Molecular Communication

Tadashi Nakano, Andrew W. Eckford
and Tokuko Haraguchi

CAMBRIDGE

Molecular Communication

This comprehensive guide, by pioneers in the field, brings together, for the first time, everything a new researcher, graduate student or industry practitioner needs to get started in molecular communication. Written with accessibility in mind, it requires little background knowledge, and provides a detailed introduction to the relevant aspects of biology and information theory, as well as coverage of practical systems.

The authors start by describing biological nanomachines, the basics of biological molecular communication, and the microorganisms that use it. They then proceed to engineered molecular communication and the molecular communication paradigm, with mathematical models of different types of molecular communication, and a description of the information and communication theory of molecular communication. Finally, the practical aspects of designing molecular communication systems are presented, including a review of the key applications.

Ideal for engineers and biologists looking to get up to speed on the current practice in this growing field.

Tadashi Nakano is an Associate Professor in the Graduate School of Engineering, Osaka University, Suita, Japan. He has authored or co-authored a series of papers on molecular communication, including the very first paper, published in 2005.

Andrew W. Eckford is an Associate Professor in the Department of Electrical Engineering and Computer Science at York University, Toronto, Canada. He has authored over 50 papers in the peer-reviewed literature, and received the Association of Professional Engineers of Ontario Gold Medal.

Tokuko Haraguchi is an Executive Researcher in the Advanced ICT Research Institute at the National Institute of Information and Communications Technology (NICT), Kobe, Japan, and a Professor with the Graduate School of Science and the Graduate School of Frontier Biosciences at Osaka University, Suita, Japan. She has authored 100 papers in biological research.

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York

Cambridge University Press in part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107023086

First published 2013

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Nakano, Tadashi, 1912–
Molecular communication / Tadashi Nakano, Andrew W. Eckford.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-02308-6 (hardback)
1. Molecular communication (Telecommunication) 2. Molecules.
3. Nanotechnology. I. Title.
TK5013.57.N35 2013
620′.5–dc23 2013009571

ISBN 978-1-107-02308-6 Hardback

Molecular Communication

TADASHI NAKANO

Osaka University, Suita, Japan

ANDREW W. ECKFORD

York University, Toronto, Canada

TOKUKO HARAGUCHI

Advanced ICT Research Institute, National Institute of Information and Communications Technology (NICT), Kobe, Japan

CAMBRIDGE
UNIVERSITY PRESS

Contents

Preface *page* xi

1 Introduction **1**

1.1 Molecular communication: Why, what, and how? 1
1.1.1 Why molecular communication? 1
1.1.2 What uses molecular communication? 2
1.1.3 How does it work? A quick introduction 3
1.2 A history of molecular communication 6
1.2.1 Early history and theoretical research 6
1.2.2 More recent theoretical research 8
1.2.3 Implementational aspects 9
1.2.4 Contemporary research 9
1.3 Applications areas 11
1.3.1 Biological engineering 11
1.3.2 Medical and healthcare applications 13
1.3.3 Industrial applications 14
1.3.4 Environmental applications 14
1.3.5 Information and communication technology applications 15
1.4 Rationale and organization of the book 15
References 16

2 Nature-made biological nanomachines **21**

2.1 Protein molecules 22
2.1.1 Molecular structure 22
2.1.2 Functions and roles 23
2.2 DNA and RNA molecules 28
2.2.1 Molecular structure 28
2.2.2 Functions and roles 30
2.3 Lipid membranes and vesicles 31
2.3.1 Molecular structure 31
2.3.2 Functions and roles 33

2.4 Whole cells 34
2.5 Conclusion and summary 35
References 35

3 Molecular communication in biological systems 36

3.1 Scales of molecular communication 36
3.2 Modes of molecular communication 37
3.3 Examples of molecular communication 38
3.3.1 Chemotactic signaling 40
3.3.2 Vesicular trafficking 41
3.3.3 Calcium signaling 42
3.3.4 Quorum sensing 44
3.3.5 Bacterial migration and conjugation 45
3.3.6 Morphogen signaling 46
3.3.7 Hormonal signaling 47
3.3.8 Neuronal signaling 47
3.4 Conclusion and summary 49
References 50

4 Molecular communication paradigm 52

4.1 Molecular communication model 52
4.2 General characteristics 54
4.2.1 Transmission of information molecules 54
4.2.2 Information representation 56
4.2.3 Slow speed and limited range 56
4.2.4 Stochastic communication 57
4.2.5 Massive parallelization 57
4.2.6 Energy efficiency 58
4.2.7 Biocompatibility 58
4.3 Molecular communication network architecture 58
4.3.1 Physical layer 60
4.3.2 Link layer 61
4.3.3 Network layer 64
4.3.4 Upper layers and other issues 65
4.4 Conclusion and summary 67
References 67

5 Mathematical modeling and simulation 71

5.1 Discrete diffusion and Brownian motion 71
5.1.1 Environmental assumptions 71
5.1.2 The Wiener process 72
5.1.3 Markov property 74

5.1.4 Wiener process with drift 75
5.1.5 Multi-dimensional Wiener processes 76
5.1.6 Simulation 77
5.2 Molecular motors 78
5.3 First arrival times 80
5.3.1 Definition and closed-form examples 80
5.3.2 First arrival times in multiple dimensions 82
5.3.3 From first arrival times to communication systems 82
5.4 Concentration, mole fraction, and counting 83
5.4.1 Small numbers of molecules: Counting and inter-symbol interference 84
5.4.2 Large numbers of molecules: Towards concentration 85
5.4.3 Concentration: random and deterministic 87
5.4.4 Concentration as a Gaussian random variable 89
5.4.5 Concentration as a random process 90
5.4.6 Discussion and communication example 92
5.5 Models for ligand–receptor systems 93
5.5.1 Mathematical model of a ligand–receptor system 93
5.5.2 Simulation 94
5.6 Conclusion and summary 95
References 95

6 Communication and information theory of molecular communication 97

6.1 Theoretical models for analysis of molecular communication 97
6.1.1 Abstract physical layer communication model 97
6.1.2 Ideal models 99
6.1.3 Distinguishable molecules: The additive inverse Gaussian noise channel 99
6.1.4 Indistinguishable molecules 100
6.1.5 Sequences in discrete time 102
6.2 Detection and estimation in molecular communication 104
6.2.1 Optimal detection and ML estimation 104
6.2.2 Parameter estimation 106
6.2.3 Optimal detection in the delay-selector channel 108
6.3 Information theory of molecular communication 109
6.3.1 A brief introduction to information theory 109
6.3.2 Capacity 110
6.3.3 Calculating capacity: A simple example 112
6.3.4 Towards the general problem 115
6.3.5 Timing channels 116
6.4 Summary and conclusion 120
References 121

7 Design and engineering of molecular communication systems 122

7.1 Protein molecules 123
7.1.1 Sender and receiver bio-nanomachines 123
7.1.2 Information molecules 124
7.1.3 Guide and transport molecules 125
7.2 DNA molecules 129
7.2.1 Sender and receiver bio-nanomachines 129
7.2.2 Information molecules 129
7.2.3 Interface molecules 130
7.2.4 Guide and transport molecules 131
7.3 Liposomes 132
7.3.1 Sender and receiver bio-nanomachines 133
7.3.2 Interface molecules 134
7.3.3 Guide molecules 135
7.4 Biological cells 136
7.4.1 Sender and receiver cells 136
7.4.2 Guide cells 142
7.4.3 Transport cells 144
7.5 Conclusion and summary 147
References 147

8 Application areas of molecular communication 152

8.1 Drug delivery 152
8.1.1 Application scenarios 153
8.1.2 Example: Cooperative drug delivery 153
8.1.3 Example: Intracellular therapy 154
8.2 Tissue engineering 156
8.2.1 Application scenarios 156
8.2.2 Example: Tissue structure formation 157
8.3 Lab-on-a-chip technology 158
8.3.1 Application scenarios 160
8.3.2 Example: Bio-inspired lab-on-a-chip 160
8.3.3 Example: Smart dust biosensors 161
8.4 Unconventional computation 162
8.4.1 Application scenarios 162
8.4.2 Example: Reaction diffusion computation 162
8.4.3 Example: Artificial neural networks 164
8.4.4 Example: Combinatorial optimizers 165
8.5 Looking forward: Conclusion and summary 166
References 166

9 Conclusion **169**

9.1 Toward practical implementation 169
9.2 Toward the future: Demonstration projects 170

Appendix Review of probability theory **172**

A.1 Basic probability 172
A.2 Expectation, mean, and variance 173
A.3 The Gaussian distribution 174
A.4 Conditional, marginal, and joint probabilities 175
A.5 Markov chains 175

Index 177

Preface

As early researchers in molecular communication, we have been amazed by the rapid expansion of the field. A decade ago, virtually nobody worked in this area; today, dozens of researchers form a multi-national research community, and over a hundred papers have been published. At the frontiers of the field, there are fundamental questions to be answered such as the relationship between information theory and biology; and disruptive innovations to be developed, such as direct manipulation of structures in the human body at a microscopic level.

Given the advances over the past few years, we believe the time is right to take stock of the field and publish a complete overview of the state of the art. In an interdisciplinary field such as this one, we hope this book can provide a needed common point of reference. Moreover, in an evolving field such as this one, we recognize that our book should not be considered the final word on the field. Indeed, in writing it we have become fully aware of the many important open problems and research questions that need to be addressed for this field to reach its potential, and we hope our book is viewed as an invitation to further research, to expand upon this exciting new discipline.

Finally, we would like to thank the many people whose work, discussions, and encouragement over the years have made this book possible: in no particular order, Akihiro Enomoto (Qualcomm), Ryota Egashira (Yahoo! Inc.), Yasushi Hiraoka (Osaka University/National Institute of Information and Communications Technology), Satoshi Hiyama (NTT DoCoMo), Takako Koujin (National Institute of Information and Communications Technology), Shouhei Kobayashi (National Institute of Information and Communications Technology), Jian-Qin Liu (National Institute of Information and Communications Technology), Michael Moore (Pennsylvania State University), Yuki Moritani (NTT DoCoMo), Kazuo Oiwa (National Institute of Information and Communications Technology), Yutaka Okaie (Osaka University), Jianwei Shuai (Xiamen University), Tatsuya Suda (Netgroup Inc.), Nariman Farsad (York University), Lu Cui (York University), Peter Thomas (Case Western Reserve University), Raviraj S. Adve (University of Toronto), K. V. Srinivas (Samsung), Sachin Kadloor (University of Illinois at Urbana-Champaign), Chris Rose (Rutgers), and Chan-Byoung Chae (Yonsei University).

1 Introduction

Historically, communications engineers have dealt with electromagnetic forms of communication: in wireline communication, electric fields move currents down a wire; in wireless communication, electromagnetic waves in the radio-frequency spectrum propagate through free space; in fiber-optic communication, electromagnetic radiation in the visible spectrum passes through glass fibers.

However, this book is concerned with an entirely different form of communication: *molecular communication*, in which messages are carried in patterns of molecules. As we shall see in this book, molecular communication systems come in many forms. For example, message-bearing molecules may propagate through a liquid medium via simple Brownian motion, or they may be carried by molecular motors; the message may be conveyed in the number and timing of indistinct molecules, or the message may be inscribed directly on the molecule (like DNA); the nanoscale properties of individual molecules may be important, or only their macroscale properties (like concentration).

Molecular communication is literally all around us: it is the primary method of communication among microorganisms, including the cells in the human body. In spite of its importance, only in the past decade has molecular communication been studied in the engineering literature. In writing this book, our goal is to introduce molecular communication to the wider community of communications engineers, and collect all the current knowledge in the field into a single reference for the sake of researchers who want to break into this exciting field.

1.1 Molecular communication: Why, what, and how?

1.1.1 Why molecular communication?

Why would engineers want to design a system involving molecular communication? To motivate this question, suppose you are given the following design problem. Your goal is to perform *targeted drug delivery*: to deliver drugs within the human body exactly where they are needed (for example, directly to malignant tumors within the body, as chemotherapy). To accomplish this goal, you have decided to use *thousands of tiny, blood-cell-sized robots* that must cooperate with each other to autonomously navigate through the body, identify tumors, and release their drugs to destroy the tumor. To

cooperate, the robots must be able to communicate – so how would you design the communication system?

This is a challenging question: as a result of their size, the devices have very small energy reserves, and must glean whatever energy they can from the environment. The devices must also operate in the body without disrupting healthy tissues, or being destroyed by the immune system prior to completing the mission. These features are consistent with the communication challenge faced by microorganisms, and these organisms have solved the problem by exchanging signals composed of molecules – that is, *molecular communication*.

As a result, for engineered systems, molecular communication is a *biologically inspired* solution to the communication problem. This communication could be engineered in two ways: first, an entirely artificial device could be designed to communicate using signaling molecules; and second, the existing molecular communication capabilities of an engineered microorganism (e.g., a bacteria with custom DNA) could be used. Remarkably, both nanoscale robots [1] and artificial bacteria [2] are within the capabilities of contemporary technology. However, nanoscale communication techniques, such as molecular communication, are needed to permit cooperation and unlock the disruptive potential of these systems.

1.1.2 What uses molecular communication?

In the previous section, we gave an example of tiny robots swimming through the human bloodstream. This example opens us up to *biological nanomachines*, or *bio-nanomachines*, one of the primary application areas of molecular communication.

For our purposes, bio-nanomachines may be defined as follows:

- *Materials.* A bio-nanomachine is made of biological materials (e.g., protein, nucleic acid, liposome, biological cell), or a hybrid of biological and non-biological materials.
- *Size.* A bio-nanomachine's size ranges from the size of a macromolecule to the typical size of a biological cell ($\sim$100 μm).[1]
- *Functionality.* A bio-nanomachine's functionality is limited to simple computation (e.g., integrating two types of input signals to produce one output signal), simple sensing (e.g., sensing only one or two types of molecule), and simple actuation (e.g., producing simple mechanical motion).

Figure 1.1 gives an overview of a molecular communication system involving bio-nanomachines [3]. In molecular communication, information is encoded onto (and decoded from) molecules, rather than electrons or electromagnetic waves. First, an information source generates information to encode onto molecules and triggers a group of sender bio-nanomachines to start propagation of information-encoded molecules.

[1] The term "nano" sometimes refers to dimensions of 1–100 nm, which is included in this definition; however, biological cells are typically much larger. Recently, the term *mesoscopic* has been used to describe dimensions that span from atomic to microbiological scales.

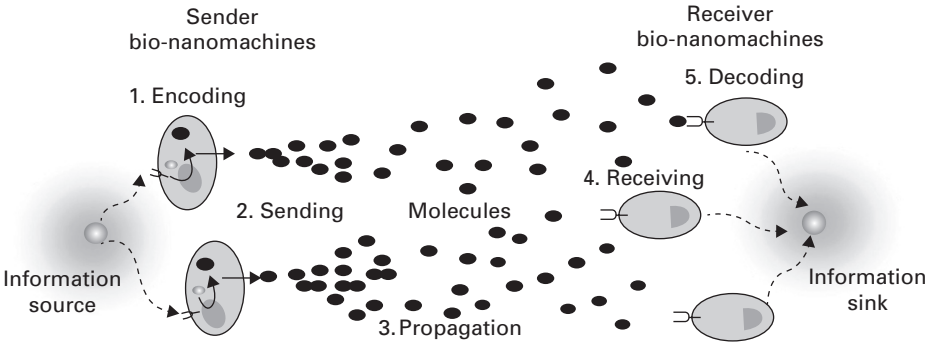


Figure 1.1 An outline of a molecular communication system incorporating bio-nanomachines [3].

Information-encoded molecules then propagate in the environment, and are detected by a group of receiver bio-nanomachines. Receiver bio-nanomachines may forward incoming molecules to next-hop bio-nanomachines or may pass them to an information sink for decoding information. We discuss this process in greater detail in Chapter 4.

Bio-nanomachines are not the only application for molecular communication – however, they are in many ways the primary motivating application, and the one that informs most of the analysis throughout this book. We give an introduction to applications in Section 1.3, and some detailed examples in Chapter 8.

1.1.3 How does it work? A quick introduction

How does molecular communication work? We spend the rest of this book answering this question, but here we give the reader a quick overview, and introduce the basic issues related to designing a molecular communication system.

First, we should be clear what we mean by "communication." We focus on artificial communication, where a manmade *message* needs to be conveyed from one point to another. A message can be discrete (like a sequence of bits, as in an IP packet), or continuous (like an analog waveform, as in AM radio), but for now, we will assume that the message is discrete. In the simplest form of communication, there are two terminals: a *transmitter*, which sends the message; and a *receiver*, which receives the message. (So far, this is general enough to include any point-to-point communication system, not just molecular communication. This setup can be generalized: in a network setting, there may be many senders and receivers, and a terminal can be both a sender and a receiver for different messages.)

To communicate, the transmitter makes a physical change to its environment, and that change must be measurable at the receiver. Again, this is true of any communication system: for instance, a wireless transmitter induces a changing EM field along an antenna, which can be detected in an antenna at the receiver. However, in molecular communication, the change must be molecular: the transmitter releases molecules into a shared medium, which propagate to (and are detected by) the receiver.

In order to convey distinct messages, each possible message is associated with a molecular *signal*: a unique pattern of molecules for each possible message, which can be distinguished at the receiver. Further, there must be a way for the receiver to *decide* which message was sent, based on the signal that it measures. For instance, say we want to send a message consisting of a single bit, 0 or 1. We can do this in many ways, but here are three possibilities:

- *Signaling with quantity.* Say we have $n > 0$ molecules available at the transmitter. We could send a 0 by releasing zero molecules, or a 1 by releasing n molecules. If the receiver observes 0 molecules, it can conclude that a 0 was sent; if it observes at least one molecule, it can conclude that a 1 was sent.
- *Signaling with identity.* Say we have two types of molecule available at the transmitter, A and B (where the receiver can distinguish A from B). We could send a 0 by releasing molecule A, or a 1 by releasing molecule B. The receiver would decide 0 or 1 if it observed A or B, respectively.
- *Signaling with timing.* Say we have a single molecule available at the transmitter. We could send a 0 by releasing that molecule right now, or we could send a 1 by waiting $t > 0$ seconds before releasing the molecule. The receiver would then decide whether 0 or 1 was sent by measuring the arrival time of the molecule.

This simple example, illustrated in Figure 1.2, encapsulates many of the general techniques that we will describe throughout the book. For example, generalizing a quantity signal, we can manipulate the concentrations of molecules in the medium. We may also wonder how to generate molecular signals; we will see throughout this book that many

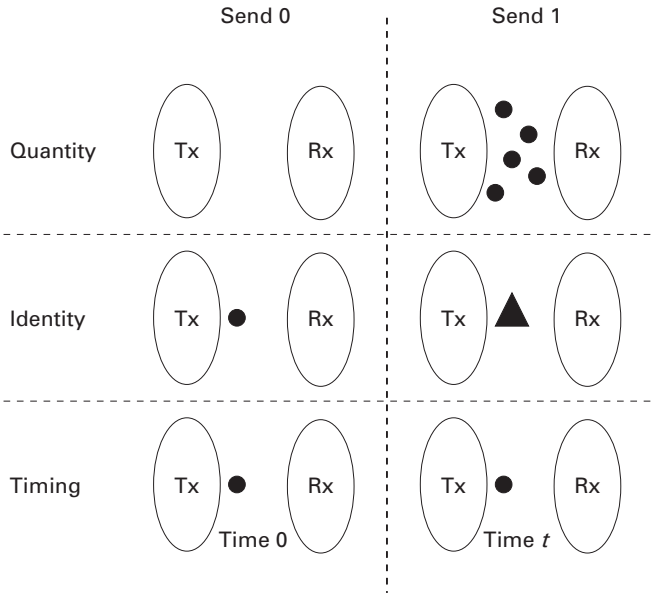


Figure 1.2 Illustration of three simple ways of generating a binary molecular signal.

biological "components" exist to emit and receive message-bearing molecules. As a result, molecular communication systems are often biologically based.

We also see that the propagation of molecules from transmitter to receiver must take place via diffusion: this could be viewed as either discrete Brownian motion, for small numbers of molecules; or continuous diffusion, for large numbers of molecules. Later, we will see that diffusion is a significant source of distortion and constraint on molecular communication systems: for instance, discrete Brownian motion might mean that message-bearing molecules are lost, or that they take an arbitrarily long time to arrive; further, continuous diffusion is a very slow process, which limits the possible rate of information transfer.

Figure 1.3 shows an example of molecular communication in the laboratory. A sender cell is stimulated at time $t = 0$, and encodes a molecular signal, using inositol trisphosphate (IP_3) and adenosine triphosphate (ATP). Here, information about the stimuli is encoded in the *type* and *number* (i.e., the concentration) of molecules. The sender cell broadcasts the molecular signal into the environment, through external pathways (in the extracellular space) or internal pathways (gap junction channels). The molecular signals diffuse through the two pathways, and receiver cells in the environment detect the molecular signals using receptors. The receiver cells then increase the concentration of intracellular molecules (e.g., calcium), and decode the signals using molecular mechanisms inside the cells. More detail of this form of molecular communication is provided in Chapter 3, as well as many other forms of molecular communication found in biological systems.

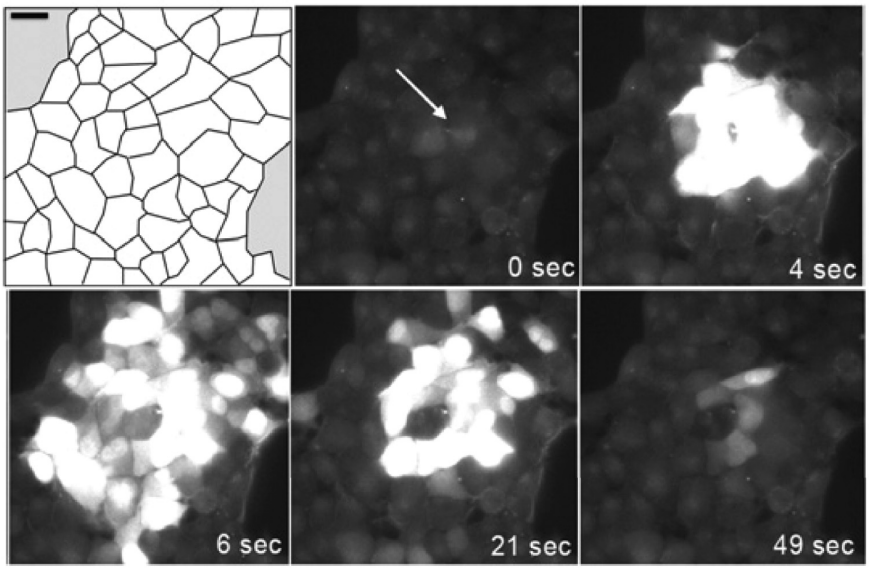


Figure 1.3 Series of images from a lab experiment. A sender cell, upon stimulation, broadcasts molecular signals and the receiver cells in the environment respond to the molecular signals [4].

1.2 A history of molecular communication

The field of nanotechnology is commonly traced back to the Nobel-laureate physicist Richard Feynman, and his famous 1959 lecture to the American Physical Society, entitled *There's Plenty of Room at the Bottom* (transcribed in [5]). Feynman argued that the laws of physics permit very small devices, far smaller than contemporary technology had managed to produce. Since 1959, Feynman's vision of extreme miniaturization has been realized in many fields, such as integrated circuitry and microscopy. Moreover, new fields of research, such as micro- and nano-electro-mechanical systems (MEMS and NEMS), were spawned to extend this miniaturization into robotics.

Meanwhile, it has long been recognized that microorganisms, including cells and bacteria, gain information from their environment by gathering chemical messengers sent by their neighbors. A simple example is quorum sensing [6], in which bacteria send molecular messages to one another in order to estimate the local population of their species; the bacteria can take action based on this estimate, such as forming a colony or seeking out larger numbers of their species. Further, the means by which cells send messages to one another and control each others' behavior is a well-studied area of biology known as *cell signaling* (see, e.g., [7]).

The engineering aspects of molecular communication have a research background that stretches back decades. In this section, we give a brief review of this field's history. We begin with a review of the (mostly theoretical) work done by early communications researchers. We then discuss more recent theoretical and implementational work, and conclude with a short review of contemporary research in this field.

1.2.1 Early history and theoretical research

Work by early researchers, such as Shannon [8] and Nyquist [9], established information theory and communication theory as mathematical disciplines. The focus was on telegraphic communication, so these theories developed (and remain) largely as subfields of electrical engineering. As abstract models, these techniques can be used in more general studies of communication, such as molecular or biological communication. However, this direction of research has remained on the fringes of information theory until recently, perhaps because Shannon himself discouraged it [10].[2]

Nonetheless, there has long been interest in information theory as a tool for explaining biological behavior, especially in terms of biomolecular interactions. To the knowledge of the authors, the first discussion of information theory in the context of biomolecular interactions occurred in [11], which analyzed the efficiency of the kidney by recognizing its operation in terms of information processing: the kidney examines molecules and makes decisions on them, either keeping them in the bloodstream or

[2] Shannon's point was not that chemistry or biology are inherently inappropriate applications for information theory, it was that the reputation of a rapidly growing field depends on scientific rigor and high-quality work. At the time, such work was found in electrical applications. Reference [10] is certainly worth re-reading, and its lessons worth remembering, as our field of molecular communication appears poised for rapid growth.

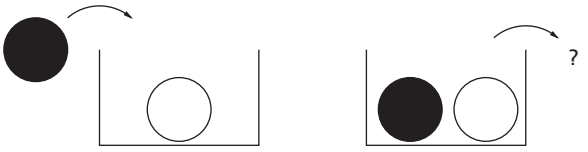


Figure 1.4 Illustration of Blackwell's "chemical channel." In the first figure, a black ball is introduced to the bag, which already contains a white ball. In the second figure, one of the two balls in the bag is selected at random and removed to form the channel output.

rejecting them as waste, in an operation reminiscent of Maxwell's demon. The key observation was that gathering molecular information has a minimum energy cost, so information processing explained the kidney's energy consumption. This work was extended by Berger [12], who showed that molecular energy efficiency could be explicitly described in terms of rate distortion theory. This result was cited as "visionary" in a book review [13].

Meanwhile, Blackwell [14] described a highly abstract channel model, where successive channel outputs are statistically dependent. This model was called the *chemical channel* by subsequent authors,[3] making it possibly the first molecular communication channel model to appear in the literature. In this model, colored balls are used to communicate: there are two colors, white and black, and the balls are otherwise identical. At the beginning of the communication session, a bag is filled with a given number of balls of unknown colors. Communication then proceeds as follows: first, the transmitter chooses a color and drops a new ball of that color into the bag, then the receiver selects a ball from the bag at random, removes it, and notes its color; this process is repeated as often as necessary to send a message.

EXAMPLE 1.1 Consider the chemical channel in Figure 1.4, where the bag initially contains one ball. You can send one bit of information for every three balls by using a *repetition code*: inserting three white balls in a row, or three black balls in a row. The receiver can tell what color the transmitter sent by picking the majority of colors out of every group of three: at most one ball will be the wrong color. This is not as good as you can do, however; capacity of the trapdoor channel is an open problem.

If the bag in this example contains many balls, then the random selection is a coarse analog to random diffusion. For instance, say we have molecules instead of balls: the transmitter inserts molecules into the channel, which diffuse randomly in the medium, and are ultimately removed by the receiver. If the molecules are perfectly mixed after each insertion, then we have something like this channel. Berger elaborated on these ideas, showing how they can be used to describe biological molecular communication in his Shannon prize lecture [16].

In its standard form, the trapdoor channel is a poor approximation to diffusion: the assumption of perfect mixing between insertions is not practical. The model can

[3] Early drafts of [15], available on arXiv, credit Thomas Cover with coining the term "chemical channel," though the claim is missing from the published version. The term "trapdoor channel" is also used.

be refined; for example, each ball can have a different probability of being selected. However, it is worth remembering that the trapdoor channel was not originally intended to model diffusion; the diffusion application came later. More recently, researchers have examined *diffusion-mediated* models that explicitly view molecular diffusion as a communication system.

Diffusion can be viewed microscopically, as a process involving individual molecules, or macroscopically, as a process involving continuous concentrations. The latter approach has the advantage of being linear: the (continuous) diffusion equation is a linear partial differential equation, so the considerable body of linear system theory for communication systems can be applied. Early work in this direction emerged from the biological literature: in [17], information theory was used to present chemical signal transduction in the retina as a communication system (to the authors' knowledge, the first explicit use of information theory in chemical signaling).[4] Building on these results, [20] simulated and analyzed a detailed linear model of a diffusion-mediated cellular transduction system, evaluating its frequency response and its information-theoretic capacity.

1.2.2 More recent theoretical research

The past five years have seen a rapid increase in information-theoretic analysis of molecular communication. The general information-theoretic model of communication is broad enough to include new methods of information transfer, including molecular communication (and we describe this general model in Chapter 6). For molecular communication, the challenge is to develop information-theoretic equivalents for the components of the model, such as the transmitter, the receiver, and the channel.

Discrete Brownian motion, modeled as a communication system, focuses on idealized models and the ultimate limits of molecular communication. This is because continuous diffusion is merely the limiting process of discrete Brownian motion, as the number of molecules becomes large. Thus, if we can find the limits of discrete Brownian motion, we have the best that can be done with molecular communication. The first work on discrete diffusion was [21], in which some "ideal" modeling assumptions were made, and the primary source of distortion in the channel was assumed to be the random propagation time of message-bearing molecules from transmitter to receiver.

It is important to note that discrete diffusion systems require processing that is far beyond contemporary technology: for one thing, these systems require sensing and manipulation of individual molecules; for another, they often assume synchronization between transmitter and receiver. However, as research into the ultimate limits of molecular communication, it is natural to consider these systems in terms of information theory.

Theoretical work has been done in other directions as well: continuous diffusion, considering the propagation of concentrations of molecules, is less efficient than discrete

[4] Information theory has been used to analyze neural coding for over fifty years, e.g. [18, 19], but not explicitly to analyze a chemical communication system.

diffusion, but feasible to implement in practice: components exist that can detect and respond to changes in concentration of a given molecular species. The capacity of such systems was considered in [22]. Biomorphic systems, as the focus of implementational work on molecular communication, are natural to analyze with information and communication theory. The aforementioned work of [20] is an example of this type of after analysis; another early work is [23], which analyzed a ligand–receptor system in discrete time.

1.2.3 Implementational aspects

The term "molecular communication," meaning an engineered communication system where messages are conveyed in patterns of molecules, was coined in the title of a 2005 paper [24]. That paper, focusing on the possible designs and uses of diffusion-based communication systems, launched a body of research on the implementation of molecular communication. These works described a variety of biological or chemical components that could be used to assemble practical systems to conduct molecular communication: in other words, this work explores the "hardware" that would form the communication system. In many cases, laboratory experiments have been performed to show the feasibility of molecular communication, or to describe potential applications.

Various subsystems for communication have been identified. As one example, the gap junction, used by cells to exchange ions, could be used by collections of cells to pass concentrations of ions. If this were done under external control, a message could be passed from one side of the collection to the other [25, 26]. As another example, liposomes (i.e., spherical vesicles that act as "packages" of molecules) can be used to exchange messages: information-bearing molecules can be encapsulated into a liposome, and passed to communication partners. This possibility was explored by [27, 28], and its feasibility was demonstrated in lab experiments in [29].

The practical problem of transporting molecules from transmitter to receiver has also been explored. Though random diffusion is one solution to this problem, there are alternatives: for example, molecular motors are used in living cells to transport molecules from one place to another. In molecular communication, motors may be used to collect message-bearing molecules (or packages of molecules, e.g., in liposomes) and transport them from transmitter to receiver [30]. Experiments validating this approach were presented in [31].

We describe some of these components and implementations in detail later in this book. An excellent review of contemporary research in implementational aspects of molecular communication is found in [32].

1.2.4 Contemporary research

Work on molecular communication has accelerated in the last five years, thanks in part to a new focus on nanoscale communication networks, or *nanonetworks* [33, 34]. Nanonetworks involve collections of very small devices that communicate and cooperate with each other, and in which essential features of the network have nanoscale dimensions. For example, swarms of nanorobots, which may be used in some of the

applications described earlier in this chapter, may form a nanonetwork to accomplish their task. Molecular communication has recently been recognized as an enabling technology for nanonetworking [35].

At the time of writing, molecular communication is increasingly popular among traditional communication engineers. As the background of these researchers is primarily theoretical and simulation-based, there has been a rapid increase in theoretical and simulation-based analysis of molecular communication. Without attempting to be comprehensive, we give four major themes of contemporary research:

- Channel modeling and noise analysis are key directions of research. Traditional communication and information theory are based on a set of mathematically precise channel models, such as the additive white Gaussian noise channel. Moreover, within each such channel, there exists a source of distortion, or "noise." However, no widely accepted general channel or noise model exists for molecular communication; depending on the scenario, it is likely that several different channel models are required. Adding to historical work on channel modeling, recent results include [36], which developed a complete end-to-end model of molecular communication based on continuous diffusion, and [37], which modeled the noise of an active-transport molecular communication system.
- The information-theoretic capacity of molecular communication, or the maximum rate at which data can be reliably transmitted, is an important open problem. The fully general problem of finding capacity is known to be difficult, but many recent papers have sought either bounds on capacity or the capacity in simplified scenarios, such as: [38], which considered continuous diffusion, simplifying the concentration to a binary variable (taking values of "high" and "low" concentration); [39], which found bounds on capacity for a general model of discrete diffusion; and [40], which considered a similar setup with generalized transmission schemes and possible molecular losses. Another direction is described in [41], which examines the symmetries in possible capacity-achieving input strategies, and bounds the general channel capacity.
- From the simulation side, system design for molecular communication is an important research topic. Compared to traditional communication, molecular communication seems less amenable to closed-form analysis and optimization; as a result, simulations are a key tool for determining the performance of the system. (Obviously, laboratory experiments are the most accurate way of determining performance, but these are significantly more expensive and difficult to perform than simulations.) A wide variety of design work has been done in simulation, such as optimization of distance-estimation techniques [42], design of channel shapes for microfluidic molecular communication [43], design of routing schemes in networks [44], and design of signaling techniques [45]. The design and analysis of simulation techniques themselves are an important open problem, and some papers are devoted entirely to that topic (e.g., [46]).
- System-level research has also attracted much recent attention. The problem described at the beginning of this chapter – operation of bio-nanomachines in the human body – was reviewed in [47], and major challenges identified. Molecular

communication for swarm nanorobotics, a related research challenge, was studied in [48, 49]. In response to surging system-level research efforts in molecular communication, the IEEE has created a standardization working group, numbered 1906.1, to standardize technologies and specifications for molecular communication [50].

We will discuss the background of many of these problems in greater detail throughout this book.

Notably, although simulation-based work is common in contemporary research, there is little work on laboratory-based implementation. Looking into the near future, there is an urgent need to validate simulation-based work in the laboratory, to ensure that the results are meaningful. This will require a close collaboration among researchers with a communication-theoretic background, and those with a biological or chemical background who have access to (and skills to use) laboratory facilities.[5]

1.3 Applications areas

Molecular communication has the potential to advance interdisciplinary applications in biological engineering, medical and healthcare, industrial, environmental, and information and communication technology areas [3, 33, 51] (Figure 1.5). In this section, we will briefly go over specific application areas where molecular communication research may advance existing methods and technologies. Some of the selected applications are discussed in more detail in Chapter 8.

1.3.1 Biological engineering

Molecular communication may benefit the area of biological engineering to analyze biological materials, engineer biological systems, and interface biological systems with manmade systems. Specific examples of relevant subareas in biological engineering include micro-electromechanical systems (MEMS) for analyzing biological samples, tissue engineering for regenerating tissues and organs, and brain–machine interfaces (BMI) for interfacing a human brain and electrical devices:

- *Micro-electromechanical systems (MEMS):* MEMS applies microtechnology to develop a small-scale system such as a lab-on-a-chip (LOC) [52, 53] or a network-on-chip (NoC) [54] for on-chip analysis of biological samples (e.g., molecules). A LOC is typically 20 micrometers to a millimeter in size and provides functionalities to manipulate molecules on a single chip, such as moving molecules from one location to the other, mixing one type of molecule with another type of molecule, and separating specific types of molecules from a mixture of molecules. For a LOC, molecular

[5] In the authors' experience, this is far easier said than done. Communications engineers have a different approach to research from biologists and chemists; for example, while the former often use abstraction to simplify a complex problem down to a "toy example," the latter often prefer to deal with the details as much as possible. Posing a question that both captures the interest, and requires the expertise, of both groups can be a challenging exercise in problem design.

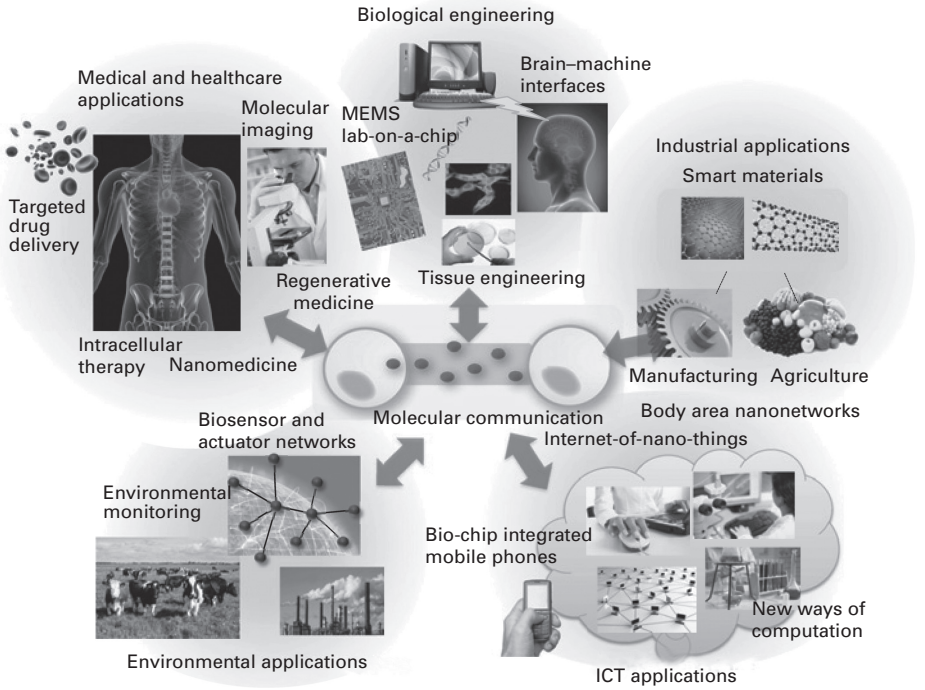


Figure 1.5 Application areas of molecular communication research

communication may provide nanoscale techniques to manipulate molecules on a chip. For instance, guide and addressing molecules considered in molecular communication (see Chapters 4 and 7) may be used to transport a molecule directionally to a specific location on a chip.

- *Tissue engineering:* Tissue engineering aims to develop a tissue structure from biological cells to restore the lost tissues of a patient's body [55]. In tissue engineering, stem cells (e.g., autologous cells) are extracted from a patient's body and cultured in vitro. Engineered extracellular matrices called scaffolds are often used as a template to guide the development and assembly of the cells into a three-dimensional tissue structure. Molecular communication may provide an additional mechanism to produce spatial patterns of molecules and thereby affect the growth and differentiation of the stem cells into specific tissue structures. For instance, bio-nanomachines (e.g., engineered stem cells) may release molecules that propagate in the environment to establish a spatial pattern of molecules, and the bio-nanomachines differentiate based on the established pattern to form a structure. The tissue structure is then implanted into the human patient to restore the lost tissues.
- *Brain–machine interfaces (BMI):* BMI provides a direct channel between the human brain and electrical devices to restore lost functions [56]. In BMI-based motor prostheses, for instance, motor signals generated in the brain are recorded through electrodes implanted in the brain and transmitted to an external device, which

interprets the motor signals to control the patient's artificial limb. Signals generated from an external device may also be transmitted to electrodes implanted in the brain, which in turn stimulate a specific part of the brain to treat a brain disease (e.g., Parkinson's disease). For BMI, molecular communication may provide a molecular or chemical means of interacting with the brain, instead of electrical ways. For such an application, bio-nanomachines may be engineered to implement a molecular communication interface to interact with a human brain as well as a man–machine interface to interact with electronic devices.

1.3.2 Medical and healthcare applications

Molecular communication may also apply to improving the ability of medical or healthcare systems. These systems may be integrated with a group of bio-nanomachines that communicate and cooperate to monitor medical conditions and perform therapy by releasing molecules. Examples of such scenarios are found in molecular imaging for gathering information for diagnosis, and targeted drug delivery and intracellular therapy for performing therapy:

- *Molecular imaging:* Molecular imaging is a technique to monitor cellular function and processes in vitro or in situ, which can be used to gather information for diagnosis. For molecular imaging, a green fluorescent protein (GFP) can be used as a reporter of gene expression, for example. Monitoring and diagnosis can be performed with GFPs by detecting the expression of particular genes in a human body, which indicates a disease or certain medical condition. GFPs can be carried by viruses, introduced in a tissue, and targeted to cancer cells in the tissue. When the tissue is illuminated, the GFP responds by emitting fluorescence and thus the location of cancer cells can be identified. Molecular communication may further improve the ability of molecular imaging, for instance, by providing coordination mechanisms for a group of bio-nanomachines (e.g., viruses carrying GFPs) to gather information about conditions from a larger area in a body, aggregate the information in situ, and transmit the aggregated information (e.g., through fluorescence) to external devices for further diagnosis.
- *Targeted drug delivery:* In targeted drug delivery, therapy on a target site (e.g., diseased cells or tumors) in a human body is performed by encapsulating drug molecules in drug delivery carriers, delivering the carriers to the target site, and releasing the drug molecules from the carriers at the target site. Targeted drug delivery therefore reduces the potential side effects of drug molecules on non-targeted sites [57, 58]. Existing research on drug delivery develops drug delivery carriers that can be targeted to a specific site in a body (e.g., tumors), where the drug molecules are released in response to specific conditions such as temperature. Molecular communication may provide alternative techniques to improve the accuracy of targeting and efficacy of therapy through the coordination of bio-nanomachines (i.e., drug delivery carriers). For instance, bio-nanomachines that identify a target site may release molecules to recruit other bio-nanomachines in the environment toward the target site, thereby, the

concentration of bio-nanomachines at the target site is increased. Also, a group of bio-nanomachines at the target site may communicate to determine the rate of drug release depending on the conditions (e.g., the number of bio-nanomachines at the target site) to achieve a sustained drug release.

- *Intracellular therapy:* Intracellular therapy is similar to drug delivery in the delivery of drug molecules to a target site, except that the target site is inside a cell (e.g., an intracellular compartment where pathogens are present) [59]. Intracellular therapy delivers drug molecules to a target site in an intracellular compartment, and it is highly challenging due to the fact that drug delivery carriers need to overcome a number of biological barriers (e.g., cell membranes and cellular processes such as endocytic events). Molecular communication may allow a group of bio-nanomachines to coordinate to reach a target site for intracellular therapy. For instance, functionally different bio-nanomachines may be introduced to modify the characteristics of a cell (e.g., permeability of the cell membrane) to bypass different types of biological barriers. These functionally different bio-nanomachines may also coordinate to detect multiple conditions to diagnose whether the cell is infected. When a cell is diagnosed as infected, bio-nanomachines carrying drug molecules (e.g., antiviral drugs) may release the drug molecules to combat the pathogens inside the cell.

1.3.3 Industrial applications

Molecular communication may also be used in some industries to produce a functional material from molecules. The agricultural industry, for instance, may benefit from food materials containing a number of bio-nanomachines through which the growth process of the food can be controlled. The manufacturing industry may also benefit from smart materials made of bio-nanomachines that have characteristics that are externally controllable or adaptive to changes in the environment. These functional materials may be several orders of magnitude larger than a single bio-nanomachine in size, and thus particular functionalities of such materials may emerge as a result of local interactions among groups of bio-nanomachines. Molecular communication is available for a group of bio-nanomachines to interact in a local environment, since molecular communication is inherently limited to the local environment. To change the functionality of a material, for instance, external stimuli (e.g., chemical, mechanical, electrical) may be applied to a material to initiate molecular communication processes among bio-nanomachines embedded in the material. Each bio-nanomachine responds to a change in a local environment, moves to a particular location, and/or transports molecules to a particular location. The movement of bio-nanomachines and molecules in local environments may lead to a change in global structure, and therefore modifies the functionality of the material.

1.3.4 Environmental applications

Molecular communication may also apply to monitoring molecules in an environment that may be contaminated with toxic or radioactive agents. To monitor a large area of

an environment, bio-nanomachines may be integrated into large or micro-scale devices (e.g., motes in wireless sensor networks) and these devices are deployed in the environment to form a large-scale biosensor network [60]. Bio-nanomachines in these devices detect molecules from the environment to provide early warning of contamination in that environment. Molecular communication may be useful to allow a group of bio-nanomachines to process molecules in a cooperative manner. For instance, one type of bio-nanomachine amplifies molecular signals from the environment, another type integrates different molecular signals to identify the location of a toxic source, another type guides the device to the location of the toxic agents, and yet another type degrades toxic molecules into a non-toxic or reusable form.

1.3.5 Information and communication technology applications

Molecular communication may also introduce a breakthrough into information and communication technology through the integration of bio-nanomachines into currently available silicon-based systems. For instance, a future mobile phone may be integrated with bio-nanomachines capable of molecular communication for on-chip analysis of biochemical signals (e.g., molecules in blood or from sweat) [61]. Such a device itself may be produced from a massive number of communicating bio-nanomachines and integrated with a human body. As another example, a dermal display screen is envisioned to be made from a population of 3 billion bio-nanomachines and embedded below the skin surface on a human body [62]. Massively distributed bio-nanomachines capable of molecular communication may also be integrated into the Internet to form the Internet-of-nano-things [63] and body-area nanonetworks [64]. In addition, molecular communication may apply to non-silicon-based computing paradigms, i.e., unconventional computation. In unconventional computation, research efforts are made to exploit physical, chemical, or biological materials to develop new computing architectures and to design algorithms for such architectures to solve computationally difficult problems. Unconventional computation has promising features such as extremely high functional complexity and large-scale parallelism that cannot be achieved with silicon-based electronic circuits. One promising approach is to use bio-nanomachines and molecular communication as the components for unconventional computation.

1.4 Rationale and organization of the book

As molecular communication is an interdisciplinary field, with elements of electrical engineering, mathematics, chemistry, and biology, it is unlikely that any new researcher in this field would have all the expertise required to quickly make a contribution. Moreover, the needed expertise is found in widely different literature across these disciplines.

Our objective in writing this book is to provide a first reference for new researchers in molecular communication, introducing them to the elements of the field; another objective is to provide enough background and references to allow new researchers to

explore the literature further on their own. Our book is directed both at traditional communications engineers who need a background in the biological principles of molecular communication, and at chemists and biologists who need exposure to information and communication theory.

The remainder of the book is organized as follows:

- In Chapter 2, we describe biological nanomachines, giving background information about their molecular structure and function.
- In Chapter 3, we describe biological molecular communication, describing how this communication method is used in nature, especially among the biological nanomachines described in Chapter 2.
- In Chapter 4, we describe the molecular communication paradigm, explaining the concept of engineered molecular communication.
- In Chapter 5, we give mathematical models for different types of molecular communication.
- In Chapter 6, we build on the mathematical models from Chapter 5 to describe the information and communication theory of molecular communication.
- In Chapter 7, building on all the material in previous chapters, we describe how to design molecular communication systems in practice.
- In Chapter 8, we give a detailed review of important applications of molecular communication.
- In Chapter 9, we present a short conclusion.

Generally, the material in Chapters 2 and 3 form the biological basis of our book, while the material in Chapters 5 and 6 form the theoretical basis. We have attempted to present the material in such a way as to assume a minimal background in either biology or communication theory, which we hope will give this book broad appeal, although the material may seem like review to some readers.

References

[1] S. Martel, "Nanorobots for microfactories to operations in the human body and robots propelled by bacteria," *Journal Facta Universitatis Series Mechanics, Automatic Control and Robotics*, vol. 7, pp. 1–10, 2009.

[2] D. G. Gibson *et al.*, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, no. 5987, pp. 52–56, 2010.

[3] T. Nakano, M. Moore, F. Wei, A. V. Vasilakos, and J. W. Shuai, "Molecular communication and networking: opportunities and challenges," *IEEE Transactions on NanoBioscience*, vol. 11, no. 2, pp. 135–148, 2012.

[4] T. Nakano and J. Shuai, "Repeater design and modeling for molecular communication networks," in *Proc. 2011 IEEE Infocom Workshop on Molecular and Nanoscale Communications,* 2011, pp. 501–506.

[5] R. P. Feynman, "There's plenty of room at the bottom," *Engineering and Science*, vol. 23, no. 5, pp. 22–36, 1960.

[6] M. B. Miller and B. L. Bassler, "Quorum sensing in bacteria," *Annual Review of Microbiology*, vol. 55, pp. 165–199, 2001.
[7] J. Hancock, *Cell Signalling*. Oxford University Press, 2010.
[8] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, pp. 379–423, Jul. 1948.
[9] H. Nyquist, "Certain factors affecting telegraph speed," *Bell System Technical Journal*, vol. 3, pp. 324–346, 1924.
[10] C. E. Shannon, "The bandwagon," *IRE Transactions on Information Theory*, vol. 2, no. 1, Mar. 1956.
[11] H. A. Johnson and K. D. Knudsen, "Renal efficiency and information theory," *Nature*, vol. 206, pp. 930–931, 1965.
[12] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice Hall, 1971.
[13] R. M. Gray, "Rate Distortion Theory: A Mathematical Basis for Data Compression (book review)," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 217–218, Jan. 1972.
[14] D. Blackwell, *Information Theory*, Modern mathematics for the engineer: Second series. New York: McGraw-Hill, 1961, pp. 182–193.
[15] H. Permuter, P. Cuff, B. V. Roy, and T. Weissman, "Capacity of the chemical channel with feedback," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2008.
[16] T. Berger, "Living information theory (Shannon lecture)," in *Proc. IEEE International Symposium on Information Theory, Lausanne, Switzerland*, 2002.
[17] P. B. Detwiler, S. Ramanathan, A. Sengupta, and B. I. Shraiman, "Engineering aspects of enzymatic signal transduction: Photoreceptors in the retina," *Biophysical Journal*, vol. 79, pp. 2801–2817, 2000.
[18] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," *Sensory Communication*, pp. 217–234, 1961.
[19] S. B. Laughlin, "A simple coding procedure enhances a neuron's information capacity," *Zeitschrift für Naturforschung*, vol. 36, pp. 910–912, 1981.
[20] P. J. Thomas, D. J. Spencer, S. K. Hampton, P. Park, and J. P. Zurkus, "The diffusion mediated biochemical signal relay channel," in *Proc. 17th Annual Conference on Neural Information Processing Systems*, 2003.
[21] A. W. Eckford, "Nanoscale communication with Brownian motion," in *Proc. Conf. on Information Sciences and Systems*, 2007, pp. 160–165.
[22] A. Einolghozati, M. Sardari, A. Beirami, and F. Fekri, "Capacity of discrete molecular diffusion channels," in *IEEE International Symposium on Information Theory*, 2011.
[23] B. Atakan and O. Akan, "An information theoretical approach for molecular communication," in *Proc. 2nd Intl. Conf. on Bio-Inspired Models of Network, Information, and Computing Systems*, 2007, pp. 33–40.
[24] S. Hiyama, Y. Moritani, T. Suda, R. Egashira, A. Enomoto, M. Moore, and T. Nakano, "Molecular communication," in *Proc. 2005 NSTI Nanotechnology Conference*, 2005, pp. 391–394.
[25] T. Nakano, T. Suda, M. Moore, R. Egashira, and K. Arima, "Molecular communication for nanomachines using intercellular calcium signalling," in *IEEE International Conference on Nanotechnology*, 2005, pp. 478–481.

[26] T. Nakano, T. Suda, T. Kojuin, T. Haraguchi, and Y. Hiraoka, "Molecular communication through gap junction channels: System design, experiments and modeling," in *Proc. 2nd International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2007, pp. 139–146.

[27] Y. Moritani, S. Hiyama, and T. Suda, "Molecular communication among nanomachines using vesicles," in *Proc. NSTI Nanotechnology Conference*, 2006, vol. 2, pp. 705–708.

[28] Y. Moritani, S.-M. Nomura, S. Hiyama, T. Suda, and K. Akiyoshi, "A communication interface using vesicles embedded with channel forming proteins in molecular communication," in *Proc. 2nd International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2007, pp. 147–149.

[29] M. Kaneda, S.-M. Nomura, S. Ichinose, S. Kondo, K. Nakahama, K. Akiyoshi, and I. Morita, "Direct formation of proteo-liposomes by in vitro synthesis and cellular cytosolic delivery with connexin-expressing liposomes," *Biomaterials*, vol. 30, pp. 3971–3977, 2009.

[30] A. Enomoto, M. Moore, T. Nakano, R. Egashira, T. Suda, A. Kayasuga, H. Kojima, H. Sakibara, and K. Oiwa, "A molecular communication system using a network of cytoskeletal filaments," in *Proc. 2006 NSTI Nanotechnology Conference*, 2006, pp. 725–728.

[31] S. Hiyama, R. Gojo, T. Shima, S. Takeuchi, and K. Sutoh, "Biomolecular motor-based nano- or microscale particle translocations on DNA microarrays," *Nano Letters*, vol. 9, pp. 2407–2413, 2009.

[32] S. Hiyama and Y. Moritani, "Molecular communication: Harnessing biochemical materials to engineer biomimetic communication systems," *Nano Communication Networks*, vol. 1, pp. 20–30, 2010.

[33] I. Akyildiz, F. Brunettib, and C. Blázquezc, "Nanonetworks: A new communication paradigm," *Computer Networks*, vol. 52, no. 12, pp. 2260–2279, 2008.

[34] S. F. Bush, *Nanoscale Communication Networks*. Artech House, 2010.

[35] L. Parcerisa and I. F. Akyildiz, "Molecular communication options for long range nanonetworks," *Computer Networks*, vol. 53, no. 16, pp. 2753–2766, Nov. 2009.

[36] M. Pierobon and I. F. Akyildiz, "A physical end-to-end model for molecular communication in nanonetworks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 4, pp. 602–611, May 2010.

[37] M. Moore, T. Suda, and K. Oiwa, "Molecular communication: Modeling noise effects on information rate," *IEEE Transactions on Nanobioscience*, vol. 8, pp. 169–190, Jun. 2009.

[38] A. Einolghozati, M. Sardari, and F. Fekri, "Capacity of diffusion-based molecular communication with ligand receptors," in *IEEE Information Theory Workshop*, 2011.

[39] L. Cui and A. W. Eckford, "The delay selector channel: Definition and capacity bounds," in *Proc. Canadian Workshop on Information Theory (CWIT)*, 2011.

[40] T. Nakano, Y. Okaie, and J. Q. Liu, "Channel model and capacity analysis of molecular communication with Brownian motion," *IEEE Communications Letters*, vol. 16, no. 6, pp. 797–800, Jun. 2012.

[41] R. Song, C. Rose, Y.-L. Tsai, and I. S. Mian, "Wireless signalling with identical quanta," in *Proc. Wireless Communication and Networking Conf. (WCNC)*, 2012, pp. 699–703.

[42] M. J. Moore, T. Nakano, A. Enomoto, and T. Suda, "Measuring distance with single-spike molecular communication feedback protocols," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3576–3587, Jul. 2012.

[43] N. Farsad, A. W. Eckford, and S. Hiyama, "Modelling and design of polygon-shaped kinesin substrates for molecular communication," in *Proc. IEEE International Conference on Nanotechnology*, 2012.

[44] P. Lio and S. Balasubramaniam, "Opportunistic routing through conjugation in bacteria communication nanonetwork," *Nano Communication Networks*, vol. 3, no. 1, pp. 36–45, Mar. 2012.

[45] M. Kuran, H. Yilmaz, T. Tugcu, and I. F. Akyildiz, "Interference effects on modulation techniques in diffusion based nanonetworks," *Nano Communication Networks*, vol. 3, no. 1, pp. 65–73, Mar. 2012.

[46] L. Felicetti, M. Femminella, and G. Reali, "A simulation tool for nanoscale biological networks," *Nano Communication Networks*, vol. 3, no. 1, pp. 2–18, Mar. 2012.

[47] B. Atakan, O. B. Akan, and S. Balasubramaniam, "Body area nanonetworks with molecular communications in nanomedicine," *IEEE Communications Magazine*, vol. 50, no. 1, pp. 28–34, Jan. 2012.

[48] P. Bogdan, G. Wei, and R. Marculescu, "Modeling populations of micro-robots for biological applications," in *Proc. 2nd IEEE Workshop on Molecular and Nanoscale Communication*, 2012.

[49] T. Nakano, "Swarming biological nano machines through molecular communication for targeted drug delivery," in *Proc. 6th International Conference on Soft Computing and Intelligent Systems/13th International Symposium on Advanced Intelligent Systems*, 2012, pp. 2317–2320.

[50] IEEE P1906.1 – recommended practice for nanoscale and molecular communication framework, http://standards.ieee.org/develop/project/1906.1.html.

[51] M. Moore, A. Enomoto, T. Suda, T. Nakano, and Y. Okaie, *The Handbook of Computer Networks*. John Wiley & Sons Inc, 2007, vol. 3, ch. Molecular communication: new paradigm for communication among nano-scale biological machines, pp. 1034–1054.

[52] P. Yager, T. Edwards, E. Fu, K. Helton, K. Nelson, M. R. Tam, and B. H. Weigl, "Microfluidic diagnostic technologies for global public health," *Nature*, vol. 442, pp. 412–418, 2006.

[53] P. S. Dittrich and A. Manz, "Lab-on-a-chip: microfluidics in drug discovery," *Nature Reviews Drug Discovery*, vol. 5, pp. 210–218, 2006.

[54] C. Teuscher, C. Grecu, T. Lu, and R. Weiss, "Challenges and promises of nano and bio communication networks," in *Fifth ACM/IEEE International Symposium on Networks-on-Chip*, 2011, pp. 247–254.

[55] L. G. Griffith and G. Naughton, "Tissue engineering – current challenges and expanding opportunities," *Science*, vol. 295, no. 5557, pp. 1009–1014, 2002.

[56] J. Clausen, "Man, machine and in between," *Nature*, pp. 1080–1081, 2009.

[57] T. M. Allen and P. R. Cullis, "Drug delivery systems: entering the mainstream," *Science*, vol. 303, no. 5655, pp. 1818–1822, 2004.

[58] J.-W. Yoo, D. J. Irvine, D. E. Discher, and S. Mitragotri, "Bio-inspired, bioengineered and biomimetic drug delivery carriers," *Nature Reviews Drug Discovery*, vol. 10, pp. 521–535, 2011.

[59] A. Ranjan, N. Pothayee, M. N. Seleem, S. M. Boyle, K. Ramanathan, J. S. Riffle, and N. Sriranganathan, "Nanomedicine for intracellular therapy," *FEMS Microbiology Letters*, pp. 1–9, 2012.

[60] R. Byrne and D. Diamond, "Chemo/bio-sensor networks," *Nature Materials*, vol. 5, pp. 421–424, 2006.

[61] S. Hiyama, Y. Moritani, and T. Suda, "Molecular transport system in molecular communication," *NTT DOCOMO Technical Journal*, vol. 10, no. 3, pp. 49–53, 2008.

[62] R. A. Freitas Jr, *Nanomedicine, vol. I: Basic Capabilities*. Landes Bioscience, 1999.

[63] I. F. Akyildiz and J. M. Jornet, "The internet of nano-things," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 58–63, 2010.

[64] B. Atakan, O. B. Akan, and S. Balasubramaniam, "Body area nanonetworks with molecular communications in nanomedicine," *IEEE Communications Magazine*, vol. 50, no. 1, pp. 28–34, 2012.

2 Nature-made biological nanomachines

Nature has evolved various forms of biological nanomachines – small-scale devices composed of chemically reacting biological molecules. Simply referred to as bio-nanomachines in this book, they consist of molecules that are abundantly found in living organisms, such as carbohydrates, lipids, proteins, and nucleic acids. They are in the nanometer to micrometer range and thus not visible to the human eye. They are machines capable of biochemical interaction with molecules. Figure 2.1 shows some examples of bio-nanomachines, including protein molecules that catalyze chemical reactions (i.e., enzymes), regulate flow of molecules (transport channels), or produce motion using chemical energy (motor proteins); deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) molecules that store genetic information; vesicles that mediate the transport of protein molecules within cells; and viruses that infect biological cells to replicate. Bio-nanomachines in this book also extend to cellular organelles that provide specific functions within cells and even whole cells that are built from billions of bio-nanomachines capable of interacting with a wide variety of molecules in the environment.

A single bio-nanomachine can be viewed as a functional unit that interacts with *molecular signals* [1, 2]. A bio-nanomachine may respond to input signals by transmitting output signals, changing its internal state, or modifying its functionality. For instance, an enzyme, a catalyst of chemical reactions, responds to specific substrate molecules by producing product molecules. A DNA molecule, a storage of genetic information, responds to molecular signals in the cell by changing its state by switching on and off particular genes. A stem cell, a biological cell found in all multicellular organisms, responds to molecular signals in the environment by differentiating into specialized cell types with particular functions (e.g., muscle cells).

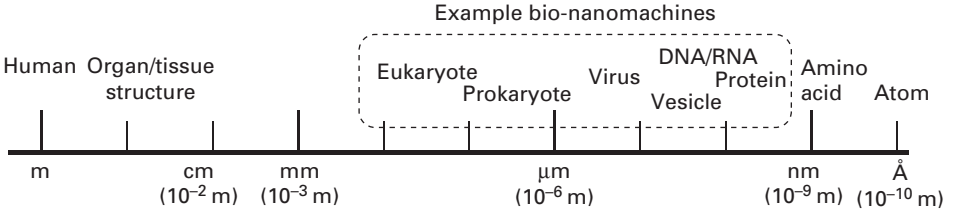


Figure 2.1 Examples and length scales of nature-made bio-nanomachines.

In this chapter, we will briefly review the naturally occurring bio-nanomachines in terms of biochemical structures and functional roles in biological systems. Specific bio-nanomachines reviewed in this chapter include protein molecules, DNA and RNA molecules, lipid membranes and vesicles, as well as whole cells that consist of large numbers of bio-nanomachines. This chapter is written for communication engineers who have no prior knowledge of bio-nanomachines. For such communications engineers, basic biology terms that appear throughout this book as well as in other molecular communication literature are in bold the first time that they appear.

2.1 Protein molecules

Proteins are one of the most basic building blocks of nature-made biological systems. We first review the biochemical structure of proteins and then look at the major roles of proteins in biological systems such as functioning as catalysts to induce chemical reactions, sensors to process molecular signals, and actuators to produce motion.

2.1.1 Molecular structure

The molecular structure of a protein molecule is a linear chain of **amino acids** (Figure 2.2 left). An amino acid is made from one amino group ($-NH_2$), one carboxyl group ($-COOH$), and a side chain called an R group specific to each amino acid. In a human being, there are twenty different types of side chains and thus twenty different amino acids, such as the well-known glutamate (Glu). Two amino acids are chained together by the chemical linkage called the **peptide bond** that is formed between the amide nitrogen atom of one amino acid and the carbonyl carbon atom of another amino acid. Similarly, many more amino acids are combined by peptide bonds to extend the linear chain. One end of such a linear chain has a free amino group called the N-terminus and the other end a carboxyl group called the C-terminus. A short chain, typically 20–30 amino acid residues, is called a **peptide** while a longer chain up to 4000 amino acid residues is called a polypeptide chain or a protein.

The protein structure described above (i.e., a linear chain of amino acids) is referred to as the primary structure. In biological cells, a protein molecule forms several levels of higher-order structure that are called the secondary, tertiary, or quaternary structure (Figure 2.2 right). The secondary structure refers to a stable local structure formed through local interactions among the amino acid residues of a polypeptide chain; the alpha helix and beta sheet are the two common secondary structures that are both formed by strong bonds between amino acid residues. The tertiary structure refers to the global structure of a polypeptide chain in which sets of secondary structures assemble together to form functionally distinct elements called **domains** (e.g., binding domains allow interactions with specific molecules). A protein molecule often consists of multiple polypeptide chains; individual polypeptide chains are called subunits. The type and number of subunits are referred to as the quaternary structure, which affects the function of the whole protein molecule.

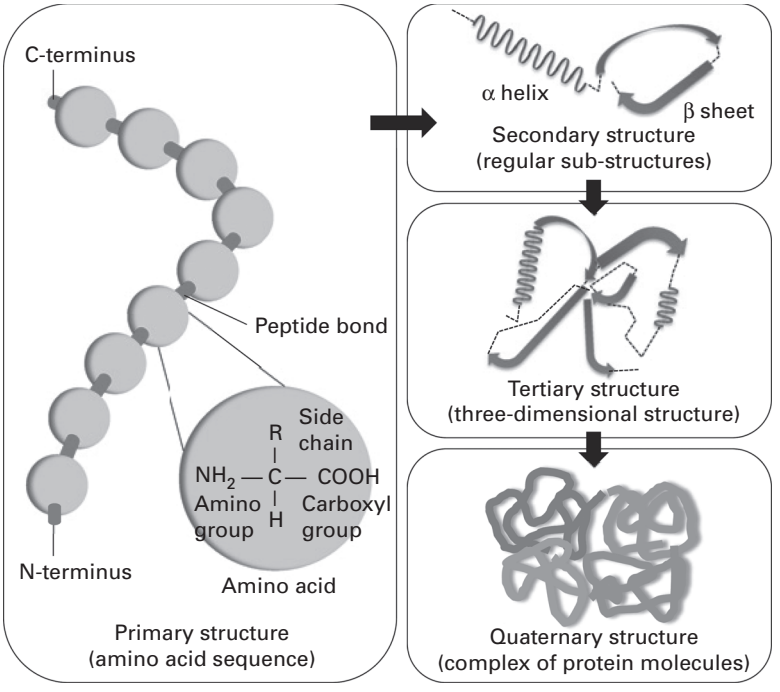


Figure 2.2 Structure of a protein molecule.

The higher-order structure or function of nearly every protein molecule can be chemically modified. One common modification is **phosphorylation** by which a phosphate group (PO_4^{3-}) is added to a specific amino acid residue of the protein molecule. The phosphorylation of a protein molecule often *activates* the molecule, meaning that the energy level of the molecule is elevated to promote a subsequent chemical reaction.

2.1.2 Functions and roles

The structure of a protein molecule is related to its function. The following sections describe three major functions of proteins: catalyzing chemical reactions, signaling and sensing molecules, and generating motion.

Enzymes catalyze chemical reactions

One major function of proteins is to catalyze chemical reactions – the proteins that function as a catalyst are called **enzymes**. An enzyme has a catalytic domain where substrates are converted into products. In its simplest form (Figure 2.3), (a) an enzyme E binds to a particular substrate S to form an enzyme-substrate complex ES; (b) ES then either dissociates back into E and S or (c) converts into E and product(s) P. Notice that the enzyme itself is not modified by the reaction. In biological systems, enzymes are involved in many chemical reactions that occur extremely slowly or are unlikely to occur without an enzyme.

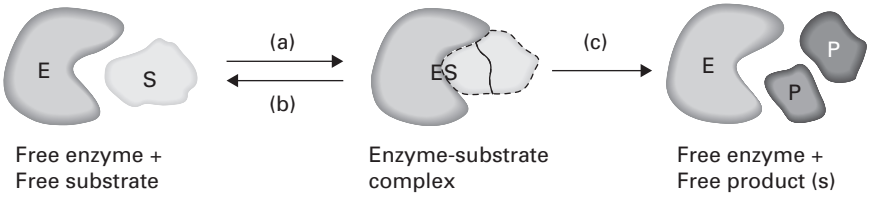


Figure 2.3 An enzymatic reaction converts substrate into product.

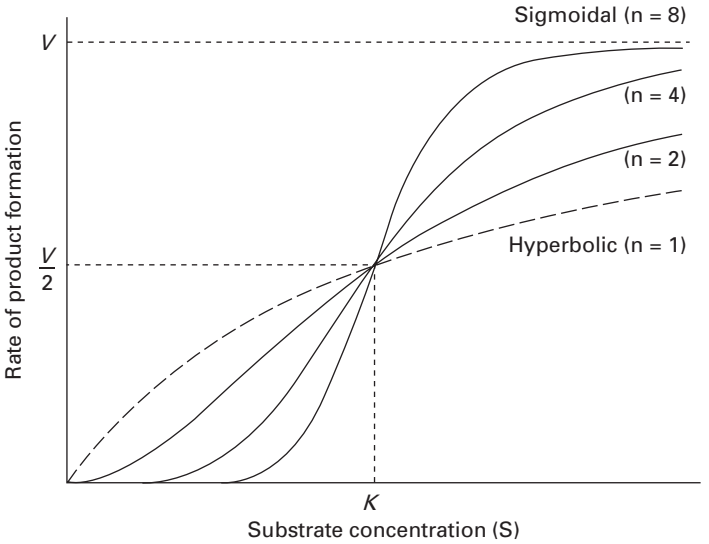


Figure 2.4 The rate of product formation as a function of the substrate concentration.

The chemical kinetics of an enzymatic reaction can be schematically described as follows.

$$E + S \rightleftarrows ES \longrightarrow E + P \tag{2.1}$$

The rate of an enzymatic reaction is described by the well-known Michaelis–Menten law, which states that the reaction rate at which the product is formed per unit time increases with the concentration of the substrate. The Michaelis–Menten law approximates the reaction rate using a simple hyperbolic function:

$$\frac{V[S]}{K+[S]}, \tag{2.2}$$

where V is the maximum reaction rate that is achieved when the reaction is saturated; $[S]$ is the concentration of the substrate; and K is the Michaelis–Menten constant, which is the substrate concentration that achieves half of the maximum rate (i.e., $\frac{V}{2}$). (See Figure 2.4 for an example case and Box 2.1 for derivation of the Michaelis–Menten equation.)

Box 2.1 Michaelis–Menten kinetics

The Michaelis–Menten equation (2.2) is derived as follows. The enzymatic reaction (2.1) involves (a) the formation of ES, (b) dissociation of ES, and (c) production of P (see Figure 2.3). By using the three reaction rate constants, k_1 for (a), k_{-1} for (b), and k_2 for (c), (2.1) is re-written as

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P$$

The time evolution of concentrations of S, E, ES, and P are then given as

$$\begin{aligned}
\frac{d[E]}{dt} &= (k_{-1} + k_2)[ES] - k_1[E][S], \\
\frac{d[S]}{dt} &= k_{-1}[ES] - k_1[E][S], \\
\frac{d[ES]}{dt} &= k_1[E][S] - (k_2 + k_{-1})[ES], \\
\frac{d[P]}{dt} &= k_2[ES].
\end{aligned}$$

The rates of formation and dissociation of ES are by convention assumed to be equal all the time ($\frac{d[ES]}{dt} = 0$); and under this assumption, we have the Michaelis–Menten equation (2.2):

$$\frac{d[P]}{dt} = \frac{V[S]}{K + [S]}$$

with

$$\begin{aligned}
V &\triangleq k_2([E] + [ES]), \\
K &\triangleq \frac{k_{-1} + k_2}{k_1}.
\end{aligned}$$

Note that V represents the maximum product formation rate and K the substrate concentration that achieves half of the maximum product formation rate (see Figure 2.4).

The enzyme kinetics considering the cooperativity, (2.3) and (2.4), can be similarly described with a set of ordinary differential equations. The product formation rate in this case is approximated under the assumption that the formation and dissociation rates are equal for ES and ESS. In general, when an enzyme with n binding sites is involved, the product formation rate is given as (2.5). Note that n is obtained from experiments and it is typically a non-integer value.

The chemical kinetics of an enzymatic reaction is more complicated when an enzyme has multiple binding sites. When an enzyme has two binding sites for the same substrate, for instance, the enzyme takes one of the three forms: free enzyme with two binding sites both unoccupied (E), the enzyme with one binding site occupied with one substrate (ES), or the enzyme with both binding sites occupied with two substrates (ESS). The chemical reaction may be described as follows.

$$E + S \rightleftarrows ES \longrightarrow E + P \tag{2.3}$$

$$ES + S \rightleftarrows ESS \longrightarrow ES + P \tag{2.4}$$

Binding of multiple substrates on an enzyme may be **cooperative**. In one case, the binding of one substrate at a binding site may increase the affinity of the substrate binding at the other site. The rate of product formation with cooperative binding appears as a sigmoidal function with a parameter n, the number of binding sites; that is,

$$\frac{V[S]^n}{K^n + [S]^n}. \tag{2.5}$$

The function is often referred to as the **Hill function** (with respect to $[S]$) and n is the Hill coefficient. As shown in Figure 2.4, an enzyme with a large n value controls the rate of product formation like a digital switch.

Ligands and receptors transduce signals

Another major function of proteins within biological cells is to transduce signals. Signaling molecules, referred to as **ligands** (e.g., hormones, growth factors), bind to specific **receptors** that recognize the type of signaling molecule with high specificity and induce chemical reactions. Proteins or peptides often function as ligands as well as receptors to form ligand–receptor systems that are commonly found throughout biological cells.

One common and important type of protein receptor is a cell-surface receptor that has three spatially different domains: extracellular, transmembrane, and intracellular (Figure 2.5). The extracellular domain projects out of the cell membrane and is exposed to the extracellular space. The extracellular domain typically has binding sites

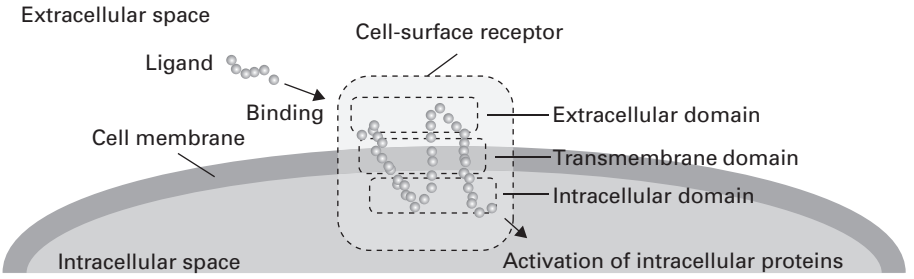


Figure 2.5 Signal transduction through cell-surface receptors.

to recognize ligands in the extracellular space. The transmembrane domain is stably located in the cell membrane and provides a mechanism to transfer signals from the extracellular domain to the intracellular domain. The intracellular domain is exposed to the intracellular space and provides a mechanism to interact with the intracellular components (e.g., proteins) to relay extracellular signals into the intracellular space.

One mechanism of signal transduction utilizes the **allostericity** of a cell-surface receptor. Allosteric reactions are a common mechanism found in many proteins, by which the binding of a molecule (i.e., a ligand) at one side of a protein (i.e., the extracellular domain) changes the conformation at the other side (i.e., the intracellular domain). A common example of this signaling involves the intracellular domain of the receptor coupled with a large trimeric protein called a guanine nucleotide-binding protein (G-protein) or coupled to an enzyme called a tyrosine kinase. When the extracellular domain of the receptor binds to a ligand, the intracellular domain changes its conformation to dissociate the subunits of the coupled G-protein or activate the coupled tyrosine kinase through phosphorylation, leading to subsequent reactions within the cell.

Another mechanism of transducing signals from the extracellular domain to the intracellular domain is for the transmembrane domain of a cell-surface receptor to be a pore structure or a **channel** through which ions and small molecules pass from the extracellular space to the intracellular space. One type of channel is an **ion channel** that opens when a ligand binds to its extracellular domain and allows ions in the intracellular and extracellular space to diffuse across the cell membrane.

Motor proteins generate motion

Another major function of proteins in biological cells is to generate motion. **Motor proteins**, also called molecular motors, such as myosin, kinesin, and dynein, are specialized at generating motion. Motor proteins are mechanochemical enzymes that convert chemical energy into a mechanical force. One prominent role of motor proteins in biological cells is performing intracellular trafficking by moving directionally along **molecular rails** (or rail proteins) made of protein polymers and transporting molecules to specific locations within cells.

A motor protein consists of functional domains including the head, neck, and tail domains (Figure 2.6). The head domain contains two heads and each head has two binding sites: one for binding to a molecular rail and the other for binding to adenosine triphosphate (**ATP**), the source of chemical energy, to generate a force. The neck domain provides a flexible link connecting the head and tail domains. The tail domain attaches to a cargo containing some molecules. In order for a motor protein to walk, one head binds to an ATP and catalyzes a chemical reaction called **ATP hydrolysis**. ATP hydrolysis detaches the phosphate from an ATP to produce an adenosine diphosphate (ADP) and at the same time releases energy. The motor protein absorbs the energy to change its conformation and moves along the rail. In one model, one head is attached to the rail, while the other head, which is detached from the rail, makes one step forward and then attaches to the rail. The first head is then detached to make another step. A repeating cycle of such motion leads to the directional movement of the motor protein from one end of the rail toward the other end.

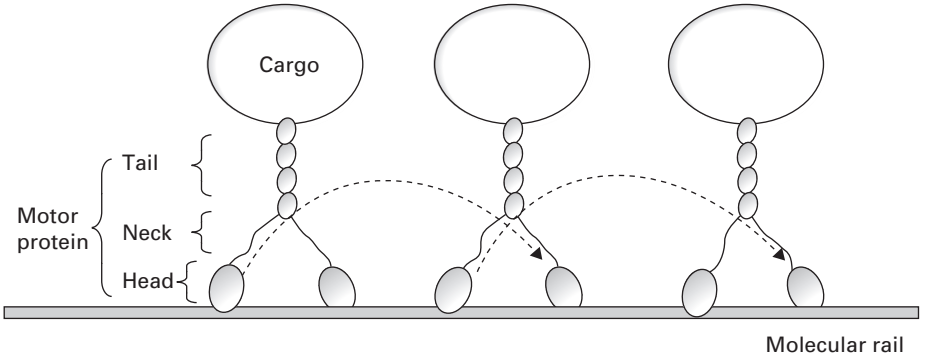


Figure 2.6 Motor proteins walk over the molecular rail.

Molecular rails that direct the movement of motor proteins form a structure within a cell. Two major types of molecular rails found in biological cells are **microtubules** and **actin filaments**. A microtubule is assembled from globular proteins called tubulins, on which kinesin and dynein motors walk. An actin filament is made from different globular proteins called actins, on which myosin motors walk. The structure of molecular rails is maintained in cells through dynamic processes. A microtubule, for instance, dynamically grows by assembly of tubulin onto its "plus" end and shrinks by disassembly of tubulin from its other "minus" end, through the process known as **dynamic instability**.

2.2 DNA and RNA molecules

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are well known as the molecules encoding the genetic information of an organism. Here we first look at the molecular structure of DNA and RNA molecules. We then give an overview of the functional roles of these molecules in the process of protein synthesis.

2.2.1 Molecular structure

DNA and RNA molecules are structurally defined as **nucleic acids**. Similar to how the primary structure of a protein molecule is described, the primary structure of a nucleic acid is a linear chain of repeating units, but instead the units are **nucleotides** (Figure 2.7A). Each nucleotide is composed of a nucleobase, a five-carbon sugar, and one phosphate group. The nucleobase is either adenine (A), cytosine (C), guanine (G), thymine (T), or uracil (U), among which A, C, G, and T are found in DNA and are called **DNA bases**, and A, C, G, and U are found in RNA and are called **RNA bases**. The five-carbon sugars are different in DNA and RNA: deoxyribose in DNA and ribose in RNA. The five-carbon sugar has five carbon atoms at positions identified as $1' - 5'$ (one prime to five prime) and the $5'$ carbon atom is bonded with the phosphate. Two adjacent nucleotides are connected through a strong covalent bond, called the phosphodiester

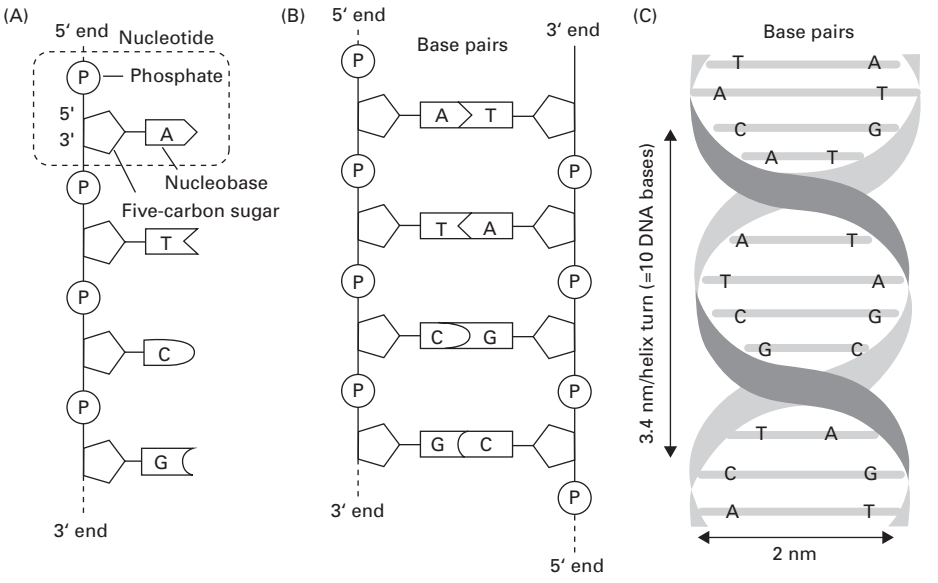


Figure 2.7 DNA. (A) Structure of a DNA molecule. (B) and (C) Two DNA molecules assemble to form a double-helical structure through the base-pairing rules.

bond, formed between the 3′ carbon atom of one nucleic acid and the phosphate of the other. The chain structure is thus directional and by convention read from the 5′ end to the 3′ end.

In cells, two DNA molecules, often referred to as **DNA strands** or **sequences**, typically assemble to form a double-helical structure as shown in Figure 2.7B–C. This double-helical structure is made through the **base-pairing rules**; A and T make one base-pair and G and C the other base-pair, each pair binds through a hydrogen bond. Notice that the two DNA strands are aligned in anti-parallel with one strand in the 3′ to 5′ direction and the other in the 5′ to 3′ direction. Also notice that the DNA bases are found inside the double-helical structure and protected by the sugar and phosphate backbones. Two DNA strands are said to be **complementary** when DNA bases in one strand forms base-pairs with DNA bases in the other. Two DNA strands may be completely complementary as shown in Figure 2.7B (i.e., 5′ATCG3′ and 3′TAGC5′), partially complementary, or not complementary at all. The double-helical structure has a diameter of about 2 nm. About 10 nucleotides are found per helix turn, which is a length of about 3.4 nm (per turn). A DNA molecule is a large molecule with millions of bases, but within cells it is compactly packed as an organized structure called a **chromosome**.

Unlike DNA molecules, RNA molecules are typically single-stranded in cells. The base-pairing rules however still apply to RNA bases, where A and U make one base-pair and G and C make the other. The base-pairing rules allow parts of an RNA molecule to fold into a local structure such as a hairpin or a loop. As a result an RNA molecule may exhibit a high degree of structural complexity, similar to protein molecules. Some RNA

molecules in fact can function as a catalyst of a chemical reaction similar to protein enzymes.

2.2.2 Functions and roles

DNA and RNA control virtually all chemical processes occurring in cells by regulating the gene expression of cells. (A **gene** is a protein coding region of a DNA molecule, and a gene is said to be *expressed* when the protein is synthesized.) In the process of gene expression, DNA functions as a storage of genetic information and RNA is involved in decoding the genetic information to synthesize proteins. The gene expression is described by the central dogma of molecular biology with two processes: DNA is copied into RNA through **transcription** and RNA synthesizes proteins through **translation**. Here we focus on the two processes, transcription and translation, and look at how DNA and RNA function to produce proteins.

Transcription is the process of copying a part of a DNA molecule into an RNA molecule, called a **messenger RNA** (mRNA). The transcription process is catalyzed by a protein enzyme, **RNA polymerase** (RNA pol), that moves in the 3′ to 5′ direction along one of the two DNA strands during the process (Figure 2.8). First, the RNA pol recognizes and binds to a specific sequence of a DNA strand called a **promoter** to initiate the transcription process. The RNA pol then starts reading the DNA strand

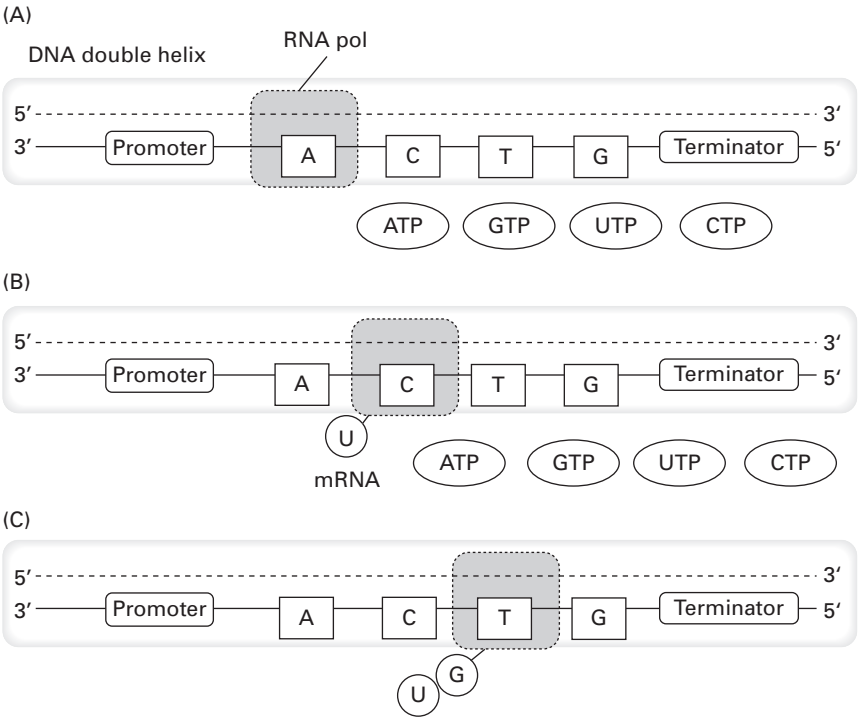


Figure 2.8 Transcription produces mRNA from a sequence of a DNA strand.

one nucleobase at a time. Following base-pairing rules for the nucleobase found in the DNA strand, the RNA pol produces a nucleobase for the mRNA by consuming a complementary nucleoside triphosphate (NTP) available in the environment; i.e., adenosine triphosphate (ATP), uridine triphosphate (UTP), cytidine triphosphate (CTP), or guanosine triphosphate (GTP). For instance, if A is found in the DNA strand, then UTP is consumed (Figure 2.8A) to produce A's base-pairing partner, U (Figure 2.8B). The RNA pol, in this way, produces and concatenates neucleobases to elongate the mRNA (Figure 2.8C). The RNA pol may end the transcription process when it encounters a specific sequence of DNA called a **terminator**. The result of the transcription process is an mRNA with a sequence complementary to a part of the DNA strand.

Translation is the process of synthesizing a protein from a sequence of mRNA produced from a transcription. The sequence of mRNA determines the primary structure of a protein where a sequence of three nucleotides called a **codon** encodes a particular amino acid. For instance, the sequence GAA encodes a glutamic acid. There are $4^3 = 64$ district codons, each of which encodes one of the 20 amino acids with a few exceptions, including a start codon (in most cases, AUG) where translation is initiated, and three stop codons where no amino acid is produced and translation is terminated. In translation, encoding of a codon into an amino acid is mediated by specific RNA molecules called **transfer RNA** (tRNA). Each tRNA contains a sequence of three nucleotides, called an anticodon, which is complementary to a particular codon according to base-pairing rules; and it also carries an amino acid specific to the anticodon. The process of translation is catalyzed by a large RNA-protein complex called the **ribosome** that has two catalytic sites to promote reactions. First, a specific type of tRNA binds to the start codon on the mRNA through the codon anticodon base-pairing (Figure 2.9A). Similarly, another type of tRNA binds to the next codon through the codon anticodon base-pairing (Figure 2.9B). The ribosome then catalyzes a chemical reaction to concatenate the amino acid of the first tRNA and that of the next tRNA through a peptide bond (Figure 2.9C). The ribosome also removes the first tRNA from the mRNA and moves in the $3'$ direction by one codon (Figure 2.9D), so that it becomes ready to bind the next tRNA. This process is repeated until the ribosome finds a stop codon where the translation is terminated and a protein is released.

2.3 Lipid membranes and vesicles

Vesicles are small *containers* of molecules enclosed by typically spherical lipid membranes. In this section, we give an overview of the structure of vesicles and how vesicles contribute to the functioning of biological cells.

2.3.1 Molecular structure

The membrane of a vesicle is made of two layers of phospholipid molecules, commonly referred to as a **lipid bilayer** (Figure 2.10). Each phospholipid molecule in a lipid bilayer has one hydrophilic head and two hydrophobic tails. The hydrophilic head

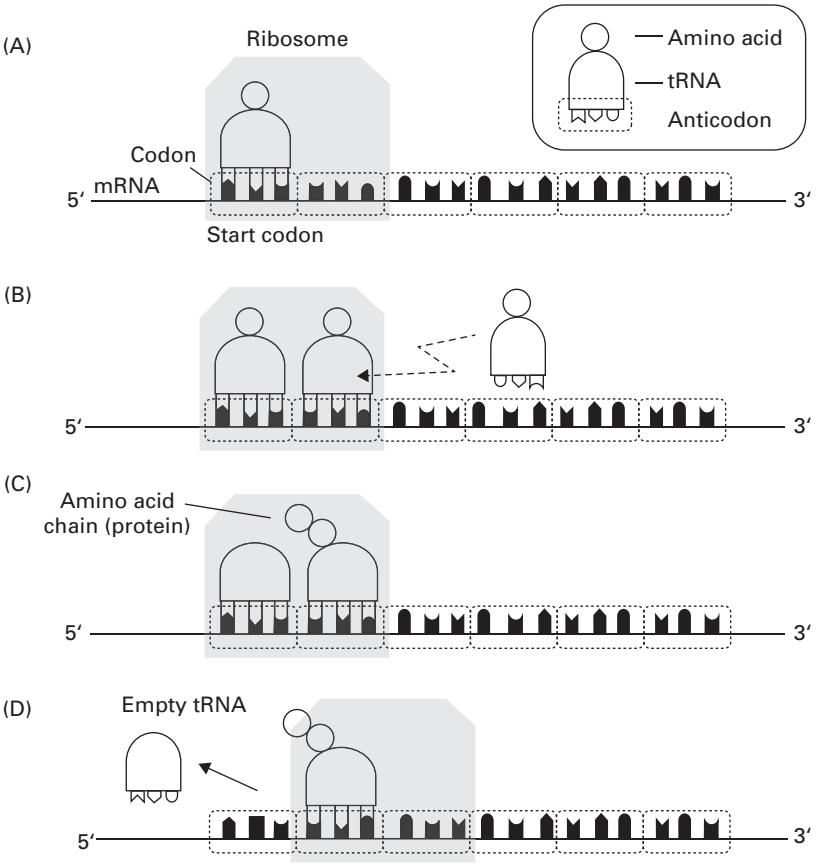


Figure 2.9 Translation produces a protein from a sequence of mRNA.

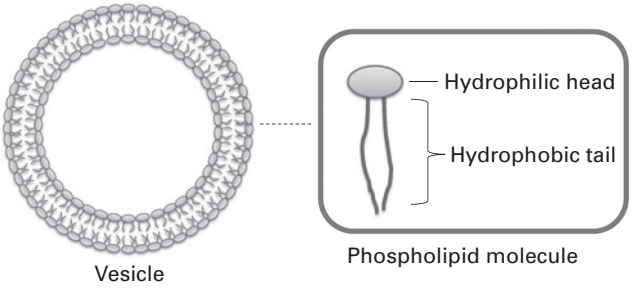


Figure 2.10 A vesicle is made of a lipid bilayer.

has a negatively charged phosphate group, and it is attracted to water molecules. The hydrophobic tail is a fatty acid hydrocarbon chain (e.g., cholesterol), and it is repelled by water molecules. The amphiphilic nature of phospholipid molecules allows the phospholipid molecules to self-assemble into a spherical structure made from a lipid bilayer,

where the two layers are aligned with hydrophobic tails against one another and with hydrophilic heads facing water molecules (i.e., facing inside the vesicle or facing the outside environment). The lipid bilayer is about 5 nm thick, and the diameter of a vesicle varies from several tens of nm to several μm. A vesicle is called a **liposome** when it is prepared artificially.

2.3.2 Functions and roles

One major function of vesicles is to contain different molecules and localize chemical reactions. Large or charged molecules (e.g., glucose molecules, ions) are generally not permeable to the lipid bilayer of a vesicle and they can be contained in the local environment made by the vesicle. A vesicle can be embedded with channel proteins through which molecules that are not permeable to the lipid bilayer are transported and become contained inside the vesicle. Chemical reactions in a vesicle can be efficient or different since molecules are localized in a small and separate environment. Most small and uncharged molecules are permeable to a lipid bilayer, and chemical reactions may also involve these lipid-bilayer permeable molecules in the environment.

In biological cells, vesicles often function as a carrier of molecules. Vesicles encapsulate molecules and motor proteins transport the vesicles and molecules from one location to another in cells (i.e., a vesicle is the cargo of the motor protein in Figure 2.6). For vesicle transport, a vesicle may be created from the membrane of an organelle through the process called **budding**, during which molecules in the organelle are stored inside the vesicle as the vesicle is created. Once the vesicle containing some molecules has been transported to a specific location, it may be integrated into the membrane of another organelle through the process called **fusion**, during which molecules in the vesicle are moved into the organelle. Since an organelle is also enclosed by a lipid bilayer, an organelle is considered a type of vesicle (Figure 2.11); in a sense, budding is the process of separating a set of molecules into two different sets, fusion is the process of combining two sets of molecules into one set, and the new sets of molecules may induce different chemical reactions.

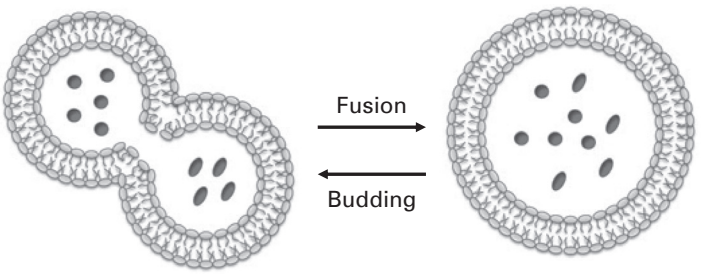


Figure 2.11 Budding and fusion of vesicles.

2.4 Whole cells

Having examined the structure and function of the three types of functional materials: proteins, DNA and RNA molecules, and vesicles – defined as bio-nanomachines in this book – we now look into how these bio-nanomachines are put together within a biological cell, i.e., a nature-made system of bio-nanomachines.

Biological cells are classified into either prokaryotes or eukaryotes depending on the complexity of their internal structure. Both types of cell are enclosed by a lipid bilayer called the **plasma membrane** and contain numerous molecules in the internal space called the **cytosol**, but they differ in how the internal space is organized. A prokaryote has a simple structure without a nucleus and organelles (Figure 2.12A). Various types of functional molecules are thus found in the cytosol as well as on the plasma membrane, including (1) cell-surface receptors for sensing molecules in the environment, (2) motor and rail proteins to give structure and shape to the cell (called the cytoskeleton) and to produce motion by, for example, a flagellum, and (3) DNA, RNA, and ribosomes for regulating protein synthesis. A representative example of a prokaryote is a bacterium such as *Escherichia coli (E. coli)*, and its typical size is 1 to 10 μm.

A eukaryote has a more complex internal structure (Figure 2.12B). A eukaryote contains various molecules in the cytosol similar to a prokaryote, but also contains membrane-bound organelles, such as a **nucleus**, **endoplasmic reticulum** (ER), **mitochondrion**, **Golgi apparatus**, and **lysosomes**. Each of the organelles provides localized space for specific molecules to promote specific biochemical reactions. For instance, the nucleus promotes gene expression including transcription and translation processes. The ER performs protein folding and transports proteins using vesicles, and the Golgi apparatus receives vesicles from the ER and modifies the proteins therein. The mitochondrion generates ATP to provide chemical energy to the cell. Lysosomes break down unnecessary molecules in the cell. Eukaryotes often form a multicellular organism in which individual cells are interconnected through cell–cell protein channels and junctions. Examples of eukaryotes are plant and animal cells. The typical size of eukaryotes

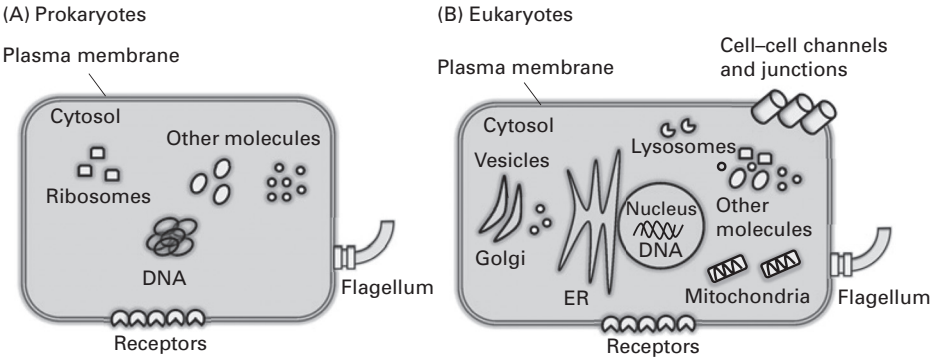


Figure 2.12 Structural organization of cells: (A) prokaryote and (B) eukaryote.

is 10 to 100 μm and they are typically larger than prokaryotes due to the increased complexity of their internal structure.

2.5 Conclusion and summary

In this chapter, we have looked at the background to cell biology with a focus on the three basic types of bio-nanomachines: protein molecules, DNA and RNA molecules, and vesicles. From a communications engineering point of view, protein molecules act as encoders or decoders of molecular signals, such as enzyme proteins that convert input signals to output signals. Protein molecules also act as molecular signals as well as transmitters (e.g., channels to diffuse molecules) and receivers (i.e., receptors). DNA stores genetic information, and operates together with proteins and RNA to decode the genetic information. Vesicles encapsulate molecular signals and function like a data packet in computer networks. When these types of bio-nanomachines are put together, a biological cell emerges with enormous complexity and functionality, where an underlying mechanism is molecular communication among bio-nanomachines – the main subject of the next chapter.

For a comprehensive coverage of cell biology, readers are referred to well-known textbooks such as [3, 4]. For communications engineers, introductory books are available that are targeted to computer scientists and engineers such as [5, 6].

References

[1] A. Regev and E. Shapiro, "Cells as computation," *Nature*, vol. 419, p. 343, 2002.
[2] D. Bray, "Protein molecules as computational elements in living cells," *Nature*, vol. 376, pp. 307–312, 2012.
[3] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Cell Biology*, 5th edn. W. H. Freeman, 2003.
[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. New York: Garland Science, 2008.
[5] W. W. Cohen, *A Computer Scientist's Guide to Cell Biology*, 1st edn. Springer, July 2007.
[6] A. T. Johnson, *Biology for Engineers*. Boca Raton, FL: CRC Press, November 2010.

3 Molecular communication in biological systems

Molecular communication occurs ubiquitously at all levels of biological systems including molecule, cell, tissue, and organ levels. In this chapter, we examine how bio-nanomachines or a system of bio-nanomachines communicate using molecules. First, we introduce two dimensions to characterize molecular communication systems: scale and mode. The scale refers to the range of distances over which bio-nanomachines communicate by propagating molecules, and it is roughly divided into intracellular, intercellular, and inter-organ levels. The mode of molecular communication systems refers to how molecules propagate between bio-nanomachines; it is either passively or actively. We then go over, from a communication engineering perspective, a number of examples of molecular communication systems found in nature. Following the previous chapter, basic biology terms are in bold in this chapter.

3.1 Scales of molecular communication

Molecular communication in a human body can be studied at three structurally different levels: intracellular, intercellular, and inter-organ levels, which are respectively, molecular communication within a cell (up to the size of a cell about 100 μm), between nearby cells (up to a population of cells, from a few μm to 10 mm or longer), and between distant cells (up to a few meters).

At the intracellular level, a number of sub-cellular bio-nanomachines within a cell communicate to sustain the life of the cell. At this level, physically separated bio-nanomachines interact directly through diffusion and collision or indirectly by propagating diffusive molecules. For example, proteins, DNA, RNA, and other molecules diffuse and interact directly through physical contact to regulate gene transcription and translation processes. Cell-surface receptors, on the other hand, interact indirectly with intracellular components by propagating diffusive molecules called **second messengers** (e.g., calcium ions); when cell-surface receptors bind to specific ligands in the extracellular space, they initiate signal transduction to generate second messengers and then the second messengers diffuse and act on target molecules in the cytosol. Cellular organelles such as ER and Golgi also indirectly interact by propagating protein-containing vesicles between them.

At the intercellular level, systems of bio-nanomachines (i.e., individual cells) communicate in the local environment to coordinate their behavior. At this level, cells

communicate in a manner similar to how bio-nanomachines communicate at the intracellular level. For instance, cells secrete diffusive molecules, the molecules propagate outside cells (i.e., extracellular pathways), and other cells react to the molecules. In many cases, molecules diffuse in a limited range in the local environment, and therefore this type of communication is called **paracrine signaling**. The molecules may also propagate cell-to-cell through internal pathways, which are physical channels established between the two neighboring cells. One common form of internal pathway is made through **gap junction channels**. Gap junction channels are formed between two apposed cells separated by a few nanometers, and the gap junction channels connect the cytosol of the two cells. Another common form of internal pathway is through tunneling nanotubes (TNTs). TNTs are made of thin membranes formed between two cells and can have a length of several cell diameters. TNTs propagate small molecules as well as vesicles and organelles between adjacent cells.

At the inter-organ level, cells in different organs (e.g., heart and liver) communicate to regulate bodily functions. At this level, cells rely on specific mechanisms to propagate molecules over a long distance. One such mechanism is the bloodstream, and this type of long-range communication is referred to as **endocrine signaling**. In endocrine signaling, cells release specific hormones into the bloodstream. The hormones then propagate and circulate throughout the body. Distant target cells detect the hormones in their extracellular fluid to regulate their cellular functions. Another mechanism for inter-organ level communication is through a nervous system, and it is referred to as **synaptic signaling**. Neurons or nerve cells are specialized cells for long-distance communication and they propagate membrane-potential differences called **action potentials** cell-to-cell among major parts of the body (e.g., the brain and the muscular system).

3.2 Modes of molecular communication

The modes of molecular communication can be categorized based on how signaling molecules propagate in the environment; namely, molecules simply diffuse or directionally propagate by consuming chemical energy. The two basic modes of molecular communication are called **passive** and **active modes** of molecular communication, respectively (Table 3.1).

The passive mode of molecular communication provides a simple method of propagating molecules within a cell and between cells. In the passive mode, molecules randomly diffuse in all available directions, making it particularly suited to environments that are highly dynamic and unpredictable. The passive mode is also suited to situations in which an infrastructure for molecular communication is not available. The passive mode, however, requires a large number of molecules to reach a distant destination. Also, due to the random movement of molecules, the time to reach a destination greatly increases. The passive mode is also not suitable for propagating large molecules in a crowded or high-viscosity environment such as the cytosolic environment.

The active mode of molecular communication provides a mechanism to directionally propagate molecules to specific locations. The active mode propagates molecules

Table 3.1. Comparison of passive and active modes of molecular communication.

Mode	Passive	Active
Propagation	Random	Directional
Probability of a molecule reaching a receiver	Low	High
Number of molecules needed to reach a receiver	Large	Small
Propagation of a large-size molecule	Difficult	Possible
Communication infrastructure	Not required	Required
Energy supply	Not required	Required

over longer distances compared with the passive mode. Macromolecules and vesicles diffuse poorly in the passive mode because of their size; but, the active mode consumes chemical energy and generates sufficient force to directionally transport large-sized molecules. In the active mode there is also a higher probability that molecules reach the destination and thus it requires fewer molecules compared with the passive mode. The active mode however often requires a communication infrastructure to be established before propagation (e.g., molecular motors require microtubules and/or filaments to move directionally). The active mode also requires a regular supply of energy to overcome the thermal noise in the environment.

How molecules propagate in the environment is also affected by many environmental factors. For instance, molecules may propagate in the environment while they react with other molecules existing in the environment. As a result, molecules may degrade and propagate only in a limited spatial range. Molecules that react in the environment may also be relayed or amplified through catalytic processes, and these molecules may propagate as a **reaction-diffusion wave** that travels at a relatively constant velocity in a broader range as in the case of synaptic signaling. Molecules may also propagate over internal pathways (e.g., gap junction channels) or external pathways (e.g., extracellular space), which affects the spatial range of propagation. Molecules may propagate in a fluid medium (e.g., bloodstream) over a long distance as in endocrine signaling.

3.3 Examples of molecular communication

Table 3.2 shows several examples of molecular communication systems found in biological systems, categorized based on the two dimensions: scale and mode. As we move to subsequent chapters, we will more precisely define a communication system. For the purpose of this chapter, let us simply describe a communication system as a system consisting of a *transmitter* (or a sender), a *receiver*, and *signal molecules* that propagate in the environment to convey a message from the transmitter to the receiver.

The following gives a quick overview of these molecular communication systems in terms of scale and mode. Additional details on these systems are provided in the rest of this chapter.

Table 3.2. Example molecular communication systems. The scale refers to the length scale or distance over which a system exists. The mode refers to either passive or active. Note that in the passive mode the diffusion coefficient D of the molecule is shown in the speed; the distance to travel during time t is proportional to $\sqrt{Dt}$ [1]. Also note that in the passive+ mode molecules react in the environment and passively diffuse as a reaction-diffusion wave; and in the passive* mode molecules passively diffuse in a fluid medium; in both cases, molecules propagate at a relatively constant velocity. Note that example measurements with respect to distance and speed vary significantly depending on the conditions.

Example	Transmitter/ Receiver	Signal molecule	Scale (Distance)	Mode (Speed)
Chemotactic signaling	Cell-surface receptor/ Flagellar motor	CheYp (Protein)	Intra cell (2 μm)	Passive (10 $\mu m^2/s$)
Vesicular trafficking	ER/Golgi	Proteins in a vesicle	Intra cell (2 μm)	Active (1 μm/s)
IP_3 signaling	Hormone receptor/ IP_3 receptor	IP_3	Intra cell (20 μm)	Passive (280 $\mu m^2/s$)
Calcium signaling	Epithelial cells	Ca^{2+} wave	Inter cell (200 μm)	Passive+ (20 μm/s)
Quorum sensing	Bacteria	AHL (Autoinducer)	Inter cell (40 μm)	Passive (1×10^{-6} cm^2/s)
Bacterial migration	F-plus bacterium/ F-minus bacterium	DNA molecule	Inter cell (50 μm)	Active (14 μm/s)
Morphogen signaling	Anchor cells/ Precursor cells	LIN-3 (EGF)	Inter-org (0.1 cm)	Passive (5×10^{-7} cm^2/s)
Hormonal signaling	Pituitary gland cells/ Thyroid gland cells	TSH (Hormone)	Inter-org (1 m)	Passive* (5 cm/s)
Neuronal signaling	Brain cells/ Heart cells	Action Potential	Inter-org (2 m)	Passive+ (100 m/s)

- Chemotactic signaling systems (Section 3.3.1) are built on transmitter and receiver proteins [2]; for instance, the cell-surface protein receptor and the flagellar motor protein communicate by propagating a diffusive intracellular protein. Some types of proteins in a bacterial species diffuse at the diffusion coefficient of 1–10 $\mu m^2/s$ in the cytosol [3]. The cytosol is about 2 μm long and proteins propagate in this length scale.
- In vesicular trafficking (Section 3.3.2), protein-containing vesicles are transported by motor proteins between two communicating organelles such as ER and Golgi in a cell [4]. Dynein motor proteins carrying a vesicle travel at the velocity of 1 μm/s with the traveling distance up to a few μm [5], which affects the length scale and speed of communication between ER and Golgi.
- In inositol 1,4,5-trisphosphate (IP_3) signaling (Section 3.3.3), cell-surface hormone receptors generate IP_3 through signal transduction and propagate the IP_3 to activate IP_3 receptors on the ER; i.e., communication between cell-surface hormone receptors

and IP_3 receptors is dependent on IP_3 diffusion in the cytosol. The diffusion coefficient of IP_3 in a cytosol is 280 $\mu m^2/s$ [6]. The diffusion occurs in the cytosol of a cell, and thus the scale length is bounded by 20 μm or so, in the case of a typical eukaryote.

- Calcium signaling (Section 3.3.3) is observed in many cell types; sending cells generate an increase in calcium concentration, the increase in calcium concentration propagates cell-to-cell as a wave, and nearby receiver cells respond to the characteristics of the wave. In epithelial cells, a calcium wave propagates in a manner similar to a reaction-diffusion wave. Its velocity is 10–30 $\mu m/s$ and the length scale can be over 10 cells (e.g., 10×20 μm) or more [7].
- In quorum sensing (Section 3.3.4), bacteria communicate by diffusing signaling molecules called autoinducers in the environment. Some types of autoinducers diffuse at the diffusion coefficient of 1×10^{-6} cm^2/s in a biofilm cluster with the radius about 40 μm [8].
- Bacterial migration and conjugation (Section 3.3.5) can be viewed as a communication system. Bacteria actively migrate by following the concentration of pheromone emitted by other bacteria. When bacteria meet they exchange DNA molecules through the process known as conjugation. Some bacteria move at 10–20 $\mu m/s$ over a distance about 50 μm in about 30 seconds [9].
- Morphogen signaling (Section 3.3.6) is observed in the developmental processes of a multicellular organism. A group of cells communicate using diffusive signaling molecules called the epidermal growth factor (EGF). Some types of EGF diffuse at 5×10^{-7} cm^2/s in the range over 500–1000 μm [10].
- In hormonal signaling (Section 3.3.7), distant organs in a body communicate by propagating hormones in the bloodstream. For instance, cells in the pituitary gland secrete hormones such as thyroid-stimulating hormone (TSH), which circulate to the target cells in the thyroid gland. A bloodstream can have a velocity around 5 cm/s [11] and circulate hormones at the length scale up to a few meters.
- Neuronal signaling (Section 3.3.8) is used by separate organs in a body to communicate. For instance, the brain and the heart signal each other through nerve cells. In nerve cells, action potentials propagate as reaction-diffusion waves at a velocity of 10–100 m/s and up to distances of over a few meters [12].

3.3.1 Chemotactic signaling

Certain biological cells move by flagellar motors to find favorable conditions containing attractant molecules (e.g., nutrients) and to avoid harmful ones containing repellent molecules (e.g., toxic molecules). This characteristic of biological cells to sense chemical conditions in the environment and to move according to the sensed chemical conditions is called **chemotaxis**. In chemotaxis, a number of cell-surface receptors act as sender bio-nanomachines or *transmitters* that activate diffusive protein molecules in the cytosol. The *receivers* of the chemical message are flagellar motor proteins that control the movement of the cell.

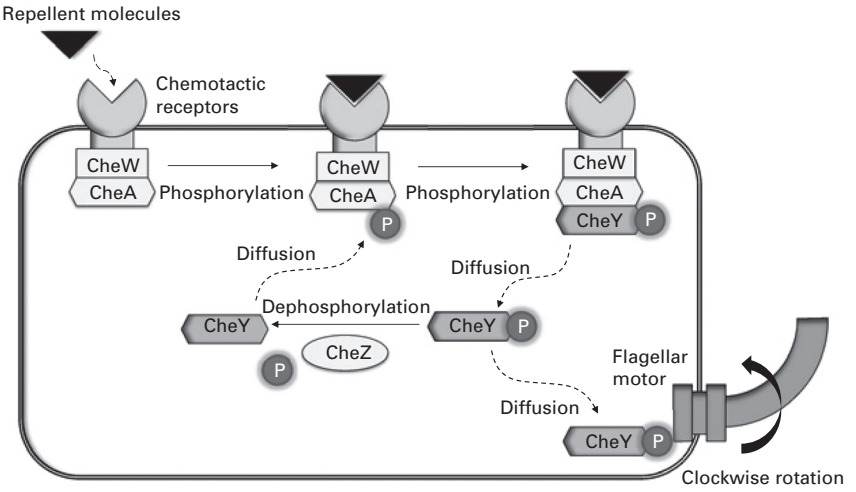


Figure 3.1 Chemotactic signaling in a bacterium.

In the case of a bacterium species *Escherichia coli* (*E. coli*), there are a number of membrane receptors such as Tar (aspartate and maltose), Trg (ribose and galactose), Tsr (serine) and Tap (dipeptides) receptors that are responsible for sensing specific molecules around the outer cell surface (Figure 3.1) [2, 13]. The binding of molecules to these receptors leads to signal transduction across the cell membrane, which eventually controls the movement of the bacterium through protein–protein interactions. For instance, the binding of repellent molecules causes an intracellular protein kinase CheA, sticking to the inner surface via the CheW protein, to become phosphorylated. The phosphate group in the phospholylated CheA (CheAp) is then passed to the diffusive CheY protein existing in the cytosol. The phosphorylated form of CheY (CheYp) then diffuses in the cytosol and binds to the flagellar motor. As a result, the flagellar motor rotates clockwise to cause the bacterium to tumble, and thereby avoids movement toward the source of the repellent molecules. However, the binding of attractant molecules to membrane receptors dephosphorylates CheA and CheY, which causes the flagellar motor to rotate counterclockwise, and thereby allows the bacterium to continue swimming toward the source of the attractant molecules. The CheZ protein in the cytosol also accelerates the dephosphorylation of CheYp to enable the bacterium to quickly respond to changes in the environment. The interactions among these proteins often result in sigmoidal responses to the concentrations or changes in concentrations of attractant and repellent molecules in the environment.

3.3.2 Vesicular trafficking

Many proteins travel between organelles (e.g., the ER, Golgi apparatus, endosome, lysosome) through the process known as **vesicular trafficking**. In vesicular trafficking, proteins are encapsulated into a vesicle or cargo at a *sending organelle*, where a

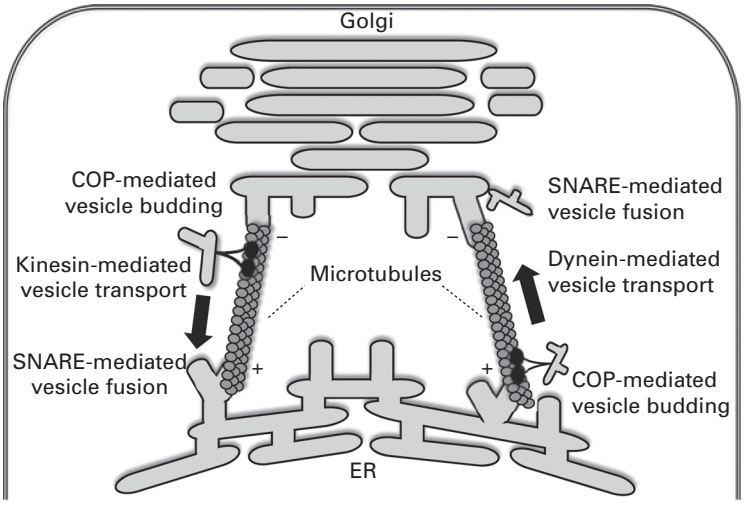


Figure 3.2 ER-to-Golgi vesicle trafficking by motor proteins.

chemical message is encoded onto the proteins. The vesicle is transported by motor proteins such as dynein and kinesin that actively move along microtubules to the *receiving organelle*. The proteins in the vesicle are then passed to the receiving organelle where they are processed (i.e., the chemical message is decoded).

The vesicular trafficking outlined above involves complex processes including the budding, transport, and fusion of vesicles (Figure 3.2) [4, 14]. In the ER-to-Golgi vesicle trafficking, for instance, protein-containing vesicles are formed through the vesicle budding process either from the ER or the Golgi apparatus mediated by specific vesicle-coating proteins called **COP** (COat Protein). The vesicles are then transported by motor proteins that move along microtubules. Since microtubules have their plus end at the ER side and the minus end at the Golgi apparatus side, the minus-end directed motor proteins, such as dynein, transport vesicles from the ER to the Golgi apparatus, and the plus-end directed motor proteins, such as kinesin, transport vesicles on the opposite pathway. Vesicles transported by motor proteins are attached to and fused with the Golgi apparatus or ER through the vesicle fusion process mediated by a specialized protein family called **SNARE** (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptors). As a result, molecules inside a vesicle are transported from one organelle to the other.

3.3.3 Calcium signaling

Calcium ions (Ca^{2+}) are the ubiquitous second messengers that regulate a large number of cellular processes, such as fertilization, differentiation, proliferation, and death, in virtually all mammalian cells [15, 16]. Within a cell, Ca^{2+} is stored in the ER and upon stimulation, it is released through calcium-release channels acting as sender bio-nanomachines or *transmitters*. A local increase of the cytosolic Ca^{2+} propagates as a

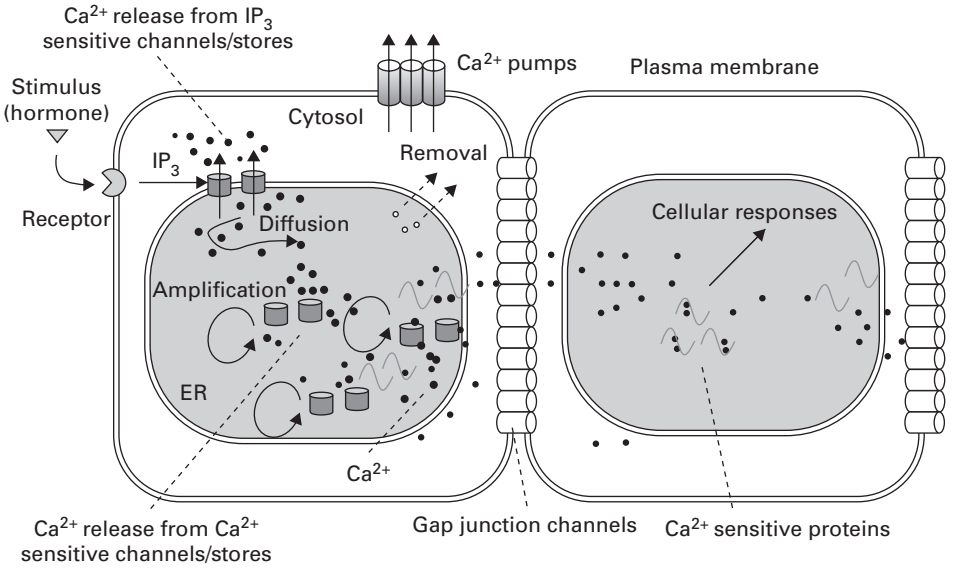


Figure 3.3 Calcium signaling.

reaction-diffusion wave in a cell and between cells. Here a chemical message is encoded onto the temporal and spatial dynamics of the Ca^{2+} concentration, such as the amplitude and frequency of calcium concentration, known as the amplitude modulation (AM) and frequency modulation (FM) of calcium signaling [17]. The temporal and spatial nature of Ca^{2+} signaling is in turn decoded by calcium-sensitive proteins such as calmodulin acting as *receivers* that further propagate chemical messages to downstream signaling pathways.

Calcium signaling is observed both within and between cells (Figure 3.3) [20]. When external stimuli (e.g., hormonal signals) are bound to cell-surface receptors, an enzyme called phospholipase C (PLC) is activated to produce calcium mobilizing molecules, inositol 1,4,5-trisphosphate (IP_3), through the hydrolysis of phosphatidylinositol 4,5-bisphosphate (PIP_2) present in the cytosol (i.e., IP_3 signaling) [18]. IP_3 then diffuses in the cytosol and causes IP_3 receptors on the ER to release Ca^{2+} from the ER. The released Ca^{2+} then freely diffuses in the cytosol. Ca^{2+} is also amplified via Ca^{2+} sensitive Ca^{2+} release channels through the process known as **calcium-induced calcium release** (CICR). In CICR, Ca^{2+} binds to Ca^{2+} release channels on the ER, which then open to release Ca^{2+} from the ER. The released Ca^{2+} further activates the nearby Ca^{2+} release channels. This positive feedback process allows a local increase of the Ca^{2+} concentration to propagate as **intracellular Ca^{2+} waves**, quickly leading to a global increase in the cytosolic Ca^{2+} concentration. The cytosolic Ca^{2+} concentration of a glial cell, 30–150 nM at the resting state, for instance, is increased to several μM within milliseconds upon excitation [19]. The increased Ca^{2+} concentration is then decreased by, for instance, plasma membrane Ca^{2+} pumps (i.e., Ca^{2+}-ATPase) that extrude cytosolic Ca^{2+} to the extracellular space, and cellular organelles that sequester Ca^{2+}. The

increase and decrease of the Ca^{2+} concentration is termed **Ca^{2+} spikes**, and they can repeatedly occur for minutes to hours under physiological conditions.

A Ca^{2+} increase generated in a cell can be propagated cell-to-cell as **intercellular Ca^{2+} waves**. There are two types of signaling pathway by which Ca^{2+} waves propagate cell-to-cell: internal and external pathways. In the case of internal pathways, Ca^{2+} and IP_3 diffuse through gap junction channels from cell to cell to propagate a Ca^{2+} increase. In the case of external pathways, a cell releases ATP outside the cell and nearby cells respond to ATP by increasing their Ca^{2+} concentration, thereby a Ca^{2+} increase propagates cell-to-cell.

3.3.4 Quorum sensing

Bacteria are unicellular organisms but they communicate and perform group behavior much like a group of cells in a multicellular organism [21]. One form of bacterial communication is called **quorum sensing**, which allows bacteria to monitor the cell-population density and coordinate their behavior based on the density. For example, quorum sensing allows a group of bacteria to synchronize their gene expression for biofilm formation and luminescence generation. In quorum sensing, each participating bacterium acts as a *sender* as well as a *receiver* of molecular communication. Each bacterium synthesizes and propagates specific molecules called **autoinducers** in the environment. The concentration of autoinducers in the environment correlates with or encodes the population density of bacteria. Each bacterium behaves in a concentration-dependent manner, and therefore the behavior of nearby bacteria can be coordinated.

Two different quorum-sensing systems are found in two different types of bacteria: gram-negative and gram-positive bacteria [22, 23]. Gram-negative bacteria use an autoinducer, acyl-homoserine lactone (AHL), that is permeable to their cell membranes; and they implement a simpler quorum sensing system using the LuxI/LuxR regulatory proteins (Figure 3.4A). In gram-negative bacteria, the LuxI protein catalyzes the formation of AHL in a cell. The AHL is permeable to the cell membrane and thus diffuses from the cell to the environment. The AHL in the environment also diffuses into a cell and binds to the LuxR protein in the cell to form a LuxR-AHL complex. When the concentration of AHL in the environment increases, the LuxR-AHL complex increases in activity level, leading to the expression of target genes in the cell.

Gram-positive bacteria have a different membrane structure and diffuse a different type of autoinducer, collectively referred to as autoinducing polypeptides (AIPs), through the membrane-bound exporter proteins (Figure 3.4B). In gram-positive bacteria, AIPs are synthesized through gene expression in a cell and transported by the membrane-bound exporter proteins across the cell membrane to the environment. The AIPs in the environment are detected by cell-surface receptors of the cells. When cell-surface receptors bind to the AIPs in the environment, they auto-phosphorylate and subsequently transfer the phosphoryl group to intracellular regulatory proteins. The regulatory proteins in the cell are then phosphorylated and activate or repress the expression of target genes in the cell.

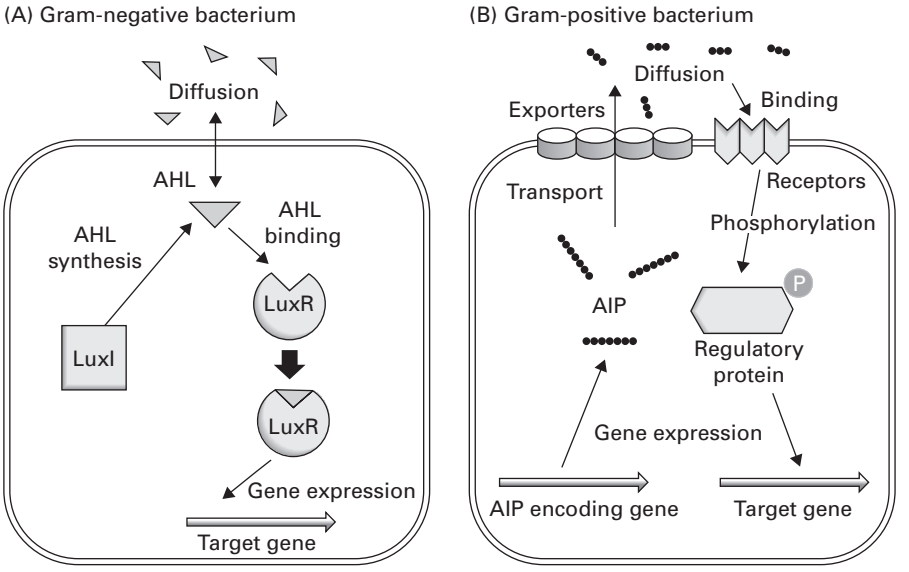


Figure 3.4 Quorum sensing in (A) gram-negative and (B) gram-positive bacteria.

3.3.5 Bacterial migration and conjugation

Bacteria actively move by performing chemotaxis within a relatively large-scale environment (e.g., an intestinal environment in a human body) and also exchange DNA molecules through the mechanism called **conjugation**. They are thus considered active carriers of DNA molecules. In bacterial conjugation, a *sending donor cell* transfers a plasmid (i.e., a chemical message encoded as a DNA sequence) to the *receiving cell*, and the receiving cell performs gene expression to decode the message – acquire a new functionality such as antibiotic resistance. Bacterial conjugation, in addition to quorum sensing, is a potential mechanism for inter-kingdom communication, through which bacteria communicate with their host cells to maintain a symbiotic environment [24, 25].

Bacterial conjugation occurs between two types of bacteria: a donor cell with the F-plasmid (i.e., a plasmid to transfer) and its recipient cell without the F-plasmid, called F-plus and F-minus bacteria respectively. To initiate bacterial conjugation, an F-minus bacterium may release pheromones (e.g., certain types of peptides) in the environment to attract an F-plus bacterium (Figure 3.5A). The F-plus bacterium, in response to the pheromones in the environment, projects a pilus from its cell surface (Figure 3.5B) and forms a channel with the F-minus bacterium. The plasmid in the F-plus bacterium is processed to produce a single-stranded DNA molecule, which is transferred to the F-minus bacterium (Figure 3.5C). The F-plus and F-minus bacteria synthesize the complementary strand of the respective single-stranded DNA molecule to reproduce the original plasmid. As a result, the F-minus bacterium is transformed into an F-plus bacterium (Figure 3.5D) to function as a donor cell.

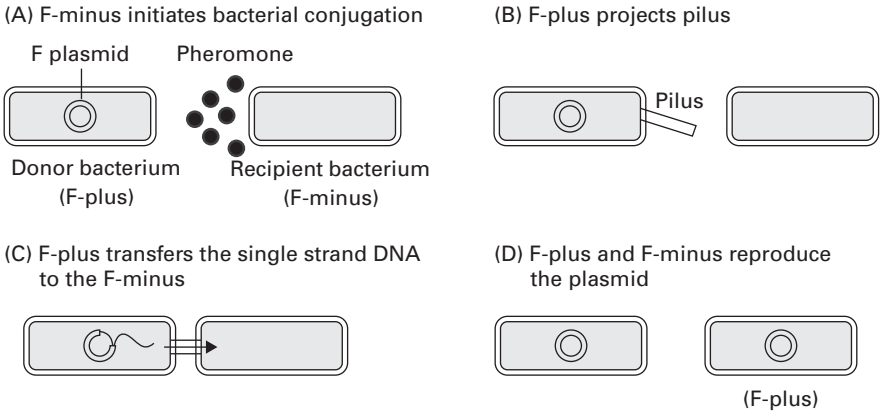


Figure 3.5 Bacteria move and exchange DNA molecules.

3.3.6 Morphogen signaling

In multicellular organisms, a group of cells coordinate their behavior through molecular communication. In developing a tissue structure, for example, a group of *sending cells* secretes diffusive molecules called **morphogens** to establish gradients of the molecules in the environment [26]. The morphogens may be encapsulated in vesicles, which propagate in the environment. The morphogens may also propagate cell-to-cell through internal pathways (e.g., gap junctional pathways). The morphogens that propagate, degrade or chemically react over time to facilitate the formation of gradients. The morphogens may also propagate as a traveling wave in a reaction-diffusion manner. The types and concentrations of morphogens also regulate the behavior of *receiving cells* in tissue development, such as proliferation, differentiation, migration, and death of the cells.

One example of morphogen signaling is found in organ development of *Caenorhabditis elegans* (*C. elegans*) (a model organism used in biology), during which cells form a series of gradients and differentiate to establish cell patterns [27,28]. In its vulval development, anchor cells release a type of morphogen, the epidermal growth factor (EGF) related signal called LIN-3, to form a gradient in the extracellular space (Figure 3.6A). The nearby vulval precursor cells have LIN-3 specific receptors and respond to LIN-3 in a concentration-dependent manner. Namely, the precursor cells close to the anchor cells receive the large amount of LIN-3 and form cell type 1, and the distant precursor cells receive the small amount to form cell type 3 (Figure 3.6B). Subsequently, type 1 cells secrete a different type of morphogen, LIN-12, in the extracellular space. The LIN-12 morphogen has a different effect and only affects type 3 cells and not type 1 cells. Nearby type 3 cells respond again in a concentration-dependent manner; type 3 cells receiving the large amount of LIN-12 form cell type 2 (Figure 3.6C). The resulting structure is a certain pattern of cells that undergo further development to produce the final organ of the *C. elegans*.

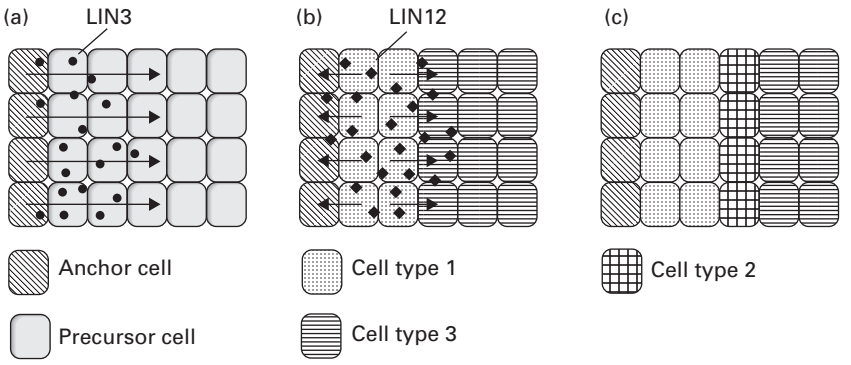


Figure 3.6 Morphogen signaling in organ development.

3.3.7 Hormonal signaling

Hormones play an important role in regulating bodily functions. In hormonal signaling, *sender cells* in one organ produce hormones and release them into the bloodstream. The hormones circulate via the bloodstream throughout the body and act on *receiver cells* in another organ, often located at a relatively longer distance (e.g., up to a few meters) from the sender cells. The receiver cells in turn release another type of hormone that acts as feedback signals to the sender cells. The feedback-based control is essential to control bodily functions, for instance, to maintain the blood glucose level.

One example of this type of molecular communication is found in the regulation of thyroxine in the human body (Figure 3.7). The overall process of thyroxine regulation represents a homeostatic system in which multiple cells communicate with each other through the concentration of a certain chemical substance in the body. Thyroxine is a hormone that invokes the process to control the basal metabolic rate of cells. The process of regulating thyroxine concentration involves two glands: the pituitary and the thyroid. Cells in the pituitary gland called thyrotropes secrete thyroid-stimulating hormone (TSH) [29]. The secreted TSH propagates through the blood circulation system to cells throughout the body. Upon receiving TSH, the epithelial cells of the thyroid gland produce thyroxine. When the thyroxine concentration becomes too high, the pituitary gland decreases the secreted amount of TSH, and this, in turn, causes epithelial cells of the thyroid gland to produce less thyroxine. Through this interactive process between the pituitary gland and thyroid gland, the thyroxine concentration is stabilized in the human body.

3.3.8 Neuronal signaling

Neurons communicate with each other to form neural networks [30]. In neuronal signaling, a pre-synaptic *sending neuron* generates an **action potential** at one end of its body and the action potential propagates through the body to the other end, where the action potential is converted to chemical signals, called **neurotransmitters**. The neurotransmitters diffuse in the **synaptic cleft**, which is a gap between the pre-synaptic

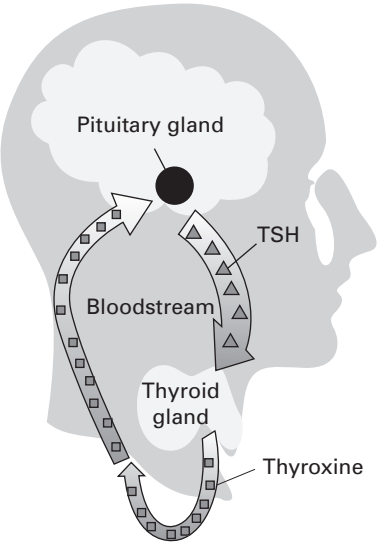


Figure 3.7 Hormonal signaling between the pituitary gland cells and thyroid gland cells.

sending neuron and the post-synaptic *receiving neuron* (Figure 3.8A). The receiving neuron may recursively generate an action potential depending on the concentration and types of neurotransmitters. An action potential propagates in a reaction-diffusion manner over long distances up to a few meters in a mammalian body.

A key molecular component involved in neuronal signaling is ion channels. Each neuron has a number of pore-forming ion channels on its membrane to regulate the diffusion of ions across the membrane. These ion channels are described as being voltage-gated since they open in response to changes in the electrical potential difference over the membrane. Ions such as Na^+, K^+, Cl^-, and Ca^{2+} are found at different concentrations in the intracellular and extracellular space. For instance, the Na^+ concentration is 12 mM inside a neuron and 145 mM outside; the K^+ concentration is 155 mM inside and 4 mM outside, both at the equilibrium or at the *resting state*. A neuron at the resting state is negatively charged with a net voltage difference over the membrane at around -70 mV.

A stimulation to one end of a neuron may cause a change in electrical potential difference. In response to this, voltage-gated Na^+ channels on the membrane open to allow a small amount of Na^+ to diffuse into the neuron, following the concentration gradient (Figure 3.8B). The membrane at the site of stimulation is then slightly depolarized, which causes more Na^+ channels to open and to recruit more Na^+ to diffuse into the neuron. The membrane at the site of stimulation is thus immediately depolarized to the positive state at around $+50$ mV. The local depolarization at the site of stimulation then leads to the depolarization of neighboring regions of the membrane. At the same time, these Na^+ channels, at the site of stimulation, close and remain inactivate for a refractory period of time; and the voltage-gated K^+ channels open to allow K^+ to diffuse out, which contributes to the polarization of the membrane at the site of simulation.

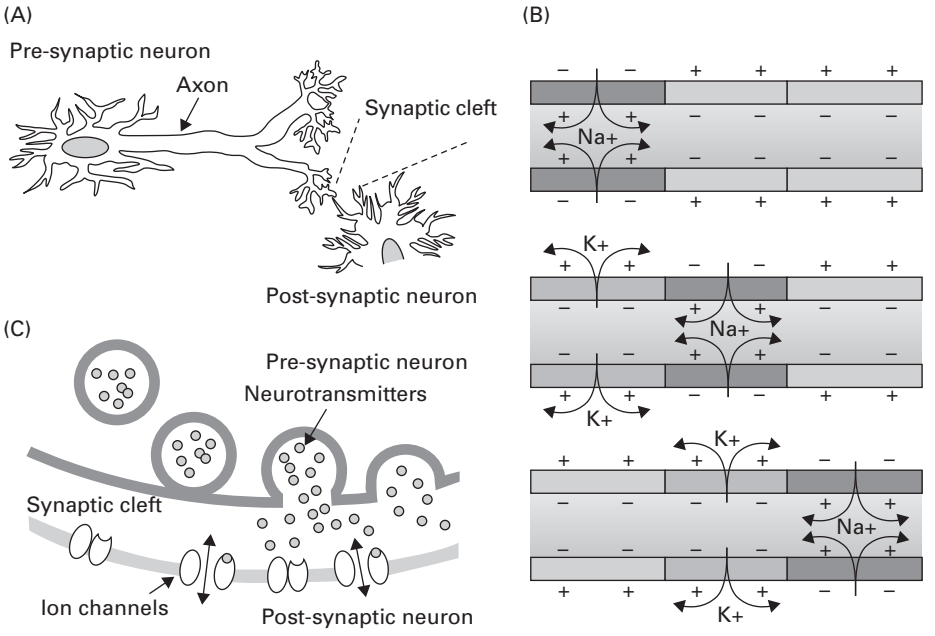


Figure 3.8 Neuronal signaling. (A) The structure of neurons. (B) Propagation of an action potential in a neuron. (C) Chemical signaling at a synaptic cleft.

A propagating action potential in a presynaptic neuron causes membrane-bounded synaptic vesicles inside the neuron to release neurotransmitters such as glutamate into the synaptic cleft (Figure 3.8C). The neurotransmitters then diffuse and bind to specific ion channels on the postsynaptic neuron. These ion channels, in response to the neurotransmitters, open and allow ions to diffuse across the membrane of the postsynaptic neuron. An electrical potential difference is then generated over the membrane in the postsynaptic neuron to initiate signaling. The neurotransmitters at the synaptic cleft are removed by specific enzymes, or by uptake by neurons or the surrounding cells.

3.4 Conclusion and summary

In this chapter, we reviewed a number of examples of molecular communication systems that are found at different scales and modes: bacterial chemotactic signaling, vesicular trafficking, IP_3 and calcium signaling, quorum sensing, bacterial migration and conjugation, morphogen signaling, hormonal signaling, and neuronal signaling. We described these systems as communication systems consisting of transmitters, receivers, and signaling molecules. Based on this initial view of molecular communication systems, we move onto subsequent chapters to develop communication system models in more detail.

References

[1] H. C. Berg, *Random Walks in Biology*. Princeton University Press, 1993.

[2] J. S. Parkinson and E. C. Kofoid, "Communication modules in bacterial signaling proteins," *Annual Review of Genetics*, vol. 26, pp. 71–112, 1992.

[3] M. B. Elowitz, M. G. Surette, P.-E. Wolf, J. B. Stock, and S. Leibler, "Protein mobility in the cytoplasm of *Escherichia coli*," *Journal of Bacteriology*, vol. 181, no. 1, pp. 197–203, 1999.

[4] A. Murshid and J. F. Presley, "ER-to-Golgi transport and cytoskeletal interactions in animal cells," *Cellular and Molecular Life Sciences*, vol. 61, pp. 133–145, 2004.

[5] T. A. Schroer and M. P. Sheetz, "Two activators of microtubule-based vesicle transport," *Journal of Cell Biology*, vol. 115, no. 5, pp. 1309–1318, 1991.

[6] N. L. Allbritton, T. Meyer, and L. Stryer, "Range of messenger action of calcium ion and inositol 1,4,5-trisphosphate," *Science*, vol. 258, no. 5089, pp. 1812–1815, 1992.

[7] M. J. Sanderson, "Intercellular waves of communication," *Physiology*, vol. 11, no. 6, pp. 262–269, 1996.

[8] P. S. Stewart, "Diffusion in biofilms," *Journal of Bacteriology*, vol. 185, no. 5, pp. 1485–1491, 2003.

[9] H. C. Berg and D. A. Brown, "Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking," *Nature*, vol. 239, pp. 500–504, 1972.

[10] R. G. Thorne, S. Hrabetova, and C. Nicholson, "Diffusion of epidermal growth factor in rat brain extracellular space measured by integrative optical imaging," *Journal of Neurophysiology*, vol. 92, pp. 3471–3481, 2004.

[11] F. Bogazzi, L. Bartalena, S. Brogioni, A. Burelli, L. Manetti, and M. L. Tanda, "Thyroid vascularity and blood flow are not dependent on serum thyroid hormone levels: studies in vivo by color flow doppler sonography," *European Journal of Endocrinology*, vol. 140, pp. 452–456, 1999.

[12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. New York: Garland Science, 2008.

[13] A. Bren and M. Eisenbach, "How signals are heard during bacterial chemotaxis: protein-protein interactions," *Journal of Bacteriology*, vol. 182, no. 24, pp. 6865–6871, 2000.

[14] N. Hirokawa, Y. Noda, Y. Tanaka, and S. Niwa, "Kinesin superfamily motor proteins and intracellular transport," *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 682–696, 2009.

[15] M. J. Berridge, P. Lipp, and M. D. Bootman, "The versatility and universality of calcium signaling," *Nature Reviews Molecular Cell Biology*, vol. 1, pp. 11–21, 2000.

[16] M. J. Berridge, D. Bootman, and H. L. Roderick, "Calcium signaling: dynamics, homeostasis and remodelling," *Nature Reviews Molecular Cell Biology*, vol. 4, pp. 517–529, 2003.

[17] M. J. Berridge, "The AM and FM of calcium signalling," *Nature*, vol. 386, no. 6627, pp. 759–760, 1977.

[18] K. Hirose, S. Kadowaki, M. Tanabe, H. Takeshima, and M. Iino, "Spatiotemporal dynamics of inositol 1,4,5-trisphosphate that underlies complex Ca^{2+} mobilization patterns," *Science*, vol. 284, pp. 1527–1530, 1999.

[19] J. W. Deitmer, A. Verkhratsky, and C. Lohr, "Calcium signaling in glial cells," *Cell Calcium*, vol. 24, pp. 405–416, 1998.

[20] T. Nakano and J. Q. Liu, "Design and analysis of molecular relay channels: an information theoretic approach," *IEEE Transactions on NanoBioscience*, vol. 9, no. 3, pp. 213–221, 2010.
[21] E. P. Greenberg, "Tiny teamwork," *Nature*, vol. 424, p. 134, 2003.
[22] B. L. Bassler, "How bacteria talk to each other: regulation of gene expression by quorum sensing," *Current Opinion in Microbiology*, vol. 2, pp. 582–587, 1999.
[23] M. B. Miller and B. L. Bassler, "Quorum sensing in bacteria," *Annual Review of Microbiology*, vol. 55, pp. 165–199, 2001.
[24] J. A. Heinemann and G. F. Sprague Jr, "Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast," *Nature*, vol. 340, pp. 205–209, 1989.
[25] D. T. Hughes and V. Sperandio, "Inter-kingdom signalling: communication between bacteria and their hosts," *Nature Review Microbiology*, vol. 6, pp. 111–120, 2008.
[26] M. Ibanes and J. C. I. Belmonte, "Theoretical and experimental approaches to understand morphogen gradients," *Molecular Systems Biology*, vol. 4, no. 176, 2008.
[27] J. S. Simske and S. K. Kim, "Sequential signalling during *Caenorhabditis elegans* vulval induction," *Nature*, vol. 375, no. 6527, pp. 142–146, 1995.
[28] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Cell Biology*, 5th edn. W. H. Freeman, 2003.
[29] C. Villalobos, L. Nunez, and J. Garcia-Sancho, "Anterior pituitary thyrotropes are multifunctional cells," *American Journal of Physiology Endocrinology and Metabolism*, vol. 287, pp. E1166–E1170, 2004.
[30] S. B. Laughlin and T. J. Sejnowski, "Communication in neuronal networks," *Science*, vol. 301, pp. 1870–1874, 2003.

4 Molecular communication paradigm

We learned in Chapter 3 that molecular communication provides a ubiquitous method by which natural bio-nanomachines communicate. In the simplest model, a group of bio-nanomachines acting as *sender bio-nanomachines* transmit information carrying signal molecules, called *information molecules* in this chapter, the information molecules propagate in the environment, and a group of bio-nanomachines acting as *receiver bio-nanomachines* chemically react to the propagating molecules. In this chapter, we first describe a simple model of molecular communication with the goal of identifying key components and processes necessary for design and analysis of molecular communication systems. We then discuss from the communication engineering point of view the need for a network architecture that extends the simple model and allows system designers to integrate a group of bio-nanomachines into a functional and robust molecular communication system.

4.1 Molecular communication model

A starting requirement in molecular communication research is to generalize communication processes and develop a basic model of molecular communication. A variety of designs and mechanisms of molecular communication appear in biological systems (e.g., natural biological cells communicate through transmission of diffusive molecules, protein-nanomachines transport materials by propagating themselves over protein filaments). Abstraction of the naturally occurring molecular communication systems may help identify key components and processes and provide a basis for design and analysis of a wide variety of molecular communication systems.

The basic model of molecular communication may be described based on Shannon's model of communication (Figure 4.1) [1, 2, 3, 4, 5]. It consists of components functioning as information molecules that represent information (or a message) to be transmitted, sender bio-nanomachines that release the information molecules, receiver bio-nanomachines that detect information molecules, and the environment in which the information molecules propagate from the sender bio-nanomachine to the receiver bio-nanomachine. It may also include other types of specialized bio-nanomachines that function as transport molecules to move information molecules, guide molecules to direct the movement of transport molecules, interface molecules that allow a transport molecule to selectively transport information molecules, and addressing molecules (not

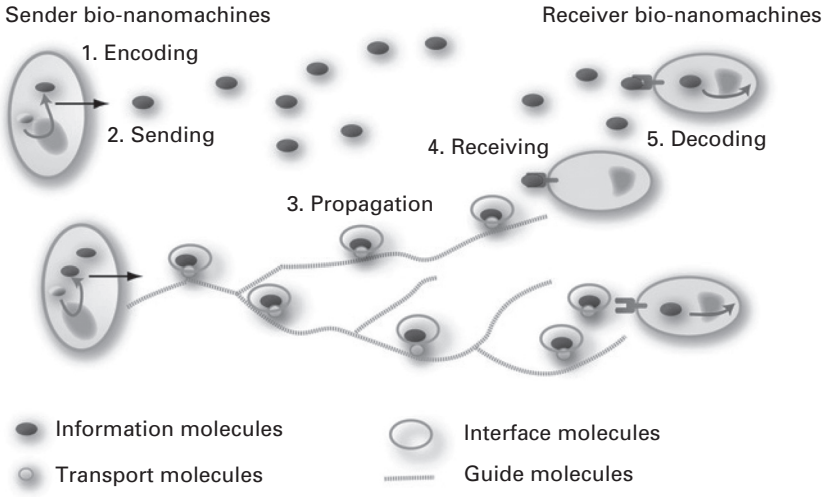


Figure 4.1 A model of molecular communication [5].

shown) that are attached to information molecules or interface molecules to specify the receiver bio-nanomachine.

The general processes of communication include (1) encoding of information into an information molecule by the sender bio-nanomachine, (2) sending of the information molecule into the environment, (3) propagation of the information molecule through the environment, (4) receiving of the information molecule by the receiver bio-nanomachine, and (5) decoding of the information molecule into a chemical reaction at the receiver bio-nanomachine:

1. *Encoding* is the process in which a sender bio-nanomachine translates information (or a message) into information molecules that the receiver bio-nanomachine can detect. Information may be encoded in various forms within the information molecules, such as in the three-dimensional structure of the information molecule (e.g., a specific type of molecule), in the specific molecules that compose the information molecules (e.g., DNA is formed by the specific sequence of nucleotides), or in the concentration of information molecules (i.e., the number of information molecules per unit volume of solvent molecules) modulated over time (e.g., a neuron can produce spikes of neurotransmitter at a particular frequency).
2. *Sending* is the process by which a sender bio-nanomachine releases information molecules into the environment. A sender bio-nanomachine may release information molecules by either unbinding information molecules from the sender bio-nanomachine (e.g., by budding vesicles from a sender bio-nanomachine if a sender bio-nanomachine is a biological cell), or by opening a molecular gate through which the information molecules diffuse away (e.g., by opening an ion channel on the membrane of a sender bio-nanomachine). A sender bio-nanomachine may also catalyze a chemical reaction that produces information molecules elsewhere.

3. *Propagation* is the process during which information molecules move from a sender bio-nanomachine through the environment to the receiver bio-nanomachine. An information molecule may passively diffuse through the environment without using chemical energy, or actively propagate, for example, by using a transport molecule (e.g., a motor protein) that consumes ATP energy. During propagation, an interface molecule may also be necessary to protect information molecules from noise in the environment. For instance, an information molecule may be contained in a vesicle-based interface molecule and propagate through the environment. The vesicle prevents the information molecule from chemically reacting with other molecules outside the vesicle.
4. *Receiving* is the process by which the receiver bio-nanomachine captures information molecules propagating in the environment. One option for a receiver bio-nanomachine to capture information molecules is to have a surface structure permeable to the information molecules. For instance, the plasma membrane of a biological cell is permeable to some types of signal molecules, and therefore the signal molecules propagating in the environment can enter the cell and directly bind to the receptors within the cell. Another option is to use surface receptors that are capable of binding with a specific type of information molecule and inducing reactions at the surface, which causes reactions within the receiver bio-nanomachine. Yet another option is to use surface channels (e.g., chemically gated channels) that allow information molecules in the environment to flow into a receiver bio-nanomachine.
5. *Decoding* is the process during which the receiver bio-nanomachine, upon capturing information molecules, reacts to the molecules. Chemical reactions for decoding may result in the production of new molecules (e.g., through an enzymatic reaction), the generation of motion, morphological changes of the receiver bio-nanomachine, and modification of chemical functionality. If a new molecule is produced and released into the environment, the receiver bio-nanomachine in turn acts as a sender bio-nanomachine, leading to multi-hop molecular communication.

4.2 General characteristics

Molecular communication utilizes biological materials and processes for communication. In the basic model (Figure 4.1), a group of bio-nanomachines communicates with another group of bio-nanomachines by propagating molecules in an aqueous environment. A more complex form of molecular communication may be possible with distributed bio-nanomachines that are interconnected via multi-hop molecular communication. In any form, as a result of using biological materials and processes, molecular communication is expected to exhibit unique features that make it distinct from the current telecommunication technology (Table 4.1).

4.2.1 Transmission of information molecules

In molecular communication, a group of bio-nanomachines communicate through transmission of information molecules. The size and structure of information molecules

Table 4.1. Telecommunication and molecular communication

Communication	Telecommunication	Molecular communication
Devices	Electronic devices	Bio-nanomachines
Signal types	Optical/electrical	Chemical
Types of information	Digital	Chemical and physical
Propagation speed	Speed of light (3×10^8 m/s)	Extremely slow (Table 3.2)
Propagation range	m – km	nm – μm (Table 3.2)
Media	Air/cables	Aqueous
Energy consumed	Electrical/high	Chemical/low
Other features	Accurate	Stochastic, massive parallelism, bio-compatible

affect how the information molecules propagate in the environment. Information molecules may need to be chemically stable and robust against environmental noise and interference with other molecules.

A small molecule with low molecular weight (e.g., $<$ 800 Da) may be used as an information molecule. Small molecules can freely diffuse in the environment and become suitable carriers for molecular communication. Small molecules can also exhibit specificity and act on receiver components such as receptors with high affinity. In addition, small molecules may be readily available and easily reused by bio-nanomachines. Small molecules can be either selected from naturally occurring molecules, or artificially synthesized. Examples of small molecules from biological systems are second messengers (e.g., Ca^{2+}, IP_3, DAG, and cyclic AMP) that relay external signals from cell-surface receptors to the target molecules inside a cell. Examples of synthetic molecules are developed in pharmacology and medicine where molecules not exceeding 800 Da are considered small and effective drugs; they can diffuse across cell membranes, bind to target receptors inside a cell, and alter the activities of the cell (e.g., inhibit cancer growth).

Macromolecules (i.e., large molecules) can also be used as information molecules. By definition in biochemistry, a macromolecule is a large polymer assembled from a number of basic units called monomers [6]. Macromolecules (e.g., a linear chain of repetitive units) fold into energetically stable structures with specific functionality depending on the number of units. Macromolecules can also consist of multiple types of units, and the order and types of units determine the structures and functions. Macromolecules also degrade easily depending on the chemical environment. As such, macromolecules exhibit a high degree of complexity in structure and function, and the complexity can be exploited to encode various information with high density. Macromolecules commonly considered in biochemistry are polysaccharides, nucleic acids, and proteins. Macromolecules can also be artificially synthesized. One example of synthetic macromolecules is carbon nanotubes that can be modified to cross cell membranes for drug delivery [7].

4.2.2 Information representation

Molecular communication propagates information molecules from a group of sender bio-nanomachines to a group of receiver bio-nanomachines, and the receiver bio-nanomachines chemically react with the propagating molecules. Information in molecular communication is represented with the physical or chemical properties of molecules, such as the type of molecules used, their three-dimensional structure (e.g., protein structure), sequence structure (e.g., DNA sequence), or concentration (e.g., calcium concentration) that the receiver bio-nanomachines are able to react to. A high density of information may be encoded in a molecular structure. For instance, a DNA sequence may store 2 bits of information within a 0.34 nm length, assuming that a DNA base is one of the four DNA bases (A, T, G, or C) and approximately 0.34 nm in length (see Figure 2.7 in Chapter 2). In addition, functional information may be encoded. For example, a DNA sequence that encodes specific proteins may be transmitted to a receiver bio-nanomachine which in turn acquires new functionality (e.g., resistance to toxic molecules) as a result of gene expression.

4.2.3 Slow speed and limited range

Molecular communication is extremely slow and restricted to a limited range when compared to current telecommunication systems. Biological systems use various methods of molecular communication, each of which has been tuned over the course of evolution to different communication needs and situations (Section 3.3). Free diffusion of molecules provides the slowest method of molecular communication. The time t required to propagate a molecule over a distance L is given as $t \approx \frac{L^2}{D}$ (i.e., the time increases with the square of the distance), where D is the diffusion coefficient of the molecule [8]. Using the typical diffusion coefficient of a protein molecule in water ($D = 100\ \mu m^2/s$), the time required to propagate over a 1 meter length ($L = 10^6\ \mu m$) is found to be approximately 10^{10} sec (about 300 years). The diffusion coefficient is dependent on the size and structure of the molecule, temperature, and the viscosity of the medium. A DNA molecule in a yeast cell propagates with the diffusion coefficient of $5 \times 10^{-4}\ \mu m^2/s$, and an intercellular signaling molecule (e.g., IP_3) with $280\ \mu m^2/s$. The propagation of these molecules in natural biological systems is thus often restricted to a sub-domain within a cell (e.g., of 20 μm in length) or a couple of cells. An alternative method of molecular communication is active transport by motor proteins. Kinesin, for instance, can travel directionally at 3–4 μm/s, enabling it to achieve distances of 50–400 mm per day. Bacterial chemotaxis also provides the active mechanism of molecular communication. *E. coli*, for instance, have several flagella that rotate to produce a biased motion at several μm/s. Another alternative method of molecular communication is through a diffusion-reaction mechanism. Calcium waves propagate at a velocity of 20 μm/s over a sheet of cells (e.g., hundreds of cells); and neurons propagate electrochemical signals (i.e., action potentials) at 100 m/s over several meters. Note that the speed and range shown above are examples and they vary depending on the conditions.

4.2.4 Stochastic communication

Molecular communication is characterized as being stochastic. The stochastic nature arises from factors such as the unpredictable movement of molecules in the environment, bio-nanomachines probabilistically reacting to information molecules, and information molecules degrading over time. The stochastic nature inherently affects the design of molecular communication systems. In order to be robust to the environmental noise, for instance, sender bio-nanomachines may release a number of molecules to increase the signal-to-noise ratio (e.g., the number of information molecules over the number of molecules in the environment that induce unintended reactions at receiver bio-nanomachines), and receiver bio-nanomachines chemically react to information molecules only when a large number of information molecules are present in the environment. Sending a large number of information molecules is highly redundant, so the degradation of some information molecules may not impact communication. In addition, molecules present in the environment may not trigger a chemical reaction at a receiver nanomachine as long as the noise from those molecules is significantly less than the signal from a large number of information molecules. Alternatively, the molecules of the same type already present in the environment, normally considered as noise, may be exploited to enhance the ability of a receiver bio-nanomachine to detect weak signals (e.g., a small number of information molecules). For instance, information molecules from prior transmissions of a sender bio-nanomachine randomly move about the environment and add random noise to the signal strength of the information molecule detected by receiver bio-nanomachines. The signal strength of information molecules transmitted from sender bio-nanomachines is probabilistically enhanced by the additive random noise, and therefore a receiver bio-nanomachine has a higher probability of detecting a weak signal.

4.2.5 Massive parallelization

Molecular communication between a group of sender bio-nanomachines and a group of receiver bio-nanomachines may represent a modern wireless communication system, called Multiple Input and Multiple Output (MIMO). MIMO uses multiple radio channels between multiple antennas at a transmitter and the receiver to improve communication performance. In molecular communication, a group of sender bio-nanomachines communicates information with a group of receiver nanomachines through multiple channels, representing a MIMO system with a high level of parallelism. For instance, a bacterium has a number of cell-surface receptors (i.e., sender bio-nanomachines) that collectively sense external stimuli, propagate molecular signals in the cytosolic environment, and control the flagellar motors (i.e., receiver bio-nanomachines) to produce a directed motion (Section 3.3.1). These sender bio-nanomachines process information massively in parallel; they release information molecules (i.e., molecular signals) through various signaling pathways independently from other sender bio-nanomachines. The receiver bio-nanomachines are also likely react to the incoming information molecules independently from other bio-nanomachines, demonstrating a

high level of parallelism. Also, a biological cell has a number of calcium channels and clusters distributed in the cytosolic environment (Section 3.3.3). These calcium channels and clusters can also be viewed as sender and receiver bio-nanomachines, forming a MIMO system. These calcium channels and clusters propagate Ca^{2+} through a spatially and temporarily arranged calcium signaling circuit to process information in parallel.

4.2.6 Energy efficiency

Molecular communication may achieve a high degree of energy efficiency by reusing materials and processes from biological systems. Motor proteins, one of the natural bio-nanomachines, convert chemical energy to mechanical work with high energy efficiency [9]. For instance, myosin V may generate a force of 2 pN to make a 36 nm step by using an energy generated from the ATP hydrolysis, 80 pN·nm. The mechanical work that myosin V performs is thus estimated as 2 pN $\times$ 36 nm $=$ 72 pN·nm, resulting in $72/80 = 90\%$ energy efficiency. It is also possible that the chemical energy necessary for molecular communication is supplied by the environment. For instance, bio-nanomachines implanted in a human body may harvest energy (e.g., glucose) from the environment, and as such require no external energy sources. Energy efficiency is an advantageous feature of molecular communication to reduce the cost of using bio-nanomachines and molecules for communication.

4.2.7 Biocompatibility

Molecular communication may also achieve a high degree of biocompatibility by reusing materials and processes from biological systems. For instance, a bio-nanomachine engineered from a biological cell (e.g., a red blood cell) may be safely injected into a human body, which can avoid unintended reactions with existing components in the human body (e.g., immune cells). Bio-nanomachines may then use encoding and decoding methods available to biological cells in the human body to directly communicate with them. The biocompatibility of molecular communication becomes a key feature to explore different applications, as we will see in the rest of this chapter.

4.3 Molecular communication network architecture

Practical molecular communication systems need to achieve application-dependent goals while remaining robust to the environment encountered in applications. A living organism from nature is an example of such a system and it exhibits enormous complexity. Applications of molecular communication previously discussed in this chapter may also require a complex system consisting of a massive number of bio-nanomachines that communicate (e.g., a dermal display to be made from 3 billion bio-nanomachines). The question is then whether it is really possible to design and engineer such complex systems.

A communication engineering approach to deal with system complexity is to define the architecture of a system [10]. In computer networks, system functionality is divided into a stack of *layers*. Each layer implements a service for the layer above by using services provided by the layer below and adding the necessary functionality. For example, one layer n may implement a service of reliable data transmission based on unreliable data transmission provided by the layer below $n-1$ and an error handling mechanism to be added within that layer n. The layered approach has the following advantages. First, it allows system designers to focus on a specific part of a large system through encapsulating the detail of other layers. Second, a change in implementation of a service in a layer does not affect the design of the rest of the layers as long as that layer provides the same service.

The architecture for molecular communication may be discussed from a computer networks perspective [5] – the TCP/IP reference model or the Internet protocol suite [10] may provide a starting point for our discussion. As an example, Figure 4.2 illustrates how information may flow from a *source* through a *router* to the *destination*, representing groups of sender bio-nanomachines, bio-nanomachines with routing functionality, and receiver bio-nanomachines, respectively. The source and router, as well as the router and destination, are within a communication range, meaning that information molecules can propagate from one to the other within a reasonable amount of time to induce the intended reactions. Similar to the TCP/IP reference model, the application layer provides a set of options to implement applications; the network and link layers provide mechanisms to transmit information over and within a communication range; and the physical layer provides biophysical mechanisms for transmission, propagation, and reception of information molecules over physical media. Briefly, at the source, the application layer initiates molecular communication by inducing a specific chemical reaction that eventually causes an intended reaction at the destination application layer. The network layer at the source selects a communication channel and the link layer ensures that the channel is available. The physical layer transmits and propagates information molecules over the selected channel to the router. The router then similarly selects a communication channel, ensures that the channel is available, and transmits

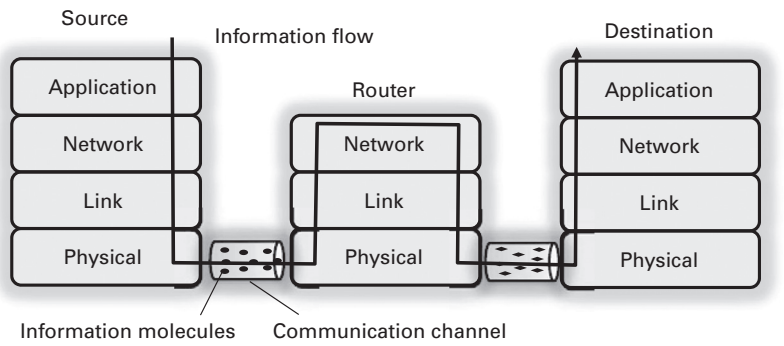


Figure 4.2 Flow of information through a molecular communication network modeled after the Internet architecture.

and propagates a type of information molecule to the destination. Finally, the destination reacts to the incoming information molecules produced by the router and initiates an application-dependent action.

In the rest of this chapter, we discuss a network architecture of molecular communication. We begin with a discussion of the physical layer responsible for transmitting information molecules over physical media. We then progress to the link layer for reliable transmission of information over a communication link between bio-nanomachines, and to the network layer for transmission of information in a network of bio-nanomachines interconnected through communication links. We also briefly discuss higher layers and other issues specific to molecular communication that may need to be considered to establish an architecture.

4.3.1 Physical layer

The physical layer of a molecular communication architecture provides the biophysical basis to address issues related to transmission, propagation, and reception of information molecules. One issue is the selection of hardware and interfaces appropriate for molecular communication by bio-nanomachines. Other issues are representing a signal using information molecules (i.e., signal modulation); propagating information molecules in the environment (signal propagation); and amplifying a signal that attenuates during propagation (signal amplification). There are also issues with performance considerations such as determining the capacity of molecular communication channels.

Hardware and interface: The physical layer determines the hardware and interfaces of bio-nanomachines. A variety of implementations will be necessary to meet the variety of application-specific constraints such as the performance, scale, and compatibility with the environment of molecular communication. Two types of hardware devices introduced earlier in this chapter include sender and receiver bio-nanomachines capable of encoding and decoding a message onto or from information molecules. Other types of bio-nanomachines introduced are used to support molecular communication between sender and receiver bio-nanomachines and include interface, transport, guide, and addressing molecules. As we saw in Chapters 2 and 3, biological systems provide a great variety of biological materials for molecular communication, such as motor proteins, gap junction channels, and self-propelling bacteria, which could be used as bio-nanomachines. In addition, bio-nanomachines can be implemented by modifying biological materials as we will see in Chapter 7. The physical layer may also provide an interface with optical or electrical systems such as wireless body sensor networks [11, 12, 13] that operate outside the molecular communication environment. Such an interface may be implemented through the use of non-biological materials (see Chapter 7).

Signal modulation: The physical layer also provides mechanisms for representing a signal using information molecules. Analogous to molecular communication among biological systems (e.g., biological cells), a bio-nanomachine in molecular communication distinguishes different types of information molecules using different types of receptors, thus each type of information molecule represents a different signal

[14, 15, 16]. In addition, the physical properties of information molecules of the same type may be modulated to represent a different signal. For instance, the amplitude, frequency, or other aspects such as timing of arrival of information molecules may be modulated to represent a different signal [17, 18, 19, 20, 21, 22, 23].

Signal propagation: The physical layer also provides mechanisms to propagate information molecules in the environment of molecular communication. An information molecule may passively diffuse through the environment without using chemical energy [17, 18, 24, 25, 26] or actively propagate by using a transport bio-nanomachine that consumes ATP energy [18, 27, 28]. Information molecules may also propagate in a fluid medium where the motion of information molecules is influenced by an externally applied force (e.g., a blood stream) [23, 28, 29]. Information molecules may also react to molecules present in the environment to propagate as a reaction-diffusion wave (e.g., calcium signaling) [19].

Signal amplification and relay: The physical layer also provides mechanisms for signal amplification and relay. The concentration of information molecules (i.e., the signal strength) decreases with distance from the sender bio-nanomachine. At longer distances, information molecules are dispersed and are unlikely to arrive at the receiver bio-nanomachine or are too low in concentration to be detected by the receiver bio-nanomachine [30]. Biological systems use positive feedback loops to amplify the signal strength and mechanisms to relay chemical messages over a long distance. For instance, the calcium concentration in a biological cell is amplified through the repetitive binding of calcium ions to protein channels, which induces the release of calcium ions from the protein channels. Since calcium ions are produced through the calcium-induced calcium release and at the same time relayed cell-to-cell, a signal (i.e., a high calcium concentration) propagates over a number of cells. Biochemical positive feedback likely becomes a key to design mechanisms for signal amplification and relay in molecular communication.

Channel capacity: The design of the physical layer determines the capacity of a communication channel. Information theory can be applied to molecular communication to model the impact of noise sources, such as the randomness in the propagation of molecules, on the channel capacity. We will come back to the detail in Chapter 6.

4.3.2 Link layer

The link layer of a molecular communication architecture provides a set of mechanisms for a group of bio-nanomachines to reliably communicate information within a communication range. A communication range can be roughly defined as an environment in which information molecules can propagate from a sender bio-nanomachine to a receiver bio-nanomachine in a reasonably short amount of time to induce biochemical reactions at the receiver bio-nanomachine. The link layer provides mechanisms for handling errors, controlling the rate of transmission (i.e., flow control), sharing a medium (media access control), addressing receiver bio-nanomachines, synchronizing clocks in bio-nanomachines, and measuring distance among bio-nanomachines.

Error handling: The link layer provides mechanisms to deal with errors that may occur when information molecules propagate from a sender bio-nanomachine to a receiver bio-nanomachine. Error handling may be required since information molecules may degrade in the environment or arrive with a large amount of jitter (i.e., dispersed arrival times and in a different order from the original transmission), which may lead to unintended receiver reactions, i.e., errors. As in biological systems, a sender bio-nanomachine may transmit a redundant number of information molecules, and the receiver bio-nanomachine reacts only when it detects a threshold number of such molecules. A larger number of information molecules increases the signal-to-noise ratio and reduces the impact of noise or fluctuation [18, 26]. Information molecules may also be encapsulated by an interface molecule to avoid degradation of the information molecules during propagation, thereby the error rate decreases. Such an interface molecule may be implemented with vesicles that are used as a container to transport materials within a biological cell. It may also be possible to embed error detection and correction codes within a pattern of transmitting information molecules. For instance, Hamming codes may be embedded in a bit sequence represented with a type of information molecule [31]. It may also be possible to embed such codes within a structure of an information molecule (e.g., a DNA sequence) [32] for error handling at a receiver bio-nanomachine.

Flow control: The link layer also provides mechanisms for flow control to adjust the rate of transmitting information molecules at a sender bio-nanomachine. A sender bio-nanomachine may need to adjust the transmission rate since the chemical kinetics at the receiver bio-nanomachine may be rapid (e.g., enzyme reactions), extremely slow (e.g., cell growth in the order of hours), or variable. Without flow control, a sender bio-nanomachine may transmit information molecules at a higher rate than the reaction rate at the receiver bio-nanomachine. Information molecules in this case may accumulate in the environment and eventually degrade, resulting in molecule loss [33]. Alternatively, a sender bio-nanomachine may transmit information molecules at a lower rate than the reaction rate at the receiver bio-nanomachine. This then degrades the throughput (e.g., the number of information molecules that react with the receiver bio-nanomachine per unit time). A potential approach to implement flow control for molecular communication is to incorporate a negative feedback loop as in biological systems (e.g., gene regulatory networks [34]). For example, a sender bio-nanomachine may produce information molecules through gene expression, and the receiver bio-nanomachine, in response to the incoming information molecules, produces repressor molecules to decrease the rate of the gene expression at the sender bio-nanomachine to achieve flow control.

Media access control: The link layer also provides mechanisms for sharing a medium among a group of bio-nanomachines to implement media access control. Without media access control, the simultaneous transmission of information molecules can cause an error such as a response from an unintended receiver bio-nanomachine. In biological systems, interference is often avoided by using a different type of molecule for each receiver bio-nanomachine, similar to using a different channel in a wireless network to avoid interference. This requires a large number of different types of molecules that do

not interfere with each other. In the case where interference does occur, media access control may apply mechanisms for multiplexing. For instance, media access control may be implemented by a group of bio-nanomachines that synchronize and transmit information molecules at different times (i.e., time division multiplexing) [35] or transmit information molecules of the same type with different characteristics to allow demultiplexing at a receiver bio-nanomachine (i.e., frequency or amplitude division multiplexing) [36].

Addressing: The link layer also provides addressing mechanisms by which a sender bio-nanomachine specifies receiver bio-nanomachines within a communication range. A sender bio-nanomachine may use addresses associated with receiver bio-nanomachines to communicate with the receiver bio-nanomachines. Such addresses can be implemented with biological materials such as a pair of complementary DNA sequences [37] – if one DNA sequence is attached to an information molecule and the other sequence to the receiver bio-nanomachine, then the information molecule binds to the receiver bio-nanomachine through the pairing of DNA sequences. A sender bio-nanomachine may also use location addresses to communicate with receiver bio-nanomachines at a specific location in a communication range. In developmental biology, a location is addressed by concentrations of molecules at the location, and the developmental process of a biological cell progresses based on the location. Similarly, for molecular communication, a location in a communication range may be addressed by a set of concentrations of molecules, and information molecules may be designed to propagate toward the location specified by a set of concentrations of molecules [38].

Synchronization: The link layer also provides mechanisms for synchronization of *clocks* in bio-nanomachines. Biological systems use biochemical clocks (e.g., circadian clocks), and bio-nanomachines may contain such clocks. Synthetic biology also demonstrates that biochemical clocks can be introduced artificially into bacterial cells [39]. Synchronization of clocks may become useful when a group of bio-nanomachines perform some function at the same time or to avoid interference through time-slotted communication [40]. In biological systems, synchronization of clocks is observed in heart cells contracting together at the same time, quorum sensing by bacteria to decide when to form a film, or sequential cell differentiation during developmental growth, indicating that the synchronization of bio-nanomachines' clocks is feasible.

Distance measurement: The link layer also provides mechanisms for measuring distance to bio-nanomachines in the environment. Distance information may be useful for tuning the distribution of bio-nanomachines, identifying the relative location of bio-nanomachines, or for optimizing the rate of transmission (e.g., for flow control). In electronic radio networks, a pair of transceivers that communicate by radio waves can use time-of-flight and signal attenuation of the radio waves to measure distance between the transceivers [41]. In molecular communication, similar techniques may apply since the expected time for a molecule to propagate increases with distance (i.e., a characteristic similar to time-of-flight) and the concentration of molecules decreases with distance (i.e., a characteristic similar to signal attenuation) [42].

4.3.3 Network layer

The network layer of a molecular communication architecture provides mechanisms for a group of bio-nanomachines to communicate information in a range larger than the communication range obtained by the link layer. The mechanisms of the network layer include distributing bio-nanomachines in the environment to form a network of bio-nanomachines interconnected through communication links (channels), selecting communication links among bio-nanomachines to transmit information from a sender bio-nanomachine to the receiver bio-nanomachine (i.e., network routing), processing information within a network of bio-nanomachines, and controlling congestion in a network of bio-nanomachines.

Network formation: The network layer provides mechanisms to form a network of bio-nanomachines, i.e., groups of bio-nanomachines interconnected through communication links. A particular network may be formed to meet application-specific goals such as delay, success rate, or energy efficiency in propagating information over the network. Mechanisms for network formation may be implemented using the distance measurement functionality provided by the link layer. For instance, bio-nanomachines may measure the distance to a bio-nanomachine in a communication range and move away from the bio-nanomachine, if it is within a specific distance. This may result in a network of bio-nanomachines separated by that distance. Since a network of bio-nanomachines represents a spatial pattern of bio-nanomachines established in the environment, mechanisms of pattern formation in biological systems may also provide promising approaches for distributing bio-nanomachines in a self-organized manner.

Network routing: The network layer provides mechanisms for routing to transmit information from a source bio-nanomachine to the destination bio-nanomachine in a network of bio-nanomachines. Without routing, the range of molecular communication is limited to a communication range achieved by a single communication link where information molecules can reliably propagate to receiver bio-nanomachines. Network routing increases the scale and complexity of molecular communication systems that may meet the needs of applications. Similar to computer networks, routing in molecular communication may involve router bio-nanomachines located in the environment, and information may be transmitted from a source bio-nanomachine via router bio-nanomachines to the destination bio-nanomachine. For instance, a source bio-nanomachine transmits (or broadcasts) information molecules with addressing molecules (e.g., a DNA sequence used to carry information and addresses) to indicate the destination bio-nanomachines. A router bio-nanomachine then receives an information molecule and applies statically defined chemical processes to determine whether the destination falls within the local network (i.e., a set of bio-nanomachines within a communication range). If the destination falls within the local network, the router bio-nanomachine does not perform any action and the information molecule continues to propagate to reach the destination bio-nanomachine. If the destination does not fall in the local network, the router bio-nanomachine selects a communication link to the next router bio-nanomachine on the path to the destination bio-nanomachine and transmits the information molecules through the link [43, 44].

Network processing: The network layer also provides mechanisms for processing information within a network of bio-nanomachines. Network processing (or in-network processing) is likely to improve the performance of molecular communication. Similar to sensor networks, a network of bio-nanomachines may consist of a large number of bio-nanomachines that are located close by and therefore sense similar information. Information in such cases may be aggregated and transmitted only from selected bio-nanomachines to an application, which may improve the energy efficiency in molecular communication. Information collected by bio-nanomachines may also be combined or averaged, and transmitted to an application to improve the accuracy of the information collected [45].

Congestion control: The network layer may also provide mechanisms to avoid congestion. Without congestion control, a network of bio-nanomachines may be overwhelmed with a large number of information molecules that are transmitted. A large number of information molecules may increase the rate of collision, reduce the speed of propagation, increase communication interference, or induce unintended reactions during propagation. Mechanisms for congestion control may be adapted from biological systems. Biological experiments indicate that a network of filaments within a cell experiences congestion under certain conditions, and motor proteins that propagate over the network slow down the speed of propagation [46]. It was observed that motor proteins avoid congestion by reducing their run length as well as increasing their rate of dissociation.

4.3.4 Upper layers and other issues

Upper layers, such as the transport and application layers, may also be included in an architecture of molecular communication. In addition, an architecture of molecular communication may need to consider key issues that cut across layers, such as energy efficiency. An architecture of molecular communication may also need to consider communications at multiple scales in time and space. Finally, we note that the standardization of an architecture is important for designing a large-scale and complex molecular communication system from bio-nanomachines.

Upper layers: Following the TCP/IP reference model, an architecture of molecular communication may include the transport and application layers. The transport layer provides mechanisms for reliably transmitting information end-to-end. This may include error handling, flow control, and in-sequence transmission of information molecules from a source bio-nanomachine to the destination bio-nanomachine, similar to the services provided by the TCP for the Internet. The application layer provides options to implement application-specific functionality. The specific functionality to be provided by the application layer will become more apparent in the future as applications are being developed.

Cross-layer architecture: The layered architecture described thus far provides advantageous features, such as allowing system designers to focus on issues in one layer by hiding the detail of other layers. Similar to wireless communication, however, a

layered architecture may be violated through a cross-layer design [47] to improve a certain system parameter such as energy efficiency in molecular communication, quality of service (QoS) (e.g., differentiated communication), or even security [48]. A cross-layer architecture may increase the design complexity, but allows a global optimization, which cannot be done with a layered architecture.

Multiscale architecture: It is also important to incorporate communications at different scales within an architecture of molecular communication. A molecular communication system may be integrated to the Internet to allow access to bio-nanomachines via the Internet [49], and this could introduce new issues related to micro-scale to macro-scale interaction. A molecular communication system may also be designed in a hierarchical manner. For instance, a group of bio-nanomachines may form a nanoscale network, a collection of nanoscale networks then forms a micro-scale network, and interactions may occur across the two very different scales. Biological systems are multiscale in nature and biological science is developing techniques for multiscale modeling [50]. However, it remains a challenge to determine how the multiscale modeling techniques may apply to extend an architecture of molecular communication or to develop generalized techniques to study the interplay among communications at different scales.

Alternative architecture: Alternative architectures may be adapted from systems biology. One promising architecture is the bow-tie architecture [51] where interactions among molecules are modeled as a fan-in fan-out network; i.e., a large number of input molecules are converted into a small number of core molecules, which are then converted into a large number of output molecules. For instance, in a cell, energy sources such as metabolites are converted into glucose molecules, and the glucose molecules are then converted into a large number of different molecules such as amino acids, sugars, and nucleotides. The bow-tie architecture is considered robust since it is simple to add a new energy source or to remove an existing energy source from the interaction network. Such a network structure identified in systems biology may serve as a foundation for an architectural design of molecular communication systems.

Standardization: The standardization of an architecture will be important at some point to facilitate the development of molecular communication systems. In computer networks, standards allow different computers developed by different vendors to communicate. In synthetic biology, a standard library for BioBrick (i.e., bio-nanomachines), called the registry of standard biological parts, is developed for engineering a synthetic living organism from a set of well-defined DNA sequences [52]. In molecular communication, the IEEE P1906.1 Standards Working Group for Nanonetworking was established in 2011 to promote the standardization of molecular communication [53, 54]. The near-term goal of the working group is to provide a definition of molecular communication, a conceptual framework for molecular communication, and common terminology for molecular communication. The long-term goal is to identify a practical architecture and a set of reusable protocols for molecular communication through design, implementation, and evaluation of molecular communication systems.

4.4 Conclusion and summary

In this chapter, we looked at the paradigm of molecular communication. We first developed an initial model of molecular communication for design and analysis of a simple molecular communication system, and illustrated components and processes involved in molecular communication. We then observed that a system of molecular communication displays unique features that make it distinct from the existing electrical or optical communication paradigms and thus applications to a variety of domains are anticipated. Finally, we discussed a possible approach to design a practical molecular communication system that is expected to be large-scale and complex. We organized system functionality into a stack of layers following the layering concept, and identified functionality (i.e., services) that should be provided by each layer. We also noticed many issues that need to be considered such as cross-layer and multiscale issues. As a result of the many issues, a considerable amount of research effort needs to be made in the future to establish a complete network architecture for molecular communication.

References

[1] S. Hiyama, Y. Moritani, T. Suda, R. Egashira, A. Enomoto, M. Moore, and T. Nakano, "Molecular communication," in *Proc. NSTI Nanotechnology Conference*, vol. 3, 2005, pp. 392–395.

[2] M. Moore, A. Enomoto, T. Suda, T. Nakano, and Y. Okaie, *The Handbook of Computer Networks*. John Wiley & Sons Inc, 2007, vol. 3, ch. Molecular communication: new paradigm for communication among nano-scale biological machines, pp. 1034–1054.

[3] I. F. Akyildiz, F. Brunetti, and C. Blazquez, "Nanonetworks: a new communication paradigm," *Computer Networks*, vol. 52, no. 12, pp. 2260–2279, 2008.

[4] S. Hiyama and Y. Moritani, "Molecular communication: harnessing biochemical materials to engineer biomimetic communication systems," *Nano Communication Networks*, vol. 1, no. 1, pp. 20–30, 2010.

[5] T. Nakano, M. Moore, F. Wei, A. V. Vasilakos, and J. W. Shuai, "Molecular communication and networking: opportunities and challenges," *IEEE Transactions on NanoBioscience*, vol. 11, no. 2, pp. 135–148, 2012.

[6] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Cell Biology*, 5th edn. W. H. Freeman, 2003.

[7] D. Pantarotto, J.-P. Briand, M. Prato, and A. Bianco, "Translocation of bioactive peptides across cell membranes by carbon," *Chemical Communications*, vol. 1, pp. 16–17, 2004.

[8] R. Phillips, J. Kondev, and J. Theriot, *Physical Biology of the Cell*. Garland Science, 2008.

[9] H. Tanaka, K. Homma, A. H. Iwane, E. Katayama, R. Ikebe, J. Saito, T. Yanagida, and M. Ikebe, "The motor domain determines the large step of myosin-V," *Nature*, vol. 415, pp. 192–195, 2002.

[10] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach*, 6th edn. Addison Wesley, 2012.

[11] I. F. Akyildiz and J. M. Jornet, "Electromagnetic wireless nanosensor networks," *Nano Communication Networks*, vol. 1, pp. 3–19, 2010.

[12] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. M. Leung, "Body area networks: a survey," *Mobile Networks and Applications*, vol. 16, no. 2, pp. 171–193, 2011.

[13] G.-Z. Yang, *Body Sensor Networks*. Springer, 2006.

[14] A. W. Eckford, "Achievable information rates for molecular communication with distinct molecules," in *Proc. Workshop on Computing and Communications from Biological Systems: Theory and Applications*, 2007, pp. 313–315.

[15] M. S. Kuran, H. B. Yilmaz, T. Tugcu, and I. F. Akyildiz, "Interference effects on modulation techniques in diffusion based nanonetworks," *Nano Communication Networks*, vol. 3, no. 1, pp. 65–73, 2012.

[16] T. Nakano and M. Moore, "In-sequence molecule delivery over an aqueous medium," *Nano Communication Networks*, vol. 1, no. 3, pp. 181–188, 2010.

[17] A. W. Eckford, "Nanoscale communication with Brownian motion," in *Proc. 41st Annual Conference on Information Sciences and Systems*, 2007, pp. 160–165.

[18] M. Moore, T. Suda, and K. Oiwa, "Molecular communication: modeling noise effects on information rate," *IEEE Transactions on NanoBioscience*, vol. 8, no. 2, pp. 169–180, 2009.

[19] T. Nakano and J. Q. Liu, "Design and analysis of molecular relay channels: an information theoretic approach," *IEEE Transactions on NanoBioscience*, vol. 9, no. 3, pp. 213–221, 2010.

[20] M. U. Mahfuz, D. Makrakis, and H. T. Mouftah, "On the characterization of binary concentration-encoded molecular communication in nanonetworks," *Nano Communication Networks*, vol. 1, no. 4, pp. 289–300, 2010.

[21] M. Pierobon and I. F. Akyildiz, "A physical end-to-end model for molecular communication in nanonetworks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 4, pp. 602–611, 2010.

[22] C. T. Chou, "Molecular circuits for decoding frequency coded signals in nano-communication networks," *Nano Communication Networks*, vol. 3, pp. 46–56, 2012.

[23] K. V. Srinivas, R. S. Adve, and A. W. Eckford, "Molecular communication in fluid media: the additive inverse Gaussian noise channel," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4678–4692, 2012.

[24] B. Atakan and O. B. Akan, "Deterministic capacity of information flow in molecular nanonetworks," *Nano Communication Networks*, vol. 1, pp. 31–42, 2010.

[25] M. Pierobon and I. F. Akyildiz, "Diffusion-based noise analysis for molecular communication in nanonetworks," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2532–2547, 2011.

[26] T. Nakano, Y. Okaie, and J. Q. Liu, "Channel model and capacity analysis of molecular communication with Brownian motion," *IEEE Communications Letters*, vol. 16, no. 6, pp. 797–800, 2012.

[27] N. Farsad, A. Eckford, S. Hiyama, and Y. Moritani, "A simple mathematical model for information rate of active transport molecular communication," in *Proc. 2011 IEEE INFOCOM Workshop on Molecular and Nanoscale Communications*, 2011, pp. 473–478.

[28] A. W. Eckford, N. Farsad, S. Hiyama, and Y. Moritani, "Microchannel molecular communication with nanoscale carriers: Brownian motion versus active transport," in *Proc. IEEE International Conference on Nanotechnology*, 2010, pp. 854–858.

[29] A. W. Eckford, "Timing information rates for active transport molecular communication," in *Proc. 4th International ICST Conference on Nano-Networks*, 2009, pp. 24–28.

[30] T. Nakano and J. Shuai, "Repeater design and modeling for molecular communication networks," in *Proc. 2011 IEEE INFOCOM Workshop on Molecular and Nanoscale Communications*, 2011, pp. 501–506.

[31] M. S. Leeson and M. D. Higgins, "Forward error correction for molecular communications," *Nano Communication Networks*, vol. 3, no. 3, pp. 161–167, 2012.

[32] F. Walsh, S. Balasubramaniam, D. Botvich, T. Suda, T. Nakano, S. F. Bush, and M. O. Foghlu, "Hybrid DNA and enzyme based computing for address encoding, link switching and error correction in molecular communication," in *Proc. Third International Conference on Nano-Networks*, 2008, pp. 28–38.

[33] T. Nakano, Y. Okaie, and A. V. Vasilakos, "Throughput and efficiency of molecular communication between nanomachines," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2012, pp. 709–713.

[34] U. Alon, *An introduction to systems biology: design principles and practices*. Chapman & Hall Mathematical and Computational Biology Series, 2007.

[35] S. Balasubramaniam, N. T. Boyle, A. Della-Chiesa, F. Walsh, A. Mardinoglu, D. Botvich, and A. Prina-Mello, "Development of artificial neuronal networks for molecular communication," *Nano Communication Networks*, vol. 2, no. 2–3, pp. 150–160, 2011.

[36] M. J. Moore and T. Nakano, "Multiplexing over molecular communication channels from nanomachines to a micro-scale sensor device," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 4158–4523.

[37] S. Hiyama, T. Inoue, T. Shima, Y. Moritani, T. Suda, and K. Sutoh, "Autonomous loading, transport, and unloading of specified cargoes by using DNA hybridization and biological motor-based motility," *Small*, vol. 4, no. 4, pp. 410–415, 2008.

[38] M. Moore and T. Nakano, "Addressing by beacon distances using molecular communication," *Nano Communication Networks*, vol. 2, no. 2–3, pp. 161–173, 2011.

[39] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335–338, 2000.

[40] M. Moore and T. Nakano, "Synchronization of inhibitory molecular spike oscillators," in *Proc. 6th International International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems (BIONETICS)*, 2011, pp. 183–195.

[41] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, vol. 34, no. 8, pp. 57–66, 2001.

[42] M. Moore, T. Nakano, A. Enomoto, and T. Suda, "Measuring distance from single spike feedback signals in molecular communication," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3576–3587, 2012.

[43] L. C. Cobo and I. F. Akyildiz, "Bacteria-based communication in nanonetworks," *Nano Communication Networks*, vol. 1, no. 4, pp. 244–256, 2010.

[44] P. Lio and S. Balasubramaniam, "Opportunistic routing through conjugation in bacteria communication nanonetwork," *Nano Communication Networks*, vol. 3, no. 1, pp. 36–45, 2012.

[45] A. Einolghozati, M. Sardari, A. Beirami, and F. Fekri, "Consensus problem under diffusion-based molecular communication," in *Proc. 45th Annual Conference on Information Sciences and Systems (CISS)*, 2011.

[46] J. L. Ross, "The impacts of molecular motor traffic jams," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5911–5912, 2012.

[47] V. Srivastava and M. Motani, "Cross-layer design: a survey and the road ahead," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 112–119, 2005.
[48] F. Dressler and F. Kargl, "Towards security in nano-communication: Challenges and opportunities," *Nano Communication Networks*, vol. 3, no. 3, pp. 151–160, 2012.
[49] I. F. Akyildiz and J. M. Jornet, "The internet of nano-things," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 58–63, 2010.
[50] G. S. Ayton, W. G. Noid, and G. A. Voth, "Multiscale modeling of biomolecular systems: in serial and in parallel," *Current Opinion in Structural Biology*, vol. 17, pp. 192–198, 2007.
[51] M. Csete and J. Doyle, "Bow ties, metabolism and disease," *Trends Biotechnology*, vol. 22, pp. 446–450, 2004.
[52] A. Arkin, "Setting the standard in synthetic biology," *Nature Biotechnology*, vol. 26, no. 7, pp. 771–774, 2008.
[53] S. F. Bush, "Wireless ad hoc nanoscale networking," *Wireless Communications Magazine*, vol. 16, no. 5, pp. 6–7, 2009.
[54] IEEE P1906.1 – recommended practice for nanoscale and molecular communication framework, http://standards.ieee.org/develop/project/1906.1.html.

5 Mathematical modeling and simulation

In this chapter, we are concerned with *mathematical models* for molecular communication systems, which allow engineers to perform mathematical analysis, design, and optimization on communication systems. Furthermore, the system model also provides a level of mathematical abstraction, which allows a communication system engineer to understand a molecular communication system along with the biochemical background provided elsewhere in this book. As a result, tools from the vast literature on communications systems may be adapted to molecular communication.

In this chapter, we review recent results in channel modeling for molecular communication. These, and related, models are used in Chapter 6 to calculate the channel capacity of molecular communication, but we present brief examples to introduce the communication problem. This chapter requires familiarity with basic probability and the Gaussian distribution; for a brief review, see Appendix.

5.1 Discrete diffusion and Brownian motion

Free molecules in a fluid propagate via *Brownian motion*, i.e., the random motion induced by collisions with the fluid's molecules. Although highly random, Brownian motion is always available, and has the advantage of zero energy cost to the user. Furthermore, Brownian motion is a very well studied phenomenon, with a rich mathematical literature. In this section, we exploit that literature to derive mathematical models for Brownian motion.

5.1.1 Environmental assumptions

Most molecular communication architectures assume that the sender bio-nanomachine (which we call the *transmitter*) and receiver bio-nanomachine (which we call the *receiver*) are connected by a fluid medium, and are located some distance apart. Further, in molecular communication, we have message-bearing signal molecules and molecules in the surrounding medium; only the former concern us, and throughout this chapter a "molecule" refers to signal molecules.

Initially, it will be convenient to make the following assumptions:

1. The transmitter is a point source of molecules, located at the origin, and is moreover the only source of molecules;

2. Molecules are immutable (i.e., they do not change identity or disappear while propagating), and their motions are independent and identically distributed (iid);
3. Once a molecule is released from the transmitter, the molecule does not interact with the transmitter in any way;
4. The receiver is a surface surrounding a connected region of points, called $\mathcal{P}$, which does not include the origin; and
5. The medium is infinite in every direction, with no barrier or obstacle except $\mathcal{P}$.

We call these the *standard assumptions* for modeling a molecular communication system. An example to illustrate these assumptions is given below.

EXAMPLE 5.1 (Transmitter-receiver system in two dimensions) Consider a two-dimensional fluid medium in accordance with the standard assumptions. Suppose the transmitter is a point source at the origin, and suppose the receiver is a region $\mathcal{P}$ located some distance away from the transmitter. Such an arrangement is depicted in Figure 5.1.

The arrangement in the figure can be imagined as a fluid environment on a two-dimensional surface, like a petri dish or a microscope slide. The transmitter may represent the point at which molecules are injected (say from a micropipette or other mechanism that exists above the plane of the figure), while the ellipsoid receiver may represent a living cell.

Broadly speaking, the remainder of the chapter is concerned with the propagation of molecules through the environment until they intersect with $\mathcal{P}$. There are various ways to mathematically model this propagation, of which we consider the Wiener process. We also give simulation algorithms for these models.

5.1.2 The Wiener process

We first consider the *Wiener process*, a simple physical model for Brownian motion that is appropriate when friction is minimal [1, 2]. Since the Wiener process is both a mathematically important model, and a model that is simple to understand and use, we will spend some time describing its properties.

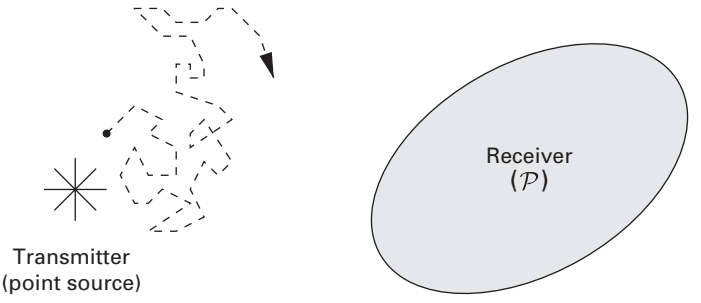


Figure 5.1 Depiction of a molecular communication system in two dimensions. The transmitter is the point source on the left, and the receiver is the connected region $\mathcal{P}$ (ellipsoid) on the right.

The Wiener process is defined in terms of the Gaussian distribution, which we give here for reference (see also Appendix). Let the notation $x \sim N(m, \sigma^2)$ signify that x is Gaussian distributed with mean m and variance σ^2; that is, x has probability density function (pdf) $f_X(x)$ given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{5.1}$$

where $\exp(t) = e^t$.

With the pdf in (5.1) in mind, let $B(t)$ represent the position of a Brownian motion at time $t \geq 0$, where $B(0)$ represents the initial position. Then $B(t)$ is a *one-dimensional Wiener process* if two criteria are satisfied:

1. For any times t_1 and t_2 (where $t_2 > t_1 \geq 0$), and some constant σ^2,

$$B(t_2) - B(t_1) \sim N(0, \sigma^2(t_2 - t_1)); \tag{5.2}$$

and

2. For two intervals $[t_1, t_2]$ and $[t_3, t_4]$, the increments $B(t_4) - B(t_3)$ and $B(t_2) - B(t_1)$ are statistically independent if the intervals do not overlap.

Intuitively, the Wiener process is a continuous-time random process with independent Gaussian-distributed increments. For instance, suppose the process starts off at $B(0) = 0$, and suppose we check the progress of the process once per second, i.e., at $B(1)$, $B(2)$, $B(3)$, and so on. Then $B(k)$, the position at time k, is a Gaussian random variable with distribution

$$B(k) \sim N(0, k\sigma^2), \tag{5.3}$$

since it is known that $B(0) = 0$. Furthermore, the increment between successive views of the process is $B(k) - B(k-1)$, which has the distribution

$$B(k) - B(k-1) \sim N(0, \sigma^2), \tag{5.4}$$

which is independent for each value of k (because the intervals do not overlap). Thus, we have a process whose instantaneous variance increases over time, but where the increments are independent and identically distributed (iid). The same is true if we sample more frequently, say at intervals of Δt, where $\Delta t < 1$ second. In this case the instantaneous distribution is

$$B(k\Delta t) \sim N(0, k\Delta t\sigma^2), \tag{5.5}$$

and the incremental distribution is $B(k\Delta t) - B((k-1)\Delta t) \sim N(0, \Delta t\sigma^2)$. Ultimately, the entire Wiener process emerges as $\Delta t \to 0$.[1]

[1] As $\Delta t \to 0$, the jumps of the Brownian motion have diminishing variance (see (5.5)), but remain independent of each other. As a result, the Wiener process has the unusual mathematical property of being everywhere continuous, but nowhere differentiable.

In physical Brownian motion, the variance parameter σ^2 is given by

$$\sigma^2 = \alpha D, \tag{5.6}$$

where D is the *free diffusion coefficient* of the molecule propagating in the given medium, and where $\alpha = 2$, 4, or 6 if the system is 1-, 2-, or 3-dimensional, respectively. The value of D is given by

$$D = \frac{k_B T}{6\pi \eta R_H}, \tag{5.7}$$

where $k_B = 1.38 \times 10^{-23}$ J/K is the Boltzman constant, T is the temperature (in K), η is the dynamic viscosity of the fluid, and R_H is the hydraulic radius of the molecule. In [3], values of D in the range 1–10 $\mu m^2/s$ were considered realistic for signaling molecules in biotechnological applications. (See also Chapter 4.)

5.1.3 Markov property

The independent increments are a key property of the Wiener process. If at a given time t, the position $B(t)$ of the process is known exactly, then the process is split into two independent intervals: all time up to t, and all time after t. Similarly, it does not matter to past values of the process what the molecule does after time t.

To describe this property, we can use the *Markov chain*. For example, let x, y, and z be dependent random variables. The joint pdf of *any* triple of random variables can be written

$$f_{X,Y,Z}(x,y,z) = f_{Z|Y,X}(z|y,x) f_{Y|X}(y|x) f_X(x). \tag{5.8}$$

The variables x, y, and z form a Markov chain if

$$f_{X,Y,Z}(x,y,z) = f_{Z|Y}(z|y) f_{Y|X}(y|x) f_X(x). \tag{5.9}$$

We say that random variables whose joint pdfs can be written like (5.9) have the *Markov property*. We can generalize this to larger numbers of variables: let $x_1, x_2, \ldots, x_m$ represent a sequence of random variables; then these random variables form a Markov chain if

$$f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = f_{X_1}(x_1) \prod_{i=2}^{m} f_{X_i|X_{i-1}}(x_i|x_{i-1}), \tag{5.10}$$

so that, given x_{i-1}, each variable x_i is conditionally independent of the past $x_1, x_2, \ldots, x_{i-2}$. The Markov property is sometimes described with the phrase: "The future is independent of the past given the present."

The conditional independence of (5.9) can be used to express the independent increments of a Wiener process. Let $t_1, t_2, \ldots, t_k$ represent a series of observation times, where $0 < t_1 < t_2 < \ldots < t_k$, and let $\{B(t_1), B(t_2), \ldots, B(t_k)\}$ be the corresponding points of a Wiener process (further, let $t_0 = 0$, and let the initial point be $B(0) = B(t_0) = 0$). In full

generality, the joint pdf of these points can be written

$$f_{B(t_1),B(t_2),\ldots,B(t_k)}(b_1,b_2,\ldots,b_k) = f_{B(t_1)}(b_1)\prod_{i=2}^{k} f_{B(t_i)|B(t_1),B(t_2),\ldots,B(t_{i-1})}(b_i|b_1,b_2,\ldots,b_{i-1}). \tag{5.11}$$

Consider the term

$$f_{B(t_i)|B(t_1),B(t_2),\ldots,B(t_{i-1})}(b_i|b_1,b_2,\ldots,b_{i-1}), \tag{5.12}$$

which describes the dependence of $B(t_1),B(t_2),\ldots,B(t_{i-1})$ on $B(t_i)$. If $B(0)=0$, and letting $t_0=0$, then for each $i\in\{1,2,\ldots,k\}$ we can write

$$B(t_i) = (B(t_i)-B(t_{i-1}))+B(t_{i-1}). \tag{5.13}$$

Thus, the increment $B(t_i)-B(t_{i-1})$ is independent of any increment prior to t_{i-1}, so given $B(t_1),\ldots,B(t_{i-1})$, $B(t_i)$ is only dependent on $B(t_{i-1})$. In other words,

$$f_{B(t_i)|B(t_1),B(t_2),\ldots,B(t_{i-1})}(b_i|b_1,b_2,\ldots,b_{i-1}) = f_{B(t_i)|B(t_{i-1})}(b_i|b_{i-1}) \tag{5.14}$$

for all i, which implies that the points of the Wiener process satisfy the Markov property. A fully detailed discussion can be found in [1, Thm. 5.12].

The Markov property simplifies the analysis of the Wiener process, and we make particular use of this property below, when we discuss simulation of the Wiener process.

5.1.4 Wiener process with drift

Consider a bio-nanomachine traversing a capillary: if the bio-nanomachine released a signal molecule into the bloodstream, that molecule's Brownian motion would be biased in the direction of the blood flow. Thus, we would need to consider the Brownian motion with drift.

The Wiener process derived in Section 5.1.2 was free of drift. That is, the instantaneous average of the Wiener process is always equal to its initial point: $E[B(t)]=B(0)$ for all $t\geq 0$. (To see this, note from (5.2) that the expected value of each increment is zero.) In a *Wiener process with drift*, the increment distribution in (5.2) is replaced with

$$B(t_2)-B(t_1)\sim N(v(t_2-t_1),\sigma^2(t_2-t_1)), \tag{5.15}$$

where v is the *drift velocity*. (To simplify the analysis, we assume that the drift velocity is constant over time, but this can be relaxed.)

Aside from the expected value, the statistical properties of the Wiener process with drift, including the Markov property, are identical to the Wiener process without drift. This is because the drift component is deterministic, and we may subtract it from the

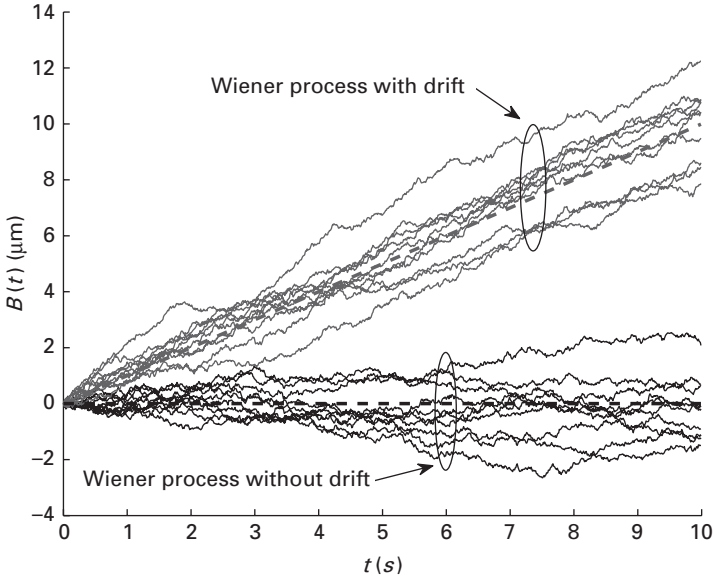


Figure 5.2 Ten realizations each of the Wiener process with and without drift. The expected value of each process is given by a thick dashed line. Parameters of the Wiener processes are given in Example 5.2.

motion: for instance, if $B(t)$ is a Wiener process with drift velocity v, then we may form a process $B'(t)$, given by

$$B'(t) = B(t) - vt, \tag{5.16}$$

which is a Wiener process without drift. The following example further illustrates Wiener processes with and without drift.

EXAMPLE 5.2 (Wiener process with and without drift) Consider two Wiener processes: one with drift velocity $v = 1$ μm/s, and the other with $v = 0$. In both cases, let $D = 1$ $\mu m^2/s$, and let $B(0) = 0$. For each Wiener process, ten realizations of the Brownian motion $B(t)$ are plotted in Figure 5.2.

5.1.5 Multi-dimensional Wiener processes

Although one-dimensional random processes are easy to analyze, the world is three-dimensional. Therefore, an accurate description of molecular motion should be described as a three-dimensional Brownian motion.

In three dimensions, we give the position of the Wiener process $B(t)$ by the triple

$$B(t) = (B_x(t), B_y(t), B_z(t)), \tag{5.17}$$

where $B_x(t)$, $B_y(t)$, and $B_z(t)$ represent the x, y, and z components of the Brownian motion, respectively; where $B_x(t)$, $B_y(t)$, and $B_z(t)$ are independent and identically distributed random processes.

If the Brownian motion has drift, we may break the drift velocity into its x, y, and z components v_x, v_y, and v_z, respectively. The random processes $B_x(t)$, $B_y(t)$, and $B_z(t)$ are independent, but possibly not identically distributed as the component velocities are different in general.

5.1.6 Simulation

The discussion leading up to (5.4) gives the basic idea behind simulating the Wiener process. Suppose time is quantized into regular intervals of Δt; then since each increment is independent and identically distributed (using the Gaussian distribution), one need only obtain a running sum of Gaussian random variables in order to simulate the points of a Wiener process.

Indeed, thanks to the central limit theorem (which states that the sum of many iid random variables approaches the Gaussian distribution), each term in a sufficiently long-running sum need not be Gaussian, which potentially simplifies the simulation. The following two-dimensional simulation algorithm is found in [3]:

1. (Initialization.) Let D represent the free diffusion coefficient of the molecule, and let Δt represent the interval between successive iterations of the simulation. Let (x_0, y_0) represent the initial position of the molecule in two dimensions. Let $i = 1$.
2. Obtain a random variable θ_i, uniformly distributed on $[0, 2\pi)$, independent of any previous $\theta_1, \theta_2, \ldots, \theta_{i-1}$.
3. Let

$$x_i = x_{i-1} + \sqrt{4D\Delta t}\cos\theta_i \tag{5.18}$$

$$y_i = y_{i-1} + \sqrt{4D\Delta t}\sin\theta_i \tag{5.19}$$

   (Note the use of $4D$, consistent with $\alpha = 4$ for two-dimensional Brownian motion.) If the molecule collides with an obstacle or wall, the collision is elastic, i.e., the path is reflected off the obstacle with the angle of reflection equal to the angle of incidence.
4. Increment i and go to 2, unless a stopping condition is reached (e.g., the molecule has arrived at the receiver).

In other words, in each interval of Δt, the molecule takes a step of exactly length $\sqrt{4D\Delta t}$ in a random direction θ. This random direction is enough, after sufficiently many iterations, for the ultimate distribution to be approximately Gaussian. Furthermore, note that the Markov property is satisfied, because the random direction is iid in each interval of Δt.

It is straightforward to include simple drift in the simulation: given velocity components v_x and v_y in the x and y directions, respectively, we replace the respective step equations (5.18) and (5.19) with

$$x_i = x_{i-1} + \sqrt{4D\Delta t}\cos\theta + v_x\Delta t \tag{5.20}$$

$$y_i = y_{i-1} + \sqrt{4D\Delta t}\sin\theta + v_y\Delta t \tag{5.21}$$

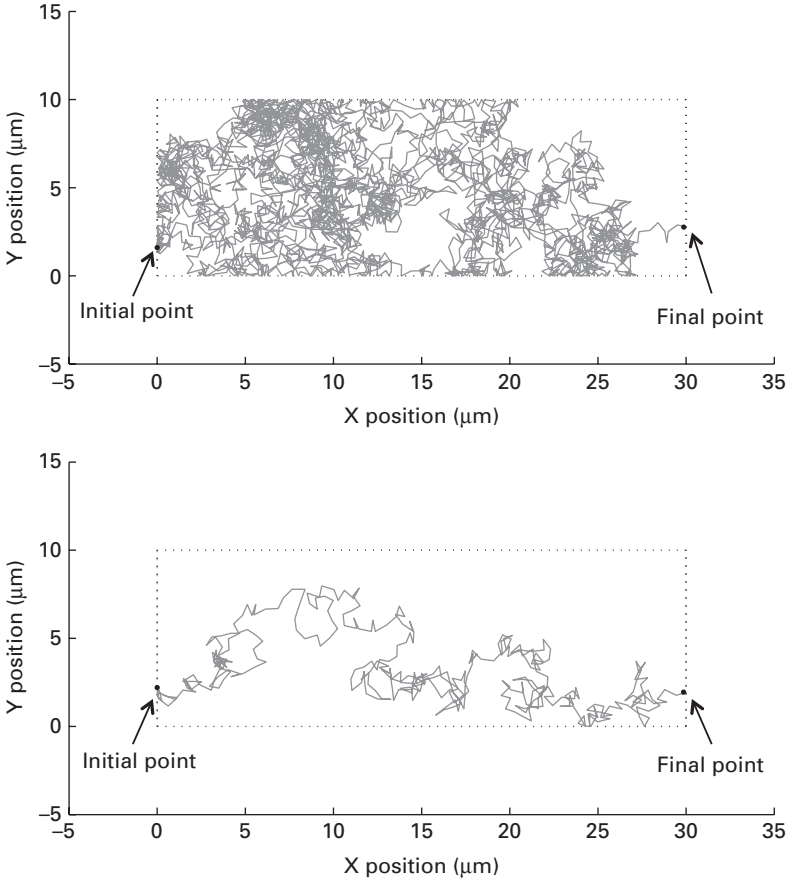


Figure 5.3 Examples of simulated Brownian motion in a box with dimensions 10 μm by 30 μm. The molecule originates at the wall on the left, and propagates until it hits the wall on the right. There is no drift in the top figure, and a drift of 10 μm/s in the bottom figure. In this simulation, $D = 10\ \mu m^2/s$, and all dimensions are in μm.

Examples of simulated Brownian motion propagation are given in the two plots of Figure 5.3. More complicated models of drift, for example having different drift velocities near the boundary, may also be incorporated [3].

5.2 Molecular motors

A major disadvantage of drift-free Brownian motion is the potentially long and highly uncertain delay in the propagation of individual molecules. Adding drift is not always a feasible solution: for example, drift is a one-way phenomenon, which would cause severe difficulties for two-way communication. Moreover, inducing drift in a microchannel is difficult because of high friction in these environments. To solve this

problem, living microorganisms use one of several biochemical reactions, known as molecular motors or motor proteins, to actively transport molecules (see Section 2.1). As pointed out elsewhere, these biochemical reactions may be used as the propagation system in molecular communication.

Unfortunately, these reactions tend to be complex, so much less is known mathematically about these systems. Thus, as we discuss in this section, much of the analysis is done by computer simulation, rather than by closed-form analysis.

An "engines down" motor propagation system consists of a layer of motors anchored to a substrate, while microtubules can freely propagate above the motors, analogous to a conveyor belt. The motion of the microtubule is largely regular, but random variations occur in its trajectory. In particular, since the microtubule is itself a molecule, it is not immune from random deviations via Brownian motion (however, it is large enough that the deviations are relatively small compared to the action of the motors).

A simulation scheme for active transport was given in [4], which replaces (5.18) and (5.19) with

$$x_i = x_{i-1} + \Delta r \cos\theta_i \tag{5.22}$$

$$y_i = y_{i-1} + \Delta r \sin\theta_i, \tag{5.23}$$

where, given the microtubule's average velocity v_{avg} and diffusion coefficient D, Δr is a Gaussian random variable with mean and variance

$$E[\Delta r] = v_{\text{avg}} \Delta t \tag{5.24}$$

$$\text{Var}[\Delta r] = 4D\Delta t. \tag{5.25}$$

Furthermore, θ_i is given by

$$\theta_i = \theta_{i-1} + \Delta\theta, \tag{5.26}$$

where, given persistence length L_p, $\Delta\theta$ is also a Gaussian random variable with mean and variance

$$E[\Delta\theta] = 0 \tag{5.27}$$

$$\text{Var}[\Delta\theta] = v_{\text{avg}} \frac{\Delta t}{L_p}. \tag{5.28}$$

In [4], the experimentally-obtained values of $v_{\text{avg}} = 0.85\ \mu\text{m/s}$, $D = 1.0 \times 10^{-3}\ \mu\text{m}^2/\text{s}$, and $L_p = 111\ \mu\text{m}$ were suggested.[2] An example trajectory is given in Figure 5.4. Though random, it is obvious from the figure that the motor's motion is much more regular than that of Brownian motion, either with or without drift.

We make two remarks on this simulation scheme. First, since the microtubules propagate over a surface, a two-dimensional simulation is sufficient to capture the entire

[2] Values have been modified to make them consistent with our exposition. Original values given at the reference were $D = 2.0 \times 10^{-3}$ and $\text{Var}[\Delta r] = 2D\Delta t$.

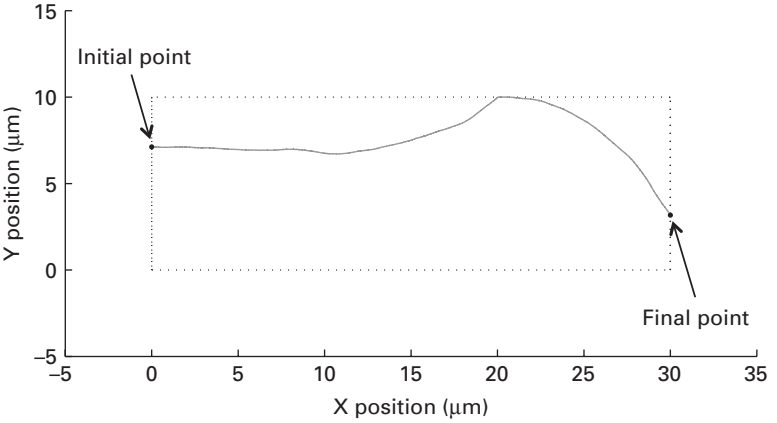


Figure 5.4 Example of simulated trajectory of a microtubule propelled by a molecular motor. All dimensions are in μm, and parameters of the motion are given in the text.

behavior of the system. Second, note from (5.22)–(5.26), the current position of the microtubule (x_i, y_i) is a function of the previous position (x_{i-1}, y_{i-1}) *and* the previous trajectory θ_{i-1}, as well as the iid random variables Δr and $\Delta\theta$. Thus, the Markov property is satisfied so long as the position (x, y) and the trajectory θ are both measured.

In the literature, results for simulated molecular communication systems have been presented in [5, 6].

5.3 First arrival times

A molecule emitted from a transmitter propagates freely in the fluid until its arrival at the receiver. Since the motion is random, the arrival of the molecule occurs at a random time. This random propagation time, called the *first arrival time*, is a significant source of distortion for almost every molecular communication system.

5.3.1 Definition and closed-form examples

Return to the standard assumptions of molecular communication from earlier in this chapter. Recalling that $\mathcal{P}$ is the connected set of points making up the receiver, let $\tau(B(t))$ represent the *first arrival time* of a Wiener process $B(t)$, defined by

$$\tau(B(t)) = \min\{t : t \geq 0, B(t) \in \mathcal{P}\}. \tag{5.29}$$

That is, $\tau(B(t))$ represents the first time that the molecule impinges on the receiver. In one dimension, the region $\mathcal{P}$ consists of the interval $[d, d']$, where $d > 0$ (since $\mathcal{P}$ excludes the origin, under the standard assumptions). In this case, (5.29) reduces to

$$\tau(B(t)) = \min\{t : t \geq 0, B(t) \geq d\}. \tag{5.30}$$

So the first arrival time for the one-dimensional Wiener process is only dependent on the distance d between transmitter and receiver. (In the next chapter, we show that the first arrival time can be viewed as *additive noise* in idealized systems that follow the standard assumptions.)

We can write $f_{\tau(B(t))}(\tau)$ to represent the pdf of $\tau(B(t))$. This pdf is given in closed form in a small number of special cases; the derivations can be found at the references. For example, for an unrestricted one-dimensional Wiener process without drift, where the receiver is located a distance d from the transmitter and the Wiener process variance parameter is σ^2, $f_{\tau(B(t))}(\tau)$ is given by [1]

$$f_{\tau(B(t))}(\tau) = \sqrt{\frac{d^2}{2\pi\sigma^2\tau^3}} \exp\left(-\frac{d^2}{2\sigma^2\tau}\right), \tag{5.31}$$

The distribution in (5.31) has a variety of names in the literature: it is most commonly called the Lévy distribution, but it is also a special case of the inverse Gamma distribution and the Pearson-V distribution.

For an unconstrained one-dimensional Wiener process with a drift velocity $v > 0$, again with the receiver located at distance d, the first arrival time distribution is a member of a family known as the *inverse Gaussian* distributions [7]. Let μ and λ represent the two parameters of this distribution, where

$$\lambda = \frac{d^2}{\sigma^2}, \tag{5.32}$$

and

$$\mu = \frac{d}{v}. \tag{5.33}$$

Then $f_{\tau(B(t))}(\tau)$ is given by

$$f_{\tau(B(t))}(\tau) = \sqrt{\frac{\lambda}{2\pi\tau^3}} \exp\left(-\frac{\lambda(\tau-\mu)^2}{2\mu^2\tau}\right). \tag{5.34}$$

If $f_{\tau(B(t))}(\tau)$ is not available in closed form, it may be obtained using a variety of numerical methods, or by *Monte Carlo* simulation, adapting the simulation technique in the previous section, as we show in the following example.

EXAMPLE 5.3 (First arrival distribution from simulations) Consider an interval $[t, t+\Delta t]$ for any initial time $t > 0$ and any interval length $\Delta t > 0$. Let $p_{[t,t+\Delta t]}$ represent the probability that a molecule's first arrival occurs during the interval $[t, t+\Delta t]$. Suppose m molecules are released, and let $m_{[t,t+\Delta t]}$ represent the number of molecules that arrive during the interval. The expected number of arrivals during the interval is $E[m_{[t,t+\Delta t]}] = mp_{[t,t+\Delta t]}$. So if $p_{[t,t+\Delta t]}$ is unknown, we can estimate it by releasing m molecules, counting the number of arrivals $m_{[t,t+\Delta t]}$, and calculating $p \simeq m_{[t,t+\Delta t]}/m$. As $m \to \infty$, the estimate $m_{[t,t+\Delta t]}/m$ should converge to p with probability 1 (by the weak law of large numbers).

To obtain the first arrival time by simulation, we use the above idea with two modifications: first, we simulate the release of m molecules, rather than actually performing the experiment; and second, we divide all time $t > 0$ into intervals of length Δt, and count the number of arrivals in each interval, thus forming a histogram.

5.3.2 First arrival times in multiple dimensions

In three dimensions, the definition of first arrival time is still given by (5.29), though the connected set of points $\mathcal{P}$ represents the region of space containing the receiver. For general regions $\mathcal{P}$, it is not possible to express the first arrival time in closed form. However, there exist special cases in which the first arrival time reduces to (5.31). Consider a region $\mathcal{P}$ defined by

$$\mathcal{P} = \{(x,y,z) : x \geq d\} \tag{5.35}$$

for some $d > 0$. This is a region partitioned into two zones: one region where $x < d$, and the other where $x \geq d$. Then (5.29) reduces to

$$\tau(B(t)) = \min\{t : t \geq 0, B(t) \in \mathcal{P}\} \tag{5.36}$$

$$= \min\{t : t \geq 0, B_x(t) \geq d\}, \tag{5.37}$$

identical to (5.30) in the x direction. The boundary between the region $\mathcal{P}$ and the complementary region $\bar{\mathcal{P}}$ is then given by the infinite plane in y,z where $x = d$. (Extensions to the y and z cases, as well as the case with drift, are straightforward.)

Now consider the effect on first arrival times. In one dimension, a boundary between one-dimensional regions is a zero-dimensional point, but a boundary between three-dimensional regions is a two-dimensional surface. Thus, first arrival time distributions in three dimensions are measured in terms of the time required to reach a *surface*, which is a generalization of reaching a point in one dimension. (It is vanishingly unlikely that a three-dimensional process will ever visit a particular point in three dimensions; thus, if one insisted on using points, the first arrival time at a given point in three dimensions is infinite with probability 1.)

5.3.3 From first arrival times to communication systems

Suppose the transmitter releases m molecules at times $\mathbf{x} = [x_1, x_2, \ldots, x_m]$, and these molecules' first arrivals occur at times $\mathbf{y} = [y_1, y_2, \ldots, y_m]$. For molecule x_i, the first arrival time y_i is given by

$$y_i = x_i + n_i, \tag{5.38}$$

where n_i is the propagation time until the first arrival. (For now we assume that the release x_i corresponds to the arrival y_i, but if the molecules are indistinguishable, the arrivals may occur in a different order, unknown to the receiver. We consider this case in Chapter 6.)

The transmitter may communicate by selecting different vectors $\mathbf{x}$ of release times. Then the receiver must distinguish the $\mathbf{x}$ that was sent by observing the corrupted version $\mathbf{y}$. This is done by calculating the conditional distribution $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x})$, the distribution of the outputs given the inputs. From (5.38), the random propagation time for each individual molecule may be viewed as *additive noise*. Let $f_N(n)$ represent the first arrival time distribution; then assuming the molecules are distinguishable, so that the output y_i corresponds to the input x_i, the pdf of y_i given x_i is written

$$f_{Y_i|X_i}(y_i|x_i) = f_N(y_i - x_i), \tag{5.39}$$

and under the standard assumptions,

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x}) = \prod_{i=1}^{m} f_{Y_i|X_i}(y_i|x_i) = f_N(y_i - x_i). \tag{5.40}$$

For example, if the Brownian motion is represented by a one-dimensional Wiener process without drift, we have (from (5.31)) that

$$f_{Y_i|X_i}(y_i|x_i) = \sqrt{\frac{d^2}{2\pi\sigma^2(y_i - x_i)^3}}\exp\left(-\frac{d^2}{2\sigma^2(y_i - x_i)}\right). \tag{5.41}$$

We conclude with an example illustrating the importance of $f_{Y|X}(y|x)$ in a communication system.

EXAMPLE 5.4 (Binary transmission) Suppose a molecular communication system is designed to transmit a single bit, as follows: x represents the release time of a molecule, where $x = 0$ to represent bit 0 and $x = 60$ s to represent bit 1. We assume that $d = 10\,\mu\text{m}$ and $\sigma^2 = 4$, that the transmitted bits 0 and 1 are equiprobable, and that the first arrival time of the one-dimensional Wiener process may be used.

Suppose we observe $y = 70$ s. Using (5.41), we may calculate $f_{Y|X}(y|x)$ as 0.0028 for bit 0 ($x = 0$), and 0.0181 for bit 1 ($x = 60$). Using Bayes' rule, we then have that $\Pr(x = 0|y = 70) = 0.136$, and $\Pr(x = 1|y = 70) = 0.864$. Thus, we pick bit 1, because it has the higher probability of being correct; furthermore, our probability of error with this choice is $P_{\text{err}} = 0.136$.

We expand on this theme in Chapter 6.

5.4 Concentration, mole fraction, and counting

Up to now, we have focused on the behavior of individual molecules, measuring exactly when they arrive at the receiver. In this section, we count the number of molecules that arrive, either in an interval of time or a region of space. This allows us to deal with more familiar chemical properties, such as concentration. It also allows us to use concepts from the conventional communications literature. Models presented in this section are closely related to the models given in [8].

5.4.1 Small numbers of molecules: Counting and inter-symbol interference

We now introduce receivers that count the number of arriving molecules, which (for small numbers of molecules) make it easier for us to consider the memory effects of the channel. For instance, molecules that are released in one time interval may arrive in a different interval, which can cause confusion at the receiver. This is similar to inter-symbol interference in conventional communication, where signal energy from a given symbol leaks into neighboring symbols.

The counting receiver breaks time into intervals of length T, and forms the vector $\mathbf{m} = [m_1, m_2, \dots]$, where m_i represents the number of molecules that arrive in the interval from time $(i-1)T$ to iT. Let $m = \sum_i m_i$ represent the total number of released molecules; these molecules can then be indexed by the set $\{1, 2, \dots, m\}$. The molecule release intervals are contained in the integer vector $\mathbf{r} = [r_1, r_2, \dots, r_m]$, where the actual release times are given by $\mathbf{r}T$.

Let $\alpha_{k,i}$ be a binary indicator variable for the arrival time of the kth molecule: $\alpha_{k,i} = 1$ if the kth molecule arrives in the ith interval (i.e., from $(i-1)T$ to iT), and $\alpha_{k,i} = 0$ otherwise. Then

$$m_i = \sum_{k=1}^{m} \alpha_{k,i}. \tag{5.42}$$

Note that $\alpha_{k,i}$ and $\alpha_{j,i}$ are independent if $k \neq j$, since they are given by the independent Brownian motions of different molecules.

We want to find the probability mass function (pmf) $p_{M_i|\mathbf{R}}(m_i \mid \mathbf{r})$. From (5.42), m_i counts the number of "successes" (i.e., arrivals in interval i) out of m "trials" (i.e., released molecules), and thus would have the binomial distribution if $\Pr(\alpha_{k,i} = 1)$ were the same for all k, but the arrival probabilities are generally different for different molecules. Instead, we explicitly derive this pmf.

Let

$$q_i = \int_{(i-1)T}^{iT} f_N(n)dn \tag{5.43}$$

represent the probability that the molecule's propagation time is between $(i-1)T$ and iT. For molecule k released at time $r_k T$, the probability that $\alpha_{k,i} = 1$ is given by

$$\Pr(\alpha_{k,i} = 1) = \int_{(i-1)T}^{iT} f_N(n - r_k T)dn \tag{5.44}$$

$$= \int_{(i-r_k-1)T}^{(i-r_k)T} f_N(n)dn \tag{5.45}$$

$$= q_{i-r_k}. \tag{5.46}$$

Then, writing the pmf of $\alpha_{k,i}$ as $p_{A_{k,i}}(\alpha_{k,i})$, a function of integers $\alpha_{k,i}$,

$$p_{A_{k,i}}(\alpha_{k,i}) = \begin{cases} 1 - q_{i-r_k}, & \alpha_{k,i} = 0; \\ q_{i-r_k}, & \alpha_{k,i} = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.47}$$

Finally, since m_i is the sum of the independent random variables $\alpha_{k,i}$, $p_{M_i|\mathbf{R}}(m_i \mid \mathbf{r})$ and is obtained by taking the m-fold discrete convolution of $p_{A_{k,i}}(\alpha_{k,i})$ for every molecule, substituting m_i for $\alpha_{k,i}$, i.e.,

$$p_{M_i|\mathbf{R}}(m_i \mid \mathbf{r}) = \bigotimes_{k=1}^{m} p_{A_{k,i}}(m_i). \tag{5.48}$$

This form is not very convenient, but it is at least possible to calculate. We illustrate this calculation in the following example.

EXAMPLE 5.5 Suppose first arrival time is inverse Gaussian distributed with $\mu = \lambda = 1$ (see Section 5.3.1), and the sampling interval is $T = 1$. Say there are three molecules, with release intervals given by $\mathbf{r} = [0, 0, 1]$ (with release times given by $\mathbf{r}T$, i.e., two molecules released at time 0, and one released at time T).

We wish to find the distribution of the number of arrivals in the 3rd interval, $p_{M_3|\mathbf{R}}(m_3 \mid \mathbf{r})$. We first need to find all necessary values of q_{i-r_k}: since there are only two possible values of r_k, 0 and 1, these are $q_{3-0} = q_3$, and $q_{3-1} = q_2$. By numerical integration, we have $q_3 = 0.068$, and $q_2 = 0.217$. Performing the convolution in (5.48), we have

$$p_{M_3|\mathbf{R}}(m_3 \mid \mathbf{r}) = \begin{cases} 0.5714, & m_3 = 0; \\ 0.3584, & m_3 = 1; \\ 0.0670, & m_3 = 2; \\ 0.0032, & m_3 = 3; \\ 0, & \text{otherwise.} \end{cases} \tag{5.49}$$

Looking back at the calculation of m_i from (5.42), note that $\alpha_{k,i}$ and $\alpha_{k,j}$ are *dependent*: if $\alpha_{k,i} = 1$, and $j \neq i$, then $\alpha_{k,j} = 0$ with certainty (because the molecule arrived at time i, and can't arrive twice). Thus, m_i and m_j are also dependent in general. Thus, $f_{\mathbf{M}|\mathbf{R}}(\mathbf{m} \mid \mathbf{r})$ is *not* given by the product of $f_{M_i|\mathbf{R}}(m_i \mid \mathbf{r})$, because the counts m_i are not statistically independent for different values of i. In general, $f_{\mathbf{M}|\mathbf{R}}(\mathbf{m} \mid \mathbf{r})$ is not easy to describe. However, an important special case is the *delay-selector channel*, in which the propagation time is upper bounded [9]: the maximum delay is kT, and $q_i = 0$ for all $i > k$. We discuss this channel further in Chapter 6.

As the interval size T becomes small, it becomes possible to release molecules at virtually any time, and to measure their arrival times with greater and greater accuracy. In the limit as $T \to 0$, we arrive at a *timing channel*, as we described earlier.

5.4.2 Large numbers of molecules: Towards concentration

If the number of molecules is large, it is convenient to deal with their count or concentration in different volumes of space, rather than the individual molecules themselves. Obviously, if the space is divided into volumes small enough to contain single molecules, this approach is equivalent to the discrete diffusion case; however, the idea is that precision may be relaxed.

We now take a slightly different approach from the preceding discussion: we now assume that the receiver operates by measuring the concentration *inside the volume* $\mathcal{P}$ (not arriving on the boundary), and that molecules pass freely across the boundary of $\mathcal{P}$. This allows us to directly measure the concentration, rather than having to infer it from the number of arrivals on the boundary.

For our purposes, concentration is defined as the mole fraction,[3] i.e., the ratio of message-bearing molecules at the measurement time T, m_T, to all molecules of all types in $\mathcal{P}$ at time T, $m_{\mathcal{P},T}$:

$$C = \frac{m_T}{m_{\mathcal{P},T}}. \tag{5.50}$$

(As a mole fraction, C is dimensionless, so m_T and $m_{\mathcal{P},T}$ can be expressed in mols, or in numbers of molecules, or any other equivalent measure as long as both are consistent.) This definition is physically straightforward, since unlike measures such as molecules per unit volume, it is invariant with respect to temperature and pressure. Further, if $m_T \ll m_{\mathcal{P},T}$, the value of $m_{\mathcal{P},T}$ is effectively constant, so measuring the mole fraction is equivalent to measuring m_T, i.e., by *counting the number of message-bearing molecules that are present.* Thus, we can continue working with molecule counts as a proxy for concentration.

Our goal in this section is to demonstrate that, for bio-nanomachines, concentration is random: deterministic approaches (e.g., using Fick's law) may give misleading results. Moreover, we will see that the uncertainty is signal-dependent, which can be modeled in a more sophisticated way than by simply adding Gaussian noise.

For simplicity, we first consider a one-dimensional diffusion, in which a "volume" is an interval. Suppose a molecule is released from the origin at time T'. The basic question is this: what is the probability that it will be in volume $\mathcal{P} = [d, d']$ at the measurement time T, where $d' > d > 0$?

From (5.2), and letting $B_{T'}(T)$ represent a Brownian motion at time T for a molecule released at time T', we have that

$$B_{T'}(T) \sim N(0, (T - T')\sigma^2), \tag{5.51}$$

where σ^2 is the variance of the underlying Wiener process. Then

$$\begin{aligned}
&\Pr\left(B_{T'}(T) \in \mathcal{P}\right) \\
&= \Pr\left(B_{T'}(T) \in [d, d']\right) && (5.52)\\
&= \int_{b=d}^{d'} \frac{1}{\sqrt{2\pi\sigma^2(T - T')}} \exp\left(-\frac{b^2}{2\sigma^2(T - T')}\right) db && (5.53)\\
&= \frac{1}{2}\mathrm{erf}\left(\frac{d'}{\sqrt{2\sigma^2(T - T')}}\right) - \frac{1}{2}\mathrm{erf}\left(\frac{d}{\sqrt{2\sigma^2(T - T')}}\right), && (5.54)
\end{aligned}$$

where $\mathrm{erf}(\cdot)$ is the error function.

[3] Sufficiently large number of molecules may be expressed in terms of moles, or mols, where 1 mol = 6.022×10^{23} molecules. This equivalence is given by the Avogadro constant, N_A.

Now consider several molecules in transit at once: suppose n molecules are released at the origin at time T', and let m_T represent the number of these molecules that arrive in $\mathcal{P}$ at time T. Using the standard assumptions, the Brownian motions of the molecules are iid, and so each molecule has an equal probability $\Pr(B_{T'}(T) \in \mathcal{P})$ of being in the volume $\mathcal{P}$ at time T. So the probability of m_T arrivals, given n initial molecules, is like the probability of m_T identically distributed "successes" out of n "trials." Unlike the previous section, since the successes are identically distributed, the pmf is given by the binomial distribution:

$$p_{M_T|N}(m_T \mid n) = \binom{n}{m_T} \Pr(B_{T'}(T) \in \mathcal{P})^{m_T} (1 - \Pr(B_{T'}(T) \in \mathcal{P}))^{n-m_T}. \tag{5.55}$$

As the number of transmitted molecules n becomes large, this probability becomes awkward to deal with: for one thing, the binomial coefficients are difficult to calculate; and for another, the probability of *exactly* a_T molecules arriving becomes very small (for any a_T). However, the probability that the number of molecules is in some range *close to the average* is very high. Moreover, we want to consider molecules released at *different* times. For these two reasons, concentration is best understood by examining the expected values of the molecular counts.

With this in mind, we calculate the mean and variance of a_T given n_T:

$$\mathrm{E}[m_T \mid n] = n \Pr(B_{T'}(T) \in \mathcal{P}) \tag{5.56}$$

and

$$\mathrm{Var}[m_T \mid n] = n \Pr(B_{T'}(T) \in \mathcal{P})(1 - \Pr(B_{T'}(T) \in \mathcal{P})), \tag{5.57}$$

respectively, which are well-known properties of the binomial distribution. Note that these calculations hold for any distribution of $B_{T'}(T)$, not just the one leading up to (5.54).

5.4.3 Concentration: random and deterministic

Following the notation from Section 5.4.1, let $\mathbf{r} = [r_1, r_2, \ldots, r_m]$ represent a sequence of times at which molecules are released, where m represents the number of molecules. (However, r_i are now real numbers that represent the exact time each molecule is released.) We continue to use T as the final measurement time, which must be greater than all elements of $\mathbf{r}$. Recall that the Brownian motion measured at time T, for a molecule released at time r_k, is $B_{r_k}(T)$. For a release at time r_i, let $\alpha_{k,T}$ represent an indicator function, where $\alpha_{k,T} = 1$ if $B_{r_k}(T) \in \mathcal{P}$, and $\alpha_{k,T} = 0$ otherwise. That is, $\alpha_{k,T} = 1$ if the molecule k is in the measurement region at the measurement time T, and is therefore counted by the receiver.

Consider the behavior of the individual molecules: if a single molecule is transmitted at time r_i, then $\alpha_{i,T}$ can be found from (5.56) and (5.57), setting $n = 1$:

$$\mathrm{E}[\alpha_{i,T} \mid \mathbf{r}] = \Pr\left(B_{r_i}(T) \in \mathcal{P}\right) \tag{5.58}$$

$$\mathrm{Var}[\alpha_{i,T} \mid \mathbf{r}] = \Pr\left(B_{r_i}(T) \in \mathcal{P}\right)\left(1 - \Pr\left(B_{r_i}(T) \in \mathcal{P}\right)\right). \tag{5.59}$$

These quantities are the expected value and variance of a Bernoulli random variable, i.e., a random variable that takes values in $\{0, 1\}$. From (5.58),

$$E[m_T \mid \mathbf{r}] = \sum_{i=1}^{m} E[\alpha_{i,T} \mid \mathbf{r}] \tag{5.60}$$

$$= \sum_{i=1}^{m} \Pr\left(B_{r_i}(T) \in \mathcal{P}\right), \tag{5.61}$$

and from (5.59),

$$\mathrm{Var}[m_T \mid \mathbf{r}] = \sum_{i=1}^{m} \mathrm{Var}[\alpha_{i,T} \mid \mathbf{r}] \tag{5.62}$$

$$= \sum_{i=1}^{m} \Pr\left(B_{r_i}(T) \in \mathcal{P}\right)\left(1 - \Pr\left(B_{r_i}(T) \in \mathcal{P}\right)\right), \tag{5.63}$$

where (5.62) follows since the $\alpha_{i,T}$ are independent for different i, which is true since the $\alpha_{i,T}$ are functions of the underlying Brownian motions $B_{r_i}(T)$; under the standard assumptions, these are independent for different molecules.

How does concentration emerge as $E[m_T \mid \mathbf{r}]$ becomes large? For this we use Chebyshev's inequality,

$$\Pr\left(|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{\mathrm{Var}[m_T \mid \mathbf{r}]}\right) \leq \frac{1}{\epsilon^2}, \tag{5.64}$$

which is true for *any* distribution on m. We can simplify this expression to gain some insight: since $\Pr(B_{T'}(T) \in \mathcal{P}) \leq 1$, by the definition of probability, each term in the sum from (5.63) is no greater than the corresponding term in the sum from (5.61). Thus,

$$E[m_T \mid \mathbf{r}] \geq \mathrm{Var}[m_T \mid \mathbf{r}]. \tag{5.65}$$

Noting that $E[m_T \mid \mathbf{r}] > 0$,

$$\Pr\left(|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{\mathrm{Var}[m_T \mid \mathbf{r}]}\right)$$

$$\geq \Pr\left(|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{E[m_T \mid \mathbf{r}]}\right) \tag{5.66}$$

$$= \Pr\left(\left|\frac{m_T - E[m_T \mid \mathbf{r}]}{\sqrt{E[m_T \mid \mathbf{r}]}}\right| > \epsilon\right), \tag{5.67}$$

where (5.66) follows since the event $|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{\mathrm{Var}[m_T \mid \mathbf{r}]}$ includes the event $|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{E[m_T \mid \mathbf{r}]}$ (see (5.65)). Substituting into (5.64),

$$\Pr\left(\left|\frac{m_T - E[m_T \mid \mathbf{r}]}{\sqrt{E[m_T \mid \mathbf{r}]}}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}. \tag{5.68}$$

From (5.68), to use Chebyshev's bound, the range of deviation of m_T away from its mean should scale with the square root of the mean. For very large $E[m_T \mid \mathbf{r}]$, this implies that m is *close to its mean with high probability*. For instance, suppose $E[m_T \mid \mathbf{r}] = 10^{12}$, so $\sqrt{E[m_T \mid \mathbf{r}]} = 10^6$. If $\epsilon = 1000$, then the probability of observing a number of molecules outside $10^{12} \pm 10^9$ (i.e., a variation of 0.1%) is less than $1/\epsilon^2 = 10^{-6}$. Compared to the size of m_T, the size of the deviation from $E[m_T \mid \mathbf{r}]$ becomes small as m_T becomes large, and the deterministic nature of concentration emerges as the deviations from the mean become negligible.

5.4.4 Concentration as a Gaussian random variable

Although Chebyshev's inequality is mathematically rigorous and applies in all circumstances, the concentration around the mean is tighter if we use the Gaussian distribution as a model for concentration.

Looking back at (5.60), this is a sum of independent random variables, and from (5.59), their variances are finite. Using the central limit theorem, we can approximate the distribution of m_T with the Gaussian distribution: that is, m_T is approximately distributed $N(E[m_T \mid \mathbf{r}], \mathrm{Var}[m_T \mid \mathbf{r}])$, with pdf given by

$$f_{M_T \mid \mathbf{R}}(m_T \mid \mathbf{r}) \simeq \frac{1}{\sqrt{2\pi \mathrm{Var}[m_T \mid \mathbf{r}]}} \exp\left(-\frac{(m_T - E[m_T \mid \mathbf{r}])^2}{2\mathrm{Var}[m_T \mid \mathbf{r}]}\right). \tag{5.69}$$

Recall that $E[m_T \mid \mathbf{r}]$ and $\mathrm{Var}[m_T \mid \mathbf{r}]$ are given by (5.61) and (5.63), respectively.

In the Gaussian distribution, the probability of deviation from the mean can be expressed in terms of the complementary error function, $\mathrm{erfc}(\cdot)$: if m_T has the pdf in (5.69), then for any positive constant η,

$$\Pr(|m_T - E[m_T \mid \mathbf{r}]| > \eta) = \mathrm{erfc}\left(\frac{\eta}{\sqrt{2\mathrm{Var}[m_T \mid \mathbf{r}]}}\right). \tag{5.70}$$

We can simplify this expression, and directly compare it to Chebyshev's inequality: letting $\eta = \epsilon\sqrt{E[m_T \mid \mathbf{r}]}$ (as in (5.67)), we have

$$\begin{aligned}
&\Pr\left(|m_T - E[m_T \mid \mathbf{r}]| > \epsilon\sqrt{E[m_T \mid \mathbf{r}]}\right) \\
&\quad = \Pr\left(\left|\frac{m_T - E[m_T \mid \mathbf{r}]}{\sqrt{E[m_T \mid \mathbf{r}]}}\right| > \epsilon\right) && (5.71) \\
&\quad = \mathrm{erfc}\left(\frac{\epsilon\sqrt{E[m_T \mid \mathbf{r}]}}{\sqrt{2\mathrm{Var}[m_T \mid \mathbf{r}]}}\right) && (5.72) \\
&\quad \leq \mathrm{erfc}\left(\frac{\epsilon}{\sqrt{2}}\right), && (5.73)
\end{aligned}$$

where the last line follows from (5.65).

The deviation from the mean, given by (5.70), is much smaller than the maximum deviation in Chebyshev's inequality: $\mathrm{erfc}(\epsilon/\sqrt{2})$ goes to zero much faster than $1/\epsilon^2$.

For example, using $\epsilon = 1000$ as in the previous example, the value of $\text{erfc}(1000/\sqrt{2})$ is essentially zero. To achieve the same 10^{-6} probability as in the Chebyshev case, we need $\epsilon = 4.89$: that is, deviations of more than about five times the square root of the mean are one-in-a-million events.

5.4.5 Concentration as a random process

So far in this section, we have considered the concentration at one particular measurement time T, given sequences $\mathbf{r}$ of molecule releases. However, the behavior of the concentration is most interesting when considered as a random process, i.e., the concentration at many different times.

Consider a set of distinct measurement times $T_1, T_2, \ldots$; how are the concentrations in $\mathcal{P}$ related at these times, given $\mathbf{r}$? For simplicity, we consider the measurement times pairwise, say T_i and T_j for $i \neq j$, and calculate the covariance of the concentrations at these two times. The covariance is important to the analysis of Gaussian random processes, so our analysis here is best used together with the Gaussian approximation from the previous section.

Let m_{T_i} represent the count of molecules in $\mathcal{P}$ at time T_i. Further, for convenience, let $\mu_{T_i} = E[m_{T_i} \mid \mathbf{r}]$. Then the covariance between m_{T_i} and m_{T_j}, written $\sigma_{i,j}$, is defined as

$$\sigma_{i,j} = E[(m_{T_i} - \mu_{T_i})(m_{T_j} - \mu_{T_j}) \mid \mathbf{r}] \tag{5.74}$$

$$= E[m_{T_i} m_{T_j} \mid \mathbf{r}] - \mu_{T_i}\mu_{T_j}. \tag{5.75}$$

The values of μ_{T_i} and μ_{T_j} can be calculated from (5.61), so here we focus on $E[m_{T_i} m_{T_j}]$. Re-using our indicator function notation from the previous section, let $\alpha_{k,T_i} = 1$ if $B_{r_k}(T_i) \in \mathcal{P}$, and $\alpha_{k,T_i} = 0$ otherwise; note the dependence on T_i rather than T. Then, similar to (5.60),

$$m_{T_i} = \sum_{k=1}^{m} \alpha_{k,T_i} \tag{5.76}$$

and

$$m_{T_j} = \sum_{k=1}^{m} \alpha_{k,T_j}. \tag{5.77}$$

Then

$$m_{T_i} m_{T_j} = \left(\sum_{k=1}^{m} \alpha_{k,T_i} \right) \left(\sum_{k=1}^{m} \alpha_{k,T_j} \right) \tag{5.78}$$

$$= \sum_{k=1}^{m} \sum_{\ell \neq k} \alpha_{k,T_i} \alpha_{\ell,T_j} + \sum_{k=1}^{m} \alpha_{k,T_i} \alpha_{k,T_j}. \tag{5.79}$$

Taking the expectation of (5.79), we have

$$E\left[m_{T_i}m_{T_j} \mid \mathbf{r}\right] = E\left[\sum_{k=1}^{m}\sum_{\ell\neq k}\alpha_{k,T_i}\alpha_{\ell,T_j} \mid \mathbf{r}\right] + E\left[\sum_{k=1}^{m}\alpha_{k,T_i}\alpha_{k,T_j} \mid \mathbf{r}\right]. \tag{5.80}$$

In the first term on the right side of (5.80), we have

$$\begin{aligned} &E\left[\sum_{k=1}^{m}\sum_{\ell\neq k}\alpha_{k,T_i}\alpha_{\ell,T_j} \mid \mathbf{r}\right] \\ &= \sum_{k=1}^{m}\sum_{\ell\neq k} E\left[\alpha_{k,T_i}\alpha_{\ell,T_j} \mid \mathbf{r}\right] &(5.81)\\ &= \sum_{k=1}^{m}\sum_{\ell\neq k} E\left[\alpha_{k,T_i} \mid \mathbf{r}\right] E\left[\alpha_{\ell,T_j} \mid \mathbf{r}\right] &(5.82)\\ &= \sum_{k=1}^{m}\sum_{\ell\neq k} \Pr(B_{r_k}(T_i)\in\mathcal{P})\Pr(B_{r_\ell}(T_j)\in\mathcal{P}), &(5.83) \end{aligned}$$

where (5.82) follows because the Brownian motions $B_{r_k}(t)$ and $B_{r_\ell}(t)$, for different molecules k and ℓ, are independent (by assumption). In the second term of (5.80), we have

$$\begin{aligned} &E\left[\sum_{k=1}^{m}\alpha_{k,T_i}\alpha_{k,T_j} \mid \mathbf{r}\right] \\ &= \sum_{k=1}^{m} E\left[\alpha_{k,T_i}\alpha_{k,T_j} \mid \mathbf{r}\right] &(5.84)\\ &= \sum_{k=1}^{m} \Pr(B_{r_k}(T_i)\in\mathcal{P} \wedge B_{r_k}(T_j)\in\mathcal{P}), &(5.85) \end{aligned}$$

where the symbol $\wedge$ represents the conjunction of two events (i.e., that both events happen). That is, the probability under the sum represents the probability of the same Brownian motion $B_{r_k}(t)$ being in the same region $\mathcal{P}$ at two different time instants. From first principles, this is given by

$$\begin{aligned} &\Pr(B_{r_k}(T_i)\in\mathcal{P} \wedge B_{r_k}(T_j)\in\mathcal{P}) \\ &= \int_{B_{r_k}(T_i)\in\mathcal{P}}\int_{B_{r_k}(T_j)\in\mathcal{P}} f(B_{r_k}(T_i)-B_{r_k}(T_j))f(B_{r_k}(T_j)), &(5.86) \end{aligned}$$

where the functions $f(\cdot)$ under the argument are pdfs of their respective arguments (omitting the subscripts). This integral is difficult to evaluate in closed form. However, it can be evaluated numerically to give the covariance.

Given the mean and variance from the above calculations, individual concentrations can be simulated as Gaussian random variables; moreover, given mean and covariance, a time series of concentrations can be simulated as a Gaussian random process.

5.4.6 Discussion and communication example

We see that molecule count (and hence concentration) is sharply concentrated around the mean for "sufficiently large" numbers of molecules: it is for these large numbers that deterministic behavior takes over, and concentration becomes a deterministic, continuous value. But our target application in this book involves bio-nanomachines at very small dimensions, which would use relatively small numbers of signaling molecules. To put this into context, we used 10^{12} molecules to illustrate Chebyshev's inequality: at room temperature and pressure, 10^{12} molecules of water would take up roughly 30 μm^3; for comparison, a single bacterium takes up roughly 1 μm^3. Thus, 10^{12} signaling molecules would occupy many times the amount of space of a single bio-nanomachine, even disregarding any molecules from the propagation medium.

On the other hand, say an average of 1000 signaling molecules are present; this would be a more realistic molecular signal from a bio-nanomachine. The Gaussian approximation suggests that the one-in-a-million range is $4.89 \times \sqrt{1000} = 155$, or about 16%. Since $\operatorname{erfc}(2.58/\sqrt{2}) = 0.01$, the one-percent range of deviation from the mean is $2.58 \times \sqrt{1000} = 82$, or about 8%. Thus, for bio-nanomachines, the random deviations in the count are significant, and it is not realistic to treat concentration as a deterministic value.

As in the previous section, here we give a short example illustrating the use of concentration as a continuous random variable: in this case, we generate a communication signal by generating different numbers (equivalently, different concentrations) of molecules. Again, we can distinguish the signals by calculating the conditional distribution of the final count m given the release times of each molecule $\mathbf{r}$; in the Gaussian example, this is given by $f_{M_T|\mathbf{R}}(m_T \mid \mathbf{r})$ in (5.69). The example also illustrates the accuracy of the Gaussian approximation.

EXAMPLE 5.6 In this example, all quantities are in arbitrary units. Let the medium be one-dimensional, with $\mathcal{P} = [1,2]$, and the transmitter wants to send one of two possible messages: to send message 0, the transmitter releases 1000 molecules at time 0 (so $\mathbf{r}_0 = [0,0,\ldots,0]$, a thousand times); to send message 1, the transmitter releases a thousand molecules at time 1 (so $\mathbf{r}_1 = [1,1,\ldots,1]$, a thousand times). We assume Brownian motion is given by a Wiener process without drift, with $\sigma^2 = 0.5$. Using (5.54), $\Pr(B(T) \in \mathcal{P}) = 0.1359$ and 0.0763 for molecules released at times 0 and 1, respectively. From (5.61)–(5.63), for the first signal we have $E[m_T \mid \mathbf{r}_0] = 135.9$ and $\operatorname{Var}[m_T \mid \mathbf{r}_0] = 117.4$; and for the second signal we have $E[m_T \mid \mathbf{r}_1] = 76.3$ and $\operatorname{Var}[m \mid \mathbf{r}_1] = 70.5$. In Figure 5.5, we plot the two Gaussian probability density functions (conditioned on message 0 and message 1 being sent), along with histograms generated by a simulation of this system.

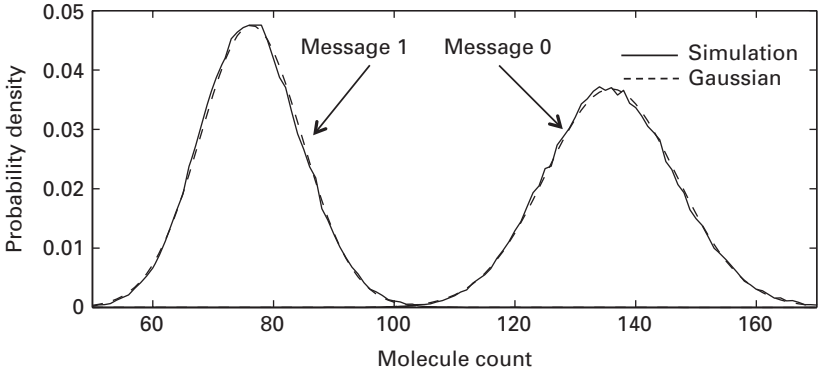


Figure 5.5 Plots of the two Gaussian probability density functions (conditioned on message 0 and message 1 being sent), along with histograms generated by a simulation of this system.

The receiver's task is to count the number of molecules in $\mathcal{P}$, and determine whether message 0 or message 1 was sent. Similarly to Example 5.4, suppose the receiver measures $m_T = 107$ molecules in $\mathcal{P}$: the Gaussian pdf values for signal 0 and 1 are given by $f_{m_T|\mathbf{R}}(m_T \mid \mathbf{r}_0) = 1.1 \times 10^{-3}$ and $f_{m_T|\mathbf{R}}(m_T \mid \mathbf{r}_1) = 5.9 \times 10^{-5}$, respectively. Thus, message 0 is more likely, and we conclude that message 0 was sent.

Again, we expand on our discussion of communication systems in Chapter 6. Note from Figure 5.5 that the Gaussian approximation is quite good for this system. However, if we had assumed that the concentration was deterministic and equal to the mean, that assumption would be quite poor: there is significant random variation in the local population of the signal molecules.

5.5 Models for ligand–receptor systems

Up to this point, we have abstracted away the biochemical machinery of receiving molecules. In this section, we give some mathematical models of ligand–receptor systems, which are widely studied in molecular communication literature.

Biochemically, cells may detect chemical messengers in their environment, known as ligands, when those ligands bind to receptor sites on the cell's surface. The binding process is dependent on the local concentration of ligands: binding is more likely to occur in the presence of higher concentration. However, the unbinding process (i.e., returning the receptor site to the unbound state, ready to receive another ligand) is independent of the surrounding concentration. This process is described in detail elsewhere in the book (see Chapter 2).

5.5.1 Mathematical model of a ligand–receptor system

The state of the receptor is binary: either bound (B), or unbound (U). The receptor's state is a random process, dependent on the past state of the receptor and on the concentration at the receptor input.

In the simplest conception of this model, the system is represented in discrete time, with states $y_i \in \{\mathsf{B}, \mathsf{U}\}$ for each $i \in \{1, 2, \ldots, \ell\}$ (where ℓ is the number of intervals in the communication session). Moreover, the concentration at time i, written x_i, is binary: high (H), or low (L). By assumption, the state y_i of the receptor is dependent only on the most recent state y_{i-1}, and on the current concentration at the receptor input x_i. If $y_i = \mathsf{U}$, then the probability of entering state B in the next time instant is concentration-dependent: α_{H} and α_{L} for input concentrations H and L, respectively. If $y_i = \mathsf{B}$, then the probability of entering state U in the next time instant is β, independent of the concentration.

Let $\mathbf{y}$ represent a vector of states, and let $\mathbf{x}$ represent a vector of concentrations. From the above description, the probability of $\mathbf{y}$ given $\mathbf{x}$ can be written

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = p(y_1 \mid x_1) \prod_{i=1}^{\ell} p(y_i \mid y_{i-1}, x_i). \tag{5.87}$$

Thus, given $\mathbf{x}$, this system is a time-varying Markov chain, with instantaneous Markov transition probability $p_{Y_i|Y_{i-1},X_i}(y_i \mid y_{i-1}, x_i)$ dependent on the input concentration x_i. This transition probability can be represented by one of two state transition probability matrices: if $x_i = \mathsf{L}$, we have

$$\mathbf{P}_{Y|X=\mathsf{L}} = \begin{bmatrix} 1-\alpha_{\mathsf{L}} & \alpha_{\mathsf{L}} \\ \beta & 1-\beta \end{bmatrix}, \tag{5.88}$$

and if $x_i = \mathsf{H}$, we have

$$\mathbf{P}_{Y|X=\mathsf{H}} = \begin{bmatrix} 1-\alpha_{\mathsf{H}} & \alpha_{\mathsf{H}} \\ \beta & 1-\beta \end{bmatrix}. \tag{5.89}$$

In these matrices, U and B are on the first and second row and column, respectively. This model was first given in [10].

Restricting concentration to a binary variable is done for simplicity. However, by making α a function of the input x_i, we can generalize these state transition probability matrices to the case where x_i takes more than two values (including the practical case where x_i takes real values).

5.5.2 Simulation

Given $\mathbf{x}$, the binding sequence $\mathbf{y}$ can be easily simulated as a time-dependent Markov chain: at each time i, we select the binding state y_i, given the previous state y_{i-1} and the current concentration x_i, according to the distribution $p_{Y_i|Y_{i-1},X_i}(y_i \mid y_{i-1}, x_i)$; this distribution is given by the state transition probability matrices (5.88)–(5.89).

In order to fully simulate the system, the input concentration $\mathbf{x}$ should also be simulated. This sequence can also be modeled as a first-order Markov chain with the transition probability matrix

$$\mathbf{P}_X = \begin{bmatrix} 1-r & r \\ s & 1-s \end{bmatrix}, \tag{5.90}$$

with elements representing L on the first row and column, and elements representing H on the second row and column. Under these circumstances, $(\mathbf{x}, \mathbf{y})$ jointly form a Markov chain with states $\{\mathsf{UL}, \mathsf{UH}, \mathsf{BL}, \mathsf{BH}\}$, with the transition probability matrix

$$\mathbf{P}_{Y,X} = \begin{bmatrix} (1-\alpha_{\mathsf{L}})(1-r) & (1-\alpha_{\mathsf{H}})r & \alpha_{\mathsf{L}}(1-r) & \alpha_{\mathsf{H}}r \\ (1-\alpha_{\mathsf{L}})s & (1-\alpha_{\mathsf{H}})(1-s) & \alpha_{\mathsf{L}}s & \alpha_{\mathsf{H}}(1-s) \\ \beta(1-r) & \beta r & (1-\beta)(1-r) & (1-\beta)r \\ \beta s & \beta(1-s) & (1-\beta)s & (1-\beta)(1-s) \end{bmatrix}. \tag{5.91}$$

Using this transition probability matrix, $\mathbf{x}$ and $\mathbf{y}$ can be jointly simulated by selecting the future state (y_i, x_i) given the current state (y_{i-1}, x_{i-1}), according to the probability $p_{Y_i,X_i|Y_{i-1},X_{i-1}}(y_i, x_i \mid y_{i-1}, x_{i-1})$.

5.6 Conclusion and summary

In this chapter, we examined the problem of mathematical modeling and simulation of molecular communication systems. There is a rich analytical literature that can be exploited to solve this problem. To summarize the material presented in this chapter:

- Individual molecules in a fluid propagate via diffusion, which may be modeled using the Wiener process. The Wiener process has an important mathematical property, known as the Markov property, which simplifies its analysis and makes simulation easier. The Wiener process may be extended to add drift or to consider multi-dimensional processes.
- As communication systems rely on the propagation of message-bearing molecules from transmitter to receiver, the first arrival time is a key feature in molecular communication systems. The first arrival time distribution may be obtained analytically in some cases, or more generally by simulation.
- If a large number of molecules are propagating, they may be modeled by their count or concentration in a given volume of space. The observed number of molecules is close to the average with high probability, and the deviation can be modeled with a Gaussian random variable (or a Gaussian process in the case of a time series). However, for bio-nanomachine applications, the number of molecules is small enough that randomness must be taken into account.
- Biologically, molecular reception is performed by ligand–receptor systems, which can be modeled using Markov chains.

In the next chapter, we use these results to mathematically analyze molecular communication systems, and place them within the context of the existing communication- and information-theoretic literature.

References

[1] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd edition. New York: Springer, 1991.

[2] S. Goldstein, "Mechanical models of Brownian motion," *Lecture Notes in Physics*, vol. 153, pp. 21–24, 1982.
[3] J. Berthier, *Microfluidics for Biotechnology*. Artech House, 2006.
[4] T. Nitta, A. Tanahashi, M. Hirano, and H. Hess, "Simulating molecular shuttle movements: Towards computer-aided design of nanoscale transport systems," *Lab on a Chip*, vol. 6, pp. 881–885, 2006.
[5] A. W. Eckford, "Timing information rates for active transport molecular communication," in *Proc. 4th International Conference on Nano-Networks*, Lucerne, Switzerland, 2009.
[6] N. Farsad, A. W. Eckford, S. Hiyama, and Y. Moritani, "Microchannel molecular communication with nanoscale carriers: Brownian motion versus active transport," in *IEEE International Conference on Nanotechnology*, 2010.
[7] R. S. Chhikara and J. L. Folks, *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Marcel Dekker, 1989.
[8] M. Pierobon and I. F. Akyildiz, "Diffusion-based noise analysis for molecular communication in nanonetworks," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2532–2547, June 2011.
[9] L. Cui and A. W. Eckford, "The delay selector channel: Definition and capacity bounds," in *Proc. Canadian Workshop on Information Theory (CWIT)*, 2011.
[10] P. J. Thomas, D. J. Spencer, S. K. Hampton, P. Park, and J. P. Zurkus, "The diffusion mediated biochemical signal relay channel," in *17th Annual Conference on Neural Information Processing Systems*, 2003.

6 Communication and information theory of molecular communication

The models introduced in Chapter 5 give us a mathematical framework to describe the elements of a molecular communication system, particularly the molecules as they traverse the medium. We now take this idea one step further, by describing the statistical interaction between two terminals as they exchange signaling molecules. Since it is the randomness of molecular motion under Brownian motion that creates uncertainty in communication, the models we gave in Chapter 5 play the role of communication noise in this chapter.

There is a rich mathematical literature on information and communication theory. Though much existing work deals with electromagnetic communication, the theories are general enough that we can apply them to molecular communication. In this chapter, we briefly describe these theories, and show how they relate to molecular communication systems. However, we will also see that there exist many open problems in this field, and that solutions are only known for simplified cases.

6.1 Theoretical models for analysis of molecular communication

In Chapter 5, we gave models for diffusion, and discussed some ways in which those models could be used to discuss communication. We now present these models more formally, to allow mathematical analysis of the communication systems. To do so, we must specify not only what is happening as molecules propagate, but also the actions of the transmitter and receiver.

Both in this section and in the chapter as a whole, we will see examples where physically unrealistic assumptions are made to simplify the analysis. Of course, the goal is to be as realistic as possible to produce the best results. However, mathematical reasoning about communication – molecular or otherwise – requires simple mathematical models, which tend to be less physically realistic for molecular communication than for conventional communication systems.

6.1.1 Abstract physical layer communication model

To simplify system design, communication engineers separate the functions of communication systems into *layers* (recall our discussion of layering in molecular communication from Chapter 4). The *physical layer* describes the interaction of the communication

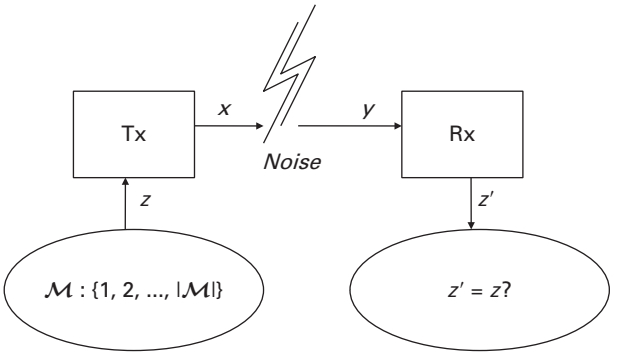


Figure 6.1 A depiction of the communication model. The message source selects a message $z \in \mathcal{M}$. The transmitter (Tx) converts the message to x, which is distorted by noise in the medium. The receiver (Rx) observes y, and uses it to guess the message z'. If $z = z'$, then communication was successful.

system with the physical world. For example, in conventional communication systems, the physical layer deals with the circuits, antennas, and other components needed to transmit information across a channel, as well as the physical features of the channel (such as bandwidth). In molecular communication, the components include the mechanisms that release and detect molecules, the medium that allows those molecules to propagate, and the chemical nature of the molecules themselves.

The physical layer can be given an abstract mathematical model, the components of which are depicted in Figure 6.1: a message z is selected from a discrete, finite set $\mathcal{M}$, and encoded for transmission across the channel. The transmitter maps the message z into allowed channel inputs x, selected from a set $\mathcal{X}$ (a process known as *modulation*). The channel and receiver transform the channel inputs into a channel output y, selected from a set $\mathcal{Y}$ (known as *demodulation*). Given y, the receiver guesses that the transmitter's message was z'. If $z' = z$, then the transmission was successful; otherwise, an error occurs.[1]

The abstract model in Figure 6.1 is sufficiently general to represent any communication system. Moreover, we have already seen that a communication system is characterized by the conditional pdf $f_{Y|X}(y|x)$, expressing the probability of observing a channel output y given a channel input x. To obtain $f_{Y|X}(y|x)$ in molecular communication, we must find a stochastic model for the physical layer communication system, and for this we can use the models from Chapter 5. However, in addition to the propagation method, we must also model the transmission and reception of a signal, which can be done in various ways. For example, in a timing communication system, the inputs, x, consist of the release times of the molecules, and the outputs, y, represent their respective arrival times at the receiver. For a counting system, the inputs, x, represent the

[1] The receiver's guess z' need not be in $\mathcal{M}$. For example, the receiver might realize that it can't make a good guess, and set z' to an element not in $\mathcal{M}$, making it clear that an error has occurred. This may be better than trying to make a wild guess.

number of molecules released, and the outputs, y, represent the number of molecules that arrive. We have seen many examples of these systems elsewhere in the book, and we will analyze some of these in this chapter.

6.1.2 Ideal models

Consider a transmitter and receiver, operating under the standard assumptions, that have the following properties:

1. The transmitter perfectly controls the initial position and release time of the molecule (i.e., the initial point of the Brownian motion).
2. The receiver perfectly measures the arrival time and position of the molecule on its first arrival at the receiver.
3. On the first arrival of the molecule, the receiver absorbs the molecule and removes it from the system.

We call this the Wiener-Ideal (WI) model: Wiener because the underlying Brownian motion is described by the Wiener process, and ideal in the sense that the WI assumptions maximize the information-theoretic capacity with respect to any other possible system operating under the standard assumptions (we will show this later).

The WI model is not particularly realistic; for example, it implies that the transmitter and receiver are perfectly synchronized. However, in addition to being ideal, these assumptions are convenient for analysis: the first two assumptions abstract away all the details of the chemistry involved in transmission and reception, while the third eliminates the need to keep track of multiple arrivals (as well as any behavior of the particle after arrival).

The WI model properties may be generalized beyond the standard assumptions: the first two properties are ideal for any system, but the last one is only ideal for systems that satisfy the Markov property. The Markov property is required because, at the first arrival of the molecule, the exact position and time are known to the receiver; therefore, given the first arrival, any subsequent observations of the molecule are conditionally independent of the molecule's past. Further, we saw in the previous chapter that additional measurements may need to be taken to ensure that the Markov property is satisfied. As a result, letting M represent the physical properties of the molecule that need to be measured to ensure that the Markov property is satisfied, we may generalize the WI model by forcing the transmitter to perfectly control M at release time, and the receiver to perfectly measure M at arrival time (in addition to the particle's position).

6.1.3 Distinguishable molecules: The additive inverse Gaussian noise channel

Consider the simplest case, in which the molecules are all distinguishable. Note that for them to be distinguishable, not only should they be chemically distinct, but the receiver needs to be able to detect the distinction. Again, this is an unrealistic assumption in practice, but it significantly simplifies the model. As we see in this section, such systems

have *additive noise models*, and we use them to introduce one analytically useful model: the additive inverse Gaussian noise (AIGN) channel [1].

In a timing channel, the message x is the release time of the molecule, selected from the set $\mathcal{X}$ of allowed release times. For example, we could let $\mathcal{X}$ be the set of nonnegative real numbers, in which case any release time $x \geq 0$ is allowed. The observation y is the arrival time of the molecule. (Further, $y = \infty$ is a possible outcome, representing the event that the molecule never arrives.)

Suppose the transmitter sends a molecule at time x; that molecule propagates via Brownian motion until the first arrival time n, at which time it is removed from the system. Thus, as we discussed in Chapter 5, the receiver observes

$$y = x + n, \tag{6.1}$$

that is, the signal x is observed in the presence of additive timing noise. Moreover, by assumption, each molecule is distinguishable (so there is no ambiguity in matching receptions y to transmissions x), and Brownian motions of different molecules are statistically independent. Thus, the additive timing noise in each use of the channel is iid.

What kind of distributions can be used for n? From an analytical perspective, closed-form distributions are obviously the most interesting, such as the inverse Gaussian distribution.

EXAMPLE 6.1 Suppose the transmitter is a point source of molecules, and the receiver is an infinite plane located distance d away. Further suppose the channel has a drift velocity of v in the direction of the receiver. Then n has the inverse Gaussian distribution, given in (5.34) in Chapter 5, and parametrized by $\lambda = d^2/\sigma^2$ and $\mu = d/v$. We call such a channel an AIGN channel.

Let $f_N(n)$ represent the AIGN distribution of the first arrival time n. If $y > x$, the conditional pdf of y given x is given by

$$f_{Y|X}(y|x) = f_N(y-x) \tag{6.2}$$

$$= \sqrt{\frac{\lambda}{2\pi(y-x)^3}} \exp\left(-\frac{\lambda(y-x-\mu)^2}{2\mu^2(y-x)}\right). \tag{6.3}$$

(If $y \leq x$, $f_{Y|X}(y|x) = 0$.) Suppose we want to send one of four possible messages, corresponding to release times of $x = 0$, 1, 2, or 3 seconds. Further let $d = \sigma^2 = 1$. Plots of $f_{Y|X}(y|x)$ are given in Figure 6.2 for several values of v.

The AIGN channel has the dual advantages of being based on a physically realistic process (Brownian motion with drift), and of having a first arrival time distribution for which many analytical results are known (see [2]).

6.1.4 Indistinguishable molecules

In the previous section, we considered single uses of each distinguishable molecule. But what happens if only one type of molecule is available? We can release a pattern of

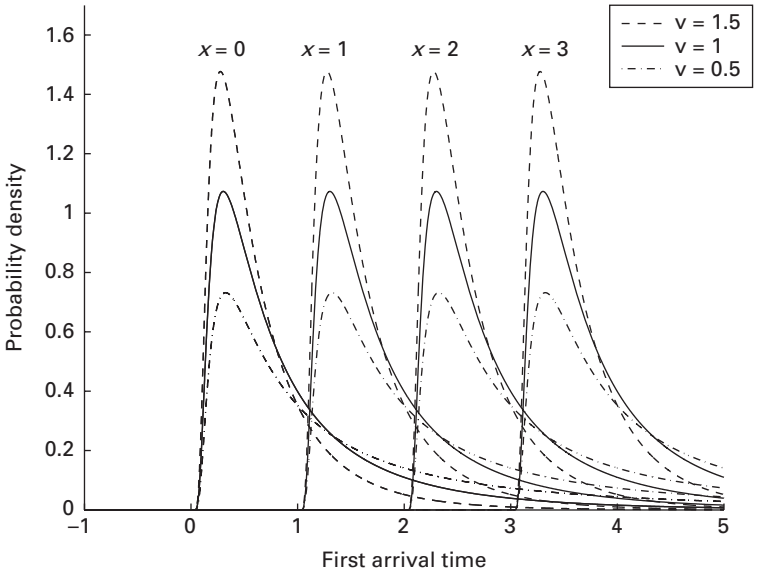


Figure 6.2 Comparison of $f_{Y|X}(y|x)$ for different values of x, in the AIGN channel.

these indistinguishable molecules at many different times to express a message. In this case, the challenge is that the molecules can arrive out of order.

Say $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ contains a vector of molecule release times: one *indistinguishable* molecule is released at x_1, one is released at x_2, and so on. (Without loss of generality, we may assume that $\mathbf{x}$ is sorted in increasing order.) If $\mathbf{n}$ is a vector of independent, identically distributed first arrival times of the Brownian motion, then the receiver observes all the elements of $\mathbf{x} + \mathbf{n}$, similarly to the distinguishable case. However, the receiver does not observe these elements in the order that they were *sent*, it observes them in the order they *arrive*, which can be different. That is, the receiver observes the vector $\mathbf{y}$, where

$$\mathbf{y} = \text{sort}(\mathbf{x} + \mathbf{n}), \tag{6.4}$$

where the sort function sorts the vector in increasing order.

We must deal with the vectors $\mathbf{x}$ and $\mathbf{y}$ together: we can no longer split them into their components, as in the distinguishable case, since a molecule released at time x_1 can arrive first (at position y_1), last (at position y_n), or anywhere in between. As a result, since the channel uses are no longer independent of each other, the channel is said to have memory. In this case, our task is to find the multivariate pdf $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, expressing the input-output probabilities in vector form. Let $\mathbf{F} = [F_{i,j}]$ represent an $n \times n$ matrix, where $F_{i,j} = f_N(y_i - x_j)$ (where $f_N(n)$ is the first arrival time distribution). Then by the Bapat-Beg theorem [3],

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \begin{cases} \text{per}(\mathbf{F}), & y_1 \leq y_2 \leq \cdots \leq y_n \\ 0, & \text{otherwise,} \end{cases} \tag{6.5}$$

where per(·) is the permanent of the matrix in the argument: letting $\mathcal{P}$ represent the set of all permutations on n letters, the permanent is defined as

$$\text{per}(\mathbf{F}) = \sum_{\pi \in \mathcal{P}} \prod_{i=1}^{n} f_{i,\pi(i)}. \tag{6.6}$$

From (6.6), there are n elements in $\mathcal{P}$, so a calculation of the permanent has, at worst, $O(n!)$ steps. Unfortunately, no known algorithm can exactly calculate the permanent in less than $O(n2^n)$ steps, making the permanent intractable for large n. At the time of writing, managing the complexity of this calculation is an important open problem.[2]

6.1.5 Sequences in discrete time

One strategy to mitigate the complexity is to limit the maximum propagation delay that a molecule can experience: this is the idea behind the delay-selector channel [5].

Similar to the models described in Section 5.4.1, the delay-selector channel is a counting channel, where the transmitter releases molecules only at the beginning of intervals, and the receiver counts the number of arrivals per interval. Such a receiver can be derived from the WI model: given the exact arrival times, the number of arrival times that occur in each interval are returned. As a result, the maximum-delay assumption in the delay-selector channel implies that molecules arrive within a given number of intervals. This is a simplifying assumption for the sake of analysis (which is why we introduce it here), but may be realistic in the presence of strong drift or with the use of molecular motors.

Using the notation from Section 5.4.1, the probability of being delayed by i intervals is q_i, and there is some maximum k such that $q_i = 0$ for all $i > k$. When the delay is small, the number of possible permutations that occur under reordering is relatively small, and can be dealt with tractably. Furthermore, for this channel, we define the inputs $\mathbf{x}$ and outputs $\mathbf{y}$ as sequences: x_i gives the number of molecules released by the transmitter at time $(i-1)T$, and y_i gives the number that arrived at the receiver between $(i-1)T$ and iT. We assume there are ℓ intervals, and the inputs are assumed to be binary (i.e., $x_i \in \{0, 1\}$) and iid. (Note that this notation is slightly different from the counting channel described in Chapter 5.)

We will deal with the simplest case, where the maximum delay k is equal to 1: a molecule arrives either in the same interval as it was transmitted, or one interval later. We use a (hidden) state variable s_i to represent the channel memory, defined as

$$s_{i+1} = x_i - y_i + s_i, \tag{6.7}$$

which is equal to the number of molecules that are delayed by one time instant. (The channel starts with $s_1 = 0$.) It should be clear from the definition of the channel that

[2] Promising recent results have shown that the permanent can be approximated using an efficient message-passing algorithm. See [4].

$s_i \in \{0,1\}$. For example, letting $\mathbf{s} = [s_1, s_2, \ldots, s_\ell]$, a valid triple of input, state, and output sequences for this channel is given by

$$\begin{array}{rcccccccccc} \mathbf{x} & = & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ \mathbf{s} & = & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ \mathbf{y} & = & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 2 & 0 \end{array}. \tag{6.8}$$

Now consider the calculation of $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$. For simplicity, first consider $s_{i-1} = 0$ (i.e., at time i, the channel is free of delayed molecules). We can write

$$p_{Y_i|S_i,X_i}(y_i|s_i = 0, x_i) = \begin{cases} 1, & x_i = 0,\ y_i = 0 \\ 1 - q_0, & x_i = 1,\ y_i = 0 \\ q_0, & x_i = 1,\ y_i = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{6.9}$$

But if $s_i = 1$, then there is a delayed molecule that *must* arrive in the interval along with y_i (because the maximum delay is 1). Thus, in general,

$$p_{Y_i|S_i,X_i}(y_i|s_i, x_i) = \begin{cases} 1, & x_i = 0,\ y_i = s_i \\ 1 - q_0, & x_i = 1,\ y_i = s_i \\ q_0, & x_i = 1,\ y_i = 1 + s_i \\ 0, & \text{otherwise.} \end{cases} \tag{6.10}$$

Furthermore, from (6.7), s_{i+1} is deterministic given y_i, s_i, and x_i, so

$$p_{S_{i+1}|Y_i,S_i,X_i}(s_i|y_i, s_i, x_i) = \begin{cases} 1, & s_{i+1} = x_i - y_i + s_i \\ 0, & \text{otherwise.} \end{cases} \tag{6.11}$$

Although $\mathbf{s}$ is not observed by the receiver, we can nonetheless write the joint probability distribution

$$p_{\mathbf{Y},\mathbf{S}|\mathbf{X}}(\mathbf{y},\mathbf{s}|\mathbf{x}) \prod_{i=1}^{\ell} p_{Y_i|S_i,X_i}(y_i|s_i,x_i) p_{S_{i+1}|Y_i,S_i,X_i}(s_{i+1}|y_i,s_i,x_i), \tag{6.12}$$

and then obtain the input-output sequence probability

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{s}} p_{\mathbf{Y},\mathbf{S}|\mathbf{X}}(\mathbf{y},\mathbf{s}|\mathbf{x}). \tag{6.13}$$

For efficient calculation, the sum in (6.13) can be distributed over the terms in s_i in (6.12), summing over terms in s_1 first, then s_2, and so on.

As a result of the distributed summation, the complexity of obtaining $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ scales with the number of possible states in s_i. In this example, there are only two possible states; however, the number of states grows exponentially with k, since the hidden state

s_i must encode not only the number of molecules that have been delayed, but when they were originally transmitted. Thus, the delay-selector channel is most useful with small k. In full generality with unlimited delays, the calculation would be comparable in complexity to the calculation of the permanent from (6.5).

6.2 Detection and estimation in molecular communication

A rudimentary question about communication is the following: if we want to transmit a message from a discrete set of messages $\mathcal{M}$, how can that message be detected, and what is the probability of error? This is a well-studied problem in the communication-theoretic literature, and here we provide a brief introduction, using some of the channel models that we have presented so far.

6.2.1 Optimal detection and ML estimation

Assume that you wanted to send a bit $x \in \{0,1\}$. (Considering the abstract physical layer model, in this section we assume that the message z is equal to the transmitted value x.) As in the previous section, the output of the channel is a random variable Y, taking values in the set $\mathcal{Y}$ of allowed channel outputs. The conditional pdf between input and output is $f_{Y|X}(y|x)$.

Determining the bit 0 or 1, given y, is a detection problem, with the objective of minimizing the probability of error (i.e., the probability that 0 is detected as 1, or vice versa). Mathematically, the detector is a function of y, so let $\hat{x}(y)$: $\mathcal{Y} \to \{0,1\}$ represent this detector. An *error* occurs if $\hat{x}(y)$ is not equal to the originally transmitted value of x: for instance, if $\hat{x}(y) = 1$ when 0 is sent, or vice versa. The probability of error, P_{err}, is the average probability of an error event. This is given by

$$P_{\text{err}} = p_X(0)\Pr(\text{error}|x=0) + p_X(1)\Pr(\text{error}|x=1) \tag{6.14}$$

$$= p_X(0)\int_{y:\hat{x}(y)=1} f_{Y|X}(y|0) + p_X(1)\int_{y:\hat{x}(y)=0} f_{Y|X}(y|1) \tag{6.15}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)\int_{y:\hat{x}(y)\neq x} f_{Y|X}(y|x) \tag{6.16}$$

(The formulation in (6.15) is useful for binary inputs, but the one in (6.16) is general enough for nonbinary inputs.) In detection problems, P_{err} is the figure of merit, and should be minimized by the detector.

The *maximum a posteriori* (MAP) detector for x, written $\hat{x}_{\text{MAP}}(y)$, is given by

$$\hat{x}_{\text{MAP}}(y) = \arg\max_{x\in\{0,1\}} f_{X|Y}(x|y), \tag{6.17}$$

where $\arg\max_x$ gives the argument x maximizing the expression (as opposed to max, which gives the maximum value). Intuitively, $\hat{x}_{\text{MAP}}(y)$ is the most probable explanation for x when y is observed.

Moreover, we can write the probability of error as

$$P_{\text{err}} = 1 - \Pr(\text{no error}) \tag{6.18}$$

$$= 1 - \int_y p_{X|Y}(\hat{x}_{\text{MAP}}(y)|y) f_Y(y). \tag{6.19}$$

Maximizing $p_{X|Y}(\hat{x}_{\text{MAP}}(y)|y)$ for each y minimizes P_{err}. Thus, the MAP detector is optimal in terms of minimizing the probability of error.

The closely related *maximum likelihood* (ML) detector calculates

$$\hat{x}_{\text{ML}} = \arg\max_{x\in\{0,1\}} f_{Y|X}(y|x). \tag{6.20}$$

Referring to (6.19),

$$p_{X|Y}(\hat{x}_{\text{ML}}(y)|y) = \frac{f_{Y|X}(y|\hat{x}_{\text{ML}}(y)) p_X(\hat{x}_{\text{ML}}(y))}{f_Y(y)}. \tag{6.21}$$

If the inputs are equiprobable (i.e., $p_X(\hat{x}_{\text{ML}}(y)) = 1/2$), then the ML and MAP solutions are the same. Further, the ML detector is usually easier to calculate. However, if the inputs are not equiprobable, then the MAP detector allows the prior probabilities $p_X(x)$ to influence its decision.

We illustrate these concepts with the following extended example.

EXAMPLE 6.2 (4-ary detection in the AIGN channel) In Chapter 5, we presented a short example about bit detection in the AIGN channel. The reader may wish to verify that all the examples in Chapter 5 implemented ML detection.

Here we extend that result to 4-ary detection, in which the receiver must discriminate among four possible messages. For consistency, we use the same setup as in Example 6.1, with $v = 1$, and we assume that all inputs are equiprobable: $p_X(x) = 1/4$ for $x \in \{0, 1, 2, 3\}$. Thus, we can use the ML detector. Suppose the receiver observes an arrival time of $y = 2.1$ s: since $x = 0, 1, 2$, or 3 s, we calculate $f_{Y|X}(y|x)$ for each of these values. Using (6.3), and noting that $\mu = \lambda = 1$, we obtain

$$f_{Y|X}(y|0) = 0.0983 \tag{6.22}$$

$$f_{Y|X}(y|1) = 0.3442 \tag{6.23}$$

$$f_{Y|X}(y|2) = 0.2198 \tag{6.24}$$

$$f_{Y|X}(y|3) = 0, \tag{6.25}$$

so in this case, $\hat{x}_{\text{ML}}(2.1) = 1$. (6.26)

Consider the set of y for which $\hat{x}_{\mathrm{ML}}(y) = 1$: this is the set of y such that $f_{Y|X}(y|1)$ is largest. It can be shown that the boundaries of this set are the values of y where: $f_{Y|X}(y|1) = f_{Y|X}(y|0)$, given by 1.1147 s; and $f_{Y|X}(y|1) = f_{Y|X}(y|2)$, given by 2.1147 s.[3] Similarly, looking at the plot for $v = 1$ in Figure 6.2, it can be shown that

$$\hat{x}_{\mathrm{ML}}(y) = \begin{cases} 0, & -\infty < y < 1.1147\,\mathrm{s} \\ 1, & 1.1147\,\mathrm{s} \le y < 2.1147\,\mathrm{s} \\ 2, & 2.1147\,\mathrm{s} \le y < 3.1147\,\mathrm{s} \\ 3, & 3.1147\,\mathrm{s} \le y < \infty. \end{cases} \tag{6.27}$$

This is consistent with the result that $\hat{x}_{\mathrm{ML}}(2.1) = 1$, as we found in the previous paragraph.

Finally, we may calculate the probability of error for this detector. Using (6.16), which applies to both the binary and nonbinary cases, we can rewrite this equation as

$$P_{\mathrm{err}} = \sum_{x \in \mathcal{X}} p_X(x) \left(1 - \int_{y:\hat{x}(y)=x} f_{Y|X}(y|x) \right). \tag{6.28}$$

(Note that the value in parentheses is the probability of error, assuming the input x was sent.) The integral may be evaluated either numerically, or using relations from [2]; either way, we find

$$P_{\mathrm{err}} = p_X(0) \cdot 0.290 + p_X(1) \cdot 0.298 + p_X(2) \cdot 0.298 + p_X(3) \cdot 0.008 \tag{6.29}$$

$$= \frac{0.290 + 0.298 + 0.298 + 0.008}{4} \tag{6.30}$$

$$= 0.224. \tag{6.31}$$

Although we presented this example for illustration only, this is a remarkably high probability of error for a digital communication system. To mitigate the errors, the system designer might spread the signals further out in time, so as to make them easier to distinguish from each other. As an exercise, the reader may wish to try this: recalculate (6.27)–(6.31) with $x \in \{0, 2, 4, 6\}$, and observe the decrease in the probability of error.

6.2.2 Parameter estimation

Up to this point, our examples have assumed that the receiver has full knowledge of all system parameters, such as the distance from transmitter to receiver, drift velocity, diffusion intensity, and so on. These parameters might be needed to implement the optimal detector (or to perform other system tasks), but they might not be available in practice. However, we may also use the ML principle to estimate these parameters.

[3] At the boundaries, e.g., where $f_{Y|X}(y|1) = f_{Y|X}(y|0)$, it doesn't matter whether $\hat{x}_{\mathrm{ML}}(y) = 0$ or $\hat{x}_{\mathrm{ML}}(y) = 1$, and either choice is acceptable. For our purposes, we will always choose the greater value.

Suppose we have a conditional pdf $f_{Y|X}(y|x)$ that depends on an unknown parameter v: to emphasize the dependence, we will write $f_{Y|X}(y|x;v)$. If x and y are given, and the pdf is treated as a function of v, then $f_{Y|X}(y|x;v)$ is called the *likelihood function*. The maximum likelihood estimate of v, written $\hat{v}_{\mathrm{ML}}$, is the maximizing value of the likelihood function:

$$\hat{v}_{\mathrm{ML}} = \arg\max_{v} f_{Y|X}(y|x;v). \tag{6.32}$$

Thus, once again, we are selecting the best fit of v for the provided inputs, x, and observations, y.

A common approach in traditional communication systems is to send a "training signal": a known signal x that can be used to estimate the parameter. In our context, the transmitter sends one or more molecules at times agreed upon in advance, and the receiver measures the arrival time in order to estimate the relevant parameter. We illustrate this in the following example.

EXAMPLE 6.3 Assume we have an AIGN channel, and the unknown parameter v is the drift velocity (all other parameters are assumed known). Substituting $\mu = d/v$ into (6.3), the likelihood function may be written

$$f_{Y|X}(y|x;v) = \sqrt{\frac{\lambda}{2\pi(y-x)^3}} \exp\left(-\frac{\lambda(y-x-d/v)^2}{2(d/v)^2(y-x)}\right). \tag{6.33}$$

If a training signal is used, x is known, and all other parameters (except v) are known by assumption. Thus, we can directly maximize (6.33) with respect to v, but a couple of observations simplify the process: first, we can drop any constants that are not functions of v (like the leading square root term) since they don't change the maximizing value; and second, $\exp(\cdot)$ is monotonic, so $\arg\max_v \exp(f(v)) = \arg\max_v f(v)$, and we can drop the exp. Thus, we have

$$\hat{v}_{\mathrm{ML}} = \arg\max_{v} -\frac{\lambda(y-x-d/v)^2}{2(d/v)^2(y-x)} \tag{6.34}$$

$$= \arg\min_{v} \frac{\lambda(y-x-d/v)^2}{2(d/v)^2(y-x)} \tag{6.35}$$

$$= \arg\min_{v} \left(\frac{v(y-x)}{d} - 1\right)^2, \tag{6.36}$$

where we switch from arg max to arg min when we remove the leading negative sign. Then

$$\frac{d}{dv}\left(\frac{v(y-x)}{d} - 1\right)^2 = 2\left(\frac{v(y-x)}{d} - 1\right)\frac{y-x}{d}, \tag{6.37}$$

and finally, setting (6.36) to zero and solving for v,

$$\hat{v}_{\mathrm{ML}} = \frac{d}{y-x}. \tag{6.38}$$

The training signal may consist of several molecules, in which case the estimate is based on all the molecules at once. Moreover, the training signal may be optimized for a given estimation task, or "blind" estimation may be used (where data is detected simultaneously with parameter estimation), but we do not consider these cases.

6.2.3 Optimal detection in the delay-selector channel

The detection task is more complicated in channels with indistinguishable molecules, as the mere calculation of the likelihood function is difficult (see (6.5)–(6.6)). However, we will use the delay-selector channel to simplify analysis.

Suppose we have an output sequence $\mathbf{y}$ from a delay-selector channel, and we want to determine the inputs that produced it. There are two possible approaches. First, we might want to find the best sequence $\mathbf{x}$ given $\mathbf{y}$; this can be found using the Viterbi algorithm [6], and maximizes the probability that the entire sequence is correct. Second, we might want to find the best individual sequence components x_i given $\mathbf{y}$; this can be found using the sum-product algorithm [7], and maximizes the probability that the decision on each x_i is correct. The difference between these two approaches is subtle, but we focus on the latter approach.

In the component-wise approach, the idea is to find, for each i,

$$\hat{x}_{i,\mathrm{MAP}}(\mathbf{y}) = \arg\max_{x_i} p_{X_i|\mathbf{Y}}(x_i|\mathbf{y}), \tag{6.39}$$

using the entire sequence $\mathbf{y}$ to find the best x_i. By Bayes' rule,

$$p_{X_i|\mathbf{Y}}(x_i|\mathbf{y}) = \frac{p_{\mathbf{Y}|X_i}(\mathbf{y}|x_i)p_{X_i}(x_i)}{\sum_{x_i} p_{\mathbf{Y}|X_i}(\mathbf{y}|x_i)p_{X_i}(x_i)}. \tag{6.40}$$

The prior distribution $f_{X_i}(x_i)$ is known, so the problem reduces to finding $p_{\mathbf{Y}|X_i}(\mathbf{y}|x_i)$. The easiest way to do this is to use the hidden state variables s_i from Section 6.1.5. Using the notation $\mathbf{x}_a^b = [x_a, x_{a+1}, \ldots, x_b]$ to represent a subvector of $\mathbf{x}$, we can write

$$\begin{aligned} & p_{X_i|\mathbf{Y}}(x_i|\mathbf{y}) \\ &= \sum_{\mathbf{x}_1^{i-1}, \mathbf{x}_{i+1}^{\ell}, \mathbf{s}} p_{\mathbf{Y},\mathbf{S}|\mathbf{X}}(\mathbf{y},\mathbf{s}|\mathbf{x}) && (6.41) \\ &= \sum_{s_i} p_{Y_i|S_i,X_i}(y_i|s_i,x_i) && (6.42) \\ &\quad \cdot \sum_{\mathbf{x}_1^{i-1}, \mathbf{s}_1^{i-1}} \prod_{j=1}^{i-1} p_{Y_j|S_j,X_j}(y_j|s_j,x_j) p_{S_{j+1}|Y_j,S_j,X_j}(s_{j+1}|y_j,s_j,x_j) && (6.43) \end{aligned}$$

$$\cdot \sum_{\mathbf{x}_{i+1}^{\ell}, \mathbf{s}_{i+1}^{\ell}} \Big(p_{S_{i+1}|Y_i,S_i,X_i}(s_{i+1}|y_i,s_i,x_i) p_{Y_\ell|S_\ell,X_\ell}(y_\ell|s_\ell,x_\ell) \cdot \prod_{j=i+1}^{\ell-1} p_{Y_j|S_j,X_j}(y_j|s_j,x_j) p_{S_{j+1}|Y_j,S_j,X_j}(s_{j+1}|y_j,s_j,x_j) \Big) \tag{6.44}$$

To understand this calculation, the distribution of the sum into (6.43) and (6.44) is made possible by gathering all the relevant terms under the respective sum. Further, notice that all the indices of summation are different between the two sums, so they can be obtained separately. Once the sums are complete, the term in (6.43) is a function of $\mathbf{y}_1^{i-1}$ and s_i, and the term in (6.44) is a function of $\mathbf{y}_{i+1}^{\ell}$ and s_i, so that the final summation in (6.42) produces the desired result. Further, the sums can be calculated efficiently in the same manner as (6.13): the sum in (6.43) is calculated by summing first over (x_1, s_1), then (x_2, s_2), and so on up to (x_{i-1}, s_{i-1}) in the "forward" direction over the time index, while the sum in (6.44) is calculated by summing first over (x_ℓ, s_ℓ), then $(x_{\ell-1}, s_{\ell-1})$ and so on down to (x_{i+1}, s_{i+1}) in the "backward" direction.

The calculation can be visualized using a powerful graphical technique called a *factor graph*, while the calculation from (6.42)–(6.44) is a special case of the sum-product algorithm, which runs over the graph. Use of this technique allows the integration of a wide variety of factor graph methods into the detector, such as graphically-defined error correcting codes, or signal-processing techniques such as the EM algorithm. These techniques are beyond the scope of this book, but a good tutorial is found in [8].

6.3 Information theory of molecular communication

In the previous section, we discussed the input-output distribution $f_{Y|X}(y|x)$ in various contexts related to molecular communication. In this section, we use these input-output distributions to calculate information-theoretic quantities, such as capacity. We begin with a brief primer on information theory and an introduction to the key concepts, for readers who may not be familiar with this area of research. For a more detailed introduction to information theory, an excellent reference is [9].

6.3.1 A brief introduction to information theory

In information theory, it is common to measure the information content of a message in *bits*. Recalling the abstract physical layer model, where $\mathcal{M}$ is a discrete, finite set that contains all the message we want to transmit, we assign each message in $\mathcal{M}$ a unique number in $\{1, 2, \ldots, |\mathcal{M}|\}$. We then use the binary expansions of these numbers to represent each possible message (leading zeros are used to ensure all messages have a binary expansion of the same length). Using this method, each message is equivalent to a binary string with $\log_2 |\mathcal{M}|$ bits. Thus, the information content of any set of finite messages can be measured in bits, even if the message is not explicitly encoded in

binary. The *information rate*, usually measured in bits per second, is $(\log_2 |\mathcal{M}|)/T$, where T is the amount of time taken to transmit a message in $\mathcal{M}$.

From our discussion in the previous section, the system designer's job is to minimize the probability of error. If a particular communication scheme can achieve a probability of error arbitrarily close to zero,[4] that scheme is said to allow *reliable communication*. The foundation of information theory, due to Shannon [10], was (in part) that reliable communication is generally possible at information rates greater than zero, and that the maximum information rate at which information can be sent across a noisy medium, while ensuring reliable communication, can be calculated for arbitrary channels. This maximum information rate is called the *capacity*.

6.3.2 Capacity

Let x represent a discrete-valued random variable taking values in $\mathcal{X}$, with probability mass function (pmf) $p_X(x)$. The *entropy* of x, written $H(X)$, is given by

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}. \tag{6.45}$$

(If $p_X(x) = 0$ for some value of x, then $p_X(x) \log_2 1/p_X(x)$ is equal to 0 by definition.)

If x is continuous-valued, with probability density function (pdf) $f_X(x)$, then $H(X)$ is called the *differential entropy*, given by

$$H(X) = \int_{x \in \mathcal{X}} f_X(x) \log_2 \frac{1}{f_X(x)}. \tag{6.46}$$

For the remainder of this section, we give the differential entropy, noting that the expressions can be changed to regular entropy by replacing pdfs with pmfs, and integrals with sums.

Now suppose there are two random variables, x and y, taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. (The entropy of y, $H(Y)$, is calculated as described in the previous paragraph, substituting y for x.) We define two quantities of X and Y together: first, the *joint entropy* $H(X, Y)$, written

$$H(X, Y) = \int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} f_{X,Y}(x, y) \log_2 \frac{1}{f_{X,Y}(x, y)}, \tag{6.47}$$

and the conditional entropy of y given x, $H(Y|X)$, written

$$H(Y|X) = \int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} f_{X,Y}(x, y) \log_2 \frac{1}{f_{Y|X}(y|x)}. \tag{6.48}$$

[4] Formally, we mean a scheme with parameters that can be set such that $P_{\text{err}} = \epsilon$, for any $\epsilon > 0$. One might think the goal would be to have a probability of error equal to zero, but this is generally not possible, as extremely unlikely events can always cause errors with nonzero probability. However, the "arbitrarily close to zero" criterion is good enough: for example, if we are using a system with a data rate of 10^9 bits per second, and can find system parameters such that $P_{\text{err}} = 10^{-30}$, then it would take about 1000 times the current age of the universe before an error occurred.

In (6.48), note that the pdf outside the log expression is the joint pdf, *not* the conditional pdf. (It is left as an exercise for the reader to show that $H(X,Y) = H(X) + H(Y|X)$.)

The mutual information between x and y, written $I(X;Y)$, can be written in three ways:

$$I(X;Y) = H(Y) - H(Y|X) \tag{6.49}$$

$$= H(X) - H(X|Y) \tag{6.50}$$

$$= H(X) + H(Y) - H(X,Y). \tag{6.51}$$

(As another exercise, the reader may wish to verify that the three right-hand-side expressions are all equal.) For a given input distribution $f_X(x)$, mutual information gives the maximum rate at which information can be transmitted, while maintaining a probability of error arbitrarily close to zero. However, since $I(X;Y)$ is dependent on $f_X(x)$, which is under the control of the system designer, we can maximize with respect to this quantity to get the capacity C:

$$C = \max_{f_X(x)} I(X;Y). \tag{6.52}$$

Thus, C is the highest rate at which reliable communication can occur.

We may extend entropy, mutual information, and capacity to vector quantities. If $\mathbf{x}$ and $\mathbf{y}$ are a channel input vector and output vector, respectively, then their marginal entropies are written $H(\mathbf{X})$ and $H(\mathbf{Y})$, respectively, and the joint entropy is written $H(\mathbf{X},\mathbf{Y})$. All these quantities are calculated in the same manner as for scalars. Mutual information can be calculated, for example, as

$$I(\mathbf{Y};\mathbf{X}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}), \tag{6.53}$$

or using an alternate expression analogous to (6.49) and (6.50). If each vector is length n, the information rate per component is given by $I(\mathbf{Y};\mathbf{X})/n$.

We close this section by proving that the WI model is information-theoretically ideal. Recall the three WI system assumptions from Section 6.1.2. The first two assumptions are obviously ideal, as they imply that measurement, control, and synchronization are all perfect. The third assumption, that arriving molecules are absorbed by the receiver, is less obviously ideal.

Consider instead two receivers:

- The first receiver does not absorb arriving molecules, but permits them to return and be counted again. Let $\mathbf{a}$ represent the vector of arrival times observed by this receiver, which contains all of the first arrival times, as well as the extra arrival times of returning molecules. For instance, if we have two molecules, one that has two arrivals at times 2.2 and 3.7, and one that has three arrivals at 1.4, 2.1, and 2.8, then $\mathbf{a} = [1.4, 2.1, 2.2, 2.8, 3.7]$.
- The second receiver does not absorb arriving molecules either, but includes a "genie" that splits the arrival times into two vectors: $\mathbf{y}$ and $\mathbf{y}'$, where $\mathbf{y}$ contained only the first arrivals, and $\mathbf{y}'$ contained the second and subsequent arrivals. Continuing our example, we would have $\mathbf{y} = [1.4, 2.2]$ and $\mathbf{y}' = [2.1, 2.8, 3.7]$.

For release times $\mathbf{x}$, the mutual information $I(\mathbf{Y};\mathbf{X})$ is clearly the same as the WI receiver, since only the first arrivals are involved. Moreover, by the Markov property of the Wiener process (from Chapter 5), the position and time of each molecule is known at all the times in vector $\mathbf{y}$, so

$$f_{\mathbf{Y}'|\mathbf{Y},\mathbf{X}}(\mathbf{y}'|\mathbf{y},\mathbf{x}) = f_{\mathbf{Y}'|\mathbf{Y}}(\mathbf{y}'|\mathbf{y}). \tag{6.54}$$

Using what we know so far about entropy and mutual information, this implies

$$\begin{aligned} &I(\mathbf{Y},\mathbf{Y}';\mathbf{X}) \\ &\quad = H(\mathbf{X}) + H(\mathbf{Y},\mathbf{Y}') - H(\mathbf{X},\mathbf{Y},\mathbf{Y}') && (6.55)\\ &\quad = H(\mathbf{X}) + H(\mathbf{Y}'|\mathbf{Y}) + H(\mathbf{Y}) - H(\mathbf{Y}'|\mathbf{Y},\mathbf{X}) - H(\mathbf{X},\mathbf{Y}) && (6.56)\\ &\quad = H(\mathbf{X}) + H(\mathbf{Y}'|\mathbf{Y}) + H(\mathbf{Y}) - H(\mathbf{Y}'|\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}) && (6.57)\\ &\quad = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}) && (6.58)\\ &\quad = I(\mathbf{Y};\mathbf{X}), && (6.59) \end{aligned}$$

where (6.57) follows from (6.54) and the definition of conditional entropy. In other words, if we know which arrivals are the first arrivals, then having the subsequent arrivals $\mathbf{y}'$ gives us no extra information. Further, since $\mathbf{a} = \text{sort}(\mathbf{y},\mathbf{y}')$, and since mutual information cannot be increased via signal processing, we have

$$I(\mathbf{A};\mathbf{X}) \leq I(\mathbf{Y};\mathbf{X}), \tag{6.60}$$

(see the data processing inequality in [9]). Thus, the third WI model assumption is ideal. The WI assumptions are physically unrealistic, but since they are information-theoretically ideal (in that relaxing any of them leads to a reduction in mutual information), they are analytically important.

6.3.3 Calculating capacity: A simple example

The following very simple and highly abstract example of a molecular communication channel is useful to illustrate the calculation of capacity.

Suppose the transmitter has *exactly one* molecule available, and can either keep or release the molecule. Also suppose the receiver can detect the arrival, or absence, of the single molecule. If a molecule is sent, it can arrive at the receiver with probability $1 - p_L$, or get lost (and never arrive) with probability p_L. There are no other molecules in the channel. Then:

- For channel inputs, the value x counts the number of molecules released by the transmitter: $x = 0$ if the transmitter *keeps* the molecule, and $x = 1$ if the transmitter *releases* the molecule. If $x \in \mathcal{X}$ represents the channel input, then $\mathcal{X} = \{0,1\}$.
- The receiver counts the number of arriving molecules: $y = 0$ if no molecules arrive, and $y = 1$ if a molecule arrives. If $y \in \mathcal{Y}$ represents the channel output, then $\mathcal{Y} = \{0,1\}$.

- If the transmitter keeps the molecule ($x = 0$), then no molecules arrive at the receiver ($y = 0$).
- If the transmitter releases a molecule ($x = 0$), then either the molecule gets lost ($y = 0$) with probability p_L, or it successfully arrives ($y = 1$) with probability $1 - p_L$.

Given the above information, we can write

$$p_{Y|X}(y|x) = \begin{cases} 1, & y = 0, x = 0 \\ 0, & y = 1, x = 0 \\ p_L, & y = 0, x = 1 \\ 1 - p_L, & y = 1, x = 1 \end{cases} \tag{6.61}$$

If p_L is small, the message at the output is a reasonable facsimile of the message at the input. For instance,

$$\begin{aligned} \mathbf{x} &= 0\,1\,0\,0\,1\,1\,0\,1\,1\,0 \\ \mathbf{y} &= 0\,1\,0\,0\,1\,0\,0\,1\,1\,0 \end{aligned} \tag{6.62}$$

is a possible pair of input sequence $\mathbf{x}$ and output sequence $\mathbf{y}$, with an error in the sixth position. Note that, unlike the delay-selector channel, there are no rearrangements of the inputs in this model, only flips from input 1 to input 0. A similar channel model was used in [11].

This is a simplistic model for molecular communication: note that the pair of processes x and y in (6.63) require that any "lost" molecules do not arrive later. However, channels with $f_{Y|X}(y|x)$ of the form in (6.61) are known as *Z channels*, a class of well-studied channels in the information- and communication-theoretic literature (see [9]).

How do we obtain p_L? Using the WI model and the first arrival time distribution, $f_N(n)$, we obtain

$$p_L = 1 - \int_{\tau}^{T} f_N(n - \tau) dn, \tag{6.63}$$

where τ is the molecule release time, and where $f_N(n)$ has the Lévy distribution, given in Chapter 5, which takes the parameter σ^2. For instance, setting $d = \sigma^2 = 1$, and setting $T = 120$ s with $\tau = 0$, we have $p_L = 0.073$.

We now use the channel model to calculate capacity. From (6.61),

$$H(Y|X) = \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p_{X,Y}(x,y) \log_2 \frac{1}{p_{Y|X}(y|x)} \tag{6.64}$$

$$= \sum_{x \in \{0,1\}} p(x) \sum_{y \in \{0,1\}} p_{Y|X}(y|x) \log_2 \frac{1}{p_{Y|X}(y|x)} \tag{6.65}$$

$$= p_X(0) \cdot 0 + p_X(1) \left(p_L \log_2 \frac{1}{p_L} + (1 - p_L) \log_2 \frac{1}{1 - p_L} \right) \tag{6.66}$$

$$= p_X(1) \left(p_L \log_2 \frac{1}{p_L} + (1 - p_L) \log_2 \frac{1}{1 - p_L} \right). \tag{6.67}$$

Define

$$\mathcal{H}(p_L) = p_L \log_2 \frac{1}{p_L} + (1-p_L)\log_2 \frac{1}{1-p_L}, \tag{6.68}$$

often called the *binary entropy function of* p_L. Then we can rewrite (6.67) as

$$H(Y|X) = p_X(1)\mathcal{H}(p_L). \tag{6.69}$$

We can find $p_Y(y)$, which is

$$p_Y(y) = \sum_{x\in\{0,1\}} p_{Y|X}(y|x)p_X(x) \tag{6.70}$$

$$= \begin{cases} p_X(1)p_L + p_X(0), & y = 0 \\ p_X(1)(1-p_L), & y = 1 \end{cases}. \tag{6.71}$$

Since $p_X(0) = 1 - p_X(1)$, note that $p_X(1)p_L + p_X(0) = 1 - p_X(1)(1-p_L)$. Thus,

$$H(Y) = \mathcal{H}\left(p_X(1)(1-p_L)\right). \tag{6.72}$$

Then the mutual information is given by

$$I(X;Y) = \mathcal{H}\left(p_X(1)(1-p_L)\right) - p_X(1)\mathcal{H}(p_L). \tag{6.73}$$

To find capacity C, the expression in (6.73) is maximized with respect to $p_X(1)$. It is shown in [12] that the maximizing value is

$$p_X(1) = \frac{(1-p_L)^{\frac{1-p_L}{p_L}}}{1 + p_L(1-p_L)^{\frac{1-p_L}{p_L}}}, \tag{6.74}$$

and substituting back into (6.73), we have

$$C = \mathcal{H}\left(\frac{(1-p_L)^{\frac{1}{p_L}}}{1 + p_L(1-p_L)^{\frac{1-p_L}{p_L}}}\right) - \frac{\mathcal{H}(p_L)(1-p_L)^{\frac{1-p_L}{p_L}}}{1 + p_L(1-p_L)^{\frac{1-p_L}{p_L}}}. \tag{6.75}$$

Even in this simple example, we can see that the expression for capacity is complicated. In many cases, closed-form capacity results are unknown, and the problem is solved numerically. As we mentioned in the introduction, this is why analysis is focused on the simple cases, though even there, capacity is not always known: for instance, the capacity is unknown for the Blackwell billiard-ball channel that we discussed in the introduction.

As a closing remark in this section, we have measured information rate in terms of the number of bits per unit time, but it is worth noting that both time and molecules are valuable resources. For example, suppose we are using intervals of length T seconds, and suppose the transmitter releases a (single) molecule in each interval with probability p_X. Then since the transmitter emits a molecule every T/p_X seconds intervals on average, we can divide the information rate (in bits per second) by T/p_X to obtain bits per molecule. Thus, capacity is expressible in terms of both bits per second and bits per molecule.

6.3.4 Towards the general problem

Here we start to formalize the general information-theoretic problem of molecular communication. Our first statement of the problem is as follows:

- For a single species of molecule, under the standard assumptions and the Wiener-Ideal model, what is the capacity of the system in terms of bits per second, and in terms of bits per molecule?

As we will see, this is not a useful way to state the problem, because the answer (both in bits per second and bits per molecule) is infinite!

Throughout this section, we make two assumptions on first arrival time $f_N(n)$. First, we assume that

$$\lim_{\eta \to \infty} \int_{n=0}^{\eta} f_N(n)\, dn = 1, \tag{6.76}$$

which implies that the particle never gets lost (it always arrives, eventually). Second, for any $\eta > 0$, we assume that

$$\int_{n=0}^{\eta} f_N(n)\, dn > 0, \tag{6.77}$$

which implies that, for any positive time, there is nonzero probability of the molecule arriving up to that time. These assumptions are true of many useful first arrival time distributions (and can be relaxed without changing anything significant), but they simplify the discussion. Moreover, throughout this section, we use the following setup:

- There are k molecules, and the communication session has length τ.
- The session length τ is broken into equal intervals of length T, and the k molecules are released from the transmitter at the beginning of an interval, chosen uniformly at random from all possible intervals.
- The receiver observes the interval in which the *first* molecule arrives, and decides that the transmitter released the molecules at the beginning of *the same interval*.

For instance, say $\tau = 10$ and $T = 1$; then the intervals are all 1 time unit long: $\{[0,1],[1,2],\ldots,[9,10]\}$, and if the receiver observes the first molecular arrival, e.g., at time 5.3, it decides that all k molecules were released at time 5. Obviously the transmitter can send $\log_2(\tau/T)$ bits to the receiver, but errors occur if the first arrival time for all k molecules are larger than T. Moreover, this is not necessarily an optimal system, and it is not our goal to design one – merely to point out that any optimal system can do better than this.

We first consider bits per molecule: suppose the transmitter and receiver had an unlimited amount of time to wait, but just one molecule ($k = 1$) – how much information can be sent? (Nothing changes if $k > 1$, so long as the number of molecules is finite.) Let $T = \log_2 \tau$ (supposing, for convenience, that τ is an integer multiple of $\log_2 \tau$). Now let $\tau \to \infty$: then $T \to \infty$, and by the first assumption, the molecule must arrive within the interval time τ with probability 1. Thus, since errors occur only if the first arrival time is greater than τ, the probability of error is arbitrarily small. Further, it

can be easily checked (e.g., using l'Hôpital's rule) that

$$\lim_{\tau\to\infty} \log_2 \frac{\tau}{T} = \lim_{\tau\to\infty} \log_2 \frac{\tau}{\log_2 \tau} = \infty. \tag{6.78}$$

Thus, our setup is capable of sending an infinite number of bits of information, with arbitrarily small error probability, using a single molecule and unlimited time; so capacity must be infinite in terms of bits per molecule.

Second, suppose the transmitter and receiver have only one time unit ($\tau = 1$), but an unlimited number of molecules (disregarding any physical constraints) – how much information can be sent? (Again, we pick $\tau = 1$ for convenience, and any finite τ can be used.) The interval τ is divided into s intervals, so that $T = 1/s$. Recall that the probability of error, P_{err}, is the probability that the first arrival of any of the k molecules *does not* occur within time T of release. This is given by

$$P_{\text{err}} = (1 - F_N(T))^k = \left(1 - F_N\left(\frac{1}{s}\right)\right)^k \tag{6.79}$$

where $F_N(n)$ is the cumulative distribution function (cdf) of the first arrival time n. For any s, $F_N(1/s) > 0$ (by assumption), so

$$\lim_{k\to\infty} p_e = \lim_{k\to\infty} \left(1 - F_N\left(\frac{1}{s}\right)\right)^k = 0. \tag{6.80}$$

Since this is true for each s, it is true for the limit as $s \to \infty$, so our setup can send an infinite number of bits with arbitrarily small error probability; so the capacity is also infinite in terms of bits per second.

These examples have little practicality (for instance, it is obviously impossible to release an unlimited number of molecules at once). However, they do illustrate that constraints are needed to obtain meaningful answers. Thus, we restate the general information-theoretic problem of molecular communication as follows:

- For a single species of molecule, under the standard assumptions and the Wiener-Ideal model:
  1. What is the capacity of the system subject to realistic waiting time and/or molecular release constraints?
  2. What is the capacity of the system in terms of bits per second per molecule?

These questions are both open. Most existing results in the literature address the first question, by applying constraints on the system.

In the remainder of this section, we examine a few special cases that have analytical solutions or bounds.

6.3.5 Timing channels

We can analyze the capacity of timing channels by viewing them as an additive noise channel, at least when the molecules are distinguishable. It should be clear that the

capacity using distinguishable molecules is always higher than that of indistinguishable molecules: using distinguishable molecules, the receiver has the option to treat them as indistinguishable by ignoring the molecule types. Thus, this analysis provides a (very loose) upper bound on the capacity using indistinguishable molecules.

Start by writing mutual information as

$$I(X;Y) = H(Y) - H(Y|X). \tag{6.81}$$

The quantity $H(Y|X)$ is given by (6.48). However, since $y = x + n$,

$$f_{Y|X}(y|x) = f_N(y-x). \tag{6.82}$$

Therefore, letting $\mathcal{X}$ and $\mathcal{Y}$ each be the set of real numbers, we can write

$$H(Y|X) = \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} f_{X,Y}(x,y) \log_2 \frac{1}{f_N(y-x)} \tag{6.83}$$

$$= \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} f_X(x) f_N(y-x) \log_2 \frac{1}{f_N(y-x)}. \tag{6.84}$$

Making the substitution $n = y - x$, and letting $H(N)$ represent the entropy of the first arrival time distribution, we have

$$\int_{x\in\mathbb{R}} f_X(x) \int_{n\in\mathbb{R}} f_N(n) \log_2 \frac{1}{f_N(n)} \tag{6.85}$$

$$= \int_{x\in\mathbb{R}} f_X(x) H(N) \tag{6.86}$$

$$= H(N). \tag{6.87}$$

That is, in additive noise channels where the addition is over $\mathbb{R}$, $H(Y|X) = H(N)$, and so

$$I(X;Y) = H(Y) - H(N). \tag{6.88}$$

This is an important simplification. The goal is to find capacity, which now reduces to

$$C = -H(N) + \max_{f_X(x)} H(Y), \tag{6.89}$$

since $H(N)$ is independent of the input distribution $f_X(x)$; moreover, for some useful distributions of n, $H(N)$ is available in closed form. Generally, there are three approaches:

1. *Solve explicitly for capacity.* An explicit, closed-form solution for the maximizing input distribution $f_X(x)$ is the most desirable outcome. However, at the time of writing, no closed-form solutions for capacity are known to exist for distinguishable molecule timing channels. Nonetheless, techniques like the Blahut-Arimoto algorithm [13, 14] may be used to obtain a quantized, numerical approximation to the capacity-achieving distribution.

2. *Upper bound: maximum entropy subject to constraints.* Given constraints on the input distribution of y, it is often possible to write the maximum possible entropy for any distribution $f_Y(y)$ subject to that constraint. Let $\hat{H}(Y)$ represent this maximum entropy: then

$$C = -H(N) + \max_{f_X(x)} H(Y) \tag{6.90}$$

$$\leq -H(N) + \hat{H}(Y), \tag{6.91}$$

providing an upper bound on capacity. As one example, suppose X has a positivity constraint and a mean constraint: $x \geq 0$, and $E[X] \leq \mu_X$ (i.e., the transmitter is only willing to wait μ_X, on average, in order to send its message). Then since $y = x + n$, we have a constraint on the mean of y: $E[Y] \leq \mu_X + E[N]$. The maximum entropy distribution under these constraints is the exponential distribution, with entropy

$$\hat{H}(Y) = \log_2 [e(\mu_X + E[N])]. \tag{6.92}$$

Thus, in general, if $E[X] \leq \mu_X$, then

$$C \leq \log_2 [e(\mu_X + E[N])] - H(N). \tag{6.93}$$

Other constraints on $H(Y)$ lead to different bounds. However, this technique is *not* valid for a *peak constraint* on y; we discuss this case below.

3. *Lower bound: explicit solution for a given input distribution.* Since the entropy $H(Y)$ is explicitly dependent on $f_X(x)$, for the moment we abuse the notation and write $H_{f_X(x)}(Y)$. Then

$$C \geq -H(N) + H_{f_X(x)}(Y), \tag{6.94}$$

with equality if $f_X(x)$ happens to be the capacity-achieving input distribution. The idea is to find $f_X(x)$ that leads to a closed-form expression of $H(Y)$. (But referring back to the general problem, note that this is a lower bound on an upper bound, which might not offer much insight, except into the general location of the upper bound.)

We illustrate these techniques with the AIGN channel. First, $H(N)$ is available in (nearly) closed form:

$$H(N) = \frac{1}{\log_e 2} \left[\log_e \left(2K_{-1/2}(\lambda/\mu)\mu \right) + \frac{3}{2} \frac{\frac{\partial}{\partial \gamma} K_\gamma(\lambda/\mu)\big|_{\gamma=-1/2}}{K_{-1/2}(\lambda/\mu)} + \frac{\lambda}{2\mu} \frac{K_{1/2}(\lambda/\mu) + K_{-3/2}(\lambda/\mu)}{K_{-1/2}(\lambda/\mu)} \right], \tag{6.95}$$

where $K_\gamma(\cdot)$ is the order-γ modified Bessel function of the second kind, and $\log_e$ represents the natural logarithm. This expression is a special case of results found in [2]. When $f_N(n)$ is inverse Gaussian, for convenience we shorten the above expression to

$$H(N) = h_{\mathrm{IG}}(\lambda, \mu). \tag{6.96}$$

Capacity of the AIGN channel is an open problem, as we noted above. However, assuming a mean constraint of μ_X, we can find both an upper and lower bound in closed form.

For the upper bound, if n is inverse Gaussian with parameters (λ_N, μ_N), then $E[N] = \mu$; thus, $E[Y] \leq \mu_X + \mu_N$. From (6.93), we have

$$C \leq \log_2 [e(\mu_X + \mu_N)] - h_{\mathrm{IG}}(\lambda_N, \mu_N). \tag{6.97}$$

For the lower bound, the inverse Gaussian distribution has the following property:

- Let x and n be inverse Gaussian random variables with parameters (λ_X, μ_X) and (λ_N, μ_N), respectively, where
$$\frac{\lambda_X}{\mu_X^2} = \frac{\lambda_N}{\mu_N^2}. \tag{6.98}$$
Then $x + n$ is inverse Gaussian with parameters
$$\left(\frac{\lambda_X}{\mu_X^2}(\mu_X + \mu_N)^2, \mu_X + \mu_N \right). \tag{6.99}$$

The first arrival time n is inverse Gaussian. If x is also inverse Gaussian with appropriately selected parameters, then $y = x + n$ will also be inverse Gaussian, by the above property. We know that the mean of x must be μ_X, by the mean constraint; thus, from (6.98), $\lambda_X = \lambda_N \mu_X^2 / \mu_N^2$, and

$$H(Y) = h_{\mathrm{IG}} \left(\frac{\lambda_N}{\lambda_N^2}(\mu_X + \mu_N)^2, \mu_X + \mu_N \right), \tag{6.100}$$

so from (6.94),

$$C \geq h_{\mathrm{IG}} \left(\frac{\lambda_N}{\lambda_N^2}(\mu_X + \mu_N)^2, \mu_X + \mu_N \right) - h_{\mathrm{IG}}(\lambda_N, \mu_N). \tag{6.101}$$

These results appeared in [1]. Moreover, these expressions are among the few known analytic bounds for capacity in timing channels.

Suppose instead we constrained the *absolute* waiting time, rather than the average waiting time. For instance, say there was some maximum time T at which a decision had to be made, and that any molecules that arrived after T were considered lost. The above approach requires modification, since it is no longer true that $H(Y|X) = H(N)$. (To see why, note that $y \neq x + n$ if $x + n > T$.) To address this issue, we modify the first arrival time n and observation y as follows: let

$$n^* = \begin{cases} n, & n \leq T \\ \infty, & n > T, \end{cases} \tag{6.102}$$

and let $y^* = x + n^*$. Then $I(X; Y^*) = H(Y^*) - H(Y^*|X)$, and we have once again that $H(Y^*|X) = H(N^*)$. Finally, since y (with a peak value of T) can be obtained as a function

of y^* (where the *noise* has a peak value of T), from the data processing inequality (see [9]) it is true that $I(X;Y) \leq I(X;Y^*)$; that is, the modified channel has higher capacity than the true channel. If closed-form expressions or bounds on $H(N^*)$ can be found, then this technique (combined with others mentioned in this section) can be used to find an upper bound on capacity (see [15]). However, techniques to find good lower bounds on capacity for this case are, as yet, unknown.

6.4 Summary and conclusion

In this chapter, we examined the application of communication and information theory to molecular communication. There is a vast analytical toolbox from these theories that can be used to investigate molecular communication. Our results from this chapter include:

- Abstract physical layer models, such as the Wiener-Ideal model, can be created for molecular communication, and they are analogous to similar models in conventional communication systems.
- Maximum *a posteriori* detection and maximum likelihood estimation can be used in molecular communication, much as they are used in conventional communication; moreover, techniques such as the Viterbi algorithm and the sum-product algorithm can also be used.
- Using the abstract physical layer model, the capacity of molecular communication can be investigated. However, without any constraints, the capacity in bits per second and bits per molecule are both infinite. For future approaches to the general problem of molecular communication capacity, we suggest analyzing the molecular efficiency, in bits per second per molecule.

From the results presented in this section, it should be clear that the application of information theory to molecular communication is at a very early stage, and there are many important open problems for researchers to address. Moreover, the broadest possible question is of course open: given a physical channel, what is the highest achievable rate of molecular communication, and what kind of a system achieves it?

There are promising research questions in both theoretical and practical directions. In the theoretical direction, solving the general problem from earlier in this chapter requires new mathematical knowledge, especially in terms of the sorting channel from (6.4): the permanent remains an extremely challenging function to work with, in spite of recent progress. (It may be possible to mitigate the complexity by seeking symmetries in the input distribution, as in [16].) In the practical direction, it remains unclear how, or even whether, bio-nanomachines can be engineered to approach the capacity. In this direction, we can seek more direct inspiration from biology, and analyze the physical systems that microorganisms use to communicate. It is reasonable to believe that these systems are optimized to approach capacity (see, e.g., [17]). As indicated by the rough results described in this section, there appear to be ample open problems to keep researchers in this field busy for the foreseeable future.

References

[1] K. V. Srinivas, A. W. Eckford, and R. S. Adve, "Molecular communication in fluid media: The additive inverse Gaussian noise channel," *IEEE Transations on Information Theory*, vol. 58, no. 7, pp. 4678–4692, Jul 2012.

[2] R. S. Chhikara and J. L. Folks, *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Marcel Dekker, 1989.

[3] R. B. Bapat and M. I. Beg, "Order statistics for nonidentically distributed variables and permanents," *Sankhya (Ser. A)*, vol. 51, no. 1, pp. 79–93, 1989.

[4] P. O. Vontobel, "The Bethe permanent of a non-negative matrix," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1866–1901, 2013.

[5] L. Cui and A. W. Eckford, "The delay selector channel: Definition and capacity bounds," in *Proc. Canadian Workshop on Information Theory (CWIT)*, 2011.

[6] G. D. Forney Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.

[7] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[8] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (2nd edn.). Wiley, 2006.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, pp. 379–423, Jul. 1948.

[11] B. Atakan and O. B. Akan, "An information theoretical approach for molecular communication," in *Proc. 2nd International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2007, pp. 33–40.

[12] S. Golomb, "The limiting behaviour of the z-channel," *IEEE Transations on Information Theory*, vol. 26, no. 3, p. 372, 1980.

[13] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[14] S. Arimoto, "An algorithm for computing the capacity of arbitrary memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[15] A. W. Eckford, K. V. Srinivas, and R. S. Adve, "The peak constrained additive inverse Gaussian noise channel," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2012.

[16] R. Song, C. Rose, Y.-L. Tsai, and I. S. Mian, "Wireless signalling with identical quanta," in *Proc. Wireless Communication and Networking Conference (WCNC)*, 2012.

[17] P. J. Thomas, "Every bit counts," *Science*, vol. 334, no. 6054, pp. 321–322, 2011.

7 Design and engineering of molecular communication systems

Recall the general model for a molecular communication system, presented in Figure 4.1 (Chapter 4), which is composed of *sender and receiver bio-nanomachines*, *information molecules*, and other molecules (or bio-nanomachines) that support communication between the sender and receiver bio-nanomachines such as *interface*, *guide*, *transport*, and *addressing molecules*. These components of molecular communication systems use biological or molecular machinery to implement specific functionalities:

- *Sender and receiver bio-nanomachines* encode and decode messages. At the sender bio-nanomachine, a molecular encoder transforms a message, such as the internal state of the bio-nanomachine or the external condition of the bio-nanomachine, into an appropriate signal. At the receiver bio-nanomachine, a molecular decoder transforms a signal into a useful state or action of the receiver bio-nanomachine.
- *Information molecules* function as signals to carry a message. An information molecule may occupy distinct states. If the encoder can set the state and the decoder can detect the state, the state can carry a message.
- *Interface molecules* encapsulate the signal and protect from noise in the environment during propagation. The signal can be passed out of the transmitting bio-nanomachine to an interface molecule during sending, and the signal can be passed from the interface molecule into the receiving bio-nanomachine during receiving.
- *Guide or transport molecules* can direct the signal from the transmitting bio-nanomachine to the receiving bio-nanomachine, in a way that is different from diffusion in free space.
- *Addressing molecules* can target the signal to a particular destination bio-nanomachine, and thus help to distinguish from among many possibilities in a network setting.

In this chapter, we examine how these components of molecular communication systems can be designed and engineered from biological materials. This chapter is divided into four sections based on the materials of interest: protein molecules, DNA molecules, liposomes, and biological cells. In each section, we describe design and engineering efforts by reviewing the state of the art of molecular communication research as well as relevant areas such as nanobiotechnology and synthetic biology. Note that each component that we describe in this chapter implements only a subset of the above functionalities.

7.1 Protein molecules

The first class of approaches to design and engineer bio-nanomachines utilizes proteins. A protein molecule is a linear polymer of amino acids that folds into a nanoscale functional structure (Section 2.1). Enzymes catalyze biochemical reactions to convert substrate molecules to product molecules and act as molecular encoders and decoders in biological systems. Ligand and receptor systems function as information molecules and decoders for the information molecules. Motor proteins can produce mechanical force and function as transport or guide molecules for molecular communication. In this section we look at the design and engineering of molecular communication components based on protein molecules.

7.1.1 Sender and receiver bio-nanomachines

Protein molecules such as enzymes function as molecular encoders and decoders for molecular communication. Enzymes transform input molecules (i.e., substrates) into output molecules (i.e., products) by catalyzing specific chemical reactions. At the sender, enzymes may encode a message into specific types of information molecules that propagate in the environment. At the receiver, enzymes detect the type of information molecule to decode the message. Encoding and decoding can be more complex when an allosteric enzyme is used. Aspartate transcarbamoylase (ATCase), for instance, has multiple binding sites and it is activated via the binding of its two substrates: carbamoyl phosphate (CP) and aspartate (Asp) [1]. The activated ATCase then catalyzes a reaction to produce N-carbamoyl aspartate. The activated ATCase is inactivated by an end product in the reaction pathway, cytidine triphosphate (CTP). Thus, the activity level of ATCase is regulated in a more complicated manner by the three molecules (CP, Asp, and CTP). Further, multiple types of enzymes can be integrated to implement Boolean logic operations such as AND, OR, and XOR [2]. In this case, multiple-input molecules can be combined and transformed into multiple-output molecules depending on the set of input molecules.

Enzymatic activities are often controlled by the concentration of input molecules (e.g., the rate of product formation increases with the concentration of the substrate molecule). Enzymatic activities are also regulated by the time-varying concentration of input molecules; for instance, calmodulin-dependent protein kinase II (CaM kinase II) responds to the frequency of Ca^{2+} spikes [3, 4]. Thus, when a number of information molecules are available, a message can be encoded into or decoded from the amplitude and frequency of oscillations of a single type of information molecule, similar to amplitude modulation (AM) and frequency modulation (FM) in radio broadcast. In addition, a message may be encoded into or decoded from both the amplitude and frequency of the concentration of a single type of information molecule. Figure 7.1A shows an example of a time-varying concentration of an input molecule and Figure 7.1B shows the response of two different protein decoders to the input. The two protein decoders show alternating activation and thus are functionally different decoders.

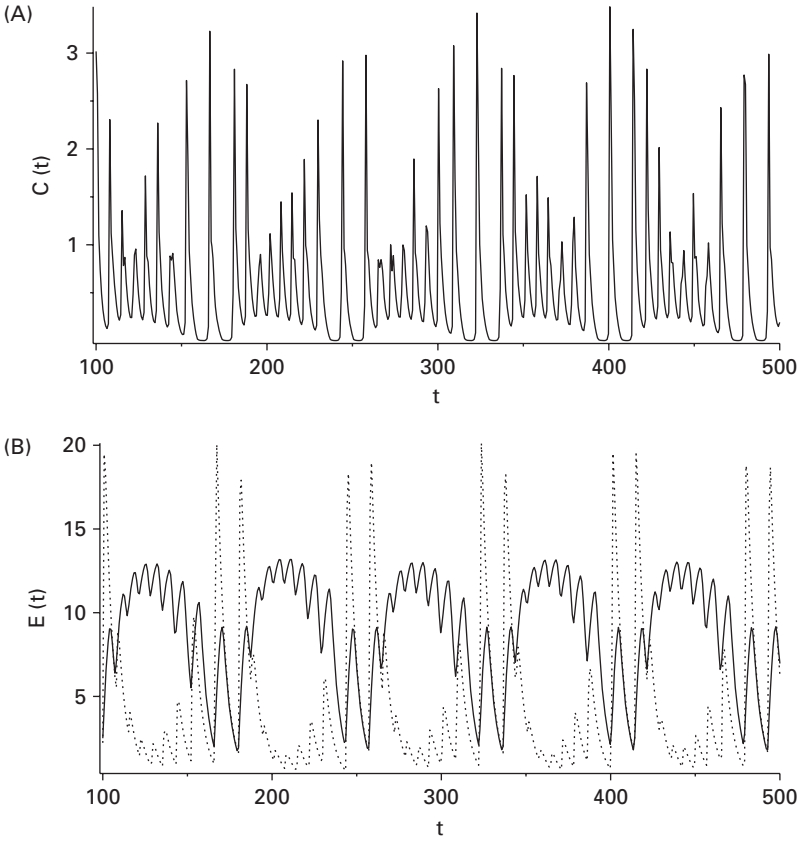


Figure 7.1 Protein decoders. (A) The concentration of the input molecule that activates enzymes. (B) The concentrations of two types of enzymes in active form, indicated by dotted and solid lines, respectively. For both types of enzymes the chemical kinetics is modeled as a sigmoidal function, and the time evolution is described as $\frac{dE(t)}{dt} = \frac{k_a C(t)^p}{K^p + C(t)^p} - k_i E(t)$, where $E(t)$ is the concentration of the active enzyme; $C(t)$ is the concentration of the input molecule at time t; k_a and k_i are the activation and inactivation rate constants; p is the Hill coefficient; and K is the concentration of the input molecule leading to the half maximum concentration of the active enzyme. Two different sets of parameters are used to simulate the activity levels of two different types of enzymes. (See [5] for detail.)

7.1.2 Information molecules

Protein molecules also function as information molecules for molecular communication. One important function and role of proteins in biological systems is their use as signaling molecules (Section 2.1). A great number of proteins available in biological systems act as signal molecules, and in addition, they can be synthesized artificially. A message can be encoded onto the structure and function of a protein molecule. Many proteins can be chemically modified through the addition of chemical groups such as phosphoryl and methyl groups, which alter their biological functions. Many proteins

thus take multiple states; e.g., phosphorylated and unphosphorylated states. If a different state produces different functionality at a receiver bio-nanomachine, it can carry a message. If a protein molecule can produce multiple functionally different states (n states) at a receiver, then the protein molecule carries up to $\log_2 n$ bits.

7.1.3 Guide and transport molecules

Protein molecules also function as transport and guide molecules for molecular communication. Motor proteins such as myosin and kinesin are capable of generating motion by hydrolyzing ATP and are responsible for transporting organelles and vesicles within a cell (Section 3.3.2). Filamentous proteins such as actin and tubulin self-assemble to form tracks of actin filaments or microtubules, along which motor proteins move to specific locations within a cell. These properties of motor and filamentous proteins can be exploited to implement transport and guide functionalities for molecular communication.

There are two design approaches for active transport mechanisms based on motor and filamentous proteins [6, 7]. In one design, filaments (e.g., microtubules) are fixed over a surface (e.g., a glass surface) to establish tracks, and motor proteins (e.g., kinesin) that may carry information molecules walk along the tracks (Figure 7.2A). Thus, motor proteins function as transport molecules and filamentous proteins as guide molecules. This is similar to the biological design of the active transport mechanism within a cell, and the materials available in a cell (e.g., motor proteins and adapter proteins) may be reused to selectively transport information molecules to specific locations. In this design, however, filamentous proteins need to be oriented correctly, and it is also difficult to control the motion of motor proteins in a complex network. In another design, the arrangement of motor and filamentous proteins is inverted; motor proteins are fixed on a surface to create a specific geometry and filamentous proteins attached with information molecules are propagated by the motor proteins according to the geometry (Figure 7.2B). Thus, filamentous proteins function as transport molecules and motor proteins as guide molecules in the inverted design. The inverted design does not require motor proteins to be oriented on a surface, since motor proteins have flexible tails to bind to filamentous proteins in the correct orientation, and it is easier to control the motion of filamentous proteins. We see specific examples of the two approaches in the following.

The first design approach (Figure 7.2A) was explored to create a network of filamentous proteins [8]. In the first design approach, as mentioned above, a network is created in a manner similar to the way a network is created within a biological cell. Sender bio-nanomachines are coated with microtubule nucleation sites (e.g., microtubule seeds) and receiver bio-nanomachines are coated with microtubule binding sites (e.g., caps such as GTP caps). When ATP is supplied, microtubules grow from a sender bio-nanomachine following dynamic instability in which the microtubules elongate and shrink stochastically via polymerization and depolymerization of tubulins (Figure 7.3A). A microtubule grows in a random direction, and upon contact with the binding site of a receiver bio-nanomachine, it becomes stabilized to form a link. The microtubules that are not

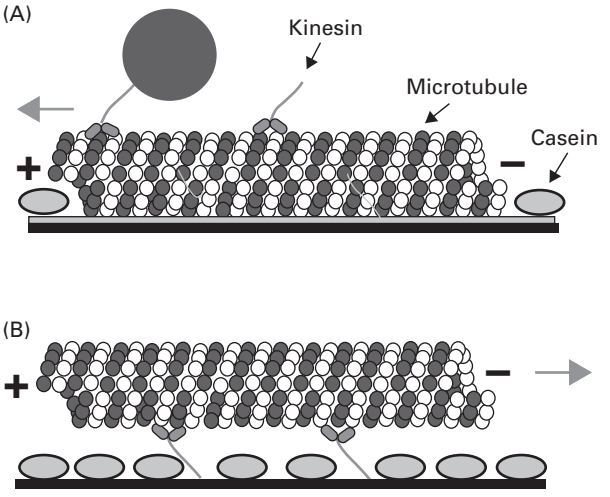


Figure 7.2 Two design approaches to protein-based active transport systems. (Casein reduces the denaturation of kinesin upon adsorption.) Adapted from [6].

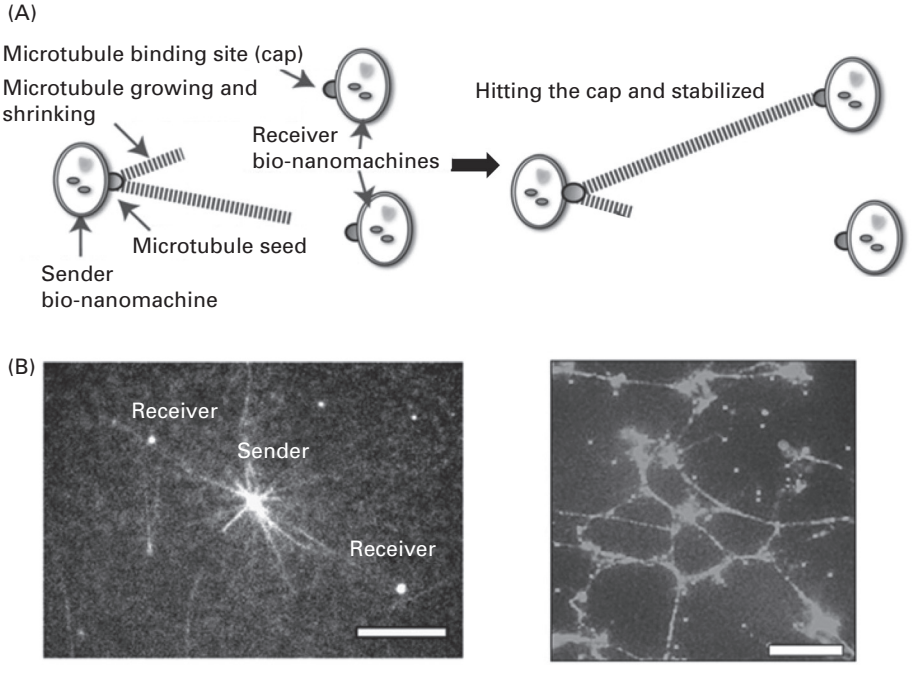


Figure 7.3 (A) Microtubule link formation processes. (B) Experimental results showing formation of (left) microtubule links between a sender bio-nanomachine and a receiver bio-nanomachine and (right) a microtubule network interconnecting multiple sender and receiver bio-nanomachines. Scale bar, 50 μm. Adapted from [8].

stabilized may continue to shrink and disappear while new microtubules may nucleate from the sender bio-nanomachine and elongate in a different direction. Through these dynamic processes, a microtubule network is formed to interconnect sender and receiver bio-nanomachines in the environment. Experiments were performed to demonstrate the formation of a microtubule link between a sender and a receiver bio-nanomachine (Figure 7.3B left) as well as that of a network (Figure 7.3B right). Experiments were also performed to demonstrate that kinesin motor proteins placed on a link move along the link. Furthermore, motor proteins not only move along links, but also exert force to move the links. The interplay between motor proteins and links leads to the dynamic organization of the network of bio-nanomachines. Types of topologies that can be formed are dependent on the environment and conditions (e.g., concentration of motors) and include asters and vortexes [9]. Once a network of bio-nanomachines is formed, any of the motors available in a cell may be used to transport molecules among bio-nanomachines on the network.

The second design approach (Figure 7.2B) was presented in, for instance, [6, 10, 11, 12]. Unlike the previous approach, a surface is coated with motor proteins and filaments are pushed by the motor proteins to propagate the filaments over the surface. This design is not naturally occurring and mechanisms are required for *loading*, *propagating*, and *unloading* of information molecules for molecular communication:

- *Loading and unloading:* One engineering approach is to use specific chemical links to bind information molecules to microtubules [6]. For instance, cargoes (i.e., vesicles containing information molecules) are coated with streptavidin, and microtubules are prepared to expose biotin on their surface. A streptavidin-coated cargo and biotinylated microtubule then bind through biotin-streptavidin linking to accomplish the loading process. For the unloading process, the chemical link between the cargo and microtubule can be broken by applying certain enzymes, changing the pH level of the environment, or illuminating UV light. The loading and unloading processes can also be accomplished by introducing DNA molecules as generic addressing molecules, as demonstrated in [11]. In this design, a microtubule attached to a single-stranded DNA (ssDNA) moves over a kinesin coated surface toward a loading site where several cargoes labeled with ssDNAs are placed (Figure 7.4A). When the microtubule passes the loading site, it selectively picks up a cargo with the ssDNA that is partially complementary to that of the microtubule through DNA hybridization (Figure 7.4B). The microtubule carrying the cargo continues to move and approaches unloading sites containing specific ssDNAs (Figure 7.4C). If the ssDNA at the unloading site is fully complementary to that of the cargo, the cargo is unloaded from the microtubule through a strand exchange (Figure 7.4D).
- *Propagation:* One engineering approach is to use surface chemistry to create a chemical track of motor proteins along which filaments move [6]. For instance, a surface can be prepared to produce a pattern of hydrophobic materials using surface chemistry. Motor proteins are then deposited and absorbed on the hydrophobic region based on the strong binding to form a track of motor proteins. A filament placed on a track of motor proteins moves along the track in a predetermined manner. Although

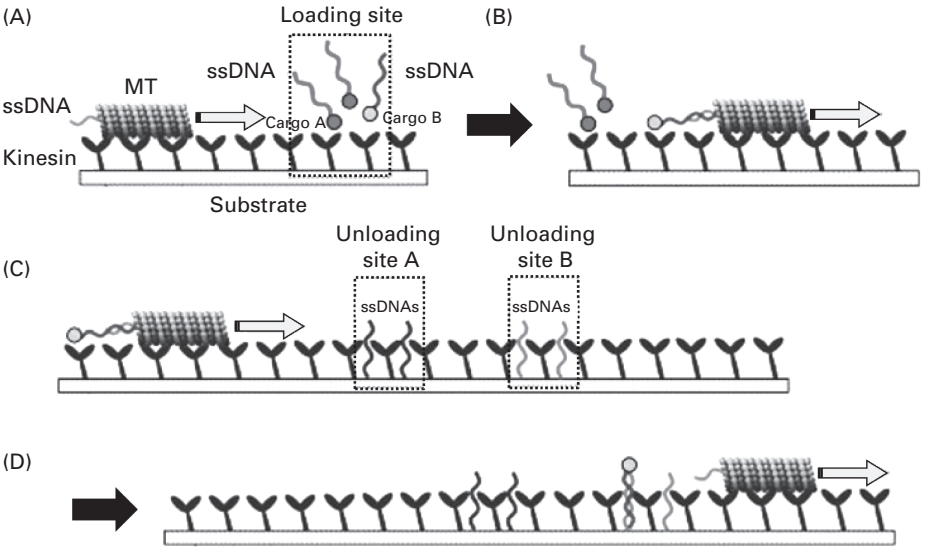


Figure 7.4 Loading, propagation, and unloading of information molecules using DNA molecules as addressing molecules. Adapted from [11].

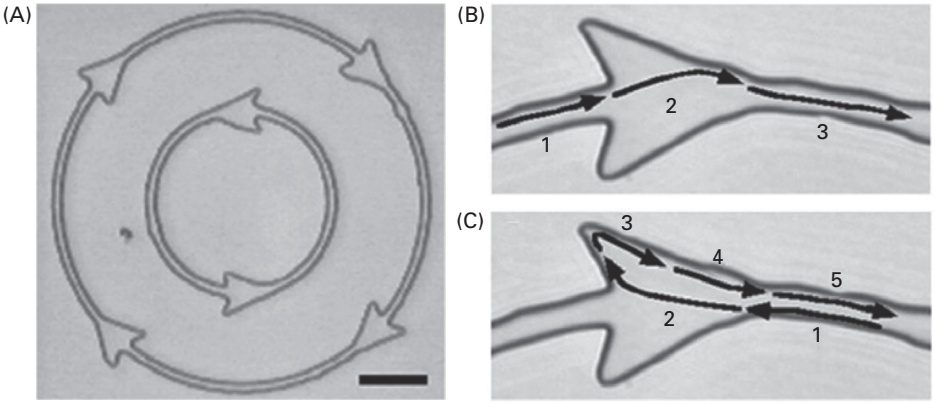


Figure 7.5 Rectifying the direction of a filament during active propagation on a patterned glass surface. Scale bar, 20 μm. Adapted from [12].

the tip of the filament is subject to random motion, the stiffness of the filament and binding to the surface cause the filament to continue in nearly the same direction. This method can be extended to achieve efficient propagation of filaments through micro-channels. As demonstrated in [12], an arrow-shaped micro-channel defined by walls can guide the direction of propagating filaments (Figure 7.5A). In the micro-channel, a filament moving in the direction of the arrow maintains its direction of movement (Figure 7.5B), while a filament moving against the direction of the arrow hits the

patterned walls and reverses its direction into the direction of the arrow (Figure 7.5C). All filaments in the micro-channel thus eventually propagate in one direction. Other methods are available to control the propagation of filaments, for instance, fluid flow, electric fields, or magnetic fields [6].

7.2 DNA molecules

The second class of approaches utilizes DNA molecules to design and engineer molecular communication components. A DNA molecule is made of a sequence of DNA bases (Section 2.2). Arbitrary DNA sequences can be readily prepared using a variety of enzymes such as DNA ligases to join two DNA sequences and restriction enzymes to cut a DNA molecule at specific sequences. The chemical features of DNA molecules are well understood; e.g., the base-pairing rule dictates that adenine (A) pairs with thymine (T) and guanine (G) with cytosine (C), and these features are exploited to construct a variety of nanoscale structures in the area of structural DNA nanotechnology [13]. As we see in this section, DNA molecules are versatile materials for implementing some components of molecular communication systems.

7.2.1 Sender and receiver bio-nanomachines

DNA molecules can function as encoders or decoders at sender and receiver bio-nanomachines for molecular communication, since DNA molecules react to diverse types of input molecules (e.g., DNA and RNA molecules, protein molecules). The chemical reactions are specific, and thus, specific sequences recognize specific types of molecules. The specificity of DNA molecules is useful for implementing the encoding or decoding functionality. For instance, an encoder may be designed as a single-stranded DNA (ssDNA) that has a sticky end for recognizing input molecules, and an output DNA sequence locked in a hairpin structure. The hairpin structure results from two complementary regions in opposite directions, and the two complementary regions of the DNA sequence bind and fold the DNA into the hairpin structure. The sticky end is a specific ssDNA which recognizes input molecules. The input molecules bind to the sticky end of the DNA to form a double helix that can be recognized and cut by a restriction enzyme. Multiple types of input molecules can be detected through each type of input molecule sequentially binding, being cut, and exposing a new sticky end. Finally, the complementary regions may be cut to release the output ssDNA sequence. DNA-based encoders or decoders can be designed to transform input molecules to produce output molecules.

7.2.2 Information molecules

DNA molecules function as information molecules for molecular communication. Following how DNA molecules are used to carry genetic information in biological systems, a message in molecular communication can be represented by a sequence of DNA bases. A DNA base is any one of the four units (i.e., A, T, G, or C), thus, a sequence of N DNA

bases can carry a maximum of $\log_2 4^N = 2N$ bits. The maximum is achieved when the receiver bio-nanomachine is able to distinguish all possible sequences. If a DNA sequence consists of 4.6 million base pairs, the amount of information it can carry is up to 9.2 megabits. The same length DNA molecule is found in an *E. coli* bacterium, which fits on an area of about 2 μm^2, and the DNA molecule encodes complex mechanisms necessary for its life. This information density is extremely high, given that in the same size area the silicon technology expected by 2014 can store only 490 bits or perform simple computation with 3 logic units [14].

A DNA-based information molecule can be designed to carry channel codes for error-handling purposes in molecular communication. An error in a DNA sequence may occur during the transmission, propagation, and reception processes of molecular communication. Errors may modify the original sequence and induce unintended reactions at receiver bio-nanomachines. In coding theory, an (n,k) block code transforms k bits of information into an n bit code with $(n-k)$ additional bits appended to the original k bits of information. If a $(3,1)$ repetition error-code is applied to a DNA sequence of *ACTG*, for instance, each DNA base is added with two repetitive DNA bases of itself, resulting in *AAACCCTTTGGG*. If a single DNA base in the sequence undergoes a substitution error during molecular communication processes, then the original sequence can be recovered by the receiver bio-nanomachine using a majority logic algorithm. Such an error-handling mechanism is thought to exist in gene translation and transcription processes in a biological cell [15].

DNA molecules can also be used as addressing molecules. In the active transport system presented in Figure 7.4, DNA sequences are used to determine the destination of information molecules in an engineered active transport system. As demonstrated in this system, an address, implemented by a DNA sequence, can be attached to an information molecule to allow the information molecule to be delivered to a specific location in the environment. A DNA-based information molecule may consist of two parts: address and payload (information). The address is implemented by the sticky end of the DNA sequence (i.e., an exposed ssDNA sequence). The payload is also implemented by a DNA sequence and may be protected in a hairpin structure. The receiver bio-nanomachine has the sticky end that is complementary to the sticky end of the information molecule, and thus, the two sticky ends form a double-helix when they chemically interact. Then, restriction enzymes in the receiver bio-nanomachine identify and remove the double-helix containing the address. The payload is then exposed from the hairpin structure and can react with the receiver bio-nanomachine.

7.2.3 Interface molecules

DNA molecules also function as interface molecules for molecular communication. DNA molecules can self-assemble into different structures that encapsulate information molecules to function as interface molecules. Information molecules may be encapsulated in the structure of a DNA-based interface molecule to modulate the functions of receiver bio-nanomachines. The DNA-based interface molecules protect the information molecules from environmental noise and prevent the information molecules

from chemically reacting with other molecules in the environment. To create structures using DNA molecules, the DNA nanotechnology exploits the sticky ends of DNA molecules (i.e., exposed ssDNA sequences). A DNA molecule can be prepared to have multiple sticky ends. Two DNA molecules with complementary sticky ends can join to form a larger combined DNA molecule. The remaining sticky ends of the combined DNA molecule may further allow the DNA molecule to bind to another DNA molecule to form an even larger structure. In this way a number of DNA molecules can repeatedly bind to form a large self-assembled structure. As an example, a two-dimensional DNA crystal is assembled from the DNA double crossover (DX) [16]. A DX consists of two DNA double helices fused to each other with parallel helix axes, forming a tile-like planar structure of about a few nanometers. A DX has four sticky ends that determine the orientations and associations with other DXs. When appropriately designed, DX molecules of about a few nanometers can self-assemble to form large-scale two-dimensional lattices about a few micrometers in length.

In another work [17], three-dimensional rhombohedral crystals have been constructed. To construct a three-dimensional structure, a three-dimensional building block is prepared from DNA molecules. The tensegrity triangle used in the construction of rhombohedral crystals contains three helical domains, each including two sticky ends, that are oriented to specific directions. It was demonstrated that the tensegrity triangles can self-assemble into rhombohedral crystals of about 250 μm in each dimension.

7.2.4 Guide and transport molecules

DNA molecules also function as guide and transport molecules for molecular communication. Several designs of DNA-based guide and transport systems, called DNA walkers in the literature, have been proposed and demonstrated [18, 19, 20, 21]. Two basic components are used: a DNA walker and a DNA track to guide the movement of the DNA walker. The DNA walker is designed with specific DNA sequences that induce motion and propagate along a DNA track that is also prepared with specific DNA sequences. This is analogous to how motor proteins interact with filamentous proteins in an engineered active transport system. The following describes the design of two types of DNA walker.

In one design [18], the movement of a DNA walker is controlled with sequentially introduced fuel DNA sequences. As shown in Figure 7.6A, the DNA walker has two legs, W1 and W2, and the DNA track exposes three branches, T1, T2, and T3. Each of the two legs and three branches exposes an ssDNA sequence containing specific DNA bases. The DNA walker progresses along the branches from T1 to T2 and finally to T3 using the following steps. First, an ssDNA sequence, A1, which is partially complementary to both W1 and T1, is added to the environment. A1 then forms a bridge to bind W1 to T1 (Figure 7.6B). Second, an ssDNA sequence, A2, which is partially complementary to both W2 and T2, is added; it then forms a bridge to bind W2 to T2 (Figure 7.6C). Third, an ssDNA sequence, D1, which is perfectly complementary to A1, is added, which then detaches W1 from T1, removing leg W1 from the DNA track

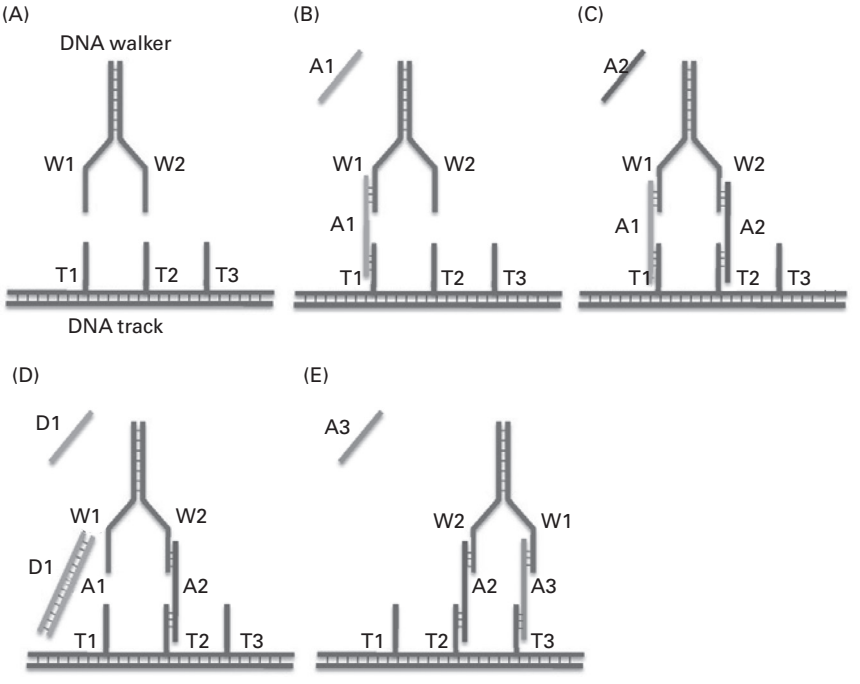


Figure 7.6 DNA walker that moves along the DNA track. Adapted from [22].

(Figure 7.6D). Finally, an ssDNA sequence A3, which is partially complementary to W1 and T3, is added; it forms a bridge to bind W1 to T3 (Figure 7.6E).

In another design [19], a DNA walker is autonomous and can move without step-by-step external instructions. In this design, specific DNA ligases and restriction enzymes are present in the environment to power the walker movement. Initially, a DNA walker is attached to the tip of a branch, A. The free end of the DNA walker (i.e., a sticky end) is then ligated to the next branch B by a DNA ligase with adenosine triphosphate (ATP), which results in the DNA walker bridging to both branches A and B. A restriction enzyme in the environment then cuts and destroys the bridge in such a way that the DNA walker remains attached to branch B but is unable to attach to branch A again. As long as the DNA ligases, restriction enzymes, and ATP are present in the environment, the walker continues to propagate unidirectionally over the track in this way.

7.3 Liposomes

The third class of approaches utilizes liposomes to design and engineer bio-nanomachines and other key components of molecular communication systems. Liposomes are artificial vesicles assembled from lipid bilayers into a sphere with a diameter from 100 nm to tens of μm (Section 2.3) and are used in various applications. For instance, in the area of artificial life, liposomes are used to study the origin of life.

Macromolecules such as DNA, RNA, and enzymes are encapsulated in a liposome to reproduce cellular behavior from a simplified setting [23]. In the area of drug delivery, liposomes are used as carriers of drug molecules (e.g., drug-carrying liposomes are injected into and circulate via blood streams [24].) As we see in this section, liposomes can be used to implement different components of molecular communication systems.

7.3.1 Sender and receiver bio-nanomachines

Liposomes can be used to implement encoding and decoding functionalities for sender and receiver bio-nanomachines for molecular communication. In [25], a liposome was integrated with functional molecules to act as a molecular encoder or decoder for molecular communication. This liposome-based bio-nanomachine, called a nano-sensory device, is designed with a liposome made from peptide lipids. Synthetic receptors and enzymes are immobilized on the surface of the bio-nanomachine and allow the bio-nanomachine to recognize specific molecules in the environment, cause enzymatic reactions on their surface, and produce amplified output molecules. Briefly, in the absence of input molecules (S) (amine signals in experiments), the mediator (M) (Cu^{2+}) present in the environment binds to the enzyme (*L*-lactate dehydrogenase: LDH) sitting on the surface of a bio-nanomachine (Figure 7.7A). The enzyme in this case is inactive (in the off-state) and produces no output molecules. When the input molecules are present the synthetic receptor (R) (hydrophobized pyridoxal derivative) strongly binds to the mediator to form a complex with an input molecule and a mediator (R-S-M complex) (Figure 7.7B). The enzyme then becomes free from the mediator and catalyzes a specific chemical reaction to convert substrate molecules present in the environment to the output molecules. The liposome-based bio-nanomachine, therefore, is capable of transforming input molecules to output molecules.

Encoding or decoding processes can also be implemented within liposomes as demonstrated in areas other than molecular communication. In a liposome-based micro-reactor system demonstrated in [26], for example, liposomes (POPC liposomes) are treated with sodium cholate to make them permeable to substrate molecules (glucose-1-phosphate) for glycogen synthesis. The substrate molecules are added to the environment and diffuse across the semipermeable membrane of the liposome into

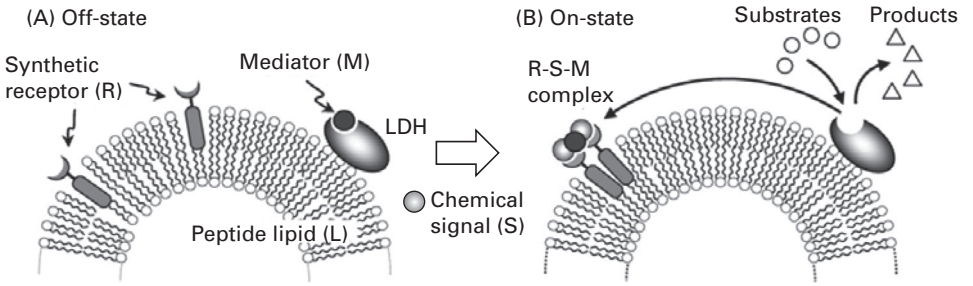


Figure 7.7 Liposome-based bio-nanomachine. Adapted from [25].

the liposome. The substrate molecules inside then react with the enzyme molecules (glycogen phosphorylase) that are trapped in the liposome to synthesize glycogen. In another work [27], for the study of the origin of life, a more complex chemical process is experimentally demonstrated within a liposome. A specific gene that encodes a pore-forming protein (α-hemolysin) is first extracted from a bacteria species *Staphylococcus aureus*. The gene is then entrapped within a liposome together with transcription and translation machinery, resulting in pore-forming proteins being expressed in the liposome, migrating to the surface of the liposome, and forming hemolysin pores on its surface. When materials for protein synthesis, such as amino acids and ATP, are added to the environment, these materials cross the hemolysin pores to enter the liposome and allow for the protein synthesis inside the liposome.

7.3.2 Interface molecules

Liposomes can function as interface molecules for molecular communication. The liposome-based interface was first proposed and designed in [28]. The designed liposome-based interface provides a generic mechanism to encapsulate and carry any type of information molecules that cannot pass through the membrane of the liposome. A sender bio-nanomachine produces vesicles containing information molecules through either fission, reproduction, or pore-formation (Figure 7.8A).

In fission, a liposome is budded from the membrane of the sender bio-nanomachine and information molecules inside the sender bio-nanomachine are encapsulated at the time of budding. In reproduction, a liposome is created through an internal

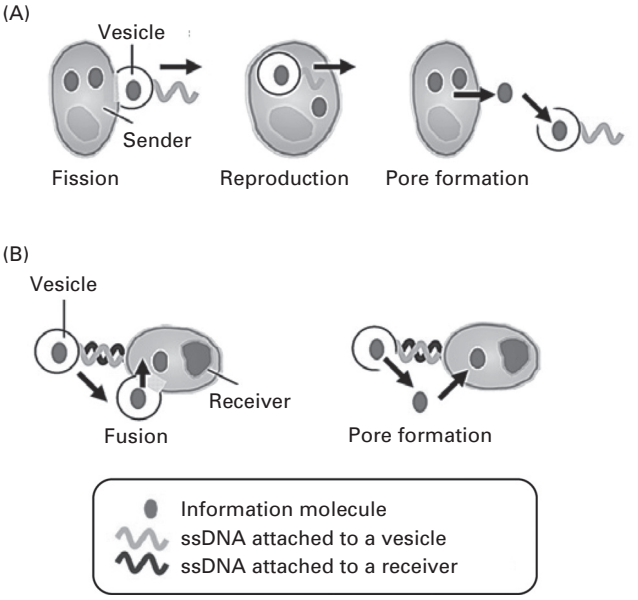


Figure 7.8 Several designs of interface molecules using vesicles. Adapted from [28].

mechanism of the sender bio-nanomachine and information molecules that are also inside are encapsulated. In pore-formation, liposomes prepared outside the sender bio-nanomachine form a pore with the sender bio-nanomachine, and information molecules inside the sender bio-nanomachine diffuse through the pore to the liposome. The receiver bio-nanomachine then receives the information molecules contained in a liposome through either fusion or pore-formation (Figure 7.8B). In fusion, the liposome is integrated into the membrane of the receiver bio-nanomachine, and the information molecules are released inside the receiver bio-nanomachine. In pore-formation, information molecules diffuse from the liposome to the receiver bio-nanomachine. Some aspects of the interface design, such as receiving through pore-formation, were experimentally investigated in [29]. Experimental results indicate that a liposome with pore-forming proteins embedded on its surface establishes pores with the pore-forming proteins expressed in a genetically modified human osteosarcoma cell, and transfers the peptides contained in the liposome through the pore to the cell.

7.3.3 Guide molecules

Liposomes can be prepared to form a network between liposome-based bio-nanomachines to implement guide functionality. Motor proteins and DNA walkers actively transport information molecules over guide molecules (Sections 7.1 and 7.2). An additional approach to propagating information molecules efficiently is to limit the physical space in which the information molecules can freely diffuse. For this, a sender bio-nanomachine may establish a physical channel with the receiver bio-nanomachine and release information molecules into the channel. Since the information molecules are contained in the channel and do not diffuse outside (e.g., an open environment), the information molecules reach the receiver bio-nanomachine in high concentration.

Such physical channels between liposome-based bio-nanomachines can be formed with lipid nanotubes [30]. In [31, 32], biological materials were selected to cause liposomes to form tubular projections to establish a liposome network. In one set of experiments, dioleoylphosphatidylcholine (DOPC) lipids were mixed with cholesterol to create a network of liposomes through lipid nanotubes. Experimental results demonstrated that micro-scale giant liposomes were assembled through natural swelling and, depending on the concentration of cholesterol, the liposomes were interconnected with lipid nanotubes with a diameter of 150 nm. In another work [33], originally spherical vesicles were deformed using a micro-injection technique to create a nanotube-vesicle network (NVN) (Figure 7.9). Fluorescent dyes (fluorescein) were then introduced into the vesicles (vesicles 1–3 in the figure), and the fluorescent dyes in selected vesicles (vesicles 2 and 3) were photo-bleached by laser illumination. The fluorescent intensity in those photo-bleached vesicles thus dropped immediately but increased as time progressed, indicating that fluorescent dyes propagated from vesicle 1 to vesicles 2 and 3. It was further demonstrated in experiments that an NVN can be made to connect vesicles filled with enzymes and vesicles with substrate molecules. They created different temporal and spatial patterns of product molecules by controlling the geometry of the NVN.

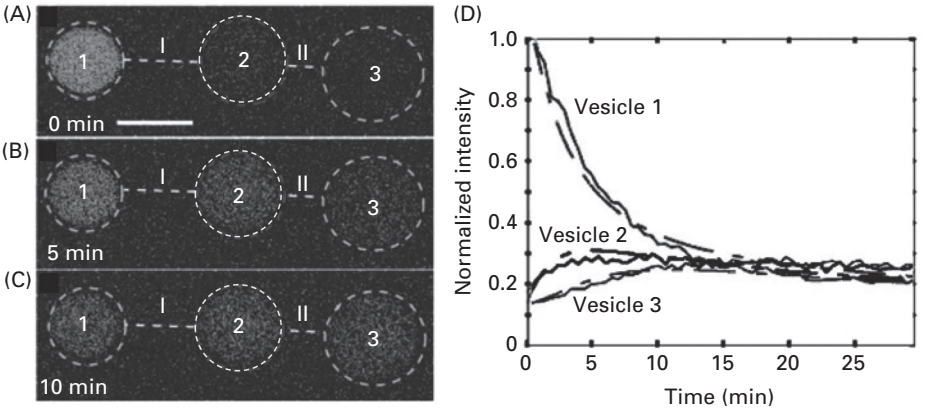


Figure 7.9 A liposome network made through lipid nanotubes. (A)(B)(C) Distribution of fluorescent dyes at time 0, 5, and 10 min, respectively, showing that the fluorescent dyes diffuse from vesicle 1 to vesicles 2 and 3. (D) Fluorescence intensity of vesicles 1, 2, and 3 (from the top in the graph). The solid lines show experimental data, and the dash-dotted lines the theoretical fit. Scale bar, 10 μm. Adapted from [33].

7.4 Biological cells

The last class of approaches exploits biological cells to design and engineer molecular communication systems. A biological cell is a system of bio-nanomachines with size typically in the range 1–10 μm in prokaryotic cells (e.g., bacteria) and 10–100 μm in eukaryotic cells (e.g., plant and animal cells) (Section 2.4). Biological cells are inherently capable of encoding and decoding chemical messages and thus are promising materials for implementing molecular communication systems.

7.4.1 Sender and receiver cells

Biological cells can be genetically modified to function as sender and receiver cells for molecular communication systems (Box 7.1). Biological cells can be engineered to transmit artificial information molecules, so that they function as senders, or to express artificial receptors for the information molecules, so that they function as receivers (Figure 7.10A). Several examples of engineered molecular communication systems are found in the literature, including:

- A cell density control system [34]. In this system, engineered bacteria communicate using the LuxR/LuxI system extracted from the marine bacterium *Vibrio fischeri*. The LuxI in the bacteria synthesizes a membrane-permeable molecule, acyl-homoserine lactone (AHL), which diffuses and accumulates in the environment. The AHL increases its concentration as the bacterial population increases and, at a high concentration, the AHL binds and activates LuxR in the bacteria to activate a killer gene, which causes programmed cell death. Therefore, this system is capable of maintaining a certain cell density through engineered molecular communication.

Box 7.1 Genetic engineering

Genetic engineering is the artificial transformation of the DNA of a cell and is widely used in many applications in biology, medicine, agriculture, and others. Briefly, genetic engineering is performed as follows. First, a DNA sequence that encodes a protein of interest, a gene, is isolated from biological cells or artificially synthesized using a DNA recombinant technology. The gene is then combined with elementary DNA sequences including a promoter and terminator region to produce a DNA construct. The promoter region of a DNA construct is recognized by an RNA polymerase and initiates the transcription of the gene, and the terminator region allows the RNA polymerase to end the transcription; the DNA sequence in between is transcribed by an RNA polymerase. The DNA construct also contains a marker gene that encodes an antibiotic resistance protein to identify the successful transformation of a cell. The DNA construct prepared this way is inserted into a cell through the process known as transformation. One method for transformation is to directly insert a DNA construct into the cell nucleus by micro-injection. Another method is lipofection in which a DNA construct is embedded into vesicles and successful merging of the vesicles with the membrane of a cell results in delivery of the DNA construct into the cell. Another method is electroporation in which an electronic shock is applied to a cell to cause its cell membrane to become permeable and the DNA construct enters the cell by diffusion across the permeable cell membrane. There are also viral vectors in which modified viruses use infection mechanisms to deliver a DNA construct into a cell. The efficiency of transformation varies depending on methods, cell types, and conditions. Selection then needs to be performed to identify and isolate successfully transformed cells from others. Those cells expressing the inserted gene are likely to express the marker gene that encodes the antibiotic resistance protein. Thus, only successfully transformed cells are likely to survive in media containing the corresponding antibiotic conditions. Normally, selected cells are further examined through other techniques such as Southern blotting and DNA sequencing to determine whether the inserted DNA sequence is indeed present in the cells.

- A pattern-forming system [35]. In this system, sender cells synthesize and diffuse the AHL, while receiver cells express green fluorescent protein (GFP) only when in a moderate AHL concentration. The concentration of the AHL is high around sender cells and decreases over distance from the sender cells. Only the receiver cells at some distance from the sender cells express the GFP. When sender cells are at one location in the environment and receiver cells are distributed throughout the environment, a ring-like pattern of GFP expression forms around the sender cells. (We see the detail of this system in Section 8.2.)
- An artificial molecular communication system [36]. Unlike the previous two systems based on engineered bacteria (prokaryotes), in this system, genetically engineered yeast cells (eukaryotes) communicate through the synthesis, transmission, and

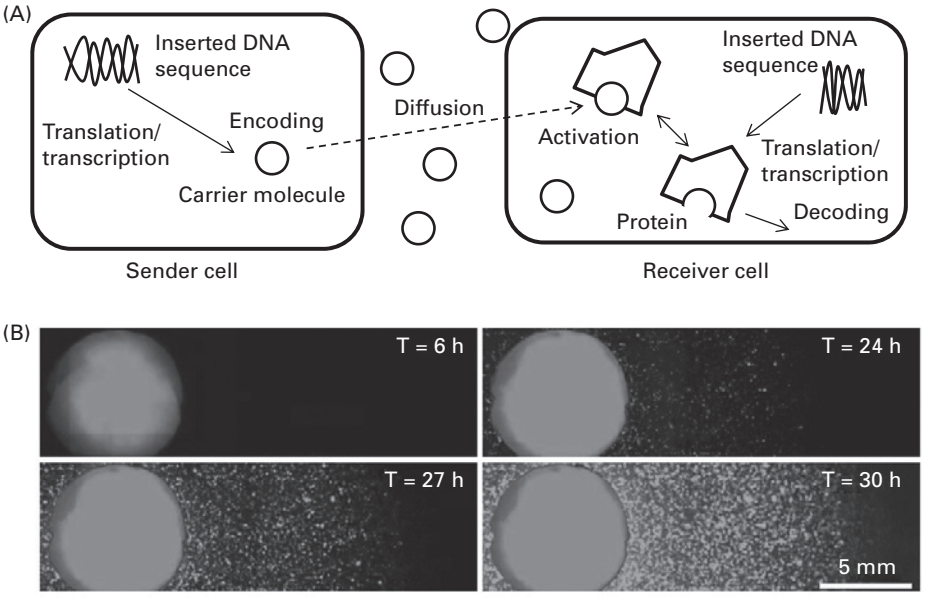


Figure 7.10 (A) A molecular communication system consisting of genetically engineered biological cells [22]. (B) Experimental results showing the location of engineered sender cells (large circles) that synthesize diffusive molecules and that of receiver cells (small dots) responding to the molecules. Adapted from [36].

reception of a plant hormone, cytokinin. Figure 7.10B shows a series of imaging with the location of sender cells that synthesize cytokinin and that of receiver cells responding to the cytokinin. The figure indicates that cytokinin diffuses in the environment to allow the sender cells and receiver cells to communicate.

In these systems, a message is encoded onto and decoded from the type and concentration of diffusive molecules (i.e., information molecules). These biological cells may be further engineered to include more complex encoding and decoding processes using several *modules* developed in synthetic biology [39, 40] such as:

- *Logic operations* [37]. The concentration of an input mRNA determines the concentration of an output mRNA to implement a logical gate. For instance, in the design of a logical NOT gate, an input mRNA is translated into the input repressor protein. In the absence of the input mRNA, an RNA polymerase recognizes the promoter region of a DNA sequence to perform transcription and produces an output mRNA (Figure 7.11A left). In the presence of the input mRNA, the repressor is expressed, binds to the promoter region, and blocks the binding of the RNA polymerase to the promoter region, thereby no output mRNA is produced (Figure 7.11A right). The gene network thus functions as a logical NOT gate since it produces the output mRNA only in the absence of the input mRNA. Logical NOT gates provide a basic building block to design a variety of gates. For instance, two types of logical NOT gates, if they produce the same output mRNA in response to different input mRNA, together

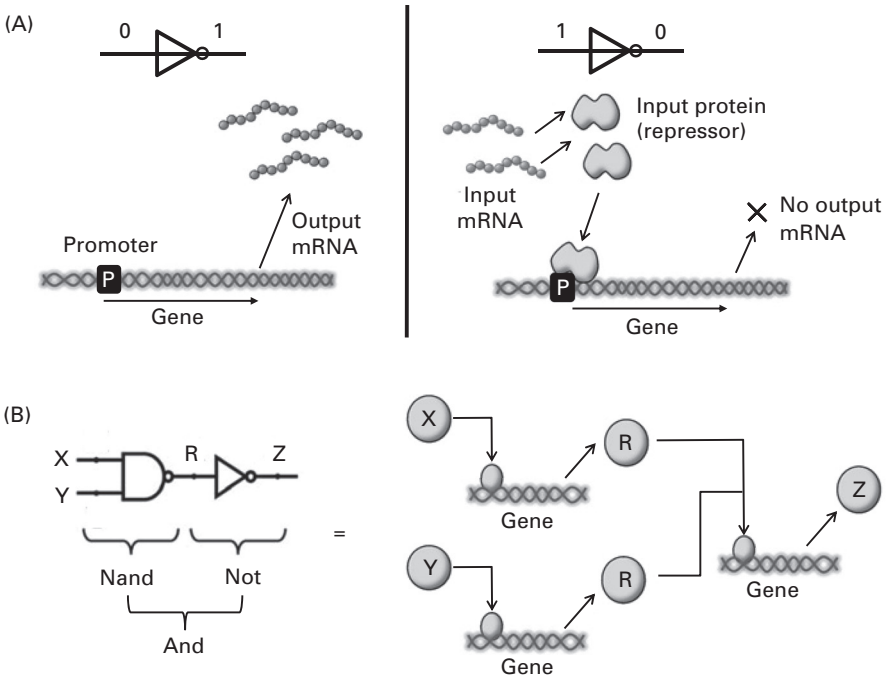


Figure 7.11 Design of logic circuits. (A) Logical NOT gate and (B) logical AND gate. Adapted from [37].

function as a NAND gate (Figure 7.11B). The logical NAND gate is then integrated with a logical NOT gate to implement a logical AND gate.

- *Toggle switches* that function as bi-stable biochemical switches [38]. A DNA sequence is prepared to contain two promoters, promoter 1 and promoter 2, that mutually repress each other through protein synthesis; each promoter transcribes the repressor of the other promoter (Figure 7.12A). The gene network achieves two stable states under appropriate settings (Figure 7.12B): one in which promoter 1 transcribes repressor 2, and the other in which promoter 2 transcribes repressor 1. This design also uses two inducers, inducer 1 and inducer 2, to switch from one state to the other. Inducer 1 (or 2) forces the expression of repressor 1 (or 2). A reporter gene encoding a fluorescent protein is also contained in the DNA sequence for observation and imaging purposes.
- *Oscillators* that cause three types of protein products to oscillate within a cell [41]. A DNA sequence is prepared to have three repressor genes *lacI*, *tetR*, and *cI* (Figure 7.13A). Each gene product represses the transcription of a different gene; the repressor protein LacI inhibits the transcription of *tetR*, TetR that of *cI*, and CI that of *lacI*. When *lacI* is being transcribed, *tetR* is blocked and *cI* starts being transcribed. As the concentration of CI increases, CI starts blocking the transcription of *lacI*. The concentrations of the three types of protein products LacI, TetR, and CI thus oscillate over time (Figure 7.13B–D).

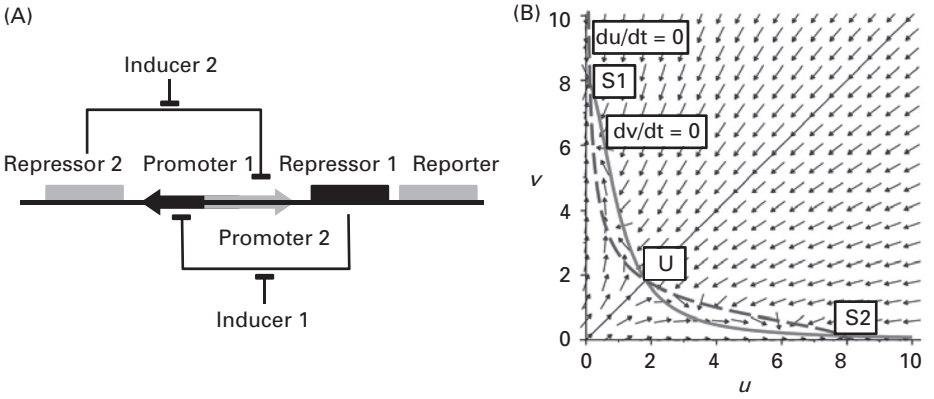


Figure 7.12 (A) Design of toggle switches [38]. (B) Bistable feature of toggle switches analyzed in a phase diagram. The interactions between the two repressors are described as $\frac{du}{dt} = \frac{\alpha_1}{1+v^{\beta_1}} - u$ and $\frac{dv}{dt} = \frac{\alpha_2}{1+u^{\beta_2}} - v$, where u and v are the concentrations of repressors 1 and 2; α_1 and α_2 are parameters that determine the effective synthesis rates of repressors 1 and 2; and β_1 and β_2 are parameters that depend on levels of repression of promoters 1 and 2. In the phase diagram, a velocity vector $(\frac{du}{dt}, \frac{dv}{dt})$ indicates how the system state (u, v) changes. The two lines drawn in the phase diagram are obtained from the equations above by setting $\frac{du}{dt} = 0$ and $\frac{dv}{dt} = 0$, and are called nullclines. The intersections of the nullclines (S_1, S_2, U in the diagram) indicate steady-state points where u and v do not move. The system at S_1 and S_2 is stable since the system returns to the original point under small perturbation (i.e., small changes in u and v), and the system at U is unstable since the system does not return to U under such perturbation. The phase diagram shows that a system starting above the $v = u$ line eventually converges to S_1 and starting below the $v = u$ line converges to S_2. ($\alpha_1 = \alpha_2 = 8, \beta_1 = \beta_2 = 2$.)

These modules can be incorporated in biological cells to implement encoding and decoding processes. For instance, logical operation modules may integrate multiple types of information molecules to produce output molecules in encoding and decoding. Toggle switches may function as a memory to implement a molecule counter (Section 5.4) for molecular communication. Oscillators may be used to control the timing of release at sender bio-nanomachines and that of reaction to information molecules at receiver bio-nanomachines to implement timing channels (Section 6.3.5).

This class of approaches may also use other cell modification techniques aside from genetic engineering to produce bio-nanomachines for molecular communication. Non-biological materials may be introduced into a cell to modify the functionality of the cell [42]. For instance, magnetic materials may be introduced into a biological cell through coating the surface of the magnetic materials with specific proteins that induce internalization into a cell. The magnetic materials in the cell may then function as an interface to an external device, by which one can control certain functions of the cell externally. An approach based on non-biological materials is useful to reduce the chance of interference with existing components in a cell.

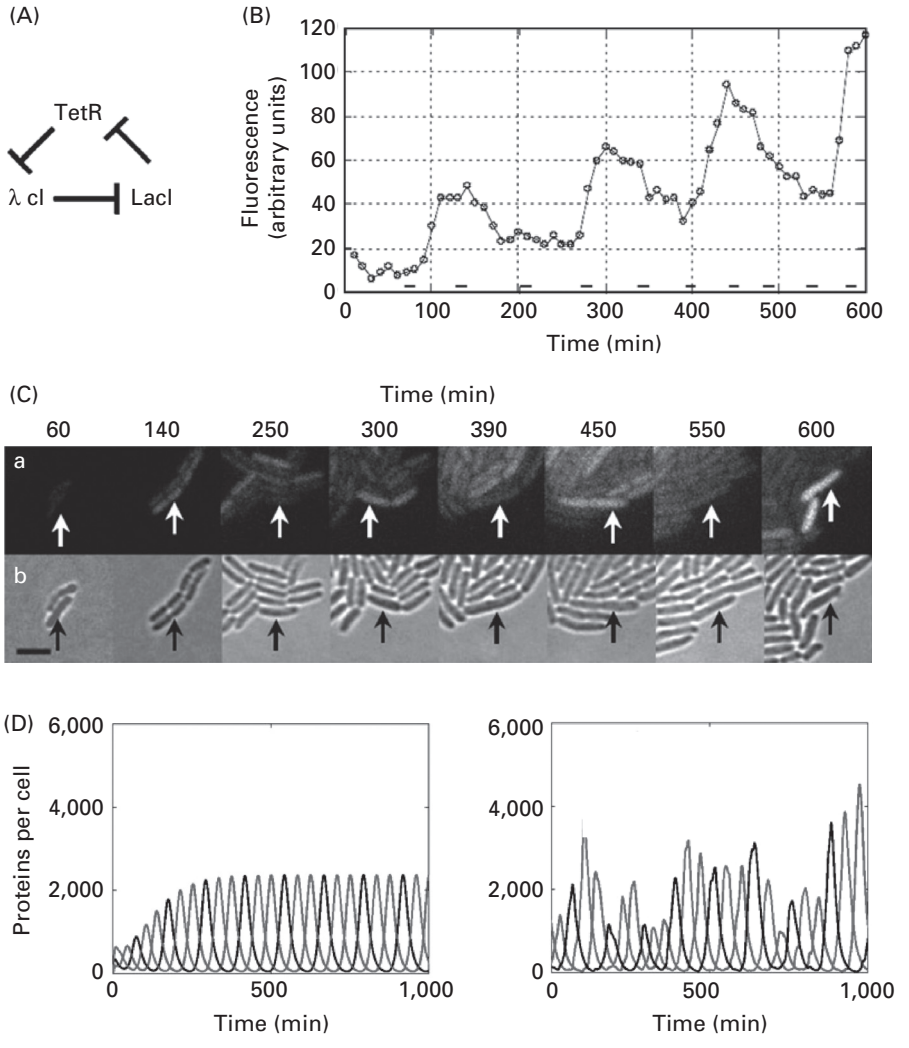


Figure 7.13 (A) Design of oscillators. (B)(C) Experimental results. Scale bar, 4 μm. (D) Simulation-based prediction of oscillatory behaviors. The dynamics of the concentrations of the three protein products are described as $\frac{dm_i}{dt} = -m_i + \frac{\alpha}{1+p_j^n} + \alpha_0$ and $\frac{dp_i}{dt} = -\beta(p_i - m_i)$, where m_i and p_i are the concentrations of mRNA and protein products synthesized from gene $i \in \{0, 1, 2\}$. Parameters α and n determine the rate of repression on gene i by the protein product synthesized from gene $j \equiv (i+1)(\text{mod}3)$. Parameter α_0 determines the basal expression rate of gene i. Parameter β determines the ratio of the protein decay rate to the mRNA decay rate. The ordinary differential equations (ODEs) can be numerically integrated to predict the concentration dynamics of the three protein products (left). A stochastic version of the ODE-based model may provide a better prediction since the number of molecules involved in gene transcription and translation is relatively small (right). Adapted from [41].

7.4.2 Guide cells

Functionally modified cells may implement guide functionality for molecular communication. Most animal cells in nature establish gap junction channels with adjacent cells to exchange and share cytosolic molecules such as ions, second messengers, and small metabolites (Section 3.3.3). For instance, a group of neurons in our body form a signaling circuit with gap junction channels that is capable of integrating various sensory inputs to activate the right set of motor outputs. And glial cells, non-neural cells in the brain, dynamically form and reconfigure a highly complicated network for storing information. Accordingly, bio-nanomachines may be designed to form gap junction channels and propagate information molecules for molecular communication. Furthermore, bio-nanomachines may molecularly communicate using a network of cells that performs networking functions such as switching and routing.

To use biological cellular networks for molecular communication, mechanisms are needed to form a network with biological cells. One engineering technique to achieve this is to develop a micro-platform for cell-patterning [44]. Biological cells placed on a micro-platform interact with molecules on the surface and either preferentially adhere and grow, or detach from the surface, and as a result, specific cellular patterns form on the surface. Figure 7.14A shows an example of this technique that is applied to align a group of cells into a straight line [43]. The biological cells in the line form gap junction channels with neighboring cells and propagate small molecules from cell to cell via the gap junction channels (Figure 7.14B). Micro-engineering techniques are also being developed for patterning cells in a cell-type specific manner or in three dimensions to create a more complex network topology. Another class of techniques to design a network with biological cells relies on the self-organization of biological cells that assemble into a network. This self-organization technique essentially represents developmental processes of a biological system. For engineered pattern formation, the biological cells may be genetically modified to control their

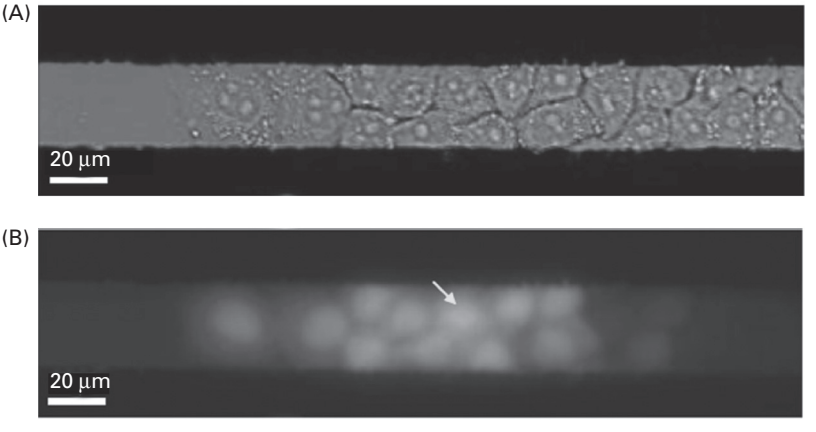


Figure 7.14 (A) Engineered cell wire. (B) Propagation of fluorescent molecules through the cell wire. Adapted from [43].

growth, migration, differentiation, and interactions. Theory for pattern formation is studied in amorphous computing [45] and experimental investigations are under way in developmental biology.

A network of cells can perform different functions that are useful for molecular communication. For instance, a computational unit can be designed from patterned cells connected through gap junction channels [46]. The computational unit has one computational cell, two input cells from where molecules diffuse into the computational cell, and one output cell to where molecules diffuse out. Computation occurs as chemical reactions occur that are dependent on the concentrations and types of molecules within the computational cell and from input cells. The result of the computation is concentrations of types of outgoing molecules that are yielded to the output cell.

A network of cells can also be used to implement other functionalities such as signal amplification. As demonstrated in [48], calcium-induced calcium release (CICR) is exploited to develop repeater cells that amplify calcium signals (Ca^{2+}). In non-excitable cells, Ca^{2+} is released via binding of both Ca^{2+} and inositol 1,4,5-trisphosphate (IP_3) to IP_3 receptors (IP_3Rs) located on the endoplasmic reticulum (ER) surface. An IP_3R has calcium-binding sites for activation to potentially enable CICR, yet such behavior is not commonly observed in non-excitable cells. However, IP_3 dictates the sensitivity of IP_3Rs to Ca^{2+}, and thus a CICR-like behavior may be artificially induced in non-excitable cells under the condition that the intracellular IP_3 concentrations are elevated. Figure 7.15A–B shows the evidence that a Ca^{2+} increase spreads from the centered cell to the neighboring cells under the condition that intracellular IP_3 levels are elevated. Upon stimulation, the centered cell increases the cytosolic Ca^{2+} concentration, and the increased Ca^{2+} concentration propagates to the neighboring cells and subsequently to their neighboring cells. A mathematical model can be used to analyze the impact of different parameters on the ability of the network of cells to amplify calcium signals (Figure 7.15C).

A network of cells may also be exploited to implement other functions such as filtering and switching [49]. Figure 7.16A illustrates one possible design of filtering and switching cells based on the selectivity and permeability of gap junction channels. The assumptions are: (1) cell A expresses Cx1 and Cx2; cell X expresses Cx1; and cell Y expresses Cx2, where Cxn denotes a specific connexin protein that constitutes one type of gap junction channel; (2) only the same type of connexin can form a channel; (3) signal molecules are able to permeate more easily through Cx1-type gap junction channels than through Cx2-type channels. Assume further that cells are patterned as shown in the figure. In this way, Cx1-type gap junction channels are formed between cells A and X and Cx2-type channels are formed between cells A and Y; and in this case, signal molecules incoming to cell A propagate preferentially to cell X. For dynamic switching, it is possible to apply an external signal to decrease the permeability of Cx1-type gap junction channels, and instead allow more signal molecules to propagate through Cx2-type gap junction channels (i.e., toward cell Y) (Figure 7.16B). Potential external signals for dynamic switching include those that activate a specific kinase inside cells, which leads to the phosphorylation of a specific type of gap junction-constituting protein (i.e., Cx1 in this case), and finally results in the decreased permeability of such channels.

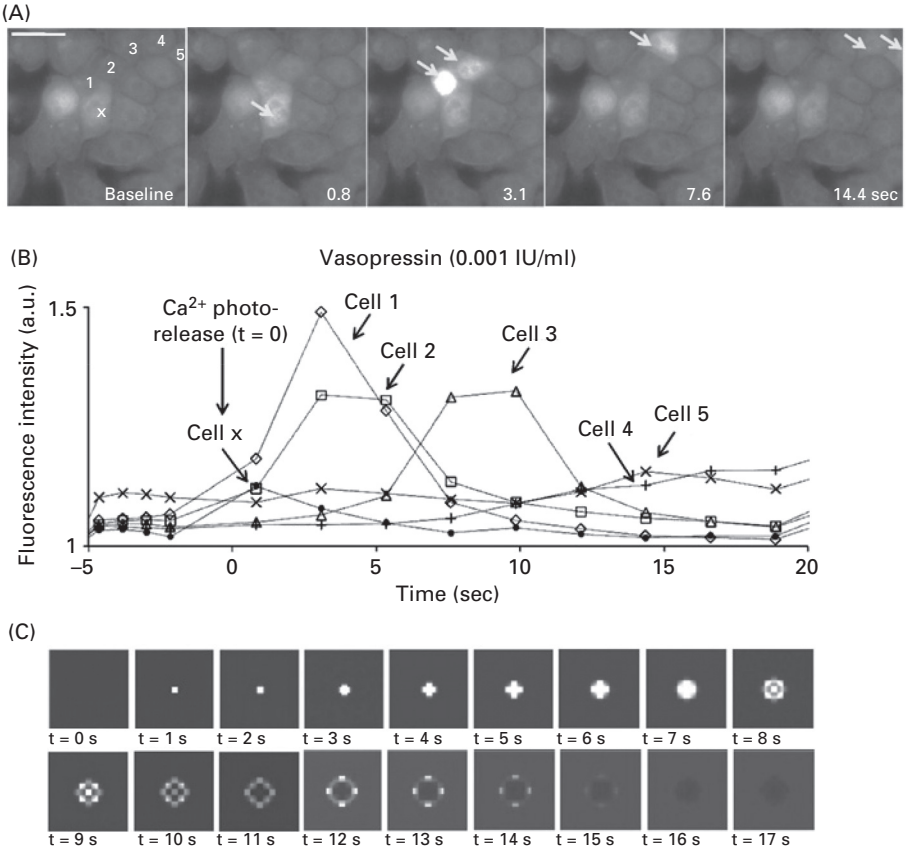


Figure 7.15 Repeater cells amplify and relay signal molecules. (A) Experimental results showing that non-excitable cells amplify and relay calcium signals under certain conditions. (B) Fluorescence intensity of cells shown in (A) indicating how the intracellular calcium concentration in each cell changes over time. (C) Simulation-based prediction and analysis. The two types of calcium dynamics, intracellular and intercellular, are considered. The intracellular calcium dynamics is controlled by the three calcium fluxes: the channel flux J_C that releases calcium from the internal calcium storage through calcium channels to the cytosol, the pump flux J_P that uptakes calcium from the cytosol and transports it into the internal calcium storage, and the leakage flux J_L that releases calcium from the calcium storage to the cytosol. The intracellular calcium dynamics is then described as $\frac{\partial c}{\partial t} = D\nabla^2 c + J_C - J_P + J_L$, where D is the diffusion coefficient of calcium in the cytosol. See [47] for more detail.

7.4.3 Transport cells

Engineered cells may also implement transport functionality for molecular communication. Motile cells, such as flagellated bacteria, actively transport molecules via chemotaxis (Section 3.3.5). A motile cell uses protein receptors to sense different types of molecules in the environment and uses flagella to move towards favorable conditions (i.e., attractants) or away from unfavorable conditions (i.e., repellents) based on chemical gradients. Motile cells, such as bacteria, can exchange molecules (e.g., genetic

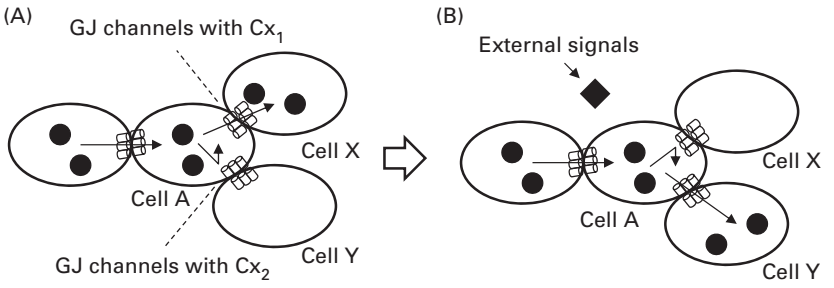


Figure 7.16 Filtering and switching with gap junction channels [22].

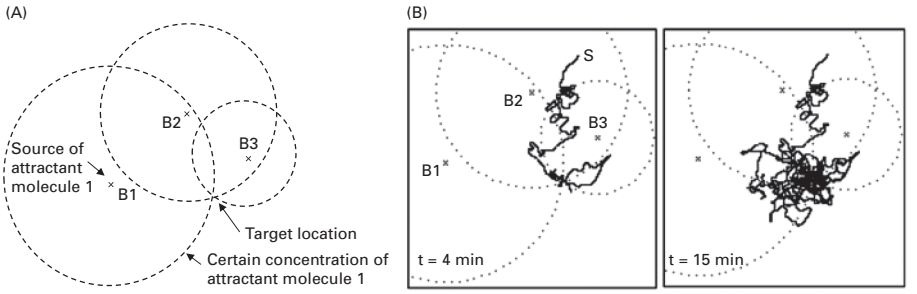


Figure 7.17 Motile cells perform addressing by a concentration profile [53].

materials) through direct cell-to-cell contact in the process called bacterial conjugation. Cell motility and its associated mechanisms provide a method to transport and exchange information molecules among bio-nanomachines in molecular communication.

In molecular communication using motile cells, sender bio-nanomachines inject information molecules into motile cells [50, 51]. Attractant or repellent molecules may be distributed in the environment to direct the movement of the motile cells toward a target location. Receiver bio-nanomachines at a target location then receive information molecules from the motile cells by, for example, exploiting the conjugation process. Motile cells might be engineered to many types of molecules in order to address target locations with a higher resolution [52]. For instance, the environment contains three different sources of attractant molecules (e.g., B1, B2, and B3 in Figure 7.17A), each releasing a different type of molecule. Each type of molecule forms a concentration that is highest around the source and decreases over distance from the source. A sender bio-nanomachine then uses the motile cell that moves toward a target location that has a specific concentration profile of the three types of attractant molecules; e.g., (x_a, x_b, x_c) where x_i $(i \in \{a, b, c\})$ indicates the target concentration of molecule i. At a given location, the motile cell measures the difference between the target concentration and current concentration (i.e., $|x_i - c_i|$ where c_i is the current concentration of molecule i). If the differences or the sum of the differences are increasing, the motile cell is moving away from the target location, and therefore tumbles more often to randomize its direction of movement. If the differences are decreasing, the motile cell is

moving towards the target location, and thus it continues its direction of movement. Over time, the motile carrier is likely to arrive at the target location specified by the concentration profile, which is expected to be the receiver location (Figure 7.17B).

Unlike the previously discussed active transport mechanisms by motor proteins or DNA walkers, motile cells do not require rails or platforms, but rather use gradients of attractant or repellent molecules as guide molecules. One approach to establish chemical gradients in the aqueous environment is to use techniques available in the area of microfluidics. In microfluidics, gradient devices have been developed to study biological processes related to chemotaxis, such as cell growth, differentiation, and death [56]. One type of gradient device uses laminar flow to generate gradient profiles. Each flow or fluid stream can consist of different chemical species and creates a chemical gradient perpendicular to its flow direction. The gradient profiles in this type of device can be controlled dynamically as well as maintained stably. Another type of gradient device relies on the free diffusion of molecules from a source to the sink via a microchannel,

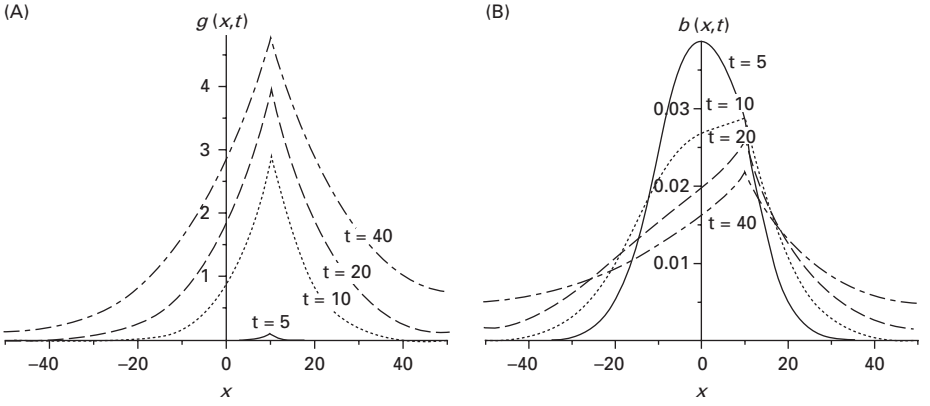


Figure 7.18 Cooperative molecule transport by motile cells [54]. The concentrations of the guide molecules (A) and motile cells (B). Here the concentrations of motile cells and guide molecules are modeled as $b(x,t)$ and $g(x,t)$ with respect to location x and time t, and their time evolution is given as $\frac{\partial b}{\partial t} = D_b \frac{\partial^2 b}{\partial x^2} - \frac{\partial}{\partial x}(\frac{Vg}{K+g} \frac{\partial g}{\partial x} b)$ and $\frac{\partial g}{\partial t} = D_g \frac{\partial^2 g}{\partial x^2} - kg + Rb\delta(x-d)$. The first term in each equation describes the diffusion of motile cells and guide molecules with the respective diffusion coefficients: D_b and D_g. In the first equation, the second term describes the biased propagation of the motile cells influenced by the concentration gradient of the guide molecules [55] where parameters V and K determine the chemotactic motility. In the second equation, the second term describes the degradation of guide molecules with a rate constant k, and the last term describes the release of guide molecules by the motile cells at the target location $x = d$ where R is the release rate of guide molecules per motile cell. Motile cells are injected at $x = 0$ and $t = 0$. They start diffusing randomly in the environment, since there are no guide molecules in the environment. Some motile cells then arrive at the target site ($d = 10$) and release guide molecules to form a concentration gradient of guide molecules in the environment (A). The concentration gradient of guide molecules attracts the motile cells to the target site. More motile cells arrive at the target site, and more guide molecules are released. Over time, many motile cells are concentrated around the target site (B). When no guide molecules are used, the concentration of motile cells becomes uniform in the environment.

and specific materials in the microchannel are used to control the gradient profile in the microchannel. For instance, the source and microchannel can be separated by semipermeable membranes with nano-to-micro scale pores, which release molecules from the source to the microchannel slowly to control the gradient profile in the microchannel. Alternatively, the microchannel may contain hydrogel (e.g., collagen and agarose) to reduce the rate of diffusion to control the gradient profile. The concentration gradients on these flow-free devices evolve over time by diffusion to reach a steady state. Compared to flow-based devices, flow-free devices allow the development of more complex patterns of chemical gradients, but make dynamic control of gradient profiles difficult.

Another approach to establish chemical gradients is for bio-nanomachines themselves to generate gradients by releasing guide molecules. Autonomous generation may be required for molecular communication systems in which external control is prohibitive. In addition to sender and receiver bio-nanomachines releasing guide molecules, motile cells may release guide molecules to guide other motile cells to target locations. For instance, a group of motile cells may be used to transport molecules to a target site that is not known to the motile cells in advance. The motile cells thus search at random. Those motile cells that successfully find the target site then start releasing guide molecules to generate a gradient. Other motile cells then move toward the target site following the concentration gradient of the guide molecules [54] (see the Figure 7.18 caption for more detail).

7.5 Conclusion and summary

In this chapter, we examined some design and engineering aspects of molecular communication systems. Promising materials include protein molecules, DNA molecules, liposomes, and biological cells as we saw in this chapter, but other materials may also be useful, such as carbon nanotubes (CNTs) [57] – the external walls of CNTs can be functionalized with biological materials (e.g., peptides and proteins) to implement functional molecules for drug delivery [58]. The choice of materials for the design and engineering of molecular communication systems depends on various factors, such as the compatibility with the environment encountered by applications, available energy sources in the environment, the cost and available methods of fabrication, autonomy and controllability of bio-nanomachines, and the communication-related performance (e.g., propagation speed) achieved by bio-nanomachines. In the next chapter, we look at specific molecular communication systems targeted to particular applications.

References

[1] D. Bray, "Protein molecules as computational elements in living cells," *Nature*, vol. 376, pp. 307–312, 2012.

[2] E. Katz and V. Privman, "Enzyme-based logic systems for information processing," *Chemical Society Reviews*, vol. 39, pp. 1835–1837, 2010.

[3] M. J. Berridge, "The AM and FM of calcium signalling," *Nature*, vol. 386, no. 6627, pp. 759–760, 1977.
[4] P. D. Koninck and H. Schulman, "Sensitivity of CaM kinase II to the frequency of Ca^{2+} oscillations," *Science*, vol. 279, pp. 227–230, 1998.
[5] A. Z. Larsen and U. Kummer, "Information processing in calcium signal transduction," *Lecture Notes in Physics*, vol. 623, pp. 153–178, 2011.
[6] H. Hess and V. Vogel, "Molecular shuttles based on motor proteins: active transport in synthetic environments," *Reviews in Molecular Biotechnology*, vol. 82, no. 1, pp. 67–85, 2001.
[7] A. Agarwal and H. Hess, "Molecular motors as components of future medical devices and engineered materials," *Journal of Nanotechnology in Engineering and Medicine*, vol. 1, no. 1, 2010.
[8] A. Enomoto, M. J. Moore, T. Suda, and K. Oiwa, "Design of self-organizing microtubule networks for molecular communication," *Nano Communication Networks*, vol. 2, no. 1, pp. 16–24, 2011.
[9] F. J. Nedelec, T. Surrey, A. C. Maggs, and S. Leibler, "Self-organization of microtubules and motors," *Nature*, vol. 389, pp. 305–308, 2011.
[10] H. Hess, C. M. Matzke, R. K. Doot, J. Clemmens, G. D. Bachand, B. C. Bunker, and V. Vogel, "Molecular shuttles operating undercover: a new photolithographic approach for the fabrication of structured surfaces supporting directed motility," *Nano Letters*, vol. 3, no. 12, pp. 1651–1655, 2003.
[11] S. Hiyama, T. Inoue, T. Shima, Y. Moritani, T. Suda, and K. Sutoh, "Autonomous loading, transport, and unloading of specified cargoes by using DNA hybridization and biological motor-based motility," *Small*, vol. 4, no. 4, pp. 410–415, 2008.
[12] Y. Hiratsuka, T. Tada, K. Oiwa, T. Kanayama, and T. Q. Uyeda, "Controlling the direction of kinesin-driven microtubule movements along microlithographic track," *Biophysical Journal*, vol. 81, pp. 1555–1561, 2001.
[13] N. C. Seeman, H. Wang, X. Yang, F. Liu, C. Mao, W. Sun, L. Wenzler, Z. Shen, R. Sha, H. Yan, M. H. Wong, P. S. Ardyen, B. Liu, H. Qiu, X. Li, J. Qi, S. M. Du, Y. Zhang, J. E. Mueller, T. J. Fu, Y. Wang, and J. Chen, "New motifs in DNA nanotechnology," *Nanotechnology*, vol. 9, pp. 257–273, 1988.
[14] M. L. Simpson, G. S. Sayler, J. T. Fleming, and B. Applegate, "Whole-cell biocomputing," *Trends in Biotechnology*, vol. 19, no. 8, pp. 317–323, 2001.
[15] E. E. May, M. A. Vouk, and D. L. Bitzer, "Classification of *Escherichia coli* k-12 ribosome binding sites," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 1, pp. 90–97, 2006.
[16] E. Winfree, F. Liu, L. A. Wenzler, and N. C. Seeman, "Design and self-assembly of two-dimensional DNA crystals," *Nature*, vol. 394, pp. 539–544, 1998.
[17] J. Zheng, J. J. Birktoft, Y. Chen, T. Wang, R. Sha, P. E. Constantinou, S. L. Ginell, C. Mao, and N. C. Seeman, "From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal," *Nature*, vol. 461, pp. 74–77, 2009.
[18] J. S. Shin and N. A. Pierce, "A synthetic DNA walker for molecular transport," *Journal of the American Chemical Society*, vol. 126, pp. 10 834–10 835, 2004.
[19] P. Yin, H. Yan, X. G. Daniell, A. J. Turberfield, and J. H. Reif, "A unidirectional DNA walker that moves autonomously along a track," *Communications of Angewandte Chemie International Edition*, vol. 43, pp. 4903–4911, 2004.

[20] J. Bath and A. J. Thuberfield, "DNA nanomachines," *Nature Nanotechnology*, vol. 2, pp. 275–284, 2007.

[21] K. Lund, A. J. Manzo, N. Dabby, N. Michelotti, A. Johnson-Buck, J. Nangreave, S. Taylor, R. Pei, M. N. Stojanovic, N. G. Walter, E. Winfree, and H. Yan, "Molecular robots guided by prescriptive landscapes," *Nature*, vol. 465, pp. 206–210, 2010.

[22] T. Nakano, "Biologically inspired network systems: a review and future prospects," *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, vol. 41, no. 4, pp. 630–643, 2011.

[23] D. Deamer, "A giant step towards artificial life," *Trends in Biotechnology*, vol. 23, no. 7, 2005.

[24] D. A. LaVan, T. McGuire, and R. Langer, "Small-scale systems for in vivo drug delivery," *Nature Biotechnology*, vol. 21, pp. 1184–1191, 2003.

[25] Y. Sasaki, Y. Shioyama, W. J. Tian, J. Kikuchi, S. Hiyama, Y. Moritani, and T. Suda, "A nanosensory device fabricated on a liposome for detection of chemical signals," *Biotechnology and Bioengineering*, vol. 105, pp. 37–43, 2010.

[26] T. Oberholzer, E. Meyer, I. Amato, A. Lustig, and P. A. Monnard, "Enzymatic reactions in liposomes using the detergent-induced liposome," *Biochimica et Biophysica Acta*, vol. 1416, pp. 57–68, 1999.

[27] V. Noireaux and A. Libchaber, "A vesicle bioreactor as a step toward an artificial cell assembly," *Proceedings of the National Academy of Sciences*, vol. 101, no. 51, pp. 17 669–17 674, 2004.

[28] Y. Moritani, S. Hiyama, and T. Suda, "Molecular communication among nanomachines using vesicles," in *NSTI Nanotechnology Conference and Trade Show*, vol. 2, 2006, pp. 705–708.

[29] M. Kaneda, S. M. Nomura, S. Ichinose, S. Kondo, K. Nakahama, K. Akiyoshi, and I. Morita, "Direct formation of proteo-liposomes by in vitro synthesis and cellular cytosolic delivery with connexin-expressing liposomes," *Biomaterials*, vol. 30, no. 23–24, pp. 3971–3977, 2009.

[30] I. Wegrzyn, H. Zhang, O. Orwar, and A. Jesorka, "Nanotube-interconnected liposome networks," *Nano Communication Networks*, vol. 2, pp. 4–15, 2011.

[31] K. Akiyoshi, A. Itaya, S. M. Nomura, and N. Ono, "Induction of neuron-like tubes and liposome networks by cooperative effect of gangliosides and phospholipids," *FEBS Letters*, vol. 534, pp. 33–38, 2003.

[32] S. M. Nomura, Y. Mizutani, K. Kurita, A. Watanabe, and K. Akiyoshi, "Changes in the morphology of cell-size liposomes in the presence of cholesterol: formation of neuron-like tubes and liposome networks," *Biochimica et Biophysica Acta*, vol. 1669, pp. 164–169, 2005.

[33] K. Sott, T. Lobovkina, L. Lizana, M. Tokarz, B. Bauer, Z. Konkoli, and O. Orwar, "Controlling enzymatic reactions by geometry in a biomimetic nanoscale network," *Nano Letters*, vol. 6, no. 2, pp. 209–214, 2006.

[34] L. You, R. S. Cox III, R. Weiss, and F. H. Arnold, "Programmed population control by cell–cell communication and regulated killing," *Nature*, vol. 428, no. 868–871, 2004.

[35] S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss, "A synthetic multicellular system for programmed pattern formation," *Nature*, vol. 434, pp. 1130–1134, 2005.

[36] M. T. Chen and R. Weiss, "Artificial cell–cell communication in yeast *Saccharomyces cerevisiae* using signaling elements from *Arabidopsis thaliana*," *Nature Biotechnology*, vol. 23, pp. 1551–1555, 2005.

[37] R. Weiss, S. Basu, S. Hooshangi, A. Kalmbach, D. Karig, R. Mehreja, and I. Netravali, "Genetic circuit building blocks for cellular computation, communications, and signal processing," *Natural Computing*, vol. 2, pp. 47–84, 2003.

[38] T. S. Gardner, C. R. Cantor, and J. J. Collins, "Construction of a genetic toggle switch in *Escherichia coli*," *Nature*, vol. 403, pp. 339–342, 2000.

[39] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss, "Synthetic biology: new engineering rules for an emerging discipline," *Molecular Systems Biology*, vol. 2, 2006.

[40] P. E. M. Purnick and R. Weiss, "The second wave of synthetic biology from modules to systems," *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 410–422, 2009.

[41] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335–338, 2000.

[42] S. Kobayashi, T. Kojidani, H. Osakada, A. Yamamoto, T. Yoshimori, Y. Hiraoka, and T. Haraguchi, "Artificial induction of autophagy around polystyrene beads in nonphagocytic cells," *Autophagy*, vol. 6, no. 1, pp. 36–45, 2010.

[43] T. Nakano, Y. H. Hsu, W. C. Tang, T. Suda, D. Lin, T. Koujin, T. Haraguchi, and Y. Hiraoka, "Microplatform for intercellular communication," in *Proc. 3rd Annual IEEE International Conference on Nano/Micro Engineered and Molecular Systems*, 2008, pp. 476–479.

[44] A. Folch and M. Toner, "Microengineering of cellular interactions," *Annual Review of Biomedical Engineering*, vol. 2, pp. 227–256, 2000.

[45] G. Homsy, T. F. Knight, and R. Nagpal, "Amorphous computing," *Communications of the ACM*, vol. 43, no. 5, pp. 74–82, 2000.

[46] T. Nakano, "Biological computing based on living cells and cell communication," in *Proc. 13th International Conference on Network-Based Information Systems (NBiS)*, 2010, pp. 42–47.

[47] T. Nakano, J. Shuai, T. Koujin, T. Suda, Y. Hiraoka, and T. Haraguchi, "Biological excitable media based on non-excitable cells and calcium signaling," *Nano Communication Networks*, vol. 1, no. 1, pp. 43–49, 2009.

[48] T. Nakano and J. Shuai, "Repeater design and modeling for molecular communication networks," in *Proc. 2011 IEEE INFOCOM Workshop on Molecular and Nanoscale Communications*, 2011, pp. 501–506.

[49] T. Nakano, T. Suda, T. Koujin, T. Haraguchi, and Y. Hiraoka, "Molecular communication through gap junction channels," *Springer Transactions on Computational Systems Biology X*, vol. 5410, pp. 81–99, 2008.

[50] L. C. Cobo and I. F. Akyildiz, "Bacteria-based communication in nanonetworks," *Nano Communication Networks*, vol. 1, no. 4, pp. 244–256, 2010.

[51] M. Gregori and I. F. Akyildiz, "A new nanonetwork architecture using flagellated bacteria and catalytic nanomotors," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 4, pp. 612–619, 2010.

[52] M. Moore and T. Nakano, "Addressing by beacon coordinates using molecular communication," in *Proc. 2011 IEEE INFOCOM Workshop on Molecular and Nanoscale Communications*, 2011, pp. 455–460.

[53] M. Moore and T. Nakano, "Addressing by beacon distances using molecular communication," *Nano Communication Networks*, vol. 2, no. 2–3, pp. 161–173, 2011.

[54] T. Nakano, M. Moore, Y. Okaie, A. Enomoto, and T. Suda, "Swarming biological nanomachines through molecular communication for targeted drug delivery," in *Proc. 6th International Conference on Soft Computing and Intelligent Systems/13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012)*, 2012, pp. 2317–2320.

[55] E. F. Keller and L. A. Segel, "Model for chemotaxis," *Journal of Theoretical Biology*, vol. 30, pp. 225–234, 1971.

[56] S. Kim, H. J. Kim, and N. L. Jeon, "Biological applications of microfluidic gradient devices," *Integrative Biology*, vol. 2, pp. 584–603, 2010.

[57] R. H. Baughman, A. A. Zakhidov, and W. A. de Heer, "Carbon nanotubes – the route toward applications," *Science*, vol. 297, pp. 787–792, 2002.

[58] A. Bianco, K. Kostarelos, and M. Prato, "Applications of carbon nanotubes in drug delivery," *Current Opinion in Chemical Biology*, vol. 9, pp. 674–679, 2005.

8 Application areas of molecular communication

We began this book with a short example about targeted drug delivery, an important application area for molecular communication; we now elaborate on this significant and motivating example for molecular communication. We also discuss other potential application areas of molecular communication, such as tissue engineering, lab-on-a-chip technology, and unconventional computation.

For each application area, we start with a brief introduction to the area and describe potential application scenarios where bio-nanomachines communicate through molecular communication to achieve the goal of an application. We then describe in detail selected designs of molecular communication systems as well as experimental results in the area.

8.1 Drug delivery

Drug delivery provides novel methodologies for drug administration that can maximize the therapeutic effect of drug molecules [1, 2]. One goal of drug delivery is to develop drug delivery carriers that can carry and deliver drug molecules to a target site in a body. Such drug delivery carriers are made from synthetic or natural particles (e.g., pathogens or blood cells) and they are typically nano to micrometer in size, so they can be injected into the circulatory system to propagate in a body. Targeting of drug delivery carriers in a body can be performed by exploiting pathological conditions that appear at a target site (e.g., tumor tissues). For instance, tumor tissues develop small gaps between nearby endothelial cells in a blood vessel, so drug delivery carriers, if small enough, can propagate through the gaps to accumulate in the tumor tissues. Targeting can also be done by using natural particulates such as pathogens and immune cells. These natural particulates are capable of detecting signaling molecules secreted from a target site (e.g., tumor tissues or diseased blood vessels) and actively migrating toward the target site. Another goal of drug delivery is to control the rate of drug release from drug delivery carriers. For this, drug delivery carriers can be made from biodegradable materials containing drug molecules, and the materials may be optimized to achieve a desired rate of drug release in a body. By achieving these goals, drug delivery has advantages over conventional drug administration, including the reduction of potential side effects of administering drug molecules by releasing drug molecules only at target sites and prolonged efficacy of drug molecules through a sustained release of drug molecules.

8.1.1 Application scenarios

Drug delivery carriers described above essentially represent bio-nanomachines that are presented in Sections 7.3 and 7.4 (e.g., liposome-based bio-nanomachines or functionally modified biological cells). For drug delivery, molecular communication may apply to (1) improve the targeting performance, (2) achieve a sustained drug release, and (3) perform a complex operation by a group of communicating bio-nanomachines. For instance, a large number of bio-nanomachines may be injected into a patient to perform a massively parallel search for a disease site. When a bio-nanomachine detects signaling molecules secreted from the disease site, it amplifies the signaling molecules to increase the concentration of the signaling molecules in the environment, which causes more bio-nanomachines to arrive at the disease site. A group of bio-nanomachines at the target site may also communicate to estimate the number of bio-nanomachines in the environment through quorum sensing (Section 3.3.4) and increase (or decrease) the rate of drug release if the number of bio-nanomachines is small (or large) to achieve a sustained release of drug molecules in the environment. A group of functionally different bio-nanomachines may also be employed to detect different environmental conditions, communicate to aggregate sensed conditions, and perform a complex operation such as logical computation to determine whether to release drug molecules.

8.1.2 Example: Cooperative drug delivery

The feasibility of applying molecular communication to drug delivery is demonstrated through a cooperative nanoparticle system for tumor targeting [3]. The molecular communication system uses two types of nanoparticles: signaling and receiving modules, namely, sender and receiver bio-nanomachines in molecular communication (Figure 8.1A). The signaling modules are first targeted to tumors where they broadcast the tumor location by activating tumor-specific endogenous biological pathways. In response to the activated tumor-specific biological pathways, the receiving modules in circulation efficiently accumulate at the tumor location. In experiments, the signaling modules are implemented as gold nanorods that can target tumors and, in response to externally applied light, produce heat to locally disrupt tumor vessels to cause tumor-specific coagulation; or implemented as engineered human proteins (tumor-targeted tissue factor) that can target angiogeneic tumor receptors to activate a coagulation pathway (Figure 8.1B). The receiving modules are implemented as magnetofluorescent iron oxide nanoworms for imaging or as a liposome for carrying drug molecules (Figure 8.1C). Communication between signaling and receiving modules is implemented by functionalizing the surface of receiving modules and allowing the receiving modules to respond to the coagulation pathway activated by the signaling modules. The receiving modules are surface modified with specific peptides that bind to fibrin (proteins generated from the activated coagulation pathway) or that act as a substrate for Factor XIII (a specific coagulation-cascade enzyme). The cooperative nanoparticle system has demonstrated a 40-fold increase in the accumulation of receiving modules at the tumor location compared to the case where only receiving modules are used.

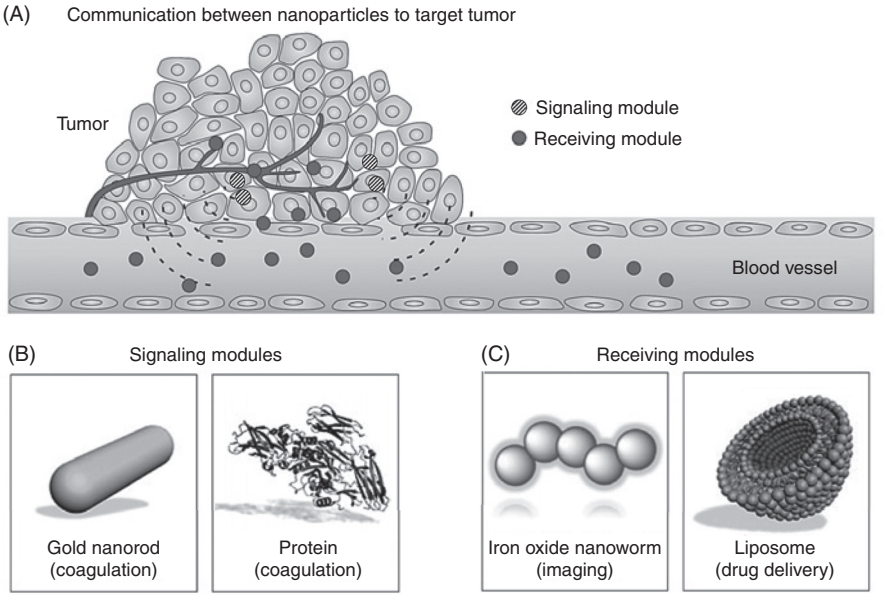


Figure 8.1 A molecular communication system for tumor targeting. Adapted from [4].

8.1.3 Example: Intracellular therapy

The feasibility of applying molecular communication to drug delivery can also be seen in the design of an autonomous molecular automaton for cancer diagnosis and therapy [5]. The molecular automaton, when applied to in vivo scenarios (e.g., within a biological cell), represents a bio-nanomachine that communicates with the DNA of a cell to form a molecular communication system (Figure 8.2). Briefly, the bio-nanomachine is implemented with a DNA molecule and composed of three modules: the input, computation, and output modules:

- The input module identifies whether disease indicating conditions are met. In the case of a prostate cancer cell, specific genes such as PPAP2B (lipid phosphate phosphohydrolase 3) and GSTP1 (glutathione S-transferase P) are underexpressed (↓), and some others such as PIM1 (proto-oncogene serine/threonine-protein kinase Pim-1) and HPN (serine protease hepsin) are overexpressed (↑). The input module in this case detects the level of the mRNA molecule transcribed from each gene, and produces transition DNA molecules used in the computation module. A transition DNA molecule implements a positive transition if a condition of prostate cancer (e.g., PPAP2B↓) is met, and otherwise, a negative transition.
- The computation module represents a molecular automaton that processes transition DNA molecules produced from the input module and indicates whether to release drug molecules. The computation module contains multiple copies of a single-stranded DNA (ssDNA) molecule in a hairpin structure. As shown in Figure 8.2, the ssDNA sequence at the right end, "drug," has known anti-cancer activity when

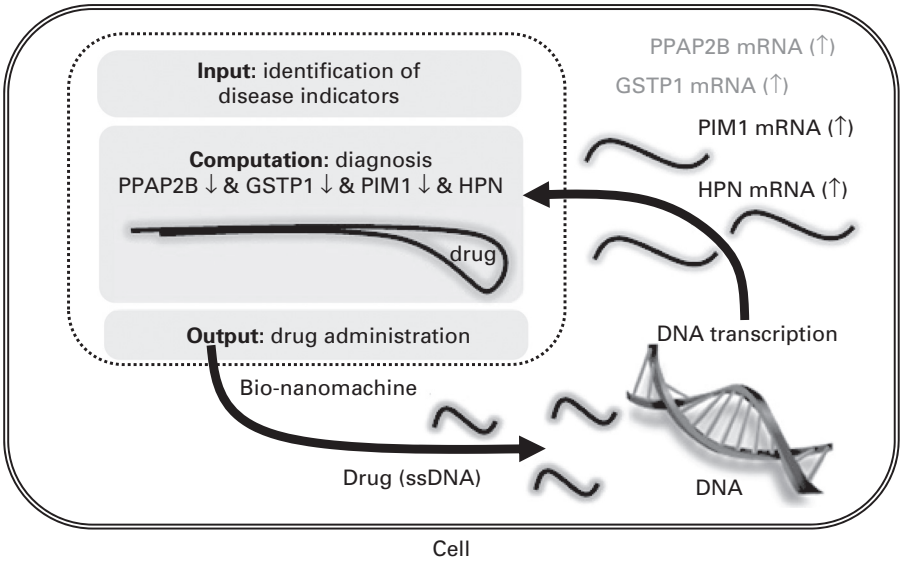


Figure 8.2 A molecular communication system for prostate cancer diagnosis and therapy. Reproduced from [5].

released from the hairpin. The ssDNA sequence at the left end is a set of diagnostic moieties for processing transition DNA molecules. A moiety is a functional group of the ssDNA molecule. In the case of the prostate cancer diagnosis, the ssDNA molecule contains four distinct diagnostic moieties, each of which is selected to react with a transition DNA molecule concerning PPAP2B↓, GSTP1↓, PIM1↑, or HPN↑. The ssDNA molecule has two states: positive and negative, each represented by a particular sequence of nucleotides in the ssDNA molecule. The ssDNA molecule is initially in the positive state, and remains in the same state, or changes to a negative state as a diagnostic molecule reacts with the corresponding transition DNA molecule. For instance, one diagnostic moiety binds to the transition DNA molecule concerning PPAP2B↓; and remains in the positive state if it is the positive transition or changes to a negative state, otherwise. When a diagnostic moiety binds to the corresponding transition DNA molecule, it is recognized and cleaved by a restriction enzyme. Thus the ssDNA molecule reacts with the transition DNA molecules concerning all four cancerous conditions in sequence and determines the final state.

- The output module combines results from multiple ssDNA molecules obtained from the computation module. In the simplest design, the number of ssDNA molecules in the positive state simply determines the number of drug molecules to release.[1] A drug molecule that is released propagates within the cell and binds to a specific

[1] In the design described in [5], another type of automaton is implemented to release drug suppressor molecules when the final state of an ssDNA molecule is negative. A drug suppressor molecule binds to and inactivates a drug molecule. Therefore, the number of ssDNA molecules in the positive state and those in the negative state determine the number of drug molecules to release.

mRNA molecule in the cell to inhibit the synthesis of the protein related to cancer activity.

8.2 Tissue engineering

Another promising area where molecular communication may apply is tissue engineering. Tissue engineering is an interdisciplinary area that applies principles of biological science, material science, and engineering to tissue and organ regeneration [6]. It forms an understanding of the developmental processes of tissues and organs; and based on the understanding, exploits living cells and biological materials to restore, maintain, or enhance tissue and organ functions. For tissue regeneration in vitro, for instance, autologous cells may be collected from a patient and cultured on engineered extracellular matrices called scaffolds. The scaffolds are fabricated from natural materials (e.g., collagen) or synthetic polymers that guide the differentiation and assembly of the cells into a three-dimensional tissue structure. The developed tissue structure is then implanted at the damaged area of the patient's body to restore lost tissues. Tissue engineering is expected to further advance with the development of stem cells as embryonic stem cells (ES cells) and induced pluripotent stem cells (iPS cells). These stem cells are promising materials for tissue engineering, especially for regenerative medicine, since they have the ability to produce more stem cells and to differentiate into diverse cell types.

Molecular communication plays an important role in tissue development (Section 3.3.6). In developing a tissue structure, a number of biological cells communicate a vast amount of information at time scales from seconds to weeks and dimensions from μm to cm to coordinate how to proliferate, differentiate, migrate, and die [7]. These biological cells synthesize growth factor molecules that propagate in the environment and bind to the cell-surface receptors of target biological cells. The types and concentrations of the growth factor molecules modulate (e.g., promote or inhibit) the proliferation, differentiation, and migration of target biological cells.

8.2.1 Application scenarios

The engineered molecular communication mechanisms described in Section 7.4 may apply to control the growth, differentiation, and movement of cells for tissue formation. Imagine for instance that iPS cells are developed from a patient and grown to form an initial structure (e.g., a sheet-like structure) in vitro. The iPS cells may be engineered to provide a man–machine interface to external devices, by which a human physician can control what types of growth factor molecules and how many growth factor molecules they release. The growth factor molecules released by the sender iPS cells may form a concentration gradient in the environment, which influences the behavior of the receiver iPS cells depending on the concentration and type of the growth factor molecules. The receiver iPS cells may differentiate into different cell types in response to the growth factor molecules through gene transcription and translation. The receiver iPS cells may also release molecules to produce cascading effects to lead to the development of a

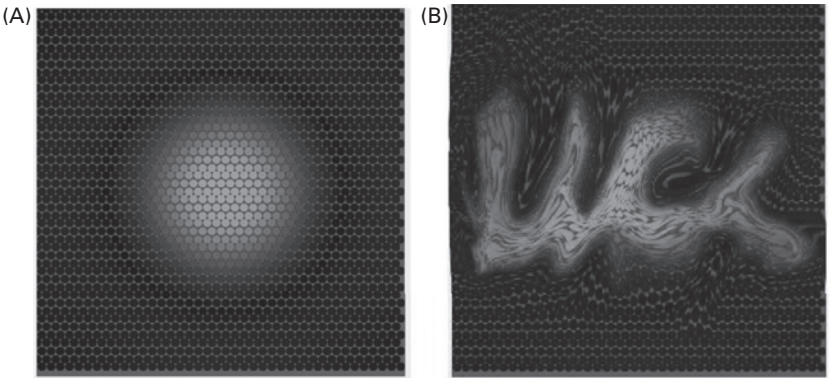


Figure 8.3 Pattern formation through molecular communication for tissue engineering. Formation of a spherical pattern (A) and three letters, UCI (University of California, Irvine) (B). Adapted from [8].

structure, similar to how tissue is developed in a natural setting. Under an autonomous scenario, iPS cells may be engineered to coordinate through molecular communication and develop into a target structure without external control. For instance, iPS cells may be engineered to form the initial structure of a molecular communication system from which a complex spatiotemporal structure emerges in a self-organized manner [8]. Two types of growth factor molecules, if they act as activator and inhibitor molecules, produce various patterns such as Turing patterns [9]. By modifying conditions, such as the release rate and diffusion coefficient of a particular type of molecule, and increasing the number of molecule types involved, different patterns can be formed. For instance, a molecular communication system (e.g., starting with a single iPS cell) may develop into a spherical pattern (i.e., of a group of iPS cells) through reaction and diffusion processes. The molecular communication system may be modified such that the molecular communication system, originally developed into a spherical form (Figure 8.3A), develops into a spatial pattern showing the three letters, UCI (University of California, Irvine) (Figure 8.3B).

8.2.2 Example: Tissue structure formation

Efforts to develop molecular communication systems for tissue engineering have been made and in part demonstrated by pioneering work in synthetic biology [10]. In the work, a molecular communication system consists of genetically engineered *E. coli* bacteria that can create patterns of differentiation based on the concentration gradient of molecules (Figure 8.4A).

The sender bacterium expresses tetracycline repressor protein (TetR) controlled LuxI. LuxI is an enzyme that catalyzes the synthesis of acyl-homoserine lactones (AHL), a membrane-permeable diffusive molecule (i.e., information molecules). The sender bacterium synthesizes TetR and produces LuxI, which catalyzes the AHL synthesis. When the sender bacterium starts synthesizing AHL, AHL diffuses in the environment

to form a concentration gradient. The AHL concentration is at its highest around the sender bacterium and decreases with distance from the sender bacterium.

The receiver bacterium expresses LuxR, a transcriptional activator protein, that responds to the AHL released by the sender bacterium. LuxR in the receiver bacterium controls the expression of the Lac repressor ($LacI_{M1}$, a product of a codon-modified lacI) and the lambda repressor (CI) that further controls the expression of a wild type Lac repressor (LacI). $LacI_{M1}$ and LacI then together control the green fluorescent protein (GFP) expression. The GFP expression in a receiver bacterium is dependent on the AHL concentration or the distance to the sender bacterium, meaning that receiver bacteria are addressed by location:

- Receiver bacterium A is close to the sender bacterium. In this case, the AHL concentration at the receiver bacterium is high. A high expression level of LuxR in the receiver bacterium leads to the expression of $LacI_{M1}$ (and CI), which represses the GFP expression.
- Receiver bacterium B is at an intermediate distance from the sender bacterium. In this case, the AHL concentration is moderate. The moderate expression level of LuxR leads to the moderate expression of $LacI_{M1}$ and CI. $LacI_{M1}$ has low repression efficiency and does not effectively repress the GFP expression. The CI has high repression efficiency and represses the wild type LacI. As a result, GFP is expressed.
- Receiver bacterium C is far from the sender bacterium. In this case, the AHL concentration is low. $LacI_{M1}$ and CI are expressed only at the basal levels, and LacI is expressed to repress the GFP expression.

The sender and receiver bacteria described above can create various patterns of GFP expression. When a sender bacterium at a particular location in a culture dish releases AHL, receiver bacteria around a particular distance from the sender bacterium location express GFP, leading to the formation of a ring with a radius equal to that distance. Further, depending on the concentrations and locations of sender bacteria, different patterns can be formed, such as an ellipse from two sender groups on center, a heart from three sender groups, and a clover from four sender groups (Figure 8.4B).

8.3 Lab-on-a-chip technology

Another promising application area of molecular communication is lab-on-a-chip (LOC) and its related systems such as micro-electromechanical systems (MEMS), microfluidic systems, and micro total analysis systems (μTAS). By rough definition [11], LOCs are called MEMS or bio-MEMS when micro-technology (e.g., photolithography) is used to integrate different functions on a single chip. LOCs are called microfluidic systems when they control and manipulate the behavior of fluids on a single chip. LOCs are also called μTAS when they provide all the necessary functions for analyzing molecules such as transport, reaction, separation, and detection of molecules on a single chip.

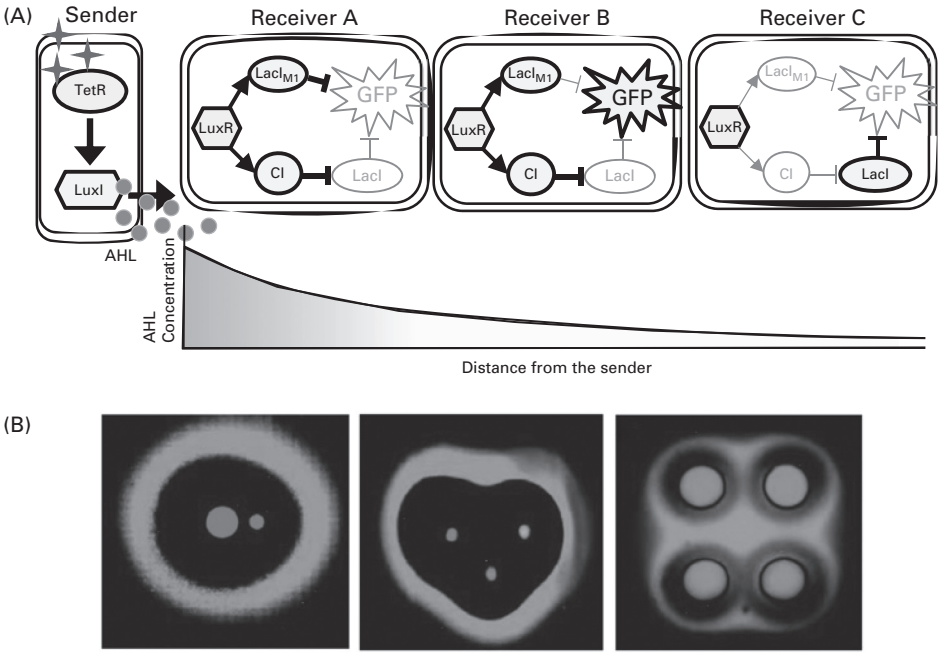


Figure 8.4 Pattern-forming molecular communication system for tissue regeneration. Reproduced from [10].

A goal of LOC and its related systems is to reduce the size of an integrated device down to a single chip size of millimeters that performs laboratory functions, such as the reaction, separation, and identification of molecules [11, 12]. The reduction of device size decreases the amount of expensive reagents that need to be consumed, and decreases costs for chemical analysis. The reduction of device size also increases the speed, efficiency, and throughput by containing chemical reactions within small volumes (e.g., sub-milliliter). The reduction of device (or component) size may also allow for an increase in functional complexity and for massive parallelization of multiple functions. The LOC technology is expected to assist portable, rapid, and complex medical diagnosis for future health care services.

One key feature of LOC technology is to create a channel network that enables integration of different functional units for reaction, separation, and detection on a single chip. The channel network may consist of multiple liquid reservoirs storing specific reactants and interconnecting channels for performing chemical functions using the reactants. For instance, a simple channel network with a Y topology may be designed, where two input reservoirs are connected to one output reservoir via a merging channel. The two input reservoirs inject the respective reactants into the merging channel, and the merging channel mixes the reactants, induces intermolecular reactions, and transports the products (e.g., by pressure) to the output reservoir. The products in the output reservoir may be mixed and labeled with fluorescent dyes for later detection, or they may be

injected into another channel (e.g., containing a separation gel) to perform separation based on their molecular mass or electrokinetics.

8.3.1 Application scenarios

The engineered molecular communication components presented in Chapter 7 may be integrated in LOCs. For instance, vesicle-based bio-nanomachines (Section 7.3) may be used as liquid reservoirs where chemical reactions take place, and engineered molecular motors (Section 7.1) may be used as a channel to transport molecules (e.g., reactants and products) between the liquid reservoirs. We will see two application examples below.

8.3.2 Example: Bio-inspired lab-on-a-chip

A conceptual framework of LOCs with molecular communication components is illustrated and partially demonstrated in [13] (Figure 8.5). The demonstrated molecular communication system consists of two giant liposomes acting as sender (a) and receiver bio-nanomachines (b), a set of reactants (c) encapsulated in a small liposome (d) to be transported, and microtubules (e) and kinesin molecular motors (f) that transport the small liposome from the sender giant liposome to the receiver giant liposome when adenosine triphosphate (ATP) is supplied. In this system, channel proteins, connexins (Cx), are embedded on the surface of the sender and receiver giant liposomes and of the small liposome. The connexins form channels that allow reactants to diffuse from the sender giant liposome to the small liposome and from the small liposome to the receiver giant liposome. The gemini-peptide lipids (GPL) are also embedded on these surfaces to assist the formation of the connexin channels between the sender/receiver giant liposomes and the small liposome. To facilitate the loading and unloading of the small liposome at the sender and receiver giant liposomes, the small liposome, microtubule, and receiver giant liposome are embedded with a single-stranded DNA sequence (ssDNA) (g). The ssDNA sequence embedded with the small liposome is partially complementary to the ssDNA sequence embedded with the microtubule, and weakly binds at the sender giant liposome. The ssDNA sequence embedded with the small liposome is completely complementary to the ssDNA sequence embedded with the receiver giant

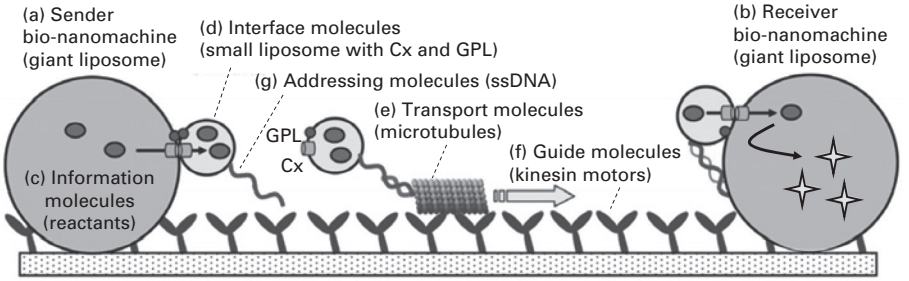


Figure 8.5 Example design of a molecular communication system for lab-on-a-chip. Adapted from [13].

liposome and strongly binds at the receiver giant liposome. The strong binding causes the small liposome to unbind from the microtubule and bind to the receiver giant liposome. The molecular communication system can thus transport a set of reactants from the sender giant liposome to the receiver giant liposome where the chemical reactions may be induced.

8.3.3 Example: Smart dust biosensors

A molecular communication system on a LOC was also demonstrated in [14]. This system, called a smart dust biosensor, uses kinesin molecular motors and microtubules for capturing, tagging, and detecting a target analyte (e.g., an antigen) in solution on a micrometer size chip. As shown in Figure 8.6, the chip is circular, 20 μm in height and 800 μm in diameter. The circular surface is divided into three radial zones: capture, tagging, and detection zones. The chip surface is coated with kinesin molecular motors that propel the microtubules from the capture zone to the tagging zone, and then from the tagging zone to the detection zone. The microtubules are surface-modified with antibodies to capture target antigens that may exist in the capture zone. The tagging zone contains fluorescent tags functionalized with second antibodies; only antigen-loaded microtubules bind to the fluorescent tags via the double antibody sandwich (i.e., microtubule-antibody-antigen-antibody-fluorescent tag). This allows the fluorescent tags in the tagging zone to be attached to and moved by the microtubules to the detection zone only in the presence of the antigens in the capture zone. In the detection zone, the microtubules overhang or are immobilized at the side wall. The excitation light is applied here and the fluorescence emitted from the fluorescent tags (if they exist)

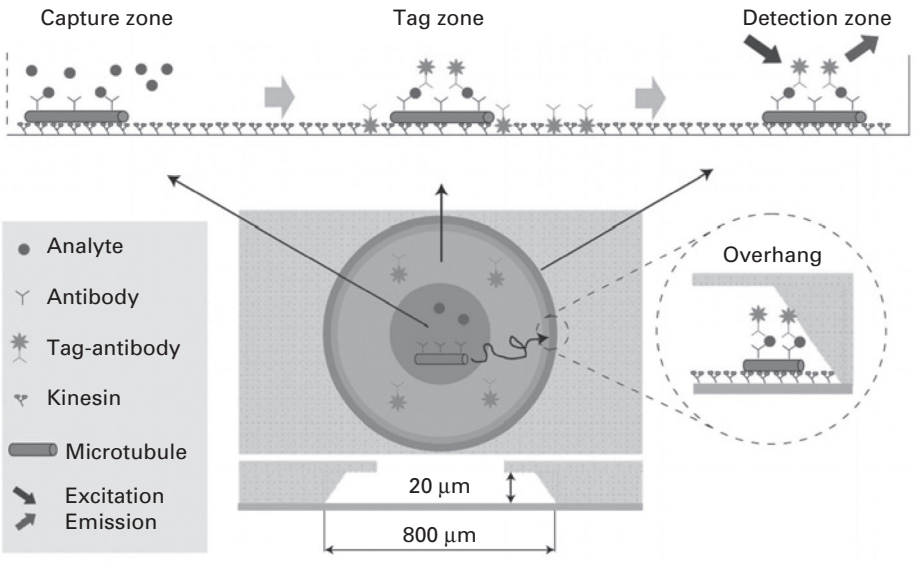


Figure 8.6 A molecular communication system built on a single chip. Adapted from [14].

is measured to identify whether the capture zone contains the antigens. The molecular communication system can thus capture, transport, and detect the target analyte that may be present in the environment.

8.4 Unconventional computation

The last application area of molecular communication discussed in this chapter is unconventional computation. Unconventional computation provides a collection of non-silicon-based computing paradigms. Today's computing architecture, which is based on silicon technology, has successfully increased computing speed and memory capacity by increasing the complexity of an electronic circuit. Stated by Moore's law, the number of transistors that can be placed on a circuit doubles approximately every 18 months. While the existing silicon technology may continue to improve for many more years (approximately 20 years), it eventually faces the physical limit when the circuit size approaches the size of a molecule, and thereafter no further improvement will be achieved [15]. In unconventional computation, research efforts are being made to exploit physical, chemical, or biological materials to develop new computing architectures and to design algorithms for such architectures to solve computationally difficult problems more efficiently and quickly. One promising approach is to use biological materials and systems (e.g., DNA molecules, enzymes, cells) since they present remarkable features such as extremely high functional complexity and large-scale parallelism that cannot be achieved with silicon-based electronic circuits. Note however that unconventional computation is not likely to replace today's computing paradigm; it is rather directed toward certain types of computational problems such as specific combinatorial problems.

8.4.1 Application scenarios

An architecture for unconventional computation may be implemented from a group of engineered bio-nanomachines. Each bio-nanomachine may function as a computational unit and communicate to perform computation collectively. A large number of bio-nanomachines may be spatially arranged to form a network for computation. A communication link in the network may be provided through molecular communication. A large number of bio-nanomachines, when they are integrated into a network, may form an unconventional computing architecture that responds to externally input molecules and produces output molecules to yield computational results. Here we explore the possibilities of developing an unconventional computing architecture using bio-nanomachines that communicate through molecular communication.

8.4.2 Example: Reaction diffusion computation

In the first hypothetical setting, a group of bio-nanomachines communicate to perform computation in a manner similar to excitable chemical systems. Excitable chemical systems, such as the Belousov-Zhabotinsky (BZ) reaction system (Box 8.1), are capable of

Box 8.1 Belousov-Zhabotinsky reactions

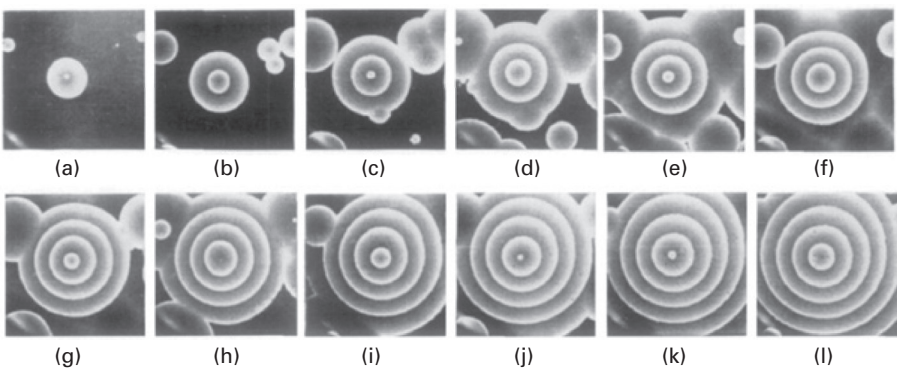
A Belousov-Zhabotinsky (BZ) reaction is known to establish a nonlinear chemical oscillator in solution [16,17]. It is caused by a set of molecules and generates chemical waves that propagate spatially and oscillate temporally in solution. A BZ reaction is caused by complex chemical processes involving oxidization and reduction of molecules; e.g.,

$$Fe^{2+} \xrightarrow{HBrO_3} Fe^{3+} \tag{8.1}$$

$$Fe^{3+} \xrightarrow{BMA} Fe^{2+} \tag{8.2}$$

where ferroin $Fe(phen)_3$ simply referred to as Fe above is oxidized in (8.1) and reduced in reaction (8.2). In a BZ reaction system, Fe is oxidized through bromic acid ($HBrO_3$) that increases its concentration through an autocatalytic process, and thus Fe^{3+} concentration increases quickly (and thus Fe^{2+} concentration decreases). Such an autocatalytic process is a key to induce a BZ reaction. (8.1) leads to the production of bromomalonic acid (BMA) to cause (8.2). BMA is then oxidized to produce bromide ion (Br^-) that functions as a strong inhibitor for (8.1). As a result, the Fe^{3+} concentration decreases and Fe^{2+} concentration increases. Since (8.2) leads to the production of $HBrO_3$, (8.1) starts again to increase the Fe^{3+} concentration (and thus decrease the Fe^{2+} concentration). Such a cycle is repeated in a BZ reaction system to continuously generate propagating and oscillating waves in solution.

The figure below illustrates series of waves observed in a BZ reaction system [17]. A wave is originated from a single point and emanating in an expanding circular pattern. The nonlinear features of BZ reaction waves are exploited to create novel devices such as a detector for the distance to a wave source [18] or a detector for the direction to the wave source [19] in addition to performing logic operations or solving computational geometry problems.



BZ reactions observed at an interval of 30 seconds. Adapted from [17].

generating and propagating waves through non-equilibrium chemical reactions. It has been demonstrated that the excitable chemical systems can perform computations such as performing logic operations or solving problems in computational geometry [20, 21, 22]. In logic operations, the presence or absence of an excitation wave at a particular location represents a binary value. A logic operation is then performed by propagating two input waves from two different locations, colliding them to produce an output wave, and propagating the output wave to a particular location. In computational geometry, a fundamental problem of creating a Voronoi diagram, given a set of locations, can be solved by initiating excitation waves from the set of locations. Since excitation waves propagate at the same constant speed, two wavefronts collide at a location with an equal distance from the locations where the two waves are initiated. The locations of collision represent the boundaries of Voronoi cells, which are thus examined to create a Voronoi diagram.

The feasibility of creating an excitable system from bio-nanomachines to perform computation can be found in biological cells such as neurons, glial cells (astrocytes), and amoeba (*Dictyostelium*) that propagate action potentials, calcium waves, and cyclic AMP (cAMP) waves, respectively. One implementation of excitable mechanisms is through calcium-induced calcium release (CICR) (Section 3.3.3). Within a biological cell, there are a number of calcium channels or clusters of calcium channels that collectively generate and propagate Ca^{2+} waves in a reaction diffusion manner. These calcium channels may be spatially arranged within microdomains of a cell, which are loosely coupled by the diffusion of Ca^{2+} to form a calcium signaling network capable of various types of computation [23]. When multiple cells are connected through communication pathways (e.g., gap junction channels) to form a network of cells, there exist multiple levels of hierarchy in calcium signaling networks (e.g., microdomain, cell, and group levels) that perform computation with high complexity.

8.4.3 Example: Artificial neural networks

In the second hypothetical setting, a group of bio-nanomachines communicate following an architecture for artificial neural networks (ANNs), a model of soft computing. In an ANN, a number of neurons are connected through synaptic connections to form a network of neurons. Each neuron receives input signals from a set of neurons connecting to the neuron, applies a sigmoidal function to integrate the input signals, and propagates output signals to another set of neurons. Neurons may function asynchronously to achieve large-scale parallelism. The synaptic connections strengthen or weaken to implement the ability of reinforcement learning.

A bio-nanomachine for such an architecture may be designed as a protein molecule with a sigmoidal behavior. A natural protein network for bacterial chemotaxis (Section 3.3.1) demonstrates the feasibility of implementing an artificial neural network of protein molecules [24]. In Section 3.3.1, the bacterial chemotaxis is shown to use a network of protein molecules that collectively measures the concentrations of external molecules (i.e., input to the network), determines the rate of change in these concentrations, and controls the flagellar motor depending on the rate of change (i.e., output from the

network). Since these protein molecules are typically large and diffuse slowly, they may not directly communicate through physical contact. The protein molecules may instead communicate by propagating diffusive molecules such as second messengers of low molecular mass. Such a communication link may strengthen or weaken when the affinity of the protein molecules for the propagating diffusive molecules changes. In natural evolution, communication links in a protein network are reinforced as a result of random mutation in protein molecules and natural selection of protein networks with the desirable affinity. For an architecture for unconventional computing, however, how such reinforcement can be implemented in a practical manner needs to be addressed.

8.4.4 Example: Combinatorial optimizers

In the last hypothetical setting, a group of bio-nanomachines communicate to form an unconventional computation architecture for combinatorial optimization. Combinatorial optimization problems are computationally difficult and conventional research is directed to design algorithms that can find reasonable solutions efficiently. An unconventional computation architecture based on bio-nanomachines may provide an alternative and scalable approach to such problems if large numbers of bio-nanomachines are available (e.g., in the Avogadro order scale) or they autonomously replicate to match the complexity and scale of a problem, and if such bio-nanomachines are capable of a massively parallel search for solutions and communicating to integrate solutions.

A possibility of developing such an unconventional computation architecture is demonstrated in the maze-solving amoeboid organism capable of finding a solution to a maze [25]. The amoeboid organism used in experiments, known as *Physarum polycephalum*, is a single-celled multi-nucleate organism. Depending on environmental conditions, such as the locations of food sources, the organism changes its structure by extending the network of protoplasmic veins, dividing the network into multiple

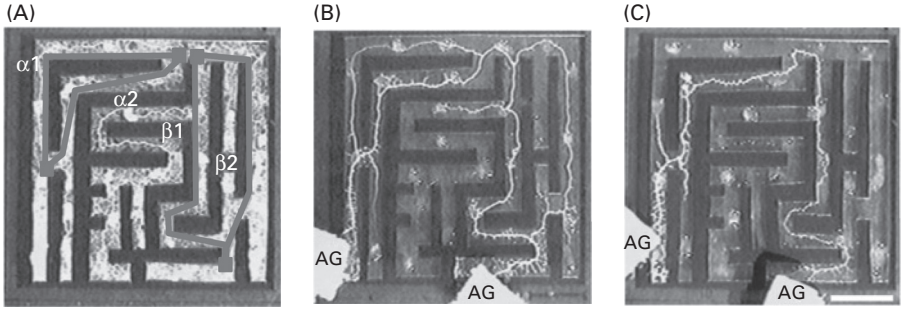


Figure 8.7 A molecular communication system of slime molds capable of finding a maze solution. Scale bar, 1 cm. Adapted from [25].

networks, or merging multiple networks into one network. Initially, small fragments of the organism are randomly placed on a plastic film representing a maze to solve (Figure 8.7A). These fragments of the organism then extend and merge to form an initial network structure containing two sets of different paths: $\alpha 1$ (41 mm) and $\alpha 2$ (33 mm); and $\beta 1$ (44 mm) and $\beta 2$ (45 mm) as shown in the figure. Their food, agar (AG), is then placed at two locations to indicate the entrance and exit of the maze (Figure 8.7B). The organism then changes the network structure to efficiently absorb the nutrients from the AG by increasing the widths of links (i.e., thicknesses of the protoplasmic veins), while keeping the network structure connected. As a result, the organism removes links to which no AG is connected. The organism also removes a link if a shorter alternative link is available, thus, $\alpha 1$ and $\beta 2$ are removed. Finally, the network structure is shaped to create the shortest path between the two AG locations (Figure 8.7C).

8.5 Looking forward: Conclusion and summary

In this chapter, we presented the state-of-the-art in several potential application areas of molecular communication: drug delivery, lab-on-a-chip, tissue engineering, and unconventional computation. From the examples presented in this chapter, molecular communication has clear potential to produce disruptive applications in a variety of fields.

However, it is also clear from the discussion in this chapter that the state-of-the-art – as futuristic as it would have seemed just a few years ago – has a long way to go before realizing the full potential of molecular communication. Notably, the communication presented in the lab-implemented examples is rudimentary, with little encoding or complexity beyond simple binary (on-off) messages. Meanwhile, contemporary theoretical research in molecular communication has shown that the mathematical theory of communication can lead to sophisticated communication strategies among bio-nanomachines.

Looking into the future, these communication strategies will be needed to develop truly advanced, disruptive, and commercially viable applications using molecular communication. As a result, it will be necessary to bring contemporary research activities – which are often performed as theoretical studies or in simulation – into wet labs, to determine their practical effectiveness.

References

[1] T. M. Allen and P. R. Cullis, "Drug delivery systems: entering the mainstream," *Science*, vol. 303, no. 5655, pp. 1818–1822, 2004.

[2] J.-W. Yoo, D. J. Irvine, D. E. Discher, and S. Mitragotri, "Bio-inspired, bioengineered and biomimetic drug delivery carriers," *Nature Reviews Drug Discovery*, vol. 10, pp. 521–535, 2011.

[3] G. von Maltzahn, J.-H. Park, K. Y. Lin, N. Singh, C. Schwoppe, R. Mesters, W. E. Berdel, E. Ruoslahti, M. J. Sailor, and S. N. Bhatia, "Nanoparticles that communicate in vivo to amplify tumour targeting," *Nature Materials*, vol. 10, pp. 545–552, 2011.

[4] Y. Wang, P. Brown, and Y. Xia, "Nanomedicine: swarming towards the target," *Nature Materials*, vol. 10, pp. 482–483, 2011.

[5] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, "An autonomous molecular computer for logical control of gene expression," *Nature*, vol. 429, pp. 423–429, 2004.

[6] L. G. Griffith and G. Naughton, "Tissue engineering – current challenges and expanding opportunities," *Science*, vol. 295, no. 5557, pp. 1009–1014, 2002.

[7] P. Tayalia and D. J. Mooney, "Controlled growth factor delivery for tissue engineering," *Advanced Materials*, vol. 21, pp. 3269–3285, 2012.

[8] S. N. Watanabe and T. Suda, "Encoding spatial patterns into a chemical regulatory network," School of Information and Computer Science, University of California, Irvine, Tech. Rep. 09-04, 2009.

[9] S. Kondo and T. Miura, "Reaction-diffusion model as a framework for understanding biological pattern formation," *Science*, vol. 329, pp. 1616–1620, 2010.

[10] S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss, "A synthetic multicellular system for programmed pattern formation," *Nature*, vol. 434, pp. 1130–1134, 2005.

[11] P. S. Dittrich and A. Manz, "Lab-on-a-chip: microfluidics in drug discovery," *Nature Reviews Drug Discovery*, vol. 5, pp. 210–218, 2006.

[12] P. Yager, T. Edwards, E. Fu, K. Helton, K. Nelson, M. R. Tam, and B. H. Weigl, "Microfluidic diagnostic technologies for global public health," *Nature*, vol. 442, pp. 412–418, 2006.

[13] S. Hiyama and Y. Moritani, "Molecular communication: harnessing biochemical materials to engineer biomimetic communication systems," *Nano Communication Networks*, vol. 1, no. 1, pp. 20–30, 2010.

[14] T. Fischer, A. Agarwal, and H. Hess, "A smart dust biosensor powered by kinesin motors," *Nature Nanotechnology*, vol. 4, pp. 162–166, 2009.

[15] T. Munakata, "Beyond silicon: new computing paradigms," *Communications of the ACM*, vol. 50, no. 9, pp. 30–34, 2007.

[16] A. N. Zaikin and A. M. Zhabotinsky, "Concentration wave propagation in two-dimensional liquid-phase self-oscillating system," *Nature*, vol. 2255, pp. 535–537, 2012.

[17] A. M. Zhabotinsky and A. N. Zaikin, "Autowave processes in a distributed chemical system," *Journal of Theoretical Biology*, vol. 40, pp. 45–61, 1973.

[18] J. Gorecki, J. N. Gorecka, K. Yoshikawa, Y. Igarashi, and H. Nagahara, "Sensing the distance to a source of periodic oscillations in a nonlinear chemical medium with the output information coded in frequency of excitation pulses," *Physical Review E*, vol. 72, no. 4, 2005.

[19] H. Nagahara, T. Ichino, and K. Yoshikawa, "Direction detector on an excitable field: field computation with coincidence detection," *Physical Review E*, vol. 70, no. 3, 2004.

[20] A. Adamatzky, B. D. L. Costello, and T. Asai, *Reaction-Diffusion Computers*, 1st edn. Elsevier Science, 2005.

[21] A. Toth and K. Showalter, "Logic gates in excitable media," *Journal of Chemical Physics*, vol. 103, no. 6, pp. 2058–2066, 1995.

[22] I. Motoike and K. Yoshikawa, "Information operations with an excitable field," *Physical Review E*, vol. 59, no. 5, pp. 5534–5360, 1999.

[23] H. G. E. Hentschel, C. S. Pencea, and A. Fine, "Computing with calcium stores and diffusion," *Neurocomputing*, vol. 58–60, pp. 455–460, 2004.
[24] D. Bray, "Protein molecules as computational elements in living cells," *Nature*, vol. 376, pp. 307–312, 2012.
[25] T. Nakagaki, H. Yamada, and A. Toth, "Maze-solving by an amoeboid organism," *Nature*, vol. 407, p. 470, 2012.

9 Conclusion

Throughout this book, we have introduced the emerging discipline of molecular communication, examining theoretical and implementational aspects. From our survey of the field, there has been tremendous progress towards understanding molecular communication in the past few years. Yet there remains much to be done in order to realize the disruptive potential of this new technology. To conclude our book, we discuss the important future challenges and open problems that remain in this field.

9.1 Toward practical implementation

Certain enabling technologies will be needed to move beyond the current state of knowledge and allow practical communication systems to be built. The following gives an outline of some of these technologies.

- **Functional design of molecular communication systems.** In Chapter 4, we discussed communication-related functionalities: encoding, decoding, information-carrying, interface, addressing, transport, and guide functionalities. In Chapter 7, we looked at how these functionalities can be implemented. Ideally, molecular communication systems – and, more generally, bio-nanomachines – should be designed in a way that is analogous to existing electronic design techniques, by selecting components that implement given functions, and interfacing them together. However, these design tools do not currently exist.
- **Theoretical framework for molecular communication.** In Chapters 5 and 6, we discussed the modeling and theoretical analysis of molecular communication systems. In spite of the large volume of work in the past several years, there remain many open problems. For example, it is not clear how accurately molecular communication systems are represented by abstract mathematical models, and realistic models for molecular communication components have yet to be developed. More fundamentally, the ultimate capacity of molecular communication is unknown, even in a crude sense: at the time of writing, even an order-of-magnitude estimate is unavailable. The ultimate goal is to establish theoretical frameworks for molecular communication among bio-nanomachines, including communication links among bio-nanomachines, and networks for bio-nanomachines. Many existing theories, including network theory (see, e.g., Chapter 4), can be used in developing this theoretical framework.

- **Manufacturing molecular communication components.** We reviewed some possible designs for molecular communication systems in Chapter 7. Most existing molecular communication systems have been designed for single experiments; however, a view towards mass manufacturing is needed to develop the masses of bio-nanomachines that are required for most applications. Thus, an important challenge is to establish engineering methods to manufacture communication components. Lithography-based top-down approaches may apply for bio-nanomachines of relatively larger (micro-scale) size, while self-assembly based bottom-up approaches may be more promising to engineer smaller-scale bio-nanomachines (e.g., in the nanometer range).
- **Incorporation into larger networks.** The applications we discussed in Chapter 8 will require the creation of interfaces for molecular communication systems. These interfaces need to take two forms: interfaces to communicate with other bio-nanomachines in the environment, and interfaces to communicate with external devices (especially macroscopic ones). For example, medical applications may require a device placed on a human body to collect information gathered by bio-nanomachines and to send control signals to bio-nanomachines imbedded in a human body. In creating the first type of interface, there is much work to be done in terms of creating practical protocols and techniques for networking on bio-nanomachines, in order to develop true nanonetworks. In creating the second type of interface (i.e., an interface to communicate with an external device), non-biological materials or components may be integrated into bio-nanomachines, and communication techniques such as magnetic, acoustic, mechanical, and temperature-based communication may be used to communicate with an external device.

9.2 Toward the future: Demonstration projects

As we have shown throughout this book, the future potential of molecular communication is enormous. An obvious next step is to develop a convincing, practical demonstration of molecular communication in a medical or industrial context: such a demonstration would signal the capabilities of molecular communication, and would serve as a bridge to the applications described in Chapter 8.

The best demonstration projects should be achievable with existing molecular communication technology. Some examples include:

- **Autonomous swarm navigation.** Can a swarm of bio-nanomachines, cooperating only with molecular communication and with no outside help, navigate a fluid maze to search for an objective? Once there, can the bio-nanomachines recognize that they have arrived at their goal? If the fluid maze simulates the bloodstream, and the objective simulates a tumor, this project would demonstrate the feasibility of targeted drug delivery.
- **Autonomous construction.** Can a swarm of bio-nanomachines, cooperating only with molecular communication and with no outside help, arrange themselves into a

complex structure? This project would emulate the development and growth of living tissue, and would illustrate the feasibility of tissue engineering (e.g., the growth of replacement tissues *in situ* inside the body); it would also demonstrate the prospect of automated micro/nanoscale construction techniques.

- **High-latency massively parallel bio-computation.** Can a swarm of bio-nanomachines perform a non-trivial computing task, where cooperation between the machines is mediated through molecular communication? It is already possible to produce bio-nanomachines capable of very simple logic operations; we now ask whether bio-nanomachines can play the role of processing cores. Processing would be very slow and subject to high latency, but this could be mitigated by massive parallel processing. This project would demonstrate the feasibility of bio-supercomputers that could grow and repair themselves.

We believe these projects to be within the reach of contemporary technology. However, we expect them to be challenging, requiring the talents of large and highly interdisciplinary groups of researchers, as well as access to sophisticated laboratory equipment. Nonetheless, we believe that demonstrating the potential of molecular communication is a crucial next step in the development of this field.

Appendix Review of probability theory

In this appendix we provide a brief review of probability theory, which may be helpful for some readers. This appendix is intended only as a quick reminder; readers who have never seen these concepts before might be better served by an undergraduate textbook on probability.

A.1 Basic probability

Suppose you have an experiment with an unknown outcome, where there are a finite number of possible outcomes: for instance, a coin flip (heads or tails), or the roll of a standard die (1 through 6). The *probability* of an outcome is the fraction of times that outcome would happen if the experiment were repeated over and over.[1]

Outcomes take values in a set called the *sample space*; for example, for a standard die, the sample space is $\mathcal{S} = \{1,2,3,4,5,6\}$. The *probability mass function* (pmf) gives the probability of each outcome: let $x \in \mathcal{S}$ represent the outcome of the roll of a fair die; then the pmf of X, written $p_X(x)$, is

$$p_X(x) = \Pr(X = x) \tag{A.1}$$

$$= \begin{cases} \frac{1}{6}, & x = 1,2,\ldots,6; \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

Two useful properties of pmfs:

1. Values in pmfs are never negative (i.e., $p_X(x) \geq 0$ for all x).
2. The sum of the probabilities of all the events in the sample space is equal to 1 (i.e., $\sum_{x \in \mathcal{S}} p_X(x) = 1$).

The second property may be interpreted as follows: some outcome in the sample space is certain to happen. Taken together with the first property, the second property implies that pmf values are never outside the range from 0 to 1.

As a short aside on notation, note that X and x are distinct: X represents the random variable, and x represents particular values that X could take in the argument of the pmf. This distinction is unimportant in most cases that we deal with in this book, and

[1] Philosophically, there are other interpretations of probability, but this view is sufficient for our purposes.

we normally use the lower case variable (though one exception is seen in (A.1) with $\Pr(X = x)$, meaning the probability that the random variable X is equal to the particular value x). However, we follow certain notational conventions with the upper case variable: in a pmf, the subscript is the capital letter form of the variable; the same is true of expected values (which we describe in the next section), as well as entropy and mutual information (which we describe in Chapter 6).

Sometimes x has a continuous sample space: for example, the first arrival time of a particle under Brownian motion is a continuous random variable, taking values in the positive real numbers. In most cases, the probability of x being *exactly equal to* some arbitrarily chosen real number is zero (think, for example, about the impossibility of a first arrival time being exactly equal to $1.000\ldots$ second, to infinite precision). However, the probability of the outcome being in a range of values may have a nonzero value. This is captured through the probability density function (pdf), written $f_X(x)$: the probability of X being on the range from a to b is given by

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx. \tag{A.3}$$

Further, the cumulative distribution function (cdf), written $F_X(x)$, is given by

$$F_X(x) = \Pr(X \leq x) \tag{A.4}$$

$$= \int_{-\infty}^{x} f_X(x)dx. \tag{A.5}$$

The pdf and cdf are nonnegative, like pmfs. However, only the cdf has a maximum value of 1; the pdf has no maximum.

A.2 Expectation, mean, and variance

The *expected value* of x, written $E[X]$, is given by

$$E[X] = \sum_{x \in \mathcal{S}} x p_X(x) \tag{A.6}$$

if x is discrete, and

$$E[X] = \int_{x=-\infty}^{\infty} x f_X(x)dx \tag{A.7}$$

if x is continuous; $E[X]$ is also called the mean, or average, of x, and is often given the symbol μ. We can take the expected value of functions of x: let $g(X)$ be an arbitrary function, then $E[g(X)]$ is given by

$$E[g(X)] = \int_{x=-\infty}^{\infty} g(x) f_X(x)dx \tag{A.8}$$

for continuous random variables. (For discrete random variables, we use the sum instead of integration and pmf instead of pdf.) Expectation is distributed over addition, and scalar multiplication comes out of expectation, for example,

$$E[\alpha X + \beta X^2] = \alpha E[X] + \beta E[X^2]. \tag{A.9}$$

We can also calculate conditional expectation. If the random variable x is dependent on y, we can write the conditional expectation as $E[X|Y]$, given by

$$E[X|Y] = \int_{x=-\infty}^{\infty} x f_{X|Y}(x|y) dx. \tag{A.10}$$

The *variance* of a random variable X is an important special case of expectation. Letting $\mu = E[X]$, variance is given by

$$\text{Var}[X] = E[(X-\mu)^2]. \tag{A.11}$$

In words, this is the average of the squared deviation from the mean. We leave it as an exercise for the reader to show that

$$\text{Var}[X] = E[X^2] - \mu^2. \tag{A.12}$$

We often use σ^2 as the symbol for variance (the square emphasizes that the variance is the squared deviation from the mean).

A.3 The Gaussian distribution

The Gaussian distribution is an important distribution, found in many applications. If x is distributed Gaussian with mean μ and variance σ^2, then the pdf of x is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{A.13}$$

(The function $\exp(\cdot)$ is the exponential function: $\exp(x) = e^x$.) The cdf of x is given by a special function known as the error function, or $\text{erf}(\cdot)$, defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(t^2) dt. \tag{A.14}$$

Then

$$F_X(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right). \tag{A.15}$$

Without $\text{erf}(\cdot)$, $F_X(x)$ can't be stated in closed form.

A.4 Conditional, marginal, and joint probabilities

If there are two or more random variables, we may discuss their joint probability. For example, let x and y represent two random variables. Their joint pdf (alternatively, pmf, if discrete) is written $f_{X,Y}(x,y)$ (alternatively, $p_{X,Y}(x,y)$). Joint probability may be interpreted as the probability of x and y occurring together.

The probability of one of the random variables alone can be formed from the joint probability: for instance, given a joint pdf $f_{X,Y}(x,y)$, we can find $f_X(x)$ by integrating over y:

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy. \tag{A.16}$$

This process is called *marginalization*, and the pdf $f_X(x)$ is sometimes called the *marginal distribution* of x.

The conditional probability is the probability of an outcome of one random variable, given knowledge of another; this quantity can be found from the joint and marginal distribution. For example, the probability of y given x, written $f_{Y|X}(y|x)$, is given by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{A.17}$$

$$= \frac{f_{X,Y}(x,y)}{\int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy}, \tag{A.18}$$

where (A.18) follows from (A.16) and (A.17). Note that (A.17) implies

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f(x). \tag{A.19}$$

A.5 Markov chains

Points of a Brownian motion form a *Markov chain*, which is a sequence of random variables with a particular structure to the joint probability. For example, let x, y, and z represent three random variables. Extending (A.19), we could write the joint probability as

$$f_{X,Y,Z}(x,y,z) = f_{Z|X,Y}(z|x,y)f_{Y|X}(y|x)f_X(x). \tag{A.20}$$

Equation (A.20) applies to any distribution of x, y, and z. However, if

$$f_{Z|X,Y}(z|x,y) = f_{Z|Y}(z|y) \tag{A.21}$$

(omitting the dependence on x), then (A.20) can be written

$$f_{X,Y,Z}(x,y,z) = f_{Z|Y}(z|y)f_{Y|X}(y|x)f_X(x), \tag{A.22}$$

and we say that the three variables form a Markov chain: each is only dependent on the preceding variable.

We can generalize a Markov chain to an arbitrary number of random variables: let a, b, c, d, ... represent a collection of random variables; then these variables form a Markov chain if their joint pdf can be written

$$f_A(a)f_{B|A}(b|a)f_{C|B}(c|b)f_{D|C}(d|c)\ldots \tag{A.23}$$

Markov chains have conditional independence: for instance, in (A.23), if I know c, that tells me all I need to know about the distribution of d: given c, knowing a and b gives no new information. This is the *Markov property*, which is sometimes stated: *given the present, the future is independent of the past.*

Index

actin filament, 28, 125
action potential, 37, 40, 47
acyl-homoserine lactone (AHL), 44, 136, 157
additive inverse Gaussian noise (AIGN) channel, 100, 118
addressing, 63
addressing molecule, 52, 122
adenine, 28
adenosine diphosphate (ADP), 27
adenosine triphosphate (ATP), 27, 31, 44
ADP, *see* adenosine diphosphate
AHL, *see* acyl-homoserine lactone
AIGN, *see* additive inverse Gaussian noise
amino acid, 22
ANN, *see* artificial neural network
artificial neural network (ANN), 164
aspartate transcarbamoylase (ATCase), 123
ATCase, *see* aspartate transcarbamoylase
ATP, *see* adenosine triphosphate
autoinducer, 44
autoinducing polypeptide (AIP), 44

bacterial conjugation, 40, 45
Belousov-Zhabotinsky (BZ) reaction, 162
bio-nanomachine, 2, 21, 53, 92, 122, 123, 129, 133, 136
BioBrick, 66
BMI, *see* brain–machine interface
bow-tie architecture, 66
brain–machine interface (BMI), 12
Brownian motion, 71–74
  first arrival time, 80–82
  with drift, 100
BZ reaction, *see* Belousov-Zhabotinsky reaction

Caenorhabditis elegans (*C. elegans*), 46
calcium-induced calcium release (CICR), 43, 143
calcium ion (Ca^{2+}), 42
calcium signaling, 40, 42
calmodulin, 43
calmodulin-dependent protein kinase II (CaM kinase II), 123
cAMP, *see* cyclic AMP
capacity, 111, 114–117
carbon nanotube (CNT), 147
cell, 34, 136
cell-surface receptor, *see* receptor
Chebyshev's inequality, 88
chemotactic signaling, 39, 40
chemotaxis, 40
chromosome, 29
CICR, *see* calcium-induced calcium release
CNT, *see* carbon nanotube
COat Protein (COP), 42
codon, 31
communications model, 98
concentration, 86, 94
congestion control, 65
connexin, 143, 160
COP, *see* COat Protein
cross-layer architecture, 65
cyclic AMP (cAMP), 164
cytosine, 28
cytoskeleton, 34
cytosol, 34

delay-selector channel, 102, 108
deoxyribonucleic acid (DNA), 28, 129
  base-pairing rule, 29
  complementarity, 29
  ligase, 132
  nanotechnology, 129
  transcription, 30
  translation, 31
  walker, 131
detection, 104–106, 108
Dictyostelium, 164
diffusion coefficient, 39, 56
DNA, *see* deoxyribonucleic acid
drug delivery, 13, 152
dynamic instability, 28, 125
dynein, 27, 39

E. coli, see Escherichia coli
EGF, *see* epidermal growth factor
embryonic stem cell (ES cell), 156

encoding, 53, 56
endocrine signaling, 37
endoplasmic reticulum (ER), 34, 42, 143
entropy, 110, 117
  conditional, 110
  differential, 110
  joint, 110
enzyme, 23, 123
epidermal growth factor (EGF), 40, 46
error handling, 62, 130
ER, *see* endoplasmic reticulum
ES cell, *see* embryonic stem cell
Escherichia coli (*E. coli*), 34, 41, 130, 157
eukaryote, 34
excitable chemical system, 164

flagellum, 34, 40
flow control, 62

gap junction channel, 37, 44, 143
Gaussian distribution, 89
Gaussian random process, 90
gene, 30, 137
genetic engineering, 137
GFP, *see* green fluorescent protein
Golgi apparatus, 34, 42
G-protein, *see* guanine nucleotide-binding protein
green fluorescent protein (GFP), 13, 137, 158
GTP, *see* guanosine triphosphate
guanine, 28
guanine nucleotide-binding protein (G-protein), 27
guanosine triphosphate (GTP), 31
guide molecule, 52, 122, 125, 131, 135, 142

Hamming code, 62
Hill function, 26
hormonal signaling, 40, 47
hormone, 47

IEEE P1906.1, 11, 66
in-network processing, 65
induced pluripotent stem cell (iPS cell), 156
information molecule, 52, 122, 129, 136
information theory, 109–120
inositol 1,4,5-trisphosphate (IP_3), 39, 43, 143
inter-symbol interference, 84
interface molecule, 52, 122, 130, 134
intracellular therapy, 14, 154
inverse Gaussian distribution, 81, 100, 119
ion channel, 27, 48
IP_3, *see* inositol 1,4,5-trisphosphate
iPS cell, *see* induced pluripotent stem cell

kinesin, 27, 125, 160, 161

Lévy distribution, 81, 113
lab-on-a-chip (LOC), 11, 158
layer, 59, 97
  link, 61–63
  network, 64–65
  physical, 60–61, 97
  upper, 65–66
ligand, 26, 93
lipid membrane, 31
liposome, 33, 132, 153, 160
LOC, *see* lab-on-a-chip
LuxR/LuxI system, 44, 136, 157
lysosome, 34

macromolecule, 55
MAP, *see* maximum *a posteriori* detection
Markov chain, 74, 94
Markov property, 74, 75, 77, 80, 99
maximum likelihood (ML), 105, 106
maximum *a posteriori* (MAP) detection, 104
media access control, 62
MEMS, *see* micro-electromechanical system
messenger RNA (mRNA), 30, 138, 154
Michaelis–Menten kinetics, 25
micro total analysis system (μTAS), 158
micro-electromechanical system (MEMS), 11, 158
microfluidics, 146, 158
microtubule, 28, 79, 125, 160, 161
MIMO, *see* multiple input and multiple output
mitochondrion, 34
ML, *see* maximum likelihood
modulation, 60
mole fraction, 86
molecular automaton, 154
molecular communication
  analytical examples, 83, 105, 107, 112
  applications, 11–15, 152–166
  characteristics, 54–58
  components, 48, 52–60, 122
  design and engineering, 122–147
  examples from biological systems, 38–49
  history, 6–11
  model, 2, 52, 71–77
  network architecture, 58–66
  process, 53–54
  simulation, 77–78
molecular imaging, 13
molecular motor, *see* motor protein
molecular rail, 27
Moore's law, 162
morphogen signaling, 40, 46
motor protein, 27, 78, 125
mRNA, *see* messenger RNA
multiple input and multiple output (MIMO), 57
mutual information, 111, 117
myosin, 27, 58, 125
μTAS, *see* micro total analysis system

nanonetwork, 9
nanotube-vesicle network, 135
network formation, 64
neuron, 47
neuronal signaling, 40, 47
neurotransmitter, 47
NTP, *see* nucleoside triphosphate
nucleic acid, 28
nucleoside triphosphate (NTP), 31
nucleotide, 28
nucleus, 34
nullcline, 140

organelle, 34

paracrine signaling, 37
parameter estimation, 106
peptide, 22
phospholipase C (PLC), 43
phosphorylation, 23
Physarum polycephalum, 165
plasma membrane, 34
plasmid, 45
PLC, *see* phospholipase C
probability of error, 105, 110, 115, 116
prokaryote, 34
promoter, 30, 138
propagation, 37, 54, 61, 72
  active mode, 37, 79
  passive mode, 37
protein, 22, 123
  allostericity, 27
  cooperativity, 26
  domain, 22

quorum sensing, 40, 44

reaction diffusion computation, 162
reaction-diffusion wave, 38, 164
receptor, 26, 41, 43, 93
restriction enzyme, 129, 132
ribonucleic acid (RNA), 28
ribosome, 31
RNA, *see* ribonucleic acid
RNA pol, *see* RNA polymerase
RNA polymerase (RNA pol), 30, 137
routing, 64

second messenger, 36, 55
signal transduction, 26
single-stranded DNA (ssDNA), 127, 129, 154, 160
soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNARE), 42
SNARE, *see* soluble N-ethylmaleimide-sensitive factor attachment protein receptors
ssDNA, *see* single-stranded DNA
stem cell, 156
sum-product algorithm, 108
synaptic signaling, 37
synchronization, 63
synthetic biology, 66, 138

targeted drug delivery, *see* drug delivery
TCP/IP, 59
terminator, 31
thymine, 28
thyroid-stimulating hormone (TSH), 40, 47
thyroxine, 47
timing channel, 100, 116
tissue engineering, 12, 156
transfer RNA (tRNA), 31
transport molecule, 52, 122, 125, 131, 144
tRNA, *see* transfer RNA
TSH, *see* thyroid-stimulating hormone
TNT, *see* tunneling nanotube
tunneling nanotube (TNT), 37

unconventional computation, 15, 162
uracil, 28
uridine triphosphate (UTP), 31
UTP, *see* uridine triphosphate

vesicle, 31
  budding, 33
  fussion, 33
vesicular trafficking, 39, 41
Vibrio fischeri, 136

WI, *see* Wiener-Ideal model
Wiener process, 72
  multi-dimensional, 76
  with drift, 75
Wiener-Ideal (WI) model, 99, 102, 111, 113, 115, 116

Z channel, 113