Intelligent Bioinformatics

Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems
Edward Keedwell and Ajit Narayanan
© 2005 John Wiley & Sons Ltd. ISBN: 0-470-02175-6

Intelligent Bioinformatics

The application of artificial intelligence techniques to bioinformatics problems

Edward Keedwell
and
Ajit Narayanan

School of Engineering, Computer Science and Mathematics
University of Exeter, UK

John Wiley & Sons, Ltd

Copyright © 2005 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The Publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark,
Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

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

Library of Congress Cataloguing-in-Publication Data

British Library Cataloguing in Publication Data

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

ISBN 0 470 02175 6

Typeset in 10.5/13.5pt Sabon by TechBooks, New Delhi, India
Printed and bound in Great Britain by TJ International Ltd., Padstow, Corwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface ix

Acknowledgement xi

PART 1 INTRODUCTION 1

1 Introduction to the Basics of Molecular Biology 3

1.1 Basic cell architecture 3
1.2 The structure, content and scale of deoxyribonucleic acid (DNA) 4
1.3 History of the human genome 9
1.4 Genes and proteins 10
1.5 Current knowledge and the 'central dogma' 21
1.6 Why proteins are important 23
1.7 Gene and cell regulation 24
1.8 When cell regulation goes wrong 26
1.9 So, what is bioinformatics? 27
1.10 Summary of chapter 28
1.11 Further reading 29

2 Introduction to Problems and Challenges in Bioinformatics 31

2.1 Introduction 31
2.2 Genome 31
2.3 Transcriptome 40
2.4 Proteome 50
2.5 Interference technology, viruses and the immune system 57
2.6 Summary of chapter 63
2.7 Further reading 64

3 Introduction to Artificial Intelligence and Computer Science 65

3.1 Introduction to search 65
3.2 Search algorithms 66
3.3 Heuristic search methods 72
3.4 Optimal search strategies 76
3.5 Problems with search techniques 83
3.6 Complexity of search 84
3.7 Use of graphs in bioinformatics 86
3.8 Grammars, languages and automata 90
3.9 Classes of problems 96
3.10 Summary of chapter 98
3.11 Further reading 99

PART 2 CURRENT TECHNIQUES 101

4 Probabilistic Approaches 103

4.1 Introduction to probability 103
4.2 Bayes' Theorem 105
4.3 Bayesian networks 111
4.4 Markov networks 116
4.5 Summary of chapter 125
4.6 References 126

5 Nearest Neighbour and Clustering Approaches 127

5.1 Introduction 127
5.2 Nearest neighbour method 130
5.3 Nearest neighbour approach for secondary structure protein folding prediction 132
5.4 Clustering 135
5.5 Advanced clustering techniques 138
5.6 Application guidelines 144
5.7 Summary of chapter 145
5.8 References 146

6 Identification (Decision) Trees 147

6.1 Method 147
6.2 Gain criterion 152
6.3 Over fitting and pruning 157
6.4 Application guidelines 160
6.5 Bioinformatics applications 163
6.6 Background 169

6.7 Summary of chapter 170
6.8 References 170

7 Neural Networks 173
7.1 Method 173
7.2 Application guidelines 185
7.3 Bioinformatics applications 187
7.4 Background 192
7.5 Summary of chapter 193
7.6 References 193

8 Genetic Algorithms 195
8.1 Single-objective genetic algorithms – method 195
8.2 Single-objective genetic algorithms – example 202
8.3 Multi-objective genetic algorithms – method 205
8.4 Application guidelines 207
8.5 Genetic algorithms – bioinformatics applications 210
8.6 Summary of chapter 217
8.7 References and further reading 217

PART 3 FUTURE TECHNIQUES 219

9 Genetic Programming 221
9.1 Method 221
9.2 Application guidelines 230
9.3 Bioinformatics applications 232
9.4 Background 236
9.5 Summary of chapter 236
9.6 References 237

10 Cellular Automata 239
10.1 Method 239
10.2 Application guidelines 245
10.3 Bioinformatics applications 247
10.4 Background 251
10.5 Summary of chapter 252
10.6 References and further reading 252

11 Hybrid Methods 255
11.1 Method 255
11.2 Neural-genetic algorithm for analysing gene expression data 256
11.3 Genetic algorithm and *k* nearest neighbour hybrid for biochemistry solvation 262

11.4 Genetic programming neural networks for determining gene – gene interactions in epidemiology 265
11.5 Application guidelines 268
11.6 Conclusions 268
11.7 Summary of chapter 269
11.8 References and further reading 269

Index **271**

Preface

It is widely recognized that the field of biology is in the midst of a 'data explosion'. A series of technical advances in recent years has increased the amount of data that biologists can record about different aspects of an organism at the genomic, transcriptomic and proteomic levels. This data is, of course, vital to advancing our knowledge. In recent years, the discipline of *bioinformatics* has allowed biologists to make full use of the advances in computer science and computational statistics in analysing this data. However, as the volume of data grows, the techniques used must become more sophisticated to cater for large-scale data and noise. Also, given the growth in biological data, there is a need to extract information that was not previously known from these databases to supplement current knowledge. Large databases may contain interesting patterns that, if identified and authenticated by further laboratory and clinical work, can lead to novel theories about the causes of various diseases and also possibly to new drugs for their treatment. The discipline of bioinformatics has reached the end of its first phase, and the motivation behind this book is to characterize the principles that may underlie second phase bioinformatics. That is, second phase bioinformatics is when the discipline, instead of being informed by just computer science and computational statistics, is also informed by artificial intelligence techniques.

As we show in this book, there are problems in bioinformatics and many other sciences that cannot be solved satisfactorily even with the fastest computers. Clearly, a more 'intelligent' approach is required to solve these increasingly difficult bioinformatics problems, such as gene expression analysis and protein structure prediction. This book attempts to address this by looking at the latest advances in artificial intelligence technology as applied to computational problems in biology. Artificial intelligence methods are often based on the ways in which humans solve search and optimization problems, or how nature has solved its own problems, for example by using the principles of 'survival of the fittest' in evolutionary computation.

This book is divided into three parts, each containing a number of chapters. These parts are designed to allow readers to access the material most relevant to them. The first part, *Introduction,* introduces the material necessary to understand the technology and biology included in the later chapters. We recognize that bioinformatics is highly cross-disciplinary and therefore some, all or none of these chapters may be relevant to the reader, depending on their background. The next part, *Current Techniques*, describes the established artificial intelligence techniques in bioinformatics including probabilistic, nearest neighbour and genetic algorithm approaches. The final part, *Future Techniques*, is intended to give the reader an impression of the latest thinking in the area of intelligent bioinformatics. Some of these approaches may not have been widely applied to problems in bioinformatics, but algorithms such as genetic programming and various hybrid approaches can be expected to make a big impact in this domain if experience in other areas of science and technology is anything to go by.

In short, this book has been written to engage and interest readers from many disciplines. Biologists are provided for in that there is a full introduction to the challenges for computer science, and computer scientists should also find the chapters on biology and bioinformatics informative. Practicing bioinformaticians are also likely to find the book enlightening, as much of the material has previously only been included in specialist publications and a collection such as this provides a single resource for many intelligent problem-solving techniques in bioinformatics. However, as with any book of this type, not every technique can be included due to space restrictions and apologies are offered to researchers whose own favourite analytical techniques are not covered in this book.

Edward Keedwell
Ajit Narayanan

Acknowledgements

The authors would like to thank everyone involved with producing this book including staff at the Department of Computer Science and Centre for Water Systems at the University of Exeter, in particular Godfrey Walters, Dragan Savic and Soon-Thiam Khu. In addition to this, we would like to thank Bjorn Olsson for his contribution to the tutorials on which this book is based, and Laetitia Jourdan for her helpful comments. Also, we would like to thank the many MSc students on the Bioinformatics programme at the University of Exeter, who contributed towards some of the material for this book. Finally we would also like to thank the editorial and production staff at Wiley, in particular Joan Marsh, Andrea Baier and Robert Hambrook for making this book possible.

We are grateful to WoltersKluwer Health for permission to adapt and re-use Figures 2.10, 6.3, 7.1, 7.2 and 7.3 and Table 5.1 from 'Artificial intelligence techniques for bioinformatics', A. Narayanan, E. C. Keedwell and B. Olsson, *Applied Bioinformatics* 2002: 1(4) 191–222.

Dedications

Ed Keedwell – This book is dedicated to my family Rob, Lyn, Rich and Loveday, to Kate, and in memory of Alex Larigo.

Ajit Narayanan – This book is dedicated to Lucy, Belinda and Kieran, my mother Janaki, my brother Ramesh and sister Seetha.