Perspectives in Neural Computing

Springer-Verlag London Ltd.

Also in this series:

Adrian Shepherd
Second-Order Methods for Neural Networks
3-540-76100-4

Dimitris C. Dracopoulos
Evolutionary Learning Algorithms for Neural Adaptive Control
3-540-76161-6

John A. Bullinaria, David W. Glasspool and George Houghton (Eds)
4th Neural Computation and Psychology Workshop, London,
9-11 April 1997: Connectionist Representations
3-540-76208-6

Maria Marinaro and Roberto Tagliaferri (Eds)
Neural Nets - WIRN VIETRI-97
3-540-76157-8

Gustavo Deco and Dragan Obradovic
An Information-Theoretic Approach to Neural Computing
0-387-94666-7

Thomas Lindblad and Jason M. Kinser
Image Processing using Pulse-Coupled Neural Networks
3-540-76264-7

Maria Marinaro and Roberto Tagliaferri (Eds)
Neural Nets - WIRN VIETRI-98
1-85233-051-1

Amanda J.C. Sharkey (Ed.)
Combining Artificial Neural Nets
1-85233-004-X

Dirk Husmeier
Neural Networks for Conditional Probability Estimation
1-85233-095-3

Achilleas Zapranis and Apostolos-Paul Refenes
Principles of Neural Model Identification, Selection and Adequacy
1-85233-139-9

Mark Girolami
Self-Organising Neural Networks
1-85233-066-X

Walter J. Freeman
Neurodynamics: An Exploration in Mesoscopic Brain Dynamics
1-85233-616-1

Paulo J.G. Lisboa, Emmanuel C. Ifeachor and Piotr S. Szczepaniak (Eds)
Artificial Neural Networks in Biomedicine
1-85233-005-8

H. Malmgren, M. Borga and L. Niklasson (Eds)

Artificial Neural Networks in Medicine and Biology

Proceedings of the ANNIMAB-1 Conference, Göteborg, Sweden, 13-16 May 2000

Springer

Helge Malmgren, BA, PhD, MD
Department of Philosophy, Göteborg University, Box 200, SE-405 30 Göteborg, Sweden

Magnus Borga, MSc, PhD
Department of Electrical Engineering, Linköping University, SE-585 83 Linköping, Sweden

Lars Niklasson, BSc, MSc, PhD
Department of Computer Science, University of Skövde, PO Box 408, SE-541 28 Skövde, Sweden

Series Editor
J.G. Taylor, BA, BSc, MA, PhD, FInstP
Centre for Neural Networks, Department of Mathematics, King's College,
Strand, London WC2R 2LS, UK

ISSN 1431-6854

ISBN 978-1-85233-289-1

British Library Cataloguing in Publication Data
Artificial neural networks in medicine and biology :
Proceedings of the ANNIMAB-1 Conference, Göteborg, Sweden,
13-16 May 2000. - (Perspectives in neural computing)
1. Neural networks (Computer science) 2. Artificial
intelligence - Medical applications 3. Artificial
intelligence - Biological applications
I. Malmgren, H. II. Borga, M. III. Niklasson, Lars F.
IV. ANNIMAB-1 (2000 : Goteborg, Sweden)
610.2'8563
ISBN 978-1-85233-289-1 ISBN 978-1-4471-0513-8 (eBook)
DOI 10.1007/978-1-4471-0513-8
Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Originally published by Springer-Verlag London Berlin Heidelberg in 2000

Typesetting: Camera ready by editors

34/3830-543210 Printed on acid-free paper SPIN 10761446

Preface

This book contains the proceedings of the conference ANNIMAB-1, held 13-16 May 2000 in Göteborg, Sweden. The conference was organized by the Society for Artificial Neural Networks in Medicine and Biology (ANNIMAB-S), which was established to promote research within a new and genuinely cross-disciplinary field. Forty-two contributions were accepted for presentation; in addition to these, 8 invited papers are also included.

Research within medicine and biology has often been characterised by application of statistical methods for evaluating domain specific data. The growing interest in Artificial Neural Networks has not only introduced new methods for data analysis, but also opened up for development of new models of biological and ecological systems. The ANNIMAB-1 conference is focusing on some of the many uses of artificial neural networks with relevance for medicine and biology, specifically:

- Medical applications of artificial neural networks: for better diagnoses and outcome predictions from clinical and laboratory data, in the processing of ECG and EEG signals, in medical image analysis, etc. More than half of the contributions address such clinically oriented issues.
- Uses of ANNs in biology outside clinical medicine: for example, in models of ecology and evolution, for data analysis in molecular biology, and (of course) in models of animal and human nervous systems and their capabilities.
- Theoretical aspects: recent developments in learning algorithms, ANNs in relation to expert systems and to traditional statistical procedures, hybrid systems and integrative approaches.

We would like to acknowledge the cooperation of a number of individuals and organisations, who have contributed to making the ANNIMAB conference and this publication possible.

- The Department of Philosophy, Göteborg University, has graciously accepted to take the financial risk for the entire project. The Foundation for Knowledge and Competence Development (KK-stiftelsen) is financially supporting both the ANNIMAB-S and the conference. Economic support for the conference has also been obtained from the Faculty of Arts, Göteborg University, and from the Swedish Council for Research in the Humanities and Social Sciences (HSFR).
- The organising committee has through numerous meetings made an invaluable effort to create an interesting and well-balanced program with several related topics, well covered by invited and submitted papers.
- We are of course indebted to the program committee and all the individual reviewers for guaranteeing the scientific quality of the conference.

- A special acknowledgement is in order to Daniel Ruhe for all his technical support, formatting checks and (maybe most noticeable) design of web pages and the cover of the proceedings. We also want to thank ARC Science Simulations for allowing us to use the copyrighted image of the earth on this cover.

We trust that readers interested in artificial neural networks, medicine or biology will find this book to be very exciting and of high quality.

Göteborg, February 4, 2000

Helge Malmgren
Magnus Borga
Lars Niklasson

ANNIMAB-1 Organising Committee

Magnus Borga
Björn Haglund
Noél Holmgren
Lars Lindström
Helge Malmgren
Filip Radovic
Martin Rydmark
Torben Kling-Petersen
Lars Niklasson
Erik Olsson
Daniel Ruhe
Christer Svennerlind
Holger Wigström

Program Committee

Rod Adams
Charles W. Anderson
Soren Brunak
Bo Cartling
Rich Caruana
Richard Dybowski
Peter Erdi
Laurene V. Fausett
Bernard Fertil
Wulfram Gerstner
Roy Goodacre
Stephen Grossberg
Torgny Groth
Juha Karhunen
Samuel Kaski
Hans Knutsson
Paulo Lisboa
Giovanni Magenes
Rainer Malaka
Hanspeter A. Mallot
Pietro G. Morasso
Ian Nabney
Mattias Ohlsson
Steve Phelps
Tomaso Poggio
John G Taylor

Contents

Invited Presentations

Medical Image Analysis

Signal Processing in Medicine

Clinical Diagnosis and Medical Decision Support

Biomolecular Applications and Biological Modelling

Learning Methods and Hybrid Algorithms

Invited Presentations

Protein β-Sheet Partner Prediction by Neural Networks

Pierre Baldi* and Gianluca Pollastri
Department of Information and Computer Science, University of California Irvine
Irvine, CA 92697-3425

Claus A. F. Andersen and Søren Brunak
Center for Biological Sequence Analysis, The Technical University of Denmark
DK-2800 Lyngby, Denmark

Abstract

Predicting the secondary structure (α-helices, β-sheets, coils) of proteins is an important step towards understanding their three dimensional conformations. Unlike α-helices that are built up from one contiguous region of the polypeptide chain, β-sheets are more complex resulting from a combination several disjoint regions. The exact nature of these long distance interactions remains unclear. Here we introduce a neural-network based method for the prediction of amino acid partners in parallel as well as anti-parallel β-sheets. The neural architecture predicts whether two residues located at the center of two distant windows are paired or not in a β-sheet structure. The distance between the windows is a third essential input into the architecture. Variations on this architecture are trained using a large corpus of curated data. Prediction on both coupled and non-coupled residues currently exceeds 83% accuracy, well above any previously reported method. Unlike standard secondary structure prediction methods, the use of multiple alignment (profiles) in our case seems to degrade the performance, probably as a result of intra-chain correlation effects.

1 Background

Predicting the secondary structure (α-helices, β-sheets, coils) of proteins is an important step towards understanding their three dimensional conformations. Unlike α-helices that are built up from one contiguous region of the polypeptide chain, β-sheets are built up from a combination of several disjoint regions. These regions, or β strands are typically 5-10 residues long. In the folded protein, these strands are aligned adjacent to each other in parallel or anti-parallel fashion. Hydrogen bonds can form between C'O groups of one strand and NH groups on the adjacent strand and vice versa with C_α atoms successively a little above or below the plane of the sheet. Hydrogen bonds between parallel and anti-parallel strands have distinctive patterns, but the exact nature and behavior of β-sheet long-ranged interactions is not clear.

While the majority of sheets seems to consist of either parallel or antiparallel strands, mixed sheets are not uncommon. A β-strand can have 1 or 2 partner strands,

*and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed (*pfbaldi@ics.uci.edu*).

H. Malmgren et al. (eds.), *Artificial Neural Networks in Medicine and Biology*

and an individual amino acid can have 0,1 or 2 hydrogen bonds with one or two residues in a partner strand. Sometimes one or several partner-less residues are found in a strand, giving rise to the so-called β-bulges. Finally, β-strand partners are often located on a different protein chain. How amino acids located far apart in the sequence find one another to form β-sheets is still poorly understood, as is the degree of specificity between side-chain/side-chain interactions between residues on neighboring strands, which seems to be very weak [13]. The presence of a turn between strands is also an essential ingredient.

Partly as a result of the exponentially growing amount of available 3D data, machine learning methods have in general been among the most successful in secondary structure prediction [2]. The best existing methods for predicting protein secondary structure, i.e. for classifying amino acids in a chain in one of the three classes, achieve prediction accuracy in the 75-77% range [3, 4, 8]. Therefore any improvement in β-sheet prediction is significant as a stand-alone result, but also in relation to secondary and tertiary structure prediction methods in general. Here we design and train a neural network architecture for the prediction of amino acid partners in β-sheets (see also [7, 15]).

2 Data Preparation

2.1 Selecting the Data

As always the case in machine learning approaches, the starting point is the construction of a well-curated data set. The data set used here consists of 826 protein chains from the PDB select list of June 1998 [5] (several chains were removed since DSSP could not run on them). All the selected chains have less than 25% sequence identity using the Abagyan-function [1]. The selection has been performed by applying the all against all Huang-Miller sequence alignment using the "sim" algorithm [6], where the chains had been sorted according to their quality (i.e. resolution plus R-factor/20 for X-ray and 99 for NMR).

2.2 Assigning β-sheet Partners

The β-sheets are assigned using Kabsch and Sander's DSSP program [9], which specifies where the extended β-sheets are situated and how they are connected. This is based on the intra-backbone H-bonds forming the sheet according to the Pauling pairing rules [11]. An H-bond is assigned if the Coulomb binding energy is below -0.5 kcal/mol. In wild-type proteins there are many deviations from Paulings ideal binding pattern, so Kabsch and Sander have implemented the following rules: a β-sheet ('E') amino acid is defined when it forms two H-bonds in the sheet or is surrounded by two H-bonds in the sheet. The minimal sheet is two amino acids long; if only one amino acid fulfills the criteria, then it is called β-bridge ('B'). Bulges in sheets are also assigned 'E' if they are surrounded by normal sheet residues of the same type (parallel or anti-parallel) and comprise at most 4 and 1 residue(s) in the two backbone partner segments, respectively.

A standard example of how the partner assignments are made is shown in figure 1. In the case of β-bridges the same rules are followed, while in the special case of β-bulge residues then no partner is assigned.

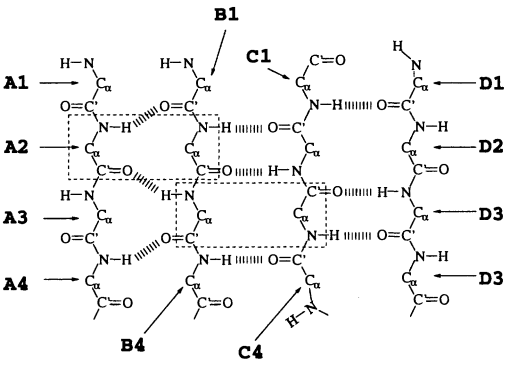

Figure 1: The assignment criteria for sheet partners are shown for two examples by the dashed boxes. That is the A sheet segment binds to the B sheet segment with a parallel sheet and residue A2 is the partner of B2. The other dashed box shows that B3 is the partner of C3, even though none of them has H-bonds in the anti-parallel B-C sheet. The other sheet partners in the example shown are: A3-B3, B2-C2, C2-D2 and C3-D3. Note that the residues A1,A4,B1,B4,C1,C4,D1,C4 are not sheet residues.

3 Neural Network Architecture

A number of different artificial neural network approaches can be considered. Because of the long-ranged interactions involved in beta-sheets, neural architectures must have either very large input windows or distant shorter windows. Very large input windows lead to architectures with many parameters which are potentially prone to overfitting, especially with sparse amino acid input encoding. Overfitting, however, is not necessarily the main obstacle because data is becoming abundant and techniques, such as weight sharing, can be used to mitigate the risk. Perhaps the main obstacle associated with large input windows is that they tend to dilute sparse information present in the input that is really relevant for the prediction [10].

Here we have used a basic two-windows approach. Since the distance between the windows plays a key role in the prediction, one can either provide the distance information as a third input to the system or one can train a different architecture for each distance type. Here, we use the first strategy with the neural network architecture depicted in Figure 2 (see also [12]). The architecture has two input windows of length W corresponding to two amino acid substrings in a given chain. The goal of the architecture is to output a probability reflecting whether the two amino acids located at the center of each window are partners or not. The sequence separation between the windows, measured by the number D of amino acids, is essential for the prediction and is also given as an input unit to the architecture with scaled activity $D/100$. As in other

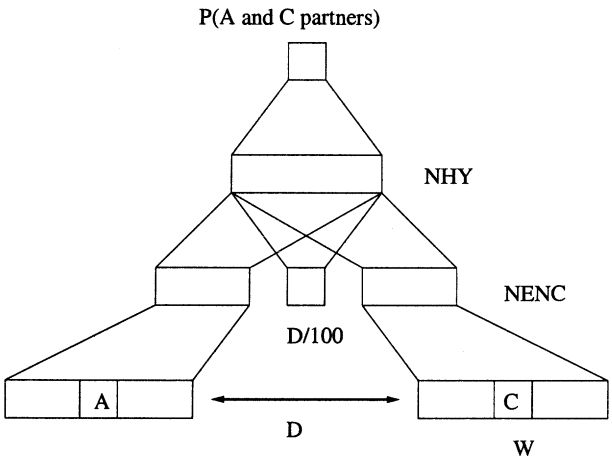

Figure 2: Neural network architecture for amino acid β-partner prediction.

standard secondary structure prediction architectures, we use sparse encoding for the 20 amino acids. Each input window is post-processed by a hidden layer comprising a number NENC of hidden units. Information coming from the input windows and the distance between the windows are combined in a fully interconnected hidden layer of size NHY. This layer is finally connected to a single logistic output unit that estimates the partnership probability. The architecture is trained by back-propagation on the relative entropy between the output and target probability distributions.

4 Experiments and Results

For training, we randomly split the data 2/3 for training and 1/3 for testing purposes. A typical split gives:

Table 1: Training set statistics, with number of sequences, amino acids, and positive and negative examples.

	Training set	Test set
Sequences	551	275
Amino acids	129119	64017
Positive ex.	37008	18198
Negative ex.	44,032,700	22,920,100

The number of negative examples (pairs of amino acids that are not partners) is of course much higher. In order to have balanced training, at each epoch we present all the 37008 positive examples, together with 37008 randomly selected negative examples at each epoch. We use a hybrid between on-line and batch training, with 50 batch