoint to the foreign body response is the formation of a fibrous tissue capsule around the device, and because this coating does not contain many blood vessels may act as a barrier between the sensor and the rest of the tissue. On the other hand, the new blood capillaries formed during the granulation tissue formation may actually help transport analytes towards the implanted sensor. These issues, especially fibrous tissue encapsulation, will also be important for implanted actuators such as drug delivery devices or electrodes designed to interact with nerves. If chronic inflammation persists and satisfactory wound healing does not occur, the only remaining option is to remove the implanted device.

6.12.3 Controlling the Tissue Response to Sensor Implantation

Protein biofouling can often be minimised by treating the outer surface of the sensor device to make it more hydrophilic [29]. This can encourage the formation of a bound layer of water to inhibit protein attachment and cell attachment. A commonly used material to make surfaces hydrophilic is polyethyleneglycol (PEG) which is also known as polyethyleneoxide (PEO). However, the attachment of certain proteins to the sensor surface may enhance the actions of inflammatory-response cells or promote angiogenesis. This can be promoted using hydrogels, a common material used to make contact lenses or to produce scaffolds for tissue engineering in regenerative medicine. Some hydrogels are synthetic and others are based on natural materials such as cellulose. They can be modified with chemicals that selectively bind to endothelial cells but inhibit fibroblast attachment, for example. Growth factors can also be incorporated onto the hydrogel surface to control tissue responses to the implant. A chemically modified hydrogel coating could thus be added to an initial coating of PEG.

Implant surfaces could also be engineered to slowly release drugs, including growth factors, perhaps from biodegradable microspheres embedded in a hydrogel matrix or also from co-implants alongside the sensor device. Two promising candidates are the vascular endothelial growth factor (VEGF) to encourage angiogenesis, and dexamethasone, which can reduce both inflammatory responses and fibrosis. VEGF could also have a negative effect of encouraging fibrous capsule formation. Nitric oxide, which occurs naturally in tissues and controls vasodilation, has also been considered as a drug to be used for controlled release from an implant surface, particularly for sensors intended to work inside blood vessels. Nitro glycerin is used for the treatment of angina because it is converted to nitric oxide within the body. Nitric oxide is also present as a signalling molecule in the wound healing response and when used in conjunction with implants has been shown to promote angiogenesis and reduce fibrosis. However, it is an oxidising agent and so may interfere with some sensors, such as an oxygen sensor, for example.

6.12.4 Regulations for and Testing of Implantable Medical Devices

The European Medicines Agency (EMA) and the British Medicines and Healthcare products Regulatory Agency (MHRA) are two agencies having responsibility for standards of safety, quality and performance of active implantable devices. They publish numerous documents which can be accessed on the internet. The definition of an *active medical device* in the European directive is: '*any medical device relying for its functioning on a source of electrical energy or any source of power other than that directly generated by the human body or gravity*'. The MHRA provides a list of standards that apply to implanted and *in-vitro* medical devices.

The International Standards Organisation has issued various standards for the biological evaluation of medical devices. In particular, those listed under ISO 10993 apply to active implantable medical devices, and mainly detail the tests required before a device can be considered for clinical use. Medical devices covered by the ISO 10993 standard are characterised by the level of tissue contact, proceeding from contact with skin, mucus membranes and compromised surfaces such as wounds, to devices that communicate between the inside and outside of the body. An important test is to check for the lack of acute cytotoxicity of the materials used in an implant. This involves culturing cells, such as fibroblasts, in contact with a sample of the material or in a fluid containing substances that have leached from the implant material. If the cells are adversely affected then the materials used may have to be changed. In many cases implants do contain toxic materials, such as in pacemaker batteries, but these have to be well encapsulated behind a hermetic seal to prevent their contact with tissue. One method of testing the inflammatory response to materials used in implants is to place the material inside a small wire mesh cage, which is then implanted under the skin of a test animal model. The levels of inflammatory-control cells within the cage are measured and compared with those from an empty control cage. If the material causes a greater inflammatory response it suggests that its biocompatibility could be a problem. When implanted sensors are tested *in vivo* they are often recovered after the experiment so that their function can be checked for degradation, such as that arising from the foreign body reaction or chemicals produced by inflammatory-response cells. It is also important to measure the thickness of the fibrous capsule around the explanted device and to perform standard histopathological tests on sections of tissue around the device regarding the type of induced cell response. An excellent review of testing biological responses induced by implant materials has been written by Anderson [30].

References

- [1] Uchiyama, S., Hasebe, Y. and Tanaka, M. (1997) L-Ascorbate sensor with polypyrrole-coated carbon felt membrane electropolymerized in a cucumber juice solution, *Electroanalysis*, **9**, 176–178.
- [2] Curulli, A., Kelly, S. O'Sullivan, C. et al. (1998) A new interference-free lysine biosensor using a non-conducting polymer film. *Biosensors & Bioelectronics*, **13**, 1245–1250.
- [3] Schöning, M. and Poghossian, A. (2002) Recent advances in biologically sensitive field-effect transistors (bio-FETs), *Analyst*, **127**, 1137–1151.
- [4] Gehring, A.G., Patterson, D.L. and Tu, S.I. (1998) Use of a light-addressable potentiometric sensor for the detection of *Escherichia coli* 0157:H17, *Analytical Biochemistry*, **258** (2): 293–298.
- [5] Chouteaua, C., Dzyadevychc, S., Durrieua, C. and Chovelonb, J.M. (2005) A bi-enzymatic whole cell conductometric biosensor for heavy metal ions and pesticides detection in water samples. *Biosensors & Bioelectronics*, **21**, 273–281.

- [6] Mikkelsen, S.R. and Rechnitz, G.A. (1989) Conductometric transducers for enzyme-based biosensors. *Analytical Chemistry*, **61**, 1737–1742.
- [7] Muhammad, Z. and Alocija, E.C. (2003) A conductometric biosensor for biosecurity. *Biosensors & Bioelectronics*, **18**, 813–819.
- [8] Weizmann, Y., Chenoweth, D.M. and Swager, T.M. (2011) DNA-CNT nanowire networks for DNA detection. *Journal of the American Chemical Society*, **133**, 33238–33241.
- [9] Park, B.W., Yoon, D.Y. and Kim, D.S. (2011) Formation and modification of a binary self-assembled monolayer on a nano-structured gold electrode and its structural characterization by electrochemical impedance spectroscopy. *Journal of Electroanalytical Chemistry*, **661**, 329–335.
- [10] Wolfbeis, O.S., Leiner, M.J.P. and Posch, H.E. (1986) A new sensing material for optical oxygen measurement, with the indicator embedded in an aqueous phase. *Microchimica Acta*, **90** (5–6), 359–366.
- [11] Marquette, C.A., Degiuli, A. and Blum, L.J. (2003) Electrochemiluminescent biosensors array for the concomitant detection of choline, glucose, glutamate, lactate, lysine and urate. *Biosensors & Bioelectronics*, **19**, 433–439.
- [12] Ko, S. and Grant, S.A. (2006) A novel FRET-based optical biosensor for rapid detection of *Salmonella typhimurium*. *Biosensors & Bioelectronics*, **21**, 1283–1290.
- [13] Ueberfeld, J. and Walt, D.R. (2004) Reversible ratiometric probe for quantitative DNA measurements. *Analytical Chemistry*, **76**, 947–952.
- [14] Peter, J., Hutter, W., Stollnberger, W. and Hampel, W. (1996) Detection of chlorinated and brominated hydrocarbons by an ion sensitive whole cell biosensor. *Biosensors & Bioelectronics*, **11**, 1215–1219, 15.
- [15] Cornell, B.A., Braach-Maksvytis, V.L.B. King, L.G. et al. (1997) A biosensor that uses ion-channel switches. *Nature*, **387** (6633), 580–583.
- [16] Song, X. and Swanson, B.I. (1999) Direct, ultrasensitive, and selective optical detection of protein toxins using multivalent interactions. *Analytical Chemistry*, **71**, 2097–2107.
- [17] Wang, J. (2008) Electrochemical glucose biosensors. *Chemical Reviews*, **108**, 814–825.
- [18] Clark, L.C. and Lyons, C. (1962) Electrode systems for continuous monitoring in cardiovascular surgery. *Annals of the New York Academy of Sciences*, **102**, 29–45.
- [19] Updike, S.J. and Hicks, G.P. (1967) The enzyme electrode. *Nature*, **214**, 986–988.
- [20] Guilbault, G.G. and Lubrano, G.J. (1973) An enzyme electrode for the amperometric determination of glucose. *Analytica Chimica Acta*, **64**, 439–455.
- [21] Frost, M. and Meyerhoff, M.E. (2006) In vivo chemical sensors: Tackling biocompatibility. *Analytical Chemistry*, **78** (21), 7370–7377.
- [22] Mao, F., Mano, N. and Heller, A. (2003) Long tethers binding redox centers to polymer backbone enhance electron transport in enzyme ‘wiring’ hydrogels. *Journal of the American Chemical Society*, **125** (16), 4951–4957.
- [23] Heller, A. and Feldman, B. (2010) Electrochemistry in diabetes management. *Accounts of Chemical Research*, **43** (7), 963–973.
- [24] Heller, A., Feldman, B.J., Say, J. and Vreeke, M.S. (Sept. 19 2000) *Method of using a small volume in vitro analyte sensor*, US Patent 6, 120, 676.
- [25] Bao, S.J., Li, C.M. Zang, J.F. et al. (2008) New nanostructured TiO₂ for direct electrochemistry and glucose sensor applications. *Advanced Functional Materials*, **18** (4), 591–599.
- [26] Liu, Q., Lu, X., Li, J. et al. (2007) Direct electrochemistry of glucose oxidase and electrochemical biosensing of glucose on quantum dots/carbon nanotubes electrodes. *Biosensors Bioelectronics*, **22**, 3203–3209.
- [27] Toghill, K.E. and Compton, R.G. (2010) Electrochemical non-enzymatic glucose sensors: A perspective and an evaluation. *International Journal of Electrochemical Science*, **5**, 1246–1301.
- [28] Pickup, J.C., Hussainia, F., Evans, N.D. et al. (2005) Fluorescence-based glucose sensors. *Biosensors and Bioelectronics*, **20**, 2555–2565.
- [29] Wisniewski, N. and Reichert, M. (2000) Methods for reducing biosensor membrane biofouling. *Colloids and Surfaces B*, **18**, 197–219.
- [30] Anderson, J.M. (2001) Biological responses to materials. *Annual Review of Materials Research*, **31** (1), 81–110.

Further Readings

- Borisov, S.M. and Wolfbeis, O.S. (2008) Optical biosensors. *Chemical Reviews*, **108**, 423–461.
- Harper, A. and Anderson, M.R. (2010) Electrochemical glucose sensors – developments using electrostatic assembly and carbon nanotubes for biosensor construction. *Sensors*, **10**, 8248–8274.

- Heller, A. and Feldman, B. (2008) Electrochemical glucose sensors and their applications in diabetes management. *Chemical Reviews*, **108**, 2482–2505.
- Homola, J. (2008) Surface plasmon resonance sensors for detection of chemical and biological species. *Chemical Reviews*, **108**, 462–493.
- Lazcka, O., Del Campo, F.J. and Munoz, F.X. (2007) Pathogen detection: A perspective of traditional methods and biosensors. *Biosensors & Bioelectronics*, **22**, 1205–1217.
- Lisdat, F. and Schafer, D. (2008) The use of electrochemical impedance spectroscopy for biosensing. *Analytical & Bioanalytical Chemistry*, **391**, 1555–1567.
- Liu, J., Cao, Z. and Lu, Y. (2009) Functional nucleic acid sensors. *Chemical Reviews*, **109**, 1948–9998.
- Pejcic, B. and De Marco, R. (2006) Impedance spectroscopy: Over 35 years of electrochemical sensor optimization. *Electrochimica Acta*, **51**, 6217–6229.
- Rahman, M.M., Ahammad, A.J.S. Jin, J.H. et al. (2010) A comprehensive review of glucose biosensors based on nanostructured metal-oxides. *Sensors*, **10**, 4855–4886.
- Sassolas, A., Leca-Bouvier, B.D. and Blum, L.J. (2008) DNA biosensors and microarrays. *Chemical Reviews*, **108**, 109–139.

7

Basic Sensor Instrumentation and Electrochemical Sensor Interfaces

7.1 Chapter Overview

This chapter begins by reviewing concepts of sensor and transducer theory before exploring the basic methods for amplifying electrical outputs from sensors. The building block of these amplifier circuits is the operational amplifier and this chapter will review the characteristics of the ideal and real op-amp and provide a tool box of standard circuits.

The second part of the chapter will focus on instrumentation for electrochemical sensors, beginning with an equivalent circuit model of the standard three-electrode cell before looking at the potentiostat which is the basic instrument used in electrochemistry. Many biosensors use ac electrochemical impedance based measurements and these techniques will also be explored.

This chapter will conclude with an examination of solid-state chemical sensors based on field effect transistor (FET) technology, and will include revision of the basics of metal-oxide-semiconductor (MOS) transistors. The most commonly used FET based sensor is the ion sensitive field effect transistor (ISFET) and this will be studied in detail. Instrumentation for measuring pH with the ISFET will be discussed, as will the variety of other biosensors that use the same mode of operation.

After reading this chapter readers will gain a refreshed or new understanding of:

- (1) the important characteristics of a sensor;
- (2) the use of operational amplifiers in sensor applications;
- (3) electrochemical sensor instrumentation, such as the potentiostat;
- (4) electrochemical impedance sensing including equivalent circuit;
- (5) the application of the ion sensitive field effect transistor.

7.2 Transducer Basics

7.2.1 *Transducers*

As discussed in Chapter 6, the basic definition of a transducer is a device which converts energy from one form to another. One example might be a motor which converts electrical energy to kinetic energy. Two subsets of transducer with specific properties are sensors and actuators. An important difference between a transducer like a motor (or its opposite a generator) and a sensor is in terms of which characteristics are most important. In a sensor the level of detection and, in most cases, the linearity of response are much more important than the efficiency of energy conversion. This might also be the case with actuators; for example, the dial on an old style analogue multimeter uses a galvanometer transducer to convert a current into the accurate movement of a needle. Probably the most useful way of defining the difference between the sensors and actuators we are interested here and the wide range of other transducers is that they transfer information rather than simply converting energy.

7.2.2 *Sensors*

Although sensors can potentially convert information from one form to any other form, this discussion will concentrate on those that produce an electrical signal in response to some physical stimulus. These can include sensors with one or more intermediate steps of transduction between the physical input to the sensor and the electrical output. This is often the case in biosensors as the biological component of the sensor is unlikely to have an output which is a direct electrical signal.

7.2.3 *Actuators*

An actuator can be thought of as the opposite of a sensor, in that it produces some physical output that is dependent on an electrical input. This is often mechanical in nature but it is not necessary to limit the definition in that way. One important use of actuators is to provide feedback in a control system, for example by opening a valve based on the input from a sensor. The development of ‘bio-actuators’ is significantly less advanced than that of biosensors but examples could include the use of muscle cells to provide mechanical actuation. One of the most desirable uses might be to combine a glucose sensor with a bio-actuator which controllably releases or produces insulin to make a feedback system for the treatment of the symptoms of diabetes and act as an ‘artificial pancreas’.

7.2.4 *Transduction in Biosensors*

Biosensors are commonly tandem sensors consisting of two or more sensing stages. The first stage is the biological element which employs some sort of molecular recognition. The selectivity of this stage may be defined by the sensor itself but there is often an additional coating which selectively allows the desired analyte to diffuse through to the sensor. This can also serve to protect the biological element from the environment and extend the lifetime of the sensor. The biological sensor itself usually does not transduce the parameter being sensed into a useful electrical signal. This is left to the second stage, and using the example of the glucose sensor described in Chapter 6, the most common type uses an electrochemical

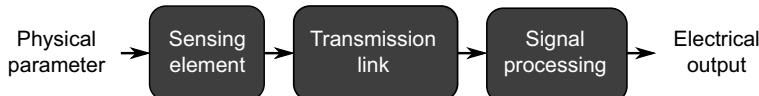


Figure 7.1 Block diagram of a traditional sensor architecture.

sensor to sense either the depletion of oxygen or the production of hydrogen peroxide caused by the enzymatic reaction.

7.2.5 Smart Sensors

In a traditional sensor architecture (Figure 7.1) the front end is the sensing or transduction element which produces an electrical signal that depends on the measured parameter. This is sent over a transmission link to the signal processing section of the system which could include amplification, analogue to digital conversion or any number of other processing schemes. This can then be transferred to a display for readout, data storage or some other further processing. The advantages of this setup are that the sensor can be quite cheap and is separated from the data acquisition and processing which can also be inexpensive, and standardised, able to be used with many different types of sensor. However, there are disadvantages, the un-amplified signal from the sensor will typically be quite weak and prone to the introduction of noise during transmission. The sensor itself is dumb and inflexible, with a response that is fixed and cannot be adjusted.

A ‘Smart’ sensor system (Figure 7.2) adds some form of signal processing, feedback or control to the front end of the device, integrated with the sensor itself. This could be anything from simple amplification of the sensor output to digitisation or more complicated processing of the signal. The point is that this happens as close as possible to the sensor itself, before transmission of the resulting signal. In the case of simple amplification this could reduce the effects of noise introduced in the transmission link, while if the sensor output is digitised you could effectively eliminate analogue noise effects after that stage. In addition to signal conditioning or processing in the smart sensor the integration with electronics could also allow the sensor to become more flexible and change its characteristics in response to the environment or to counteract drift. The main disadvantage is to make the sensor itself more complex and expensive, it could also have the opposite effect in terms of flexibility by fixing the attributes of the initial signal amplification at an early stage.

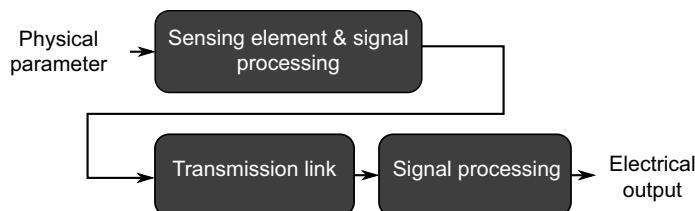


Figure 7.2 Block diagram of a *smart* sensor architecture.

7.2.6 Passive vs. Active Sensors

There are two main classes of sensor, that operate purely on the transduction of physical parameters into electrical signals. The first is a self-generating or active transducer where the electrical output signal is directly generated from the energy of the physical input. One example is a photovoltaic cell which produces a voltage that is proportional to the intensity of the light falling on it.

The other type of sensor is a passive or modulating transducer, which requires a separate power source to operate. In this type the sensing element simply modulates the flow of electrical energy supplied by the source. For example, in a passive sensor for light intensity the photovoltaic cell is replaced by a photoconductive cell or light dependent resistor. Other examples of modulating sensors include the piezoresistive strain gauge and the thermistor. Impedance based modulation of an electrical signal can be capacitive or inductive in nature as well as being purely resistive.

7.3 Sensor Amplification

Regardless of the parameter that a sensor is designed to measure or the sensing modality used, one common attribute of most sensors is that they do not produce large signals at the output. Therefore, the first stage in any sensor interface will normally be an amplifier to increase the signal.

7.3.1 Equivalent Circuits

Hopefully the idea of representing or modelling a sensor by an equivalent circuit will be quite familiar to readers of this book. The sensor is basically a source of an electrical signal, typically low voltage and/or low current, with an output resistance which can be quite significant. Obviously we could exchange the Thévenin equivalent circuit shown in Figure 7.3, for a Norton equivalent consisting of a current source and parallel resistance. In this circuit, V_S is small and R_S is large which will obviously have implications on the next stage in the sensor system, which would typically be an amplifier.

The equivalent circuit of an amplifier should also be a familiar concept. Figure 7.4 is a simple model of a voltage amplifier with voltage gain G , an input resistance R_{in} and an output resistance R_{out} . For the best performance R_{in} should be as large as possible while R_{out}

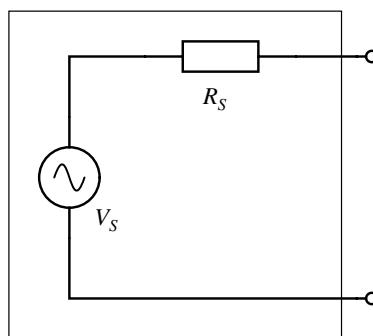


Figure 7.3 Equivalent circuit of a sensor system with a voltage output.

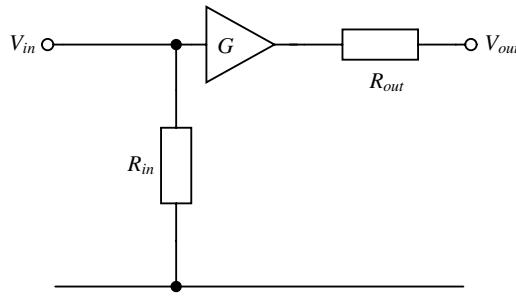


Figure 7.4 Equivalent circuit of a voltage amplifier.

should be as small as possible. There are other ways to represent the amplifier, and Figure 7.5 shows more of a black box model where the input resistance is now simply a load on the input while the output side becomes a Thévenin equivalent circuit with a voltage source outputting the input voltage, V_{in} , multiplied by the amplifier voltage gain, G , and the output resistance R_{out} .

If this amplifier model is combined with the sensor equivalent circuit in Figure 7.3, the output resistance of the sensor forms a voltage divider with the input resistance of the amplifier, giving an input voltage, V_{in} , which is some fraction of the voltage output by the sensor:

$$V_{in} = V_s \frac{R_{in}}{R_s + R_{in}}. \quad (7.1)$$

In addition there will be a second voltage divider on the output of the amplifier if it is driving a resistive load R_L . Then the actual output voltage across the load will be:

$$V_{out} = \frac{G \left(V_s \frac{R_{in}}{R_s + R_{in}} \right) R_L}{R_{out} + R_L}. \quad (7.2)$$

It is clear that to get the best output voltage we need the amplifier input resistance to be much larger than the sensor output resistance (i.e. $R_{in} \gg R_s$) and the amplifier output resistance to be smaller than the load ($R_{out} \ll R_L$). If that is the case we can approximate the output $V_{out} = GV_s$.

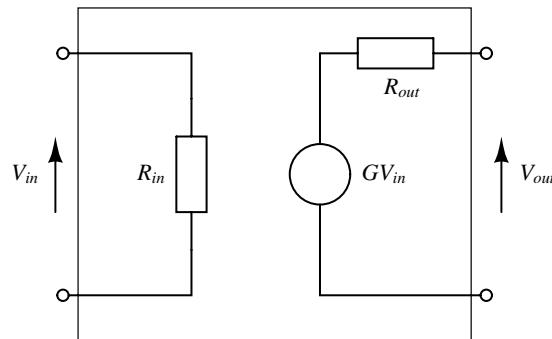


Figure 7.5 Alternative equivalent circuit of a voltage amplifier.

7.4 The Operational Amplifier

7.4.1 Op-Amp Basics

The operational amplifier (or op-amp) is probably one of the most widely used components in analogue design due to its flexibility and general utility. It is an integrated circuit meaning that the whole circuit, consisting of many transistors, resistors and capacitors is integrated onto a single silicon chip. The circuit symbol is shown in Figure 7.6. It is a differential amplifier where V_{out} is proportional to the difference between the two inputs, multiplied by a large number, A , usually referred to as the open loop gain:

$$V_{out} = A(V_+ - V_-). \quad (7.3)$$

There are a couple of useful features of the ideal op-amp, as a circuit element, that drove their development. Firstly, it has no current into the inputs and secondly the output can drive any current that is required. Clearly a physical implementation of an op-amp device will not actually be capable of this, but in practice the input current for a good op-amp will be in the pico ampere range while the output resistance will be very low, especially compared to the sensors we are considering. The op-amp requires power supply connections but these are rarely shown in circuit diagrams. Unfortunately, this means it is easy to forget them, which is a problem as the output cannot exceed these values and is usually slightly less. Equally, the inputs should not go beyond the power supply levels either, but what happens if you do try to beat these limits? Applying even a relatively small differential voltage to the inputs of an op-amp without any negative feedback could theoretically give an output voltage of many hundreds or thousands of volts. This is physically impossible and on an old style op-amp such as the 741 the output will only go to within 2 V of the voltage supply. Therefore, Equation (7.3) for V_{out} in terms of the open loop gain and the input voltages only applies if V_{out} is not saturated, that is it is in the range defined by the supply voltages.

If we assume V_{out} is not saturated then it must be much smaller than the open loop gain A and so the difference between the input voltages must be very small. This is one of the so-called Golden Rules of op-amp circuit design, that the inputs, V_+ and V_- are effectively equal. Taking this assumption that the inputs are at equal voltages what will happen to V_+ if the V_- input is connected to ground? The answer is that it becomes what is commonly referred to as a ‘virtual earth’. The next question is what does this mean for the input

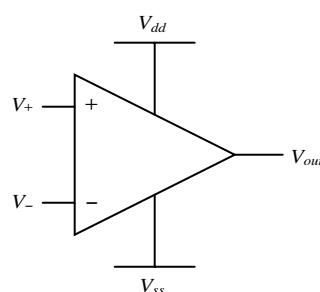


Figure 7.6 Circuit schematic symbol for an operational amplifier.

resistance of the op-amp? To answer this we should consider a model of an op-amp where the differential input voltage ($V_+ - V_-$) that is being amplified is applied across some input resistance. If the two inputs always have the same value then the current into or out of the input terminals will be zero and so the effective input resistance will be infinite. However, this is something of a circular argument and, although this gives us another ‘Golden Rule’, it only applies in certain circumstances.

To summarise, there are three Golden Rules for op-amp circuit design and analysis:

- (1) no current flows into the inputs, V_+ and V_- ;
- (2) the input voltages are always equal, that is $V_+ = V_-$;
- (3) the op-amp output can drive any current that is required.

In reality these only apply when external components are used to provide negative feedback to the op-amp. What these rules do allow you to do is to analyse the design of any standard op-amp circuit with help from Ohm’s law and nodal analysis.

7.4.2 Non-inverting Op-Amp Circuit

There are two basic designs for an op-amp circuit, both with negative feedback, meaning there is a connection between the op-amp output and the inverting input. In the non-inverting amplifier, Figure 7.7, V_+ is the input voltage V_{in} . Using the second golden rule, V_- is also equal to V_{in} . In order to determine V_{out} we analyse the currents into the V_- node assuming that there is no current into the op-amp inputs. Following this through gives the transfer function for the non-inverting amplifier:

$$\frac{V_{out}}{V_{in}} = \frac{R_1 + R_2}{R_1}. \quad (7.4)$$

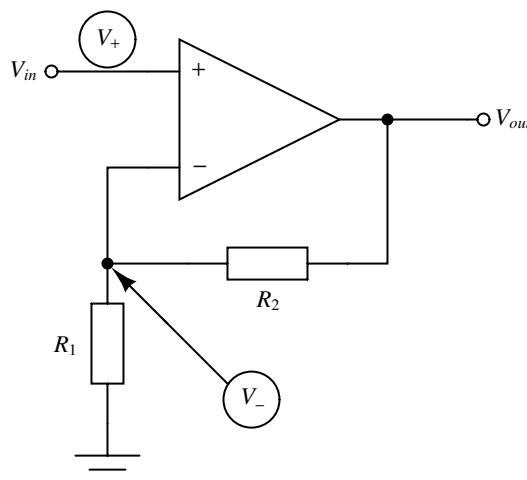


Figure 7.7 Non-inverting op-amp circuit.

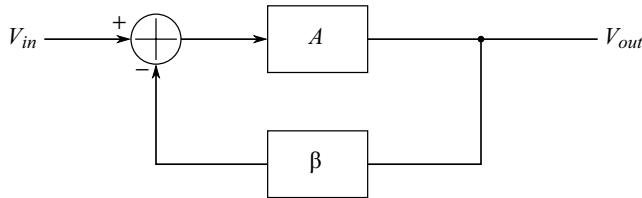


Figure 7.8 Feedback diagram for a non-inverting op-amp circuit.

The feedback diagram shown in Figure 7.8 is an alternative representation of this amplifier. The differential stage is represented by an adder where the minus sign on one input represents the subtraction of the inverting input from the non-inverting input. The open-loop gain of the op-amp is A while the gain (attenuation) of the feedback loop, which forms a potential divider, is β . Then the transfer function is:

$$\frac{V_{out}}{V_{in}} = \frac{A}{1 + A\beta} = \frac{A}{1 + A \frac{R_1}{R_1 + R_2}}. \quad (7.5)$$

The transfer function of this feedback system becomes the same as equation (7.4) if the loop-gain $A\beta \gg 1$, which is usually the case when the open-loop op-amp gain is very large. The minus sign associated with negative feedback is equivalent to 180° phase shift in the feedback loop. If there is no additional phase shift from the loop gain then the system will be very stable and there is little chance of getting positive feedback. However, the loop gain will typically have some frequency dependence and at some frequency it will also have a phase shift of 180° and the total phase around the loop will be 360° . There will be positive feedback and oscillation if the magnitude of the loop gain is ≥ 1 at this point.

7.4.3 Buffer Amplifier Circuit

If we take a non-inverting amplifier and connect the output directly to the input it becomes a buffer amplifier (Figure 7.9) or voltage follower. We are effectively shorting out R_2 while R_1 can be ignored thanks to golden rule 3, namely that the output can drive any current. In the buffer, V_{out} is equal to V_{in} but the combination of the high input impedance and low output

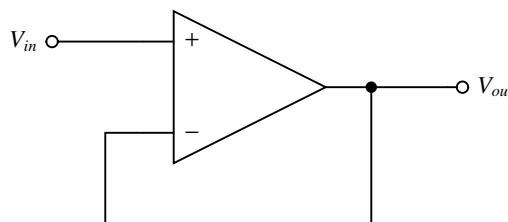


Figure 7.9 Buffer amplifier circuit.

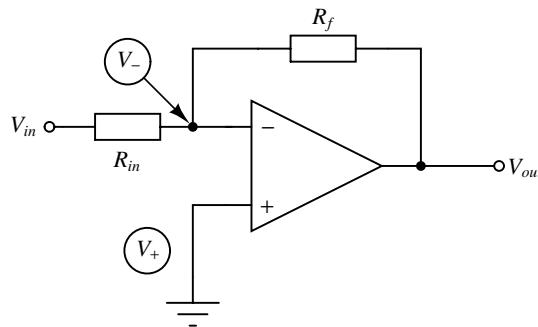


Figure 7.10 Inverting amplifier circuit.

impedance makes it a very useful element. This circuit can be used in a sensor amplifier system to ‘buffer’ the high output impedance of the sensor stage from the input stage of the amplifier.

7.4.4 Inverting Op-Amp Circuit

The second main class of op-amp circuit is the inverting amplifier (Figure 7.10). Again we have negative feedback with a resistor connecting the output to the inverting input. The V_+ connection is grounded so that, following Golden Rule 2, the V_- node is also grounded. This arrangement is usually referred to as a virtual earth/ground. As before, we can analyse the currents into the V_- node and derive Equation (7.6) for the transfer function of this amplifier. The minus sign in this equation indicates that it will invert the input signal.

$$\frac{V_{out}}{V_{in}} = -\frac{R_f}{R_{in}}. \quad (7.6)$$

The input resistance of the inverting amplifier is R_{in} , and is therefore finite and likely to be relatively low if the gain V_{out}/V_{in} is high. A possible use of the buffer circuit described in Section 7.4.3 is to precede an inverting amplifier and provide a low impedance output to drive it. It may also be ideal for a modulating impedance sensor where $R_{sensor} = R_{in}$. The summing amplifier is a useful variation on the inverting amplifier which has multiple inputs. The output is a weighted sum of the input voltages, where the weights are effectively the feedback resistance divided by the input resistor for each input.

7.4.5 Differential Amplifier Circuit

Although the op-amp is itself a differential amplifier, in practice it cannot be used in a differential mode as anything other than a comparator without applying feedback. The differential amplifier (Figure 7.11) uses 4 external resistors, and if $R_1 = R_2 = R_3 = R_4$ the output voltage will be the difference between the two inputs:

$$V_{out} = V_2 - V_1. \quad (7.7)$$

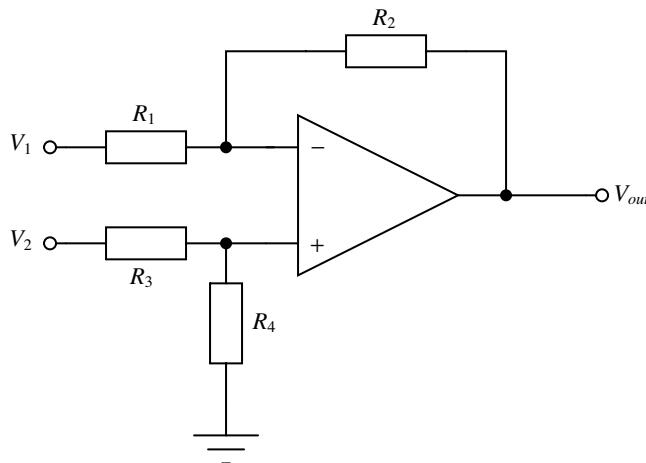


Figure 7.11 Differential amplifier circuit.

Gain can be obtained by making R_2 and R_4 larger than R_1 and R_3 . If we make sure that $R_2 = R_4$ and $R_1 = R_3$ then:

$$V_{out} = \alpha(V_2 - V_1) \quad (7.8)$$

$$\alpha = \frac{R_2}{R_1} = \frac{R_4}{R_3}. \quad (7.9)$$

Having equal values for the resistor ratios is very important otherwise the gain will be unequal between the two inputs. This brings us to some very important characteristics of op-amps and the circuits they are used in, which are the common mode gain and the Common Mode Rejection Ratio (CMRR). The ideal op-amp will have zero common mode gain meaning that if the inputs are connected together the output will be 0 V regardless of the common mode input voltage. Real op-amps will have a very high CMRR, even the old 741 has a value of around 70 dB. Using negative feedback will tend to correct for common mode errors but there could still be transient problems when the input changes quickly. The differential amplifier will have its own CMRR which will depend on the accuracy of the resistor values.

7.4.6 Current Follower Amplifier

Using the golden rules described in Section 7.4.1, the inverting input of the current follower (Figure 7.12) is at the same voltage (earth) as the non-inverting input. Assuming there is no current into the input all of the current will flow through the resistor. Therefore the output voltage will be proportional to the input current:

$$V_{out} = -IR. \quad (7.10)$$

As the transfer function, V_{out}/I_{in} , has units of resistance/impedance it is often known as a transimpedance amplifier. A common use of this circuit is to amplify the output of a photo-diode optical detector, which looks like a current source with high resistance and a

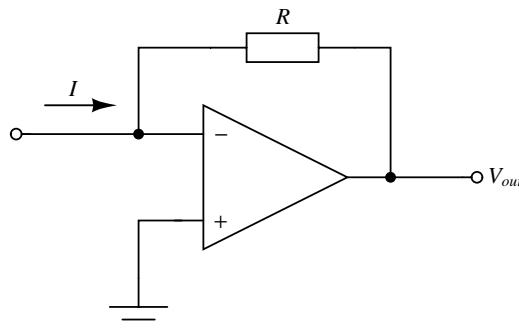


Figure 7.12 Current follower circuit.

significant capacitance. Capacitive impedance on the input combines with the feedback resistor to form a low pass filter that can induce instability. To compensate for this a small capacitance can be placed in parallel with the feedback resistor.

7.5 Limitations of Operational Amplifiers

7.5.1 Resistor Values

The resistors used with an ideal op-amp could be of any value as it is often just the ratio that is important. In reality they will typically be in the $\text{k}\Omega$ range between 1 and 100 $\text{k}\Omega$. If the resistor values are too low we might find that the current flow, and power dissipation, is too high for the op-amp to drive properly. It can also lead to problems with interfacing with other stages, for example if they have low input resistances. If the resistor values are too high then there may be unacceptable levels of thermal noise generated and there may be significant voltage drops at inputs where in a real device the current is non-zero. In addition to these dc effects there may also be issues with the frequency response of interface circuits if the resistors are not carefully chosen.

7.5.2 Input Offset Voltage

If both inputs are grounded then the output of an ideal op-amp will be 0 V. In a real device with no negative feedback the output will typically be found to be saturated to one of the voltage rails. This is due to a small internal offset voltage between the two inputs which can be compensated for with an external potential divider, usually connected to the negative power supply. This is connected to two pins on the package of an op-amp chip, usually marked offset-null or balance.

7.5.3 Input Bias Current

As noted in Section 7.4.1, the ideal op-amp has zero current into the inputs but in practise this will be some small but non-zero figure. The actual value will depend on the op-amp design, those using FETs in the input differential stage will have lower bias currents. This really matters where there is significant series resistance on the inputs as it can lead to

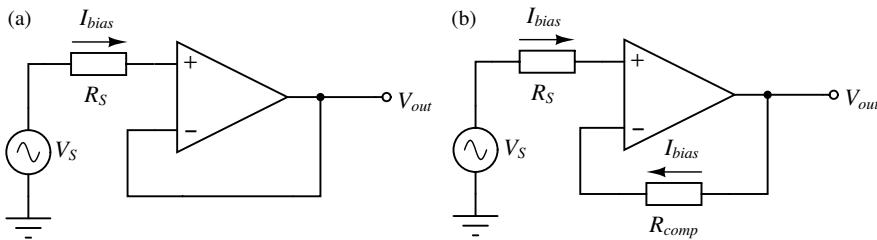


Figure 7.13 (a) An uncompensated buffer connected to a sensor (b) A buffer circuit which has been compensated for input bias current induced voltage drops.

undesirable voltage drops. This is illustrated in Figure 7.13(a) which shows an uncompensated buffer connected to a sensor with an output resistance, R_s . In this circuit there will be a voltage drop across this resistance with the result that the output voltage is not equal to the sensor output. It is compensated for by adding a resistance, of the same value as R_s , into the feedback connection to the inverting input. The result of this is that there is a similar bias current induced voltage drop on both inputs, as shown in Figure 7.13(b). This approach assumes that the input bias current is the same for both inputs, but even if there is a small difference it should be better than not compensating. Similar compensation can be added to inverting and non-inverting amplifier designs. In those cases the compensation resistor, connected to the non-inverting input of the op-amp, will be the parallel combination of the resistors in the feedback network.

7.5.4 Power Supply

The point was made in Section 7.4.1 that the output voltage of the operational amplifier is constrained by the power supply range. This can lead to ‘clipping’ of the output signal if the amplitude of the input signal is too large. More modern op-amps are better at this than the old 741 which could only get within about 2 V of the power rails. Those that drive the output within a few hundredths of a volt of the power supply voltage, such as the LM6142 from National Semiconductor, are commonly referred to as ‘rail-to-rail’ op-amps.

7.5.5 Op-Amp Noise

Using large resistors in an amplifier design will generate thermal noise but there is also the intrinsic noise of the amplifier itself to deal with. This is made up from a number of different sources and will contain both white and pink spectral components. The noise is often specified for a particular device with a graph indicating the white noise level at high frequencies and the ‘ $1/f$ corner’ frequency below which the $1/f$ pink noise dominates. This can be used, along with knowledge of the bandwidth of the amplifier, to calculate a noise level.

7.5.6 Frequency Response

The open-loop gain of an op-amp will depend on the frequency of the signal applied. The maximum gain typically only applies to dc inputs and the actual cut off frequency (-3 dB point) can be pretty low, maybe only a few tens of Hertz. By applying negative feedback and lowering the gain we actually increase the frequency range for the circuit. The open loop

gain will roll off at the standard first order rate of 20 dB/decade so, given a value for the open loop cut off frequency it should be possible to work out the cut off frequency, f_c , for a given open-loop gain. An important characteristic of an op-amp is the ‘Gain Bandwidth Product’. This is fixed and indicates that increasing gain will therefore reduce the bandwidth.

7.6 Instrumentation for Electrochemical Sensors

7.6.1 The Electrochemical Cell (Revision)

As discussed in detail in Chapter 5, Section 5.4.3, the standard setup for electrochemical measurements requires three electrodes. The working electrode (WE), where the reaction of interest is occurring, has a potential defined versus the reference electrode (RE). The reference electrode should have a stable potential under the conditions of the experiment. It is desirable to get RE as physically close to WE as possible but if it is too close it can affect mass transport and the electric field at the electrode surface, resulting in an uneven current distribution. The counter electrode (CE) supplies all the current and is usually very much larger in area than the working electrode so it can supply plenty of current and not affect the WE reaction. It can also affect the electric field and current distribution at the working electrode so it can be advantageous to have CE surrounding WE.

It can be difficult when coming from an electrical engineering background to understand exactly why these three terminals are required. The reference electrode defines a fixed potential in the electrolyte and in order for this to be stable there should be no current flow through RE which might lead to a reaction that changes the potential. The WE is often connected to ground but it is the potential difference between the WE and the solution that is important in electrochemistry. The effective WE potential is set by the reference electrode and this is usually quoted as being ‘vs. SCE’ or similar, where the acronym refers to the reference electrode type and allows you to work out the characteristic potential. SCE in this case is a ‘Saturated Calomel Electrode’, described in Chapter 5, Section 5.5.2. As discussed above, the RE cannot supply significant current and still be a stable reference so we need another low impedance terminal, that is the counter electrode, which supplies the current required to support the electrochemical reaction at the WE.

7.6.2 Equivalent Circuit of an Electrochemical Cell

Figure 7.14 shows the standard symbol for the 3-terminal cell and the equivalent circuit. The resistance R_s represents the electrolyte solution and will depend on the ionic conductivity of the medium and the distance between the CE and the RE/WE. Ideally, the reference electrode will be placed as close as possible to the WE in order to minimise the uncompensated solution resistance R_u . Otherwise this will lead to an error in the potential (with value iR_u) between the RE and WE. Next we have the reference electrode resistance R_{ref} which can be very high depending on the type of electrode used. Ideally there will be no current flow in the reference electrode as this will contribute to any error in the potential applied to the WE. The RE also has a parasitic capacitance C_{ref} that will effect transient operation. This parasitic capacitance may be due to the connecting leads or the structure of the electrode itself.

C_{dl} represents the sizeable capacitance of the electrical double layer at the surface of the working electrode. This along with R_u and R_s will affect transient operation giving a

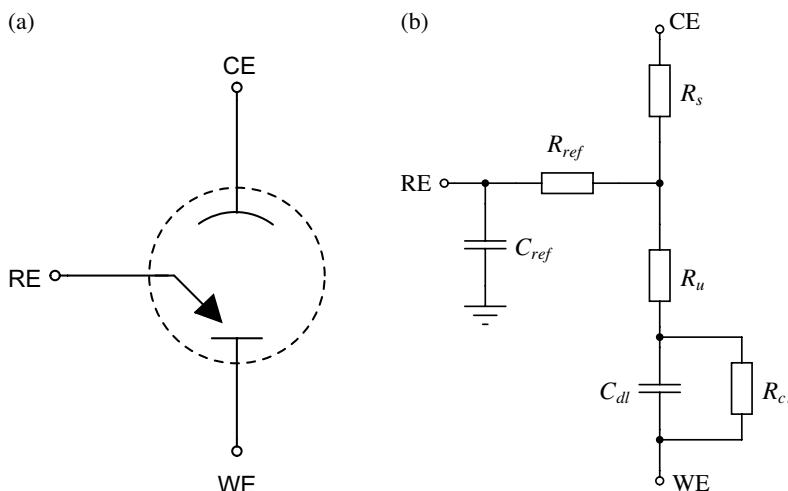


Figure 7.14 (a) Circuit symbol for a 3-terminal electrochemical cell. (b) Equivalent circuit for a 3-terminal electrochemical cell.

charging behaviour for step changes of the potential applied to the WE. R_{ct} represents the conduction at the WE when an electrochemical charge transfer reaction occurs.

7.6.3 Potentiostat Circuits

The potentiostat is the basic instrument used in electrochemical experiments. It controls the potential between the reference and working electrodes while sourcing whatever currents are required through the counter electrode. A standard potentiostat circuit is shown in Figure 7.15.

The summing point marked with 'S' is at virtual earth so the input potential e_m will be dropped across the first bias resistor R . Assuming there is no current into the input of op-amp OA-1 then the potential across the other bias resistor R will also be e_{in} and so the voltage on the inverting input of OA-2 will be $-e_{in}$. The effective feedback loop between the counter and reference electrodes through OA-1 and OA-2 will serve to keep the potential of the RE at $-e_{in}$ with respect to circuit ground. Meanwhile, the fact that the reference electrode is connected to the non-inverting input of OA-2 means that there will be little or no current flow. The working electrode is held at a virtual earth by the current follower circuit of OA-3. Therefore the effective WE potential (vs. RE) is $e_{WE} = e_{in}$, and OA-1 will supply the required current to sustain electrochemical reactions and keep this potential stable. The input bias currents of OA-2 need to be very small otherwise current through the reference electrode can give a potential error and so FET based input stages should be used. The working electrode current (i_{WE}) is measured at the output of OA-3 and the WE potential can be checked at the inverting input of OA-2. Many potentiostats using this arrangement for current measurement have a range of different values of R_{out} which can be selected depending on the value of the current to provide more or less gain to the current convertor. Low input bias current is useful in OA-3 as it reduces errors when using a high-value feedback resistor.

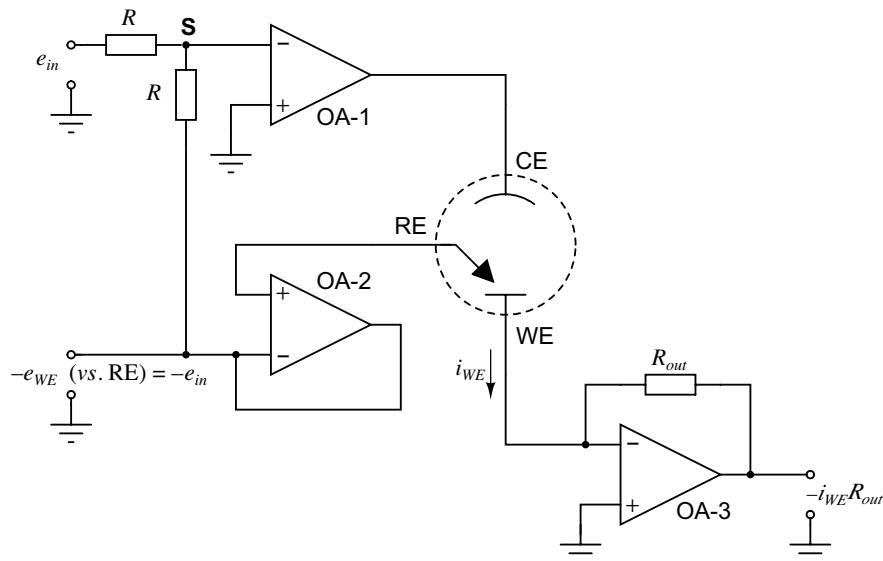


Figure 7.15 Typical potentiostat circuit.

The alternative potentiostat circuit shown in Figure 7.16 has the working electrode connected directly to ground rather than to a virtual ground formed by the input of an op-amp. The input side (RE/CE) is practically identical to the potentiostat in Figure 7.15, so the potential on the working electrode (vs. ref) is e_{in} . The difference here is that the working electrode current (i_{WE}) is measured by placing a resistor R_f in the feedback loop of the control amplifier (OA-1) and measuring the voltage across it. This requires a differential amplifier and we could use the circuit described in Section 7.4.5 though buffering will be required

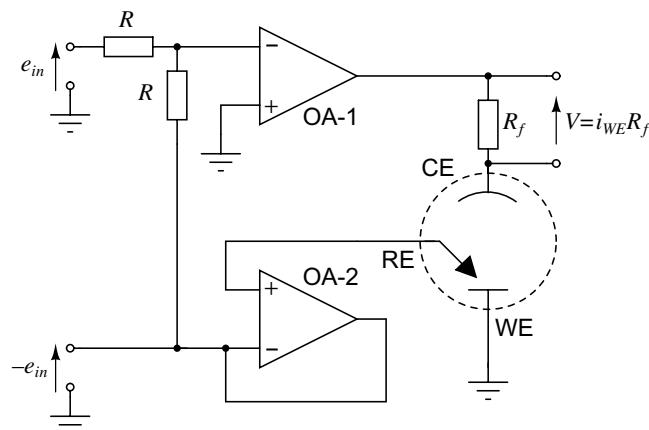


Figure 7.16 Alternative potentiostat circuit.

on the inputs to avoid changing the current flowing through the cell. Good measurements will require a high common mode rejection ratio and so it may be better to use a more advanced differential amplifier, known as an instrumentation amplifier.

7.6.4 Instrumentation Amplifier

The instrumentation amplifier circuit shown in Figure 7.17 may look quite daunting until we break it down and analyse it using the rules established for op-amp design. Basically it is a buffered differential amplifier with additional resistors in the buffering stage and the most important of these is R_{gain} . Begin by assuming that all the other resistors have the same value (R). Then, due to feedback in the buffer stages, the voltages at points 1 and 2 will become equal to the input voltages V_1 and V_2 . This means there is a voltage across R_{gain} which is equal to the difference between the input voltages and the current flow through R_{gain} will be proportional to this difference:

$$I_{R_{gain}} = \frac{V_1 - V_2}{R_{gain}}. \quad (7.11)$$

Following the golden rule that there is no current into the op-amp inputs we know that the current through R_{gain} will also flow through the two resistors marked R_1 so that the voltages at points 3 and 4 will be:

$$V_3 = V_1 + I_{R_{gain}}R \quad (7.12)$$

$$V_4 = V_2 - I_{R_{gain}}R. \quad (7.13)$$

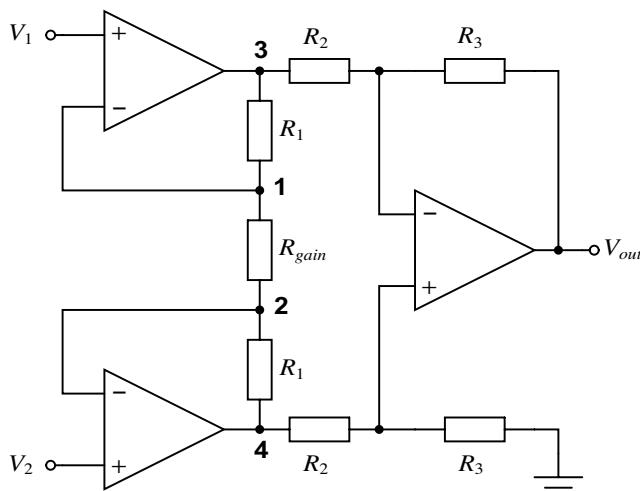


Figure 7.17 Instrumentation amplifier circuit.

These are then the inputs for the differential amplifier stage and assuming all resistors are equal then $V_{out} = V_4 - V_3$. If we substitute Equations (7.12) and (7.13) into this expression and also substitute for the current in R_{gain} using (7.11) we obtain the following relationship:

$$V_{out} = (V_2 - V_1) \left(1 + \frac{2R}{R_{gain}} \right). \quad (7.14)$$

This is an equation for the output of a differential amplifier with a voltage gain (A_d):

$$A_d = 1 + \frac{2R}{R_{gain}}. \quad (7.15)$$

This might seem a little complicated compared to the standard differential amplifier but it has the advantage of a high impedance input and a gain which is completely controlled by a single resistance (R_{gain}) which could be a variable component. Think about why you would prefer to adjust gain in a differential amp like this rather than by changing other resistor values. For example, what is the gain if R_{gain} is removed (open circuit)?

It is possible to achieve a larger gain by adjusting the other resistor values, but these need to be well matched to give a good common mode rejection ratio. For that reason, instrumentation amplifiers like this are not usually built using discrete components. Instead it would typically be a single integrated circuit package with high value integrated resistors R_1 to R_3 that have been laser trimmed to give the best possible matching between pairs. R_{gain} may still be an external component to allow the gain to be adjusted. These devices can be expensive but will give the best possible results, particularly for sensors with high output impedances.

7.6.5 Potentiostat Performance and Design Considerations

Returning to the discussion of potentiostats, there are a number of important design considerations. The first of these concerns the control amplifier (OA-1) that drives the current through the cell through the connection to the counter electrode. This needs to be able to source sufficient current for the measurement being undertaken and also be able to supply enough voltage to drive this current through the cell. There is little that can be done about the voltage as it will be set by the power supply but it will make the use of rail to rail op-amps desirable. If the current requirements are too high a power amplifier can be added to boost the current. Note that a similar amplifier will probably be required in the output stage if a current follower is used to measure the working electrode current. Current boost could be essential in non steady-state measurements as the capacitance of the electrical double layer can be very large. An example showing how large currents and voltages may be required in transient operation can be found among the sample problems at the end of the chapter.

So far this discussion has not considered the transient response of the op-amps used to build the potentiostats. The ideal op-amp has an output that responds instantly to changes in the input. An alternative way to think about it is that the open loop gain is independent of frequency. Obviously, this will not be the case in a real device where the open loop gain has a low-pass response with some cut-off frequency. This will determine the time constant of the potentiostat response to a step change in the input voltage. As mentioned previously the open loop gain has quite a low cut-off frequency but the potentiostat circuits are mainly built

with voltage followers with a unity gain and so they can have relatively wide bandwidths if stability issues can be avoided.

Obviously the cell will also contribute to the transient response of the measured system. As mentioned previously the electrical double layer capacitance (C_{dl}) can be sizeable and during measurement it is typically charged through the solution resistance. Reducing C_{dl} by having a small surface area for the WE and reducing the solution resistance R_s by putting CE and WE close together or increasing electrolyte conductivity can improve the transient response. This, combined with the potentiostat time constant determined by the op-amp characteristics, will determine how fast the measurements can be made and what types of signals can be applied.

The main effect of the uncompensated resistance R_u between RE and WE is on the charging of the double layer capacitance of the WE when a change is made to the input voltage. As mentioned previously this could require significant current and voltage to be supplied by the control amplifier to the CE to make it work. Putting R_{ct} in parallel with the double layer capacitance to represent the Faradaic reaction is an added complication.

When current flows there will always be a voltage drop iR_u so that the WE potential is never exactly the same as the input e_{in} . If we need to reduce the effect of the uncompensated resistance R_u between the RE and WE there are a number of schemes but a common electronic method involves feeding some fraction of the current measurement output back to the input. This is added to the input voltage to compensate for the R_u voltage drop. This is never usually completely compensated as it typically leads to stability problems due to the effective positive feedback. One crude method used in the past would involve increasing the feedback until the potentiostat starts to oscillate then backing off to about 80% of this value. However, this is all rather imprecise and other methods have been developed to directly measure R_u , such as current interruption. Here a constant current is set up through the cell then reduced as instantaneously as possible. When the current step occurs the WE potential decays in some complex manner related to discharging of the double layer capacitance through Faradaic processes. However, the ohmic drop across R_u decays instantaneously as well so we can estimate it from the transient potential curve. The trouble here is achieving the instantaneous interruption of current but if this is done quickly enough then it will be possible to monitor R_u without upsetting the experiment.

A more modern way of doing this is in a computer controlled potentiostat is to apply a small step in potential ($\Delta E \approx 50$ mV), while the cell is biased in a potential region where there is no Faradaic reaction occurring at the WE. If this is the case the only current flowing will be a transient as the electrical double layer capacitance C_{dl} of the WE is charged through R_u . Automated analysis of the data can be used to extract R_u and C_{dl} and then to adjust the value of the feedback from the WE current output to provide R_u compensation while simultaneously monitoring for instability or oscillation.

When the internal resistance of the cell is small (~ 0.1 Ω) the contact resistance (R_c) to the working electrode can dominate if it is of a similar level. This is very possible with micro-scale electrodes connected to a standard bench potentiostat but might not be so important in a smart system with integration. If the cell currents are high then the voltage drop iR_c can be significant. There will also be contact resistances at the RE and CE but they are less important. There should be little or no current through the reference electrode and the CE will

simply require a larger voltage from the control amplifier to make everything work. R_c can be compensated by adding a fourth lead that connects to the WE but draws no current. Effectively this will involve placing a voltmeter, with a high internal resistance, in parallel with the WE connection to measure iR_c , which can then be compensated in a similar way to iR_u . However, it is more likely that the currents for low resistance cells with microelectrodes will be so small that the potential drop will be negligible.

Like all control and instrumentation circuits using negative feedback, the stability of the system depends on the phase shift around the feedback loop. If this ever becomes 180° then it can switch to positive feedback which leads to instability and oscillation if the gain is ≥ 1 . The phase shift can come from either the potentiostat frequency response itself or from the impedance of the cell, which includes the parasitics of the reference electrode. The cell impedance can be difficult to control, but the time constant of the RE parasitics can be improved through careful electrode design and by applying a driven shield to the reference connection. Noise can also cause problems here and placing the whole experiment in a Faraday cage to reduce electromagnetic interference is a possibility.

7.6.6 Microelectrodes

Many modern biosensors using electrochemical transduction are of a type known as ‘Ultra-microelectodes’ (UME). That is often defined as a electrode with a characteristic dimension less than $20 \mu\text{m}$. This value is related to the length of the diffusion layer that forms at an electrode surface under steady state conditions. With a macro electrode this leads to a reduction in current as the reaction becomes diffusion controlled. The reactive ions are effectively depleted at the surface of the macro-electrode and the reaction rate becomes controlled by mass transport of the chemicals from the bulk solution to the surface. In a microelectrode a hemispherical diffusion field forms which means in practice that there is always sufficient transport of reactive species to the surface to sustain a steady but very small current flow. Currents become proportional to the radius rather than surface area and so the effective current density is very high. These microelectrodes are ideal for studying fast reactions as the small current drops and reduced charging times due to the small area make it easier to sweep the potential quickly.

7.6.7 Low Current Measurement

Low current measurements on UMEs require special considerations for the measurement system. Measuring very low currents ($\text{nA} - \text{fA}$) requires a large feedback resistance in a current follower amplifier and an op-amp with an input bias current that is as small as possible. This measurement will also be susceptible to noise causing erroneous currents and the current follower has particular problems with high frequencies and large values of R_f . The working electrode has some capacitance associated with it and there is likely to also be some capacitance from the input stage of the op-amp. This forms a low-pass circuit with the feedback resistor causing problems with bandwidth and stability. This is often dealt with by putting a capacitor in parallel with the feedback resistor.

If the current is low then rather than using a large feedback resistor we could use a current amplifier as shown in Figure 7.18. This is made up from a current follower followed by an inverter. The voltage at inverter input will be $i_{WE}R_f$ as expected while the voltage at the inverter output is $i_{WE}R_f$. Assuming that the output resistor is connected to another current

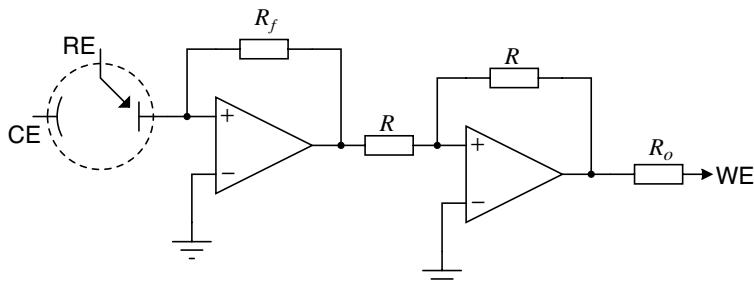


Figure 7.18 Low current amplifier circuit.

follower, such as the WE connection of a potentiostat, then the current through the output resistor R_o will be:

$$I_{R_o} = i_{WE} \frac{R_f}{R_o}. \quad (7.16)$$

Overall, the circuit amplifies the WE current with a gain of R_f/R_o . An important design consideration is to make sure that the first op-amp has a very low input current bias to reduce errors. When the currents are very low the current measurement may need to be done with a very sensitive multimeter such as an electrometer which has very high input resistance and low input bias current. These kinds of low level measurements are very common in measurements of advanced semiconductor devices and one of the main test equipment manufacturers produces a free handbook which covers many of the issues. Details of how to get this are provided at the end of this chapter.

The low currents involved in electrochemistry with UME can effectively eliminate effects of R_u . In fact we can often do away with the reference electrode completely and just use two electrodes: one WE and a combined RE/CE. As long as the currents are very low there should still be a constant potential at the RE/CE. This is particularly good for integrated microelectrode systems or experiments using arrays of working electrodes. It also reduces the complexity of the electronics considerably; so that a simple function generator can drive RE/CE with a current amplifier/follower measuring the current.

7.7 Impedance Based Biosensors

7.7.1 Conductometric Biosensors

Conductometric biosensors operate by detecting variation in electrical conductivity, often represented by changes in the ionic concentration local to the sensor, that are induced by some enzymatic reaction. Interdigitated microfabricated electrodes are coated with a gel which contains the immobilised enzyme, and changes in conductivity depending on the detection reaction. Therefore, the enzyme is the biological sensing element, while the electrodes transduce the conductivity change in the gel into a measurable resistance. To measure this type of sensor we could use a resistance to voltage converter. This could be realised using the inverting amplifier shown in Figure 7.10, where the input voltage is fixed at V_{ref} and the

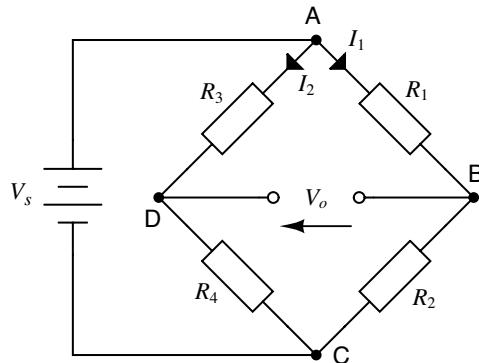


Figure 7.19 A resistive Wheatstone bridge.

conductometric sensor (R_S) replaces the input resistance R_{in} . The output of this circuit will be:

$$V_{out} = -V_{ref} \frac{R_f}{R_S}. \quad (7.17)$$

Therefore, the output voltage is inversely proportionate to the resistance of the sensor, that is it is proportional to the conductance.

Another common method for resistance to voltage transduction is the Wheatstone bridge and it is worth reviewing the principle behind this circuit. Referring to the circuit shown in Figure 7.19, if we assume the output potential $V_o = 0$ V then the voltage drops across R_1 and R_3 must be equal:

$$V_{AD} = V_{AB} \Rightarrow I_1 R_1 = I_2 R_3. \quad (7.18)$$

Similarly, the voltage drops across R_2 and R_4 must also be equal:

$$V_{DC} = V_{BC} \Rightarrow I_1 R_2 = I_2 R_4. \quad (7.19)$$

If we divide these together then we get an equation stating that the ratios of the resistors in one side of the bridge must equal those in the other half:

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}. \quad (7.20)$$

This is a balanced bridge circuit but a small difference in one resistor will unbalance the bridge so that $V_o \neq 0$ V.

In order to work out the voltage V_o for an unbalanced bridge we need to know the voltages at points D and B, which will be the source voltage V_s minus the voltages across R_1 and R_2 . These are calculated as voltage dividers:

$$V_{AB} = \frac{V_s R_1}{R_1 + R_2} \quad (7.21)$$

$$V_{AD} = \frac{V_s R_3}{R_3 + R_4}. \quad (7.22)$$

Then $V_o = (V_s - V_{AD}) - (V_s - V_{AB}) = V_{AB} - V_{AD}$ which works out to be:

$$V_o = V_s \left(\frac{R_1}{R_1 + R_2} - \frac{R_3}{R_3 + R_4} \right) \quad (7.23)$$

$$V_o = V_s \frac{R_1 R_4 - R_2 R_3}{(R_1 + R_2)(R_3 + R_4)}. \quad (7.24)$$

If V_o is zero we can work back from this to the resistor ratios shown before. However, what if one resistor is a sensor with a resistance that has changed by some value δR_1 ? The new value for the resistor is then $R_x = R_1 + \delta R_1$ and if this is substituted into Equation (7.24) the result is:

$$V_o = V_s \frac{(R_1 + \delta R)R_4 - R_2 R_3}{(R_1 + \delta R + R_2)(R_3 + R_4)}. \quad (7.25)$$

Now if we begin with a balanced bridge where all the resistors, including the sensor, have the same value, that is $R_1 = R_2 = R_3 = R_4 = R$ then (7.25) can be simplified to:

$$V_o = V_s \frac{\delta R}{4R + 2\delta R} \approx V_s \frac{\delta R}{4R}. \quad (7.26)$$

The final approximation only applies if the assumption is made that $\delta R \ll R$ and the result is that the output voltage V_o is proportional to the change in resistance. Therefore, this circuit can be considered as a resistance to voltage converter.

Buffering and amplification of the output signal from the Wheatstone bridge can be achieved using a buffered differential amplifier or the instrumentation amplifier discussed earlier. The resistors in the circuit can also be more general impedances (Z_1 to Z_4) and the bridge can be driven with an ac signal. For example, this could be used to measure an unknown capacitance. The phase and amplitude of the output voltage will contain information about the impedances in an unbalanced circuit.

7.7.2 Electrochemical Impedance Spectroscopy

Electrochemical impedance spectroscopy (EIS) is a standard measurement of the impedance at the surface of a working electrode in electrochemistry. This involves applying a small sinusoidal signal, via a potentiostat, at some set dc level (V_0) and measuring the resulting current. The input signal is:

$$V(t) = V_0 + |V| \sin(\omega t) \quad (7.27)$$

while the measured current is:

$$I(t) = I_0 + |I| \sin(\omega t + \phi). \quad (7.28)$$

The important parameters are the dc level (V_0) which sets the point on the IV curve for the particular electrochemical reaction, the magnitude of the input voltage $|V|$ and the frequency $\omega = 2\pi f$. If the input amplitude is small then the response should be approximately linear

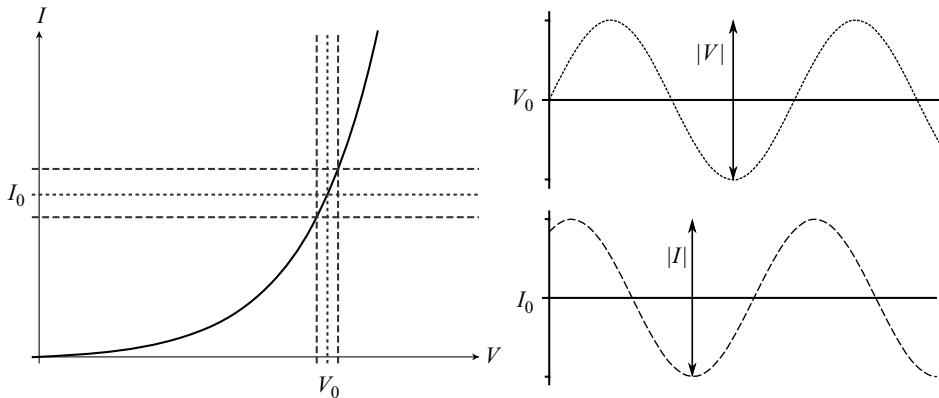


Figure 7.20 On the left is the IV curve for a sensor being measured using electrochemical impedance spectroscopy. An ac voltage with a dc level V_0 and an amplitude $|V|$ is the input (upper right graph) while the output is a phase shifted, alternating current with amplitude $|I|$ centred at I_0 (lower right graph).

and the output will be a sinusoidal current at some dc level, I_0 , as shown in Figure 7.20. The impedance $Z(\omega)$ as a function of frequency can then be calculated using Ohms Law:

$$Z(\omega) = \frac{V(t)}{I(t)}. \quad (7.29)$$

Unless the impedance is purely resistive it will be a complex quantity with a magnitude, $|Z|$, and phase shift ϕ in polar coordinates:

$$Z(\omega) = |Z(\omega)|e^{j\phi(\omega)}. \quad (7.30)$$

Similarly in a cartesian form it will have a real part Z_r and an imaginary part Z_j :

$$Z(\omega) = Z_r(\omega) + Z_j(\omega). \quad (7.31)$$

We can move between the two forms using the formulae below:

$$|Z| = \sqrt{Z_r^2 + Z_j^2}, \phi = \tan^{-1} \left(\frac{Z_r}{Z_j} \right) \quad (7.32)$$

$$Z_r = |Z|\cos(\phi), Z_j = |Z|\sin(\phi). \quad (7.33)$$

The complex impedance will be dependent on the frequency of the sinusoidal signal used to make the measurement, measurements are made over a range of frequencies to obtain an impedance spectra. This will contain a great deal of information about the electrode surface and is the basis of EIS biosensing.

7.7.3 Complex Impedance Plane Plots and Equivalent Circuits

One of the most common ways of representing EIS data is a plot of the complex impedance plane where the x-axis is the real part of Z while the y-axis is the imaginary part of Z .

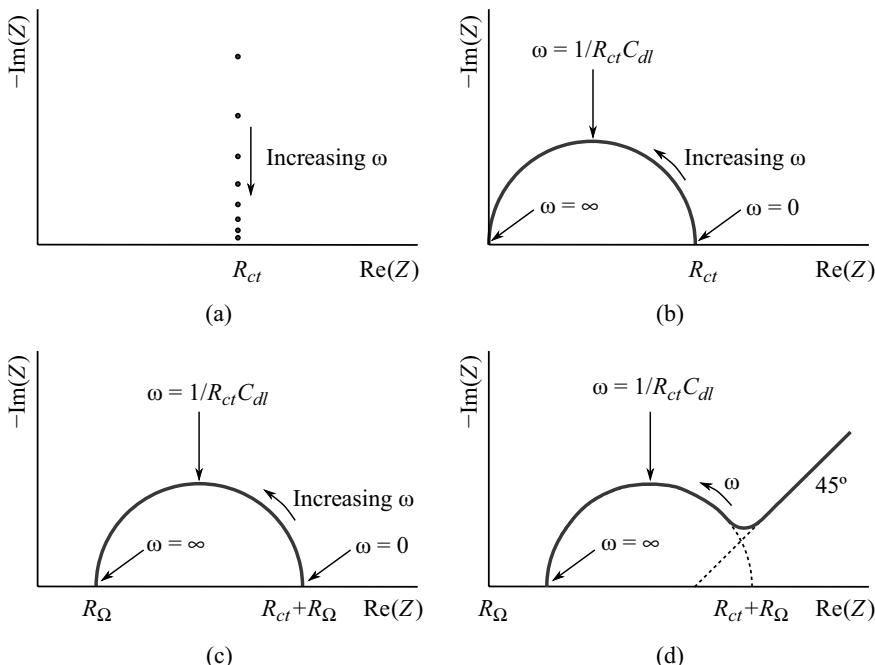


Figure 7.21 Impedance plane plots for the following circuits: (a) resistor and capacitor in series, (b) resistor and capacitor in parallel, (c) resistor and capacitor in parallel with added series resistance, (d) the Randles equivalent circuit in Figure 7.22.

Generally we know that for EIS measurements Z_j is usually a negative, capacitive, value and so we plot $-\text{Im}(Z)$. Each point on the curve will represent a measurement made at a different frequency.

A simple example is a series combination of resistor and capacitor (Figure 7.21(a)). This has a constant real part and a imaginary part which is inversely proportional to frequency. This might represent an electrochemical cell operated in a potential region where there is no Faradaic reaction. The resistor is the series resistance of the solution, or the uncompensated resistance in a 3-terminal cell, and the capacitance is that of the electrical double layer at the working electrode.

The second example (Figure 7.21(b)) is that of a resistor in parallel with a capacitance, which is equivalent to a double layer capacitance and a charge transfer resistance at a WE. At low frequencies the capacitor is basically an open circuit so the impedance is simply the resistance R_{ct} . At very high frequencies the capacitor will short out the resistance and the impedance will tend towards zero. The rest of the plot resembles a semicircle and the highest point on this curve represents the frequency where $\omega = 1/R_{ct}C_{dl}$ making it possible to estimate both the resistance and capacitance from this curve.

Adding a resistor in series with this circuit, to represent the solution resistance of the electrochemical cell, will shift the response along the real axis (Figure 7.21(c)). Now the high-frequency intersection with the real axis is at R_Q , where the double layer capacitance

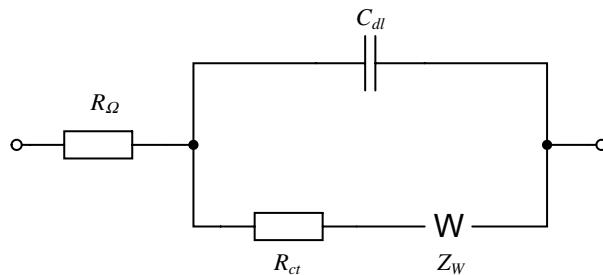


Figure 7.22 Randles equivalent circuit for an electrochemical impedance spectroscopy experiment.

shorts out R_{ct} , while the low frequency intersection is at $R_\Omega + R_{ct}$. Capacitance information can again be obtained from the frequency of the measurement which gives the maximum imaginary part, that is the top of the semicircle.

The most complete model is sometimes referred to as a Randles equivalent circuit and it includes a ‘Warburg Impedance’ which represents the effects of diffusion in the electrochemical cell. This is explored in detail in Chapter 5 from the viewpoint of the physical phenomena involved. As with the previous model, at high frequencies the impedance is dominated by the solution resistance R_Ω while at lower frequencies the double layer charges up and a reaction at the electrode occurs (R_{ct}). As this reaction continues the reactive species are depleted at the electrode surface and so a diffusion layer forms. Eventually this will act as a fundamental limiting factor on the reaction. This appears as an increase in the impedance at low frequencies which has a constant phase angle of 45° and will tend towards an effectively infinite impedance for dc measurements. This is the limiting case for a macro-electrode where the reaction effectively stops once the reactive species are completely depleted at the electrode surface. The impedance plane plot for this model is shown in Figure 7.21(d) while the equivalent circuit, with the Warburg impedance Z_W can be seen in Figure 7.22. As there is no low frequency intersection with the real axis, a semi-circle is fitted to the plot obtained by EIS measurements in order to estimate the value of $R_\Omega + R_{ct}$.

7.7.4 Biosensing Applications of EIS

A very good review of the use of conductometric and impedimetric biosensors was published by Katz and Willner in 2003 [1]. The main reason for using EIS in biosensors is that it provides very accurate measurement of changes to a surface caused by molecular attachments. There are two main forms of electrodes for EIS biosensor applications. The first is the interdigitated electrode (IDE) which can measure changes in the in-plane impedance between the two electrodes. This is typically non-Faradaic in nature and will show changes in the dielectric or conductive properties of the material coating the electrodes. The second type is a functionalised electrode where the impedance between the electrode and the solution can be measured. Selective attachment of analyte molecules to a biosensor coating will change the EIS response which can be measured as changes in R_{ct} or C_{dl} . Figure 7.23 shows the expected changes in EIS spectra for an electrode in a bare state (a), then with a sensing layer (b) and finally after antibody attachment (c), as described in [1].

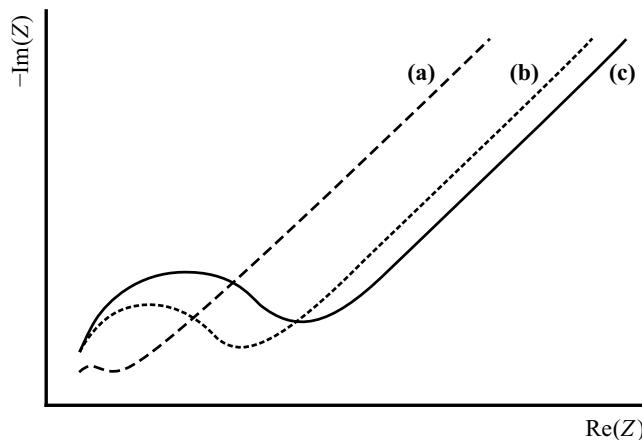


Figure 7.23 Representative Faradaic impedance spectra for a biosensor: (a) bare electrode before coating; (b) coated with a biosensing layer; (c) after specific attachment of analyte to the biosensing layer.

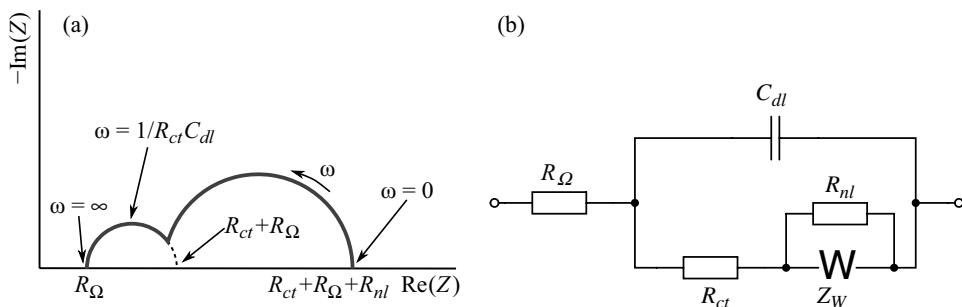


Figure 7.24 (a) Impedance plane plot for a microelectrode, (b) modified Randles equivalent circuit for a microelectrode to achieve improved mobilities.

Microelectrodes do not show the same response as macroelectrodes, particularly at low frequencies where hemispherical diffusion means there are no reaction limiting diffusion issues. This requires a modified Randles circuit with a resistor (R_{nl}) in parallel with the Warburg impedance. This means that the impedance will tend towards a constant resistance at very low frequencies, which represents the stable current seen at dc with a microelectrode. Figure 7.24 shows the modified Randles circuit along with a model impedance plane plot for this type of microelectrode experiment.

7.8 FET Based Biosensors

7.8.1 MOSFET Revision

The Metal Oxide Semiconductor Field Effect Transistor (MOSFET) is one of the most ubiquitous electronic components in the world with most modern computing devices containing several billion. Although most electrical engineers will have at least some familiarity with

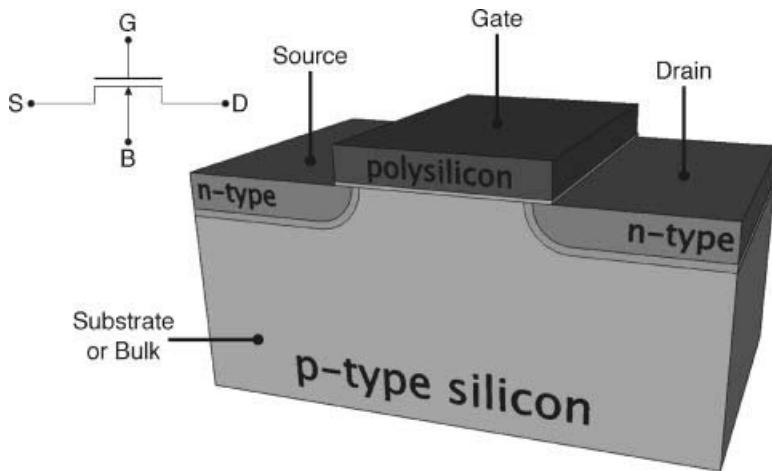


Figure 7.25 Schematic image of an n-channel, polysilicon gate MOSFET, along with the symbol used for such a device in a circuit diagram.

the theory of operation of the MOSFET it is worth reviewing the important characteristics of the device before looking at how this structure can be used to implement biological and chemical sensors. A basic knowledge of semiconductor physics, such as the difference between p-type and n-type doping, is assumed in the following discussion.

The basic structure of an MOS transistor is shown in Figure 7.25 and in fact this device should probably be given a different name as the gate is fabricated from polycrystalline silicon (polysilicon) rather than metal. The substrate is p-type silicon into which are diffused or implanted two n-type regions, while the depletion regions between these are shown in red. The n-type areas are the source and drain of the transistor and are typically formed by an implantation through a mask formed by the polysilicon gate electrode in a self-aligned process. Separating the gate electrode from the silicon substrate is a very thin dielectric layer, normally of silicon dioxide. This structure is not quite the same as an advanced CMOS (Complementary MOS) device from a modern digital electronic process where the gate length may only be 25–30 nm. In those devices the gate has now generally switched to a metal rather than a polysilicon structure, there are multiple implants to create complicated source-drain regions, and the substrate may be silicon-on-insulator or use Si-Ge to create strain in the channel to achieve improved mobilities.

The easiest way to think of the MOSFET is as a switch where conduction between the source and drain contacts is dependent upon the voltage applied to the gate electrode. However, this is a four terminal device and the potential applied to the bulk or substrate also has an effect. In the n-channel device shown in Figure 7.25 the n-type source and drain regions are separated by the p-type channel region and no current will flow unless a very high voltage is applied between source and drain (V_{DS}), causing the transistor to break down. If the voltage on the gate, which is usually referenced to source potential as V_{GS} , is increased from 0 V up to some device dependent value called the threshold voltage (V_T), then the charge on the gate will begin to repel mobile positive charge in the channel leading to the formation of a depletion region. For gate-source voltages above the threshold value the attraction of

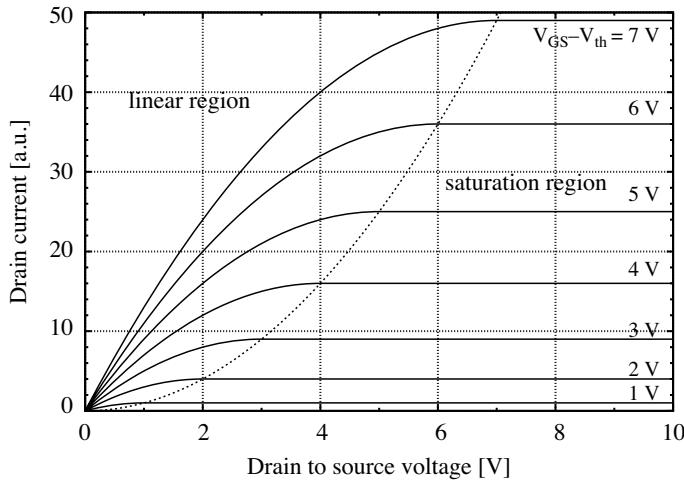


Figure 7.26 Transistor IV characteristic graph for an n-channel MOSFET. Note this is a model characteristic and is not representative of a real device.

negative mobile charge carriers causes an inversion layer or channel to form under the gate. This connects the source and drain allowing current to flow and the transistor is now considered to be on.

There are two main regions of operation for a transistor if the subthreshold or weak-inversion region, where the transistor is almost on, is ignored. This is important for some modern analogue electronics where the exponential characteristic can be exploited but is less important for the analysis of FET based sensors that will follow. The standard transistor characteristic using the first-order model can be seen¹ in Figure 7.26 and the boundary between the regions of operation is marked with a dotted line. In the linear region of operation, where $V_{GS} > V_T$ and $V_{DS} < (V_{GS} - V_T)$, the current flow between the source and drain (I_{DS}) can be estimated with the following equation:

$$I_{DS} = \beta \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right], \quad (7.34)$$

where β is a device dependent variable calculated using:

$$\beta = \mu C_{ox} \frac{W}{L}, \quad (7.35)$$

where μ is the mobility of the majority charge carrier (electrons for an n-channel transistor), C_{ox} is the gate oxide capacitance and W and L are the channel dimensions.

In the saturation region, where $V_{GS} > V_T$ and $V_{DS} > (V_{GS} - V_T)$, the large drain-source voltage causes the channel region to narrow near to the drain and ‘pinch-off’ so that the inversion layer no longer connects the source and drain. However, the large electric field across the pinch-off region still enables the source drain current to flow. In the ideal transistor

¹ Graph created using public domain gnuplot code from http://en.wikipedia.org/wiki/File:IvsV_mosfet.png.

model the the drain current then becomes independent of V_{DS} , but in a real device, particularly one with a short channel, there is some variation associated with the modulation of the effective channel length with the drain voltage. The equation for the current through the transistor channel becomes:

$$I_{DS} = \frac{\beta}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS}), \quad (7.36)$$

where λ is a device dependent channel length modulation term that leads to an increase in the drain current with drain-source voltage.

It is useful to look at the different parameters which can contribute to the threshold voltage of a MOSFET. This will become essential when trying to understand the operation of FET based chemical and biological sensors later in this chapter. The equation for the threshold voltage of a n-channel MOSFET is:

$$V_T = V_{FB} + 2\phi_F + \frac{\sqrt{2\varepsilon_s q N_a (2\phi_F + V_{BS})}}{C_{ox}}, \quad (7.37)$$

where V_{FB} is the flatband voltage of the transistor, a very important characteristic which represents the voltage applied to the gate that results in no charging of the channel region. This means that if V_{GS} is equal to the flatband voltage the effective carrier density in the channel will be the same as in the bulk substrate. The second term in the threshold equation is twice the bulk potential of the silicon substrate which is defined as:

$$\phi_F = \frac{kT}{q} \ln \frac{N_a}{n_i}, \quad (7.38)$$

where kT/q is the thermal voltage, N_a is the p-type doping level of the substrate and n_i is the carrier density of intrinsic silicon. The final term in the threshold voltage equation defines the potential across the gate oxide due to the depletion layer charge under the gate. This includes parameters such as the silicon permittivity (ε_s), the gate oxide capacitance (C_{ox}), the substrate doping (N_a) and the applied voltage between the source and bulk (V_{BS}).

The flatband voltage is dependent on the difference in work function between the material of the metal/polysilicon gate and the silicon substrate and varying this is the principle by which many of the sensors examined later in the chapter will operate. The equation for the flatband voltage is:

$$V_{FB} = \Phi_{MS} - \frac{Q_f + Q_{ox}}{C_{ox}}, \quad (7.39)$$

where Φ_{MS} is the potential due to the work function difference, while the second term is the potential due to the trapped charge (Q_f) at the boundary between the gate oxide and the silicon substrate and trapped charge within the gate oxide layer itself (Q_{ox}).

7.8.2 The Ion Sensitive Field Effect Transistor

The ion sensitive field effect transistor (ISFET) is an attempt to create a miniaturisable, solid state, ion selective electrode and was originally developed by Prof. Piet Bergveld at the

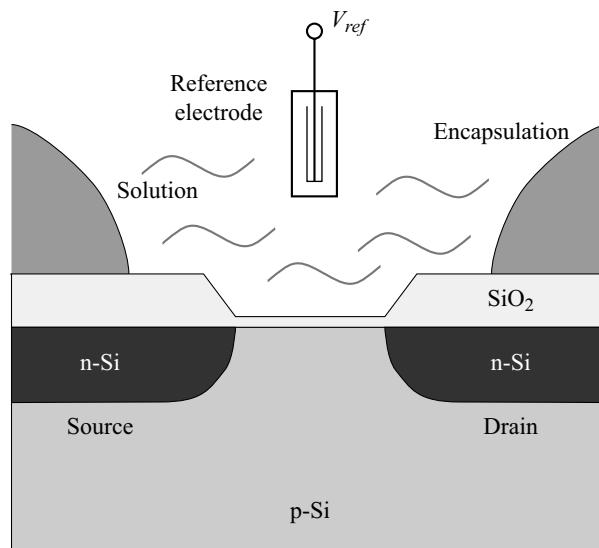


Figure 7.27 Schematic structure of a bare gate ISFET pH sensor.

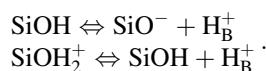
University of Twente in the 1970s [2]. It is a standard MOSFET with the gate metal/polysilicon removed and replaced by a combination of the solution being measured and an electrochemical reference electrode (Figure 7.27). The gate dielectric, commonly SiO_2 , acts as the ion sensitive membrane where the surface potential is controlled by the pH of the solution.

In the ISFET the effective gate source voltage is controlled by the reference electrode connection to the solution but the characteristics of the solution will also affect the threshold voltage. The expression for the flatband voltage of the ISFET is:

$$V_{FB} = E_{ref} + \Psi_0 + \chi_{sol} - \Phi_{Si} - \frac{Q_f + Q_{ox}}{C_{ox}}. \quad (7.40)$$

The first three terms here are the contribution from the gate electrode side of the device, equivalent to Φ_M in a MOSFET. This includes figures for the potential of the reference electrode (E_{ref}) and the solution dipole potential (χ_{sol}), but the most important for operation as an ion sensor is the surface potential Ψ_0 . The final two terms are due to the work function of the silicon and the trapped oxide charges as described in the discussion of Equation (7.39).

The interface between the thin gate oxide over the channel region of the transistor and the solution will consist of hydroxyl (OH) groups, which can either accept or donate protons (H^+ ions) from the solution. They are ‘amphoteric’ sites, meaning they can exist in acidic, basic and neutral forms. The balance of these charges at the surface will depend on the pH in the solution and will effectively change the surface potential of the oxide. The equilibrium reactions between the surface and the solution are:



The surface acts as a buffer for changes in the pH of the bulk solution. When this increases the surface will donate protons, becoming more negatively charged, whereas when the pH reduces the surface will accept protons and become more positively charged. The theory is developed in more depth than is appropriate for this book in [3], which may be of interest to some readers. In the simplest form, the surface potential at the oxide solution interface (Ψ_0) is dependent on the pH of the bulk solution (pH_B) with the following function:

$$\frac{\delta\Psi_0}{\delta pH_B} = -2.3 \frac{kT}{q} \alpha. \quad (7.41)$$

This is Nernstian in nature but includes a sensitivity parameter α which can vary between 0 and 1. The closer to 1 it is then the more Nernstian the pH response, that is it will show a similar characteristic to a traditional glass membrane pH sensor electrode, where the surface potential has the following dependence on pH: $\Delta\Psi_0 = -59.2 \text{ mV pH}^{-1}$ at 298 K (see Chapter 5, Section 5.3.5). Theoretically the sensitivity factor α will have the following formula:

$$\alpha = \frac{1}{\frac{2.3 kTC_{dif}}{q^2 \beta_{int}} + 1}. \quad (7.42)$$

This is a function of physical constants (k , T & q) as well as the variables C_{dif} and β_{int} which represent firstly the double-layer capacitance at the solution/oxide interface and secondly the buffer capacity of the oxide surface.

The capacitance C_{dif} is a result of the electrical double-layer that forms at the surface of any object when it is submerged in a liquid. In the most common model of this is the Gouy-Chapman-Stern model described in Chapter 3, Section 3.16.1. This is made up from two parallel layers of charges as illustrated in Figure 7.28. The inner part is a tightly bound layer

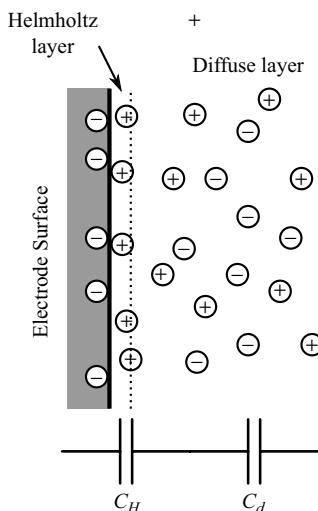


Figure 7.28 Schematic diagram of the arrangement of charge in an electrical double layer according to the Gouy-Chapman-Stern model.

of ions at the oxide surface referred to as the compact Stern or Helmholtz layer. This is surrounded by a diffuse layer of charges that balance the surface potential or charge of an electrode or other charged surface when in contact with a solution. These are effectively two capacitances in series, one with a fixed value representing the Stern/Helmholtz layer (C_H) and one representing the diffuse layer (C_d) which will vary with the solution concentration. The characteristic thickness of the double-layer is defined by the Debye length which is inversely proportional to the square root of the concentration. Therefore, the double layer reduces in thickness as the concentration increases and, as it is considered to have a constant permeability, then the capacitance of the layer will also increase. Therefore, C_{dif} will vary with changes in the ionic concentration of a solution that are unrelated to pH, and this will change α and affect the ISFET performance.

The intrinsic buffer capacity (β_{int}) is a measure of the ability of the oxide surface to accept or donate protons to and from the solution. It should be maximised to increase the sensitivity by making closer to unity. It turns out that SiO_2 is not the best material to use for an ISFET as the buffer capacity is relatively low. This means the sensitivity is sub-Nernstian and the sensor will be more sensitive to changes in C_{dif} due to variation of ionic concentration. Other materials such as silicon nitride (Si_3N_4), aluminium oxide (Al_2O_3) and tantalum pentoxide (Ta_2O_5) are better for a variety of reasons and have higher β_{int} and α . One reason for the better performance of the aluminium and tantalum oxides might be the greater number of oxygen atoms involved in the oxides, meaning a greater density of amphoteric sites at the surface. Silicon nitride in an aqueous solution will develop an oxidised surface with the hydroxyl sites required for pH sensing.

The result of the pH sensitivity of the flatband voltage is a similar dependence in the threshold voltage of the device. If the other parameters are constant then the threshold voltage will have the same dependence on pH as the surface potential:

$$\frac{\delta V_T}{\delta \text{pH}_B} = -2.3 \frac{kT}{q} \alpha \quad (7.43)$$

This means that the ISFET can be controlled by the solution pH, though setting it up for a measurement can be quite complicated and will be discussed later in this chapter.

7.8.3 ISFET Fabrication

There are a number of issues to be considered when deciding how to fabricate ISFET sensors. Standard CMOS often uses a self aligned process where the gate electrode is fabricated before the source and drain regions. These are then produced by implantation of dopants to form the source and drain, with the gate electrode and field isolation acting as a mask. This makes integrating an ISFET in this process a problem as it is almost impossible to remove the gate electrode afterwards without damaging the thin gate oxide. What is required is either a custom process that uses some other method to define the S/D regions or a completely different way to make ISFETs. Probably the most common method that has been developed to make CMOS compatible integrated ISFETs is to use the metallisation layers available in the process to effectively bring the gate connection up to the surface of the integrated circuit. There the top passivation layer of silicon nitride or silicon oxy-nitride is ideal to act as a pH sensitive membrane without having to alter the process. Figure 7.29 is a schematic cross-section through a CMOS compatible ISFET, sometimes known as an extended gate field

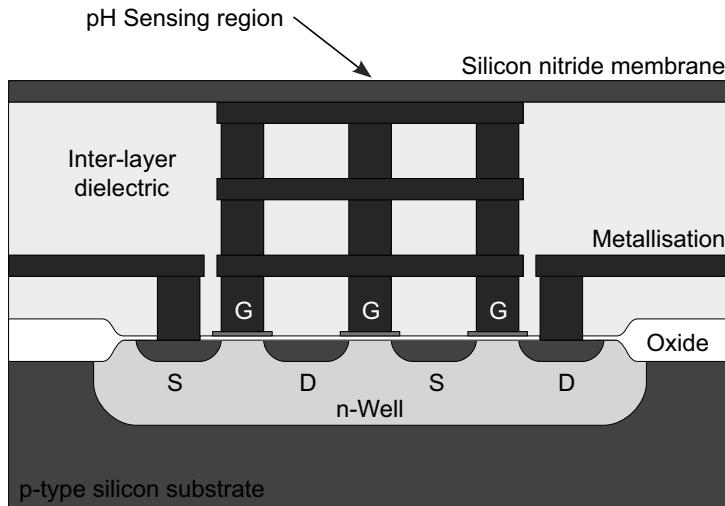


Figure 7.29 Schematic structure of a bare gate ISFET pH sensor.

effect transistor (EGFET), similar to a device described in [4]. The original concept for this type of EGFET structure was first published in [5].

7.8.4 ISFET Instrumentation

The standard method for measuring the ISFET is to bias it in the linear region of operation. V_{DS} is set to be some small, constant value so that the transistor channel is not pinched-off. Then, the current through the device, I_{DS} , is also controlled and held at a constant value. Equation 7.34, for the drain current in the linear region, can be rearranged to give an equation for the gate-source voltage, V_{GS} :

$$V_{GS} = V_T + \frac{1}{V_{DS}} \left(I_{DS} + \frac{V_{DS}^2}{2\beta} \right). \quad (7.44)$$

If I_{DS} and V_{DS} are constant then V_{GS} will be dependent on V_T and if this increases due to a change in the pH of the solution, then V_{GS} will also have to increase. Achieving this experimental arrangement requires some sort of electronic feedback and an example originally developed by Bergveld is shown in Figure 7.30 [6].

This circuit, often referred to as an ISFET amplifier, is similar in part to an instrumentation amplifier where the ISFET replaces the resistor that sets the gain of the circuit. The input to the instrumentation amplifier is set by the resistor R_{DS} and the constant current I_{in} , which also controls the voltage across the ISFET such that $V_{DS} = I_{in} R_{DS}$. If the threshold voltage V_T drops due to an increase in pH then the effective resistance of the ISFET channel will decrease and the output of the instrumentation amplifier, which is inversely proportional to this, will increase. If this is now larger than V_{ref} the output of the output op-amp will increase feeding a current through R_S that will raise the source voltage. This then reduces the effective gate-source voltage, which is defined by the reference electrode in the solution, and raises the effective channel resistance. This feedback will ensure that the source drain current I_{DS}

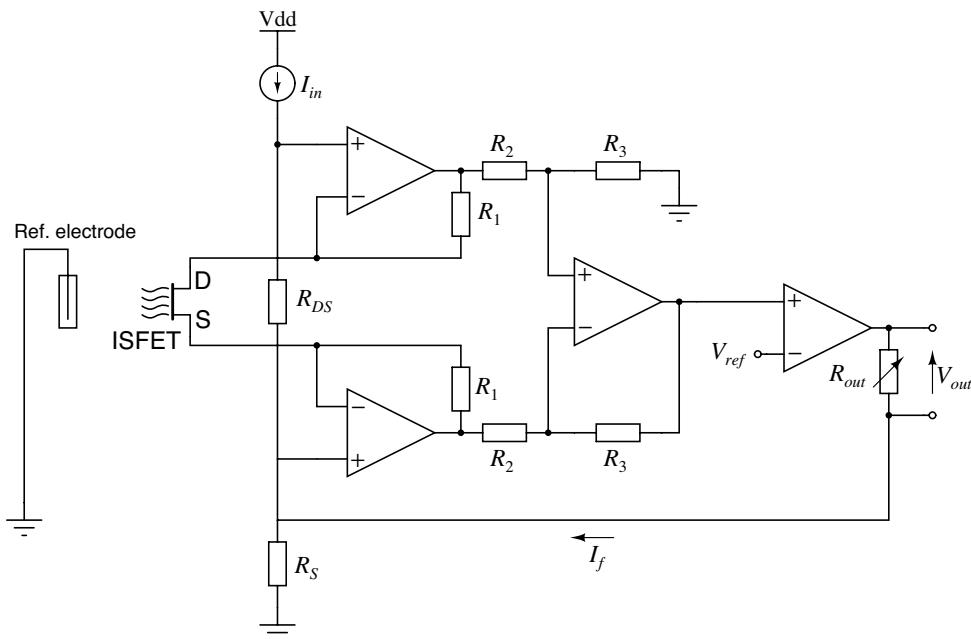


Figure 7.30 ISFET amplifier circuit based on the principle of the instrumentation amplifier.

of the ISFET is a constant value, controlled by V_{ref} . The result of this is that the change in the source voltage will be the same as the change in V_T but with an opposite sign while the output will be this value amplified by the ratio of R_{out} and R_S . The output of the circuit will be:

$$V_{out} = I_f R_{out} = \delta V_S \frac{R_{out}}{R_S} = -\delta V_T \frac{R_{out}}{R_S}. \quad (7.45)$$

Setting up this system is relatively straightforward. The ISFET sensor is placed into a buffer with a known pH, typically a pH of 7, before the value of V_{ref} is adjusted to set the output voltage to 0 V. This effectively sets the current through the ISFET to whatever value is required to make the instrumentation amplifier output equal to V_{ref} . Then, the sensitivity of the output is set by adjusting R_{out} and this calibration could be performed with another known buffer solution. For example, the ISFET could be placed in a solution with a pH of 4 and R_{out} could be adjusted to give $V_{out} = 3$ V for a sensitivity of 1 V pH⁻¹.

7.8.5 The REFET

Correct operation of an ISFET requires a good reference electrode and these are difficult to produce as a miniaturised, integrated device. Unfortunately there is no solid-state reference electrode. There are integrated quasi-reference electrodes which can operate in a limited range of conditions and one of the most common for biosensing applications is the Ag/AgCl electrode. The REFET is not a FET based reference electrode but a possible way to do without one in ISFET sensing applications.

The REFET is a device identical to the ISFET to be used in the sensor application but with some additional layer over the pH sensing area which blocks the transport of protons to and

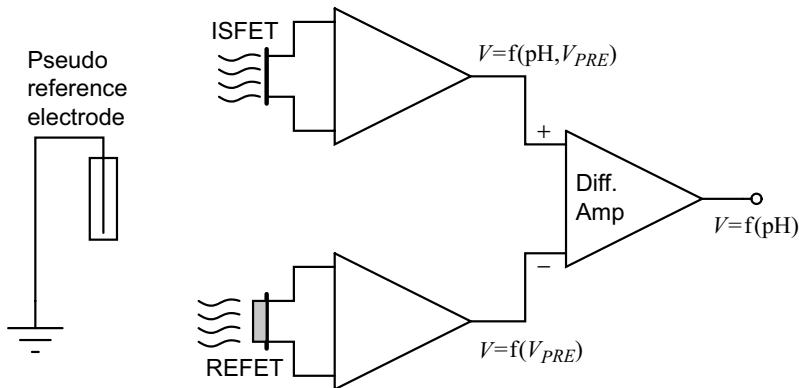


Figure 7.31 Block diagram of an ISFET/REFET measurement system [7].

from the surface. This will potentially allow the use of a pseudo-reference electrode which does not have a known, fixed potential. A possible example of this is a platinum electrode. The output of a source drain follower using the ISFET will be a function of the pH and the pseudo-reference electrode potential V_{PRE} while the output of an identical circuit using a perfect REFET will simply be a function of V_{PRE} . Feeding these into a differential amplifier, as shown in Figure 7.31, will result in an output that is solely a function of the solution pH [7]. Unfortunately there have been many problems in producing an ideal REFET where the application of the blocking layer does not affect the matching between the two sensors. It will typically change the threshold voltage of the REFET and can also lead to drift if the blocking layer can become contaminated by the solution being measured.

7.8.6 ISFET Problems

There are also problems, or rather issues to be taken into consideration, with ISFETs in general. They are obviously temperature dependent through the inclusion of the thermal voltage in many terms that feed into the response so they should be used in conjunction with a temperature sensor. Drift is also a huge problem with silicon nitride sensing layers tending to change to include more oxygen over time when they are in aqueous solutions. The typical solution for this is to measure the drift and correct for it over time which is not very elegant and controllable although it seems to be an accepted method, particularly for the few commercial ISFET sensors available on the market. Another general problem with electronic sensors in aqueous solutions is making sure of good sealing and encapsulation of electrical contacts to prevent corrosion and other impacts on the stability and operation of the device.

7.8.7 Other FET Based Sensors

The ISFET can be made sensitive to other ions by changing the gate dielectric to a different material, or by applying it as a coating to a standard pH ISFET. It is also possible to make an FET sensor for non-ionic chemistries with a sensing layer that changes the local pH or changes the charge distribution at the gate surface so that it can be sensed with a FET.

Enzyme based FET sensors will use a similar operating principle and may use the same type of enzymes used to make conductivity sensors, such as urease. The enzymatic reaction

must produce protons for sensing with a pH sensitive ISFET. A REFET type differential measurement can be obtained by combining a standard ISFET with the ENFET. There are issues with the sensitivity being very non-linear, making quantitative measurements difficult, and there are methods using electrochemical feedback to adjust the local pH in the ENFET to make it more controllable [7].

Other FET-based sensors in the literature use biological sensing methods varying from antibody attachment in an immuno-FET, DNA binding or right up in scale to whole cell sensing. Binding of charged molecules like DNA is particularly suitable for FET based sensing and a review of bio-FET sensors can be found in [8].

Problems

- 7.1. Two important characteristics of a sensor system are accuracy and precision.
 - (a) Explain these terms with the help of a diagram.
 - (b) Think about how you might attempt to determine them for a particular sensor. What methods would you use?
 - (c) How could you improve the accuracy and precision of a measurement?
- 7.2. A glass electrode based potentiometric pH sensor operating in its linear range has a response to pH defined by the following equation:

$$E = E^0 - \frac{2.303 RT}{F} \text{pH(V)}.$$

- (a) Explain the terms of this equation.
- (b) Draw the equivalent circuit of a glass electrode with resistance R .
- (c) The characteristic response of a typical glass electrode pH metre is shown in Figure 7.32. The linear region helpfully includes the most useful range of pH values.

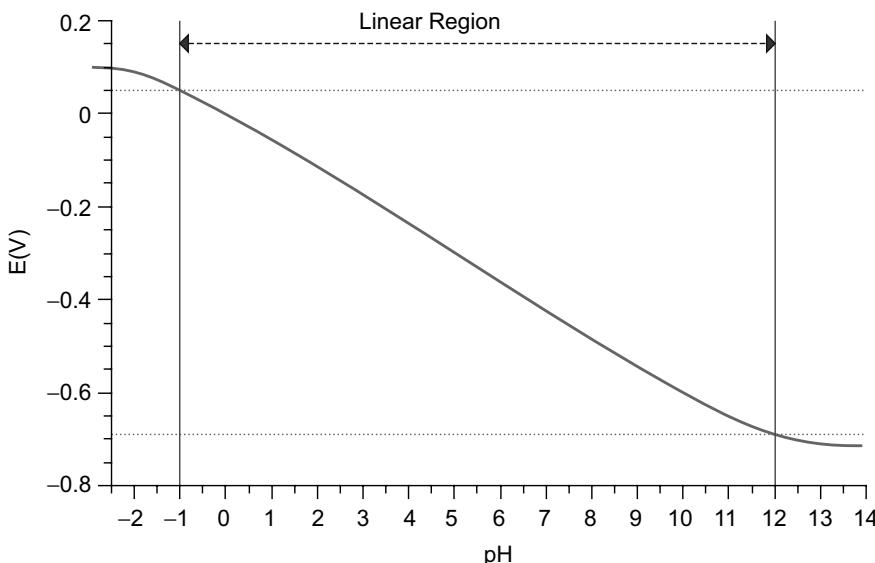


Figure 7.32 Characteristic transfer function of a glass electrode pH sensor.

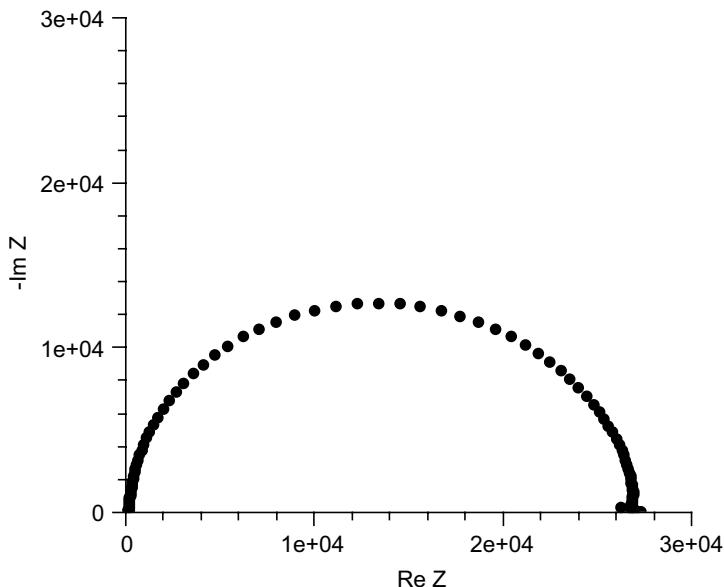


Figure 7.33 Electrochemical impedance spectroscopy data from an organic light emitting diode [9].

Assuming that the sensor will be connected to a voltmeter and that the specification for the next stage requires that the output be equal to the actual pH, design an amplifier for this electrode using an operational amplifier. Note that the output resistance of a glass electrode pH sensor may be as high as $500\text{ M}\Omega$.

- (d) If the op-amps used in your amplifier circuit have input bias currents of 1 pA what is the error on the output? What would it be if the bias current was as high as 1 nA ?
- 7.3. What voltage and current will be required to be supplied to the counter electrode of an electrochemical cell in order to charge the double layer capacitance C_{dl} of the working electrode by 1 V in 10^{-5} seconds? Refer to the equivalent circuit in Figure 7.14(b) and assume that: $C_{dl} = 10\text{ }\mu\text{F}$, $R_s = 100\text{ }\Omega$, $R_u = 0\text{ }\Omega$.
- 7.4. You have been given a conductometric biosensor consisting of interdigitated micro-electrodes coated with a enzyme which reduces in resistivity when exposed to a specific analyte. Suggest a suitable circuit to convert the resistance to a voltage. Explain your choice with reference to the characteristics of the sensor and the resistance to voltage converter. Assuming the sensor has an initial resistance R_0 of $1\text{ k}\Omega$ that varies with δR of up to 10 in the presence of the analyte, design a circuit to produce a voltage signal in the range of 0 to 5 V.
- 7.5. Figure 7.33 shows some real data from an EIS experiment. This is not an electrochemical sensor but an organic light emitting diode which operates in a similar manner [9]. The plot resembles that for a resistor in parallel with a capacitor as in Figure 7.21(b). Estimate the values for the resistance and capacitance in the equivalent circuit if the frequency at the maximum value of $\text{Im}(Z)$ is 1.1 kHz .

References

- [1] Katz, E. and Willner, I. (2003) Probing biomolecular interactions at conductive and semiconductive surfaces by impedance spectroscopy: Routes to impedimetric immunosensors, DNA-Sensors, and enzyme biosensors. *Electroanalysis*, **15** (11), 913–947.
- [2] Bergveld, P. (1970) Development of an ion-sensitive solid-state device for neurophysiological measurements. *Biomedical Engineering, IEEE Transactions on*, **BME-17** (1), 70–71.
- [3] Bergveld, P. (2003) ISFET, theory and practice. IEEE Sensor Conference Tutorial, pages 1–26.
- [4] Hammond, P.A., Ali, D. and Cumming, D.R.S. (2004) Design of a single-chip pH sensor using a conventional 0.6- μ m CMOS process. *Sensors Journal, IEEE*, **4** (6), 706–712.
- [5] Bausells, J., Carrabina, J., Errachid, A. and Merlos, A. (1999) Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. *Sensors & Actuators: B. Chemical*, **57** (1–3), 56–62.
- [6] Bergveld, P. (1981) The operation of an ISFET as an electronic device. *Sensors and Actuators*, **1**, 17–29.
- [7] Bergveld, P. (2003) Thirty years of ISFETOLOGY. *Sensors & Actuators: B. Chemical*, **88** (1), 1–20.
- [8] Schonning, M.J. and Poghossian, A. (2006) Bio fets (field-effect devices): State-of-the-art and new directions. *Electroanalysis*, **18**, 1893–1900.
- [9] Cummins, G., Underwood, I. and Walton, A.J. (2011) Electrical characterization and modelling of top-emitting PIN-OLEDs. *Journal of the Society for Information Display*, **19** (4), 360.

Further Readings

- Wilson, J.S. (ed.) (2005) *Sensor Technology Handbook*, Elsevier, ISBN 978-0750677295.
- Fraden, J. (2010) *Handbook of modern sensors: Physics, designs, and applications*, Springer, ISBN 978-1441964656.
- Southampton Electrochemistry Group (2001) *Instrumental Methods in Electrochemistry*, Woodhead Publishing, ISBN 978-1898563808.
- Bard, A.J. and Faulkner, L.R. (2001) *Electrochemical Methods - Fundamentals and Applications*, John Wiley & Sons, Ltd. ISBN 978-0471043720.
- Keithley Instruments (2004) *LowLevel Measurements Handbook*, 6th edn, Keithley Instruments, Inc. Free book request from <http://www.keithley.com/promo/wb/259>.

8

Instrumentation for Other Sensor Technologies

8.1 Chapter Overview

This chapter begins by detailing the reasons why temperature measurement is important for biosensing applications. It then describes methods for temperature sensing using resistive temperature detectors and silicon p-n junction diodes. These are considered to be the most relevant methods for temperature calibration of biosensors, particularly when microminiaturisation is required.

The second section of the chapter concentrates on different methods for mechanical sensing using three different modes of measurement: piezoresistive, piezoelectric, and capacitive sensing. Examples of applications in biosensing or biomedical measurement are included. The section on capacitive sensing leads naturally into an introduction to switched capacitor based instrumentation.

Fluorescent sensing is key to many biological and biomedical applications and the third part of this chapter will review the theory of fluorescence before describing photodetectors that can be used to transduce optical information into electrical signals. This section finishes with a discussion of a lab-on-a-chip biosensor which uses fluorescent readout and electrochemical actuation for detection of specific DNA sequences.

The final part of this chapter examines instrumentation for electrophysiology, and in particular the methods used to measure electrical activity in neurons. These techniques are extremely important for researchers attempting to understand the function of the nervous system and to develop treatments for neurological diseases.

After reading this chapter readers will gain a refreshed or new understanding of:

- (1) the requirement for temperature sensing in biosensor applications;
- (2) the application of resistors and junction diodes as electronic temperature sensors;

- (3) different modes of operation for mechanical sensors and the different instrumentation used to measure them
- (4) basics of optical transduction in biosensor technology;
- (5) applications of electronic sensors in neuroscience and electrophysiology.

8.2 Temperature Sensors and Instrumentation

8.2.1 Temperature Calibration

One very important reason to integrate temperature measurements alongside other sensors is that in most cases there will be some cross-correlation between the output of a sensor and the temperature. This can be as simple as thermal variations in the value of a resistive sensor or the dependence of the chemical reactions in a biosensor on the ambient temperature. In addition some of the biological components of a biosensor system, such as enzymes, will often have a very narrow range of temperatures over which they operate. This means that the temperature sensors do not need to have a wide range but do need to be sensitive and accurate within this range.

Some common temperature sensors or instruments such as a glass thermometer or a bimetallic strip are not well suited to miniaturisation. In addition it is often useful to choose a transducer for temperature that is similar in nature to the sensor it will be used alongside. For example if the transduction method in a biosensor involves measuring a change in resistance it might be helpful if the temperature sensor is also resistive.

8.2.2 Resistance Temperature Detectors

Probably the most basic electrical or electronic temperature sensor is a simple resistive transducer which exploits the temperature dependence of the resistivity of most conductive materials. The basic formula for the resistance (R_T) of a Resistance Temperature Detector (RTD) at a temperature T is:

$$R_T = R_0(1 + \alpha(T - T_0)). \quad (8.1)$$

This will depend on the resistance (R_0) at some known calibration point T_0 . This is often 0 °C but it could simply be the midpoint of the temperature range of interest. The most important factor is the temperature coefficient α which will depend on the resistor material and may be positive or negative. Metals in a solid state will generally have a positive coefficient of temperature. The units of α are K⁻¹.

One of the most common materials for RTDs is platinum and sensors made with this are usually referred to as platinum resistive thermometers or PRTs. Platinum is used because it is stable and proof against most contamination and corrosion that can affect the temperature coefficient, which will depend on the purity of the RTD material. Pt can be deposited in thin films and micromachined to produce narrow resistive tracks, for example on the surface of an integrated circuit. One standard type of PRT has a purity of platinum with a characteristic sensitivity of $\sim 0.385 \Omega \text{ K}^{-1}$ when made into a resistor with a value of 100 Ω at 0 °C.

RTDs can be made as a thin film but a more common construction for a discrete sensor has a rigid former, made of a ceramic for example, around which a platinum wire is wound. More modern sensors may have coils of wire within a rigid, insulating support. The purpose

of the former is mainly to support the wire and prevent the resistance of the RTD being altered by thermal expansion or other mechanical forces.

The variation of the resistance of a platinum RTD is not completely linear and a more complete description of the characteristic is given by the Callendar-Van Dusen equation, which comes in two forms depending on the temperature range being investigated:

$$R_T = R_0[1 + AT + BT^2 + (T - 100)CT^3] \quad (8.2)$$

for temperatures $-200^\circ\text{C} < T < 0^\circ\text{C}$ and:

$$R_T = R_0[1 + AT + BT^2] \quad (8.3)$$

for $0^\circ\text{C} < T < 661^\circ\text{C}$.

These equations contain higher order coefficients and these can be found by calibrating the individual resistor. The accepted values for a ‘pt100’ RTD with a standard temperature coefficient are:

$$A = 3.908310^{-3}\text{K}^{-1}$$

$$B = -5.77510^{-7}\text{K}^{-2}$$

$$C = -4.18310^{-12}\text{K}^{-4}.$$

Over the most typical temperature ranges the resistance is effectively linear and the 1st order equation (8.1) can be used with $\alpha = 0.00385\text{ K}^{-1}$ and $R_0 = 100\Omega$ at 0°C .

A common method to convert the resistance variation of an RTD sensor into a useful voltage signal is to place it into a Wheatstone bridge circuit (see Chapter 7, Section 7.7.1) as shown in Figure 8.1.

The variation in the output voltage δV_o will be approximately proportional to the change in the resistance δR of the RTD. The main problem with this approach is that the resistances of the leads connecting the RTD are included in the measurement. That is not a significant issue when setting up the circuit, as the other resistors in the bridge can be tuned to balance the additional resistance of the leads. The problem arises when the sensor is used, as the lead resistance will also have some dependence on the temperature which will appear as an error in the temperature measurement if it varies significantly.

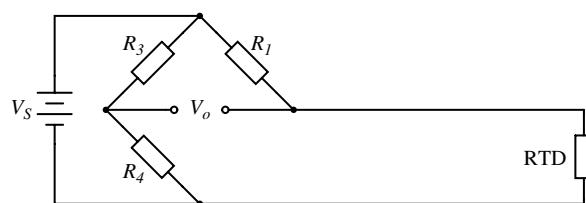


Figure 8.1 Resistive temperature detector in a Wheatstone bridge circuit.

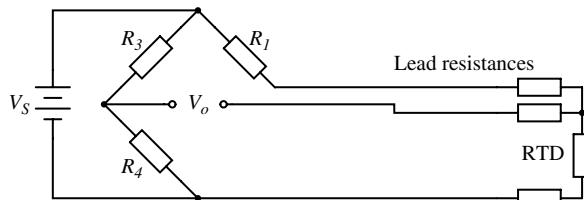


Figure 8.2 RTD bridge circuit with 3-wire compensation for connection resistance.

There are a number of different ways of compensating for the connection resistance. One of the most common methods involves using an extra lead to bring the output connection right down to where it connects to one side of the resistor (Figure 8.2). This lead should not carry much current unless the bridge is very unbalanced, and the outer resistances will cancel each other out in the bridge assuming the leads have well matched resistances. A more elaborate, 4-wire connection scheme adds a pair of dummy leads into the R_1 section of the bridge (Figure 8.3) but this still depends on the leads being matched and is costly in terms of the amount of cabling required.

If accurate current sourcing and voltage measurement is available it is possible to do away with the bridge entirely and instead use a Kelvin or four-terminal measurement setup. Here the measurement current is forced through the resistive sensor while the voltage is measured with separate connections made as close as possible to the ends of the resistor (Figure 8.4). If the voltmeter has a high input impedance then little current will flow in these connections and the voltage measured will be solely due to the resistance of the sensor. This technique is named for Lord Kelvin (William Thompson) after his invention of the Kelvin or Thompson double bridge which applies a similar technique to the resistive Wheatstone bridge circuit.

There are a number of important issues that need to be considered when using an RTD to measure temperature. Firstly, it is a passive sensor meaning that current needs to be supplied in order to make the measurement. If the current used is too high this can cause self-heating of the resistor and an error in the measurement. This implies that low currents are required but that will mean that there is a requirement to measure or amplify relatively small voltages. Secondly, the sensitivity of an RTD is the product of R_0 and the temperature coefficient α , suggesting that it might be useful to use a large value of resistance. Typically, a standard RTD sensor has $R_0 = 100 \Omega$ at 0°C but it may be possible when designing an integrated sensor to use a larger resistance. However, this would obviously be balanced by an increase

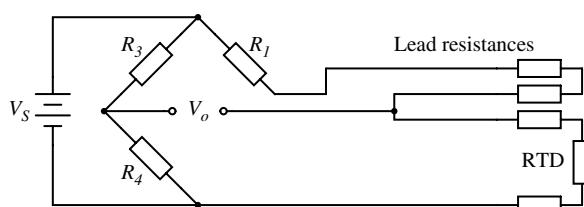


Figure 8.3 RTD bridge circuit with 4-wire compensation for connection resistance.

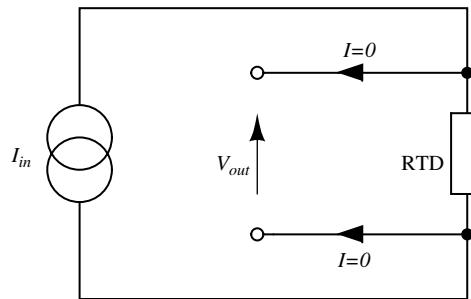


Figure 8.4 Resistance temperature detector with four-terminal Kelvin measurement.

in thermal noise from the RTD. Thirdly, all RTDs will have some sort of time constant reflecting how quickly they react to changes in the temperature of the measured environment. This will depend on the thermal conduction properties of the material of the resistor and of the packaging used. Heat sinks such as the wires connecting the RTD may also have an effect. Finally, the long-term stability of the sensor will depend on preventing contamination of the resistor material and avoiding strain gauge effects due to applied mechanical forces. The resistance to contamination or corrosion of platinum is one reason why it is chosen for this application.

Thermistors are similar to RTDs in that they are temperature dependent resistors but the difference is that they are usually not made from metal but from a conducting ceramic, polymer or semiconductor. Unlike metal RTDs they can have either a positive or negative thermal coefficient and can also have a non-linear response. For example, most semiconductors will have a negative thermal coefficient of resistance because heating will tend to increase the levels of free charge carriers. Thermistors with highly non-linear responses can be useful for application as temperature dependent switches.

8.2.3 p-n Junction Diode as a Temperature Sensor

Solid-state electronic temperature sensors represent an alternative solution for the measurement of temperature in integrated, microscale biosensing applications. The simplest method is to use a forward biased p-n junction diode as a sensor. Most readers should be somewhat familiar with the Shockley diode equation given below, which defines the current in the device as a function of the voltage bias, V_D , the reverse bias saturation current, $I_{D(sat)}$, the temperature, T , and some other physical constants:

$$I_D = I_{D(sat)} \left(e^{\frac{qV_D}{nkT}} - 1 \right). \quad (8.4)$$

If, instead of applying a bias voltage a constant current is forced through the diode the equation can be rearranged to give the voltage across the diode, assuming that $I_D \gg I_{D(sat)}$:

$$V_D \approx \left(\frac{nkT}{q} \right) \ln \left(\frac{I_D}{I_{D(sat)}} \right). \quad (8.5)$$

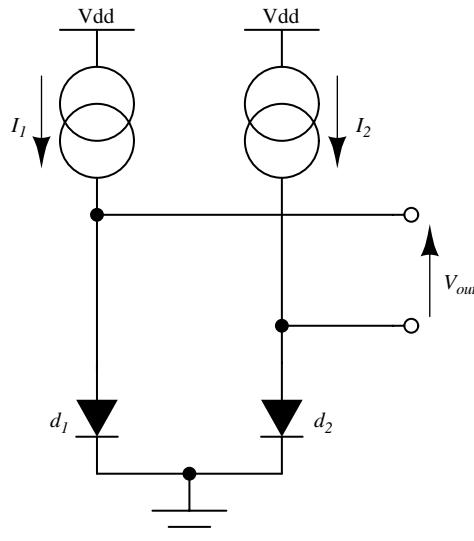


Figure 8.5 Proportional To Absolute Temperature (PTAT) sensor using p-n diodes.

This appears at first glance to have a positive, linear dependence on temperature, due to the T in the term for the thermal voltage, but this does not take into account the temperature dependence of the saturation voltage $I_{D(sat)}$. The overall effect of this is that the diode voltage is proportional to temperature with a sensitivity of about -2 mV K^{-1} for a typical silicon p-n junction.

One disadvantage of the single diode as a sensor for temperature is that, although the sensitivity is well known, without calibration it is difficult to say what the temperature actually is. Small variations in the diode fabrication process can lead to significant variations in the saturation current. A differential measurement of two diodes with different currents can provide an absolute temperature measurement with a positive linear coefficient. If two different currents (I_1, I_2) are forced through two similar diodes as shown in Figure 8.5 the resulting difference in voltage drop across them is:

$$V_{out} = \Delta V_D = \frac{kT}{q} \ln\left(\frac{I_1}{I_2}\right). \quad (8.6)$$

Equation 8.6 can be rearranged to give an expression for the absolute value of temperature in terms of the voltages, currents and physical parameters. The sensitivity of this system depends upon the ratio of the two currents, and a typical value for this is 10 : 1, which gives an overall sensitivity of $198 \mu\text{V K}^{-1}$. If the ratio of the currents is too high then significant errors can occur due to different levels of self-heating in the devices. It is also possible to take a single diode and alternately switch two different currents through it. The amplitude of the resulting ac voltage will have the same proportional to absolute temperature (PTAT) dependence as the circuit shown in Figure 8.5.

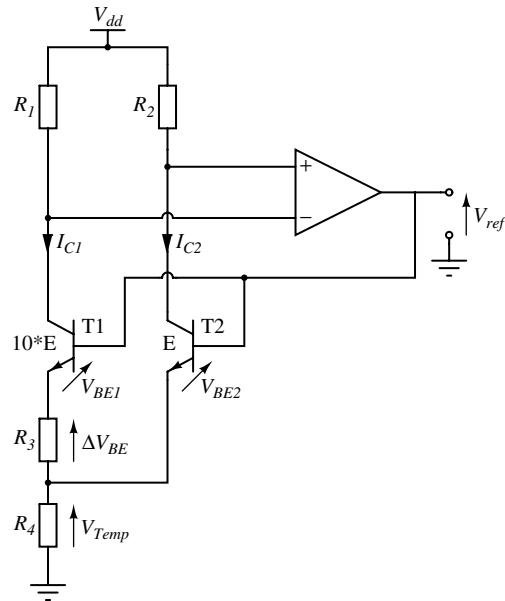


Figure 8.6 Brokaw bandgap voltage reference circuit.

The temperature dependences of silicon p-n junctions can be exploited to make a voltage reference circuit with an output that is insensitive to temperature. By combining parts with positive and negative temperature coefficients it is possible eliminate the temperature dependency. These circuits are often referred to as bandgap references because the output is set by the physical property of the silicon bandgap potential. Figure 8.6 shows an example of one standard voltage reference known as a Brokaw reference circuit [1]. The feedback loop made up from the operational amplifier and the other components will hold the voltages across the two equal resistances R_1 and R_2 equal so that the collector-emitter currents (I_{CE1} , I_{CE2}) are the same in both transistors. The two transistors will have different emitter areas with a ratio of 10 : 1 so that the base emitter voltages, which depend on the current densities, will be different. These junction voltages (V_{BE1} , V_{BE2}) will have a negative coefficient of temperature while the difference between V_{BE1} and V_{BE2} (ΔV_{BE}) will have a positive coefficient as in the diode based temperature sensor. The voltage across R_3 will be:

$$V_{R_3} = \Delta V_{BE} = \frac{kT}{q} \ln\left(\frac{J_{CE2}}{J_{CE1}}\right), \quad (8.7)$$

where J_{CE1} and J_{CE2} are the current densities in the transistors. A temperature dependent voltage can be read out from the voltage across R_4 which is:

$$V_{R_4} = 2 \frac{R_4 kT}{R_3 q} \ln\left(\frac{J_{CE2}}{J_{CE1}}\right). \quad (8.8)$$

The output voltage $V_{ref} = V_{R_4} - V_{BE2}$. Careful choice of R_3 and R_4 will balance the positive and negative temperature coefficients to give a constant output V_{ref} that is independent of temperature and is close to the bandgap voltage of silicon, typically around 1.25 V.

8.3 Mechanical Sensor Interfaces

This section will examine three different types of mechanical transduction with application in biosensing but it should be understood that there are a number of other modes of operation for micromechanical sensors and actuators, such as the use of magnetism. The three different transduction methods discussed here are:

- (1) piezoresistive effect;
- (2) piezoelectric effect;
- (3) capacitive coupling.

8.3.1 Piezoresistive Effect

Piezoresistance is probably the simplest way to transduce movement into an electrical signal. It is a change in the resistance of a piece of material that is subjected to a mechanical force and was discovered by Lord Kelvin around 1856. It is also referred to as the strain gauge effect and most electrically readable strain gauges will operate in this way. It lends itself to microfabrication as thin film deposited or diffused resistors make good strain gauges. There is also a greater effect in semiconducting materials than there is in metallic conductors.

A typical strain gauge is fabricated in a thin metal layer on a flexible support, probably an insulating polymer. A typical design is illustrated in Figure 8.7 where the gauge will be sensitive to strain applied along the long axis of the structure. Stretching it in this direction will put the conducting material into tension, making it longer and thinner, and therefore increasing the resistance. Similarly putting it into compression, will make the conducting material shorter and wider in cross-section so that the resistance decreases.

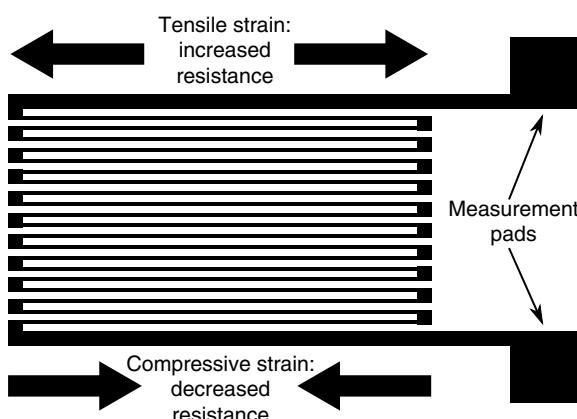


Figure 8.7 Schematic design for a piezoresistive strain gauge.

A piezoresistive sensor is characterised by its gauge factor, GF ; this is a dimensionless parameter defining the relative change in the resistance, ΔR , divided by the applied strain, ε :

$$GF = \frac{\Delta R}{\varepsilon R}. \quad (8.9)$$

The basic formula for resistance is $R = \rho L/A$ and from this is possible to define the fractional change in resistivity for a strain gauge:

$$\frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} - \frac{\Delta A}{A} + \frac{\Delta L}{L}. \quad (8.10)$$

The first term refers to the change in the material resistivity, ρ , due to the strain. The piezoresistive coefficient π is defined as:

$$\pi = \frac{\Delta \rho / \rho}{E \varepsilon}, \quad (8.11)$$

where E is the Young's modulus of elasticity for the material and ε is the applied strain ($\varepsilon = \Delta L/L$). Another important parameter is Poisson's ratio, ν , which defines the change in the transverse strain due to an applied axial strain. In most materials when they are stretched in one direction the cross sectional area will be reduced in the transverse directions and it is possible to define Poisson's ratio with this formula:

$$\frac{\Delta A / A}{\Delta L / L} = -2\nu. \quad (8.12)$$

By combining Equations (8.10) to (8.12) with the expression for the gauge factor (8.9) it is possible to define GF in terms of the material properties (ν , π and E):

$$GF = 1 + 2\nu + \pi E. \quad (8.13)$$

In most metals the Poisson ratio is between 0 and 0.5 while the change in resistivity with strain is negligible. This means that the maximum GF for a typical metal strain gauge is ~ 2 . With a good design strains of up to 10% can be measured but this requires a firm connection between the gauge and the structure being measured. For example, slippage in the glue used to stick down the gauge can reduce the coupling of strain and therefore the sensitivity.

The biggest issue to control in strain gauges is the effect of temperature as all materials used in strain gauges will have a resistivity that varies with temperature, as we have seen in Section 8.2.2. In addition there will be effects from thermal expansion of the gauge materials and this can add to the mechanical strain being measured. It is possible to make the gauge out of an alloy like constantan which has very low thermal coefficients of resistance and expansion.

There are a number of different ways to arrange strain gauges into a Wheatstone bridge, which can also serve to reduce temperature effects. The simplest is to replace two of the resistors in one side of the bridge with strain gauges but only one of the gauges should be fixed to the surface being measured. The other will be a dummy gauge that experiences the same thermal environment as the active strain gauge but does not change with applied force. The other two resistors should be chosen to have the same resistance of the unstrained

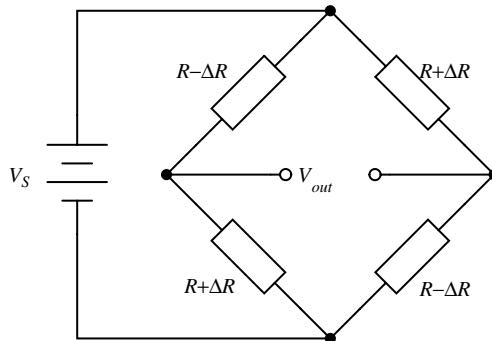


Figure 8.8 Arrangement of strain gauges in a Wheatstone bridge circuit for maximum sensitivity.

gauges but may not be located near to the measurement. In an alternative setup is possible to use four gauges at the same time but arrange them so that two, opposite, gauges will be compressed, reducing R and two will be in tension, increasing R . This gives a much larger imbalance to the bridge for the same input so the resulting output voltage sensitivity to strain will be maximised. This is illustrated in Figure 8.8 which will have an output voltage:

$$V_{out} = V_S \frac{\Delta R}{R}. \quad (8.14)$$

It is very important to ensure that the gauges are properly arranged in the bridge circuit so that they do not cancel each other out and give no output at all. A mistake like this is the source of ‘Murphy’s Law’ which is named for the US Airforce Engineer, Capt. Edward A. Murphy Jr. The full story of how this infamous law, which is often misstated as ‘If anything can go wrong it will’ but was originally ‘If it can happen, it will happen’, was coined is set out in [2,3].

The piezoresistive effect is generally greater in semiconductors than in metals due to effects of strain on the mobility of charge carriers in the semiconductor. This means that the piezoresistive coefficient π dominates in Equation (8.13) resulting in gauge factors of 100 or more in single crystal silicon diffused resistors. The direction of the strain is very important in these devices as the value of π will depend on the orientation of the crystal planes in these devices. Polysilicon piezoresistors will display less dependence on orientation but with the trade-off of a lower gauge factor. These mobility effects of strain are exploited in modern microelectronics. For example in advanced CMOS technologies the lattice mismatch between Si and SiGe is used to create strain in the channel region of transistors to improve the mobility.

8.3.2 Applications of Piezoresistive Sensing

Silicon or polysilicon strain gauges are ideal for integration into a MEMS sensor system and this section presents two examples with possible medical application.

The first of these is a pressure sensor consisting of a micromachined cavity sealed with a flexible membrane. This could be fabricated from a variety of different materials depending on the mechanical properties required. Pressure differences between the sealed cavity and the surrounding atmosphere will lead to flexing of the membrane, which can be detected

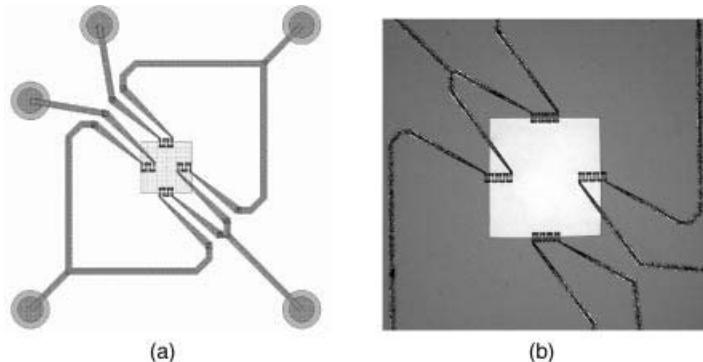


Figure 8.9 (a) Schematic layout of part of a MEMS pressure sensor with piezoresistive sensing. (b) Photomicrograph of a similar MEMS sensor.

using polysilicon strain gauges in a Wheatstone bridge arrangement. The layout of the device is shown in Figure 8.9 alongside an image of the fabricated structure.

The polysilicon piezoresistive strain gauges are short vertical tracks arranged in serpentine meanders. In Figure 8.9a there are four resistors in each arm of the bridge while in Figure 8.9b there are 8 resistors in each arm. The gauges are laid out so that two sets of resistors straddle the edge of the membrane (the central square in Figure 8.9a) while the other two are parallel to the edge and should not see significant stress. This is a type of ‘half-bridge’ sensor interface where only two resistors change with applied stress while the other two should be unaffected. Therefore, if the change in resistance in the gauges that are under stress is ΔR and $\Delta R \ll R$ then the output voltage from the bridge will be:

$$V_{out} = V_S \frac{\Delta R}{2R}. \quad (8.15)$$

The Wheatstone bridge wiring includes a break in one arm which allows the addition of a variable resistor, as shown in Figure 8.10a, to balance the bridge at some set pressure. It

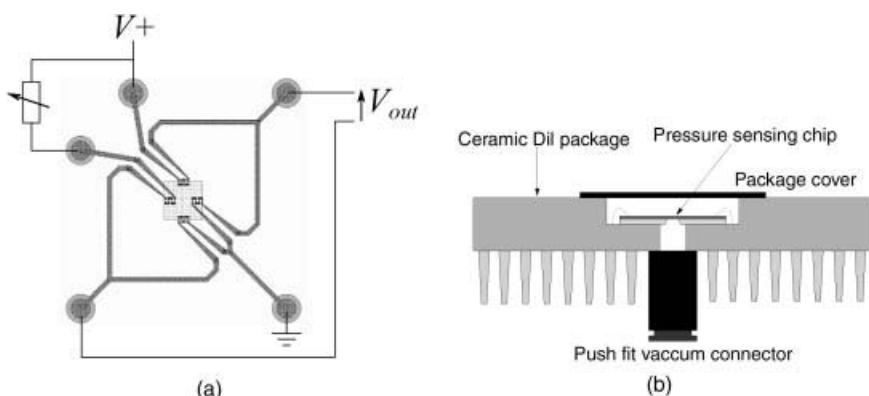


Figure 8.10 (a) Use of a variable resistor in a pressure sensor Wheatstone bridge to balance the circuit. (b) Schematic cross-section through a packaged pressure sensor.

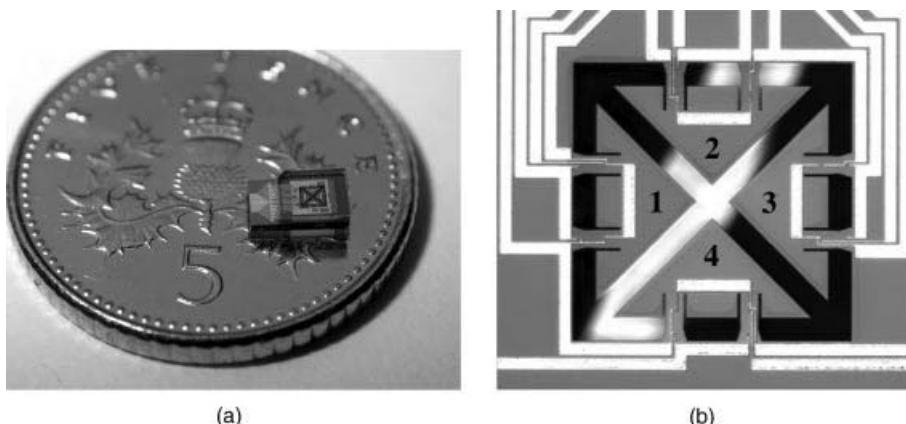


Figure 8.11 (a) Packaged MEMS accelerometer placed on a five penny coin (approximately 18 mm diameter) to show scale (courtesy of C. Lowrie and M. Desmulliez, Heriot-Watt University). (b) Microscope image of a MEMS accelerometer (reproduced from [5] with permission of the authors).

could then be connected up to a differential amplifier or instrumentation amplifier to boost the output signal. Figure 8.10b shows how the chip was packaged by bonding it into a standard integrated circuit package which has had a hole bored through it with a high speed drill. The pressure can then be adjusted by applying a vacuum to the back of the device via a standard push fit connector. Micro-miniaturised pressure sensors such as this have many possible medical applications, particularly if they can be used monitor fluid pressures such as blood. This type of sensor can be integrated with other MEMS sensors and with CMOS electronics to provide amplification of the sensor outputs [4].

The second example of piezoresistive sensing is a 3-axis MEMS accelerometer intended for use as a motion sensor capable of monitoring the activity of a beating heart. The specific application is described in more detail in [5,6] but essentially the device would be stitched to the outer wall of the heart during bypass surgery to monitor for post-operative complications. The requirement was for a device that would be small enough to be removed when it was no longer needed, without the requirement for further invasive surgery, by pulling it out along the track of the wires used to connect it to the outside of the patients body. The complete packaged device is shown in Figure 8.11a and measures 5 mm by 3 mm by 1.5 mm. This includes hermetically sealed, biocompatible packaging which protects the moving parts of the sensor structure.

The MEMS accelerometer used in this case consists of four moving proof masses fabricated in a silicon on insulator (SOI) substrate. An SOI wafer consists of a standard thick silicon substrate, several hundred micrometers thick, topped with a thin buried oxide (BOX) layer and a thin device layer of high quality single crystal silicon. In advanced CMOS electronic processes this Si layer is used to fabricate devices with increased resistance to lock-up and radiation hardness but the BOX layer can also be used as an etch stop for bulk micro-machining. In this accelerometer device the moving masses are etched out from the bulk substrate from the back of the wafer and are suspended from thin cantilevers fabricated in

the device layer of silicon. The movement of each of the proof masses is monitored using strain gauges fabricated as implanted resistors in the silicon cantilevers. There are 2 cantilevers for each mass and two resistors on each cantilever. This means there are 16 piezoresistors in total, arranged as four separate Wheatstone bridges. Figure 8.11b shows a typical accelerometer with triangular proof masses. The white metal tracks are interconnect for the piezoresistors which are not visible in this image.

In plane acceleration of the MEMS sensor will cause differential movement of opposite masses. So, in the case when there is movement in the X direction from right to left, mass 1 will move up while mass 3 moves down (refer to Figure 8.11b). Motion in the opposite direction would obviously cause the opposite effect while acceleration in Y would involve movement of masses 2 and 4. Out of plane accelerations in the Z direction would cause simultaneous motion of all four of the proof masses in the same direction.

Sensing of this motion using implanted resistors in the single crystal silicon device layer is significantly more complicated than in the polysilicon resistors used in the pressure sensor. As was previously mentioned the effect of applied strain on the piezoresistive coefficient is the dominant effect in strain gauges fabricated in semiconducting materials. With careful choice of doping levels and crystal plane orientation, it is possible to fabricate piezoresistive sensors in silicon where the resistance increases when strain is applied in the longitudinal direction but decreases by a similar amount when the same strain is applied in the transverse direction. Figure 8.12 illustrates a possible layout for a simple implanted piezoresistor and indicates the different directions of applied strain. If the piezoresistive coefficients in the longitudinal (π_L) and transverse (π_T) are such that $\pi_L \approx -\pi_T$ then for a given applied tensile strain of a sensor orientated longitudinally will be:

$$R_L = R + \Delta R. \quad (8.16)$$

While the resistance of a sensor orientated transversely to the applied tensile strain is:

$$R_T = R - \Delta R. \quad (8.17)$$

Each of the proof masses in the heart monitor device has two resistors in each cantilever arm and these are connected up into four Wheatstone bridges so that there are two bridges for the pair of X masses (1, 3) and two for the Y masses (2, 4). The piezoresistive sensors are

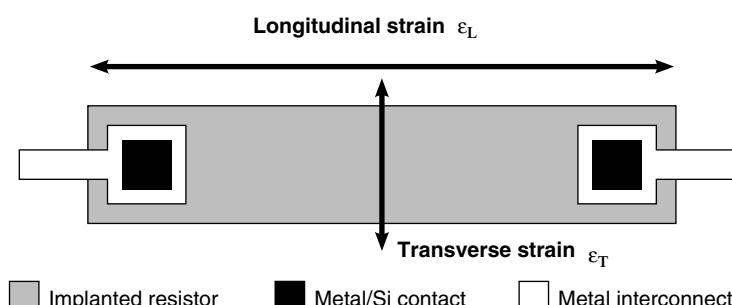


Figure 8.12 Schematic plan view of a diffused or implanted piezoresistor fabricated in single crystal silicon indicating the directions of longitudinal and transverse strain, defined relative to the direction of current flow in the resistor.

arranged so that they are either orientated in the same direction as the strain induced in the cantilever when a mass moves (longitudinal) or at right angles (transverse). Finally, one of the bridge circuits for each pair of masses is designed to be sensitive to in-plane motion and insensitive to out-plane acceleration and vice versa. Taking the example of the sensor circuit for in-plane motion along the X axis, a longitudinal resistor (R_L) will increase in resistance when in tension and decrease in compression while a transverse resistor (R_T) has the opposite characteristic. Figure 8.13 shows the arrangement of longitudinal and transverse resistors in the in-plane Wheatstone bridge. For in-plane acceleration from left to right the cantilever arms of mass 1 will bend upwards, compressing the piezoresistors, while the arms of mass 3 will bend downwards so the sensing resistors will be under tension. The result of this, assuming that the changes in resistance are all equal to ΔR and the nominal value is R the values of the resistors in the bridge will be:

$$R_{L1} = R - \Delta R \quad (8.18)$$

$$R_{T1} = R + \Delta R \quad (8.19)$$

$$R_{L3} = R + \Delta R \quad (8.20)$$

$$R_{T3} = R - \Delta R, \quad (8.21)$$

resulting in an output voltage:

$$V_{out} = V_S \frac{\Delta R}{R}. \quad (8.22)$$

Similarly if the acceleration is out of plane, that is the device is moving into or out of the page looking at Figure 8.11b, both of the cantilever arms will be deflected downwards so that all of the resistors are under tension. Then the resistances in the in-plane sensing Wheatstone

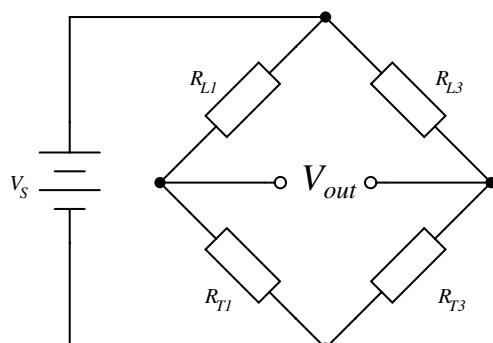


Figure 8.13 Arrangement of strain gauges in the X axis, in-plane Wheatstone bridge circuit from a MEMS accelerometer.

bridge will be:

$$R_{L1} = R + \Delta R \quad (8.23)$$

$$R_{T1} = R - \Delta R \quad (8.24)$$

$$R_{L3} = R + \Delta R \quad (8.25)$$

$$R_{T3} = R - \Delta R. \quad (8.26)$$

These resistance changes will balance out and so the resulting output voltage due to this out-of-plane acceleration will be:

$$V_{out} = 0. \quad (8.27)$$

Therefore this circuit is insensitive to out-of-plane acceleration but sensitive to in-plane accelerations along the X axis. By simply rearranging the connections of the Wheatstone bridge it would be possible to create a similar circuit which was sensitive to out-of-plane accelerations but not to in-plane movement.

8.3.3 Piezoelectric Effect

While piezoresistors are passive sensors requiring external power the piezoelectric effect can be used to make active sensors, and actuators. The phenomenon of piezoelectricity is the ability of some materials to generate an electrical potential when mechanically stressed. This is a two way effect and electrical inputs can cause a piezoelectric material to change shape. One of the most common examples in medicine is ultrasound where piezoelectric transducers both produce the ultrasound and sense the returning echoes.

Piezoelectric materials are typically crystalline, and contain fixed dipole charges which have some form of polarisation, that is the dipole charges are lined up in the same directions. Applying mechanical forces can move these charges around generating the measured potentials. The important point is that useful signals are only produced by changing the applied force, ‘dc’ mechanical signals do not produce any useful output. Piezoelectric materials range from natural substances like quartz crystal and bone, to synthetic compounds like lead zirconate titanate (PZT) and lithium niobate. Aluminium nitride is another interesting piezoelectric material as it should be possible to deposit thin films using microfabrication techniques for CMOS compatible mechanical actuation.

Figure 8.14 illustrates the effect of applying stress to crystalline quartz, which results in a redistribution of charge in the material.

8.3.4 Quartz Crystal Microbalance

Quartz was the first piezoelectric material to be widely used, originally in the form of natural quartz crystals that were carefully cut to give the desired characteristics. The first, and still probably the most widely used piezoelectric sensor, was the quartz crystal microbalance (QCM), which consists of a very thin slice of quartz with thin film electrodes on either side, an example is shown in Figure 8.15. These are driven with an ac electrical signal which sets up a resonant standing wave in the piezoelectric crystal.

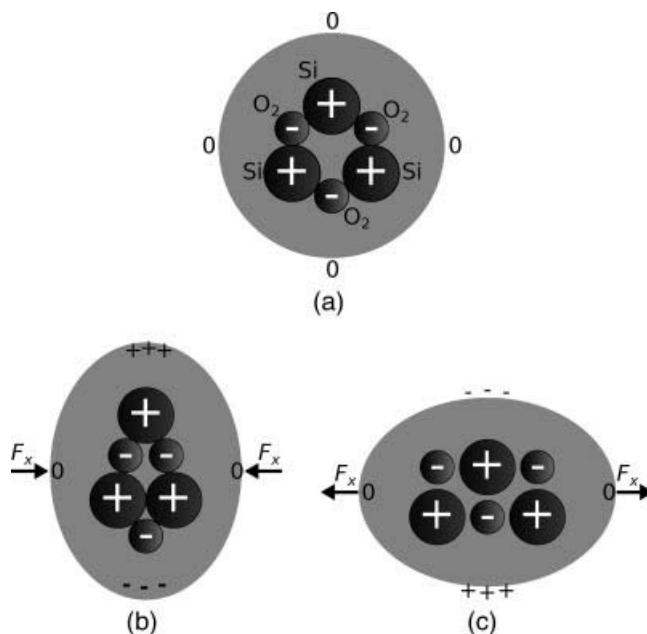


Figure 8.14 (a) Arrangement of positive (Si) and negative (O₂) charges in crystalline quartz. (b) Applying compressive stress redistributes the charge leading to a potential difference across the crystal. (c) Applying tensile stress has the opposite effect.



Figure 8.15 Quartz crystal microbalance, diameter 1.25 mm and thickness around 100 μm . Designed for use as a thickness monitor in a metal deposition system.

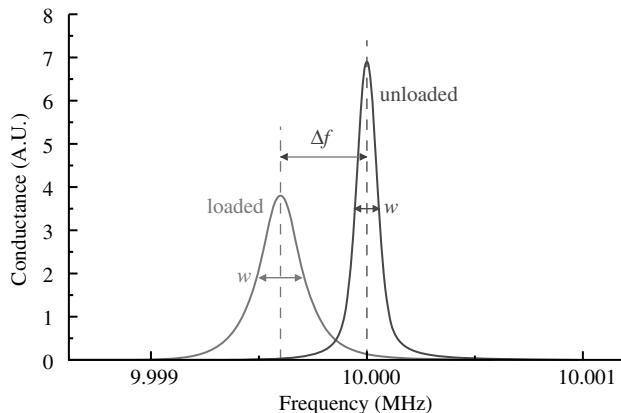


Figure 8.16 Indicative plot showing a typical response of a QCM. Loading causes a change in the resonant frequency f and may also change the bandwidth w (Adapted from http://en.wikipedia.org/wiki/File:QCM_principle.gif).

Figure 8.16 illustrates a common method of characterising resonant devices like the QCM by measuring the conductance or admittance while sweeping the frequency of the applied ac voltage. The conductance will show a characteristic peak at the resonant frequency f with a bandwidth w . The quality factor Q gives a measure of the dissipation of the resonant energy and is defined as:

$$Q = \frac{f}{w}. \quad (8.28)$$

The resonant frequency is sensitive to the mass of the QCM and a common application of this sensor is as a thickness monitor in a thin film deposition process. If the material deposited is very thin or has similar mechanical properties to the QCM then the frequency will change in proportion to the deposited mass. If the material deposited is soft or viscoelastic then the Q factor will change as well as the energy dissipation is increased. Measurement of biological materials in a liquid environment will generally increase the energy dissipation. CQM-D is a patented technique from a company called Q-Sense,¹ which measures both the resonant frequency and the dissipation factor $D = 1/Q$. This makes the QCM sensitive to the mechanical properties of the deposited film as well as just the mass and is targeted at biological sensing. Dissipation sensing is important when investigating chemical attachments in a liquid environment where there is significant damping. The sensitivity of the QCM is such that it can measure molecular binding events, for example in an immunosensing application using antibody/antigen binding or detection of specific DNA attachment.

The equivalent circuit of a QCM at a frequency close to resonance is shown in Figure 8.17. The lower of the two parallel arms of the circuit has a capacitance C_P representing the physical capacitance of the parallel plate structure formed by the QCM, plus any parasitic effects

¹ <http://www.q-sense.com/>.

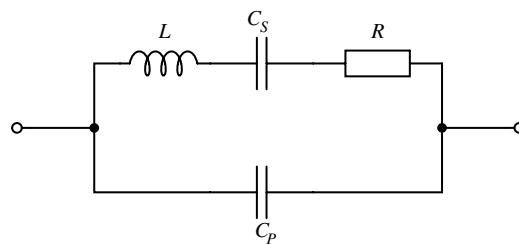


Figure 8.17 Equivalent circuit for a quartz crystal microbalance.

from the connections. The upper arm represents the resonant properties of the piezoelectric crystal. The inductance L and capacitance C_S cancel out at the resonant frequency to leave the resistor R which represents the dissipative energy losses. If the resistance is low that is equivalent to a high quality factor. Loading the QCM by attaching something to the surface can potentially change any and all of the resonant circuit elements.

The initial paper on the QCM-D technique describes the instrumentation used to measure dissipation in a quartz crystal microbalance driven at its natural resonant frequency [7]. Prior to this the dissipation was often measured by monitoring the amplitude of the oscillation signal in the QCM, which is inversely proportional to D . However, this places limitations on how the microbalance is driven, for example there can be no automatic gain control which would tend to correct for dissipation. The solution detailed by Rodahl *et al.* used a setup similar to that shown in Figure 8.18. Here the QCM is driven into resonance and once it is oscillating stably a computer controlled relay disconnects the input signal. Simultaneously,

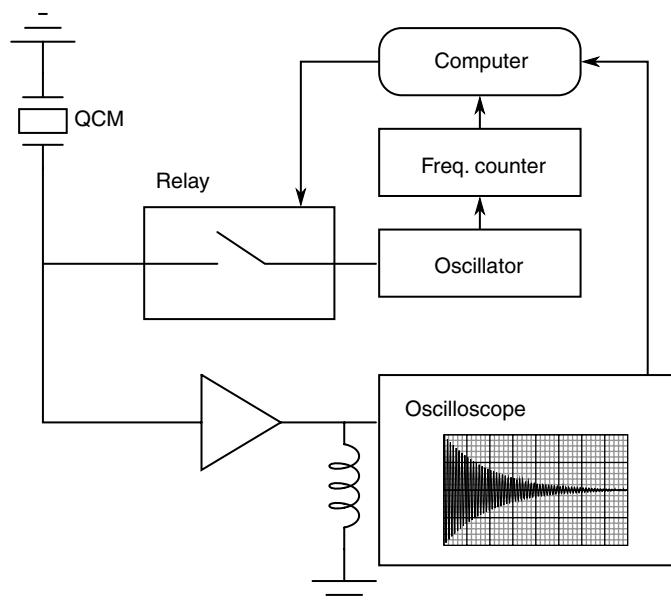


Figure 8.18 Instrumentation for dissipation measurements from a quartz crystal microbalance sensor, based on [7].

an oscilloscope is triggered to capture the output signal from the QCM as the oscillations decay. The rate of decay of the signal provides a measure of the dissipation factor of the microbalance setup. The inductor after the buffer in the sensing part of the circuit will act as a high pass filter, removing any dc offset introduced when the relay disconnects the drive oscillator.

8.3.5 Surface Acoustic Wave Devices

The quartz crystal microbalance is a bulk acoustic wave device meaning that the vibrations occur throughout the material. There is another class of piezoelectric resonant device that makes use of acoustic waves confined to the surface of a piezoelectric material. The Surface Acoustic Wave (SAW) shown in Figure 8.19 is known as a Rayleigh wave. This is the type of vibration that causes damage far from the epicentre of an earthquake but can also work with much smaller, faster vibrations in thin film materials. The movement here is purely at the surface, the vibrations only reach about one wavelength into the material and the amplitude of Rayleigh waves is normal to the surface.

Surface acoustic waves can be excited in a piezoelectric substrate using an interdigitated transducer electrode (IDT) fabricated in a thin metal film on the surface of the substrate. The device shown in Figure 8.20 is known as a SAW delay line where the waves are transmitted along the surface between a transmission IDT connected to an ac signal generator, and a receiver IDT connected to a load. Another type might be a single transducer sending waves towards metal structures on the surface that reflect the SAW energy back to the input. They are often used to create high quality-factor filters and other RF components for use in mobile phones and other devices where the small size and high Q is important. Mobile devices with wireless connectivity rely heavily on SAW components.

The resonant frequency (f) of the SAW device will depend on the acoustic wave velocity (v) of the piezoelectric material and the pitch (d) of the IDT with the relationship in Equation (8.29), and typical SAW characteristic frequencies are in the range of 30–500 MHz.)

$$f = \frac{v}{d}. \quad (8.29)$$

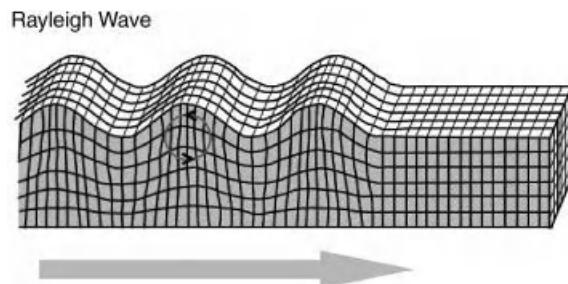


Figure 8.19 Illustration of the propagation of a Rayleigh surface acoustic wave (Reproduced from <http://earthquake.usgs.gov/learn/glossary/>).

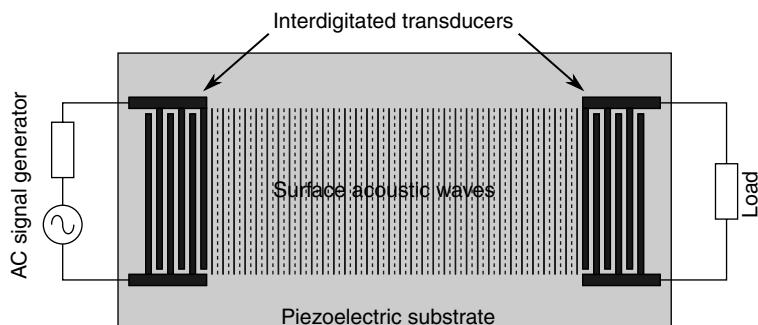


Figure 8.20 Schematic plan view of a SAW delay line.

The SAW delay line can be turned into a sensor, and its acoustic properties, sensitive to parameter being sensed. The primary interaction mechanisms are those that affect the frequency or the delay between the transducers by changing the wave velocity, the IDT distance, or both. Temperature, strain, pressure, force, and the properties of a sensing layer added to the device are examples of quantities that can be measured. In particular, the accumulated surface mass produces a decrease in frequency so a SAW device can be used as a deposition monitor in a similar way to the QCM.

Biosensors can require a different type of SAW as Rayleigh waves will be dissipated through viscoelastic coupling with a liquid medium. Shear-Horizontal SAW waves require the crystal planes of the piezoelectric material to be orientated such that the induced waves are parallel rather than normal to the surface. Another alternative is the 'Love wave' mode of SAW, where a top coating of a material with low acoustic wave velocity is applied to the surface. This acts as a wave guide and prevents attenuation into the substrate or the surrounding liquid.

Most SAW biosensors operate by detecting mass loading as molecules attach to the prepared surface. By running two transducers in parallel, where only one has the chemically/biologically selective surface, you can compensate for cross-correlations with temperature or strain. As the devices have a characteristic resonant frequency, they can be driven into oscillation by an amplifier connected between input and output. The sensor output is the difference in frequency between the reference and sensor device. The reference device may have an additional blocking layer on top to prevent non-specific attachment. This experimental system is illustrated in Figure 8.21.

SAW devices are also widely used in passive radio frequency identification (RFID) tags. With an antenna attached to the IDT they can pick up pulses of electromagnetic radiation from an RF transmitter. This excites a pulse of SAW energy that is reflected back by reflecting elements spaced like a barcode. The antenna retransmits the reflected pulses and by measuring the timing the 'barcode ID' can be calculated. A wireless SAW sensing device could have a single reflector and the delay of the pulse would carry information about the state of the sensor. A differential measurement could be made possible by using a double ended SAW sensor with the IDT in the centre. Only one SAW transmission area would

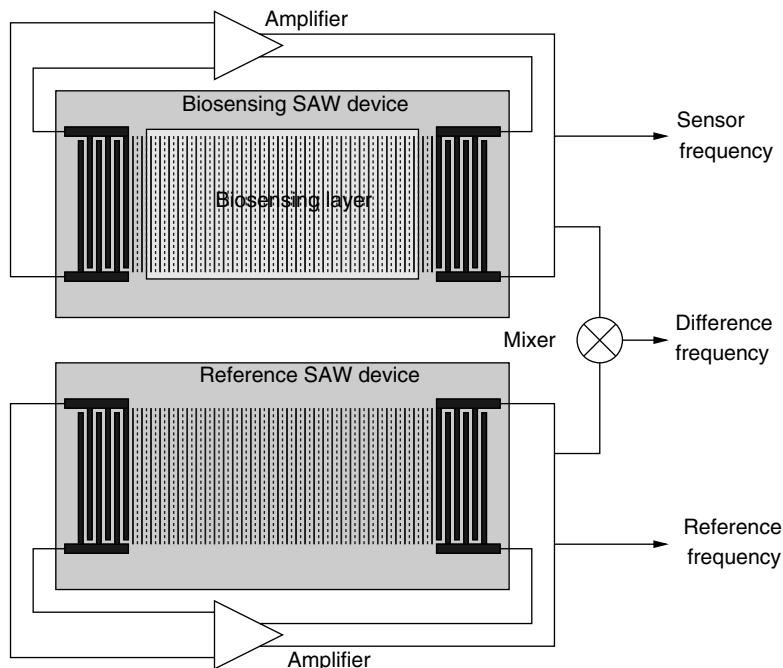


Figure 8.21 Possible arrangement for a SAW biosensing system which uses a pair of SAW delay lines, where one has a specific biochemical receptor layer while the other device acts as a reference to reduce effects of confounding factors.

be sensitive to the measured parameter so there would be also be a reference pulse to make the timing changes easier to interpret.

8.3.6 Capacitive Sensors

An alternative to piezoelectric or strain gauge transduction in mechanical sensors is the use of capacitance. This often involves something like a parallel plate capacitor where the movement is sensed as a change in the effective thickness of the gap between the plates. They have advantages of high sensitivity, if the gap is on the microscale, and low power operation compared to resistive sensing. The example shown in Figure 8.22 is of a pressure sensor that measures the difference in pressure between a reference chamber and a measurement chamber. Any difference will change the position of a moving membrane that forms one plate of a pair of complementary capacitors. This means that as one capacitance increases, as the membrane moves closer, the other will reduce. The capacitors are arranged in a bridge driven by an ac signal and the output voltage will depend on the ratio of the sensor capacitors.

Capacitive sensing is the basis of some of the most successful MEMS devices, such as the surface micromachined accelerometers originally developed by Analogue Devices. These typically consist of a moving mass connected to the substrate through flexible springs, as shown on the left hand side of Figure 8.23. Along the edges of this proof mass are electrode

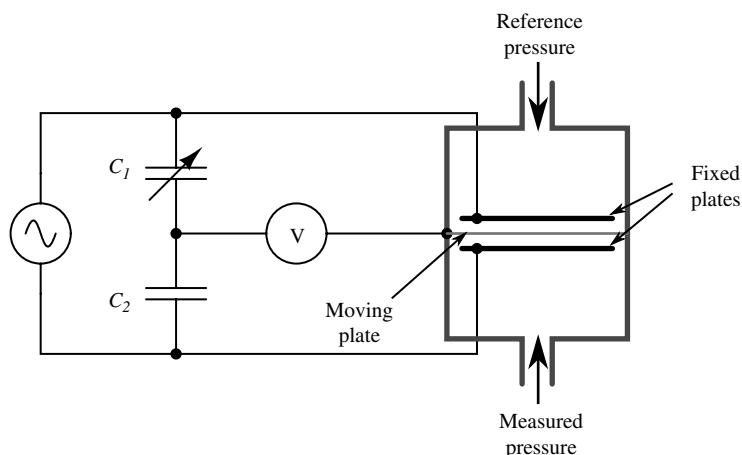


Figure 8.22 Schematic representation of a pressure sensor using a capacitive bridge circuit for transduction.

fingers between pairs of fixed plates which form capacitors (C_{S1} , C_{S2}) to be used in sensing the displacement of the mass. When an acceleration is applied, as in the right-hand side of Figure 8.23, the mass will move in the opposite direction. This differentially changes the two capacitances and the movement can therefore be sensed by measuring the difference.

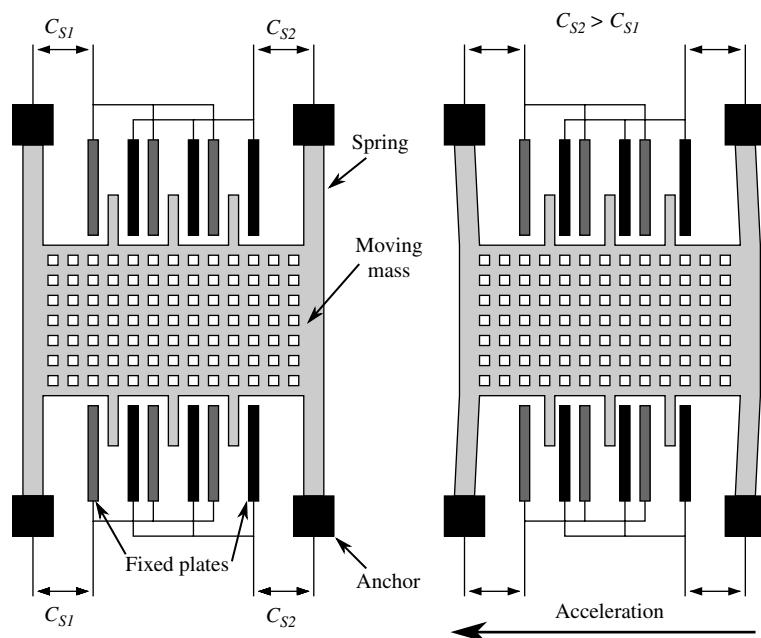


Figure 8.23 Operation of a surface micromachined MEMS accelerometer with capacitive sensing.

8.3.7 Capacitance Measurement

The values of the complementary capacitors (C_{S1} , C_{S2}) in a differential capacitive sensor can be represented in terms of the area A , nominal separation d , and the change in the position of the central plate x . We can also define the nominal capacitance C_0 , when there is no movement.

$$C_{S1} = \varepsilon \frac{A}{d+x} = \frac{C_0}{1 + \frac{x}{d}} \quad (8.30)$$

$$C_{S2} = \varepsilon \frac{A}{d-x} = \frac{C_0}{1 - \frac{x}{d}} \quad (8.31)$$

$$C_0 = \varepsilon \frac{A}{d}. \quad (8.32)$$

These equations will only apply when the displacement x is relatively small compared to d , as large deflections cause non-linearities. The increase in capacitance will lead to an increase in the electrostatic attraction between the plates, which is balanced by the mechanical spring force that tries to centre the moving mass. The mechanical design issues are beyond the scope of this discussion but a detailed introduction can be found in [8].

The small signal current in a capacitor can be represented by the following equation:

$$i = \frac{\delta(CV)}{\delta t} = C \frac{\delta V}{\delta t} + V \frac{\delta C}{\delta t}. \quad (8.33)$$

This equation has two parts to it, the first representing the familiar method of measuring a relatively constant capacitance with an ac voltage, where the current depends on the derivative of voltage with time. The second is possibly less familiar and suggests that you can measure a changing capacitance by applying a dc voltage. In terms of capacitance based sensors, the first term represents the measurement of a relatively slow variation of capacitance using a high-frequency ac voltage. This is the case in an accelerometer or pressure sensor where $f_c \ll f_v$. The second case would be a capacitance that changes very quickly, for example in a MEMS microphone with capacitive feedback, where a dc voltage can be applied and will produce a current signal dependent on the rate of change of capacitance. This discussion will focus on the measurement of slowly varying capacitances through the application of ac drive voltages.

The displacement in a differential capacitive sensor can be sensed by measuring the current output of the circuit shown in Figure 8.24. An ac voltage with a frequency $f(\omega = 2\pi f)$ and amplitude v to the top of the pair of capacitors and the inverse of the voltage signal ($-v$) to the bottom. The central point ‘A’ is effectively grounded through the low impedance ammeter. The current through the upper capacitor C_{S1} into the measured node will be:

$$i_{C_{S1}} = j\omega C_{S1} v, \quad (8.34)$$

while the current through C_{S2} into the measured node is:

$$i_{C_{S2}} = -j\omega C_{S2} v. \quad (8.35)$$

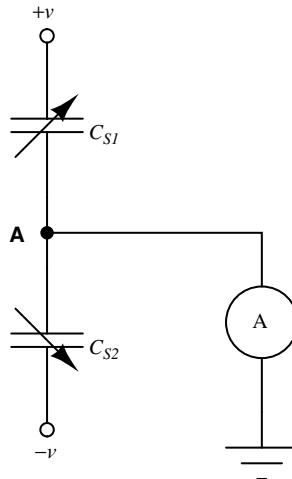


Figure 8.24 Measurement circuit for differential capacitive sensors.

If Equations (8.34) and (8.35) are added together and the expressions for the capacitances in terms of the displacement (8.30) and (8.31) are substituted in then the result is:

$$i = j2\omega VC_0 d \frac{x}{d^2 - x^2}. \quad (8.36)$$

This is a function of the displacement but there are second-order terms as well that can be an issue if it is not the case that $x \ll d$. If this is the case then the term for x^2 can be ignored and the equation simplifies to:

$$i = j2\omega VC_0 \frac{x}{d}. \quad (8.37)$$

This current could be measured with the current follower or transimpedance amplifier examined in Chapter 7, Section 7.4.6 but there are other methods of transduction for capacitive sensors that may be more appropriate depending on the application.

8.3.8 Capacitive Bridge

An alternative to current measurement is to use a capacitive bridge circuit (see Figure 8.25a) as in the sensor in Figure 8.22. In a half bridge arrangement two of the capacitances will vary with the displacement of the sensor in a push-pull arrangement:

$$C_1 = \frac{C_0}{1 - \frac{x}{d}} \quad (8.38)$$

$$C_2 = \frac{C_0}{1 + \frac{x}{d}}. \quad (8.39)$$

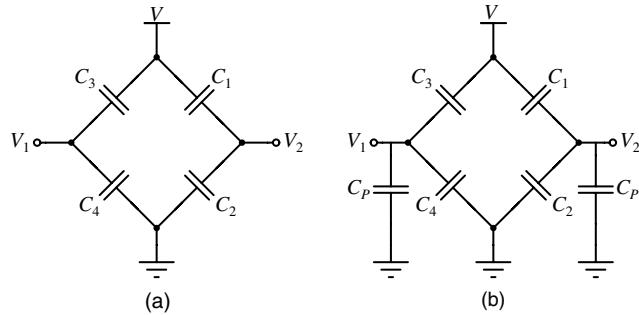


Figure 8.25 (a) A bridge circuit for capacitive sensors without parasitic capacitances. (b) Similar circuit with parasitics on output terminals.

While C_3 and C_4 will both be equal to the nominal capacitance C_0 . Then the output voltages (V_1, V_2) will be:

$$V_1 = V \frac{C_3}{C_3 + C_4} = \frac{V}{2} \quad (8.40)$$

$$V_2 = V \frac{C_1}{C_1 + C_2} = V \frac{\frac{C_0}{1 - \frac{x}{d}}}{\frac{C_0}{1 - \frac{x}{d}} + \frac{C_0}{1 + \frac{x}{d}}} \quad (8.41)$$

The overall output voltage of the bridge is then:

$$V_{out} = V_2 - V_1 = V \frac{x}{2d}. \quad (8.42)$$

As with the resistive bridge circuits the use of a full bridge arrangement, where all four of the capacitances vary with displacement, will double the sensitivity of the system if the two sides of the bridge are working in opposition.

Parasitic capacitances can be a problem with this type of measurement as these can be significantly larger than the actual capacitances being sensed. A parasitic (C_p) to ground on both of the bridge outputs (see Figure 8.25b) will have the effect of lowering the effective sensitivity of the capacitive sensor. Using the same half bridge arrangement as before but defining the differential sensing capacitors as:

$$C_1 = C_0 + \Delta C \quad (8.43)$$

$$C_2 = C_0 - \Delta C. \quad (8.44)$$

The output voltage without the parasitic capacitances (Figure 8.25a) will be:

$$V_{out} = V_2 - V_1 = V \frac{\Delta C}{2C_0}. \quad (8.45)$$

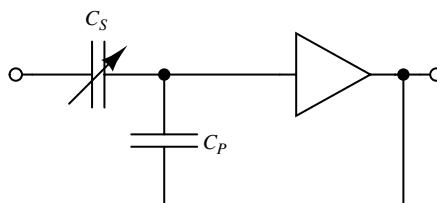


Figure 8.26 Bootstrap method for eliminating the effect of parasitic capacitance C_P on the measurement of sensor capacitance C_S .

While the output voltage with the parasitics will be:

$$V_{out} = V_2 - V_1 = V \frac{\Delta C}{2C_0 + C_P}. \quad (8.46)$$

There are a number of different possible sources of parasitic impedances, including the electrical connections made to the sensor, capacitances in the active components of the instrumentation and parasitic capacitances in the actual MEMS structure, for example between the suspended moving elements and a conductive substrate. If the conductive plane which has parasitic capacitances to a measured capacitor is not grounded then effectively you have the series combination of the parasitics in parallel with the sensing capacitor. Changes in the parasitics will directly affect the measurement which is obviously undesirable.

Bootstrapping is a common method of dealing with parasitics which attempts to eliminate any current flow in them. With a parasitic capacitance on a signal line this can be achieved by driving the other side of the capacitor with the same signal through a buffer amplifier (Figure 8.26). This technique is applied in triaxial cables, commonly used for low level signals, where the inner conducting core is surrounded by a guard conductor that is driven with the same signal as the core. This is subsequently surrounded by a grounded shield conductor as in a coaxial cable.

The current measurement method mentioned previously is another useful way to reduce the effects of parasitic capacitance between a signal line and ground. A low impedance current meter or a virtual ground, like the input of a operational amplifier configured as a current follower, will have the same effect as bootstrapping by removing the voltage drop across the parasitic.

There are a number of non-ideal effects in the fabrication or the subsequent operation of capacitive MEMS sensors, and some of these may be reduced or compensated for by using force feedback control. Here, the output from the sensor interface is fed back, through a large resistance, to the central output terminal of the sensor. This negative feedback applies a balancing electrostatic force which serves to keep the moving mass centred and the displacement as small as possible. The output should then be a linear function dependent on the electrostatic force applied to stop the mass from moving under acceleration.

8.3.9 Switched Capacitor Circuits

There are a number of different reasons to favour capacitors over resistors for biasing in amplifier design, particularly in integrated circuits. Where high-value resistors are

required, for example in current to voltage converters for low current measurement, the integrated components may be prohibitively large and there are other problems such as thermal noise, component matching and temperature dependence. Capacitor ratios may be set relatively easily in integrated circuit processes, though the absolute values will vary, and they are less prone to temperature dependence or thermal noise. Capacitors generally have a smaller footprint area than equivalent resistors and do not represent a dc load, which may be important for CMOS based op-amps where the output impedance can be significant.

Simply replacing resistors in an op-amp circuit with an equivalent ratio of capacitors could work for some signals but they will be open circuit for low frequency voltages meaning there is no dc feedback to the inverting input. This is not compatible with stable op-amp circuit design and so the alternative is to use a switched capacitor amplifier such as that shown in Figure 8.27. This is a simplified circuit that does not take into account parasitics or other non-ideal behaviours but it is useful for illustration of the approach. There are three switches, which are driven by two non-overlapping and complementary clock signals as shown in Figure 8.28. This is a non-inverting amplifier, although it resembles the layout of the inverting amplifier, and has a gain set by the capacitor ratio C_1/C_2 . The best way to understand it is to look at the circuit during each of the clock cycles. In cycle Φ_1 , where s_1 and s_3 are closed and s_2 is open, capacitor C_1 is connected to V_{in} at one end and the inverting op-amp input, that is a virtual earth, at the other end. If the previous stage can supply sufficient current then by the end of cycle one the capacitor will be fully charged to V_{in} and the charge will be: $Q_{C_1} = V_{in}C_1$. The other capacitor (C_2) is discharged and the amplifier output $V_{out} = 0$ V.

At the start of cycle Φ_2 , s_1 and s_3 will open while s_2 closes. Now the op-amp would like both ends of C_1 to be at the same potential (ground) and so it will be discharged. The current required to do this will flow through C_2 and so the charge is transferred between the two capacitors such that: $Q_{C_2} = Q_{C_1}$. At the end of this charge transfer, which should be complete by the end of cycle Φ_2 , the voltage on the output will be:

$$V_{out} = \frac{Q_{C_2}}{C_2} = \frac{Q_{C_1}}{C_2} = V_{in} \frac{C_1}{C_2}. \quad (8.47)$$

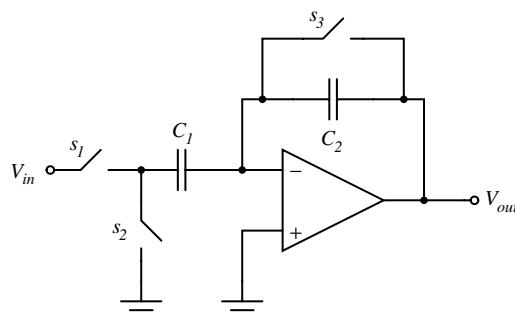


Figure 8.27 Switched capacitor amplifier with non-inverting function.

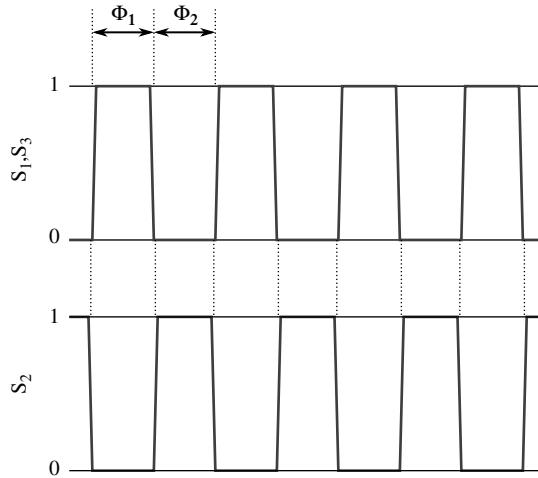


Figure 8.28 Non-overlapping clock signals to control a switched capacitor amplifier.

Therefore, this is a non-inverting amplifier with the gain set by the capacitor ratio. Another way to think about this circuit is as an amplifier with sample and hold functionality where the input signal is sampled at the time point where we change from cycle Φ_1 to cycle Φ_2 .

The circuit in Figure 8.29, is similar to the previous switched capacitor amplifier but is more proof against parasitic capacitances. It can be used as a capacitance to voltage converter to measure the unknown capacitance C_x . As before there are two, non-overlapping, switching cycles Φ_1 and Φ_2 . The parasitic capacitors C_{p1} , C_{p2} and the parasitic leakage conductance G_p are not dealt with in detail here and a more complete description can be found in [9]. In short, the parasitic capacitances are shorted out by the parallel switches in order to reduce their effects on the charge transfer. In any case, when the Φ_1 switches are closed C_x is charged to V_{ref} , while the feedback capacitor C_f is discharged. Then in cycle Φ_2 , C_x is discharged and the accumulated charge is transferred onto C_f . Again, the output voltage at the end of this cycle will depend on the ratio of the two capacitors:

$$V_{out} = V_{ref} \frac{C_x}{C_f}. \quad (8.48)$$

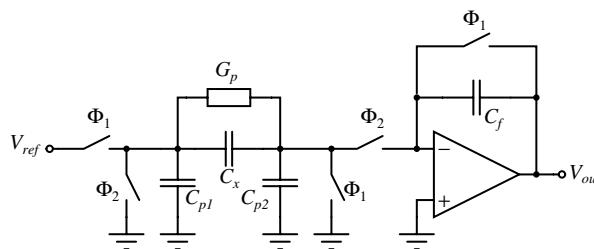


Figure 8.29 A switched capacitor circuit capacitor circuit with an output that depends on the value of the unknown capacitance C_x .

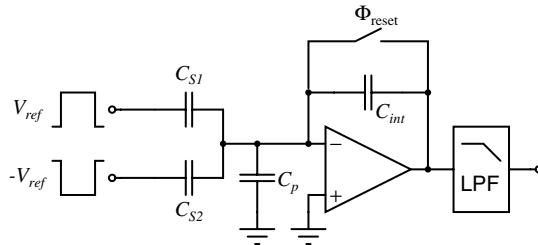


Figure 8.30 Switched capacitor integrating amplifier which could be used to measure a differential capacitive sensor.

The previous circuit allowed the measurement of a single unknown capacitance but Figure 8.30 shows an alternative scheme for the measurement of a differential capacitive sensor using a switched capacitor integrating amplifier. The operation of this is as follows. The sensing capacitors are driven with square waves that are not only 180° out of phase but also dc shifted so that V_{ref} goes from, for example, 0 to 5 V and $-V_{ref}$ goes from 0 to -5 V. Therefore the capacitors will be charged in different directions and the overall charging current will be dependent on the difference between them. The excess charging current, if $C_{S1} \neq C_{S2}$, has to flow through C_{int} , assuming no current flow into the op-amp inputs. Therefore, C_{int} will integrate this charging current in each cycle and the output voltage will depend on the difference $\Delta C = C_{S1} - C_{S2}$ and the value of C_{int} :

$$V_{out} = V_{ref} \frac{\Delta C}{C_{int}}. \quad (8.49)$$

The reset switch will discharge C_{int} once each cycle so that the starting value of V_{out} is always zero. The reset switch is driven with a signal that has half the frequency of the inputs so that it captures one whole cycle of charge from each capacitor. The output is low pass filtered to remove the modulation introduced by the V_{ref} and $-V_{ref}$ signals. Dealing with all of the parasitics in a capacitive sensor would require a much more complicated switched capacitor amplifier than can be sensibly covered by the current work. An example of such a circuit can be found in [10].

8.4 Optical Biosensor Technology

8.4.1 Fluorescence

This section will examine a number of different sensor technologies but generally focusses on optical sensors using fluorescence as this is by far the most widely used technique in biosensing. Figure 8.31 shows a Jablonski energy diagram illustrating the phenomenon of fluorescence. Briefly, a fluorescent material absorbs a photon which excites an electron from a ground state to a higher energy level. When it decays back down to the ground state, after some delay, it can re-emit a photon at a lower energy. Figure 8.31 is a relatively simplified version of the Jablonski diagram, which does not show all the different ways in which an excited electron can lose energy, many of which do not lead to emission of a photon by fluorescence. The main point to take away is that the emitted photon will have a lower energy

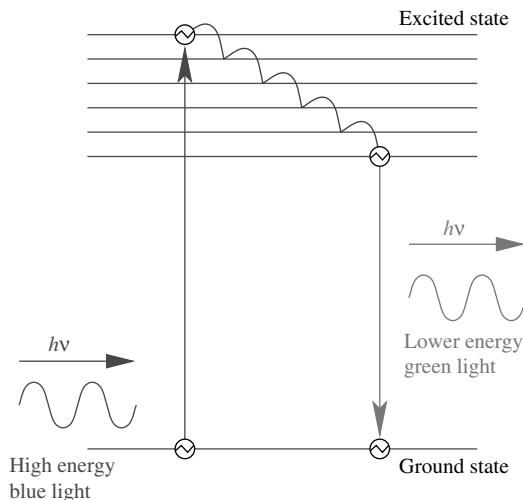


Figure 8.31 Energy band diagram illustrating the phenomenon of fluorescence.

than the absorbed photon. The energy of a photon is expressed by this formula:

$$E = h\nu, \quad (8.50)$$

where h is Planck's constant and ν is the frequency of the light. If the emitted photon has a lower energy then it will necessarily have a different wavelength λ as:

$$\nu = \frac{c}{\lambda}, \quad (8.51)$$

where c is the speed of light. This change in the wavelength of the emitted fluorescent photons compared to the absorbed excitation light is referred to as the Stokes shift after George Stokes who first described fluorescence. In addition to the change in wavelength there will also be a time delay between the excitation and emission. The average time between the absorption of a photon and the subsequent emission is the fluorescence lifetime. This can be quite different for dyes with very similar emission wavelengths and can be used as the basis of biosensing techniques. The final value that characterises a fluorescent material is the quantum yield or efficiency, which is the ratio of photons emitted versus the number of photons absorbed, and is always less than 1.

For many years fluorescence has been the most widely used tool for qualitative measurements in biology and biochemistry. There are two principal methods in which fluorescence is used in cell biology and biosensing. Firstly there are extrinsic fluorescent labels or fluorophores which have been deliberately introduced to the biological system. These can be attached to whatever is being investigated, for example an antibody, and once it attaches to a cell or antigen then the fluorescent emission can be used to detect this specific binding event. The alternative to this is intrinsic fluorescence, where certain cells or organisms either naturally produce fluorescent proteins like the now common Green Fluorescent Protein (GFP) or have been genetically modified to express it.

Fluorescence is not only used at a cell level, one of the most common and important uses of fluorescence in biosensing is in microarrays. Glass slides or chips may be prepared with arrays of microscopic spots of biosensing probe molecules, for example these could be specific lengths of single stranded DNA or antibodies. The target molecules then have to be labelled with specific fluorophores and will attach to their complementary probes on the surface. By measuring intensity of the fluorescence from each spot the presence of the target molecules in the sample under test may be determined. However, the fact that the targets need to be labelled is a limitation on the process and the reason why research into 'label-free' biosensing techniques is so important for the future.

One of the most common fluorescence based analytical techniques in cell biology is Fluorescence Activated Cell Sorting (FACS). This is a type of flow cytometry; a high throughput measurement of cells which involves flowing them past a sensor then sorting them based on the result, as illustrated in Figure 8.32. The starting point for FACS is a heterogeneous population of cells that are labelled with one or more fluorophores based on the expression of some marker on the cell surface or within the cell. Next the cells are confined into a narrow sheath flow channel so that they are separated into a stream of single cells. At this stage they are illuminated with a laser and the resulting fluorescence is measured, often along with an

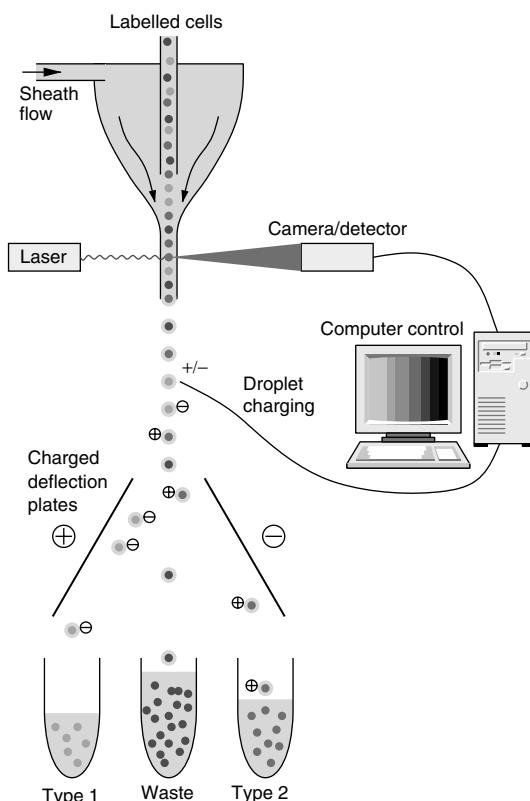


Figure 8.32 Schematic diagram illustrating the function of a fluorescence activated cell sorting system.

estimation of the size of the cell. Next the suspended cells are dispensed from narrow nozzles designed to produce droplets with no more than one cell in each. Depending on the results of the fluorescence measurements these droplets have an electric charge applied to them which means that they can subsequently be deflected by charged plates into a number of different reservoirs. FACS is a very fast and accurate way to gain statistical information about a cell population but it does require that the cells are labelled, and the high flow rates used are likely to damage the cells through shear stress. This means that FACS is not suitable for some cell types such as stem cells to be used in clinical applications. In many cases it can be considered to be a destructive technique and there is considerable interest in other characterisation or sorting methods for cells which are label free and do not cause damage.

There are a number of different techniques for fluorescence microscopy but the following are some of the most commonly used methods:

- **Spectroscopy:** A basic method where the intensity of light emission is measured over a range of different wavelengths to get information about the different fluorescent materials in the sample.
- **FLIM (fluorescence lifetime imaging microscopy):** This technique looks at the lifetime of the fluorescence rather than the intensity and has some advantages which will be covered later in the chapter
- **TIRF (total internal reflection fluorescence):** This makes use of evanescent field effects for very sensitive measurements. TIRF effects are important in fibre optic based sensors covered later in the chapter.
- **FRET (Förster resonance energy transfer):** This exploits pairs of fluorescent dyes and can be used for sensitive molecular measurements. A case study later in the chapter will explore this more fully.

There are a number of ways in which the fluorescence of a fluorophore can be affected by the environment or by the measurement itself. Photobleaching is a consequence of exposing the dye to the high energy excitation radiation which can gradually alter the chemical structure of the dye leading to a reduction of fluorescence. This can be exploited for certain measurements or reduced by making sure that excitation is kept to short pulses or with reduced power. Quantum dots are a special case as their fluorescence is a consequence of their physical structure and they are not generally subject to bleaching. Quenching is a reduction of the fluorescence caused by environmental parameters, usually chemical in nature. This can therefore form the basis of a chemical sensor.

8.4.2 Optical Fibre Sensors

Many of the techniques for optical analysis or measurement of biological and chemical materials, functions and reactions are suitable for adaption for use in sensors based on optical fibre technology (or other optical waveguides). In addition there are a variety of physical sensors which can also make use of optical fibres. Optical fibres have low losses and so they can be used to locate a remote sensor a significant distance from the optical detection and light source. When compared to a wireless sensor the fibre sensor still needs a physical link but there are also advantages in certain environments due to the fact that there is no need for an electrical connection for power or signal transfer. For example, this can be a significant

advantage for sensors in liquids where corrosion is a problem or where sparks could lead to risks of explosions. In addition, optical fibres have little or no problem with interference and have a high sensitivity and bandwidth.

Optode (occasionally referred to as optrode) sensors are physical or, more commonly, chemical transducers which are interrogated optically. These may or may not use optical fibres as part of the sensor system but many do, and there are two main forms that this can take. In an extrinsic sensor the fibre simply guides the light to and from the sensing element but in an intrinsic sensor the fibre itself becomes part of the transduction. This could happen through a mechanical interaction such as the bending of the fibre altering the light path or through some other type of interaction, possibly chemical in nature, which changes the propagation of light in the fibre.

In an extrinsic optical sensor, where the fibre acts as a guide for the light to and from the actual sensor, there are many different possible sensing modalities. The light could pass through the sensing element where the absorbance changes with the parameter being sensed or it could reflect off a surface. Membrane pressure sensors with the bandwidth to measure pressure waves from explosions are an example of a physical sensor using reflection, as is a gas sensor with a membrane that expands and bends as the analyte is absorbed into it. Some materials may change in reflectivity due to a chemical interaction with the environment or produce fluorescence.

One example of an optode using fluorescence is that of an oxygen sensor which exploits the quenching effect of oxygen on ruthenium based fluorophores. Sensing layers can be immobilised on the tip of an optical fibre in a structural matrix which is then coated with an oxygen permeable membrane such as Teflon. A good example of this is a fibre sensor which uses an oxygen sensitive coating on the tapered tip of a multimode fibre which can carry both the excitation and emission wavelengths. The tapered tip increases the interaction of the light with the fluorophore and this aids the sensitivity of the sensor [11].

The operating principle of an optical fibre makes use of the phenomenon of total internal reflection at the boundary between the core of the fibre and the surrounding cladding. Assuming the light hits the interface between the core and cladding at an angle shallower than the critical angle for the materials, then there will be total reflection and efficient transmission of the light down the fibre. However, there is also a so-called evanescent wave that extends beyond the boundary but decays exponentially. This has been described previously in Chapter 4, Section 4.3.5. The effective extent of the near field evanescent wave is typically less than the wavelength of the light but this is enough for it to interact with the material close to the interface and, for example, excite fluorescence in a sensing layer as in TIRF. This sensing layer could be biological in nature, such as an antibody probe which fluoresces when labelled targets are attached. The sensing region in an intrinsic fibre sensor can be distributed over the length of the fibre or confined to specific points. The penetration depth of the evanescent wave is very low so if the fibre becomes coated with a fouling layer, which will depend on the environment, it can stop it working properly. On the other hand, sensing using this evanescent illumination is considered to be very sensitive to changes in the very thin layer into which it penetrates.

8.4.3 Optical Detectors

There are many different methods by which the light from an optical biosensor can be transduced into a meaningful electronic signal. One of the simplest photonic transducers is the

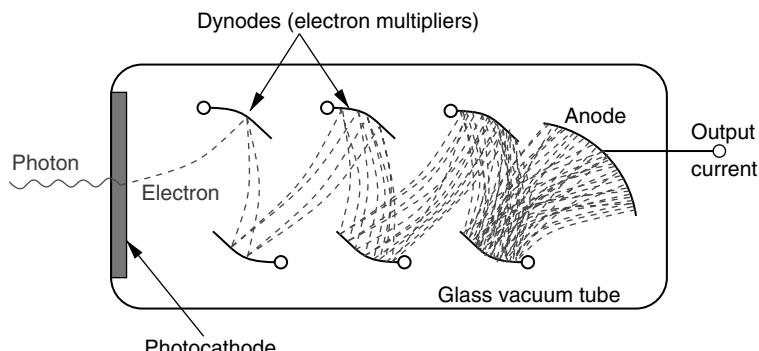


Figure 8.33 A schematic representation of a photomultiplier tube which can act as a sensitive optical detector.

photoresistor or light-dependent resistor. This is a conductometric sensor based on a semi-conducting material. Exposing this to light leads to the generation of mobile charge carriers in the semiconductor which decreases the resistivity of the material. Photoresistors are relatively crude but can provide useful information about the overall intensity of light, for example as an exposure metre for a camera. The wavelength of light to which the photoresistor is most sensitive may be controlled by doping the semiconductor, which will change the energy required to raise an electron to the conduction band.

The photomultiplier tube (PMT) illustrated in Figure 8.33, is an extremely sensitive single photon detector and still the best choice for many low level measurements, regardless of the fact that it relies on pre-solid-state, vacuum tube technology. The PMT consists of a sealed glass tube under high vacuum with a number of delicate metal electrodes inside. At one end is a photocathode that uses the photoelectric effect to produce an electron when struck by a photon. This is accelerated towards a dynode electrode held at a more positive voltage. This acceleration gives the electron additional energy so that when it hits the dynode it can excite more than one electron to be emitted. These are then accelerated towards a sequence of additional dynodes, each of which is held at a higher positive voltage than the last. Finally, at the end of the tube is an anode which receives a pulse of current due to the multiplication of the electrons which happens at each of the dynodes. This amplification effect means that the PMT is extremely sensitive as each photon can potentially produce an output current made up of a large number of electrons. Therefore it has a quantum efficiency, meaning the number of electrons out per photon input, which is greater than unity. A common use for the PMT in biomedical imaging is in a positron emission tomography (PET) system. Positron annihilation produces gamma radiation which can be converted to visible light with a scintillator. The PET scanner uses a ring of PMT detectors with scintillators in front of the photocathode.

There are a number of solid state devices that can be used as optical detectors and have advantages of small size and high speed of operation. The photodiode illustrated in Figure 8.34 is a p-i-n structure though p-n diodes can also be used as optical sensors. Photodiodes are operated in reverse bias, meaning that there is only a small leakage current and a wide depletion region. When they are exposed to light, photons of sufficient energy can be absorbed to generate electron hole pairs. If this happens in the depletion region the carriers are swept apart by the electric field before they can recombine and a photo-generated current

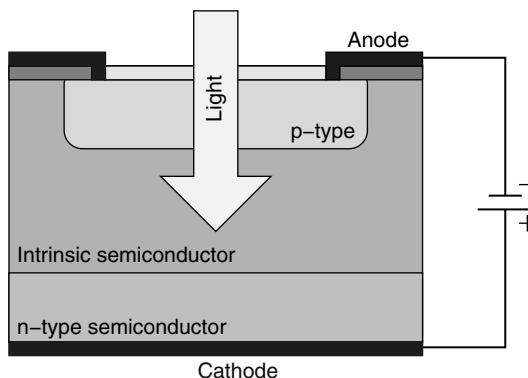


Figure 8.34 Cross sectional schematic illustration of a p-i-n photodiode structure.

will flow. The advantage of a p-i-n structure over a p-n diode is the wider effective depletion region meaning a higher quantum efficiency and a lower capacitance which can increase the bandwidth.

In a standard photodiode the quantum efficiency is always less than unity because each photon can only excite at most one electron-hole pair. However, it is possible to design photodiode structures which can provide amplification of the carriers generated by photon absorption. The single photon avalanche photodiode (SPAD) is designed to be biased well into breakdown, where normally a high current would be expected. In a SPAD though this does not occur until a carrier pair is generated in the depletion region at which point a spike of current will occur through an avalanche breakdown process which is similar to the electron multiplication in a PMT. SPADs are biased with a resistance in series, as shown in Figure 8.35, which can either be an actual resistor or an appropriately biased transistor. When the SPAD goes into breakdown the current flows through this resistance which reduces the voltage across the device and quickly brings it out of the avalanche breakdown region of operation. This is known as passive quenching and a SPAD biased in this way operates in a ‘Geiger-mode’ where each photon will generate a spike of a few picoseconds duration on the

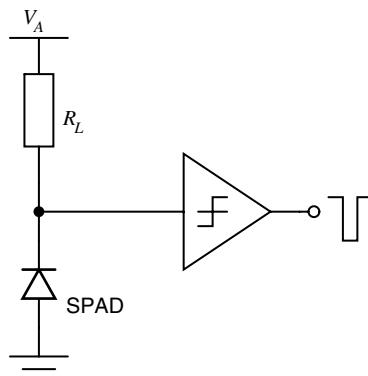


Figure 8.35 Simplified passive quenching circuit for a single photon avalanche diode (SPAD) operated in a Geiger mode.

output. These spikes can be counted to provide a number related to the intensity of the light which has a very high dynamic range. This range will be between the ‘dark-count’, a noise floor dependent on the spontaneous generation of carriers in the device, and some maximum level where photons are arriving too close together to be able to detect each one. The fact that the device can accurately measure the time of arrival of each photon means that SPADs can potentially be used for three-dimensional imaging through ‘time of flight’ measurement [12] or as part of a fluorescence lifetime imaging microsystem [13].

The charge coupled device (CCD) was originally developed in the 1960s as a memory device and consists of an array of gate electrodes separated from a semiconducting substrate by a thin dielectric. Packets of charge can be stored in doped channels under the electrodes and moved between adjacent pixels in a row by changing the voltage on the gates. Although it was conceived as solid state memory it was soon found that charge could be generated through the photoelectric effect when the array is exposed to light. In a modern digital camera a two-dimensional CCD array is exposed through a lens, and a colour filter, to give a pattern of charge relating to the light exposure. This is then read out by moving the charge packets along a row and measuring the intensity of each pixel at the end. The CMOS camera is a more modern alternative to the CCD camera which is common in certain applications, such as mobile devices, due to the smaller size. High end camera sensors still tend to be CCD based as the integrated circuits on a CMOS camera usually mean that the proportion of each pixel taken up by the light sensitive area is lower. A CMOS camera chip is typically made up of an array of photodiodes integrated with electronics for row and column addressing. SPAD sensors may also be integrated with CMOS electronics to provide readout, addressing and even time resolved photon counting in each individual pixel.

8.4.4 Case Study: Label Free DNA Detection with an Optical Biosensor

This section will discuss a biosensor system which uses a combination of an artificial DNA structure that operates a switch, fluorescence based readout and a control mechanism using electrochemistry and microfabricated electrodes. Therefore, it combines many of the different techniques covered in this book.

Förster resonance energy transfer or FRET has been mentioned previously in Section 8.4.1 and its application in biosensing is described in Chapter 6, Section 6.9. It relies on the use of a pair of fluorophores where the emission spectrum of one dye (donor) overlaps with the excitation spectrum of the other (acceptor). Under suitable conditions, excitation of the donor fluorophore leads to energy transfer to the acceptor and emission of a different wavelength of light than would be expected from the donor alone. This is illustrated in Figure 8.36 which shows FRET between cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP). The energy transfer effect only happens over short distances, the Förster length is typically a few nm, which is significantly less than the wavelength of the light used to excite the donor. The acceptor effectively quenches the donor, reducing the energy that would cause it to fluoresce.

FRET can be exploited in a biosensor if it is possible to arrange for the dyes to be brought together to indicate the presence or absence of whatever is being measured. For example in an immuno-assay where the antibody probe is labelled with one fluorophore and the target is labelled with the other. The sensor discussed here avoids the need to label the target molecule through the use of a specific DNA structure known as a Holliday junction (HJ). This is a

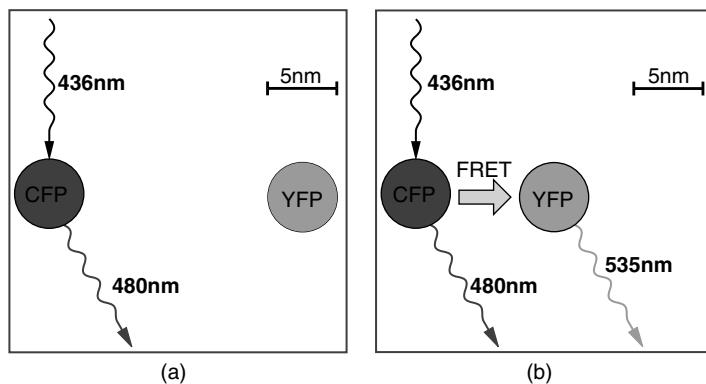


Figure 8.36 (a) When fluorophores are separated by more than the Förster distance there is no energy transfer and fluorescence emission from the donor (CFP) is seen. (b) When the dye molecules are close together FRET occurs and emission from the acceptor fluorophore can be observed.

4-way junction of DNA strands that is usually artificial but does occur naturally in certain biological processes. The biosensor exploits the fact that this molecule can be switched between two different conformations as illustrated in Figure 8.37. It is believed that the change in the HJ from the open to the closed conformation occurs when it is exposed to cations, which act to screen charges in the HJ that normally prevent the collapse of the structure.

The HJ can be turned into a sensor by attaching FRET fluorophores to the arms so that they are kept apart in the open conformation and brought together in the closed. This means that the FRET signal can be turned on and off by controlling the concentration of cations in the environment of the HJ molecule. Figure 8.38 shows the way in which the emission spectrum of the FRET will change depending on the HJ conformation. With no FRET (open conformation) the spectrum is dominated by the emission peak from the donor fluorophore but when the HJ is closed this peak is lowered due to quenching and a second peak can be observed indicating fluorescent emission from the acceptor. The conformation of the

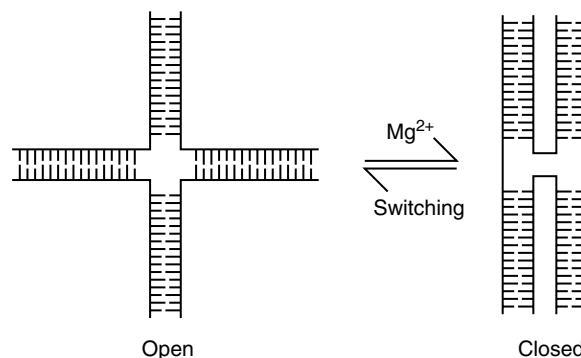


Figure 8.37 Schematic representation of a Holliday junction which can switch between two different conformations.

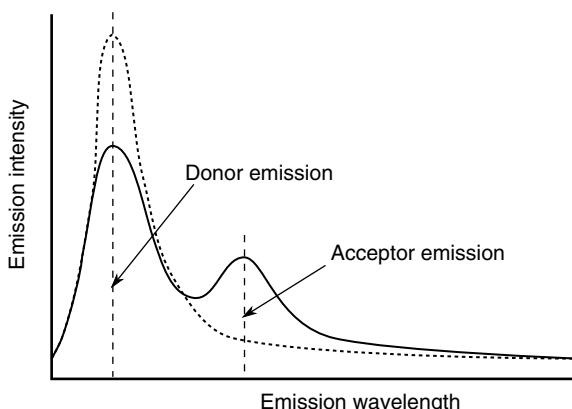


Figure 8.38 Representative graph of emission spectra from a pair of fluorescent dyes on an HJ with (closed) and without (open) FRET occurring.

molecule can be measured with a ratiometric technique where the intensity of the fluorescence is measured at the peak emission wavelengths of the two dyes. The ratio between these intensities will provide information about the level of FRET and therefore the shape of the HJ.

The HJ can be applied as a biosensor which is capable of detecting a specific sequence of nucleotides through the creation of an incomplete HJ molecule made up from only three strands of DNA. In this molecule, two arms of the probe junction will be incomplete and should have complementary sequences to the single stranded target DNA. Hybridisation of the probe and target will complete the Holliday junction structure. The incomplete probe-HJ will already have FRET fluorophores attached so the target DNA does not need to be labelled. The probe-HJ will not switch to allow FRET and so we can determine the presence of the target DNA sequence by actively switching the completed Holliday junctions and measuring the change in the fluorescence ratio. This type of biosensor has been demonstrated using changes in the concentration of Mg^{2+} ions to switch the completed HJ molecule [14]. In addition, it has been shown that a similar sensor using fluorescence lifetime measurements to detect the FRET effect is capable of detecting changes to a single nucleotide in the target molecule [15].

Those publications demonstrate the technique using changes in the bulk concentration of ions in the solution around the HJ molecules but it is also possible to use electrochemical methods to change the ion concentration. This opens up the possibility of a biosensor micro-system using this technique where the switching ion concentration is controlled with micro-fabricated interdigitated electrodes. These are fabricated in platinum before one half of the electrode pair is electroplated with silver and converted into a Ag/AgCl reference/counter electrode. The other half, the working electrode, is coated with an electrochemically active polymer film which is loaded with magnesium ions. Applying positive voltages to the ion switching film (WE) versus the RE causes the release of the Mg^{2+} ions while a negative voltage will lead to reabsorption of ions from the solution into the film.

In the prototype system used in initial experiments the incomplete HJ probes were spotted onto glass coverslips in a similar process to that used in making DNA microarrays. These are then fixed over the microelectrodes with an adhesive spacer in between. The chambers above

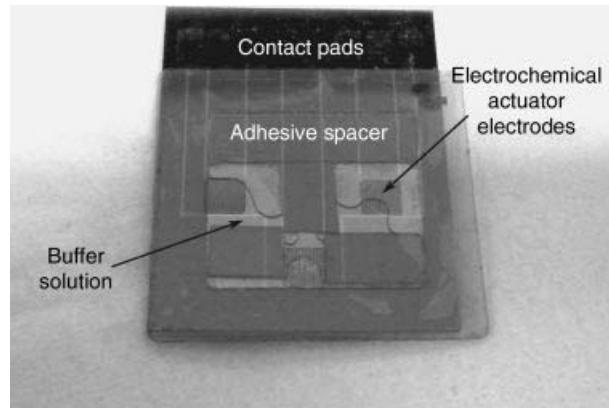


Figure 8.39 Photograph of an assembled lab on a chip system for the study of Holliday junction based DNA sensors using on-chip electrochemical switching.

the electrodes are filled with a solution, which acts as a buffer for the electrochemistry, before they are sealed. The buffer solution also contains either a complementary oligonucleotide to complete the HJ or a non-complementary target which will act as a control. Figure 8.39 shows an assembled device and full results can be found in [16].

8.5 Transducer Technology for Neuroscience and Medicine

This section will cover an important area of research and development in biomedical engineering, where microsystems technology is applied to the problems of interfacing electronics with nerves or brain tissue. Chapter 3 has introduced the electrogenic behaviour of certain cells and tissues, the most important being the neuron. This section will continue with a review of the different methods that are available for measuring the activity of neurons and interfacing with them in the laboratory. However, it is important to first describe the structure of the neuron in more detail.

8.5.1 The Structure of a Neuron

Figure 8.40 shows the structure of a typical neuron, the basic cell of the nervous system. The main cell body is often referred to as the soma and contains the nucleus and the rest of the cellular machinery that keeps the cell alive and working. The neuron is electrically excitable, as described in Chapter 3, and has inputs and outputs. The inputs are the dendrites which receive signals from other neurons, while the output is via the axon. This is a long cellular process and each neuron can only have one, though it usually has multiple branches and terminals at the other end where it connects to the dendrites of other neurons. Axons can be extremely long, sometimes even extending the entire length of the organism they are part of. In order to speed the transmission of signals along the axon, most neurons in vertebrates have myelination of the axon process. The myelin sheath is an insulating layer formed by glial cells that reduces the capacitance of the cell membrane allowing the action potential to move more quickly. The regular gaps in the myelination are known as nodes of Ranvier and contain large concentrations of ion channels. The action potential effectively jumps from one node to

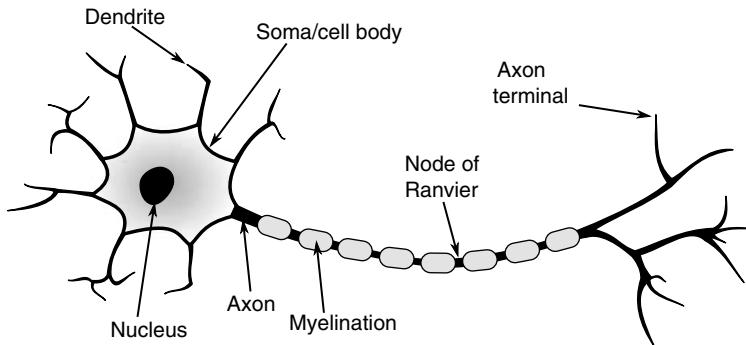


Figure 8.40 Schematic diagram of the structure of a neuron.

the next as it is transmitted down the axon, in a process known as saltatory conduction. Diseases which cause the loss of myelination, such as multiple sclerosis, can have a serious debilitating effect on the nervous system by affecting the transmission of signals down axons.

8.5.2 Measuring and Actuating Neurons

One of the first methods of electrically interacting with neurons to be developed was the voltage clamp, which involves inserting electrodes into the cell to control the potential across the membrane without interference from ion channel effects. This was first achieved with the squid giant axon, which is around 1mm in diameter, as it was large enough for the insertion of the macroscale electrodes available in the 1950s. The reason for the large size of the squid giant axon is that most invertebrates like molluscs do not have myelination of their axons meaning that a large diameter is required for fast transmission of action potentials over significant distances. This experiment allowed the researchers involved to monitor the opening and closing of ion channels during an action potential for the first time. As technology developed it became possible to fabricate microscale glass capillary electrodes to perform similar experiments in much smaller cells, such as mouse neurons. The voltage clamp technique is illustrated in Figure 8.41. It involves the insertion of two glass capillary electrodes through the cell membrane into an axon where they make contact with the cytoplasm. One of the internal electrodes, to the right of the figure, is connected to the high impedance input of a differential amplifier with a low gain. It is used to sense the membrane potential V_m between the internal electrode and an external potential sensing electrode. This potential is fed into one input of another differential amplifier along with a control voltage V_{in} . The output of this high gain amplifier is connected to the other capillary electrode, which will force a current into the cell in order to try to make the measured membrane potential equal to the control voltage. The current required will balance the net flow of charge into or out of the cell through the ion channels in the cell membrane. Therefore this technique can be used to study voltage gating of these ion channels by modulating the membrane potential and recording the membrane current, I_m . This current can be monitored with a pair of external sensing electrodes by measuring the voltage drop in the solution across a short distance. This voltage drop will be dependent on the resistance of the solution, R_s so that:

$$V = I_m R_s. \quad (8.52)$$

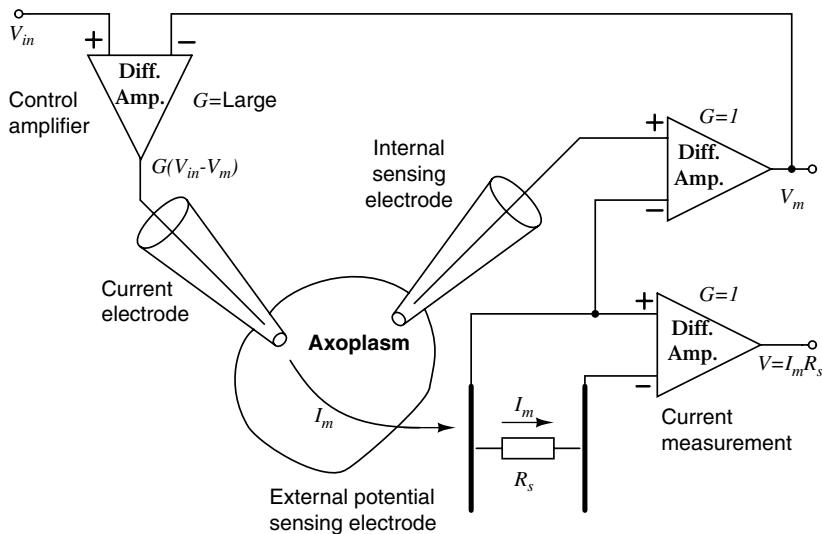


Figure 8.41 Schematic diagram of a voltage clamp setup for electrophysiology.

It should be clear from Chapter 7, Figure 7.15, that this technique, and the instrumentation used, is analogous to the potentiostat in electrochemistry.

An alternative method for the investigation of the electrophysiological behaviour of neurons and other electrogenic cells is the current clamp, which is illustrated in Figure 8.42. A single

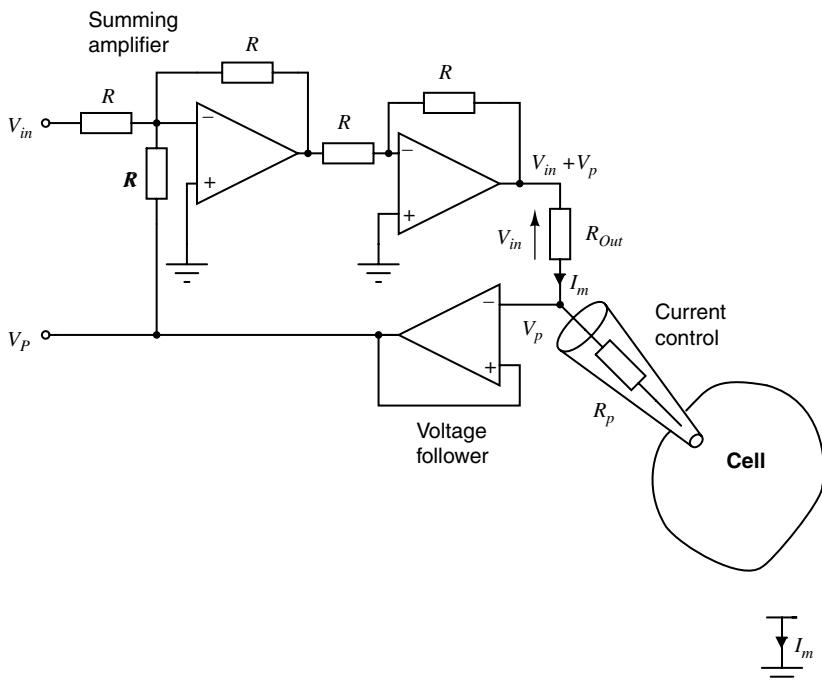


Figure 8.42 Schematic representation of a current clamp setup for electrophysiology.

capillary electrode is inserted into the cell, and is used to inject a controlled current. This could be used to model the excitation of an ion channel for example. The injected current is generated by controlling the voltage across the resistor R_{out} , and this is achieved by sensing the probe voltage V_p relative to the earthed culture solution with a voltage follower. This probe voltage is added to the control voltage V_{in} with a summing amplifier, and this feedback will try to keep the voltage across R_{out} equal to V_{in} . The current flow into the cell will then be:

$$I_m = \frac{V_{in}}{R_{out}}, \quad (8.53)$$

which can either be monitored by placing a current follower circuit, similar to that used to sense working electrode current in a potentiostat, on the solution reference electrode or by adding a differential amplifier to measure the voltage across R_{out} . The probe voltage V_p is not exactly the same as the membrane potential of the cell as there is an additional voltage drop due to the resistance of the capillary electrode, which can be considerable.

The patch clamp technique, as described in Chapter 3, Section 3.15, involves the use of a very fine-drawn glass needle, filled with an electrolyte solution and including a suitable measurement electrode. A highly accurate micropositioning system is used to bring the needle into contact with a neuron while the cell is viewed through a microscope. Suction is applied to the needle to create a tight seal to a patch of the cell membrane. This is often referred to as a ‘giga-ohm seal’ or ‘gigaseal’ because the resistance between the inside of the needle and the solution surrounding the cell should be very high. The patch clamp technique potentially allows the measurement of the current through a single ion channel as it is manipulated by changing the cellular environment. For example the needle may be preloaded with neurotransmitter chemicals to activate ligand-gated ion channels or the potential on the electrode may be manipulated to control voltage gated channels. If the pressure used to suck the cell onto the needle is increased beyond some critical point it can break the membrane and allow access to the cell interior for whole cell measurements. This is destructive to the cells as the electrolyte in the needle gradually replaces the cytoplasm, meaning that the measurements are only useful for a few minutes after the membrane is broken. Other patch clamp techniques may involve tearing a small piece of cell membrane away so that it can be studied separately, either with the original inside surface facing outwards or the other way around. As with the whole cell measurements, these ‘inside-out’ and ‘outside-out’ patch techniques are typically destructive to the cell.

An example of a possible instrumentation setup for patch clamp measurements, often referred to as a head-stage amplifier, is shown in Figure 8.43. Here the current is sensed as the voltage developed across the feedback resistor R_f in an op-amp current follower circuit. The non-inverting terminal of the op-amp is connected to an input voltage signal V_{in} and, due to the feedback, this will control the potential on the pipette electrode. The input voltage and the output of the op-amp are then fed into a differential amplifier and the output of this, assuming unity differential gain, will be equal to the voltage across R_f , which is proportional to the current being measured.

The planar patch clamp is a variation on the patch clamp where the connection to the cell is made via a micromachined aperture in a flat surface, for example a silicon

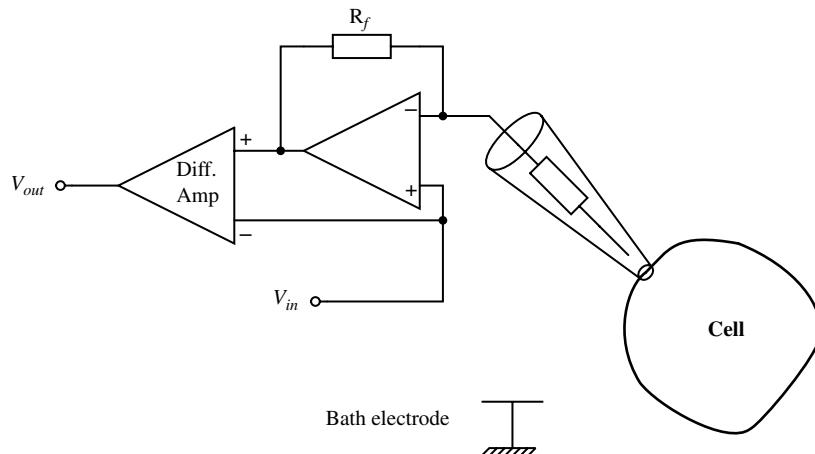


Figure 8.43 Schematic representation of a patch clamp setup for electrophysiology.

substrate. The cells are sucked down onto the aperture by applying negative pressure to the electrolyte on the other side of the aperture, where an electrode is used to control the voltage and measure current as in the standard patch clamp. Use of microfabrication techniques could potentially enable mass production of identical, optimised patch clamp apertures with obvious advantages over the standard bespoke fabrication of glass pipettes which requires considerable skill. Other advantages of the planar patch clamp include the possibility of using arrays of apertures to monitor connected networks of neurons, and integration with microfluidics and microelectronics. One of the major difficulties with a planar patch clamp is achieving a reliable, high-impedance seal between the aperture and the cell. Researchers have tried various methods to ensure this, including different materials to ensure a smooth surface and the fabrication of a protruding lip but results are often inconsistent [17].

8.5.3 Extracellular Measurements of Neurons

As an alternative to intracellular and patch clamp methods, electrodes in proximity to neurons and other electroactive cells can measure behaviour such as the generation and propagation of action potential spikes. This technique measures transient potentials in the extracellular space induced by ion currents but the resulting signals are significantly weaker than those that can be measured inside cells. Therefore the method depends on the use of sensitive, low noise amplifiers and could benefit from an approach where amplification is integrated close to the measurement electrodes. For *in vivo* measurements the electrodes may be very fine wires of compatible metals such as platinum, stainless steel or titanium coated in a suitable insulator. The area of the electrode will depend on the diameter of the wire and how much of the insulation is removed from the tip. If the electrodes are on a similar scale to the cells they can potentially measure the activity of a single neuron, but in many cases there will be crosstalk from other cells in the vicinity.

Microelectrode arrays (MEA) are a common tool in neurophysiological research looking at cultured neuron cells or sectioned brain tissue. Typically, these are planar arrays of individually addressable electrodes capable of extracellular recording and stimulus of cells. Common systems have passively addressed arrays of over 60 electrodes, which could potentially require a measurement system with the same number of low noise amplifiers for simultaneous, multichannel recording. However, the use of multiplexing and sampling at a much faster rate than the signals of interest could obviously allow more electrodes to be measured using a single measurement system. Another alternative which would potentially allow much larger arrays of electrodes would be the integration of electronics for addressing and amplification. The main disadvantage of this, beyond the increased cost of the electrode array, is that silicon substrates are not transparent which precludes the use of an inverted microscope to view the cells.

Electrode arrays can be used to both excite action potentials and measure the response. Forcing current through the electrode can stimulate nearby cells or alternatively a voltage pulse can set up large transient currents across the electrode double-layer capacitance for a similar effect. If an action potential is induced and propagates to neurons elsewhere in the tissue section the response can be measured using other electrodes in the array.

Extracellular microelectrode array measurements typically require low-noise, voltage amplification with very low input bias currents and built in band-pass filtering to remove unwanted signals. Commercially available implementations often include separate head stage amplifiers for each electrode in the array and these may be built into the equipment used to make electrical contact to the MEA chip so that they are as close as possible.

Problems

- 8.1. One of the most straightforward methods of monitoring temperature is to use the thermal variation of a resistor.
 - (a) Suggest a problem with the use of a standard Wheatstone bridge setup in a remote application like an implanted sensor.
 - (b) Can you suggest a possible solution?
 - (c) The alternative might be to locate the resistor at the remote sensing position while the instrumentation is in a stable environment. Why does this lead to a problem?
 - (d) Why does a Kelvin measurement improve measurements of small resistances?
- 8.2. A platinum based strain gauge with an unstrained resistance (R_G) of $1\text{ k}\Omega$ has a gauge factor (GF) of 2.
 - (a) If the gauge is securely fixed to a measured structure undergoing a strain (ϵ) of 10%, what will be the maximum change in resistance (ΔR)? Use Equation (8.9) for the gauge factor.
 - (b) You have four gauges to measure a structure with a maximum strain of 1%. How would you arrange the gauges in a Wheatstone bridge in order to maximise the output voltage?
 - (c) What will the output voltage range be if the bridge is supplied with $V_S = 5\text{ V dc}$?
 - (d) Assuming that it is only possible to attach two of the 4 gauges to the structure under test, how would you do this and what effect would it have on the sensitivity?
(Note: there are two possible answers.)

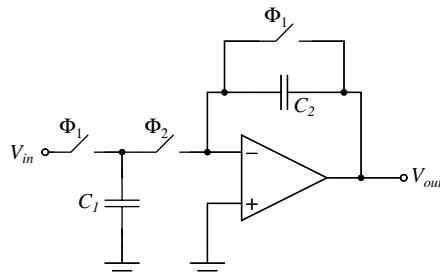


Figure 8.44 Switched capacitor amplifier.

- (e) How could you add to the Wheatstone bridge circuit in order to allow small mismatches in the gauge resistors to be corrected?
- 8.3. In the following question, refer to Section 8.3.8, Figure 8.25 and Equations (8.38) to (8.42) for a half-bridge capacitive sensing circuit.
- Assuming that it is possible to rearrange the circuit so that all four capacitances are sensors such that $C_4 = C_1$ and $C_3 = C_2$ what is now the equation for the output voltage $V_{out} = V_2 - V_1$?
 - If the area of one of the differential capacitors is $400 \times 400 \mu\text{m}$ and the separation $d = 1 \mu\text{m}$ calculate the amplitude of the output voltage from the bridge when the deflection $x = 10 \text{ nm}$ and the input voltage has an amplitude of $v = 5 \text{ V}$.
- 8.4. The circuit shown in Figure 8.44 is controlled by two non-overlapping clock signals as in Figure 8.28. With the help of sketches of the circuit during the two different phases, describe the operation of the amplifier and state the transfer function V_{out}/V_{in} .
- 8.5. Figure 8.43 shows a circuit that could be used in patch clamp measurements of the ion current (I_{ch}) through a single ion channel.
- Derive an expression for the output voltage V_{out} .
 - If I_{ch} is in the pA range suggest a suitable value for R_f .
 - Assume that the resting value of the transmembrane potential is -70 mV and the patch clamp has been successfully located over a voltage gated potassium (K^+) ion channel. How could this ion channel be opened?
 - What will be the result in terms of the ion channel current?

References

- [1] Brokaw A simple three-terminal IC bandgap reference. *IEEE Journal of Solid-State Circuits*, **9** (6), 388–393. (1974).
- [2] Spark, N.T. (May 2006) *A History of Murphy's Law*, Periscope Film, ISBN: 978-0978638894.
- [3] Spark, N.T. (February 2003) The Fastest Man on Earth, *Annals of Improbable Research*, <http://improbable.com/archives/paperair/volume9/v9i5/murphy/murphy0.html>.
- [4] Chaehoi, A., Begbie, M., Weiland, D. et al. (2011) Piezoresistive sensors development using monolithic CMOS MEMS technology. *Sensors & Transducers Journal*, **11**, 1–9.
- [5] Lowrie, C., Desmulliez, M.P.Y., Hoff, L. et al. (2009) Fabrication of a MEMS accelerometer to detect heart bypass surgery complications. *Sensor Review*, **29** (4), 319–325.
- [6] Lowrie, C. (2010) *A Three-Axis Accelerometer for Measuring Heart Wall Motion*, PhD thesis, Heriot Watt University.
- [7] Rodahl, M., Höök, F., Krozer, A. et al. (1995) Quartz crystal microbalance setup for frequency and Q-factor measurements in gaseous and liquid environments. *Review of Scientific Instruments*, **66** (7), 3924.

- [8] Kaajakari, V. (2009) *Practical MEMS*, Small Gear Publishing.
- [9] Bracke, W., Puers, R. and Van Hoof, C. (June (2007)) *Ultra Low Power Capacitive Sensor Interfaces*, Springer Verlag.
- [10] Yazdi (2004) Precision readout circuits for capacitive microaccelerometers. *Sensors*, 2004. Proceedings of IEEE, vol. 1, pp. 28–31.
- [11] Jorge, P.A.S., Caldas, P. Rosa, C.C. et al. (2004) Optical fiber probes for fluorescence based oxygen sensing. *Sensors & Actuators: B. Chemical*, **103** (1–2), 290–299.
- [12] Walker, R.J. and Richardson, J. (2011) A 12896 pixel event-driven phase-domain-based fully digital 3D camera in 0.13m CMOS imaging technology. *Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2011 IEEE International, pp. 410–412.
- [13] Rae, B.R., Yang, J. McKendry, J. et al. (2010) A vertically integrated CMOS microsystem for time-resolved fluorescence analysis. *IEEE Transactions on Biomedical Circuits and Systems*, **4** (6), 437–444.
- [14] Buck, A.H., Campbell, C.J. Dickinson, P. et al. (2007) DNA nanoswitch as a biosensor. *Analytical Chemistry*, **79** (12), 4724–4728.
- [15] McGuinness, C.D., Nishimura, M.K.Y. Keszenman-Pereyra, D. et al. (2010) Detection of single nucleotide polymorphisms using a DNA Holliday junction nanoswitch – a high-throughput fluorescence lifetime assay. *Molecular BioSystems*, **6** (2), 386–390.
- [16] Smith, S. et al. (2012) *Demonstration of a Device for Label-Free DNA Detection Through Nanoswitching*. In Preparation.
- [17] Curtis, J.C., Baldwin, K. Dworak, B.J. et al. (2008) Seal formation in silicon planar patch-clamp micro-structures. *Journal of Microelectromechanical Systems*, **17** (4), 974–983.

Further Readings

- Molecular Devices (2008) or Axon Instruments (1993) *The Axon Guide: A Guide to Electrophysiology & Biophysics Laboratory Techniques*, Various editions available online.
- Standen, N.B., Gray, P.T.A. and Whitaker, M.J. (eds) (1994) *Microelectrode Techniques. The Plymouth Workshop Handbook*, Company of Biologists, ISBN: 978-0948601491.

9

Microfluidics: Basic Physics and Concepts

9.1 Chapter Overview

Microfluidics is the science of fluid flow in structures that have at least one dimension in the microscale (between 1 micrometre and 1 millimetre). Although the terms are often used interchangeably, *Lab-on-a-Chip* is used to describe devices that integrate *several* laboratory processes on a single chip, whereas *Micro-Total-Analysis Systems* (μ TAS) are considered to integrate *all* laboratory processes required for an analysis on a single chip. For both cases fluid flow in one or more channel networks, fabricated into or from a solid substrate, is an essential element of the analytical or preparative function of the device.

From this chapter readers will gain an appreciation of:

- (i) the basic molecular physics and properties of gases and liquids;
- (ii) Pascal's and Laplace's laws applied to static fluids;
- (iii) Bernouilli's and Poiseuille's laws applied to fluid flow in microchannels;
- (iv) application of Kirchhoff's electrical circuit laws to microfluidic networks;
- (v) Navier-Stokes and related continuity equations to describe fluid flow;
- (vi) the consideration of fluid physics at the continuum, meso and molecular scales;
- (vii) the processes controlling molecular diffusion in fluids;
- (viii) surface tension as a significant force in microfluidics.

9.2 Liquids and Gases

Fluids, unlike solids, do not resist shearing forces such as those acting at a solid surface. They continue to deform as long as the force is applied and assume the shape of the solid boundary, whereas solids can resist such shear and maintain an unsupported shape. As shown in Figure 9.1, fluid motion is controlled by the interaction and internal shear between fluid layers.

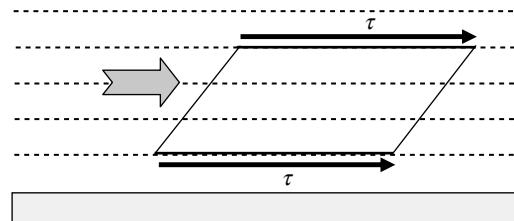


Figure 9.1 Fluids deform under the action of a shearing force (τ). The fluid can be considered as laminae parallel to a surface. Each fluid lamina applies a shear force τ to the next one, and is in turn sheared by those it touches.

Although they respond very differently to changes in pressure and temperature, the term ‘fluid’ includes both liquids and gases. Gases can be expanded and compressed more easily than liquids due to the lower density and larger spacing between molecules. At the molecular scale ($\sim 10^{-9}$ m) the interaction between layers involves collisions of many molecules. At the macroscale scale ($> 10^{-4}$ m) the physical properties of a fluid result from the statistical averages of such molecular interactions. The effects of individual molecular collisions can be ignored and we can deal with the liquid’s bulk, or continuum, properties.

9.2.1 Gases

The molecules in a gas are widely spaced and interactions between them (apart from collisions) are weak, especially at low pressures. An increase of temperature increases the kinetic energy of the molecules, mass transfer between gas layers increases and viscosity increases. In gases, except for extremely high pressures, viscosity is independent of pressure. At a sufficiently low pressure, where intermolecular interactions are negligible, all gases obey the *ideal gas law*

$$PV = nRT, \quad (9.1)$$

where P is the pressure, V the volume, n the amount (moles) of substance of gas molecules, and T the absolute temperature. In this equation R is the gas constant given by $R = kN_A$, where k is Boltzmann’s constant = 1.38×10^{-23} J K $^{-1}$, and N_A is Avogadro’s constant = 6.022×10^{23} mol $^{-1}$. The *ideal gas law* follows from experiment (e.g. Boyle’s Law) and Avogadro’s Hypothesis (formulated in 1810) that:

Equal volumes of gases at the same temperature and pressure contain the same number of molecules.

Thus, a mole of hydrogen (2 gm) and a mole of oxygen (32 gm) at the same temperature and pressure occupy the same volume. At standard temperature and pressure (STP: 273.15 K, 100 kPa) this volume is 22.414 litres.

9.2.1.1 Mean Free Path between Molecular Collisions in a Gas

Although on average the molecules in a gas are widely spaced apart, they are in constant motion and often collide with each other. If we treat each molecule as a hard sphere of

diameter d , the effective collision area for two colliding molecules can be taken as a circle of diameter $2d$. The effective cross-sectional collision area A_c for a molecule is thus

$$A_c = \pi d^2.$$

The frequency of collisions between sets of two molecules will depend upon their relative velocity \vec{v}_{rel} of approach. For two molecules having random velocities \vec{v}_1 and \vec{v}_2 this is given by their vector difference (see Figure 9.26 for an example)

$$\vec{v}_{rel} = \vec{v}_1 - \vec{v}_2.$$

The magnitude of the relative velocity is given by

$$v_{rel} = \sqrt{\vec{v}_{rel} \cdot \vec{v}_{rel}} = \sqrt{(\vec{v}_1 - \vec{v}_2) \cdot (\vec{v}_1 - \vec{v}_2)} = \sqrt{\vec{v}_1 \cdot \vec{v}_2 - 2\vec{v}_1 \cdot \vec{v}_2 + \vec{v}_2 \cdot \vec{v}_1}.$$

Velocities \vec{v}_1 and \vec{v}_2 are random and uncorrelated, and because the same average velocity $\langle \vec{v}_{rel} \rangle$ is associated with each molecule we have

$$\langle \vec{v}_{rel} \rangle = \sqrt{\vec{v}_1^2 + \vec{v}_2^2} = \sqrt{2}\langle v \rangle.$$

Over time t a collision cross-section associated with one molecule moving with an average velocity $\langle v \rangle$ will travel a path length $\langle v \rangle t$. The volume of collision space swept through during this time will be $\pi d^2 \langle v_{rel} \rangle t = \pi d^2 \sqrt{2} \langle v \rangle t$. The mean free path length L_{mfp} can then be taken as the path length $\langle v \rangle t$ divided by the number of molecular collisions:

$$L_{mfp} = \frac{\langle v \rangle t}{\pi d^2 \sqrt{2} \langle v \rangle t N_v} = \frac{1}{\pi d^2 \sqrt{2} N_v}.$$

N_v is the number of molecules per unit volume, and can be found from Avogadro's number and the ideal gas law given by Equation (9.1):

$$N_v = \frac{n N_A}{V} = \frac{P N_A}{R T} = \frac{P}{k T}.$$

The mean free path length between collisions of molecules in a gas is therefore

$$L_{mfp} = \frac{k T}{\sqrt{2 \pi P d^2}}. \quad (9.2)$$

In the worked Example 10.3 in Chapter 10 we find that the average distance of around 140 nm between collisions of nitrogen molecules in nitrogen gas is more than 550-times their molecular diameter (~ 0.25 nm) and some 40-times larger than their average molecular separation distance of 3.3 nm. This demonstrates that molecules in a gas can travel (in straight paths) over significant distances at the molecular scale before they collide with another molecule.

9.2.2 Liquids

The molecules in a liquid are closer together than for a gas, and cohesive forces such as those arising from intermolecular van der Waals interactions (see Chapter 1, Section 1.2.5) give rise to viscous effects. Glass and molten polymers are highly viscous because their large molecules get entangled. Water has a higher viscosity than liquids such as benzene or alcohols because of its network of cohesive hydrogen-bonds. An increase of temperature, and hence of molecular kinetic energy, reduces these cohesive forces and hence the viscosity. An increase of molecular kinetic energy will also facilitate an increased molecular interchange between the fluid layers, which will increase viscosity. This produces a relatively small effect and so the net result is that liquids show a reduction in viscosity for an increase in temperature. With increasing pressure the energy required for relative movement of molecules is increased, and the viscosity is increased.

9.3 Fluids Treated as a Continuum

When treated as a continuum, the properties of a fluid such as density, pressure and velocity remain constant at any defined point and changes in these properties due to molecular motions are taken to be negligible. The physical properties of fluids can be defined as continuous functions of time and space.

9.3.1 Density

This is defined as the mass contained within a unit volume, and is computed as the product of molecular mass m and the number of molecules N per unit volume V :

$$\rho = \frac{Nm}{V}.$$

Molecular mass ‘ m ’ is the mass of the molecule given by $m = M_w m_u$, where M_w is the molecular weight (molecular mass relative to $^{12}\text{C} = 12$) and m_u is the atomic mass unit (1.6606×10^{-27} kg). We can interpret the ideal gas law of Equation (9.1) as stating that pressure is linearly proportional to the product of temperature and density.

9.3.2 Temperature

Temperature relates to the translational kinetic energy E of N molecules in a particular volume domain – each molecule having velocity v_j and mass m :

$$E = \sum_{j=1}^N \frac{1}{2} m v_j^2.$$

The kinetic theory of gases uses statistical mechanics to relate the temperature to the average kinetic energy of the atoms in the system. At bulk scales the number of molecules is (almost) infinite and their average squared velocities can be assumed to follow the Maxwell

distribution to be described in Section 9.3.4. In one dimension, the average kinetic energy of the molecules is related to the temperature as:

$$\langle E \rangle = \frac{1}{2} m \langle v_x^2 \rangle = \frac{kT}{2}.$$

For a three-dimensional domain we have:

$$\langle E \rangle = \frac{1}{2} m \langle v_x^2 + v_y^2 + v_z^2 \rangle = \frac{3}{2} kT. \quad (9.3)$$

This relationship is an important aspect of our understanding of gas pressure.

9.3.3 Pressure

Pressure is the force imparted by collisions of molecules against a unit area of surface. Consider, as depicted in Figure 9.2, a single perfectly elastic molecule, of mass m , bouncing rapidly back and forth with velocity v_x between two solid surfaces distance L apart. The motion of the particle is assumed to remain on the same path parallel to the x-axis between the two surfaces. What is the resulting force exerted on each of the two surfaces?

We will use Newton's law (force = mdv/dt) in the form

$$\text{force} = \text{rate of change of momentum}.$$

The momentum lost Δp by a single molecular collision with the right-hand surface of Figure 9.2 is given by

$$\Delta p = p_{\text{initial}} - p_{\text{final}} = mv_x - (-mv_x) = 2mv_x.$$

The time Δt between successive collisions is

$$\Delta t = 2L/v_x.$$

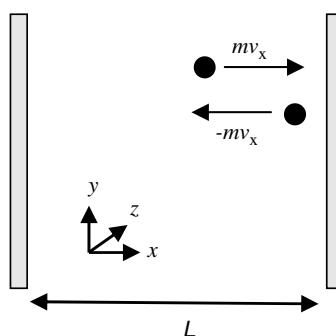


Figure 9.2 A perfectly elastic molecule bouncing between two solid boundaries with velocity v_x along a path parallel to the x-axis.

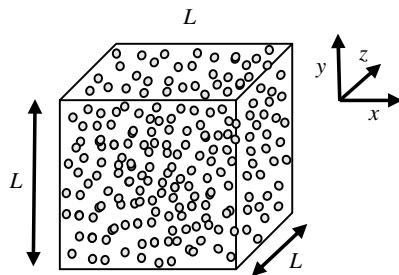


Figure 9.3 A large number of molecules in motion within a cubic box and continually colliding with the internal walls.

The force F is given by

$$F = \Delta p / \Delta t = mv_x^2 / L.$$

The same force will be exerted on the left-hand surface. We now generalise the situation to a large number of molecules colliding continuously with the walls of a cubic box of side L , as shown in Figure 9.3.

The force acting on each of the six walls of the cubic box is:

$$F = \left(m \sum_j [v_{jx}^2 + v_{jy}^2 + v_{jz}^2] \right) / L = \left(m \sum_j (v_j^2) / 3L \right). \quad (9.4)$$

We assume that the molecules are equally likely to be moving in any direction, so that the average value of v_{jx}^2 is the same as that of v_{jy}^2 or v_{jz}^2 . The average squared velocity of the N molecules is given by:

$$\langle v^2 \rangle = \frac{1}{N} \sum_j v_j^2. \quad (9.5)$$

Defining pressure p as the force per unit area, and noting that the product of cross-sectional area A and length L gives the volume V , from Equations (9.4) and (9.5) we have:

$$p = \frac{F}{A} = \frac{m \sum_j v_j^2}{3AL} = \frac{Nm\langle v^2 \rangle}{3V}. \quad (9.6)$$

The *macroscopic* pressure of a gas therefore relates directly to the average kinetic energy per molecule. From Equations (9.2) and (9.6) we have

$$p = \frac{1}{3} \rho \langle v^2 \rangle,$$

thereby describing pressure as a function of density and kinetic energy of molecules.

From Equations (9.1) and (9.6) we have

$$\frac{1}{2}m\langle v^2 \rangle = \frac{3pV}{2N} = \frac{3nRT}{2N}.$$

In this equation we can replace n , the moles of gas, with $n = N/N_A = Nk/R$, so that

$$\frac{1}{2}m\langle v^2 \rangle = \frac{3}{2}kT,$$

to give the result quoted in Equation (9.3).

9.3.4 Maxwell Distribution of Molecular Speeds

Equation 9.6 connects the dynamics of the system at the molecular level to the macroscopic pressure, which represents a physical property averaged over a very large number of molecules. The model used to derive this equation is simplistic. It uses Newton's law and assumes that the molecules do not interact through van der Waals forces, for example, or collide with each other. The molecules are depicted as moving along straight paths between their collisions with the solid surfaces, and that they behave as perfectly elastic bodies – which they are not – during a collision process. For a system of gas particles to reach thermal equilibrium with their neighbouring molecules and enclosing walls, inelasticity is in fact required to provide the necessary energy exchanges. Maxwell understood that at equilibrium the distribution of molecular speeds remains constant with time. Also, if we ignore the effects of gravity, the molecules will on average be evenly distributed throughout containers such as that shown in Figure 9.3. Therefore, we do not require details of the position and velocity of each molecule as a function of time.

In Section 9.3.3 the term *velocity* was employed because both the direction and speed of molecular motion was specified. The term *speed* on the other hand is a scalar not a vector quantity – it describes how fast a molecule is moving but not the direction of motion. When considering an assembly of molecules moving in three-dimensional space, for any given molecular speed there will be many possible velocity vectors. We need therefore to consider

$$v = \left[v_x^2 + v_y^2 + v_z^2 \right]^{1/2}. \quad (9.7)$$

Equation 9.7 represents a sphere of radius v centred at the origin. Thus, for a particular speed v all of the possible velocity vectors lie on the surface of a sphere of radius v . As v is made larger, the sphere increases in size and the number of possible velocity vectors increases in proportion to $4\pi v^2$. However, this situation cannot go on indefinitely because at some stage there must be fewer and fewer molecules to be found proceeding to higher speeds. Our task is to find out how the speeds of all the N molecules in the cubic box are distributed in this spherical volume of velocity space (see Figure 9.4).

It can be shown statistically that if there are N molecules in a defined volume of gas in velocity space, the fluctuation of the number with time is of the order one part in $N^{1/2}$. In Chapter 10, Example 10.3, we determine that a cubic volume $10 \mu\text{m} \times 10 \mu\text{m} \times 10 \mu\text{m}$ of nitrogen gas at STP contains around 10^{10} nitrogen molecules. The corresponding number

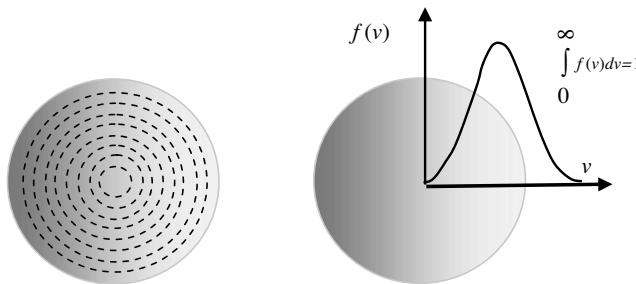


Figure 9.4 For a particular molecular speed v all of the possible velocity vectors lie on the surface of a sphere of radius v . The distribution function $f(v)$ describes the probability of finding a molecule of speed between v and $v + dv$.

fluctuation for even this small volume of gas is of the order one part in 10^5 . For a macroscopic volume of gas we can therefore assume the molecular density to be smoothly distributed in both ordinary and velocity space.

We will define $f(v_x)$ to be the probability distribution function of velocities, so that $f(v_x) dv_x$ is the *fraction* of the total number of molecules in a fixed volume to have a velocity between v_x and $v_x + dv_x$. The sum of all these possible fractions must be 1:

$$\int_{-\infty}^{\infty} f(v_x) dv_x = 1.$$

For a volume of gas containing N molecules, the number of molecules having velocity between v_x and $v_x + dv_x$ is $N f(v_x) dv_x$. Extending this to three-dimensional velocity space, the number of molecules of velocity lying between v_x and $v_x + dv_x$, v_y and $v_y + dv_y$, and v_z and $v_z + dv_z$, is:

$$N f(v_x) f(v_y) f(v_z) dv_x dv_y dv_z.$$

Because the probability distribution function must depend only on the total speed of a molecule and not on the separate velocity components, Maxwell deduced that

$$f(v_x) f(v_y) f(v_z) = F(v_x^2 + v_y^2 + v_z^2). \quad (9.8)$$

Thus, the *product* of the distribution functions manifests itself as the *sum* of the velocity components. Solutions for Equation (9.8) take the form

$$f(v_x) = Ae^{-Bv_x^2}. \quad (9.9)$$

The number of molecules in a spherical elemental shell of volume $dv_x dv_y dv_z$ can thus be given as:

$$\begin{aligned} N f(v_x) f(v_y) f(v_z) dv_x dv_y dv_z &= NA^3 e^{-B(v_x^2 + v_y^2 + v_z^2)} dv_x dv_y dv_z \\ &= NA^3 e^{Bv^2} dv_x dv_y dv_z \end{aligned} \quad (9.10)$$

For very small speed increments dv , the volume of velocity space between two spheres of radius v and $v + dv$, both centred at the origin, is equal to $4\pi v^2 dv$. The fractional distribution function $f(v)dv$ is thus the result given by Equation (9.10) multiplied by the factor $4\pi v^2 dv/N$:

$$f(v)dv = 4\pi v^2 A^3 e^{-Bv^2} dv.$$

The sum of all the fractions of molecules in velocity space must together add up to 1:

$$\int_0^\infty f(v) dv = \int_0^\infty 4\pi v^2 A^3 e^{-Bv^2} dv = 1. \quad (9.11)$$

The final result we require makes use of standard integrals to solve for Equation (9.11) together with Equation (9.3) to find the average kinetic energy per molecule:

$$\frac{1}{2}m\langle v^2 \rangle = \frac{3}{2}kT = \int_0^\infty \frac{1}{2}mv^2 f(v) dv.$$

This gives $B = m/2kT$ and $A = 4\pi(m/2\pi kT)^{3/2}$ to be used in Equation (9.11), so that Maxwell's probability distribution function for the molecular speeds is given by:

$$f(v) = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} v^2 e^{-mv^2/2kT}. \quad (9.12)$$

Examples of this speed distribution for nitrogen gas at three different temperatures are presented in Figure 9.5. At the lower speeds the function increases from zero at a parabolic rate to a maximum value and then decreases exponentially with increasing speed. The area under each curve must, according to Equation (9.11), equal 1.

The maximum speed value can be obtained by differentiating Equation (9.12), equating this to zero and solving for v_{\max} . The result we obtain is:

$$v_{\max} = \sqrt{(2kT/m)}.$$

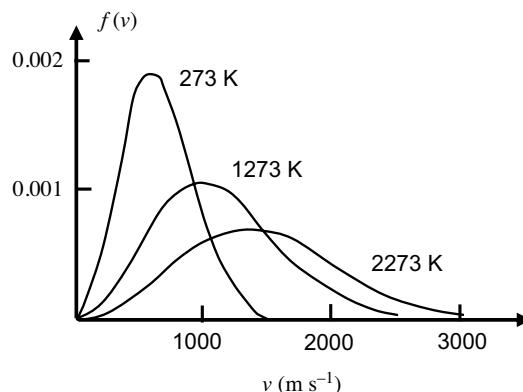


Figure 9.5 The Maxwell speed distribution function for nitrogen gas at three different temperatures.

The mean (arithmetic average) speed $\langle v \rangle$ is given by:

$$\langle v \rangle = \sqrt{(8kT/\pi m)}. \quad (9.13)$$

Equation 9.13 can be used to determine the mean time between collisions of the nitrogen molecules, as well as the collision frequency (see Example 10.5 in Chapter 10). The root-mean-square speed, associated with the mean kinetic energy, is given by:

$$v_{\text{RMS}} = (\langle v^2 \rangle)^{1/2} = \sqrt{(3kT/m)}.$$

The kinetic energy E is equal to $mv^2/2$, and so from equation (9.12) the speed distribution can also be given as

$$f(v) = 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} v^2 e^{-E/kT}.$$

The Maxwell distribution function is only valid for atoms or molecules in the gaseous state. The laws of classical physics can be applied to particles. However, molecules or atoms in the liquid or solid state have to obey the laws of quantum statistics, in which case only certain energy values, rather than a continuous distribution, are permitted.

9.3.5 Viscosity

Viscosity is the measure of the effort required to deform a fluid and is often discussed in terms of *Couette* flow, corresponding to the situation shown in Figure 9.6 where a fluid is contained between a moving plate and a parallel stationary plate. The action of rubbing a cream lotion into skin corresponds to applying a shear stress to the lotion, which will then experience Couette flow. In a macro-scale Couette flow device the fluid velocity immediately next to a surface will equal the velocity of that surface. This is referred to as *zero slip*. If the fluid is a *Newtonian* fluid, such as water, the fluid velocity will change smoothly from zero at the stationary surface to the velocity of the moving surface. In other words the spatial gradient of the fluid velocity dv/dy is a constant.

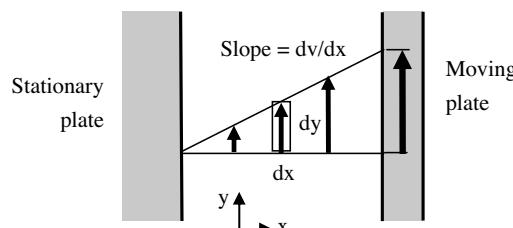


Figure 9.6 Couette flow induced in a fluid contained between a moving plate and a stationary surface. For a Newtonian fluid the fluid velocity v will change smoothly from zero at the stationary surface to the velocity of the moving plate. The spatial velocity gradient dv/dx is a constant.

The *dynamic viscosity* η is defined as the proportionality constant between the shear stress τ applied to the fluid and the resulting rate of shear strain. The shear strain is defined as the ratio dy/dx of the lateral deformation dy to the thickness dx of the layer being displaced, and the rate of shear strain is simply $(dy/dx)/dt = dv/dx$ (namely, the induced velocity gradient). For a Newtonian fluid we therefore have the following relationship between stress, rate of shear strain, and the dynamic viscosity:

$$\tau = \eta dv/dx. \quad (9.14)$$

We can interpret this relationship as indicating that for $\eta = 1 \text{ Pa.s}$ and $\tau = 1 \text{ Pa}$, the mobile plate moves in one second a distance equal to the thickness of the fluid layer between the plates. Values of the dynamic viscosity for some liquids and gases are given in Table 9.1.

The viscosity of a Newtonian fluid depends only on temperature and concentration (if diluted with another miscible fluid). For some fluids, particularly molten polymers or biological fluids such as blood, their viscosity depends also on the internal stress. These are classed as non-Newtonian fluids. Their viscosity decreases with an increase of the rate of the applied shear stress $d\tau/dt$ applied to a fluid flowing between two parallel surfaces, one moving at a constant velocity and the other one stationary, and is defined by:

$$d\tau/dt = v/h,$$

where v is the velocity of the moving surface, and h is the distance between the two parallel surfaces.

Non-Newtonian fluids exhibit viscoelastic behaviour (shear-thinning) and some require an initial shear stress that must be applied before they begin to flow – as for the case of whole blood, for example. Viscoelastic fluids exhibit a relaxation time, typically ranging from milliseconds to seconds, given by the reciprocal of the critical shear rate. The critical shear rate corresponds to the shear threshold at which the viscosity begins to change or, for the case of molten polymers, where the polymer chains make the transition from a coiled to a stretched configuration.

Table 9.1 Viscosity of some liquids and gases (293 K unless specified)

Liquid	Viscosity (Pa.s)	Gas	Viscosity (Pa.s)
Water	1.0×10^{-3}	Water vapour (373 K)	1.3×10^{-5}
Blood (whole) (310 K)	$3 \sim 4 \times 10^{-3}$	Air	1.8×10^{-5}
Blood (plasma) (310 K)	1.5×10^{-3}	Argon	2.1×10^{-5}
Ethyl alcohol	1.2×10^{-3}	Helium	1.9×10^{-5}
Glycerine	1.49	Methane	2.0×10^{-5}
Oil (light)	0.11	Nitrogen	1.8×10^{-5}
Oil (heavy)	0.66	Oxygen	2.0×10^{-5}

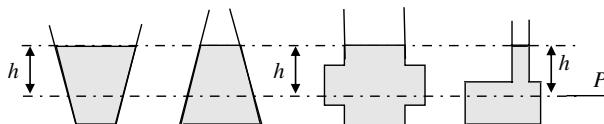


Figure 9.7 The static fluid pressure P at a given depth h does not depend upon the total mass or total volume of the liquid, and so the shape of the vessel is irrelevant. The static fluid pressure is given as $P = \rho gh$, where ρ is the fluid density and g the acceleration of gravity.

9.4 Basic Fluidics

9.4.1 Static Fluid Pressure

The pressure exerted by a static fluid arises from the weight of that fluid and so depends only upon the fluid depth h , its density ρ , and the acceleration of gravity g :

$$P_{\text{static fluid}} = \frac{\text{weight}}{\text{area}} = \frac{mg}{A} = \frac{\rho V g}{A} = \rho gh. \quad (9.15)$$

Because the fluid pressure at a given depth h does *not* depend upon the total mass or total volume of the liquid, the shape of the fluid container is also not relevant. Examples of this are given in Figure 9.7.

Pressure is measured in units of Pascals, but it is also not unusual to find pressures expressed in column height units (e.g. mm Hg), reflecting the height dependence shown in Equation (9.15).

9.4.2 Pascal's Law

Pascal's law is expressed as:

The pressure exerted anywhere in an enclosed, incompressible, static fluid is transmitted equally in all directions throughout the fluid.

This follows from the above description of static fluid pressure given by Equation (9.15), namely that the change in pressure between two elevations is due to the weight of the fluid between the elevations regardless of the geometry of the container.

Pascal's law can be interpreted to indicate that any *change* in pressure applied at any given point of the fluid is transmitted *undiminished* throughout the fluid. As shown in Figure 9.8 this makes possible a large multiplication of force and forms the basis for the operation of a hydraulic press that provides the means to lift a heavy weight with a small force.

In Figure 9.8a the fluid pressure P_2 at the base of the vessel is $P_2 = P_1 + \rho gh$. The force F_2 acting on the base is given by

$$F_2 = P_1 A_2 + (\text{force arising from weight of fluid}).$$

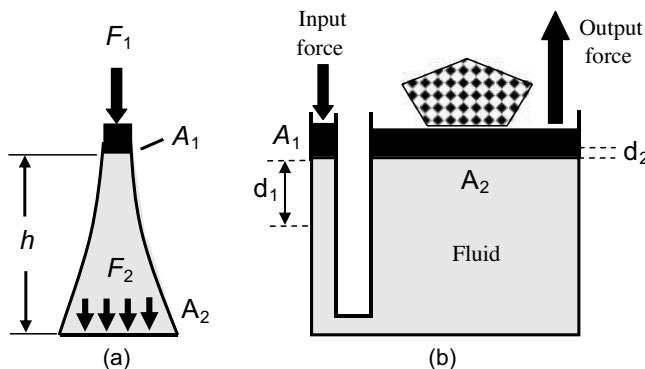


Figure 9.8 Demonstrations of Pascal's Law: (a) The pressure in a vessel is transmitted equally throughout a fluid; (b) The operation of a hydraulic press relies on the fact that any change in pressure applied at any given point of a fluid is transmitted undiminished throughout the fluid.

Because $A_2 \gg A_1$ there is an amplification of force. This is demonstrated in the basic hydraulic press depicted in Figure 9.8b. If we ignore frictional losses

$$\text{Input Force} \times \text{distance } d_1 = \text{Output Force} \times \text{distance } d_2.$$

$$\text{Thus } d_1 = (\text{Output Force}/\text{Input Force})d_2 = (A_2/A_1)d_2$$

which is equivalent to the action of a *fluidic lever*.

9.4.3 Laplace's Law

Laplace's law states that

The tension on the wall of a cylindrical chamber is the product of the pressure times the radius of the cylinder (or half that value for a spherical chamber).

Thus, a vessel of large radius will require a larger wall tension than one of smaller radius to withstand a given internal fluid pressure. Also, for a given vessel radius and internal pressure, a spherical vessel will have half the wall tension of a cylindrical vessel. This is demonstrated in Figure 9.9.

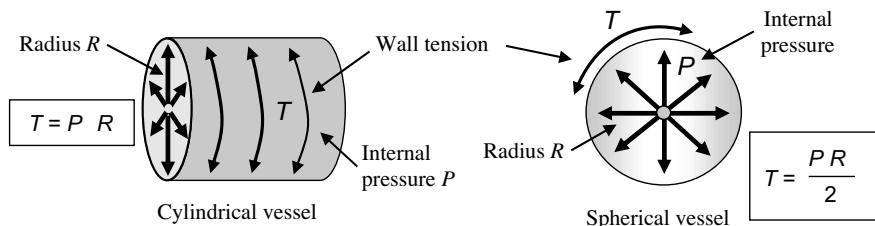


Figure 9.9 Laplace's Law informs us that the tension on the wall of a cylindrical vessel is twice that for a spherical vessel of the same radius and internal pressure.

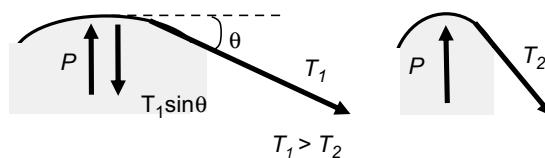


Figure 9.10 Wall tension T increases with vessel radius because for a fixed internal pressure P , the counter component of the wall tension $T\sin\theta$ must equal P .

An explanation of why wall tension increases with radius is given in Figure 9.10. Basically, if the fluid pressure remains constant then the inward component ($T\sin\theta$) of the wall tension must remain the same. If the wall curvature ($\sin\theta$) is less, the total wall tension must increase in order to obtain the same inward component of tension.

The flow of blood in arteries and veins is a good example of Laplace's law in action. The larger arteries of the body are subject to higher wall tensions than the smaller arteries having comparable blood pressures. Arteries are reinforced by fibrous bands to strengthen them against the risks of an aneurysm (capillaries with their very thin walls rely on their small radii). If an artery wall develops a weak spot and expands as a result, the expansion subjects the weakened wall to even more tension. The weakened vessel may continue to expand in what is called an aneurysm, and lead to rupture of the vessel. This is why aneurysms require prompt medical attention.

9.5 Fluid Dynamics

We will now consider the flow of fluids in channels and the forces that act on them.

9.5.1 Conservation of Mass Principle (Continuity Equation)

Fluid flowing steadily through a pipe of reducing cross-sectional area is shown in Figure 9.11. We can assume that all of the fluid mass passing through area A_2 will exit the pipe and be measured as Q in units of gm s^{-1} . If ρ_2 is the density of the fluid flowing through A_2 the value for Q is equal to $A_2\rho_2v_2$, where v_2 is the effective velocity of the fluid flow through A_2 .

No fluid can exit or enter the pipe between areas A_1 and A_2 , and so from the conservation of mass principle the mass of fluid crossing each section of the pipe per unit time must be the same:

$$\begin{aligned} \text{Fluid Flow Rate through } A_1 &= \text{Fluid Flow Rate through } A_2 \\ A_1\rho_1v_1 &= A_2\rho_2v_2 \end{aligned}$$

This relationship, which takes the form of the *equation of continuity*, can be expressed as the *law of conservation of mass* in fluid dynamics:

$$A\rho v = \text{Constant.}$$

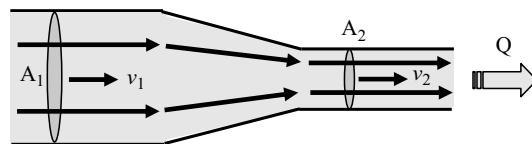


Figure 9.11 Fluid flow through a pipe whose cross-sectional area reduces from A_1 to A_2 . The rate of fluid flow Q through the pipe can be determined as either volumetric flow ($\text{dm}^3 \text{s}^{-1}$) or mass flow (gm s^{-1}). The conservation of mass principle dictates that $A_1 v_1 = A_2 v_2$ where v_1 and v_2 are the average fluid velocities through A_1 and A_2 , respectively. Because $A_1 > A_2$, then $v_1 < v_2$.

We will only consider fluids that are incompressible, and so the density of the fluid will be constant ($\rho_1 = \rho_2$). The conservation of mass principle can thus be written as

$$A_1 v_1 = A_2 v_2$$

or

$$Av = \text{Constant.}$$

The reduction in pipe area shown in Figure 9.11 indicates that $v_2 > v_1$.

Example 9.1

Water flows through a channel of circular cross-section at a rate of 0.1 mL per minute. The channel has the same geometrical profile as that depicted in Figure 9.11, with the radius constricting down from 100 to 60 μm . Calculate the effective velocity of fluid flow through areas A_1 and A_2 .

Solution:

$$\begin{aligned} \text{Volumetric flow } Q &= 0.1/60 = 1.67 \times 10^{-3} \text{ mL s}^{-1} = 1.67 \times 10^{-9} \text{ m}^3 \text{ s}^{-1} \\ A_1 &= \pi(10^{-4})^2 = 3.14 \times 10^{-8} \text{ m}^2. \end{aligned}$$

and so

$$v_1 = Q/A_1 = 1.67 \times 10^{-9} \text{ m}^3 \text{ s}^{-1} / (3.14 \times 10^{-8} \text{ m}^2) = 5.3 \times 10^{-2} \text{ m s}^{-1}.$$

From the principle of conservation of mass

$$\begin{aligned} A_1 v_1 &= A_2 v_2 \\ \therefore v_2 &= (3.14 \times 10^{-8} \text{ m}^2)(5.3 \times 10^{-2} \text{ m s}^{-1}) / (3.14 \times (6 \times 10^{-5})^2) \\ &= 0.15 \text{ m s}^{-1}. \end{aligned}$$

9.5.2 Bernoulli's Equation (Conservation of Energy)

Referring to Figure 9.11, the principle of conservation of mass informs us that the fluid velocity is greatest in the part of the pipe having the smaller cross-sectional diameter. In passing from the large cross-sectional area to the smaller one the fluid velocity increases. This corresponds to an acceleration of the fluid mass, which in turn requires an unbalanced force in the form of a pressure gradient exerted on the fluid by the walls of the pipe. As depicted in Figure 9.12, the pressure P_1 in the large area of the pipe must be greater than P_2 in order to accelerate the fluid. Likewise, if the fluid flow is reversed, P_1 must exceed P_2 in order to bring about deceleration of the fluid.

Bernoulli's principle states that, for *viscous free* fluid flow, an increase of the fluid velocity occurs simultaneously with a decrease in fluid pressure or a decrease in the fluid's potential energy. This principle can be applied to various types of fluid flow and quantified using various forms of what is known as Bernoulli's equation. A simple form of this equation is valid for incompressible fluids and for compressible gases moving at speeds well below the velocity of sound in a particular gas. This equation can be derived from the principle of conservation of energy, which states that in a steady fluid flow the sum of all forms of mechanical energy remains constant along a flow line. The fluid possesses kinetic energy due to its motion, and because of its location in the earth's gravitational field it also possesses potential energy. Work is also being done on the fluid due to the static pressure acting on it. If there are no frictional losses we can apply the law of conservation of energy and write Bernoulli's equation as:

$$P + \rho gh + \frac{1}{2}\rho v^2 = \text{constant}, \quad (9.16)$$

where P is the static pressure, h the height above some reference level, v the velocity, ρ the density, and g the acceleration due to gravity at any chosen elemental volume in the fluid flow line. The term $(\frac{1}{2}\rho v^2)$ is known as the dynamic pressure, and the total pressure is the sum of the static pressure P and this dynamic pressure. The sum of the elevation h and static pressure head ($P/\rho g$) is known as the hydraulic head.

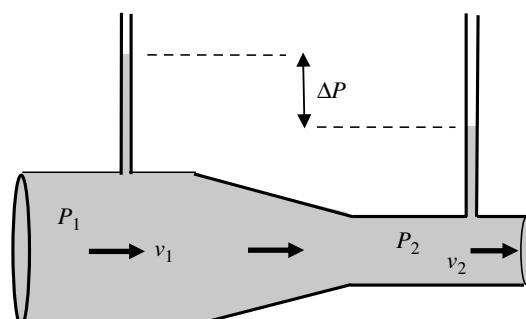


Figure 9.12 Fluid flowing through a constriction. A pressure drop ΔP will occur across the constriction, with $P_1 > P_2$ because $v_1 < v_2$.

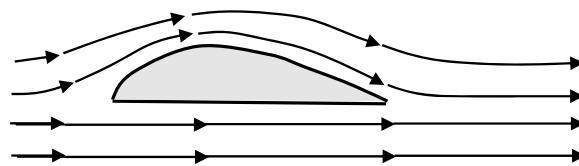


Figure 9.13 The fluid velocity immediately above this object is greater than that below it. According to Bernoulli's principle the fluid pressure below the object will be greater than that above it, and will produce a lift force.

A consequence of Bernoulli's principle is demonstrated in Figure 9.13, which shows an aerofoil-shaped object in flowing fluid. The fluid flows more rapidly over the top surface than over the lower surface. Faster fluid flow implies a lower pressure, so that the pressure will be greater on the bottom surface of the aerofoil and produce an upward lift force. This is the principle used in the design of aircraft wings and propeller blades, for example. An inverted version of the shape shown in Figure 9.13 will result in a downward force, which is an effect used in racing car spoilers.

Example 9.2

Water flows through a horizontal channel of geometrical profile shown in Figure 9.12. The fluid velocity v_1 before reaching the constriction is 0.1 ms^{-1} , and the fluid exit velocity v_2 is 15 ms^{-1} . The exit port is open to the atmosphere. Neglecting frictional losses, calculate the pressure P_1 in the main channel. (The density of water is 1000 kg m^{-3} , and atmospheric pressure is 100 kPa .)

Solution:

Because frictional losses can be ignored we will employ Bernoulli's equation (9.16). The channel is horizontal and so we will ignore potential energy differences arising from changes of fluid height. Equation 9.15 therefore takes the form:

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

We have (in mks units) $\rho = 1000 \text{ kg m}^{-3}$, $v_1 = 0.1 \text{ ms}^{-1}$, $v_2 = 15 \text{ ms}^{-1}$, $P_2 = 100 \text{ kPa} = 10^5 \text{ N m}^{-2}$.

$$\begin{aligned} P_1 &= P_2 + \frac{1}{2}\rho(v_2^2 - v_1^2) \\ &= 10^5 + 500(225 - 10^{-2}) = 2.125 \times 10^5 \text{ N m}^{-2} (\sim 2.1 \text{ atm}). \end{aligned}$$

We can also calculate the value for ΔP in Figure 9.12:

$$\Delta P = P_1 - P_2 = 1.125 \times 10^5 \text{ N m}^{-2} (\sim 1.1 \text{ atm}).$$

Example 9.3

Measurements of the flow of air around the object shown in Figure 9.13 reveal that the upper and lower flow velocities are 0.3 m s^{-1} , and 0.25 m s^{-1} , respectively. Calculate the pressure difference ΔP between the bottom and top surface.

Solution:

From Equation (9.15)

$$\begin{aligned}\Delta P &= P_1 - P_2 = \frac{1}{2}\rho(v_2^2 - v_1^2) \\ &= 500(0.09 - 0.0625) = 13.75 \text{ N m}^{-2}.\end{aligned}$$

The upward force F acting on this object will be $A.\Delta P$, where A is the lower surface area of the object.

9.5.3 Poiseuille's Law (Flow Resistance)

Bernoulli's principle assumes the fluid flow is not influenced by viscous forces. In fact, for the case of smooth, turbulence free, fluid flow the viscous shearing forces shown in Figure 9.1 will determine the fluid velocity profile across a channel. There will be zero fluid slip at the surfaces of the channel walls and the flow velocity will increase towards the centre line of the channel. The consequence of this is that in order to pump a viscous fluid along a channel a pressure difference ΔP must be applied between its inlet and outlet, irrespective of any changes of the channel diameter. In the 1840s Poiseuille experimentally and then theoretically derived the following relationship for fluid flow in pipes of circular-cross section:

$$\Delta P = \frac{8\eta LQ}{\pi r^4}, \quad (9.17)$$

where L is the length of the pipe, r its internal radius, and η is the dynamic viscosity of the fluid. This is also known as the Hagen-Poiseuille relationship in recognition of the contributions made by Hagen.

The flow resistance R of a channel is defined from the relationship

$$Q = \Delta P/R_f, \quad (9.18)$$

where Q is the volumetric flow rate. From Equations (9.17) and (9.18) the flow resistance of a channel of circular cross-section is thus given as

$$R_f = \frac{8\eta L}{\pi r^4}. \quad (9.19)$$

In practice, fluidic channels of either a rectangular or semicircular cross-section are easier to fabricate than those of circular cross-section (e.g. by placing a flat plate on top of a rectangular or rounded trench). The fluidic resistance of a rectangular channel with a high aspect ratio (i.e. width $w \gg$ height h) can be calculated using the formula

$$R_f = \frac{12\eta L}{wh^3}. \quad (9.20)$$

For a channel of semicircular cross-section defined by a radius of curvature r

$$R_f = \frac{64\eta L}{3r^4}. \quad (9.21)$$

Thus, for any specified channel geometry, the flow resistance is directly proportional to the viscosity of the fluid. From Table 9.1 we can see that the viscosity values for gases are considerably smaller than those for liquids. Bernoulli's approximation that fluid flow is not influenced by viscous forces may thus be adequate for the flow of gases, but should not be adopted when considering the flow of liquids. For example, the flow resistance of water will be about 50-times greater than for the flow of a gas along the same channel, and this difference increases to a factor of $\sim 10^5$ for the flow of a highly viscous fluid such as glycerine.

Example 9.4

A circular channel of diameter 20 μm and length 1 cm is designed to simulate a blood capillary for the purpose of biomedical research. A pump is to be used to flow blood through this channel at a flow rate of $10^{-2} \mu\text{L s}^{-1}$ into an analysis reservoir open to atmospheric pressure (100 kPa).

- (a) What pressure must this pump produce to achieve this flow rate?
- (b) Calculate the pressure required if the channel diameter is reduced to 10 μm .

Solutions:

- (a) From Equation (9.19) and the viscosity given for blood in Table 9.1:

$$R_f = (8 \times 3.5 \times 10^{-3} \times 10^{-2}) / (\pi \times (10^{-5})^4) = 8.9 \times 10^{15} \text{ Pa m}^{-3} \text{ s.}$$

Pressure ΔP drop along channel to give $Q = 10^{-2} \mu\text{L s}^{-1}$ ($10^{-11} \text{ m}^3 \text{ s}^{-1}$):

$$\Delta P = QR_f = (10^{-11} \text{ m}^3 \text{ s}^{-1})(8.9 \times 10^{15} \text{ Pa m}^{-3} \text{ s}) = 89 \text{ kPa.}$$

Total pressure output required of pump = $\Delta P + \text{atmospheric pressure} = 189 \text{ kPa.}$

- (b) Channel diameter halved from 20 to 10 μm

$$R_f \propto r^{-4} \therefore \text{new value for } R_f = (8.9 \times 10^{15}) / (1/2)^4 = 1.4 \times 10^{17} \text{ Pa m}^{-3} \text{ s.}$$

$$\Delta P = (10^{-11} \text{ m}^3 \text{ s}^{-1})(1.4 \times 10^{17} \text{ Pa m}^{-3} \text{ s}) = 1424 \text{ kPa.}$$

Total pressure output required of pump = $\Delta P + \text{atmospheric pressure} = 1524 \text{ kPa.}$

9.5.4 Laminar Flow

Figure 9.6 depicts a fluid moving in laminas with successively higher velocity. The flow velocity is zero in the vicinity of a stationary wall and increases away from the stationary wall. The fluid flow is a function of the x-coordinate and not of the y- and z-directions. This

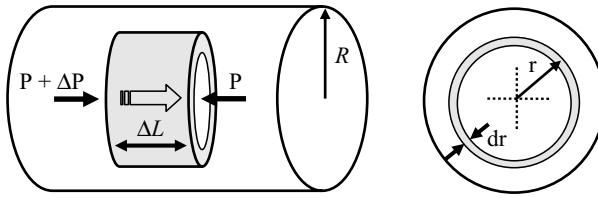


Figure 9.14 Laminar flow in a cylindrical pipe can be depicted as a series of concentric ‘stream tubes’ of length ΔL whose velocities increase as a function of the distance ($R-r$) from the pipe wall towards the centre axis of the pipe.

is termed as laminar flow. As depicted in Figure 9.14 we can envisage laminar flow in a pipe of circular cross-section to take the form of concentric, thin-walled, tubes of fluid whose velocities increase from zero at the pipe wall to a maximum at the centre line of the pipe. The flow is directed along the pipe’s axis and there are no pressure gradients across the pipe diameter. A shear stress τ exists between each tube and increases by $d\tau$ for each tube. A pressure drop between the ends of the fluid tube is required to overcome the shear stress. It is normally assumed that the pressure declines uniformly with distance down the fluid stream, so the pressure gradient $\Delta P/\Delta L$ is assumed to be constant.

Consider the elemental fluid tube shown in Figure 9.14, of length ΔL , radius r and thickness dr . If τ is the shear stress per unit area acting on the surface of this tube, the shear force F_s is given by

$$F_s = 2\pi r \Delta L \tau$$

From Equation (9.14) $\tau = \eta dv/dx = -\eta dv/dr$ ($x = R - r$)

to give

$$F_s = -2\pi r \Delta L \eta dv/dr.$$

At equilibrium this shear force will balance the force acting on the ends of the fluid tube as a result of the pressure difference ΔP :

$$\Delta P \pi r^2 = -2\pi r \Delta L \eta dv/dr$$

to give

$$dv = -\frac{\Delta P}{2\eta \Delta L} r dr.$$

The velocity v of a fluid tube at any radius r is found by integrating between the limits $v=0$ ($r=R$) and $v=u$ for $r=r$:

$$\int_0^v dv = -\frac{\Delta P}{2\eta \Delta L} \int_R^r r dr$$

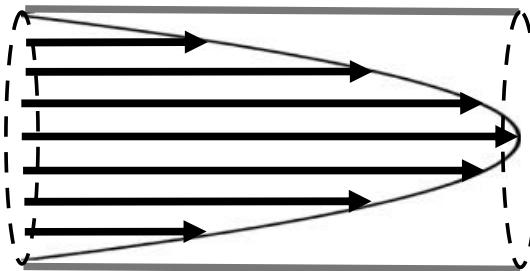


Figure 9.15 Laminar flow exhibits a parabolic fluid velocity profile, as described by Equation (9.22). The fluid velocity is zero at the channel wall and reaches a maximum at the centre line of the channel.

from which

$$v(r) = -\frac{\Delta P}{2\eta\Delta L}(r^2 - R^2) = \frac{\Delta P}{2\eta\Delta L}(R^2 - r^2). \quad (9.22)$$

The fluid velocity profile across the pipe is therefore parabolic, as shown in Figure 9.15, with zero velocity at the pipe walls and a maximum velocity along the central axis (at $r = 0$). The maximum velocity is given as

$$\bar{v} = \frac{\Delta p R^2}{4\eta\Delta L}.$$

The mean velocity $\langle v \rangle$ is the averaged velocity in the cross-section

$$\langle v \rangle = \frac{1}{\pi R^2} \int_0^R v(r) 2\pi r dr = \frac{\Delta p R^2}{8\eta\Delta L},$$

which corresponds to half the maximum value. The volumetric flow rate Q is given by the product of the mean velocity and the cross-sectional area:

$$Q = \frac{\Delta p R^2 \pi R^2}{8\eta\Delta L} = \frac{\pi R^4 \Delta p}{8\eta\Delta L},$$

corresponding to the Hagen-Poiseuille relationship of Equation (9.17).

We should note that the Hagen-Poiseuille relationship is derived assuming that the walls of the pipe or channel are perfectly smooth, so that the fluid flow has a unique axial component and no transverse components. If the walls are sufficiently rough to induce 3-dimensional components of fluid flow near the wall surfaces the pressure drop will tend to be greater than that predicted by Equation (9.17) and the fluid flow resistance will also be larger.

9.5.5 Application of Kirchhoff's Laws (Electrical Analogue of Fluid Flow)

Equation 9.18 can be written in the form

$$\Delta P = Q R_f.$$

This takes the same form as Ohm's law that relates the current I generated along an electrical conductor of resistance R_e as a result of the application of a voltage difference ΔV between the two ends of the conductor

$$\Delta V = I R_e.$$

This implies that we can employ Kirchhoff's rules, as applied to the analysis of electrical networks, to analyse fluid flow in fluidic networks.

Example 9.5

A T-junction for the micro-mixing of fluids is shown in Figure 9.16. Fluid flow J_3 exits into a reservoir open to atmospheric pressure (100 kPa).

- Derive an equation for calculating the fluid flow J_3 in terms of the channel fluidic resistances R_1 , R_2 and R_3 , and the pressures applied by syringe pumps P_1 and P_2 .
- Use this equation to calculate the fluid flow J_3 that would exit into a chamber at atmospheric pressure for $R_1 = R_2 = 10^{14} \text{ Pa m}^{-3} \text{ s}$; $R_3 = 2 \times 10^{15} \text{ Pa m}^{-3} \text{ s}$, and where pumps P_1 and P_2 exert pressures of 200 kPa and 150 kPa, respectively.

Solutions:

- Apply Kirchhoff's laws to the electrical analogue of the fluidic T-network shown in Figure 9.17:

1. Current Law (algebraic sum of currents at a junction is zero)

$$J_3 = J_1 + J_2 \quad (\text{i})$$

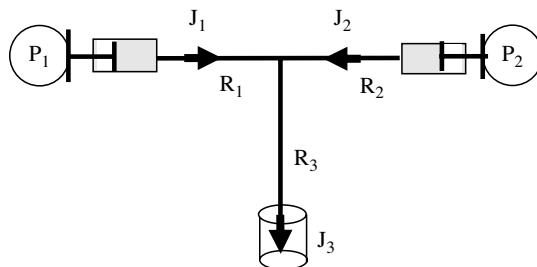


Figure 9.16 A microfluidic T-junction for mixing fluids pumped from two pressurised reservoirs (P_1 and P_2) through channels of fluidic resistances R_1 , R_2 and R_3 . Fluid flow J_3 is collected into a tube open to atmospheric pressure.

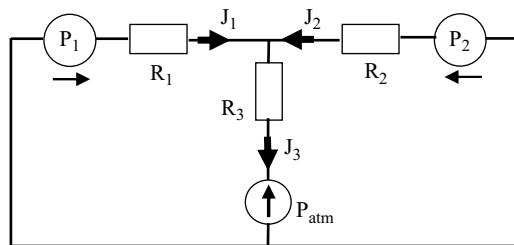


Figure 9.17 The electrical circuit analogue for the microfluidic T-junction shown in Figure 9.16.

Voltage Law (Algebraic sum of voltage drops around a closed circuit is zero)

$$\Delta P_1 = J_1 R_1 + J_3 R_3 \quad (\Delta P_1 = P_1 - P_{\text{atm}})$$

$$\text{to give } J_1 = (\Delta P_1 - R_3 J_3) / R_1 \quad (\text{ii})$$

$$\Delta P_2 = J_2 R_2 + J_3 R_3 \quad (\Delta P_2 = P_2 - P_{\text{atm}})$$

$$\text{to give } J_2 = (\Delta P_2 - R_3 J_3) / R_2 \quad (\text{iii})$$

Substitute J_1 obtained from (ii) and J_2 obtained from (iii) into (i):

$$J_3 = (\Delta P_1 - R_3 J_3) / R_1 + (\Delta P_2 - R_3 J_3) / R_2.$$

Rearranging:

$$J_3 = (\Delta P_1 R_2 + \Delta P_2 R_1) / (R_1 R_2 + R_2 R_3 + R_1 R_3) \quad (\text{iv})$$

In Equation (iv) we have defined $\Delta P_1 = P_1 - P_{\text{atm}}$ and $\Delta P_2 = P_2 - P_{\text{atm}}$

Therefore $\Delta P_1 = 200 - 100 = 100 \text{ kPa}$; $\Delta P_2 = 150 - 50 = 50 \text{ kPa}$.

Substituting the given values for R_1 , R_2 , and R_3 into (iv):

$$J_3 = (1.5 \times 10^{19} \text{ Pa}^2 \text{ m}^{-3} \text{ s}) / (4.1 \times 10^{29} \text{ Pa}^2 \text{ m}^{-6} \text{ s}^2) = 3.7 \times 10^{-11} \text{ m}^3 \text{ s}^{-1}$$

$$= 37 \text{ nL s}^{-1}.$$

This example demonstrates how the Kirchhoff's laws that are used to analyse current flows in electrical circuits can also be used to control and design for liquid flow in fluidic networks.

9.6 Navier-Stokes Equations

The Navier-Stokes Equations are widely used to describe the behaviour of fluids in terms of continuous functions of space and time. They encapsulate the three conservation laws of

mass, energy and momentum, and are considered in terms of flux rather than changes of their instantaneous values. In mathematical terms this is represented as partial derivatives of the dependent variables.

The calculation of fluid velocities and pressures at the macroscopic scale is based on the assumption that the fluid can be treated as a continuum. Apart from fluid velocity v and pressure P , for the most general situation that includes compressible and incompressible fluids we also require knowledge of the density ρ , viscosity η , specific heat C_p and temperature T of the fluid. Pressure and temperature characterise the energy state and number of molecules present in a given volume of fluid. If the pressure and temperature do not vary too greatly within this volume element, analytical functions can be derived that relate the density, viscosity and specific heat to the pressure and temperature. In a 3-dimensional system we are therefore left with five unknowns, namely P , T , v_x , v_y and v_z . These five unknowns are related by a system of equations that describe:

- the conservation of mass,
- the conservation of momentum,
- the conservation of energy.

The equations describing these three conservation laws are often referred to as the Navier-Stokes equations, but it is more correct to reserve this description to the equations that describe conservation of momentum. Conservation of energy usually concerns heat flow in fluid systems in which a temperature gradient is created by an energy source or sink associated with chemical reactions or heating and cooling devices. For most microfluidic flows in lab-on-chip devices the temperature is constant, in which case the conservation of energy equation is redundant. We will thus focus on the derivations of the conservation of mass and conservation of momentum equations.

9.6.1 Conservation of Mass Equation

To simplify the situation we will consider, as depicted in Figure 9.18, a 2-dimensional element (Δx , Δy) using the Cartesian coordinate system, with fluid velocities u and v in the x - and y -directions, respectively. We will then generalise to the 3-dimensional case.

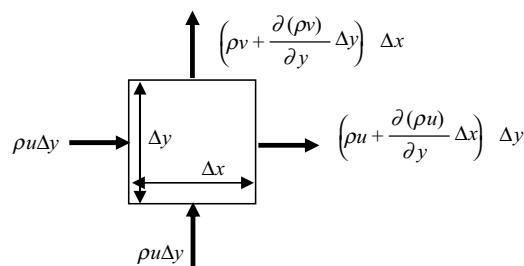


Figure 9.18 Conservation of fluid mass for a volume element $\Delta x \Delta y$.

For the system of fluid flow shown in Figure 9.18 the conservation of mass is given by

$$\frac{\partial(\rho\Delta x\Delta y)}{\partial t} = \rho u \Delta y + \rho v \Delta x - \left[\rho u + \frac{\partial(\rho u)\Delta x}{\partial x} \right] \Delta y - \left[\rho v + \frac{\partial(\rho v)\Delta y}{\partial y} \right] \Delta x.$$

Dividing by $\Delta x \Delta y$ we obtain

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} = 0,$$

which can be written as

$$\frac{\partial \rho}{\partial t} + \frac{u \partial \rho}{\partial x} + \frac{\partial \rho}{\partial y} + \rho \left[\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right] = 0. \quad (9.23)$$

Defining the operator D/Dt in 3-dimensional Cartesian coordinates as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z},$$

we can write Equation (9.23) in the vector form

$$\frac{D\rho}{Dt} + \rho \nabla \cdot \vec{V} = 0, \quad (9.24)$$

where \vec{V} is the velocity vector (u, v, w). We are concerned mainly with incompressible liquids, in which case terms such as $\partial \rho / \partial t$, $\partial \rho / \partial x$ and $D\rho / Dt$ are zero, and density ρ remains constant. Equations 9.23 and 9.24 thus reduce, for the 3-dimensional case, to

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad (9.25a)$$

and

$$\nabla \cdot \vec{V} = 0. \quad (9.25b)$$

9.6.2 Conservation of Momentum Equation (Navier-Stokes Equation)

The change of momentum in a fluid element is given by the balance between the inlet and outlet fluid momentum, and the tangential and normal stresses acting on that element. These are considered separately in Figures 9.19 and 9.20 for the 2-dimensional case.

For Newtonian fluids the tangential stress τ and normal stress σ are given as

$$\tau_{xy} = \eta \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \quad (9.26a)$$

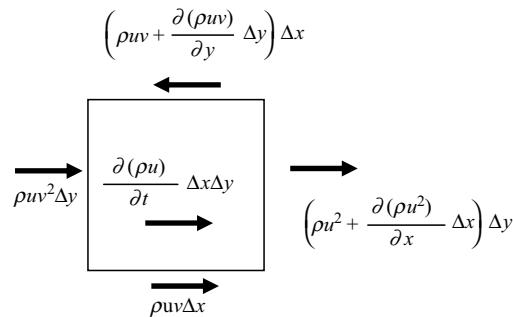


Figure 9.19 Inlet and outlet fluid momentum for a fluid element in the x -direction.

and

$$\sigma_x = P - 2\eta \frac{\partial u}{\partial x} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right). \quad (9.26b)$$

Summing the forces shown in Figure 9.20 in the x -direction, and using the mass conservation equations (9.25) we obtain

$$\rho \frac{Du}{Dt} = -\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + F_x.$$

Combining this result with Equation (9.26) gives the Navier-Stokes equation

$$\rho \left(\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} \right) = -\frac{\partial P}{\partial x} + \eta \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) + F_x. \quad (9.27a)$$

Extending this to 3-dimensions

$$\rho \left(\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} \right) = -\frac{\partial P}{\partial y} + \eta \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right) + F_y \quad (9.27b)$$

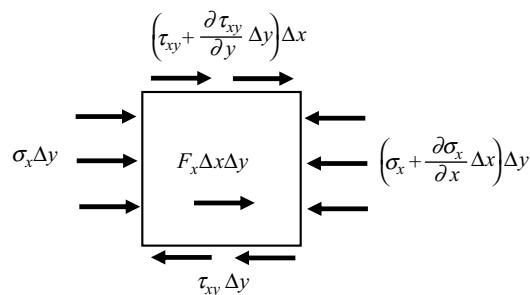


Figure 9.20 The normal and tangential stresses acting on the volume element shown in Figure 9.19.

$$\rho \left(\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} \right) = - \frac{\partial P}{\partial z} + \eta \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right) + F_z \quad (9.27c)$$

and in vectoral form

$$\rho \frac{D\vec{V}}{Dt} = -\nabla P + \eta \Delta \vec{V} + \vec{F}, \quad (9.27d)$$

where \vec{V} is the velocity vector (u, v, w) and \vec{F} is the force per unit volume acting on the element ($\Delta x, \Delta y, \Delta z$).

9.6.3 Conservation of Energy Equation

To derive this equation we identify either a source or sink of heat S_H , and specify the specific heat C_p and heat conductivity k of the liquid. The specific heat is defined as the amount of heat Q per unit mass required to raise the temperature of a substance by one degree Celsius,

$$Q = C_p m \Delta T.$$

The thermal conductivity of a substance is defined in terms of the quantity of heat Q conducted per unit time Δt down a unit temperature gradient ΔT in a direction normal to a surface of unit area ΔA . The heat conduction must arise only from the temperature gradient, and not from a secondary heat source or chemical reaction, for example:

$$k = Q \Delta T / (\Delta t \Delta A) = -Q / (\partial T \partial n)$$

The specific heat of water is $4.186 \text{ J gm}^{-1} \text{ K}^{-1}$, and its thermal conductivity is ~ 0.6 watts $\text{m}^{-1} \text{ K}^{-1}$.

In 3-dimensional Cartesian coordinates the conservation of energy equation is

$$\rho C_p \left(\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} + w \frac{\partial T}{\partial z} \right) = \frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left(k \frac{\partial T}{\partial y} \right) + \frac{\partial}{\partial z} \left(k \frac{\partial T}{\partial z} \right) + S_H. \quad (9.28)$$

9.7 Continuum versus Molecular Model

In Chapter 10 the *Knudsen Number* Kn (a dimensionless parameter that compares the characteristic dimensions of a fluidic system to the molecular spacing) will be shown to be an important indicator of whether continuum or molecular physics should be used to describe fluid properties. For dimensions larger than one micron or so, we have $\text{Kn} < 0.1$ and the continuum model can be used (although the concept of zero fluid slip at a surface might require modification). At this scale we can assume that the fluidic domain contains an infinite number of molecules, and that it is homogeneous at all scales. As depicted in Figures 9.21 and 9.22, the fluid can be divided up or decomposed into an infinite number of identical

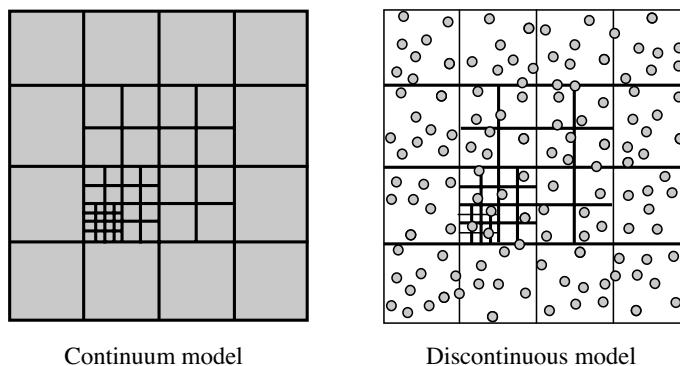


Figure 9.21 If the characteristic dimensions of a fluidic system are large compared to the molecular spacing, we can treat it as a continuum in which its properties are continuous and infinitely divisible. At nanometric dimensions the fluid is discontinuous – some elements of it will contain molecular mass and energy and some will not.

domains and its physical properties are continuous functions of space and time. At nanometric dimensions we are obliged to consider the fluid in terms of a finite number of molecules. When divided up, even into a finite number of elements, some elements will contain molecular mass and energy and some will not. A nanometric domain has spatially discontinuous properties.

We will now summarise the methods employed to simulate the physics of relevance to the two extreme cases of the continuum and molecular scales. This will assist an understanding of how to model the *mesoscale* where $0.1 < \text{Kn} < 10$.

9.7.1 Solving Fluid Conservation Equations

Two widely used computational fluid dynamic methods for simulating fluid properties are outlined in Figure 9.23, and are known as the *Finite Difference Method* (FDM) and the *Finite Element Method* (FEM). With FDM, instead of derivatives being computed over infinitesimal elements, increments of finite width are used as an approximation. The partial

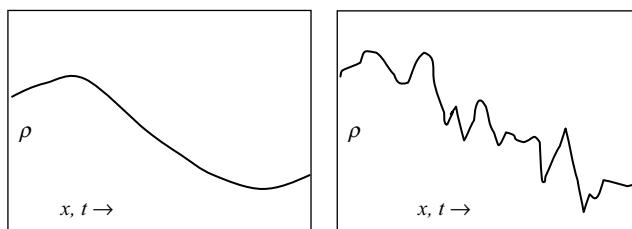


Figure 9.22 The physical properties (e.g. density, viscosity, temperature) of a continuum are continuous functions of space and time. A nanometric domain will exhibit statistical variations in its physical properties arising from a finite number of molecules in the domain.

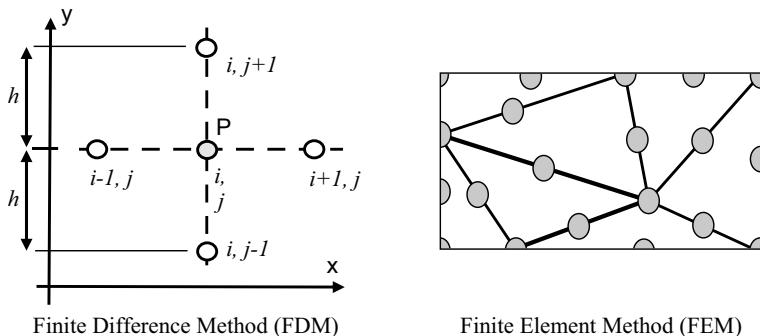


Figure 9.23 Two common computational methods for simulating fluid properties are the Finite Difference Method (FDM) and the Finite Element Method (FEM). With FDM simple algebraic equations yield values of fluid transport variables at discrete points in the flow field. With FEM the flow domain is divided into a finite number of cells, and the governing Navier-Stokes equations are evaluated at nodes placed at the corners or sides of these elements.

differential (Navier-Stokes) equations can thus be replaced with simple algebraic equations, and by solving (either iteratively or by matrix inversion) yield values of fluid transport variables at discrete points in the flow field.

Examples of such equations, with reference to Figure 9.23, are:

Calculations at Point P:

$$\text{Forward Difference: } \left(\frac{\partial p}{\partial y} \right)_{i,j} = \frac{p_{i,j+1} - p_{i,j}}{h}$$

$$\text{Backward Difference: } \left(\frac{\partial p}{\partial y} \right)_{i,j} = \frac{p_{i,j} - p_{i,j-1}}{h}$$

$$\text{Central Difference: } \left(\frac{\partial p}{\partial y} \right)_{i,j} = \frac{p_{i,j+1} - p_{i,j-1}}{2h} \quad \left(\frac{\partial p}{\partial y} \right)_{i,j} = \frac{p_{i,j+1} - p_{i,j-1}}{2h}$$

FDM can only be used to solve fluid flows (or heat, electric current flows) having simple boundaries. More complicated situations can be solved using FEM which divides the fluid continuum into a finite number of cells or elements. Nodes placed at the corners or sides of these elements are used to fractionate and evaluate the governing equations in an integral form using weighting functions. As shown in Figure 9.24 for a heat flow simulation there is a trade off between the number of nodes chosen (hence calculations) and the resulting accuracy.

The size of the cells or elements can vary so as to focus the most computational effort on those regions of the flow field, or any other dimensional field, where rapid changes of the physical quantity are expected. An example of this is shown in Figure 9.25 for the modelling of the electric field generated at a pin electrode. The size of the elements is made smaller around the tip of the electrode.

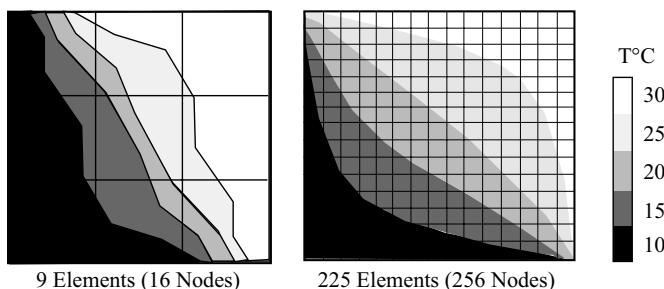


Figure 9.24 As shown in this example of a heat flow simulation using FEM, there is a trade off between the number of elements chosen (hence number of computations required) and the accuracy achieved. The heat source and sink are situated in the top right and bottom left corners, respectively.

An extension of FEM is the *Finite Volume Method* (FVM) which divides up the flow domain into elemental control volumes surrounding a node. The flow parameters are treated as fluxes between control volumes, and continuity is maintained in each element. Because of the continuum approximation these volume elements must be uniform throughout and infinitely divisible.

9.7.2 Molecular Simulations

At the nanometric scale the characteristic length of a fluid flow approaches that of the diameters of individual molecules, and so we must account for motions of individual molecules. The basic mechanism assumed in most molecular simulations is that the force acting on any molecule is determined by the movements of its neighbours, and interacts with them according to Newton's three laws of motion.

Molecular simulations rely on representative particles interacting with each other, where each particle has individual properties that determine its next position and momentum

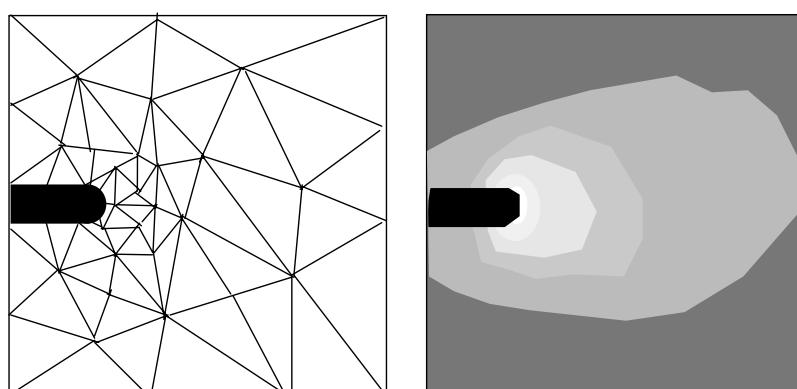


Figure 9.25 In this FEM model of the electric field generated by a pin electrode the size of the elements is reduced in the region where the field is expected to change the most.

after each interaction. There are two approaches – namely deterministic (where the outcome can theoretically be worked out) or stochastic (having elements of unpredictability and chance).

9.7.2.1 Deterministic Simulations

These fall into two categories, according to whether the molecules are represented as hard or soft spheres. In the *Hard Sphere Model* collisions between molecules are modelled as binary interactions, occurring instantaneously, where they exchange linear momentum. No long-range interactions are assumed. The simulation advances one step at a time, to the next event or collision, and is based on the assumption that all spheres have an initial position and velocity, and that between collisions they travel at a constant speed and direction, such that their positions at any time can be calculated. This is depicted in Figure 9.26.

In Figure 9.26 the position of a sphere at any time is given by

$$r_i = r_i(t_0) + (t_1 - t_0)v_i(t_0)$$

where r_i , and v_i are the position (centre) and velocity of sphere i , t_0 is the start time and t_1 is the new time.

The time to collision depends on the relative velocity v_{12} of spheres 1 and 2. If the condition $v_{12}r_{12} < 0$ is met, then the two spheres are heading towards each other. Collision occurs when:

$$r_1(t) - r_2(t) = \sigma$$

and the value of θ shown in Figure 9.26 determines whether or not a collision of two spheres will occur.

In the *Soft Sphere Model* the molecules are considered to interact by exerting van der Waals forces on each other. As described in Chapter 1 these forces are composed of long-range attractive interactions and short-range but strongly repulsive ones. These interactions occur continually, with each molecule having a ‘zone’ (of radius ~ 3 molecular diameters) within which other molecules are influenced, as depicted in Figure 9.27. This differs from the hard sphere model, where spheres interact only when contact is made. The resultant force

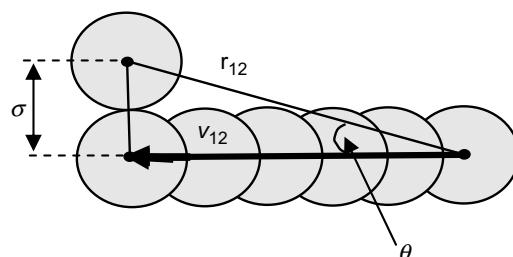


Figure 9.26 Molecular collisions using the *Hard Sphere Model* are treated as instantaneous binary interactions. Between collisions the molecules are assumed to travel in straight paths with a constant relative velocity v_{12} , and long-range interactions are absent. See main text for further details.

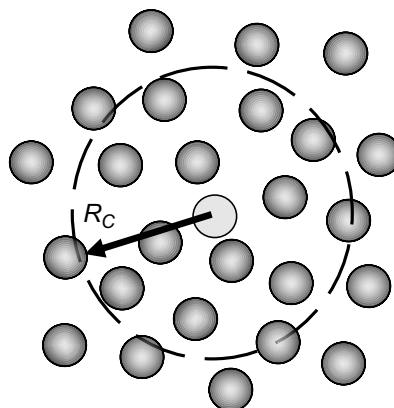


Figure 9.27 In the *Soft Sphere Model* of molecular interactions each molecule interacts continually with others lying within a critical zone of radius R_c . The interactions are often modelled as short- and long-range van der Waals forces.

arising from the repulsive and attractive force is often modelled using the Lennard-Jones 6–12 potential described in Chapter 1 and Figure 1.2.

9.7.2.2 Stochastic Simulations

These simulations incorporate an element of randomness into the model. This is often achieved using Monte Carlo simulations involving repeated random sampling of molecular perturbations. A simple example of a stochastic procedure is given in Figure 9.28, where the value of π is calculated by ‘probing’ a square domain containing an arc. The area inside the arc is estimated from the number of points inside its constraints (squares) and the area of

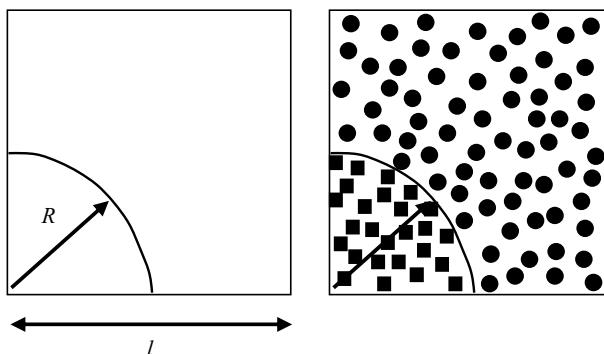


Figure 9.28 In this method for estimating the value of π , a square domain containing an arc is probed using randomly distributed test points. As the number of random sample points increases, the total number of points (squares + circles) divided by the number of points (squares) inside the arc approaches the value for π .

the square is given by the total number of points (squares + circles). An estimation of π is obtained using the following logic:

$$\frac{\text{Area of Arc}}{\text{Area of Square}} = \frac{\text{No. of squares}}{\text{No. of circles} + \text{squares}}$$

$$\pi = \frac{\text{No. of squares}}{\text{No. of circles} + \text{squares}} \cdot \frac{4l^2}{R^2}$$

As an initial step in molecular dynamics modelling, the overall energy of a system of molecules is calculated. One molecule in this system is chosen at random, with a defined probability of being selected. This molecule is assigned a small perturbation, for example either of its position or speed, and the new system energy is calculated. If the new system energy is smaller than the initial system energy, this added perturbation is accepted. If the new system energy increases, the perturbation is accepted according to a probability law (e.g. one based on Boltzmann statistics). Rejected perturbations are ignored. This procedure is repeated until the system reaches ‘equilibrium’.

9.7.3 Mesoscale Physics

The mesoscale region, defined as a scale between the micro- and molecular-scale, covers the change in physics between the continuum approximation and discontinuous molecular models. The lower limit of the mesoscale can be taken to be around 100 atomic or molecular diameters. The upper limit, corresponding to where the continuum approximating laws are violated, is not so well defined. For example, a rarefied gas might invalidate the use of continuum physics up to scales of $\sim 10 \mu\text{m}$, whereas for a dense liquid the continuum laws could be valid down to scales below $1 \mu\text{m}$. Travis *et al.* [1] have modelled the velocity profile and heat flux profile of an atomic liquid in a narrow channel, using molecular dynamics and Navier-Stokes equations. For a channel width of 5.1 molecular diameters the two simulations of the velocity profiles differed significantly. The heat flux profile did not agree with that predicted by Navier-Stokes hydrodynamics, but exhibited significant oscillations located about one molecular diameter from the walls. However, classical Navier-Stokes behaviour was approached for a channel width greater than 10 molecular diameters.

As an example of experimental investigations at the boundary between the micro- and mesoscale, Pfahler *et al.* [2] constructed three channels of rectangular cross-section ranging in area from 7200 to 80 square microns. In the relatively large flow channels the experimental observations were in rough agreement with predictions from the Navier-Stokes equations but significant deviations were found for the smallest of the channels. Mala and Li [3] studied water flow through microtubes with diameters ranging from 50 to $254 \mu\text{m}$. Results in rough agreement with conventional continuum theory were obtained for the large diameters, but not for the smaller diameters.

Either a top-down or a bottom-up approach can be adopted to model the mesoscale.

9.7.3.1 Top-Down Approach

Gases are well understood in terms of kinetic theories and are thus better suited than fluids to describe the approach from continuum physics to molecular dynamics. We will progress down in scale from where the fluid can be considered to be continuous, infinitely divisible and in thermodynamic equilibrium. In the next chapter we will see that the first stage of the

breakdown of the continuum approximation for gases occurs at a Knudsen number greater than 0.001, where areas of high gradient such as boundaries, cannot maintain the continuous distribution of macroscopic properties. This is a result of the deviation from thermodynamic equilibrium, where there are insufficient collisions in the system for the energy to propagate smoothly in areas near boundaries. The low number of molecular collisions with the boundary means that the velocity and temperature of the solid and fluid are no longer the same at the interface. This causes a violation of the no-slip condition at a moving surface, as well as the no-jump-in temperature condition, assumed in continuum physics. To account for this initial violation we can describe the slip and no-slip conditions by relating the difference in velocity between the wall and fluid ($u_{\text{fluid}} - u_{\text{wall}}$) to the strain rate at the wall ($\partial u / \partial y$)_{wall}:

$$u_{\text{fluid}} - u_{\text{wall}} = L_s \left(\frac{\partial u}{\partial y} \right)_{\text{wall}},$$

with L_s being the slip length. This slip condition can be included in the continuum approximation, using either a simulated or experimentally observed value for L_s . For normal continuum conditions L_s is so small that the fluid and wall move at the same speed (no-slip condition), but as the Knudsen number of the system increases above 0.001 the slip effects become more pronounced. The amount of slip that is allowed depends on the roughness of the surface over which the fluid is flowing and the interaction rate between the fluid and solid molecules. We can expect this model to fail as the surface roughness of the wall approaches the mean free path of the fluid.

The transition between the continuum and molecular regions for liquids goes through the same stages as for gases, but there is no parameter to act as a guide throughout the transition. The Knudsen number cannot be defined, as there is no concept of mean free path for liquid flows – the molecules are in a constant state of collision and move over much shorter distances (comparable to the molecular diameter). As shown in Figure 9.29, computational modelling can be used in which finite element meshes are systematically reduced in scale to ‘handshake’ a molecular dynamic region, either using Monte Carlo simulations [4], or avoiding such simulations [5]. Nie *et al.* [6] have developed a hybrid multiscale method for simulating micro- and nanoscale fluid flows. Continuum Navier-Stokes equations are used in

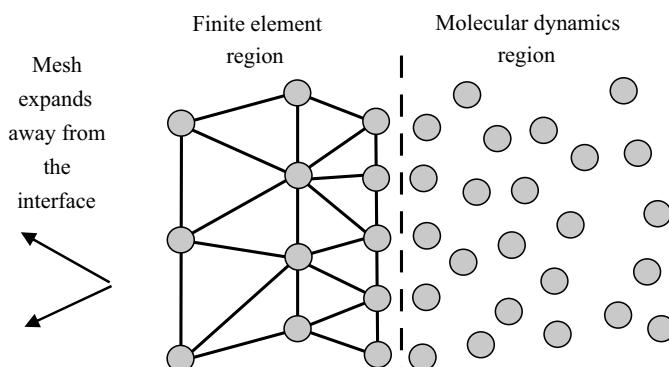


Figure 9.29 In the top-down approach to model the interface between the micro- and mesoscale, a finite element mesh is scaled down to meet molecular sites.

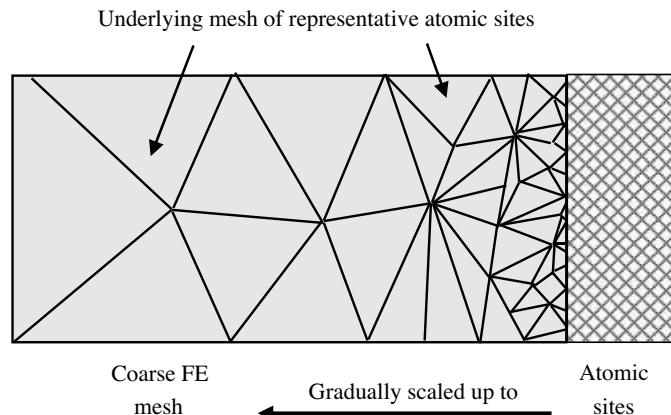


Figure 9.30 In this bottom-up approach to mesoscale simulation, a mesh of atomic sites at constant density is covered by a finite element mesh that gradually scales up from the atomistic to a quasi-continuum scale.

one flow region and molecular dynamics in another. The spatial coupling between continuum equations and molecular dynamics is achieved through constrained dynamics in an overlap region.

9.7.3.2 Bottom-Up Approach

A serial approach can be adopted, where a very small molecular simulation builds up to describe the physical relationships behind a large-scale fluid system. Such approaches can require supercomputers operating in parallel. An extreme example of this is the global simulation described by Clementi [7], which begins with the building of molecular liquids from nuclei and electrons using quantum mechanics, proceeding next to multibody interaction potentials, again by quantum mechanics, followed by the application of Monte Carlo and molecular dynamics to study the motions and collective properties of water molecules. This is then extended to the realm of fluid dynamics by considering a flow along a channel with or without obstacles, to finally consider the tidal movements in Buzzards Bay, New England!

However, in general the main issue faced with bottom-up methods is the scaling up of information from the molecular level and the removal of degrees of freedom from the system to minimise computational effort. One approach is to use the scheme depicted in Figure 9.30. The simulation involves two layers. One layer contains molecular information and covers the whole system. This is overlaid with an equivalent finite element mesh that gradually scales up, from nodes corresponding to atomic sites close to the region of interest, to elements containing many molecules. Because the large-scale elements contain many molecules, their energies local to each other have approximately similar values. In a method known as the *quasi-continuum technique* the approximation is made that the local neighbourhood of molecules can be represented by the value of just one molecule [8]. This can dramatically reduce the number of computations required in the simulation – but can only be applied to the coarse FE meshes, and breaks down when the effective mesh element contains only a few molecules.

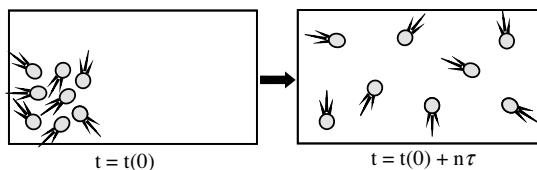


Figure 9.31 Randomising collisions, at a rate of $1/\tau \text{ s}^{-1}$, result in the even distribution of molecules through a process that is called diffusion.

9.8 Diffusion

From Equation (9.3) we know that a molecule in thermal equilibrium with a surrounding fluid of absolute temperature T has an average kinetic energy of $3kT/2$, and thus an average velocity $(3kT/m)^{1/2}$ associated with motion along each of the three axes in a 3-dimensional volume. Diffusion is the random migration of molecules or small particles from multiple collisions arising from the kinetic motion of neighbouring molecules. A schematic of this process is shown in Figure 9.31, where a cluster of gas molecules is shown occupying the corner of an otherwise empty container. As an oversimplification, only to demonstrate the process involved, we will assume that the time τ and average mean free path length between collisions remains constant. The rate of randomising collisions is thus $1/\tau$. After a sufficiently large number n of such collisions the molecules will become evenly distributed in the container (after time $n\tau$).

From their independent analyses of Brownian motion (the buffeting of macroscopic particles through collisions with fluid molecules) Einstein and Smoluchowski derived the following expression for the diffusion coefficient D

$$D = \frac{L_{mfp}^2}{2\tau}, \quad (9.29)$$

where L_{mfp} is the mean free path length given in Equation (9.2). An excellent discussion of the origins and validity of this so-called Einstein-Smoluchowski equation has been given by Isla [9]. From the worked Examples 10.3 and 10.5 of Chapter 10 we find that for nitrogen gas at room temperature and atmospheric pressure $L_{mfp} = 14.4 \times 10^{-10} \text{ m}$, and $\tau = 3.1 \times 10^{-10} \text{ s}$. From Equation (9.29) we obtain an estimate for D of $3.3 \times 10^{-9} \text{ m}^2 \text{ s}^{-1}$. Einstein also demonstrated that for macroscopic particles exhibiting Brownian motion in a fluid, the particle's diffusion coefficient can be expressed as

$$D = \frac{kT}{6\pi\eta a}, \quad (9.30)$$

where ' a ' is the particle's effective hydrodynamic radius and η is the fluid viscosity. This is called the Stokes-Einstein equation, and its origin and validity has also been discussed by Isla [9]. For particles suspended in water, the effective hydrodynamic radius is defined as the radius of a rigid uncharged sphere which exhibits the same hydrodynamic behaviour as the solvated molecule in solution. This should therefore include water of hydration which is too firmly bound to the particle's surface to participate in the viscous shearing process as it

Table 9.2 Approximate diffusion coefficients for some biologically relevant particles in water at 293 K. Values for the mean diffusion distance (diffusion layer thickness), defined by Equation (9.31), are given for time intervals of 1 ms and 10 s

Particle	Diffusion coefficient $\text{m}^2 \text{s}^{-1}$	Diffusion layer thickness (m)	
		10^{-3}s	10s
Small ions	2×10^{-9}	2×10^{-6}	2×10^{-4}
Sugar molecules	5×10^{-10}	1×10^{-6}	1×10^{-4}
Small proteins (e.g. lysozyme)	1×10^{-10}	4.5×10^{-7}	4.5×10^{-5}
50-base pair DNA	2.5×10^{-11}	2.2×10^{-7}	2.2×10^{-5}
Large proteins (e.g. collagen)	7×10^{-12}	1.2×10^{-7}	1.2×10^{-5}
Virus	4×10^{-12}	9×10^{-8}	9×10^{-6}
5000-base pair DNA	1×10^{-12}	4.5×10^{-8}	4.5×10^{-6}
Bacteria	2×10^{-13}	2×10^{-8}	2×10^{-6}

moves through the aqueous medium. Equation 9.30 was derived on the assumption that the solute molecule is large compared to the solvent. Nevertheless the equation has been experimentally confirmed for suspended particles with radii as small as 5 nm, and for large colloidal particles with suspension volume fractions up to 3%. It can also provide good approximations for the diffusion of molecular species in water. Thus sucrose ($a \approx 0.5 \text{ nm}$) can be estimated to have a diffusion coefficient in water ($\eta = 1 \times 10^{-3} \text{ Pa}$) at 298 K of $\sim 3.9 \times 10^{-10} \text{ m}^2 \text{s}^{-1}$, which can be favourably compared to the value of $5.2 \times 10^{-10} \text{ m}^2 \text{s}^{-1}$ cited in Table 10.4 of Chapter 10. Approximate values of diffusion coefficients for some biologically relevant particles are given in Table 9.2.

A description of particle diffusion can be made in terms of a one-dimensional random walk, often described in terms of the ‘drunken sailor’ problem outlined in Figures 9.32 and 9.33.

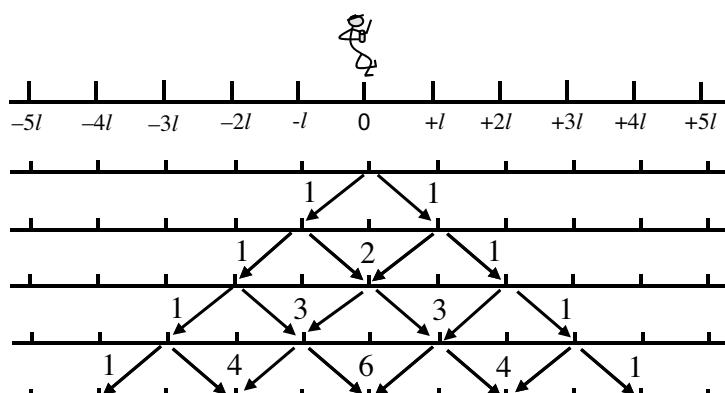


Figure 9.32 The probability distribution for a 1-dimensional random (drunken sailor) walk is given by the factorials of the binomial coefficients as displayed by the rows in Pascal’s triangle (Pascal’s pyramid for 3D).

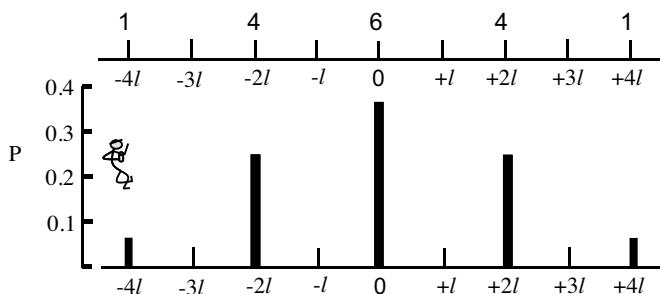


Figure 9.33 The probability distribution for a 1-dimensional random walk after time $t = 4\tau$. After a period of 4τ the probability of the sailor standing straight ahead of his original location is $6/16 = 0.375$.

At each new step forward the drunken sailor is equally likely to stagger one step to the left as he is to the right. We can use this analogy to describe the resulting random direction that a molecule follows after colliding with another molecule.

After a number of random steps the spatial distribution of particles along a one-dimensional axis takes the form of a probability distribution described by the factorials of the binomial coefficients. Applying Stirling's approximation for these factorials, then for a sufficiently large number of thermal collisions we can represent the probability distribution as a Gaussian or normal distribution. In one-dimension the probability $P(x)dx$ of finding a particle between x and $x + dx$ at time t is given by Isla [9] as:

$$P(x)dx = \frac{1}{(4\pi Dt)^{1/2}} e^{-x^2/4Dt} dx.$$

The mean displacement $\langle x^2 \rangle$ of the particle is thus given by

$$\langle x^2 \rangle = \frac{\int_0^\infty x^2 dP}{\int_0^\infty dP} = 2Dt.$$

An alternative way to derive this equation is to employ Equation (9.29) as follows:

$$\langle x^2 \rangle = nL_{mfp}^2 = \frac{t}{\tau} L_{mfp}^2 = 2Dt$$

to give

$$\langle x \rangle = \sqrt{2Dt}. \quad (9.31)$$

We can consider $\langle x \rangle$ as the mean diffusion length for a molecule interacting through collisions with neighbouring molecules. Values for this diffusion length are given in Table 9.2 for times of 1 ms and 10 s. We can see that small sugar molecules like glucose will diffuse a distance of around 1 μm after 1 ms, and 0.1 mm after 10 s. These can be significant distances in microfluidic systems. For particles of the size of bacteria, however, the corresponding diffusion lengths are much less at 20 nm and 2 μm , respectively.

Diffusion of molecules and particles tends to occur down their concentration gradient – also referred to as diffusion gradients (see Figure 9.34).

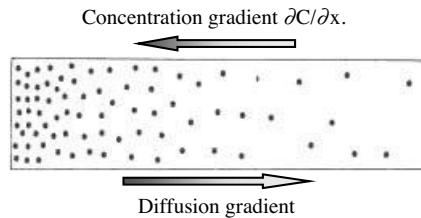


Figure 9.34 Molecules tend to diffuse down a concentration gradient (also termed as a diffusion gradient).

This diffusion process can be described by Fick's 1st equation of diffusion:

$$J_x = -D \frac{\partial C}{\partial x},$$

which states that the net flux J_x (moles $\text{m}^{-2} \text{s}^{-1}$) of diffusing molecules or particles is proportional to the concentration gradient and diffusion constant of the molecule/particle (the negative sign indicates that the molecules diffuse *down* the concentration gradient). Unless the concentration gradient is artificially maintained (e.g. with a continuous source and sink of the molecules or particles) the factor $\partial C/\partial x$ will change as a function of time. This leads to Fick's 2nd equation:

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}.$$

This equation can be used (with the appropriate boundary conditions) to determine how a nonuniform distribution of molecules or particles will redistribute itself as a function of time. Diffusion along a microfluidic channel is effectively a one-dimensional problem. In this case the solutions of Fick's 2nd equation are:

$$\frac{\partial C}{\partial x} = \frac{C_0}{(4\pi Dt)^{1/2}} e^{-x^2/4Dt} \quad \text{and} \quad \frac{\partial C}{\partial t} = -\frac{x}{2t} \frac{\partial C}{\partial x}. \quad (9.32)$$

An example of the application of Equations (9.32) is given in Figure 9.35, to show how a 0.1 nL aliquot of 1 M methanol ($D = 1.6 \times 10^{-5} \text{ cm}^2 \text{s}^{-1}$) diffuses after its injection into a channel, of height and width 100 μm , filled with stationary water.

As shown in Figure 9.35, within a period of 5 seconds the leading edges of the diffusing methanol reach a distance of $\sim 250 \mu\text{m}$ either side of the injection port, and this extends 500 μm after 10 seconds. For a fixed location along the channel, the concentration of diffusing methanol rises and then falls, rather like a tidal wave passing a buoy (see Figure 9.36).

In Chapter 10 we will learn that it is very difficult, if not impossible, to induce turbulent flow in channels of micron-scale dimensions. The flow is inevitably laminar, of the form shown in Figure 9.15. In Figure 9.37 a fluidic Y-junction (equivalent to the T-junction of Figure 9.16) is used to flow together two liquids into a third channel of diameter 100 μm . A practical question to ask is how long should the third channel be to achieve complete mixing of the two liquids, and to what extent is mixing influenced by the rate of fluid flow? In the absence of mechanical stirring, the only way for the merging liquid streams to mix is through

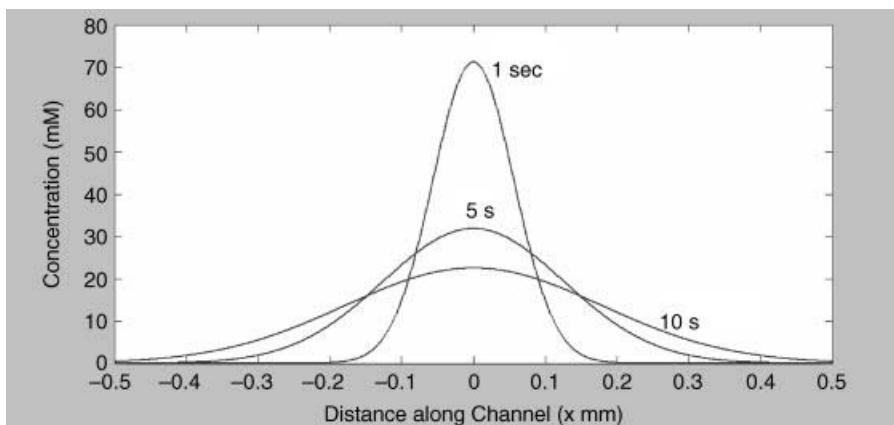


Figure 9.35 The temporal concentrations of methanol (1 M, 0.1 nL, aliquot injected at time $t = 0$ into location $x = 0$) as it diffuses along a channel of static water of width and height 100 μm .

the diffusion of their constituent molecules across the interface between the travelling liquids. The profile of this interface will broaden and dissipate with time along the channel rather like the concentration profiles shown in Figure 9.35. If a flow rate of 5 μL per minute is chosen, then as shown in Figure 9.37 no discernable mixing of the two fluid streams occurs after a distance of 2.5 mm along the third channel. A 10-fold reduction of the flow rate to 0.5 μL per minute does result in some mixing of the two fluid streams, but it is clear that the channel length will need to extend much further before complete mixing occurs. The long channel lengths required for the thorough mixing of laminar flow streams can be accommodated in lab-on-chip designs using the serpentine geometry shown in Figure 9.37.

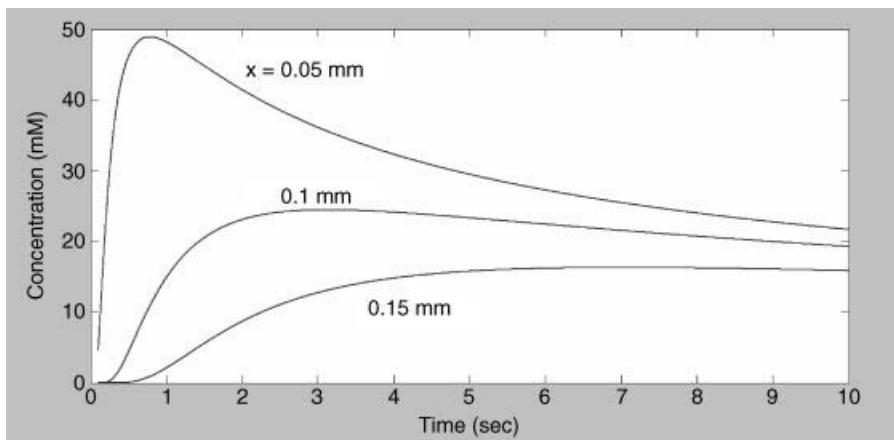


Figure 9.36 At any fixed location along the channel the concentration of diffusing methanol shown in Figure 9.35 appears as a passing wave.

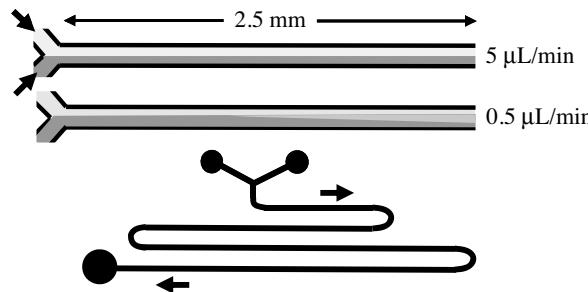


Figure 9.37 Modelling (top) of liquid streams flowing together via a Y-junction into a channel of radius 100 μm . Mixing of the fluids is evident for a fluid flow rate of 0.5 $\mu\text{L}/\text{min}$, but not at 5 $\mu\text{L}/\text{min}$. A serpentine geometry (bottom) is often used in lab-on-chip devices to accommodate the long channel lengths required for the mixing of laminar fluid streams.

9.9 Surface Tension

Surface tension is a significant and useful force in microfluidic devices. The origin of the surface force known as surface tension is shown in Figure 9.38, which depicts a free surface between liquid and air. A molecule in the fluid bulk experiences mutually attractive forces with neighbouring molecules. Van der Waals forces are usually the most dominant, and for aqueous solutions hydrogen-bond forces are also significant. A molecule at the surface is attracted by a reduced number of neighbours and so is in an energetically unfavourable state. The creation of a new surface is thus energetically costly, and fluids will act to minimise their surface area. This is the driving force for small volumes of fluid to assume a spherical shape, as for example trickles of water breaking up into spherical drops to minimise the total surface area.

If U is the total cohesive energy per molecule in the fluid bulk, then this is halved to a value of $U/2$ for a molecule located at a flat surface. The surface tension created per unit area of surface is directly related to this cohesive energy reduction. For a characteristic molecular dimension R , the effective molecular area is R^2 and the surface tension is $U/(2R^2)$. Surface tension is thus directly proportional to the intermolecular attraction and inversely

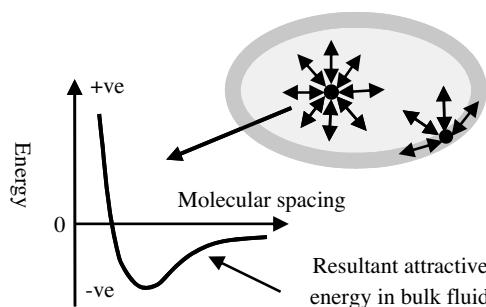


Figure 9.38 The molecular origin of surface tension is the fact molecules in bulk liquid experience mutual attractive forces. The net attractive force on a surface molecule is less, and so a surface molecule is in an unfavourable energy state.

Table 9.3 Values of surface tension T (liquid surface with air) at 20 °C

Liquid	T (mN m ⁻¹)	Liquid	T (mN m ⁻¹)
Ethanol	22.3	Glycerine	63.1
Soap solution	25.0	Water	72.8
Olive oil	32.0	Mercury	465

proportional to the molecular size. Surface tension values for some common fluid–air interfaces are given in Table 9.3. Of note in this table is the fact that water has a significantly larger surface tension than alcohol, soap solutions and oils, reflecting not only the relatively small size of the water molecule but also the cohesive energy supplied by hydrogen-bonds in water. Liquid metals (e.g. mercury) exhibit the highest surface tension values.

Surface tension T is defined as the ratio of the surface force F to the length d along which the force acts ($T = F/d$) and thus has units of force per unit length (equivalent to energy per unit area) and acts tangentially to the free surface. Because its action is confined to the interface it does not appear in the Navier-Stokes equations, which deal with pressure gradients within a fluid bulk. A total force γdl will act on a surface line element dl . If the surface line element is a closed loop, and the surface tension uniform, the net surface tension force acting on the loop is zero. If surface tension gradients are present, a net force on the surface element may result and distort it through an induced flow of surface liquid. Surface tension gradients can arise from the presence of a surfactant.

9.9.1 Surfactants

Chemical compounds that lower the surface tension of a liquid are known as surfactants. Detergents, soap, fatty acids and fatty alcohols are common examples. They can be used to stabilise mixtures of oil and water, for example, by reducing the surface tension at the interface between the oil and water molecules. Their molecular structures often consist of a hydrophilic head and a hydrophobic tail group, so that their location at a free liquid surface can be energetically favourable. Gradients in surfactant concentration result in surface tension gradients. A simple, but instructive demonstration of this can be observed by coating one end of a matchstick with soap or liquid detergent. When placed carefully on water the matchstick will be propelled across the surface away from the dissolving surfactant.

9.9.2 Soap Bubble

The pressure difference between the inside and outside of a soap bubble can be derived by representing the bubble as two hemispheres. This is depicted in Figure 9.39. The net pressure acting to push the top hemispheres away is balanced by the surface tension T acting around the circumference of the lower hemisphere (noting that there is an outer and inner surface to the bubble). If we ignore the film thickness compared to the bubble radius r , and the weight of fluid comprising the bubble film, then the net upward and downward forces are given by:

$$F_{up} = (P_i - P_o)\pi r^2; \quad F_{down} = 2T(2\pi r).$$

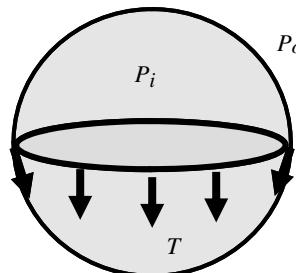


Figure 9.39 The pressure difference ($P_i - P_o$) between the inside and outside of a soap bubble can be derived by representing the bubble as two hemispheres. The net pressure acting on the upper hemisphere is balanced by the surface tension T acting around the circumference of the lower hemisphere.

Equating these two forces we obtain the result:

$$(P_i - P_o) = \frac{4T}{r}.$$

The $1/r$ relationship in this equation indicates that it is more difficult to inflate a small balloon than a larger one. This is of anatomical relevance. Oxygen exchange in the lungs takes place across the membranes of small balloon-like structures called alveoli. Their surface tension is lowered nearly 15-fold by a surfactant coating so that a very small pressure differential is sufficient to inflate them with air. Elastic recoil of the alveoli assists with their deflation (exhalation).

If instead of a soap bubble we consider an air bubble in a liquid (or a fluid droplet in air) there is just the one (outer) surface to take into account. Accordingly, the pressure difference is halved: $P_i - P_o = 2T/r$.

9.9.3 Contact Wetting Angle

A drop of liquid on a solid surface is shown in Figure 9.40. As shown in this figure, there are three interfaces, namely the solid–liquid, the liquid–air and the solid–air interface. A line on the solid surface (the xy -plane) defines the boundary separating these three interfacial areas. The contact angle θ is defined as the angle formed at this three phase boundary between the

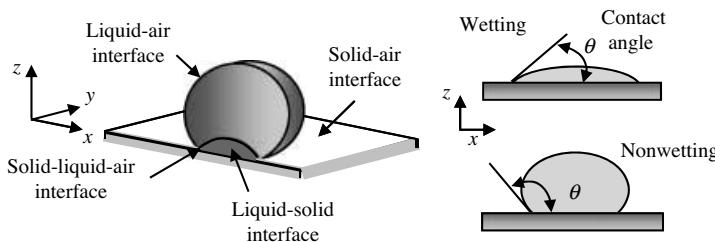


Figure 9.40 The contact angle θ of a drop of liquid on a solid surface is defined as the angle formed between the tangent to the liquid surface and the xy -plane at the boundary between the three interfaces (liquid–solid, liquid–air, solid–air). A summary of the factors influencing the contact angle in terms of the surface energy and wettability of the surface is given in Table 9.4.

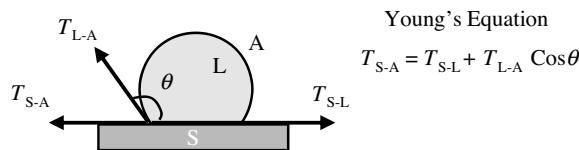


Figure 9.41 At equilibrium the horizontal components of the surface free energies balance, and this is expressed in the form of Young's equation.

tangent to the liquid surface and the xy-plane. A tension exists in each interface, and different values for these result in different liquids adopting different contact angles relative to different solid surfaces. An equilibrium situation exists when the horizontal components of the surface free energies balance. From Figure 9.41 the equilibrium condition is readily seen to be described by the following relationship, known as Young's equation:

$$T_{S-A} = T_{S-L} + T_{L-A} \cos\theta,$$

where T_{S-A} , T_{S-L} and T_{L-A} are the surface tensions at the solid–liquid, the liquid–air and the solid–air interfaces, respectively, and θ is the contact angle defined above.

A liquid with low surface tension (low surface energy) resting on a solid surface of higher surface energy will spread out on the surface forming a contact angle θ less than 90° . The liquid is said to wet the surface – if the liquid is water we say the surface is hydrophilic. If the surface energy of the liquid exceeds that of the solid, the liquid will form a bead and θ will have a value between 90° and 180° . In this case we have a nonwetting liquid relative to the surface, corresponding to a hydrophobic surface when considering aqueous liquids. These aspects are summarised in Table 9.4.

Surface wetting is an important aspect of printing. The surface energy of the ink when wet should be lower than that of the surface to be printed. In this case the contact angle will be low and the ink will spread evenly and adhere strongly to the surface. However, if the surface energy of the ink is greater than that of the surface the contact angle will be large – the ink will form globules and not spread evenly.

9.9.4 Capillary Action

In a sufficiently narrow capillary of circular cross-section (radius a), the interface between a fluid and the capillary surface forms a meniscus that is a portion of the surface of a sphere,

Table 9.4 A summary of the parameters associated with the contact angle of a liquid with a solid surface

$>90^\circ$	Contact angle	$<90^\circ$
Low	Solid's surface free energy	High
Poor (hydrophobic)	Wettability	Good (hydrophilic)
Poor	Adhesiveness	Good

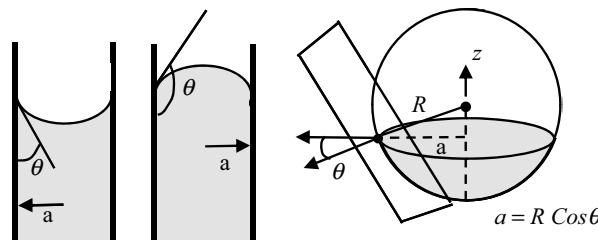


Figure 9.42 The contact angle θ between a sphere and a tangent plane is the angle between the normal to the sphere at the point of tangency and to the plane perpendicular to the z -axis.

and has radius R given by:

$$R = \frac{a}{\cos \theta}.$$

This geometrical relationship is shown in Figure 9.42. The contact angle θ depends on the free surface energies of the fluid and the capillary surface with which it is in contact. The pressure jump ΔP across this surface is:

$$\Delta P = \frac{2T}{R}.$$

The pressure difference is thus given by:

$$\Delta P = \frac{2T \cos \theta}{a}.$$

To maintain hydrostatic equilibrium, the induced capillary pressure is balanced by a change in height (h) of the fluid. As shown in Figure 9.43, this height change can be positive or negative, depending on whether the contact angle is less or greater than 90° . At equilibrium we have the condition:

$$\frac{2T \cos \theta}{a} = h \rho g \quad (9.34)$$

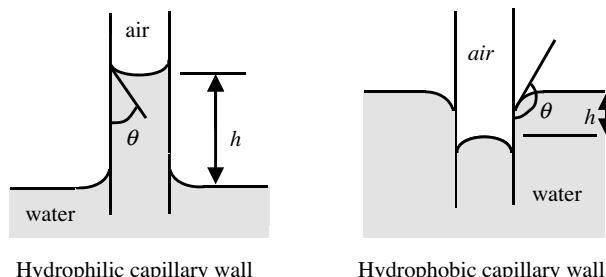


Figure 9.43 To maintain hydrostatic equilibrium the capillary pressure is balanced by the height h of a fluid in a capillary. The height change depends on the magnitude of the contact angle.

where ρ is the density of the liquid in the capillary and g is the gravitational acceleration constant. From Equation (9.34) we obtain the capillary height h as:

$$h = \frac{2T\cos\theta}{\rho g a}. \quad (9.35)$$

Capillarity manifests itself in many ways. Paper towels absorb water through capillarity. When burning a candle, the melted wax rises up the wick due to capillarity. In biology, though blood is pumped throughout the body, capillary action distributes blood in the smallest blood vessels – called *capillaries*. Many species of water-walking insects utilise the high water tension of water. They deflect the free surface of water, thus generating curvature forces that bear their weight, which is typically of the order 1~10 mg.

9.9.5 Practical Aspects of Surface Tension for Lab-on-Chip Devices

- The energy associated with surface tension can be used to drive liquids through microfluidic devices. By treating the surfaces of microchannels to be hydrophilic, water will be driven through any sized channel (typical of lab-on-chip devices) without any applied pressure. This flow is driven by the attractive energy between the water and the channel wall surface.
- Air bubbles pose a big problem in microfluidic devices because of their small radius of curvature r . The relationship $P_i - P_o = 2T/r$, derived in Section 9.9.2 for a bubble in a fluid, informs us that the smaller the radius, the larger will be the pressures required to remove them from a fluidic channel. In the initial wetting of a hydrophilic device, air can become trapped where a wide channel narrows down to a smaller one, for example. Air bubbles can also form after the device is wet, if air spontaneously comes out of solution. Water has a very high surface tension, about 3-times higher than other liquids (see Table 9.3). A strategy to reduce air bubbles is to initially wet the device with a liquid, such as alcohol, that has a lower surface tension. Then water can be fed in behind the other liquid without exposing any air/water interface. This reduces by a factor of around three the force due to surface tension that must be overcome to push air bubbles through a constriction.

Problems

- 9.1. A syringe pump, exerting an excess pressure of 0.1 bar (10 kPa) between the inlet and outlet of an open-ended fluidic micromixer chip, produces a volumetric water flow of 100 $\mu\text{L}/\text{min}$. Calculate the resistance to water flow of this device.
(Give your answer in units of $\text{Pa m}^{-3} \text{s}$.)
- 9.2. (a) Calculate the fluidic resistance of a channel of width 100 μm , height 10 μm , and length L of 2 cm, when filled with an aqueous solution of viscosity (η) 10^{-3} Pa s .
(b) What pressure drop must be applied along this channel to establish a volumetric fluid flow rate of 0.1 $\mu\text{L/sec}$?
- 9.3. (a) Derive an equation to calculate the pressure P_J at the junction of the microfluidic Y-connector, shown in Figure 9.44, in terms of the two input fluidic resistances R_1 and R_2 , the resistance R_o of the output channel, and the applied pressures P_1 and P_2 . The output flow J_o is open to atmospheric pressure.

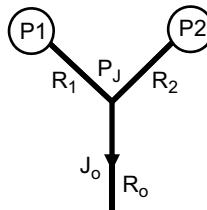
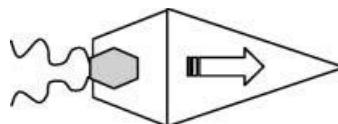


Figure 9.44 The microfluidic Y-connector system for self-study Problem 9.3.

- (b) Use this equation to calculate the fluid flow J_o through the output channel for values of P_1 and P_2 of 120 and 150 kPa, respectively. The circular channels have a diameter of 100 μm , and the two input channels are each of length 5 mm. The output channel is of length 1 cm and exits into a chamber maintained at atmospheric pressure (100 kPa).
- 9.4. The figure below depicts the top view of a toy boat that uses soap as the propelling agent. The soap is contained in the shaded box at the rear of the boat, and the arrow indicates the direction of motion of the boat.



- Explain the origins of the net propulsive force acting on the boat when it is placed on water.
- 9.5. Calculate the height to which water will rise in a glass tube of radius 50 μm . The following values are to be used:

$$T = 7.3 \times 10^{-2} \text{ N m}^{-1}, \theta = 30^\circ, \rho = 1000 \text{ kg m}^{-3}, g = 9.8 \text{ m s}^{-2}.$$

References

- [1] Travis, K.P., Todd, B.D. and Evans, D.J. (1997) Departure from Navier-Stokes hydrodynamics in confined liquids. *Physical Review E*, **55** (4), 4288–4295.
- [2] Pfahler, J., Harley, J. and Bau, B. (1989) Liquid transport in micron and submicron channels. *Sensors & Actuators*, **22**, 431–434.
- [3] Mala, Gh.M. and Li, D. (1999) Flow characteristics of water in microtubes. *International Journal of Heat and Fluid Flow*, **20**, 142–148.
- [4] Garcia, A.L., Bell, J.B., Crutchfield, W.Y. and Alder, B.J. (1999) Adaptive mesh and algorithm refinement using direct simulation Monte Carlo. *Journal of Computational Physics*, **154**, 134–155.
- [5] Abraham, F.F. (2000) Dynamically spanning the length scales from the quantum to the continuum. *International Journal of Modern Physics*, **11** (6), 1135–1148.
- [6] Nie, X.B., Chen, S. and Robbins, W.N.E. (2004) A continuum and molecular dynamics hybrid method for micro- and nano-fluid flow. *Journal of Fluid Mechanics*, **500**, 55–64.
- [7] Clementi, E. (1988) Global scientific and engineering simulations on scalar, vector and parallel LCAP-type supercomputers. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences*, **326**, 445–470.
- [8] Ransing, R., Dyson, P., Williams, P.M. and Williams, P.R. (2008) Chapter 2, in *Fluid Properties at Nano/Meso Scale*, John Wiley & Sons, Ltd, Chichester.
- [9] Isla, M.A. (2004) Einstein-smoluchowski diffusion equation: a discussion. *Physica Scripta*, **70**, 120–125.

Further Readings

- Bruus, H. (2008) *Theoretical Microfluidics*, Oxford University Press.
- Nguyen, N.-T. and Wereley, S.T. (2008) *Fundamentals and Applications of Microfluidics*, 2nd edn, Artech House Inc., Boston.
- Ransing, R., Dyson, P., Williams, P.M. and Rhodri Williams, P. (2008) *Fluid Properties at Nano/Meso Scale*, John Wiley & Sons, Ltd, Chichester.

10

Microfluidics: Dimensional Analysis and Scaling

10.1 Chapter Overview

This chapter describes how dimensional analysis and the use of dimensional parameters can provide the means for simplifying complex physical problems. These tools elucidate the basic relationships between the dependent and independent variables of a physical effect or system and in this way can assist in the planning of experiments and the presentation of measured data designed to better understand them.

After reading this chapter readers will gain an understanding of:

- (i) microscale phenomena and scaling laws for microfluidic systems;
- (ii) the basics of fluid flow at scales above and below the mesoscale;
- (iii) how nondimensional parameters (e.g. Knudsen, Peclet, Reynolds number) can be applied to practical flow problems;
- (iv) how the dominant physical forces and effects at the micro- and nanoscales can be applied to the design of microfluidic systems.

10.2 Dimensional Analysis

Dimensional analysis can be used to reduce complex physical problems to more simple forms prior to attempting quantitative solutions. The basis of this analysis relies on the fact that physical laws are independent of arbitrarily chosen units of measurement, and that they also observe the *concept of similarity*. Forces acting on a particle do not change if we alter the unit of length from microns to miles, or the unit of time from nanoseconds to years, and so forth. In physical terms *similarity* refers to equivalence between two disparate phenomena. For example, under certain conditions there is a direct relationship between the forces acting on a log floating down a river and forces acting on a bioparticle in a microfluidic channel. We will also learn later in this chapter that the ‘ball of fire’ created by a nuclear explosion is similar to the behaviour of an air bubble rising in water! Dimensional

analysis can elucidate the particular conditions required for the similarity of such phenomena, as well as the relationships between them.

An important aspect of dimensional analyses for microfluidics is an appreciation of how physical quantities scale as a function of the characteristic length L of a system. A simple, but instructive, application of scaling is to ask the question ‘Is it better to be a mouse or an arctic hare in the Arctic tundra?’ The parameter of interest is the ratio of heat loss to body mass. Heat loss is proportional to surface area (L^2) and body mass is proportional to volume (L^3). The ratio of heat loss to body mass thus scales as L^{-1} . The heat loss suffered by a mouse (length ~ 4 cm) will thus be ten-times larger than that experienced by a hare of length ~ 40 cm. The energy input (food) required by a mouse will scale accordingly. A polar bear (length ~ 2 m) would have an even more comfortable time than a mouse, by a factor of around 50 : 1. A scaling law therefore expresses the variation of physical quantities with a change in the size L of the system, but maintaining all other quantities such as temperature, pressure and time, for example, constant. We have just considered the scaling of heat loss with body mass in terms of the ratio of a surface effect to a volume effect. We could also consider surface forces such as viscosity, capillary pressure and surface tension, and compare these to volume forces such as inertia and gravity. The basic scaling law for the relative importance of these two classes of force can be expressed as

$$\frac{\text{Surface Forces}}{\text{Volume Forces}} = \frac{L^2}{L^3} = L^{-1}.$$

Thus, as L tends to zero the ratio of surface to volume forces approaches infinity. Capillary pressure and surface tension effects scale as L^{-1} , so that with sufficient reduction of size they can overcome gravitational forces. This is why small diameter fluidic channels (capillaries) can pump water to the top of 100 m tall redwood trees, and how some small insects can walk on water, for example. The electric field generated between electrodes of constant voltage difference ΔV also scale as L^{-1} , where L is the electrode gap. This has been exploited in the application of dielectrophoresis (DEP) described in Chapter 3. The DEP force acting on a particle is proportional to the product of the applied electric field and the local field gradient, and so has dimensions of $(VL^{-1})(VL^{-2}) = (V^2L^{-3})$. Thus, if the effective volume enclosed by the electrodes is reduced 1000-fold, the same DEP force is exerted for a 100-fold reduction of the applied voltage.

A simple example of dimensional analysis is shown in Figure 10.1. The objective is to derive a proof for the theorem of Pythagoras, namely that in any right-angled triangle the square of the hypotenuse is equal to the sum of the squares on the other two sides.

We proceed by recognising that the area A_t of the whole triangle (area $A_1 +$ area A_2) shown in Figure 10.1 is a function of angle θ and c^2 :

$$A_t = f(\theta) c^2$$

We also have

$$A_1 = f(\theta) b^2 \text{ and } A_2 = f(\theta) a^2$$

From the relationship

$$A_t = A_1 + A_2$$

Cancelling $f(\theta)$:

$$f(\theta) c^2 = f(\theta) b^2 + f(\theta) a^2$$

$$c^2 = b^2 + a^2$$

QED

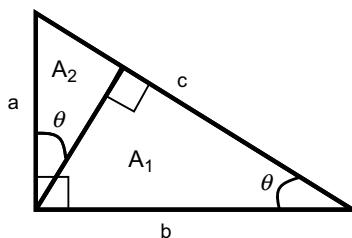


Figure 10.1 Proof of the Pythagorean Theorem using dimensional analysis.

10.2.1 Base and Derived Physical Quantities

Listed in Table 10.1 are the seven base quantities defined for the Système International (SI system), namely: length, time, mass, temperature, current, number of elementary particles, and luminous intensity. The units of length, time and mass are the metre (m), the second (s) and the kilogram (kg), respectively. The SI temperature unit is the kelvin (K), which is defined as the fraction 1/273.16 of the thermodynamic temperature of the triple point of water. The unit for current is the ampere (A) and is defined as the current which, when passed through each of two parallel and infinite, conductors placed one metre apart in vacuum, produces a force of $2 \times 10^{-7} \text{ N m}^{-1}$ on each conductor. The number of elementary particles (e.g. ions, molecules, cells, etc.) is counted in mole units, with one mole equal to Avagadro's number (6.02×10^{23}). The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency $540 \times 10^{12} \text{ Hz}$ and that has a radiant intensity in that direction of $1/683$ watt per solid angle (steradian). 540 THz corresponds to a wavelength of 555 nm, at which the human eye is most sensitive to light.

Other physical quantities can be derived and expressed in terms of the base quantities listed in Table 10.1. For example, force is made a derived quantity by writing Newton's law as $F = ma$. This use of Newton's law is solely for the purpose of deriving the dimensions of force. (Dimensions are not the same as units. For example, the physical quantity, velocity, may be measured in units of metres per second, miles per hour and so on; but regardless of the units used, speed is always a length divided by a time, with dimensions Lt^{-1} .) A particular force being considered, such as that used to define the ampere, need not involve a mass

Table 10.1 The base quantities in the SI system of units

Quantity	SI name	SI Symbol
Length (L)	metre	m
Time (t)	second	s
Mass (M)	kilogram	kg
Temperature (T)	kelvin	K
Electric Current (I)	ampere	A
Amount of substance	mole	mol
Luminous Intensity	candela	cd

being accelerated. An important theorem in dimensional analysis is the *Dimensional Homogeneity Theorem* [1] which states that any physical quantity Q is dimensionally a power law monomial of the form:

$$Q = kM^aL^b t^c I^d T^e, \quad (10.1)$$

where the coefficient k and exponents a, b, c, d and e are real dimensionless numbers whose values distinguish one type of derived physical quantity from another. No alternative form will represent a physical quantity – *all* monomial derived quantities have this power-law form. Some derived physical quantities are presented in Table 10.2.

10.2.2 Buckingham's π -Theorem

The most important exercise in dimensionless analysis is to identify a *complete* set of *independent* variable quantities q_1, q_2, \dots, q_n that determine the value of a *dependent* variable quantity Q in a defined physical effect or process. Q will be a *dependent* variable if its value is determined uniquely by the set of independent variable quantities q_1, \dots, q_n . We can represent this relationship in the form:

$$Q = f(q_1, q_2, \dots, q_n). \quad (10.2)$$

Table 10.2 Some derived physical quantities, with their defining equation or law, and dimensions

Quantity	Defining equation/law	Dimension	Dimension (SI units)	Name
Area	$A = \int dx dy$	L^2	m^2	
Volume	$V = \int dx dy dz$	L^3	m^3	
Velocity	$v = dx/dt$	$L t^{-1}$	ms^{-1}	
Acceleration	$a = d^2x/dt^2$	$L t^{-2}$	ms^{-2}	
Mass density	$\rho = M/V$	ML^{-3}	$kg\ m^{-3}$	
Concentration	$mole/V$	$mol\ L^{-3}$	$mol\ m^{-3}$	
Force	$F = Ma$	$ML t^{-2}$	$kg\ m\ s^{-2}$	newton (N)
Stress/pressure	$p = F/A$	$ML^{-1} t^{-2}$	$kg\ m^{-1}\ s^{-2}$ $(N\ m^{-2})$	pascal (Pa)
Dynamic viscosity	$\eta = p/(dv/dy)$	$ML^{-1} t^{-1}$	$kg\ m^{-1}\ s^{-1}$	poiseuille
Work/energy	$W = \int F dx$	$ML^2 t^{-2}$	$kg\ m^2\ s^{-2}\ (Nm)$	joule (J)
Surface tension	$T = W/A$	Mt^{-2}	$kg\ s^{-2}\ (Nm^{-1})$	
Power	$P = dW/dt$	$ML^2 t^{-3}$	$kg\ m^2\ s^{-3}\ (J\ s^{-1})$	watt (W)
Frequency	$f = 1/t$	t^{-1}	s^{-1}	hertz (Hz)
Charge	$Q = \int I dt$	It	$A\ s$ (C)	coulomb
Electromotive force (voltage)	$E = P/I$	$ML^2 t^{-3} I^{-1}$	$kg\ m^2\ s^{-3}\ A^{-1}$	volt (V)
Capacitance	Q/E	$M^{-1} L^{-2} t^4 I^2$	$kg^{-1}\ m^{-2}\ s^4\ A^2$	farad (F)
Resistance	E/I	$ML^2 t^{-3} I^{-2}$	$kg\ m^2\ s^{-3}\ A^{-2}$	ohm (Ω)
Conductance	I/E	$M^{-1} L^{-2} t^3 I^2$	$kg^{-1}\ m^{-2}\ s^3\ A^2$	siemens (S)

According to Equation (10.1) all the quantities q have dimensions of the form:

$$[q] = M^a L^b t^c I^d T^e$$

Quantities q_1, \dots, q_n will form a *complete* set if no other quantity can influence the value of Q , and *independent* if by changing any one of their values this does not alter the value of any other member of the set. As an example of this exercise, consider an experiment to determine the terminal velocity of various solid spherical particles rising or falling freely through various fluids. Our understanding of the basic hydrodynamics involved informs us that the terminal velocity U (our dependent quantity) is attained when the gravitational buoyancy or settling force is equal to the viscous drag force acting on the particle. This would suggest that U depends on the mass (m) and diameter (D) of the particle, the density ρ_p of the particle and fluid (ρ_f), and the viscosity η of the fluid (we assume that the gravitational acceleration constant g and temperature remain constant). The five quantities m , D , ρ_p , ρ_f , and η form a *complete* set – but not an *independent* one. If we define the diameter and density of the particle, the value for the mass m will also be defined. Therefore, either m or ρ_p must be excluded, to give a total of four independent variables in the set q_1, \dots, q_n . If, for example, we require five data points for each determination of U as a function of each of the four independent variables, we would have to make 5^4 (i.e. 625) measurements. We can reduce the number of required measurements to 125 by making the difference in the particle and fluid density ($\Delta\rho$) as one of the variables, to replace the two variables ρ_p and ρ_f . An important application of dimensional analysis is that it leads to a significant reduction of the number of required experiments, whilst also providing an indication of how the experiments should be designed and the results presented. The way to achieve this is by application of Buckingham's π -Theorem [2].

Sonin [3] gives a clear description of the physical basis of dimensionless analysis and Buckingham's π -Theorem, and shows that an alternative form of Equation (10.2) is:

$$\pi = f(q_1, q_2, \dots, q_k; \pi_1, \pi_2, \dots, \pi_{n-k}), \quad (10.3)$$

where $\pi_1, \pi_2, \dots, \pi_{n-k}$ is a complete dimensionless subset of the original q_1, q_2, \dots, q_n given in Equation (10.2). The values of these dimensionless quantities are independent of the sizes of the base units, but the values of q_1, \dots, q_k do depend on base unit size. From the principle of *Dimensional Homogeneity* [1] that any physically meaningful equation must be dimensionally homogeneous, then quantities q_1, \dots, q_k must in fact be absent from Equation (10.3). In other words we have a dimensionless equation of the form:

$$\pi = f(\pi_1, \pi_2, \dots, \pi_{n-k}). \quad (10.4)$$

Equation 10.4 encompasses Buckingham's π -Theorem, which states:

For a given physical effect or process where there are n physically relevant variables that can be described by k fundamental dimensions, there are a total of $n-k$ independent, dimensionless, quantities (or 'Pi groups') $\pi_1, \pi_2, \dots, \pi_{n-k}$. The behaviour of the effect or process can be described by dimensionless Equation (10.4).

This theorem can be applied as follows:

1. Clearly define the problem and list the n variable parameters of importance.
2. Identify the *dependent* variable of interest.

3. Express each variable in terms of [M] [L] [t] [I] [T] dimensions. The number of fundamental units corresponds to the value for k .
4. Determine the required number of dimensionless parameters ($n - k$).
5. Form a dimensionless parameter π by multiplying the dependent variable by the remaining variables, each raised to an unknown exponent.
6. Solve for the unknown exponents.
7. Repeat this process if $(n - k) > 1$.
8. Express the result as a relationship among the dimensionless parameters.
9. Compare with experimental data.

As a demonstration of the procedure we will consider a particularly extreme case of the *concept of similarity*. The British physicist Sir Geoffrey Taylor employed dimensional analysis [4] to estimate the energy yield of the first atomic bomb explosion (Trinity, New Mexico, 16 July 1945) when this was still classified information. The concept of similarity he used was that the buoyancy of hot air created in the ‘ball of fire’ of the explosion would behave like a rising bubble in water (until the hot air suffers turbulent mixing with the surrounding cold air). An example of one of the photographs he used in this analysis is shown in Figure 10.2.

Taylor assumed that the radius R of the ‘equivalent bubble’ depends on the energy yield E of the explosion, the time t after detonation, and the initial density ρ of the air. He found that the only parameter with dimensions of length that can be constructed from these quantities is:

$$R = k \left(\frac{Et^2}{\rho} \right)^{1/5}. \quad (10.5)$$

Based on official high-speed photography of the Trinity atom bomb test Taylor used his formula to deduce that the bomb’s yield was 16.8 kilotons of TNT – a result close to the

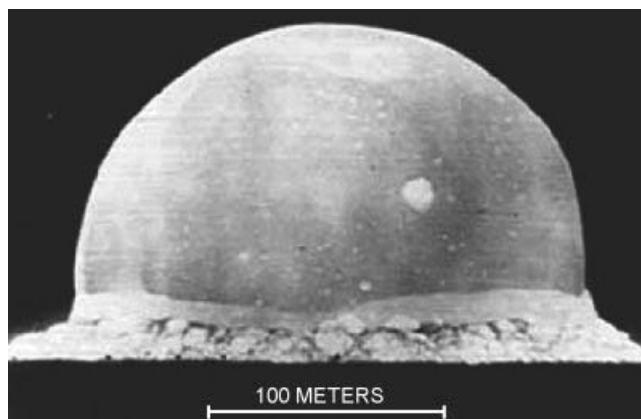


Figure 10.2 A photograph of the ‘ball of fire’ produced by the Trinity atom bomb taken 0.015 seconds after detonation. The luminous globe rises like a large bubble in water. (G. Taylor, *Proc. Roy. Soc. Lond. A201: 175–186*, 1950. Reproduced with permission.)

official ‘top-secret’ value of 18 kilotons. Equation 10.5 can be derived as follows:

Assuming that the radius R of the atomic ‘cloud’ depends on the explosive energy E , time t after detonation and the initial density ρ of the air, we have:

$$R = f(E, t, \rho).$$

This equation contains four variable parameters, to give $n=4$ in Buckingham’s π -Theorem. We now construct the following table containing these variables and define their dimensions with the aid of Table 10.2:

R	E	t	ρ
L	ML^2t^{-2}	t	ML^{-3}

In this table there are three different dimensions (M, L, t) to give $k=3$, and $n-k=1$. Buckingham’s π -theorem thus informs us that only one dimensionless parameter is required to describe the initial atomic blast. This dimensionless parameter π is formed by multiplying the dependent variable R by the remaining parameters, each raised to an unknown exponent:

$$\pi = R(E^a t^b \rho^c). \quad (10.6)$$

In terms of the dimensions of the quantities in equation (10.6) we have:

$$M^0 L^0 t^0 = (L)(ML^2t^{-2})^a(t)^b(ML^{-3})^c.$$

We now solve for a, b and c using the method of exponents:

$$\begin{aligned} M: \quad 0 &= a + c, \\ L: \quad 0 &= 1 + 2a - 3c, \\ t: \quad 0 &= -2a + b. \end{aligned}$$

to give $a=-1/5$; $b=-2/5$; $c=1/5$. Equation 10.6 therefore takes the form:

$$\pi = R \left(\frac{\rho}{Et^2} \right)^{1/5}$$

or

$$R = k \left(\frac{Et^2}{\rho} \right)^{1/5}. \quad (10.7)$$

In Equation (10.7) the initial air density ρ is known and the bomb’s energy yield E is also fixed. Thus, in the time range for which the luminous globe produced by the explosion behaves like a rising bubble in water, we should expect $R^{5/2} \propto t$. Thus, a plot of $2.5 \log R$ against $\log t$ should produce a straight line. This plot is shown in Figure 10.3.

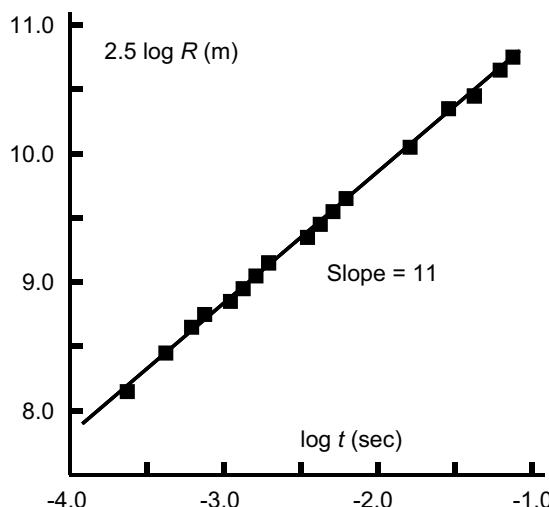


Figure 10.3 A logarithmic plot, based on equation (10.12), of the radius of the luminous globe produced by the Trinity atom bomb as a function of time after detonation. (Derived from the dimensionless analysis performed by G. Taylor, *Proc. Roy. Soc. Lond.* A201: 175–186, 1950. Reproduced with permission.)

The straight line in Figure 10.3 corresponds with

$$R^5 t^{-2} = 6.67 \times 10^{-8} \text{ m}^5 \text{ s}^{-2}.$$

Based on this result Taylor determined the value of the atom bomb's energy yield to be 7.14×10^{13} J. Taking 1gm of TNT as liberating 1000 calories (4.2 kJ), then this energy is equivalent to 16.8 kilotons of TNT. Because of the assumed similarity with a rising air bubble, this energy does not include that part of the energy radiated beyond the 'ball of fire'. This may partly account for the slight discrepancy between Taylor's result and the official one of 18 kilotons TNT disclosed many years later.

The following worked example is of relevance to the design of microfluidic systems:

Example 10.1

Measurements of the pressure drop ($\Delta p = p_1 - p_2$) along a microfluidic channel of fixed length and depth are to be made as a function of the channel widths w_1 and w_2 for different values of the fluid density ρ and the volumetric fluid flow Q ($\mu\text{L/sec}$). The geometry of the channel is shown in Figure 10.4.

- (a) Construct a table containing the experimental variables and their dimensions.
- (b) Derive the required dimensionless parameters.
- (c) A multitude of different plots of the experimental data can be drawn (e.g. Δp as a function of widths w_1 and w_2 , for various values of flow rate and fluid density). Explain how the data can be presented to produce a single 'universal' plot.

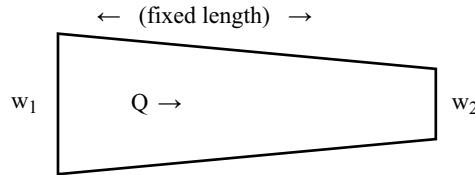


Figure 10.4 Fluid flow Q along a channel of fixed length and depth, and variable width.

Solution:

- (a) Assume $\Delta p = f(Q, \rho, w_1, w_2)$ and construct a table of the experimental variables:

Δp $ML^{-1}t^{-2}$	Q L^3t^{-1}	ρ ML^{-3}	w_1 L	w_2 L
-------------------------------	--------------------	---------------------	--------------	--------------

- (b) From this table $n = 5$ (variables) and $k = 3$ (dimensions). From Buckingham's π -Theorem there are $n-k = 2$ required dimensionless parameters.

It is often possible by inspection of a table of variables to deduce the form of one of the required dimensionless parameter. In this case we have an obvious one, namely the ratio w_1/w_2 (or w_2/w_1). For the first dimensionless parameter we choose:

$$\pi_1 = w_1/w_2. \quad (10.8)$$

To find the 2nd dimensionless parameter we assume that

$$\pi_2 = \Delta p(Q^a \rho^b w_1^c);$$

that is,

$$M^0 L^0 T^0 = (ML^{-1}T^{-2})(L^3T^{-1})^a(ML^{-3})^b(L)^c.$$

Solving for a, b and c using the method of exponents:

$$\begin{aligned} M: \quad 0 &= 1 + b, \\ L: \quad 0 &= -1 + 3a - 3b + c, \\ t: \quad 0 &= -2 - a. \end{aligned}$$

Solving these equations gives: $a = -2$; $b = -1$; $c = 4$.

$$\text{So that : } \pi_2 = \frac{\Delta p \cdot w_1^4}{Q^2 \rho}$$

or

$$\Delta p = k \left(\frac{Q^2 \rho}{w_1^4} \right). \quad (10.9)$$

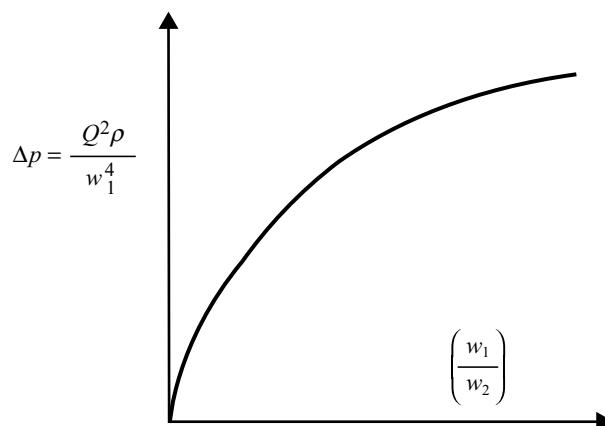


Figure 10.5 A plot of $\Delta p (= Q^2 \rho/w_1^4)$ as a function of (w_1/w_2) takes the form of a smooth *universal plot*.

(We could also choose $\pi_2 = \Delta p(Q^a \rho^b w_2^c)$ which would yield $\pi_2 = \frac{\Delta p \cdot w_2^c}{Q^a \rho^b}$.) (Attempts to find a dimensionless parameter of the form: $\pi_2 = \Delta p(Q^a \rho^b)$ will fail.)

- (c) To produce the complete family of experimental plots of the five variable parameters Δp , Q , ρ , w_1 and w_2 will require a very large number of measurements (if only four data points are to be taken, this would still require $4^5 = 1024$ data points). From Equation (10.4)

$$\pi_1 = f(\pi^2, \dots, \pi_{n-k}),$$

so that, from equations (10.8) and (10.9), we can write:

$$\Delta p = \frac{Q^2 \rho}{w_1^4} = f\left(\frac{w_1}{w_2}\right).$$

A plot of $\frac{Q^2 \rho}{w_1^4}$ versus $\left(\frac{w_1}{w_2}\right)$ should produce a single *universal* curve as depicted in Figure 10.5.

Not only has this dimensionless analysis provided the means for greatly simplifying the presentation of experimental data, it also assists in the planning of the experiments by revealing the basic relationships between the dependent and independent variables.

10.3 Dimensionless Parameters

To be able to design a well-functioning microfluidic device requires an understanding of how fluids flow within channels having micro- and nanoscale dimensions. Although fluids are collections of discrete molecules, each with individual properties, it is usually helpful to define velocity, temperature, density and pressure as the statistically based properties of a continuous material. This may no longer be the case when device dimensions are reduced.

Table 10.3 Values of viscosity and surface tension for various liquids at 293 K

Liquid	η (Pa s)	T_s (N m ⁻¹)
Water	1.002×10^{-3}	7.275×10^{-2}
Blood (37 °C)	$3 \sim 4 \times 10^{-3}$	5.5×10^{-2}
Ethanol	1.074×10^{-3}	2.21×10^{-2}
Methanol	5.94×10^{-4}	2.27×10^{-2}
Mercury	1.55×10^{-3}	47.2×10^{-2}
Benzene	6.04×10^{-4}	2.89×10^{-2}
Chloroform	6.96×10^{-4}	2.75×10^{-2}
Glycerol (100%)	1.41	6.4×10^{-2}

As an example, if the physical dimensions of a fluidic channel become so small that there are statistically few molecules near the channel walls at any time, the assumption of zero slip boundary conditions discussed in Chapter 9 can no longer be assumed. Fluid flow behaviour can be better appreciated and predicted by employing dimensionless parameters such as the Knudsen, Peclet, Reynolds, and Bond numbers. These parameters can reveal which physical mechanisms should be used to manipulate the flow and create the functionality desired for a particular microfluidic device. Such ‘top level’ understanding can be obtained without the need to perform in-depth analyses.

Useful physical data to be used in the application of dimensionless parameters are presented in Tables 10.3–10.6.

10.3.1 Hydraulic Diameter

Most dimensionless parameters include a length scale in their definition. This characteristic dimension L varies with the geometrical shape of the fluidic channel. The concept of a wetted or hydraulic diameter is generally chosen as the appropriate length scale, and two

Table 10.4 Diffusion coefficients for various molecules and ions in water at 298 K

Molecule	D ($10^{-9} \text{ m}^2 \text{s}^{-1}$)
Water	2.26
Sucrose	0.52
Methanol (CH ₃ OH)	1.6
Glycine	1.06
NaCl	1.7
H ⁺	9.3
OH ⁻	5.3
Na ⁺	1.33
K ⁺	1.96
Cl ⁻	2.03

Table 10.5 Diffusion coefficients for various macromolecules and particles in water at 293 K. (Derived using the Stokes-Einstein relation: $D = kT/(6\pi\eta a)$ where ‘ a ’ is the hydrodynamic radius of a spherical particle.)

Macromolecule	D ($\text{m}^2 \text{s}^{-1}$)
Ribonuclease	1.2×10^{-10}
Lysozyme	1.0×10^{-10}
Serum albumin	5.9×10^{-11}
Haemoglobin	6.9×10^{-11}
Urease	3.5×10^{-11}
Collagen	6.9×10^{-12}
Viruses, bacteria, cells	$10^{-13} \sim 10^{-16}$

common examples of this are described in Figure 10.6. The hydraulic diameter D_H is defined as:

$$D_H = \frac{4 \times \text{Area}}{\text{Wetted Perimeter}}. \quad (10.10)$$

For the case of the channel of circular cross-section D_H is equal to the channel diameter D . For rectangular channels:

$$D_H = \frac{2wh}{(w+h)} = \frac{2}{(1/h + 1/w)},$$

where, as shown in Figure 10.6, h is the channel height and w its width.

For a channel of triangular cross-section and equal sides (a)

$$D_H = \frac{2a \sin 60^\circ}{3} = \frac{a}{\sqrt{3}}.$$

Equilateral triangle cross-sections can be etched into silicon substrates, but are not commonly found in glass or plastic substrates. Two cross-sections that are common in microfluidic devices are the trapezoidal and rounded trapezoidal ones shown in Figure 10.11 for self study problem number 4. The effective hydraulic radius value ($D_H/2$) obtained for a non-standard cross-section channel (e.g. a trapezoid) can be used with Equations (9.17) and (9.18) of Chapter 9 to provide a *rough* estimate of the pressure drop required to produce a

Table 10.6 Diffusion coefficients for some molecules in water and air at 293 K. (As a rough guide, a molecule’s diffusion coefficient is $\sim 10^4$ -times greater in air than in water)

Molecule	In water ($10^{-9} \text{ m}^2/\text{sec}$)	In air ($10^{-5} \text{ m}^2/\text{sec}$)
H_2O	2.26	2.5
CO_2	1.6	1.6
O_2	2.0	2.0

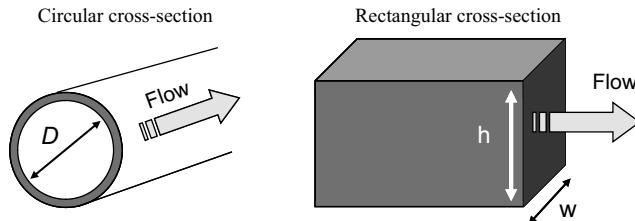


Figure 10.6 The appropriate length scale for a fluidic channel is often given as its wetted or hydraulic diameter (e.g. D or $2wh/(w + h)$).

desired volume flow rate through such a channel. Equations 9.19 to 9.21 should be used to provide a more accurate estimate for circular, semi-circular or rectangular cross-sections.

10.3.2 The Knudsen Number

The *Knudsen Number* is a dimensionless parameter that compares the characteristic dimensions of a microfluidic device to the intermolecular spacing (mean free path between molecular collisions) of the fluid. It provides an important test of the validity of the continuum approximation, and is defined as:

$$Kn = L_{mfp}/L,$$

in which L_{mfp} is the mean free path and L is the characteristic length of the flow-field. L can be taken as the hydraulic diameter defined by Equation (10.10), or the gradient of a bulk property such as density ($\rho/|d\rho/dx|$). From Chapter 9, Equation (9.2), the mean free path, for an ideal gas, is given by the formula:

$$L_{mfp} = k T / (\sqrt{2\pi} P d^2), \quad (10.11)$$

where k is the Boltzmann constant, T is absolute temperature, P is absolute pressure, and d is the molecular diameter. Typical molecular diameters fall in the range $0.2 \sim 0.3$ nm. For liquids, since the molecules are always in a collision state, the mean free path is roughly equivalent to the molecular diameter. We can appreciate the difference at the molecular level between a liquid and a gas by noting that one dm³ of liquid nitrogen weighs ~ 800 gm, whilst at STP gaseous nitrogen weighs ~ 1.2 gm per dm³. At the molecular level this informs us that on average a nitrogen molecule in the gas phase occupies ~ 670 -times more space than it does as a liquid. The average centre-to-centre separation of molecules in the liquid state is a little larger than its molecular diameter d , and so the average separation of molecules in a gas will be $\sim 670^{1/3}d$, namely $\sim 8.8d$. We can imagine a gas consisting of molecular spheres randomly distributed in space, with an average separation close to ten-times their molecular diameter. The molecules will have a distribution of velocities given by the Maxwell-Boltzmann distribution described in Chapter 9, and the distance they travel before colliding with another molecule is given by Equation (10.11).

The range of the Knudsen number for gaseous systems is shown in Figure 10.7. As a rough guide we can adopt the continuum model if the characteristic scale of our system is more than 1000-times larger than the molecular mean free path length. At the other end of

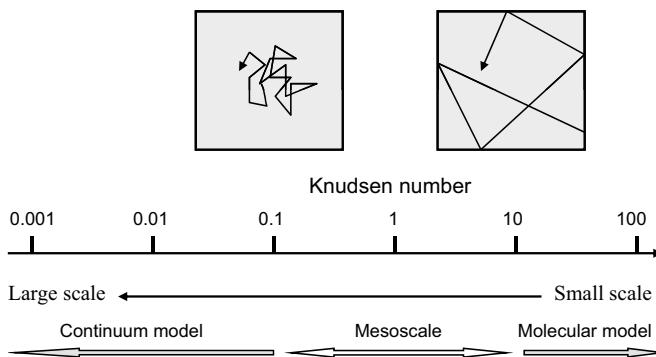


Figure 10.7 The range of Knudsen numbers for gaseous systems, to show the range ($\text{Kn} < 0.1$) where the continuum model and zero slip boundary conditions apply, and the range ($\text{Kn} > 10$) where discontinuous, dynamic, molecular flow dominates. Between these two ranges there is the meso region.

the scale, where the molecular mean free path length is 10-times larger than the characteristic length of the system, the molecular particles will collide with the physical boundaries of the system more often than they will with one another. We now have to consider the dynamics of the individual molecular particles – we cannot treat them as a homogeneous medium. In the Knudsen number range between the continuum and molecular model we operate in the mesoscale, and methods for modelling at this scale were discussed in Chapter 9.

The following ranges of the Knudsen number provide a rough guide as to when we can treat a gaseous fluid as a continuum or as an assembly of discrete molecular particles:

- $\text{Kn} < 0.001$: The continuum model and zero slip boundary conditions are appropriate.
- $< \text{Kn} < 0.1$: The continuum model holds, but there is finite slip at boundaries.
- $< \text{Kn} < 10$: The *mesoscale* region between the continuum approximation and a model involving discontinuous, dynamic, molecular physics.
- $\text{Kn} > 10$: The continuum approximation is invalid and a particle-based method, such as Monte Carlo simulations, should be used.

The following examples should assist with an understanding of these concepts:

Example 10.2

- How many molecules of nitrogen (N_2) are to be found in samples of pure nitrogen gas enclosed in vessels of the following dimensions at STP (293 K, $1.013 \times 10^5 \text{ Pa}$)?
 - $10 \mu\text{m} \times 10 \mu\text{m} \times 10 \mu\text{m}$,
 - $10 \text{ nm} \times 10 \text{ nm} \times 10 \text{ nm}$.
- Estimate the average separation of the nitrogen molecules in these two vessels.
- Calculate the average distance between collisions of the nitrogen molecules (assume a molecular diameter d of 0.25 nm).

Solutions:

- (a) To solve this we remind ourselves from Chapter 9 of the Gas Law ($pV = nRT$) that follows from Avogadro's law, which in turn states that:

Equal volumes of gases at the same temperature and pressure contain the same number of particles.

The volume occupied by a mol (gram molecular weight) of any gas at STP is thus 22.414 dm^3 . Based on Avogadro's number there are $6 \times 10^{23}/22.4 = 2.7 \times 10^{22}$ nitrogen (N_2) molecules, or any other gas molecule, per dm^3 at STP. (The estimate of the number of hydrogen molecules in the Earth's exosphere is 8000 per dm^3 .)

- (i) A vessel of dimensions $10 \times 10 \times 10 \mu\text{m}$ has a volume of 10^{-12} dm^3 , and thus contains 2.7×10^{10} nitrogen molecules.
- (ii) A vessel of dimensions $10 \times 10 \times 10 \text{ nm}$ has a volume of 10^{-21} dm^3 , and thus contains 27 nitrogen molecules.
- (b) 1 dm^3 contains 2.7×10^{22} molecules. If we assume that each molecule occupies a cube of side l , then:

$$l^3(2.7 \times 10^{22}) = 1, \text{ to give } l = 3.34 \text{ nm.}$$

(We assume cubic molecular spaces because spheres cannot be packed together to fill all space.)

- (c) Using Equation (10.11) to calculate the mean free path length between collisions we obtain a value for L_{mfp} of 144 nm (i.e. ~ 44 -times larger than the average molecular separation distance of 3.34 nm calculated for (b) above).

Example 10.3

- (a) What is the size of container below which the properties of a gas within it (at STP) should not be modelled as a continuum?
- (b) What size of container will correspond to a transition between the meso scale and where a gas (at STP) within it should be modelled in terms of molecular dynamics?

Solutions:

- (a) The value for the Knudsen number Kn is given by the ratio (mean free path length L_{mfp})/(cube side l). A value for $\text{Kn} < 0.001$ represents the limit where the continuum model is valid. Thus, the minimum value for $l = 144 \text{ nm}/0.001 = 144 \mu\text{m}$.
- (b) The transition between the meso and molecular scale occurs for $\text{Kn} \approx 0.1$. The transition dimension is thus given by $l \approx 1.4 \mu\text{m}$.

Example 10.4

- Calculate the mean molecular speed $\langle v \rangle$ of nitrogen molecules in nitrogen gas at STP. Assume an atomic mass unit for N₂ of 28.
- Calculate the mean time between collisions of the nitrogen molecules and the collision frequency.

Solutions:

- From Equation (9.13) in Chapter 9 the mean speed is given by:

$$\langle v \rangle = \sqrt{\frac{8kT}{\pi m}}$$

in which m is the atomic mass, calculated by taking the periodic table value multiplied by the atomic mass unit (amu) of value given by 1 amu = 1.66×10^{-27} kg.

$$\text{Thus } \langle v \rangle = [(8 \times 1.38 \times 10^{-23} \times 293)/(28\pi \times 1.66 \times 10^{-27})]^{1/2} = 470 \text{ m s}^{-1}.$$

- The mean time τ between collisions is given by:

$$\tau = L_{mfp} / \langle v \rangle = 144 \text{ nm} / 470 \text{ m s}^{-1} = 0.31 \text{ ns.}$$

The collision frequency is given by $1/\tau = 3.3 \times 10^9 \text{ s}^{-1}$.

10.3.3 The Peclet Number: Transport by Advection or Diffusion?

The constant motion of molecules in fluids ensures that, when one fluid is placed adjacent to a second fluid, its molecules proceed into that second fluid in a process called diffusion. When we employ the continuum concept, instead of calculating each individual motion, we calculate the average motion of a statistically significant number of molecules. It then becomes convenient to separate the actual diffusion process into two conceptual transport mechanisms: a molecular process modelled as a statistical random walk that is proportional to the degree of kinetic energy in the system and an advective process in which molecules are carried along by the average velocity of the flow. The common practice is to restrict the word *diffusion* to describe the first process and label the second process *advection* (*convection* if heat is being transferred). The relative importance of these two conceptual transport mechanisms is given by the Peclet Number, the ratio of advection and diffusion:

$$\text{Pe} = vL/D,$$

in which v is the fluid velocity, D is the diffusion coefficient of the solute in the solvent, and L is the characteristic dimension of the fluid passage (Figure 10.6). When L is so small that the Peclet number is less than 1000, molecular diffusion becomes an important mechanism for mixing. Stirring may be appropriate for mixing in a macroscale device, but a diffusion-based approach should be used in a low Pe device.

10.3.4 The Reynolds Number: Laminar or Turbulent Flow?

All fluid flow, whether around an object, in pipes, or in a river, can be broadly classified as either laminar or turbulent. These two flow regimes behave markedly differently, with

significant implications for mass and heat transport. Whether fluid flow is laminar or turbulent depends on the relative importance of the inertial forces ($\rho v^2/L$) versus viscous forces ($\eta v/L^2$) in the flow (i.e. ratio of the momentum of the fluid and the friction force imparted by the channel walls). This ratio is defined as the Reynolds Number:

$$Re = \frac{(\rho v^2/L)}{(\eta v/L^2)} = \frac{\rho v L}{\eta}$$

Re (originally proposed by Osborne Reynolds [5] in which v is the bulk velocity of the flow, ρ is fluid density, and η is the fluid's *dynamic viscosity*) can also be expressed as vL/ν , where ν is the *kinematic viscosity* ($\nu = \eta/\rho$) with units of m^2/s . The characteristic length L can be taken as the diameter of a fluid channel or pipe, or the diameter of a spherical object in a fluid stream.

A low Reynolds-number flow is a laminar, or layered, flow in which fluid streams flow parallel to each other and mix only through advective and molecular diffusion. Laminar flow is dominated by viscous forces and has fluid velocity at all locations invariant with time when boundary conditions are constant. There is advective mass transport only in the direction of fluid flow. An excellent example of laminar flow can be found with some toothpastes (Figure 10.8). Several brands have two or more different components, typically varying in both colour and composition. When such toothpaste is squeezed out of its tube, the colors do not mix because the paste's high viscosity ensures a low Re and thus laminar flow.

In contrast, a high Reynolds-number flow is a turbulent flow in which inertial forces dominate and various parts of the fluid exhibit motions that are simultaneously random in both space and time. Significant advective mass transport occurs in all directions. This is the kind of flow we can see in rapidly flowing mountain streams, or when we vigorously stir cream into coffee, for example. This difference in the behaviour of laminar and turbulent flow is depicted in Figure 10.9.

The transition between laminar and turbulent flow typically occurs above $Re = 2000$, though some experiments [6] suggest transition in gas flows in microchannels may occur at Re as low as 400. A flow can be identified as laminar or turbulent by either experimental or computational methods. From experimental data laminar flow is identified by a linear

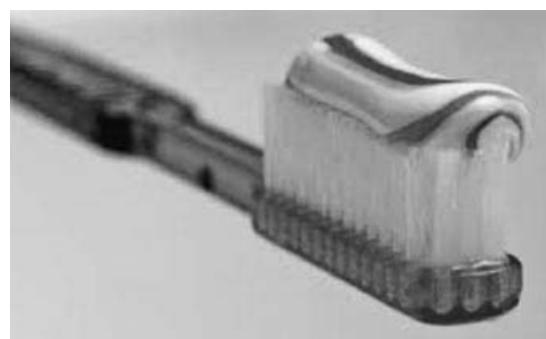


Figure 10.8 Laminar flow in toothpaste, characterised by a low Reynolds number resulting from high viscosity. Mixing of the components takes place by molecular diffusion – a very slow process.

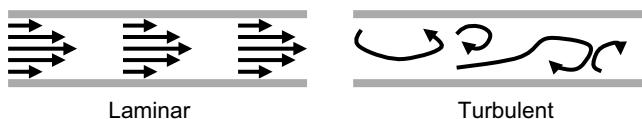


Figure 10.9 Schematic representation of laminar flow ($Re < \sim 2000$) and turbulent flow ($Re > \sim 2000$).

proportionality between the log of the pressure drop in the channel and the volume flow rate, i.e. a straight line on a log-log plot of pressure loss versus flow rate. If the flow transitions to turbulence at higher flow rates, the same linear proportionality no longer holds and the slope of the line changes at that flow rate, as depicted in Figure 10.10.

Transition to turbulence can also be identified using numerical techniques such as the finite element or finite volume methods described in Chapter 9 to simulate the flow. Turbulence can be defined as irregular flow with random variation of flow properties (e.g. velocity, pressure, etc.) in both time and space coordinates simultaneously. Hence, a numerical simulation based on solving the appropriate conservation of mass and momentum equations will not converge to a steady solution if the flow is randomly varying. Time-averaging of the flow properties or some other technique must be used to model a turbulent flow. By these means, the flow in a microchannel can be accurately differentiated as laminar or turbulent, and analysed accordingly.

10.3.5 Reynolds Number as a Ratio of Time Scales

If a flat plate at rest receives a step-function impulse of force, causing it to move in its own plane at velocity v , a fluid boundary layer will develop at the plate surface due to the nonslip of fluid at this boundary. If L is the characteristic length scale, then the characteristic time τ_c for transport of material by convection down the resulting fluid flow is $\tau_c = L/v$. The boundary layer will widen at a rate proportional to the fluid viscosity. The *kinematic viscosity* ν ($\nu = \eta/\rho$) has units of $m^2 s^{-1}$ so that the characteristic time τ_c for a

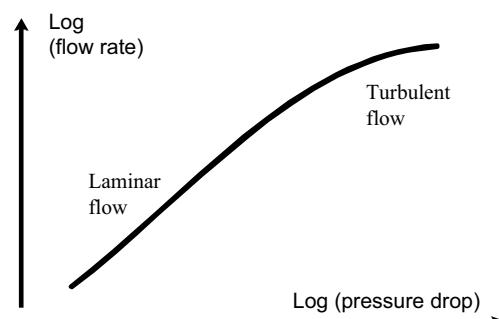


Figure 10.10 Laminar flow is identified by a linear proportionality between the logarithm of the pressure drop in a channel and the volume flow rate of the fluid.

viscosity controlled effect to be transmitted normal to the fluid flow can be given as $\tau_{visc} = L^2/v$. The ratio of viscous to convective time scales is

$$\frac{\tau_{visc}}{\tau_c} = \frac{(L^2/v)}{(L/v)} = \frac{\rho v L}{\eta} = Re.$$

Thus the Reynolds number is a measure of the viscous and convective time scales. A large Reynolds number means that viscous effects propagate slowly into the fluid. This is the reason why boundary layers are thin in high Reynolds number flows because the fluid is being convected along the flow direction at a much faster rate than the spreading of the boundary layer, which is normal to the flow direction.

10.3.6 The Bond Number: How Critical is Surface Tension?

Another flow characteristic that becomes important in microscale channels is the interfacial tension between gas and liquid phases, or between immiscible fluids. In flow in porous media the Capillary number, a ratio of viscous and interfacial tension forces, is important. For droplet breakup the Weber number, a ratio of inertial and interfacial tension forces, is the useful parameter (see next section).

The Bond number is defined as:

$$Bo = \Delta\rho g L^2 / T_s,$$

which represents a ratio of gravity and interfacial tension forces, in which $\Delta\rho$ is the density difference between the two fluids, g is the acceleration of gravity, T_s is surface tension, and L is the appropriate length scale (e.g. wetted diameter of a capillary, or the contact radius).

Because L is squared the Bond number decreases rapidly as the flow passage is reduced in size to the microscale. A high Bond number indicates that the system is relatively unaffected by surface tension effects; a low number (typically less than one) indicates that surface tension dominates. A low Bond number flow is more likely to respond to change in surface energy than change in elevation of the free surface between the phases. This is the reason that a liquid will rise in a capillary tube in spite of the gravitational force acting on it. Intermediate Bond numbers indicate a nontrivial balance between the two effects.

A characteristic length λ , known as the capillary length, is defined as

$$\lambda = \sqrt{\frac{T_s}{\Delta\rho g}}$$

and corresponds to the curved surface or meniscus length of a droplet or fluid in contact with a surface. The Bond number therefore compares the characteristic dimension of the fluid element to the capillary length ($Bo = L^2/\lambda^2$). For water at STP the capillary length λ is $\sim 2 \times 10^{-3}$ m. A fluid droplet of low Bond number ($Bo \ll 1$) deposited on a substrate will appear as a rounded, hemispherical drop, whereas a large drop of fluid ($Bo \gg 1$) will appear more like a flat fluid disk.

The Bond number can also be configured to accommodate forces other than gravitational ones. Electrowetting on a dielectric (EWOD) is described in Chapter 3. The deformation of a

droplet when exposed to an electric field E can be related to the electric Bond number, given by

$$Bo_e = \epsilon E^2 L / T_s,$$

where ϵ is the dielectric permittivity of the liquid.

10.3.7 Capillary Number: Relative Importance of Viscous and Surface Tension Forces

In small-scale flows where the effect of surface tension is important, the Capillary number Ca is defined as the ratio of the viscous (elongational) force to the surface tension force acting at an interface between a liquid and a gas, or between two immiscible liquids (e.g. oil and water):

$$Ca = \frac{\eta v}{T_s},$$

where η is the dynamic viscosity, v is the characteristic velocity and T_s is the surface or interfacial tension between the two fluid phases. The shear stress is given by $\eta v / L$ and the capillary force as T_s / L . The ratio of these two forces gives the Capillary number. The characteristic velocity v could be the rate (m s^{-1}) of emergence of oil through an oil-water saturated porous material, or the rate of shear or elongation of a fluid emerging from a nozzle or constriction. As the emerging fluid stream elongates and reduces in diameter, the capillary force (T_s / L) increases until it breaks up the fluid stream into droplets. This occurs at a critical Capillary number typically of the order $0.1 \sim 0.01$.

10.3.8 Weber Number: Relative effects of Inertia and Surface Tension

The Weber number We can be used to predict the disruption of the flow of small volumes or thin films formed between the interface between two immiscible fluids or between a fluid and a gas. The quantity determined by the Weber number is the ratio of the fluid's inertial force and surface tension forces:

$$We = \frac{\rho v^2 L}{T_s} \left(= \frac{\rho v^2}{T_s / L} \right)$$

where L is typically the thickness of a fluid film or the diameter of a fluid droplet. The factor ρv^2 corresponds to dynamic pressure and T_s / L to capillary pressure. When surface tension forces dominate, a fluid element is likely to take the form of a spherical droplet (i.e. having a convex interface). If inertial forces dominate an interface can assume a rippled structure or form localised concave surfaces that can eventually disrupt the fluid's structural form.

A fluid jet that rapidly 'atomises' into a fine spray of spherical droplets corresponds to a low Weber number. Consider the 'splash' caused by an object falling into a water-air surface. If the splash mainly takes the form of droplets of water emerging from the surface, we can assume that surface tension forces are dominant and that the Weber number is low. However, if we observe intricate fluid-air surfaces and nonspherical droplets we can assume that a high Weber number is in action.

10.3.9 Prandtl Number: Relative Thickness of Thermal and Velocity Boundary Layers

The Prandtl number Pr is defined as:

$$Pr = \frac{\eta C_p}{k},$$

where C_p is the specific heat at constant pressure and k is the coefficient of thermal conduction. It is the ratio of momentum diffusivity (kinematic viscosity) to thermal diffusivity, and can be related to the thickness of the thermal and velocity boundary layers (it is in fact the ratio of these two thicknesses). When $Pr = 1$, these boundary layers are equal and coincide. When Pr is small, heat diffuses very quickly compared to the fluid velocity (momentum), so that the thickness of the thermal boundary layer is much bigger than the velocity boundary.

10.4 Applying Nondimensional Parameters to Practical Flow Problems

As an instructive exercise we will predict the flow behaviour of water vapour and a liquid in a channel of width 1 mm and height 0.05 mm (50 μm) at STP. We will specify a flow rate of 0.05 $\mu\text{L s}^{-1}$.

10.4.1 Channel Filled with Water Vapour

To decide to what extent we can assume a continuum model, we calculate the Knudsen number:

$$Kn = L_{mfp}/L, \quad \text{where } L_{mfp} = k T / (\sqrt{2\pi} P d^2),$$

$k = 1.3806 \times 10^{-14}$ nJ/K; diameter of a water molecule $d = 0.25 \times 10^{-6}$ mm.

If we assume absolute pressure $P = 1.013 \times 10^5$ Pa, $T = 293$ K, then the mean free path is $L_{mfp} = 1.44 \times 10^{-4}$ mm.

By using the smallest dimension L , the channel height of 0.05 mm, gives $Kn = 0.0029$.

We have $Kn < 0.1$, so according to the discussion of Knudsen number ranges in 10.3.2 we can utilise the continuum approximation in this case – but will need to account for a finite slip between the water vapour and the channel walls.

10.4.2 Channel Filled with a Dilute Electrolyte at 293 K

We are now dealing with a liquid, so we approximate L_{mfp} with the molecular diameter d . This gives $Kn = 0.00025$, which allows both the continuum approximation and zero slip boundary conditions. Will diffusion be a major factor? We can answer this by calculating the Peclet number

$$Pe = vL/D.$$

The diffusion coefficient for NaCl in water is $D = 1.74 \times 10^{-3}$ mm²/s. Since the width is an order-of-magnitude larger than the height, the hydraulic diameter concept suggests a characteristic length that is approximately twice the smaller dimension, or $L = 0.1$ mm.

The bulk fluid velocity v is given by the ratio of the flow rate and the flow area, so $v = 10^{-3} \text{ m s}^{-1}$. Thus, the Pecllet number is $Pe = 57.5$, suggesting that diffusion is an effective mass transport mechanism in this case. (Indeed, a diffusion front of NaCl would cross the channel height in less than 2 seconds.)

The Reynolds number will show if the flow is laminar, turbulent, or possibly in transition. Assuming essentially water properties, the fluid density is $\rho = 0.001 \text{ g}/\mu\text{L}$ and dynamic viscosity is $\eta = 0.001 \text{ Pa s}$.

Thus, $Re = \rho v L/\eta = 0.1$, and we certainly have laminar flow. Two streams carrying different solutes would flow side-by-side in this channel with their components mixing only by diffusion. The Bond number is given by:

$$Bo = \rho g L^2 / T_s.$$

We have $g = 9810 \text{ mm/s}^2$ and surface tension for water $T_s = 72 \mu\text{N/mm}$, giving $Bo = 0.0014$. This suggests that during filling of the channel, gravity will be a weak mechanism compared to the capillary effect.

These examples demonstrate that before we perform any complex analysis or sophisticated numerical simulation of the flow we can predict much about the characteristics of microfluidic flow. In these examples we know that we have a highly laminar flow in which solutes mix only by diffusion. The wetting of the channel will depend on surface energy, and not on elevation change for example. In addition, we also know how changing channel dimensions or fluid properties will impact this flow behaviour.

10.5 Characteristic Time Scales

Various characteristic time scales may be defined for microfluidic systems, and the following four are commonly used:

10.5.1 Convective Time Scale

This is the time taken for a perturbation to propagate in a liquid:

$$t_c = L/v,$$

where L is the characteristic dimension and v is the velocity of the liquid. If the fluid is responding to shear stress (see Section 9.3.5 in Chapter 9) the convective time scale is also given by the reciprocal of the shear rate:

$$t_c = 1/(dv/dx).$$

10.5.2 Diffusion Time Scale

This is the time taken for a physical perturbation to propagate (diffuse) in the fluid:

$$t_D = L^2/v,$$

where $v = \eta/\rho$ is the kinematic viscosity (η is the dynamic viscosity and ρ the density of the fluid).

10.5.3 Capillary Time Scale

This is the time taken for a physically perturbed interface to regain its original shape against viscous opposition:

$$t_T = (\eta L)/T_s,$$

where T_s is the surface tension.

10.5.4 Rayleigh Time Scale

This is the time scale of a physically perturbed interface induced by inertial and surface tension forces:

$$t_R = \sqrt{(\rho L^3 / T_s)}.$$

10.6 Applying Micro- and Nano-Physics to the Design of Microdevices

Researchers developing microfluidic devices are frequently confronted by issues that are directly related to the fundamental physics that apply at the micro- and nanoscales. Some practical examples include:

- Fluids that are brought together in a microfluidic circuit do not mix easily. Any mixing that does occur arises mainly from the diffusion of solutes across the boundaries between separate laminar flows.
- Solute particles that are heavier than the surrounding fluid settle to the channel bottom very quickly.
- The devices tend to have a very large surface-to-volume ratio, thus presenting the opportunity for particles to stick to a large surface area for a given volume.
- A small drop of fluid placed in the inlet of a microfluidic device can evaporate very rapidly.
- In microdevices, capillary force and surface energy effects are large forces compared to gravity. Depending on their direction and nature, they may move fluids upwards and sideways, or even block downward fluid movement.
- Small fluid volumes will almost immediately take on the temperature of the environment, and cool down or heat up very quickly.
- Liquids flowing in microchannels often do not have sharply defined frontal and trailing boundaries. This is caused by capillary action, and can cause spatial dispersion during fluid flow that hinder the ability to accurately define the timing of reactions or analyses.

However, at the same time these seemingly undesirable effects can be converted into extremely powerful tools: Examples include:

- Flow is usually laminar, allowing the parallel flow of several layers of fluid. This can be turned to advantage in the design of separation and detection devices based on laminar fluid diffusion interfaces.
- At microdimensions, diffusion becomes a viable approach to move particles, mix fluids, and control reaction rates. Typical low molecular weight biomolecules can diffuse more

than 10^{-3} m s^{-1} at 298 K in aqueous solutions. This allows the establishment of controlled concentration gradients in flowing systems, as well as complete equilibration of the molecule across a 100 μm channel in less than one minute, for example.

- Unaided by centrifugation, sedimentation becomes a viable means to separate dispersed particles by density across small channel dimensions. For example, red blood cells will sediment in a 100 μm deep channel in about 1 minute and generate a 50 μm layer of plasma in the process.
- In microchannels, the diffusion distance can be made extremely small, particularly if fluid streams are hydrodynamically focused. Thus, diffusion-controlled chemical reactions occur more rapidly than in comparable macroscopic reaction vessels. For example, microfluidic immunoassays can be completed in less than 25 seconds, as opposed to more typical immunoassay reaction times of 10 minutes or more.
- Evaporation of small quantities of fluids can be extremely rapid because of a typically large surface-to volume ratio. This effect can be used for the concentration of sample particles.
- Active particle transportation and separation methods, such as capillary electrophoresis, show greatly enhanced separation performance in small channels.

Other positive characteristics of microfluidic devices that are derived from economics, convenience, and safety include:

- Plastic microfluidic structures can be mass-produced at very low unit cost, allowing them to be made disposable.
- Microfluidic devices are amenable to high throughput by processing multiple samples and assays in parallel. For example, massively parallel processing can speed DNA, RNA, protein, immunologic, and other tests to reduce time intervals for drug discovery and medical diagnosis.
- Microdevices require only small volumes of sample and reagents, and produce only small amounts of waste, which can often be contained within the disposable device.
- The small scale of the various components of microfluidic systems allows them to be integrated into total-analysis systems (Micro-TAS) capable of handling all steps of the analysis on-chip, from sampling, sample processing, separation and detection steps to waste handling. This integration also makes complex analyses potentially simpler and safer to perform.
- It is possible to design passive fluidic devices that utilise inherent properties of the fluid and its microenvironment (capillary force, evaporation, wicking, heat transfer, diffusion, etc.) for fluid movement, mixing, heating, cooling, and catalysing chemical reactions. Thus, disposable standalone devices can be designed that require no external power source or instrumentation, yet still perform many, if not all, of the functions typically associated with full-scale automated chemical analysis devices containing pumps, mixers, heat elements, readout signals utilising fluorescence, for example. The phenomenon of ‘wicking’ occurs when a fluid and channel wall have a higher degree of affinity, as for example an aqueous solution and a hydrophilic surface. This effect can be used to modify fluid frontal boundaries through surface treatments such as the hydrophobic coating of a channel wall.

Problems

- 10.1. Determine the dimensions of the following quantities:
 (a) $\text{Sin}(30^\circ)$, (b) $\pi \cdot \log_{10} 2$, (c) Thermal conductance (defined as the quantity of heat that passes in unit time through a plate of particular area and thickness when its opposite faces differ in temperature by one kelvin).
- 10.2. For $y = e^{kt}$, where t is time, what are the dimensions of the factor 'k'?
- 10.3. Determine if the following equations are dimensionally correct:
 (a) $P = \sqrt{\rho g d}$
 where P is pressure, ρ is density, g is gravitational acceleration, d is depth below surface fluid level.
- (b) $\log_e(N_1/N_2) = Vgd(\rho_1 - \rho_2)/kT$
 where N_1 and N_2 are number of particles, V is volume, g is gravitational acceleration, d is distance, ρ_1 and ρ_2 are densities, k is Boltzmann's constant (Joules per Kelvin), T is absolute temperature.
- 10.4. Calculate the hydraulic diameter for the two quite common channel cross-sections shown in Figure 10.11.
- 10.5. The drag force on a spherical particle of radius r immersed in a fluid of flow velocity v and dynamic viscosity η is given by Stokes' Law as:

$$F = 6\pi r v \eta.$$

Use this equation to determine the dimensions of dynamic viscosity.

- 10.6. The force on a body immersed in a flowing fluid depends on a channel dimension L , as well as the fluid density ρ , viscosity η and velocity v . Express the relationship between these parameters in the form:

$$\pi_1 = f(\pi_2, \pi_3, \dots).$$

One of these π parameters takes the form of a well-known dimensionless parameter. Name this parameter and explain how its magnitude can be helpful in fluid flow analysis.

- 10.7. The drag force acting on a rough sphere is a function of its diameter D , the average depth k of grooves made into its surface, the density ρ , viscosity η and velocity v of the fluid. Express the relationship between these parameters in the form:

$$\pi_1 = f(\pi_2, \pi_3, \dots).$$

- 10.8. A flowing fluid will exert a force on any object that it encounters. Assume, under the conditions of interest, that this drag force F depends on the speed v of the fluid relative

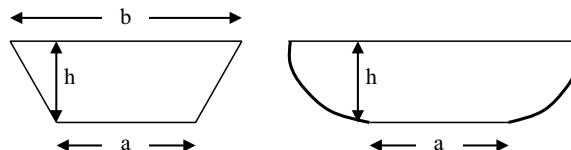


Figure 10.11 Self-study Problem 10.4: Calculate the hydraulic diameter for the trapezoidal (left) and rounded trapezoidal (right) channel cross-sections.

to the object, the fluid density ρ , the viscosity η of the fluid, and the size of the body (expressed in terms of its forward-facing cross-sectional area A).

- (a) How many dimensionless parameters can be formed to analyse this effect?
- (b) Derive these dimensionless parameters.
- (c) Express your result in the form $\pi_1 = f(\pi_2, \dots, \pi_{n-k})$
- (d) Experiments are performed, using different fluids, on a spherical object as a function of its diameter, and for different fluid speeds, viscosity and density. The data is plotted as π_1 vs π_2 . How many different curves appear in such a plot?
- (e) What do we gain by using Dimensionless Analysis?

References

- [1] Bridgman, P.W. (1931) *Dimensional Analysis*, 2nd edn, Yale University Press, New Haven.
- [2] Buckingham, E. (1914) On Physically Similar Systems; Illustrations of the Use of Dimensional Analysis. *Physical Review*, **4**, 345–376.
- [3] Sonin, A.A. (2001) *The Physical Basis of Dimensional Analysis*, 2nd edn, MIT, Cambridge, MA.
- [4] Taylor, G. (1950) The Formation of a Blast Wave by a Very Intense Explosion. II. The Atomic Explosion of 1945. *Proceedings of the Royal Society of London*, **A201**, 175–186.
- [5] Reynolds, O. (1883) An experimental investigation of the circumstances which determine whether the motion of water in parallel channels shall be direct or sinuous and of the law of resistance in parallel channels. *Philosophical Transactions of the Royal Society*, **174**, 935–982.
- [6] Wu, P. and Little, W.A. (1983) Measurement of friction factors for the flow of gases in very fine channels used for micro-miniature Joule-Thomson refrigerators. *Cryogenics*, **23**, 273–277.

Appendix A

SI Prefixes

Prefix	Symbol	Factor	Prefix	Symbol	Factor
yotta	Y	10^{24}	deci	d	10^{-1}
zetta	Z	10^{21}	centi	c	10^{-2}
exa	E	10^{18}	milli	m	10^{-3}
peta	P	10^{15}	micro	μ	10^{-6}
tera	T	10^{12}	nano	n	10^{-9}
giga	G	10^9	pico	p	10^{-12}
mega	M	10^6	femto	f	10^{-15}
kilo	k	10^3	atto	a	10^{-18}
hecto	h	10^2	zepto	z	10^{-21}
deca	da	10^1	yacto	y	10^{-24}

Appendix B

Values of Fundamental Physical Constants

Quantity	Symbol	Value
Elementary charge	e	$1.60217(6) \times 10^{-19} \text{ C}$
Electric constant	ϵ_0	$8.85418(8) \times 10^{-12} \text{ F m}^{-1}$
Magnetic constant	μ_0	$4\pi \times 10^{-7} \text{ N A}^{-2}$
Speed of light in vacuum	c	$2.99792(4) \times 10^8 \text{ m s}^{-1}$
Planck constant	h	$6.62606(9) \times 10^{-34} \text{ J s}$
Rest mass of electron	m_e	$9.10938(2) \times 10^{-31} \text{ kg}$
Rest mass of proton	m_p	$1.67262(2) \times 10^{-27} \text{ kg}$
Boltzmann constant	k	$1.60217(6) \times 10^{-23} \text{ J K}^{-1}$
Avogadro constant	N_A	$6.02214(2) \times 10^{23} \text{ mol}^{-1}$
Faraday constant	F	$9.64853(4) \times 10^4 \text{ C mol}^{-1}$
Molar gas constant	R	$8.31447(2) \text{ J K}^{-1} \text{ mol}^{-1}$

Appendix C

Model Answers for Self-study Problems

C.1 Chapter 1

1. A chemical reaction reaches equilibrium when the rates of the forward and reverse reactions become equal and the concentrations of the reactants and products do not change with time.
2. We will assume the following molecular mass values:

$$\text{H} = 1; \text{O} = 16; \text{Na} = 23; \text{Mg} = 24.3; \text{K} = 39.1; \text{Cl} = 35.5.$$

- (a) Dissolve 0.59 gm NaCl into 100 mL water will give 0.1 M NaCl.
- (b) Dissolve 3.73 gm KCl in 1 liter of water, take 10 mL of this solution.
- (c) 9.53 gm of anhydrous MgCl₂ dissolved into 100 mL will give 1 M.
- (d) 14.93 gm of the MgCl₂ hydrate dissolved into 100 mL will give 1 M.
- (e) 0.373 gm of KCl dissolved into 100 mL of water gives 20 mM. Pipette 10 μL of this solution into 1 liter of water to give 2 μM KCl.

For practical preparations that involve the weighing of chemicals and the determination and dispensing of small solution volumes, the following guidelines are useful:

- A good, calibrated, analytical balance can weigh a 100 gm sample to an accuracy of ± 0.1 mg. (The more common top-loading balance can weigh 100 gm to a repeatability of ± 1 mg.)
- Class A calibrated glass cylinders can be used to an accuracy of $\sim 1\%$ (e.g. a 100 mL cylinder has graduations of 1 mL).
- Pipettes can dispense volumes ranging from 1 μL to 10 mL to an accuracy of $\sim 1\%$ (and better).

3. Two key properties of water that make it a good solvent for ionic and polar molecules are the polar form of the water molecule and its ability to form networks of hydrogen-bonds. This gives water a large relative permittivity (dielectric constant) value and excellent properties as a solvent for ionic salts because it can weaken the electrostatic interactions

involved in ionic bonds to the point where the ionic crystal lattice dissociates into free ions. The ability of water to dissolve nonionic substances that possess polar chemical bonds (e.g. sugars and alcohol molecules) arises from it being able to form hydrogen bonds with such molecules. Organic molecules that are nonpolar and nonionic are not soluble in water.

4. A hydrogen bond is formed when an electropositive hydrogen atom is partially shared by two electronegative atoms, which for the case of water are oxygen atoms. The hydrogen in an H-bond can be thought of as a proton that has partially dissociated from a donor atom, allowing it to be shared with a second (acceptor) atom. An H-bond is strongest when a straight line can be drawn through the centres of its three participating atoms ($\text{O}-\text{H}\cdots\text{O}$). Common H-bonds in protein structures are of the form $\text{N}-\text{H}\cdots\text{O}$ or $\text{N}-\text{H}\cdots\text{N}$, and water can weaken such bonds by forming competing H-bonds with the N or O atoms in these bonds.

Without its hydrogen-bond structure water would exist as a gas at room temperature rather than as a liquid.

5. On the assumption of 100% dissociation, then:
 - (a) $\text{pH} = -\log(1) = \text{pH } 0$. (b) $\text{pH} = -\log(10^{-6}) = \text{pH } 6$. (c) 10 mM NaOH dissociates to give $[\text{OH}^-] = 10^{-2}\text{M}$. At room temperature $[\text{H}^+][\text{OH}^-] = 10^{-14}$, to give $[\text{H}^+] = 10^{-12}$ and a pH value of pH 12.
6. A pH value of pH 2.4 corresponds to a proton concentration

$$[\text{H}^+] = \text{antilog}(-2.4) = 3.98 \times 10^{-3} \approx 4 \times 10^{-3} \text{ M.}$$

If fully dissociated $[\text{H}^+]$ should equal 1 M. A concentration of 4×10^{-3} M therefore indicates that the acid is only 0.4% dissociated. Acetic acid is thus a weak acid.

7. We can calculate the $\text{p}K_a$ value using the Henderson-Hasselbalch equation (1.11):

$$\text{p}K_a = \text{pH} - \log \frac{[\text{A}^-]}{[\text{HA}]} = 4.8 - \log \frac{[0.087]}{[0.01]} = 4.8 - 0.94 = 3.86.$$

8. We can define a base concentration factor α as:

$$\alpha = \frac{[\text{A}^-]}{[\text{A}^-] + [\text{HA}]} = \frac{[\text{base}]}{([\text{base}] + [\text{acid}])}.$$

Writing the Henderson-Hasselbalch equation in terms of α and differentiating, we obtain:

$$d\text{pH} = d\text{p}K_a + d \log \frac{\alpha}{1 - \alpha} = \frac{1}{2.3} \ln \frac{\alpha}{1 - \alpha} = \frac{d\alpha}{2.3\alpha(1 - \alpha)}.$$

$$\text{Thus: } \frac{d\text{pH}}{d\alpha} = \frac{1}{2.3\alpha(1 - \alpha)}.$$

The product $\alpha(1 - \alpha)$ is a maximum, and thus $d\text{pH}/d\alpha$ a minimum, at $\alpha = 0.5$. This corresponds to the situation $[\text{A}^-] = [\text{HA}]$, which from the Henderson-Hasselbalch equation corresponds to a value of pH equal to the $\text{p}K_a$. The buffering capacity is thus a maximum

at the pK_a . As a practical example, take the case of a solution of acetic or lactic acid (pK_a values of 4.76 and 3.86, respectively, and conjugate bases of acetate and lactate, respectively) to which is added a strong base such as sodium hydroxide. The resulting plots of pH against added hydroxide will be of the forms of the curves shown in Figure 1.5. At values of pH removed from near the pK_a value, the pH changes very rapidly with additions of small amounts of the hydroxide. However, near the pK_a the addition of the hydroxide base results in only small changes in pH. In other words, solutions of acetic and lactic acid are buffered at values of pH near to their pK_a values.

C.2 Chapter 4

1. (a) The dye absorbs at wavelengths extending from red to yellow and extending into green, and also from violet into the ultraviolet. The dye therefore reflects and appears blue. (In fact the dye is methylene blue).
- (b) The molar extinction coefficient at 290 nm is $\sim 4 \times 10^4 \text{ M}^{-1} \text{ cm}^{-1}$. The absorbance at this wavelength is given by the Beer-Lambert law:

$$A = \varepsilon l[C] = 4 \times 10^4 \times 1.0[10 \times 10^{-6}] = 0.4.$$

i. % Transmittance: $T\% = 100 \text{ antilog}(-A) = 39.8\%$.

Therefore, percentage of incident radiation absorbed at 290 nm = 60.2%.

ii. % Transmittance at 420 nm $\approx 100\%$.

Therefore, percentage of incident radiation absorbed at 420 nm $\approx 0\%$.

iii. The molar extinction coefficient at 670 nm is $\sim 8 \times 10^4 \text{ M}^{-1} \text{ cm}^{-1}$. The corresponding absorbance = 0.8.

% Transmittance: $T\% = 100 \text{ antilog}(-A) = 15.85\%$.

Therefore, percentage of incident radiation absorbed at 670 nm = 84.15%.

iv. From the definition $\varepsilon = 2.61 \times 10^{20} \sigma$, the effective cross-sectional area σ of the methylene blue molecule at 670 nm = $3 \times 10^{-16} \text{ cm}^2$.

2. (a) $d = 52.66 \text{ nm}$.

(b) $0.37 E_o$; $0.003 E_o$.

(c) Typically, the binding of an analyte to a biosensing agent (immobilised on the surface of a prism, glass slide or nanoparticle) is detected as the induced fluorescence of the analyte.

3. Absorbance $A = -\log T = -\log(0.3) = 0.52$.

From the Beer-Lambert law:

$$\varepsilon = A/(l[C]) = 0.52/(0.5 \times 0.1) = 10.4 \text{ M}^{-1} \text{ cm}^{-1}.$$

From the definition:

$$\varepsilon = 2.61 \times 10^{20} \sigma$$

the effective cross-sectional area σ for the Cu^{2+} ion = $4 \times 10^{-20} \text{ cm}^2$.

4. If A and AR have the same effective cross-sectional areas at resonance, then the ratio of their peak absorptions at wavelengths λ_{AR} and λ_R is:

$$\frac{a(\lambda_{AR})}{a(\lambda_R)} = \frac{[AR]}{[R]}.$$

Substituting this equality into Equation (4.9) we have:

$$[A] = \frac{1}{K_{eq}} \frac{a(\lambda_{AR})}{a(\lambda_R)}.$$

From Figure 4.22 the ratio of the two absorptions $\approx 0.8/0.3 \approx 2.7$.

With $K_{eq} = 2.2 \times 10^2 \text{ M}^{-1}$, the corresponding analyte concentration $[A] = 12.2 \text{ mM}$.

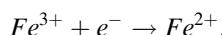
C.3 Chapter 5

1. (a) E° : Standard reduction potential; n : number of electrons involved in the redox reaction; α_{Ox} and α_R : activities (equal to the concentrations if low) of the oxidised and reduced species.
 (b) The factor corresponds to a voltage and has units of volts. Designating $n = 1$ indicates that the redox reaction involves a single electron transfer in the reaction.
2. The Nernst Equation applies to the reaction:



in which Ox is the oxidised species and R the reduced species of the redox couple.

Thus, for the reaction:



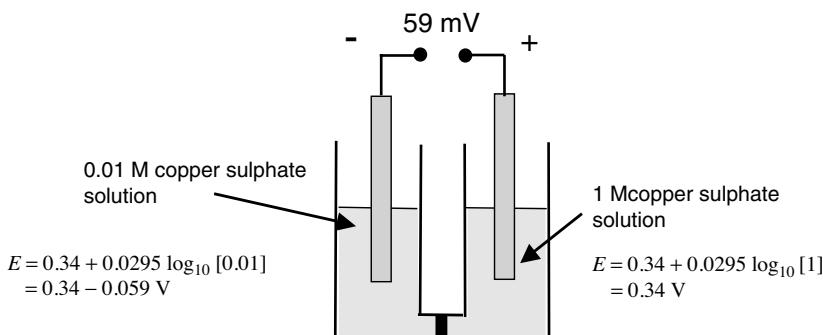
equations (c) and (d) are correct.

3.

$$E = 0.34 + 0.0295 \log_{10} \frac{[Cu^{2+}]}{[Cu]} = 0.34 + 0.0295 \log_{10} [Cu^{2+}]$$

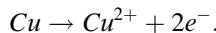
(the copper electrode has a constant concentration (activity) of 1).

4.



- (a) From the two relevant Nernst equations, we obtain a potential difference of 59 mV, with the polarity as shown above.
- (b) In the $\frac{1}{2}\text{-cell}$ with 1M copper sulphate the reaction direction is: $Cu^{2+} + 2e^- \rightarrow Cu$. Copper ions deposit on the electrode and acquire electrons (are reduced). Copper atoms are deposited onto the electrode surface.

In the $\frac{1}{2}$ -cell with 0.01M copper sulphate the reaction direction is reversed:



Copper atoms dissolve into solution and give up electrons (are oxidised).

5. A reference electrode should have the following properties:

- It should be easy to make,
- It must provide a stable potential,
- It must be polarisable (ideally).
- It should have a low temperature coefficient.

6. (a) The equation: $R_{ct} = \frac{RT}{nFI_o}$ contains the following parameters:

R : Universal gas constant ($8.31 \text{ J K}^{-1} \text{ mol}^{-1}$);

T : Absolute temperature (Kelvin);

F : Faraday Constant ($9.648 \times 10^4 \text{ C mol}^{-1}$);

n : Number of electrons involved in the charge transfer reaction;

I_o : The net charge transfer exchange current.

The derived units for R_{ct} are [$\text{J A}^{-2} \text{ s}^{-1}$] = [V.A.A^{-2}] = [V.A^{-1}] = Ohms.

- (b) R_{ct} is small when the exchange current I_o is large, corresponding to a high charge transfer rate.
- (c) Apply a small sinusoidal voltage ($\leq 5 \text{ mV pk.}$) across the working and counter electrode and measure the working electrode current. To ensure that there is a near linear current-voltage relationship, the area of the counter electrode should also be much larger than that of the working electrode.
- (d) 126 ohms.
7. (a) $2.6 \text{ k}\Omega$ and 20Ω - both obtained by estimating the low frequency intercept of the semicircular arc on the Z' axis (x-axis).
- (b) Apart from the impedance to current flow related to charge transfer at the electrode-solution interface, there is also effective impedance related to diffusion-controlled mass transfer. The left-hand Nyquist plot is indicative of an electrode reaction that is dominated by the kinetic control (electron tunnelling) of the overall charge transfer reaction at all the frequencies of measurement. For the right-hand plot we have a mass transport controlled impedance at the lower frequencies (the straight line portion of the plot) and kinetic control at higher frequencies.

C.4 Chapter 7

1. (a) The **accuracy** of a measurement defines how close the measured value is to the true value, while the **precision** describes how the measured value changes with repeated measurements. You could just use a graph of the distribution of the results showing the accuracy (error in the mean) and the precision (spread of results) but a nice analogy involves a comparison with hits on a target in archery as shown in Figure C.1

In the figure on the left the hits are accurate but imprecise, that is, they are centred on the bullseye but are well spread out. In the figure on the right the hits are more precise but less accurate, that is, they are grouped closely together but are not well aimed at the centre of the target.

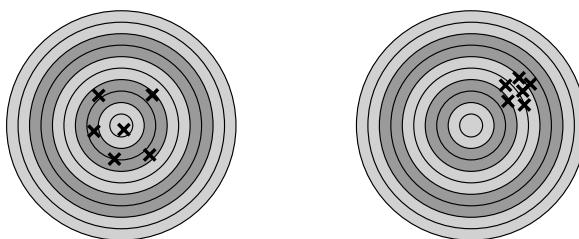


Figure C.1 Illustration of the accuracy and precision of a measurement.

- (b) Determining the accuracy and precision of a sensor will both require multiple measurements, typically made of a known ‘standard’ which will obviously change depending on what the sensor is measuring. For example with a pH sensor you would need one or more buffer solutions of known pH. Measurements need to be made with the same conditions and you need to be reasonably sure that the parameter being measured is not changing with time. Ideally the measurements would be made quickly to avoid any problems with drift in the sensor. The actual value of the standard isn’t really important for measurement of precision, you just need to make enough measurements of an otherwise unchanging variable to determine the variance or standard deviation of the results. Determination of the accuracy will also require multiple measurements, unless of course you know for sure that the measurement is extremely precise, but the important figure here is the mean value of the parameter. The accuracy will be defined by the offset between the known value of the standard and the mean of the set of measurements. Of course these measurements will often only give you an idea of the accuracy and precision at one measurement point and it would be a good idea to have two or more standards to measure which are spread out along the range of values you expect to be using the sensor to measure.
- (c) This question could be more properly stated to ask how you would deal with these issues in a real sensor. Basically if the precision is poor then you need to make multiple measurements and take the average of these as your answer. The way to deal with an offset or inaccuracy is to measure this offset against a standard (or multiple standards) and calibrate the sensor instrumentation to remove this. If the offset is not a constant this might need to be done at several points.
2. (a) The parameters of the pH electrode equation are as follows:
- E^0 - Characteristic potential for the particular combination of glass electrode and reference electrode used
 - R - Universal gas constant = $8.314472(15) \text{ J K}^{-1} \text{ mol}^{-1}$
 - T - Absolute temperature (K)
 - F - Faraday constant = $9.64853399(24) \times 10^4 \text{ C mol}^{-1}$
- Supplementary question: Where does the figure ‘2.303’ come from and can you simplify the equation for $T = 298 \text{ K}$ (25°C)?

(b) See Figure C.2:

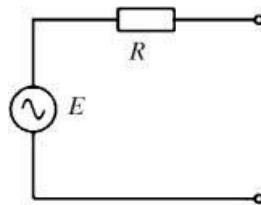


Figure C.2 Equivalent circuit of a sensor with voltage output E and output resistance R .

(c) In the linear region of Figure 7.32 the output voltage varies by about 0.75 V over a pH range of 13, with a negative slope. Therefore the sensitivity is around -0.06 V pH^{-1} (Nernst type response). In order to get an output resolution of 1 V pH^{-1} we need an amplification factor G of -16.67 .

Because the output resistance of the pH sensor is very high, there is a problem with using an inverting amplifier which has a finite input impedance. One solution is to use a voltage follower or buffer between the sensor and the inverting amplifier as shown in Figure C.3.

The gain of this circuit is given by the following equation:

$$G = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{-R_f}{R_{\text{in}}}$$

For a gain of -16.67 using standard resistor values you could use:

$$R_f = 20 \text{ k}\Omega \text{ and } R_{\text{in}} = 1.2 \text{ k}\Omega$$

but there are obviously many other possible solutions.

For a pH of 7 we can estimate the output voltage of the sensor to be around -0.42 V . The output of the amplifier will therefore be approximately 7 V which is a good confirmation that it's operating correctly.

(d) The effective error in the output voltage will be equal to the input bias current (I_b) multiplied by the feedback resistance R_f . So that:

$$V_{\text{out}} = GV_{\text{in}} - I_b R_f$$

Therefore the error will be 20 nV for 1 pA bias current and $20 \mu\text{V}$ for 1 nA bias current.

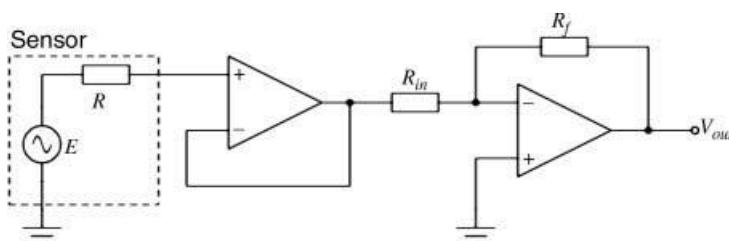


Figure C.3 Buffered inverting amplifier for a high output impedance sensor.

3. The current required to charge a $10 \mu\text{F}$ capacitor by 1 V in 10 ms is:

$$i = C \frac{dV}{dt} = 1 \text{ A}$$

Then, in order to drive 1 A through a 100Ω resistor, a voltage of 100 V on the output of the amplifier connected to the counter electrode will be required. These are probably unrealistic numbers.

4. Using a Wheatstone bridge with the sensor taking the place of R_1 , balance the bridge by making all the other resistors $1 \text{k}\Omega$. If we use a 5V supply for V_s and ∂R_1 is 10Ω then V_o will be $\sim 12.5 \text{ mV}$ using equation (7.26). This can be amplified using a buffered differential amplifier with a gain $G = 400$. Alternatively the gain could be split between two amplifier stages or an instrumentation amplifier could be used. The output of the Wheatstone bridge may be positive or negative depending on which way round the differential amplifier is connected.
5. R_{ct} can be estimated to be around $27 \text{k}\Omega$, then we can calculate the double layer capacitance using $\omega = 2\pi f = 1/R_{ct}C_{dl}$. It should be around 5.4nF .

C.5 Chapter 8

1. (a) The other resistors used in the bridge will also have some temperature dependence which will be confounded with that of the RTD if care is not taken. It would be very hard to control temperatures in an implanted device like that.
- (b) It is unlikely but if the temperature dependence of the other resistors can be made to be much much smaller than the measured device then this could almost be ignored.
- (c) The resistance of the electrical connections to the RTD become a problem. These could be balanced at one particular temperature but they may also have some temperature dependence that will cause an error in the measurement.
- (d) Current and voltage are separated in a four-terminal Kelvin measurement meaning that the resistance of the electrical connections should not impact on the measured resistance.
2. (a) $\Delta R = GF\epsilon R_G = 2 \times 0.1 \times 1000 = 200 \Omega$.
- (b) This is obviously quite large and reflects the rather unrealistically high strain value.
- (c) The gauges will be applied to the structure so that two are in compression (R_{C1}, R_{C2}) and two are in tension (R_{T1}, R_{T2}) these will be arranged so that they are opposite each other as in figure C.4.
- (d) Assuming the same gauge factor as in (a) ΔR will be 20Ω . Then on one side of the bridge the divider voltage will be:

$$V_1 = V_B \frac{R_{C1}}{R_{C1} + R_{T1}} = 5 \times \frac{980}{980 + 1020} = 2.45 \text{ V.}$$

While on the other side it will be:

$$V_2 = V_B \frac{R_{T2}}{R_{C2} + R_{T2}} = 5 \times \frac{1020}{980 + 1020} = 2.55 \text{ V.}$$

Therefore V_{out} will be 100mV.

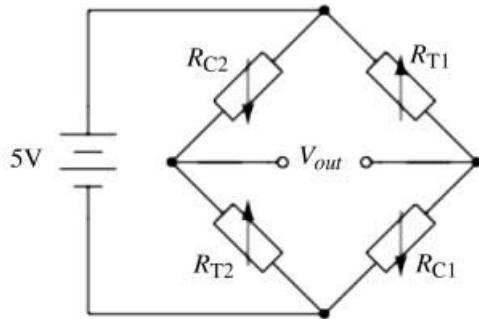


Figure C.4 Wheatstone bridge arrangement using 4 strain gauges.

- (e) If it is possible then attach one in compression (R_{C1}) and one in tension (R_{T1}) while the other two resistances will be standard resistors with values chosen to balance the unstrained gauges. The sensitivity will be halved so that the output for a strain of 1% will be 50mV. Alternatively if the gauges can only be used in compression then use R_{C1} and R_{C2} as in figure A8.1. Again the other two resistors will be unstrained and the output sensitivity will be halved compared to the full bridge implementation.
 - (f) A variable resistor could be added to one arm of the circuit. This could be adjusted to give a zero output when the gauges are unstrained.
3. (a) The equations for the output voltages are as follows:

$$V_1 = V \frac{C_2}{C_2 + C_4} = V \frac{\frac{C_0}{x}}{1 + \frac{d}{\frac{C_0}{x}}} = V \frac{\frac{C_0}{x}}{1 + \frac{d}{\frac{C_0}{x}} + \frac{C_0}{1 - \frac{x}{d}}}$$

$$V_2 = V \frac{C_1}{C_1 + C_3} = V \frac{\frac{C_0}{1 - \frac{x}{d}}}{1 - \frac{d}{\frac{C_0}{1 - \frac{x}{d}}} + \frac{C_0}{1 + \frac{x}{d}}}$$

$$V_{out} = \frac{V}{\frac{C_0}{1 - \frac{x}{d}} + \frac{C_0}{1 + \frac{x}{d}}} \left(\frac{C_0}{1 - \frac{x}{d}} - \frac{C_0}{1 + \frac{x}{d}} \right)$$

$$V_{out} = V \left(\frac{1 - \left(\frac{x}{d}\right)^2}{2} \right) \left(\frac{2 \frac{x}{d}}{1 - \left(\frac{x}{d}\right)^2} \right) = V \frac{x}{d}$$

- (b) As the output is independent of C_0 the output amplitude will be 0.05V
4. Figures C.5 and C.6 show the two different switch positions. In phase $\Phi 1$ (Figure C.5) C_1 charges to V_{in} so $Q_{C1} = V_{in} C_1$. Meanwhile C_2 is discharged. When the switches change to

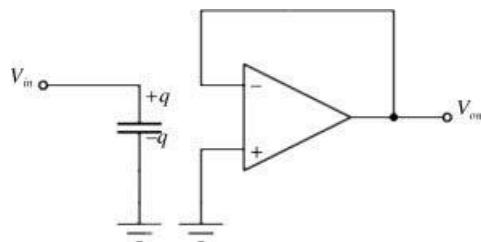


Figure C.5 SC circuit in phase $\Phi 1$.

phase $\Phi 2$ (Figure C.6) C_1 will discharge and the current flow will charge up C_2 but the direction of this will mean that:

$$V_{out} = \frac{Q_{C_2}}{C_2} = \frac{-Q_{C_1}}{C_2} = -V_{in} \frac{C_1}{C_2}.$$

So the transfer function is:

$$\frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2}$$

and this is an **Inverting** switched capacitor amplifier.

5. (a) Control voltage V_{in} is applied and the resulting ion current I_{ch} flows through R_f assuming the op-amp has very low input current. So the positive input to the diff. amplifier will be $V_+ = V_{in} - I_{ch}R_f$. The negative input of the diff. amplifier is connected to V_{in} so the output will be:

$$V_{out} = V_+ - V_- = V_{in} - I_{ch}R_f - V_{in} = -I_{ch}R_f$$

- (b) R_f needs to be very large, for example a $G\Omega$ resistor would give output voltages in the mV range. Possible problems may include the time constant as a large resistor like that typically has a parallel parasitic capacitance.
(c) Potassium channels are opened by applying a negative voltage step to simulate what happens when the membrane depolarises during an action potential event.

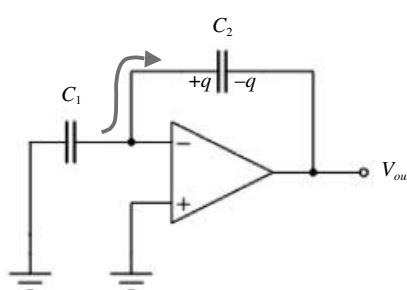


Figure C.6 SC circuit in phase $\Phi 2$.

- (d) When it opens the K^+ ions will flow out of the cell, which is equivalent to conventional current flowing out of the cell and through R_f . The output of the differential amplifier will be a negative voltage that is proportional to this current.

C.6 Chapter 9

- $R = 10 \text{ kPa}/(10^{-4}/60) = 6 \times 10^9 \text{ Pa L}^{-1}\text{s} = 6 \times 10^{12} \text{ Pa m}^{-3} \text{ s}$ (units of Pressure/Volumetric flow)
- This channel has a high aspect ratio (i.e. width $w \gg$ height h). The appropriate equation to calculate the fluidic resistance R of the rectangular microchannel is:

$$R = \frac{12 \eta L}{wh^3}.$$

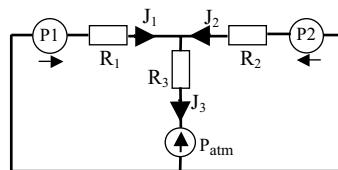
- (a) For a channel of width 100 μm , height 10 μm , and length L of 2 cm, filled with a solution of viscosity (η) 10^{-3} Pa s, the corresponding value for R is:

$$R = 12(10^{-3} \text{ Pa s})(2 \times 10^{-2} \text{ m})/[(10^{-4} \text{ m})(10^{-5} \text{ m})^3] = 24 \times 10^{14} \text{ Pa m}^{-3} \text{ s}.$$

- (b) The pressure drop required to give a flow of 0.1 $\mu\text{l/sec}$ is:

$$\Delta P = JR = (10^{-10} \text{ m}^3 \text{ s}^{-1})(24 \times 10^{14} \text{ Pa m}^{-3} \text{ s}) = 240 \text{ kPa}.$$

- (a) Apply Kirchhoff's rules to the electrical analogue of the fluidic Y-connector shown below:



$$J_3 = J_1 + J_2. \quad (1)$$

$$\Delta P_1 = J_1 R_1 + J_3 R_3. \quad (2)$$

(Define $\Delta P_1 = P_1 - P_{\text{atm}}$.)

$$\Delta P_2 = J_2 R_2 + J_3 R_3. \quad (3)$$

(Define $\Delta P_2 = P_2 - P_{\text{atm}}$.)

Substituting J_1 obtained from (2) and J_2 obtained from (3) into (1):

$$\text{we obtain } J_3 = (\Delta P_1 - R_3 J_3)/R_1 + (\Delta P_2 - R_3 J_3)/R_2$$

$$\text{to give } J_3 = (\Delta P_1 R_2 + \Delta P_2 R_1)/(R_1 R_2 + R_2 R_3 + R_1 R_3). \quad (4)$$

- (b) With $R_1 = R_2 = 10^{14} \text{ Pa m}^{-3} \text{ s}$, $R_3 = 2 \times 10^{15} \text{ Pa m}^{-3} \text{ s}$, $P_1 = 200 \text{ kPa}$, $P_2 = 150 \text{ kPa}$, and $P_{\text{atm}} = 100 \text{ kPa}$

$$\Delta P_1 = 100 \text{ kPa}, \Delta P_2 = 50 \text{ kPa}.$$

From Equation 4

$$\begin{aligned} J_3 &= (1.5 \times 10^{19} \text{ Pa}^2 \text{ m}^{-3} \text{ s}) / (4.1 \times 10^{29} \text{ Pa}^2 \text{ m}^{-6} \text{ s}^2) = 3.7 \times 10^{-11} \text{ m}^3 \text{ s}^{-1} \\ &= 37 \text{ nl s}^{-1}. \end{aligned}$$

4. The origin of the propulsive force is the surface tension (units of force per unit length) acting tangentially to the boat–water interface. The surface line element can be drawn as a closed loop around the boat. For a uniform surface tension the net surface tension force acting around this loop is zero. If, however, a surface tension gradient is produced a net force will act on the surface element and distort it through an induced flow of surface liquid. Surface tension gradients can arise from the presence of a surfactant, such as soap.
5. In a sufficiently narrow capillary of circular cross-section (radius r), the interface between a fluid and the capillary surface forms a meniscus (surface tension T) that is a portion of the surface of a sphere with radius R . The pressure jump ΔP across this surface is:

$$\Delta P = \frac{2T}{R}$$

The radius of the sphere will be a function only of the contact angle θ , which in turn depends on the exact properties of the fluid and the solid with which it is in contact:

$$R = \frac{r}{\cos \theta},$$

so that the pressure difference may be written as: $\Delta P = \frac{2T \cos \theta}{r}$.

To maintain hydrostatic equilibrium, the induced capillary pressure is balanced by a change in height, h , of the fluid (this height change can be positive or negative, depending on whether the contact angle is less or greater than 90°). At equilibrium:

$$\frac{2T \cos \theta}{r} = h \rho g \text{ to give } h = \frac{2T \cos \theta}{\rho g r}.$$

Using this formula, the height to which water will rise in a glass tube of radius $50 \mu\text{m}$, with $T = 7.3 \times 10^{-2} \text{ N m}^{-1}$, $\theta = 30^\circ$, $\rho = 1000 \text{ kg m}^{-3}$, $g = 9.8 \text{ m s}^{-2}$, is given by:

$$h = (2 \times 7.3 \times 10^{-2} \cos 30^\circ) / (10^3 \times 9.8 \times 5 \times 10^{-5}) = 0.26 \text{ m.}$$

C.7 Chapter 10

1. (a) & (b) dimensionless.
(b) Thermal conductivity k (Watt per Kelvin per metre) has units of $\text{M L t}^{-3} \theta^{-1}$.
Thermal conductivity and conductance are analogous to electrical conductivity σ and conductance G ($G = \sigma A/L$). Thermal conductance ($= kA/L$) has units of $\text{M L}^2 \text{ t}^{-3} \theta^{-1}$.
2. $[k] = \text{t}^{-1}$.

3. (a) Not dimensionally correct. $[P] = M \cdot L^{-1} \cdot t^{-2}$, whereas $[\sqrt{\rho g d}] = M^{1/2} \cdot L^{-1/2} \cdot t^{-1}$.

(b) Dimensionally correct.

$$[\text{Log}_e(N_1/N_2)] = 1 \text{ (i.e. dimensionless).}$$

$$[Vgd(\rho_1 - \rho_2)] = M \cdot L^2 \cdot t^{-2}; [kT] = M \cdot L^2 \cdot t^{-2} \therefore [Vgd(\rho_1 - \rho_2)/kT] = 1.$$

4.

$$(a) D_H = \frac{4 \times \text{Area}}{\text{Wetted Perimeter}} = \frac{4 \times \left(ah + \frac{(b-a)h}{2} \right)}{b + a + 2\sqrt{h^2 + [(b-a)/2]^2}} = \frac{2h(a+b)}{b + a + 2\sqrt{h^2 + [(b-a)/2]^2}}$$

$$(b) D_H = \frac{4(ah + \frac{\pi}{2}h^2)}{2a + 2h + \pi h} = \frac{h(2a + \pi h)}{(a + h + \pi h/2)}.$$

$$5. \eta = \frac{F}{6\pi rv} \quad [\eta] = \frac{MLt^{-2}}{L^2t^{-1}} = ML^{-1} t^{-1}.$$

6. Assume $F = f(L, v, \rho, \eta)$.

F	L	v	ρ	η
MLT^{-2}	L	LT^{-1}	ML^{-3}	$M L^{-1} T^{-1}$

$n = 5$ (dimensional parameters); $k = 3$ (dimensions).

From Buckingham's π -Theorem, number of dimensionless parameters: $n-k=2$.

Select independent 'repeating' variables: L , v , and ρ .

Combine these with remaining 'non-repeating' variables: F and η .

$$\pi_1 = \eta(L^a v^b \rho^c) \quad M^0 L^0 T^0 = (ML^{-1} T^{-1})(L)^a (LT^{-1})^b (ML^{-3})^c.$$

$$M : 0 = 1 + c. \quad \dots \rightarrow c = -1$$

$$L : 0 = -1 + a + b - 3c. \quad \dots \rightarrow a = -1$$

$$T : 0 = -1 - b. \quad \dots \rightarrow b = -1$$

$$\pi_1 = \frac{\eta}{Lv\rho} \quad \text{or} \quad \pi_1 = R = \frac{\rho v L}{\eta}.$$

that is Reynolds number (inertial to viscous forces) - important factor in all fluid flow problems – especially regarding laminar vs. turbulent flow considerations.

$$\pi_2 = F(L^a v^b \rho^c).$$

On solving, as above, by exponent method $\rightarrow \pi_2 = \frac{F}{L^2 v^2 \rho}$,

$$\pi_2 = f(\pi_1) \rightarrow \frac{F}{L^2 v^2 \rho} = f(R) \dots$$

Dimensionless force is a function of Reynolds number.

7. Assume $F = f(D, v, \rho, \eta, k)$

F	D	v	ρ	η	k
MLT^{-2}	L	LT^{-1}	ML^{-3}	$M L^{-1} T^{-1}$	L

$n = 6$ (dimensional parameters); $k = 3$ (dimensions).

Number of dimensionless parameters: $n-k=3$.

Select independent ‘repeating’ variables: D , v , and ρ .

Combine these with remaining non-repeating variables: F , η , and k .

Solving via exponent method:

$$\pi_1 = \frac{F}{D^2 v^2 \rho} \quad \pi_2 = \frac{\eta}{D v \rho} \quad \pi_3 = \frac{k}{D},$$

$$\frac{F}{D^2 v^2 \rho} = f\left(\frac{\eta}{D v \rho}, \frac{k}{D}\right).$$

8. (a) Assume $F = f(A, v, \rho, \eta)$.

F	A	v	ρ	η
MLT^{-2}	L^2	LT^{-1}	ML^{-3}	$ML^{-1}T^{-1}$

$n = 5$ (variables); $k = 3$ (dimensions).

From Buckingham’s π -Theorem, number of dimensionless parameters = $n-k = 2$.

$$(b) \pi_1 = F(A^a v^b \rho^c) \quad M^0 L^0 T^0 = (MLT^{-2})(L^2)^a (LT^{-1})^b (ML^{-3})^c.$$

$$\text{Solving for exponents: } \pi_1 = \frac{F}{Av^2 \rho}$$

(In fluid dynamics, the factor $\frac{F}{\frac{1}{2}(Av^2 \rho)}$ is defined as the Drag Coefficient.)

$$\pi_2 = v(A^a \eta^b \rho^c) \quad M^0 L^0 T^0 = (LT^{-1})(L^2)^a (ML^{-1} T^{-1})^b (ML^{-3})^c.$$

(We choose v as a dependent variable.)

$$\text{Solving for exponents: } \pi_2 = \frac{v \rho \sqrt{A}}{\eta} \text{ (this is related to the Reynolds Number).}$$

If we try $\pi_2 = A(\eta^a v^b \rho^c)$ we still obtain a valid result: $\pi_2 = \frac{Av^2 \rho^2}{\eta^2}$.

$$(c) \frac{F}{Av^2 \rho} = f\left\{\frac{v \rho \sqrt{A}}{\eta}\right\}.$$

(d) All the data should lie on one single curve.

(e) We can identify the important parameters; we do not have to conduct experiments for every single size of particle or fluid flow velocity; our results will work for different fluids (ρ and η); guides investigations of small-scale versions of large systems; any consistent set of units will work.

Index

- absorbance (optical), 166, 329
accelerometer, 308, 318
acetylcholine receptor, 121
acid dissociation, 21,
acids, 18
action potential, 68, 335, 339
active sensor, 262, 311
active transport across membrane, 97, 99
activity coefficient, 6
activity coefficient of ions in solution, 79, 84,
 86, 93
activity gradient, 93
adipocyte, 32
ADP, 14
advection, 406
aerofoil, 359
amino acid, 35
amperometric biosensor, 220, 228
amperometry, 200
amphipathic molecule, 31
ampholyte, 128
amphoteric, 288
amplification, 262, 308, 339, 340
analytical noise, 224
anode, 180
antibody, 34, 57, 68, 216, 240, 283, 313, 326,
 329, 332
antibody immobilisation, 238, 242
antibody-antigen interaction, 239, 240, 244,
 313, 326
antigen, 34, 57, 68, 240, 313, 326
anti-stokes shift, 153, 159
apoptosis, 57
applying Kirchhoff's laws to fluid flow, 364
aquaporin, 54, 95, 109
ATP, 13
ATP synthesis, 178
avidin-biotin system, 218
Avogadro's hypothesis, 344
Avogadro's number, 4, 344
bacteria, 58
bacteria growth, 59
base physical quantities, 392
bases, 19
B-cell, 57, 68
Beer-Lambert law, 165, 242
Bernoulli's equation, 358
biocompatibility, 252
BioFET, 234, 294
biomimetic sensor, 245
biosensor, 215
biosensor format, 218
biosensor pH response, 228
biosensor temperature response, 228, 298
biosensor, 1st, 2nd and 3rd generation, 232,
 248, 250
blood composition, 56
blood cells, 34, 55
blood group, 34
blood viscosity, 353, 361
bond number, 409, 412
bootstrapping, 332
bottom-up approach, 377

- British Medicines Healthcare Regulatory Agency, 255
Brokaw bandgap voltage reference, 303
Buckingham's π -theorem, 394
buffer amplifier, 266, 322
buffers, 24
Butler-Volmer equation, 196
- calcium channel (pump), 118
capacitive sensors, 317
capillary action, 386
capillary characteristic length, 409
capillary electrophoresis, 132
capillary number, 410
capillary time scale, 413
carbohydrates, 32
cardiac action potential, 119
catalyst, 12
cathode, 180
CCD, 332
cell chemical components, 30
cell culture, 63
cell cycle, 54
cell signalling, 42
cell wall, 58
cell-cell communication, 66
cells, 51
cells, electrical properties, 105
cells, transmembrane ion distribution, 104, 114
central dogma of molecular biology, 50
characteristic time scales, 409, 412
charge-transfer resistance, 209, 272, 282
chemical bonds, 2, 6, 9
chemical concentrations, 4
chemical equilibrium constant, 10, 12, 14, 19
chemical reaction, 9
chemiluminescence, 150, 243
chlorophyll, 153
chromophore, 154, 164
chromosome, 46, 50, 54
CMOS, 290, 332
CMOS Camera, 332
codon, 49
common mode rejection ratio, 268
complementary colour, 154
complex plane impedance plot, 211, 213, 240, 281
concanavalin A, 218
concept of similarity, 391, 396
condensation reaction, 33, 38, 43
conductometric biosensor, 220, 237, 278
- conservation of energy, 358, 366, 369
conservation of mass, 356, 366
conservation of momentum, 366, 367
contact wetting angle, 385
continuity equation, 356
continuum model of fluid, 346, 404
continuum versus molecular model, 369
convective time scale, 409, 412
corrosion of iron, 180
Couette flow, 352
Coulomb energy, 78
Coulomb potential, 76
Coulomb's law, 74
counter electrode, 195, 202, 210, 271
current amplifier, 277
current clamp, 237
current follower, 268, 320, 322, 338
cyclic voltammetry, 197
cytometry, 327
cytoplasm, 52, 336, 338
cytoskeleton, 53
- Daniell cell, 186, 193
Debye screening length, 81
Debye-Hückel equation, 86
dendrite, 335
density, 346
derived physical quantities, 393
deterministic simulations, 373
diabetes, 247
dielectric permittivity of water, 78, 88
dielectrophoresis, 137, 392
differential amplifier, 267
diffusion, 94, 99, 277, 378, 406
diffusion coefficient, 379, 401, 402
diffusion gradient, 381
diffusion time scale, 412
diffusion, facilitated, 97, 99
dimensional analysis, 391
dimensional homogeneity theorem, 394
dimensionless parameters, 394, 400, 411
dipole-dipole interaction, 8, 88
DNA, 43, 332
DNA length, 45, 46, 50
DNA replication, 46
DNA sensor, 245, 332
DNA-RNA transcription, 47, 50
Donnan equilibrium, 100
double layer capacitance, 282, 340
drug discovery, 123

- dynamic pressure, 358
dynamic viscosity, 353, 360, 410
- E. coli*, 15
Einstein-Smoluchowski equation, 378
electrical double layer, 282
electrochemical cell, 180, 271
electrochemical gradient, 99, 117
electrochemical half-cell, 179
electrochemical impedance spectroscopy, 208, 280
electrochemistry, 178
electrode potential, 184
electrode reactions, 180,
electrolysis, 181
electromagnetic radiation, 148
electron spin resonance (ESR), 163
electron transfer mediator, 232
electron transfer reaction, 179, 194
electronegativity scale, 6
electronic spectroscopy, 152, 153
electronic transitions between energy levels, 150, 153, 155, 158, 163
electro-osmosis, 129
electrophoresis, 124
electrophoretic mobility, 124
electroplating of copper, 180
electrostatic potential, 77, 80, 84
electrowetting on dielectric (EWOD), 143
energy, 1
enthalpy, 11, 78
entropy, 11, 78, 90
enzyme cofactor, 229, 231, 249
enzyme immobilization, 217, 238
equilibrium constant, 68, 169
erythrocyte, 56
eukaryotic cell, 52, 54
European Medicines Agency, 255
evanescent wave, 160, 164, 329
- FACS, 327
Faradaic reaction, 276
fat cell, 32
fatty acid, 30
feedback, 260, 264, 265
Fermi energy, 183, 194, 200
ferri/ferrocyanide redox couple, 231
FET, 284
Fick's equations, 381
field effect transistor, 284
finite difference method, 370
- finite element method, 370
finite volume method, 372
flatband voltage, 287, 288
flavoenzyme, 229
FLIM, 328
flow resistance, 360
fluid conservation equations, 370
fluid dynamics, 356
fluid pressure, 354
fluid shearing force, 344
fluidic junction, 364, 365, 381, 389
fluidic lever, 355
fluorescence, 150, 162, 165, 325
fluorescence intensity, 242
fluorescence lifetime, 326
fluorescence microarrays, 327
fluorophore, 151, 160, 242, 327–329, 332–334
Förster resonance energy transfer (FRET), 164, 242, 244, 247, 332–334
Frank-Condon principle, 155
FRET, 164, 242, 244, 247, 328, 332–334
fungal cell, 60
- gap junction, 69
gas constant, 344
gas viscosity values, 353
gated-ion sensor, 246
gauge factor, 305
Gaussian (normal) distribution, 380
genetic code, 49
genome, 45
Gibbs free energy, 11, 14, 86, 113, 181, 186
glucose, 33
glucose sensor, 230, 247–251, 260
glycoprotein, 53
golden rules (for operational amplifiers), 265
Gouy-Chapman-Stern equation, 83, 289
Gram stain, 58
gramicidin channel, 121
granulocytes, 57
green fluorescent protein (GFP), 151, 165, 326
Grotthuss mechanism, 17
- haemoglobin, 42, 56
Hagen-Poiseuille relationship, 360, 363
hard sphere model, 373
heart muscle membrane, 118
Helmholtz-Smoluchowski equation, 126
Henderson-Hasselbalch equation, 21
histone, 46, 52

- HIV, 51, 56, 94, 151
Holliday junction, 333
hydration forces, 91
hydration shell, 78, 87
hydraulic diameter, 401
hydraulic press, 354
hydrodynamic radius, 378
hydrogen bond, 16, 90, 241, 383
hydrophobic forces, 90
hydroxyl group, 17, 24, 288
hypertonic solution, 94
hypotonic solution, 94, 96
- ideal gas law, 344
ideal polarised electrode, 195, 201
immune response, 253
immune system, 57
immunoglobulin, 34, 42
impedance spectroscopy, 170, 280
impedimetric biosensor, 237, 283
implantable sensor, 252
infrared spectroscopy, 152, 156
instrumentation amplifier, 274
interdigitated transducer, 315
inverting amplifier, 267, 323
ion channel, 89, 115, 123, 335, 338
ion channel conductance, 114, 120
ion channel dysfunction, 124
ion electrical mobility, 17
ion pump, 54
ion selective electrode, 233, 235, 287
ion transport across membrane, 94
ion-dipole interaction, 86
ionic bond, 7
ionic double layer, 79
ions in membrane, 88
ions in protein, 90
ions in water, 78
ion sensitive field effect transistor, 287, 288
ISFET, 287, 288
ISFET fabrication, 290
ISFET instrumentation, 291
isoelectric focussing, 127
isoelectric point, 20
isotonic solution, 94, 96
- Kelvin measurement, 300
kinematic viscosity, 407, 408
kinetic theory of gases, 346
Knudsen number, 369, 376, 403, 411
- lab-on-chip device, 136, 143, 335, 343, 382, 388
laminar flow, 361, 407
Laplace's law, 355
law of mass action, 10, 19
laws of thermodynamics, 11, 78
Lennard-Jones potential, 8, 88, 374
leukocyte, 57
light dependent resistor, 330
lipids, 29
liquid junction potential, 207
liquid viscosity values, 353
liquids, 346
luminescence, 150
lymphocyte, 56, 57, 253
- macrophage, 58, 253
mass transfer, 197
Maxwell distribution of molecular speeds, 349, 351
MEA, 340
mediated amperometric biosensor, 231
membrane action potential, 107, 116, 165, 335, 339
membrane capacitance, 109, 141
membrane equilibrium potential, 98, 105, 111, 114
membrane excitation, 108
membrane hyperpolarisation, 117
membrane ion channel, 108, 335, 338
membrane mosaic model, 115
membrane resistance, 108
membrane, active electrical response, 108
membrane, passive electrical response, 108
meso scale, 375, 404
messenger rna, 49
microdevice design, 411, 413
microelectrode array, 340
microelectrodes, 277
mitochondria, 179
mobility, 286
molar absorption coefficient, 168
molar mass, 4
molar solution, 5
molarity, 5
molecular mass, 346
molecular mean free pathlength, 345, 403
molecular model, 404
molecular simulations, 372
molecular speed, Maxwell distribution, 349
molecule kinetic energy, 347, 352

- monocyte, 58
Monte Carlo simulation, 374
MOSFET, 284
Murphy's Law, 306
myelin, 335
myocyte, 118

Navier-Stokes equations, 365
Nernst equation, 113, 192, 198, 233
Nernst potential, 112
Nernstian, 289
nerve (axon) membrane, 117, 335
neuron, 68, 335
neuron action potential, 165, 335, 339
neutrophils, 252
Newtonian fluid, 352, 367
Newton's law, 347, 393
non-inverting amplifier, 265, 323
nonpolar bond, 7
nuclear magnetic resonance (NMR), 162
nucleic acid, 43
nucleosome, 46
nucleotide, 44, 334

oligopeptide, 39
op-amp limitations, 269
operational amplifier, 264
optical fibre sensor, 328
optode, 329
optometric biosensor, 219, 325
osmolarity, 91
osmole, 91
osmosis, 95, 99
osmotic pressure, 91
osmotic properties of cells, 103
over-potential, 196
oxidation, 178, 180

pacemaker cell, 118
parasitic capacitance, 321
partition coefficient, 95
Pascal's law, 354
Pascal's triangle and pyramid, 379
passive sensor, 262, 300
patch-clamp, 122, 338
Peclet number, 406, 411
peptide bond, 38
pH, 19, 234, 288–294
pH electrode, 234, 235

pH scale, 20
phagocyte, 34, 55, 57, 68, 252
phospholipid, 31
phosphorescence, 150
photobleaching, 328
photodiode, 330
photometric biosensor, 242, 325
photomultiplier tube, 330
piezoelectric effect, 311
piezoresistive effect, 304
pK, 21
Planck's constant, 148
plant cell, 60
plasma, 56
plasma membrane, 53
plasmids, 52
platelets, 56
platinum resistive thermometer,
p-n junction diode, 301
Poiseuille's law, 360
Poisson equation, 79
Poisson's ratio, 305
Poisson-Boltzmann equation, 80, 84
polar bond, 7
polypeptide, 40
potassium channel, 117
potentiometric biosensor, 233, 245
potentiostat, 272, 337
Prandtl number, 411
pressure, 347
pressure sensor, 306, 317
principle of superposition, 75, 138
prion, 62
prokaryotic cell, 52, 54
protein folding process, 90
protein molecular weight determination, 127
protein structure, 40
proteomics, 42
proton mobility, 17
protons, 17
PRT, 298
PTAT sensor, 302

QCM, 311
QCM equivalent circuit, 314
QCM instrumentation, 314
quantum mechanical theory, 149, 157, 158, 162
quartz crystal microbalance, 311
quenching, 328

- Ramachandran plot, 39
Raman spectroscopy, 153, 159
Randles equivalent circuit, 283
random (drunken sailor) walk, 379
ratiometric detection, 165, 170, 242, 244
Rayleigh time scale, 413
Rayleigh waves, 315
red blood cell, 56
redox reaction, 178
reduction, 178, 180
reduction-oxydation reaction, 178
reference electrode, 201, 203, 271, 288, 292
regenerative medicine, 64
resistance temperature detector, 298
retrovirus, 50
Reynold's number, 406, 412
ribosome, 40, 52
RNA, 47, 49, 52
rotational spectroscopy, 152, 157
RTD, 298

saccharide, 33
salt bridge, 186, 236
salt dissociation in water, 78
saturated-calomel electrode, 205, 271
SAW, 315
scaling, 392
second law of thermodynamics, 78, 113
self-assembled monolayers (SAMS), 239
sensor accuracy, 222
sensor bandwidth, 227
sensor decision limit, 224
sensor detection limit, 224
sensor drift, 222
sensor dynamic range, 226
sensor hysteresis, 227
sensor noise, 221
sensor precision, 222
sensor resolution, 227
sensor response time, 226
sensor selectivity, 221
sensor sensitivity, 220
sensor transfer function, 220
serpentine channel, 383
shear stress, 362
silver-silver chloride reference electrode, 204, 292
similarity, 391, 396
single photon avalanche detector, 331
smart sensor, 261
Snell's law, 160

soap bubble, 384
sodium channel, 117
sodium-potassium pump, 54, 98
soft sphere model, 373
soma, 335
SPAD, 331
specific heat, 411
spectroscopy, 151
squid axon, 122, 336
stability, 277
standard conditions, 188
standard electrode potential, 185
standard free energy change, 12
standard hydrogen electrode, 187
standard reduction potential, 187
stem cell, 64, 66
stochastic simulations, 374
Stokes shift, 153, 159, 326
Stokes viscous force, 124
Stokes-Einstein equation, 378
strain gauge, 262, 304
sugars, 32
surface acoustic wave, 315
surface enhanced Raman spectroscopy, 159
surface plasmon resonance, 163, 219
surface tension, 383, 401
surface wetting, 386, 388
surfactants, 384
switched capacitor circuits, 322–325
synapse, 68

T-cell, 57, 68
temperature definition, 346
temperature sensor, 293, 298–304
theristor, 301
thermodynamics, Laws of, 11
three-electrode system, 201, 271
threshold voltage, 285
TIRF, 328
tissue engineering, 64
titration curve, 22, 23
tonicity, 91
top-down approach, 375
total internal reflection fluorescence (TIRF), 160, 328
transfer coefficient, 196
transfer function, 265–268
transimpedance amplifier, 268, 320
translation of DNA code, 40
transmittance, 166

- T-type calcium channel, 119
turbulent flow, 407
turgor pressure, 96

ultra-microelectrode, 277

vaccine, 61
van der Waals force, 8, 87, 148, 241, 373, 383
van der Waals radius, 7, 39
vibrational spectroscopy, 152, 156
virus, 61
viscosity, 352, 401
visible EM spectrum, 149
voltage amplifier, 263
voltage clamp, 121, 336
voltage follower, 266
voltage gated channel, 117
voltage reference, 303

wall tension, 355
Warburg impedance, 209, 283–284
water, 15
water dipole moment, 77, 86
water dissociation, 19
Weber number, 410
Wheatstone bridge, 279, 299, 305, 310
white blood cell, 57
working electrode, 195, 202, 210, 233, 271
wound healing, 252

X-chromosome, 50

Y-chromosome, 50
Young's equation, 145, 386
Young's modulus, 305

zeta potential, 126, 131