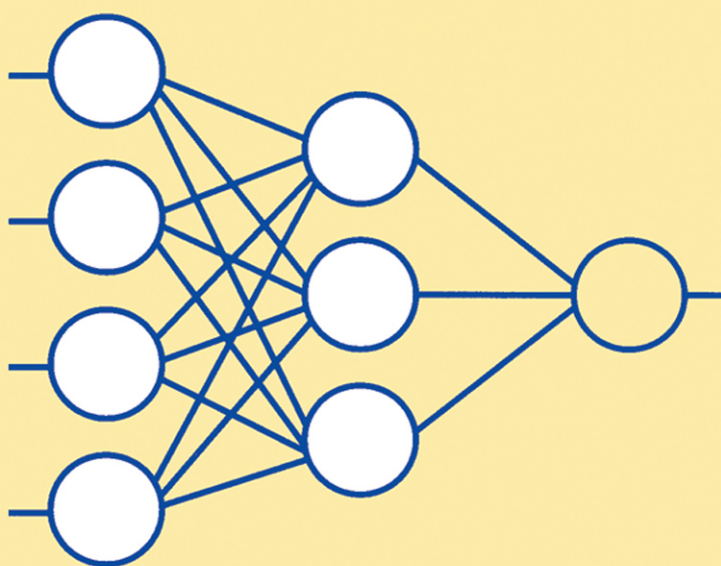
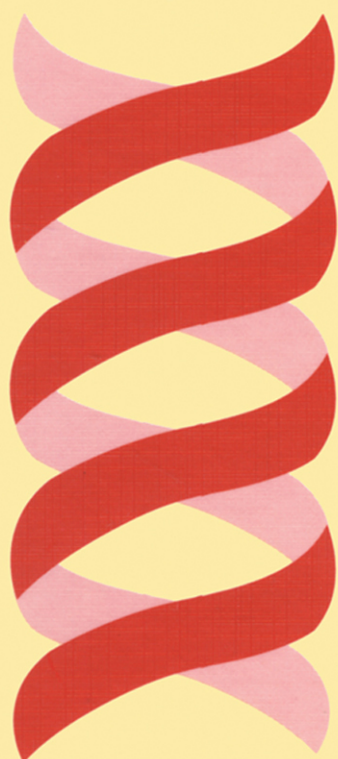


METHODS IN COMPUTATIONAL
BIOLOGY AND BIOCHEMISTRY

SERIES EDITOR: A.K. KONOPKA

Volume
1

Neural Networks and Genome Informatics



Cathy H. Wu
Jerry W. McLarty

ELSEVIER

Neural Networks and Genome Informatics

METHODS IN COMPUTATIONAL BIOLOGY AND BIOCHEMISTRY

Volume 1

Series Editor

A.K. KONOPKA

Maryland, USA



ELSEVIER

Amsterdam - Lausanne - New York - Oxford - Shannon - Singapore - Tokyo

Neural Networks and Genome Informatics

C.H. Wu
J.W. McLarty

University of Texas Health Center at Tyler
Department of Epidemiology and Biomathematics
11937 U.S. Highway 271
Tyler, TX 75708-3154
USA



2000

ELSEVIER

Amsterdam - Lausanne - New York - Oxford - Shannon - Singapore - Tokyo

ELSEVIER SCIENCE Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB, UK

© 2000 Elsevier Science Ltd. All rights reserved.

This work is protected under copyright by Elsevier Science, and the following terms and conditions apply to its use:

Photocopying

Single photocopies of single chapters may be made for personal use as allowed by national copyright laws. Permission of the Publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.

Permissions may be sought directly from Elsevier Science Global Rights Department, PO Box 800, Oxford OX5 1DX, UK; phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.co.uk. You may also contact Global Rights directly through Elsevier's home page (<http://www.elsevier.nl>), by selecting 'Obtaining Permissions'.

In the USA, users may clear permissions and make payments through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone: (978) 7508400, fax: (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 171 631 5555; fax: (+44) 171 631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works

Tables of contents may be reproduced for internal circulation, but permission of Elsevier Science is required for external resale or distribution of such material. Permission of the Publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage

Permission of the Publisher is required to store or use electronically any material contained in this work, including any chapter or part of a chapter.

Except as outlined above, no part of this work may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the Publisher.

Address permissions requests to: Elsevier Global Rights Department, at the mail, fax and e-mail addresses noted above.

Notice

No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2000

Library of Congress Cataloging in Publication Data

A catalog record from the Library of Congress has been applied for.

British Library Cataloguing in Publication Data

A catalogue record from the British Library has been applied for.

ISBN: 0 08 042800 2

Ⓢ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

This book is dedicated to:

My late father, my husband, daughter, and son, for their inspirations. – CHW.

The late Allen B. Cohen, whose passion for science, generous spirit, inherent fairness and integrity and gentle leadership helped create a wonderfully productive environment in which to work. - JWM.

This page intentionally left blank

Preface

The resurgence of interest in artificial neural networks fortunately coincided with the emergence of new technology in molecular biology and the explosion of information about the genomes of humans and other species. Many important problems in genome informatics have been successfully addressed with artificial neural networks, and a vast literature has developed within the last two decades. The purpose of this book is to introduce molecular biologists and other informatics scientists to artificial neural network technology and terminology; to review the major neural network applications in genome informatics; to address the important issues in applying neural network technology to informatics; and to identify significant remaining problems.

Part I of this book gives an overview of applications of artificial neural network technology. Part II contains a tutorial introduction to the most commonly used neural network architectures, network training methods, and applications and limitations of the different architectures. Part III reviews the current state of the art of neural network applications to genome informatics and discusses crucial issues such as input variable selection and preprocessing. Finally, Part IV identifies some of the remaining issues and future directions for research, including integration of statistical rigor into neural network applications, hybrid systems and knowledge extraction.

Acknowledgements

This work would not have been possible without the support of the National Library of Medicine, the National Biomedical Research Foundation in Georgetown, Washington, D.C., and the University of Texas Health Center at Tyler. The authors are grateful for their helpful discussions with Dr. Hongzhan Huang at the National Biomedical Research Foundation and for the expertise and efforts of Dr. Karen Sloan and Sara Shepherd.

Cathy H. Wu
Jerry McLarty
Georgetown and Tyler, May 1999

This page intentionally left blank

Contents

PART I.....	1
Overview	1
CHAPTER 1.....	3
Neural Networks for Genome Informatics.....	3
1.1 What Is Genome Informatics?	3
1.1.1 Gene Recognition and DNA Sequence Analysis	4
1.1.2 Protein Structure Prediction.....	8
1.1.3 Protein Family Classification and Sequence Analysis	9
1.2 What Is An Artificial Neural Network?.....	10
1.3 Genome Informatics Applications.....	11
1.4 References.....	12
PART II	17
Neural Network Foundations	17
CHAPTER 2.....	19
Neural Network Basics.....	19
2.1 Introduction to Neural Network Elements.....	19
2.1.1 Neurons	19
2.1.2 Connections between Elements.....	20
2.2 Transfer Functions.....	21
2.3.1 Summation Operation.....	21
2.3.2 Thresholding Functions	22
2.3.3 Other Transfer Functions.....	24
2.4 Simple Feed-Forward Network Example	25
2.5 Introductory Texts.....	26
2.6 References.....	27
CHAPTER 3.....	29
Perceptrons and Multilayer Perceptrons.....	29
3.1 Perceptrons	29
3.1.1 Applications	29
3.1.2 Limitations.....	33
3.2 Multilayer Perceptrons	33
3.2.1 Applications.....	36
3.2.2 Limitations.....	38
3.3 References.....	38
CHAPTER 4.....	41
Other Common Architectures	41

4.1 Radial Basis Functions.....	41
4.1.1 Introduction to Radial Basis Functions	41
4.1.2 Applications	44
4.1.3 Limitations	46
4.2 Kohonen Self-organizing Maps.....	46
4.2.1 Background	47
4.2.2 Applications	48
4.2.3 Limitations	50
4.4 References	50
CHAPTER 5	51
Training of Neural Networks	51
5.1 Supervised Learning	51
5.2.1 Training Perceptrons.....	52
5.2.2 Multilayer Perceptrons.....	55
5.2.3 Radial Basis Functions.....	58
5.2.4 Supervised Training Issues.....	59
5.3 Unsupervised Learning	62
5.4 Software for Training Neural Networks.....	63
5.5 References	63
PART III	65
Genome Informatics Applications.....	65
CHAPTER 6	67
Design Issues – Feature Presentation.....	67
6.1 Overview of Design Issues.....	67
6.2 Amino Acid Residues	68
6.3 Amino Acid Physicochemical and Structural Features.....	69
6.4 Protein Context Features and Domains.....	71
6.5 Protein Evolutionary Features	73
6.6 Feature Representation.....	74
6.7 References	76
CHAPTER 7	79
Design Issues – Data Encoding.....	79
7.1 Direct Input Sequence Encoding	79
7.2 Indirect Input Sequence Encoding.....	81
7.3 Construction of Input Layer	83
7.4 Input Trimming	84
7.5 Output Encoding.....	86
7.6 References	86
CHAPTER 8	89
Design Issues – Neural Networks.....	89
8.1 Network Architecture.....	89
8.2 Network Learning Algorithm	91