

The Editorial Policy for Proceedings

The series Lecture Notes in Physics reports new developments in physical research and teaching - quickly, informally, and at a high level. The proceedings to be considered for publication in this series should be limited to only a few areas of research, and these should be closely related to each other. The contributions should be of a high standard and should avoid lengthy rehashings of papers already published or about to be published elsewhere. As a whole, the proceedings should aim for a balanced presentation of the theme of the conference including a description of the techniques used and enough motivation for a broad readership. It should not be assumed that the published proceedings must reflect the conference in its entirety. (A listing or abstract of papers presented at the meeting but not included in the proceedings could be added as an appendix.) When applying for publication in the series Lecture Notes in Physics the volume's editor(s) should submit sufficient material to enable the series editors and their referees to make a fairly accurate evaluation (e.g. a complete list of speakers and titles of papers to be presented and abstracts). If, based on this information, the proceedings are tentatively accepted, the volume's editor(s) whose name(s) will appear on the title pages, should select the papers suitable for publication and have them refereed (as for a journal) when appropriate. As a rule discussions will not be accepted. The series editors and Springer-Verlag will normally not interfere with the detailed editing except in fairly obvious cases or on technical matters.

Final acceptance is expressed by the series editor in charge in consultation with Springer-Verlag only after receiving the complete manuscript. It might help to send a copy of the authors' manuscripts in advance to the editor in charge to discuss possible revisions with him. As a general rule, the series editor will confirm his tentative acceptance if the final manuscript corresponds to the original concept discussed. If the quality of the contribution meets the requirements of the series, and if the final size of the manuscript does not greatly exceed the number of pages originally agreed upon, the manuscript should be forwarded to Springer-Verlag shortly after the meeting. In cases of extreme delay (more than six months after the conference) the series editors will check once more the timeliness of the papers. Therefore, the volume's editor(s) should establish strict deadlines, or collect the articles during the conference and have them revised on the spot. If a delay is unavoidable, one should encourage the authors to update their contributions if appropriate. The editors of proceedings are strongly advised to inform contributors about these points at an early stage.

The final manuscript should contain a table of contents and an informative introduction accessible also to readers not particularly familiar with the topic of the conference. The contributions should be in English. The volume's editor(s) should check the contributions for the correct use of language. At Springer-Verlag only the referees will be checked by a copy-editor for language and style. Grammatical or technical shortcomings may lead to the rejection of contributions by the series editors. A conference report should not exceed a total of 500 pages. Keeping the size within this bound should be achieved by a careful selection of articles and not by imposing an upper limit to the length of the individual papers. Editors receive jointly 50 complimentary copies of their book. They are entitled to purchase further copies of their book at a reduced rate. As a rule no reprints of individual contributions can be supplied. No royalty is paid on Lecture Notes in Physics volumes. Commitment to publish is made by letter of interest rather than by signing a formal contract; Springer-Verlag secures the copyright for each volume.

The Production Process

The books are hardbound, and the publisher will select quality paper appropriate to the needs of the author(s). Publication time is about ten weeks. More than twenty years of experience guarantee authors the best possible service. To reach the goal of rapid publication at a low price the technique of photographic reproduction from a camera-ready manuscript was chosen. This process shifts the main responsibility for the technical quality considerably from the publisher to the authors. We therefore urge all authors and editors of proceedings to observe very carefully the essentials for the preparation of camera-ready manuscripts, which we will supply on request. This applies especially to the quality of figures and halftones submitted for publication. In addition, it might be useful to look at some of the volumes already published. As a special service, we offer free of charge *AT&T* and *TeX* macro packages to format the text according to Springer-Verlag's quality requirements. We strongly recommend that you make use of this offer, since the result will be a book of considerably improved technical quality. To avoid mistakes and time-consuming correspondence during the production period the conference editors should request special instructions from the publisher well before the beginning of the conference. Manuscripts not meeting the technical standard of the series will have to be returned for improvement.

For further information please contact Springer-Verlag, Physics Editorial Department II, Tiergartenstrasse 17, D-69121 Heidelberg, Germany.

John W. Clark Thomas Lindenau
Manfred L. Risting (Eds.)

Scientific Applications of Neural Nets

Proceedings of the 194th W.E. Heraeus Seminar
Held at Bad Honnef, Germany, 11-13 May 1998



Springer

A handwritten signature in black ink, appearing to be 'M. L. Risting', with the date '1998' written below it.

Editors

John W. Clark
Department of Physics
Washington University
St. Louis, MO 63130, USA

Thomas Lindenau
Manfred L. Ristig
Institut für Theoretische Physik
Universität zu Köln
D-50937 Köln, Germany

Library of Congress Cataloging-in-Publication Data

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Scientific applications of neural nets : proceedings of the 194th W. E. Heraeus Seminar, held at Bad Honnef, Germany, 11 - 13 May 1998 / John W. Clark ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 1999
(Lecture notes in physics ; Vol. 522)
ISBN 3-540-65737-1

ISSN 0075-8450
ISBN 3-540-65737-1 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1999
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by the authors/editors
Cover design: design & production, Heidelberg

SPIN: 10644319 55/3144 - 5 4 3 2 1 0 - Printed on acid-free paper

Preface

The chapters of this volume are based on invited reviews presented at the 194th W. E. Heraeus Seminar, which was devoted to the topic "Scientific Applications of Neural Nets." The workshop was organized to foster communication between scientists who are active in this highly interdisciplinary research area and have made important contributions to the development and implementation of neural-network algorithms suited to the analysis and solution of problems at the leading edge of science.

Recent years have seen a rapid expansion of neural-network applications in many areas of science, including physics, astronomy, geoscience, chemistry, biology, and linguistics. A growing list of interesting examples includes the deployment of neural-network models for event analysis in experimental high-energy physics; star/galaxy discrimination; control of adaptive optical systems; prediction of nuclear properties; fast interpolation of potential energy surfaces in chemistry; classification of mass spectra of organic compounds; protein-structure prediction; analysis of DNA sequences; and design of pharmaceuticals. This intense activity has produced new insights into regularities and mechanisms as well as substantial progress toward the creation of quantitative and reliable predictive tools. At the Heraeus Seminar, speakers discussed applications based on feedforward connectionist systems taught by example, recurrent attractor nets, and self-organizing feature maps, as well as hybrid systems. The presentations ranged over a spectrum of fields:

- Astronomy - Adaptive optical systems have been invented to improve the clarity of images formed by ground-based telescopes from light passing through the turbulent atmosphere of the earth. Neural networks are used to monitor and instruct the operation of these systems by wave-front sensing, reconstruction, and prediction, with the goal of approaching the diffraction limit of resolution.
- Nuclear physics - The production of a host of new nuclei at radioactive-beam facilities has stimulated interest in predictive global models of nuclear properties. As a promising alternative to traditional theory-rich modeling, feedforward neural nets and higher-order probabilistic perceptrons are being applied to statistical analysis of the existing nuclear database. In recent work, notable successes in the development of neural-network predictions have been recorded in the modeling of atomic masses,

Evaluation Teaches Neural Networks to Predict Protein Structure

Burkhard Rost

LION Bioscience AG, Im Neuenheimer Feld 517, 69120 Heidelberg, Germany,
Columbia, Dep. of Biochem. and Molecular Biophys., 630 West 168th Str, New
York, N.Y. 10032, USA,

EMBL, 69 012 Heidelberg, Germany; rost@embl-heidelberg.de.

<http://www.embl-heidelberg.de/~rost/>

Abstract. In the wake of the genome data flow, we need - more urgently than ever - accurate tools to predict protein structure. The problem of predicting protein structure from sequence remains fundamentally unsolved despite more than three decades of intensive research effort. However, the wealth of evolutionary information deposited in current databases enabled a significant improvement for methods predicting protein structure in 1D: secondary structure, transmembrane helices, and solvent accessibility. In particular, the combination of evolutionary information with neural networks proved extremely successful. The new generation of prediction methods proved to be accurate and reliable enough to be useful in genome analysis, and in experimental structure determination. Moreover, the new generation of theoretical methods is increasingly influencing experiments in molecular biology. Neural networks have been applied to many pattern classification problems. Here, I reviewed applications to the problem of predicting protein structure from protein sequence. Initially, methods were designed as a 'quick and dirty' demonstration that artificial intelligence-based machines could solve real-life problems. At that stage, biologists typically reached higher levels of accuracy when using their expertise than computer scientists when using their machines. However, more thorough investigations enabled to include the information used by experts into neural network-based tools. Now, some tools are - on average - as accurate as the best experts; and experts using such tools often arrive at even more accurate predictions. Thus, several neural network-based methods have eventually contributed significantly to advancing the field of bio-informatics, and some are clearly influencing molecular biology.

Key words: protein structure prediction, evolution, neural networks.

1 Introduction

Proteins constitute life's machinery. The first bacterial genome was sequenced in 1995 [1]; the first eukaryote (yeast) followed in 1996 [2]. Meanwhile, more than ten other genomes have been published [3], and the human genome (200-times larger than yeast) is expected to be sequenced as one of the first milestones in the next millennium. Why bother? Since genomes contain the blueprint for all parts of life's machinery. The machinery itself consists of proteins that perform all important tasks in organisms (catalysis of biochemical

reactions, transport of nutrients, recognition, and transmission of signals). Proteins are formed by joining 20 different amino acids (dubbed residues, when joined in proteins) into a stretched chain. In water, the chain folds into a unique three-dimensional (3D) structure (Fig. 1; introduction to protein structure: [4]).

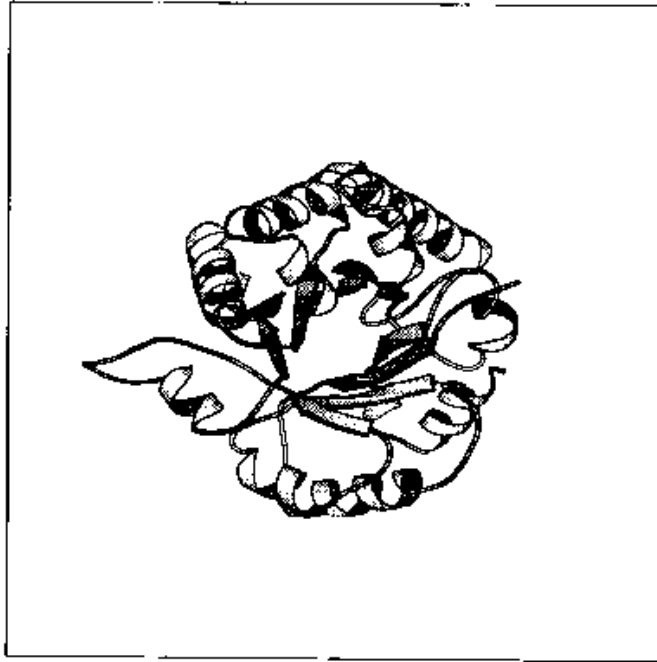


Fig. 1. Representation of the 3D structure of *Leishmania mexicana* triose phosphate isomerase (TIM, Protein Data Bank [58] code 1amk). The trace of the protein chain in 3D is plotted schematically as a ribbon. Strands are indicated by arrows (yellow), helices by open coiled-tubes (red). Graph made with RASMOI. (Roger Sayle, ras@32425@ggr.co.uk), and Molscript (Peer Kraulis, <http://www.avatar.se/molscript>, [60]). The TIM-barrel is named after the barrel formed by the strands in the centre of the molecule. The enzyme is found with a similar structure in most of the known life-forms, and thus represents a billion year's old perspective at the complexity of the shapes of life.

“ *Sequence determines structure determines function.* The world of proteins is governed by shape: interactions between proteins are mediated by the 'key-hole' principle, i.e., two proteins interact when they fit to one another like a

key into a hole. Thus, protein structure determines protein function. What determines structure? All information about the native structure of a protein is coded in the amino acid sequence, plus its native solution environment [5]. Can we decipher the code, i.e., can we predict 3D structure from sequence? In principle, we could; in practice, such approaches are frustrated by the difficulty of the task resulting from the high complexity of protein structure formation [6]. For over 40 years, there has been an ardent search for methods predicting protein structure from sequence (reviews: [7,8,6]; books: [9]). Many methods were found which looked initially very promising - but always the hope has been dashed [10]. The most successful predictions are achieved by experts starting combining machine-based predictions with their intuition and expertise [11].

How can neural networks predict protein structure? In practice, the most successful structure predictions extract patterns from data bases of known protein structures. Neural networks comprise a particular tool for pattern recognition and classification [12,13]. To which extent do neural networks contribute to predicting protein structure, in practice? Initially, researchers applied black boxes, and searched improvements through optimising the internal free parameters (training speed, network architecture). Later, researchers have opened the black boxes by extracting, or implementing rules, by carving specific knowledge into the networks, and by using networks to detect errors or outliers in data bases. More recently, the full potential of the tool has been explored by combining neural networks with evolutionary information. Now, applications of neural networks are amongst the most widely used methods in everyday's bioinformatics.

Here, I sketched neural network based methods (PHD series) for the prediction of 1D aspects (secondary structure, transmembrane helices, solvent accessibility) of protein structure. The methods illustrated that (1) neural networks as black-boxes failed to improve prediction accuracy, (2) neural networks were sufficiently flexible to carve expertise from biology into the tool, (3) the quantum leap in prediction accuracy achieved in the 90's unearthed from implementing evolutionary information into neural networks, (4) and that the new generation of prediction methods is extremely useful in assisting, facilitating, and speeding-up experiments in molecular biology.

2 Carving Biology into Neural Networks

2.1 Conventional Prediction of Secondary Structure

Simplifying the structure prediction problem: The rapidly growing sequence-structure gap (number of known protein structures vs. number of known protein sequences) has enticed theoreticians to solve simplified prediction problems [8]. An extreme simplification is the prediction of protein structure in one dimension (1D), as represented by strings of, e.g., secondary structure, and residue solvent accessibility. Theoreticians are lucky not only because the

1D prediction problem is not only the task they can accomplish best, but in that even partially correct predictions of 1D structure are useful, e.g., for predicting protein function, or functional sites.

Basic idea of secondary structure prediction: The usual goal of secondary structure prediction methods is to classify a pattern of adjacent residues as either H (a-helix), E (for extended b-strand), or L (for loop, i.e., no regular structure). The principal idea underlying most secondary structure prediction methods is the fact that segments of consecutive residues have preferences for certain secondary structure states [4,14]. Thus, the prediction problem becomes a pattern-classification problem tractable by pattern recognition algorithms. The goal is to predict whether the residue at the centre of a segment of typically 13-21 adjacent residues is in a helix a strand, or in none of the two regular structures.

First and second generation prediction methods: The first generation of 1D prediction methods was based on physico-chemical principles, expert rules, and statistics of single residues [15-17,4]. The second generation incorporated the influence of residues adjacent to the residue for which 1D structure was predicted (local information). These secondary structure prediction methods shared three major shortcomings: (1) prediction accuracy was limited to about 60% accuracy (percentage of residues predicted correctly in either of the three states H, E, L), (2) strands were predicted at typically < 40% accuracy, (3) predicted secondary structure segments were, on average, only half as long as observed segments. Methods were tailored to overcome one of these problems (long-range information: [18,19]; strand accuracy: [20]; length: [21]). However, the basic assumption was that these problems originated from using only local information (13-21 adjacent residues). It was assumed that, in general, 65% of the secondary structure formation is determined by local interactions, and that strands are dominantly determined by long-range interactions [22].

2.2 Improving Secondary Prediction by Neural Networks

No improvement by simple network: A simple tool that classified sequence stretches into three secondary structure states was a neural network (more precisely a multi-layered feed-forward network) [23-25]. Input was the sequence vector composed of 13-21 residues; output the secondary structure state of the central residue (Fig. 2). However, this simple device was not better than any other good prediction method. In particular, none of the three problems (prediction accuracy limited to 60%, strand accuracy around 40%, short segments) of conventional methods could be solved by such a device [26]. (However, due to inappropriate choices of the test sets this was not revealed by the first publications [27].)

Better prediction of strand by balanced training: Prediction accuracy for each of the three secondary structure classes approximately mirrored the observed occurrence of these classes in the training set [28,14]. In particular,

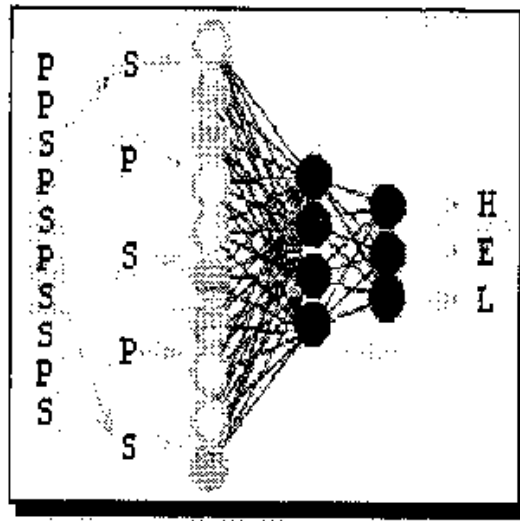


Fig. 2. Simple neural network for secondary structure prediction. For simplification the protein sequence given consists of two amino acid types (S and P). The protein sequence is translated into patterns by shifting a window of w adjacent residues (shown $w = 5$; typical values in practice are $w = 13-21$) through the protein. The output of the network is uniquely determined. Suppose the output would be: 0.2, 0.4, 0.5 for the three output states (H, E, L). For known examples the desired output is also known (1, 0, 0 if the central residue is in a helix). Consequently, the network error is given by the difference between actual network output and desired output. The only free variables are the connections. Training or learning means changing the connections such that the error decreases for the given examples. A training set typically comprises some 30,000 examples. If training is successful, the patterns are correctly classified.

only 21% of the correctly predicted residues belonged to the class E. Looking at the training dynamics of the network revealed that the network learned H, and L ten times faster than E. Consequently, the idea was to improve the prediction for strand residues by simply increasing the frequency in presenting strand residues during training. Thus, instead of presenting in 1000 iteration time steps 220 examples for E, 310 for H and 470 for L (according to database distribution, dubbed unbalanced training), now at each time step one example for each class was used for training (balanced training). (1) All three classes were predicted almost equally well [28]. (2) Overall accuracy decreased, as the loop residues that were predicted more accurately by the unbalanced network comprised almost 50% of all residues. However, a balanced network

proved that the inferior prediction of strand did NOT result primarily from long-range interactions, but from a technical problem.

Better prediction of segment length by 2nd level network: The average length of a helix is about 10 residues. However, helices predicted by the network were, on average, four residues long. The reason was that the network failed to learn correlating the secondary structure state of adjacent residues. The fact that, e.g., helices span over, at least, three residues was obscured by the particular training dynamics necessary to avoid unwanted database bias: examples presented in time steps t_1 and t_2 were chosen at random from the training set (and, thus, were usually not adjacent in sequence). This problem was corrected by introducing a 2nd level (structure-to-structure) network [28,14]. The input of this 2nd level network was the output of the 1st level (sequence-to-structure) network; the output was the secondary structure state of the central residue (Fig. 3). The 2nd level network had almost no effect in terms of overall accuracy. However, the average predicted helix extended over more than seven residues, i.e., predictions appeared considerably more protein like than for the 1st level network [28,14].

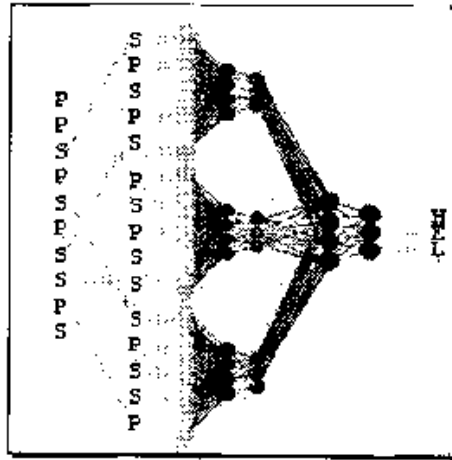


Fig. 3. Second level neural network [14]. (1) The window of w adjacent residues is shifted through the protein (here $w = 5$). For each window secondary structure is predicted for the central residue (shown three windows with central residues S, P, S). (2) The prediction of this first level network is fed into a second level network. This is again realised by shifting a window of w adjacent predictions through the protein (for the second level $w = 3$). The final prediction of secondary structure is valid for the central residue of the second window (here a P).

Better overall accuracy by averaging over many networks: Networks classify patterns separating them by lines. A particular training run results in a particular classification associated with a particular error. Part of this error usually is random noise. Furthermore, unbalanced, and balanced training occasionally yielded quite different predictions, in detail. Which to choose? The answer was to average over both networks, and to attempt reducing the random noise by generating even more differently trained networks (over $2 \times 2 = 4$ networks: 1st level: balanced, unbalanced, 2nd level: balanced, unbalanced). This 3rd level average over different networks improved prediction accuracy by 1-2 percentage points, and elegantly combined differently focused specialists [28-30].

Several problems solved, but accuracy still rather low: Incorporating facts about protein structures into the specific choice of the training dynamics and the combination of many independent neural networks solved two of the problems of conventional prediction methods (inaccurate prediction of strand, short segments). However, the overall prediction accuracy was still limited to about 65% [20,26]. Long-range information was not incorporated into the method. Increasing the window size (number of adjacent residues in protein fragment fed into the network) failed, as the signal-to-noise ratio increased considerably for longer windows. This problem was also reflected by that the networks did hardly use higher-order correlations in the input information: networks with and without hidden layers performed almost equally well [25,26].

2.3 The Issue of Appropriate Cross-Validation

Publishing optimistic results? The fact that the early applications of neural networks did not use higher order information was obfuscated by the inappropriate choices of the tests sets. Interestingly, the problem did not arise from 'computer geeks' intruding into the unknown space of biology. Rather, in the early 90's protein structure prediction experts just started to become aware of the importance of the issue of appropriately testing. Part of the problem is 'social' (better results, better journals, more grants). For example, the history of secondary structure prediction has partly been a hunt for highest accuracy scores, with over-optimistic claims by predictors seeding the scepticism of potential users. However, the CASP experiments (protein structures are predicted before the structures are known, then predictors meet every second year and assess how well they did [31,11]) have illustrated: exaggerated claims are more damaging than genuine errors. Even a prediction method of limited accuracy can be useful if the user knows what to expect. For the editors of scientific journals this implies that no protein structure prediction method should be published that has not been sufficiently cross-validated. This raises a difficult question: how to evaluate prediction methods?

Full jack-knife test: The prediction of protein structure is an excellent example to illustrate the traps of 'positive thinking'. Given a data set of N

proteins of known structure. The ideal testing is by jack-knifing: take $N - 1$ proteins for training, one for testing, and repeat N times. This recipe is simple, but does it suffice? Not, without ascertaining additional constraints. (1) No information of the performance on the one protein used for testing each time ought to steer the training. That is, developers do not want to adjust free parameters (network architecture, training speed, start conditions, training stop) such that the performance on the test protein is optimal. (2) The $N - 1$ training proteins and the one test proteins ought to be distinct. For most protein structure prediction methods the definition of 'distinct' is simple: the percentage of sequence identity between the one test protein and any of the $N - 1$ training proteins ought to be below 30%. If the two proteins have a higher level of pairwise sequence identity, we know from experience that they will adopt similar structure. Thus, we can predict structure by simply aligning the two proteins (much better and simpler than by any fancy device).

Complete cross-validation: The full jack-knife (N repeats of the experiment with splitting the N proteins of known structure into $N - 1/1$) is often prohibited by limited computer resources. A simple alternative to full jack-knifing is the complete cross-validation: the N proteins are split into $N - N/F$ training and N/F test proteins, this is repeated F times. For say $F = 10$, we refer to this as a complete ten-fold cross-validation. Again, the same constraints apply to each of the separations: (1) no information from F test proteins used for training, and (2) no overlap between training and testing set. In practice, we first have to play around with the neural networks to find a combination of network parameters for which we believe the method would be optimal. How can we thus avoid to look at what we pretend to be unknown (the structure for the testing proteins)? The solution brings forward a third data set, the validation set. Now the N proteins are split into $(N - N/F - V)$ training proteins, V validation proteins, and N/F test proteins. The V validation proteins are used to find the optimal parameters for the neural networks (and the conditions between training and testing set now apply to all three sets).

Number of cross-validation experiments not important: Are higher values of F (number of cross-validation experiments) better than lower ones? We often find an affirmative answer to this question in the literature. However, the exact number of F is not important provided the test set is representative, comprehensive and the cross-validation results are NOT miss-used to again change parameters. Developers usually have an interest to choose F as high as possible as that permits maximal use of the available information. However, the choice of F is of no meaning for the user provided the cross-validation is done appropriately.

Comprehensive data sets: All available unique proteins should be used for testing prediction methods (currently about 1000). The reason for taking as many proteins as possible is simply that proteins vary considerably in structural complexity; certain features are easy to predict, others harder. Thus,

we can easily select a large subset from the database of known structures for which any prediction method looks much more accurate than any other. This problem also raises the issue of comparing apples and oranges: no matter which data sets are used for a particular evaluation, a standard set for which results are published by others should also be included. Finally, the field of protein structure prediction offers an additional benefit: every month we witness the publication of about 50 novel protein structures. Thus, after having finalised the manuscript describing prediction methods, developers can simply apply the tool to all the new structures that were unknown by the time they started to invent their method.

3 Profiting from the Experiment Evolution

3.1 The Wealth of Evolutionary Information

Variation in sequence space: The exchange of a few residues can already destabilise a protein [32]. This implies that the majority of the $20N$ possible sequences of length N form different structures. But, has evolution created such an immense variety? Random errors in the DNA sequence lead to a different translation of protein sequences. These 'errors' are the basis for evolution. Mutations resulting in a structural change are not likely to be accepted, since the protein can no longer function appropriately. Furthermore, the universe of stable structures is not continuous: minor changes on the level of the 3D structure may destabilise the structure (due to high complexity). Thus, residue exchanges conserving structure are statistically unlikely. However, the evolutionary pressure to conserve function has led to a record of this unlikely event: structure is more conserved than sequence [33-35]. Indeed, all naturally evolved protein pairs that have 35 of 100 pairwise identical residues have similar structures [36,37]. But, the attractors of protein structures are larger, even: the majority of protein pairs of similar structures has levels of below 15% pairwise sequence identity [38,39].

Long-range information in multiple sequence alignments: The residue substitution patterns observed between proteins of a particular family, i.e., changes that conserve structure, are highly specific for the structure of that family. Furthermore, the substitutions realised by evolution, implicitly also carry information about long-range interactions: suppose residues i and $i + 100$ are close in 3D, then the types of amino acids that can be exchanged (without changing structure) at position i are constrained by that their physico-chemical characteristics have to fit the amino acid types at position $i + 100$. Indeed, correlated mutations permit to predict inter-residue contacts [40].

Feeding profiles of residue exchanges into the networks: The simplest way to use evolutionary information was as following [14]. (1) A sequence of unknown structure (U) was aligned against the database of known sequences (i.e. no information of structure required!). (2) Proteins that had significant sequence identity to U to assure structural similarity [36,37] were extracted

and re-aligned by the multiple alignment algorithm MaxHom [41]. (3) For each position the profile of residue exchanges in the final multiple alignment was compiled, and was used as input to the 1st level sequence-to-structure network.

3.2 Secondary Structure Prediction (PHDsec)

Significant improvement in overall accuracy: Using evolutionary information in the simple way described improved prediction accuracy already from about 65% to over 70%. Further incorporation of specific information compiled from the multiple alignments [14] yielded a further improvement to levels of about 72% accuracy (system dubbed PHDsec). This number represented an average over a distribution (some proteins were predicted more accurately than others), with an approximate Gaussian form, and a standard deviation of about 10% [14]. The neural network system described, here, was the first to surpass the magic line of 70% accuracy [26], and proved four years after its implementation still to be the most accurate method at the Asilomar prediction contest in 1996 [42].

Predicting prediction accuracy: I failed to distinguish proteins predicted well from those predicted poorly based on their sequence characteristics. However, the strength of the prediction (measured as the normalised difference between the output unit with the highest and the one with the next highest value) provided an extremely useful index for the reliability of the prediction for each residue [14], and for the likelihood that the prediction for the entire protein was below, or above the average of 72% [14,42]. This allows in practice to focus on regions predicted with higher reliability.

3.3 Transmembrane Helix Prediction (PHDhtm)

Important class problematic for determining 3D: Even in the optimistic scenario that in the near future most protein structures will be either experimentally determined [39], one class of proteins will still represent a challenge for experimental determination of 3D structure: transmembrane proteins. The major obstacle with these proteins is that they do not crystallise, and are hardly tractable by NMR spectroscopy. Consequently, structure prediction methods are even more needed for this protein class than for globular water-soluble proteins. Fortunately, the prediction task is simplified by strong environmental constraints on transmembrane proteins: the lipid bilayer of the membrane reduces the degrees of freedom to such an extent that 3D structure formation becomes almost a 2D problem. Two major classes of membrane proteins are known: proteins that insert helices into the lipid bilayer, and proteins that form pores by a barrel of β -strands.

Failure of PHDsec to predict transmembrane helices: The neural network system designed to predict secondary structure for globular proteins failed in predicting transmembrane helices. Hence, the networks were trained again on

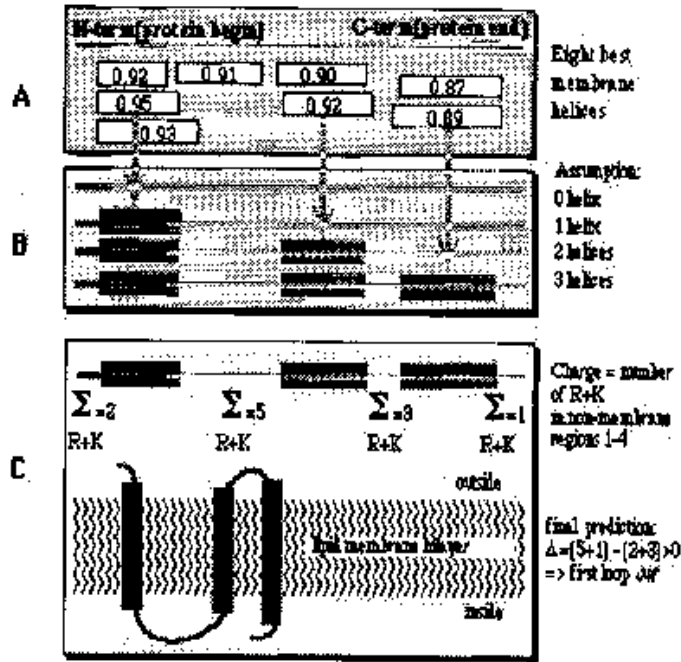


Fig. 4. Post-processing neural network output. A system of neural networks was trained to predict the location of transmembrane helices. The network output was treated as an 'energy-landscape' through which the best path was chosen given the constraint that transmembrane helices have a minimal (span of lipid bilayer) and a maximal length (longer energetically unfavourable). Thus, the normalised network outputs were used as input to a dynamic programming algorithm that found the model (number and locations of HTM's) representing the best path through all possible models consisting of HTM's between 18 and 25 residues by optimising the compatibility of the model with the neural network outputs. The final refined model output from the dynamic programming was used to apply the positive-inside rule: positively charged amino acids (Arginine: R, and Lysine: K) are more often observed inside [59]. **A**: pool of all possible membrane helices; **B**: successively building the prediction that is most compatible with the neural network output, given the assumption that the protein contains 0, 1, 2, 3 helices; **C**: final assignment of the helix orientation (topology) by the charge difference of the non-membrane regions.

proteins with transmembrane helices. Largely, the resulting prediction system (PHDhtm) was similar to the one used for predicting secondary structure for globular proteins [14]. One difference was the number of output units. PHDhtm distinguished two states: T (transmembrane helix), and non-T (i.e. a globular region). Again information from multiple alignments improved prediction accuracy significantly [14]. The final prediction system was, at least, as accurate as the best alternative prediction schemes [43,8,6].

Problems of PHDhtm: The system described so far had a major drawback: the 2nd level structure-to-structure network predicted too long membrane helices. This was corrected by introducing cut-off filters that chopped too long segments into several shorter ones. This procedure was relatively sensitive to parameter choices (when and where to cut). Furthermore, the number of transmembrane segments predicted overall was relatively often wrong.

Finding the optimal path through the network output: The problem of predicting transmembrane helices was ideal to incorporate additional aspects of globular information. This was realised by the following algorithm (Fig. 4) [44,45]. (1) The neural network (PHDhtm) output was converted to preferences. These preferences constituted an energy landscape predicted by the network. (2) The optimal path through this landscape was searched by dynamic programming. (3) The space of all possible predictions was limited by a minimal (18), and a maximal (25) length of transmembrane segments considered. (4) The final refined model was used to additionally predict the orientation of the transmembrane helices with respect to the cell (dubbed topology). (5) The average PHDhtm preference for the best transmembrane region possible was used to distinguish proteins bound to a membrane, and globular proteins.

Significant improvements by post-processing network output: The final system (PHDtopology) achieved a significantly higher performance accuracy than the simple neural network-based system (PHDhtm) [44]: for about 89% of all membrane proteins all segments were predicted correctly, and for 86% of all proteins all segments, and the topology were correctly predicted (compared to 82% for PHDhtm). Furthermore, the number of false positives (globular proteins predicted to contain membrane spanning regions) fell from above 4% to below 2%. (Note: this number is extremely important to analyse entire genomes [44].)

3.4 Solvent Accessibility Prediction (PHDacc)

Important step towards predicting SD? If secondary structure segments could be predicted sufficiently accurately, they may be arranged in space as rigid bodies to yield a model for 3D prediction [46]. One criterion for assessing each arrangement could be to use predictions of residue solvent accessibility. The solvent accessibility of a residue embedded in a protein structure can be described in several ways [47-49]. The simplest is a two-state description

distinguishing between residues that are buried (relative solvent accessibility $< 16\%$) and exposed (relative solvent accessibility $\times 16\%$). The classical method to predict accessibility is to assign either of the two states, buried or exposed, according to residue hydrophobicity [50-52].

Evolutionary information improves prediction accuracy: Solvent accessibility at each position of the protein structure is evolutionarily conserved within sequence families [53,54]. This fact was used to develop another neural network method for predicting accessibility from multiple alignment information (PHDacc) [53,14]. For this method, I skipped the 2nd level network since accessibility was hardly correlated between adjacent residues. The network output comprised ten units. Unit n , for $n = 0, \dots, 9$ coded for a relative accessibility A in the interval $n^2 \times A < (n+1)^2$. This encoding reflected the observation that in protein structures residues flip more easily between 70% and 100% relative accessibility than between 0% and 5%. The final network system predicted about 75% of the residues correctly in either of the two states buried, or exposed. This was more than five percentage points higher than for methods not using alignment information.

4 Conclusions: Do Neural Networks Help Biology?

Structure prediction: work in progress... Native 3D structures of proteins are encoded by a linear sequence of amino acid residues. To predict 3D structure from sequence is a task challenging enough to have occupied a generation of researchers. Have we finally succeeded? The bad news is: no, we still cannot predict structure for any sequence. The good news are: we have come closer, and growing databases facilitate the task.

Predictions in 1D: significant improvement by larger databases: The rich information contained in the growing sequence and structure databases enables improving the accuracy of 1D predictions. Here I sketched, how evolutionary information input to neural network systems yielded better predictions of secondary structure, solvent accessibility, and transmembrane helices. These predictions of protein structure in 1D are significantly more accurate, and more useful than five years ago.

Conditions to become useful: In the field of structure prediction we have witnessed blooming over-optimism [10], as well as, more, and less intended cheating. The Asilomar meetings [31] to some extent are succeeding in separating the chaff from the wheat. However, Asilomar does not change the basic formula: when you develop a prediction method you ought to spend more than 70% of the time on appropriate evaluation of the performance [55,8]. The sustained levels of prediction accuracy published for the PHD methods were, supposedly, one of the major reason for their success. Another important issue is that of making the method available. Molecular biologists do NOT have the time to become experts in running programs. Thus, methods should be easy-to-use, and available via the internet [56,57].

Learning from evolution to help studying evolution... Mastering protein structure prediction has several impacts on the advance of biology. Firstly, prediction methods have assisted the experimental determination of proteins structures. Secondly, predictions helped unravelling unknown protein functions, and improving our understanding of the mechanisms of partially known functions. Thirdly, prediction methods allow to separate the chaff from the wheat: we can explore which biological information improves the performance of neural networks and which doesn't. This is a simple, yet effective, means of shedding light into the understanding protein structure formation, and protein function (since we fail to predict structure and function, we thoroughly do NOT understand the underlying principles!).

What next? Most breakthroughs in protein structure prediction were achieved over the last six years. Thus, although we still cannot solve the general prediction problem, progress has been made. In general, however, we could ask the question - is it worth persevering with structure prediction, given that it is clearly such a difficult task? The answer is: yes. The methods which have spun off from structure prediction have already given us considerable insight into the first four complete genomes. Perseverance with structure prediction will yield fruit in about five years time when the human genome will be known.

Acknowledgements

Thanks - in alphabetical order - to all those who contributed ideas, and helped with motivating discussions: Michael Braxenthaler (Hoffman-LaRoche, New Jersey), Søren Brunak (CBS, Copenhagen), Rita Casadio (Univ., Bologna), Sean O'Donoghue (EMBL, Heidelberg), Piero Fariselli (Univ. Bologna), Terry Gaasterland (Univ. Chicago), Gunnar von Heijne (Univ. Stockholm), Tim Hubbard (Sanger, Hinxton), Rainer Kühnen (Univ. Heidelberg), Chris Sander (Millenium, Boston), Michael Scharf (Take5, Heidelberg), Reinhard Schneider (LION, Heidelberg), Manfred Sippl (Univ. Salzburg), Sara Solla (Western Univ., Chicago), Anna Tramantano (IRBM, Rome), Alfonso Valencia (CNB, Madrid), Gerrit Vriend (EMBL, Heidelberg). Thanks to Friedrich von Bohlen (LION, Heidelberg) for financial support.

References

1. Fleischmann, R. D., et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (1995) 490-512
2. Coffeau, A., et al.: Life with 6000 genes. *Science* 274 (1996) 546-567
3. Gaasterland, T.: Genome sequencing projects. WWW document (<http://www.mcs.anl.gov/home/gaasterl/genomes.html>), Univ. Chicago (1998)
4. Bründerlin, C., Toose, J.: *Introduction to Protein Structure*. New York, London: Garland Publ. (1991)

5. Anfinsen, C. B.: Principles that govern the folding of protein chains. *Science* 181 (1973) 223-230
6. Rost, B., O'Donoghue, S. I.: Sisyphus and prediction of protein structure. *CABIOS* 13 (1997) 345-356
7. Barton, G. J.: Protein secondary structure prediction. *Curr. Opin. Str. Biol.* 5 (1995) 372-376
8. Rost, B., Sander, C.: Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* 25 (1996) 113-136
9. Doolittle, R. F.: *Computer methods for macromolecular sequence analysis*. San Diego: Academic Press (1990)
10. Homig, B., Cohen, F. E.: Adding backbone to protein folding: why proteins are polypeptides. *Folding & Design* 1 (1996) R17-R20
11. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., Pedersen, J. T.: Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins Suppl* 1 (1997) 2-6
12. Arbib, M.: *The handbook of brain theory and neural networks*. Cambridge, MA: Bradford Books/The MIT Press (1995)
13. Fiesler, E., Beale, R.: *Handbook of Neural Computation*. New York: Oxford Univ. Press (1996)
14. Rost, B.: PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* 266 (1996) 525-539
15. Schulz, G. E., Schirmer, R. H.: *Principles of Protein Structure*. Heidelberg: Springer (1979)
16. Kabach, W., Sander, C.: How good are predictions of protein secondary structure? *FEBS Lett.* 155 (1983) 179-182
17. Fasman, G. D.: *Prediction of protein structure and the principles of protein conformation*. New York, London: Plenum (1989)
18. Maxfield, F. R., Scheraga, H. A.: Improvements in the Prediction of Protein Topography by Reduction of Statistical Errors. *Biochem.* 18 (1979) 697-704
19. Zvelebil, M. J., Barton, G. J., Taylor, W. R., Sternberg, M. J. E.: Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* 195 (1987) 957-981
20. Gascuel, O., Golmard, J. L.: A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *CABIOS* 4 (1988) 357-365
21. Kabach, W., Sander, C.: Segment83. unpublished (1983)
22. Garnier, J., Levin, J. M.: The protein structure code: what is its present status? *CABIOS* 7 (1991) 133-142
23. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Laurrup, B., Nørskov, I., Olsen, O. H., Petersen, S. B.: Protein secondary structure and homology by neural networks. *FEBS Lett.* 241 (1988) 223-228
24. Qian, N., Sejnowski, T. J.: Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* 202 (1988) 805-834
25. Holley, H. L., Karplus, M.: Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sc. U.S.A.* 86 (1989) 152-156
26. Rost, B., Sander, C.: Secondary structure prediction of all-helical proteins in two states. *Prot. Engin.* 6 (1993) 831-836

27. Rost, B., Sander, C., Schneider, R.: Progress in protein structure prediction? *TIBS* 18 (1993) 120-123
28. Rost, B., Sander, C.: Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232 (1993) 584-599
29. Rost, B., Sander, C.: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sc. U.S.A.* 90 (1993) 7558-7562
30. Rost, B., Sander, C.: Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19 (1994) 55-72
31. Moulton, J., Pedersen, J. T., Judson, R., Fidelis, K.: A large-scale experiment to assess protein structure prediction methods. *Proteins* 23 (1995) ii-iv
32. Dao-pin, S., Söderlind, E., Baase, W. A., Wozniak, J. A., Sauer, U., Matthews, B. W.: Cumulative site-directed charge-change replacements in bacteriophage T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability. *J. Mol. Biol.* 221 (1991) 873-887
33. Chothia, C., Lesk, A. M.: The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5 (1986) 823-826
34. Doolittle, R. F.: *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences.* Mill Valley California: University Science Books (1986)
35. Lesk, A. M.: *Protein Architecture - A Practical Approach.* Oxford, New York, Tokyo: Oxford University Press (1991)
36. Sander, C., Schneider, R.: Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9 (1991) 56-68
37. Rost, B.: Twilight zone of protein sequence alignments. *J. Mol. Biol.* (1998)
38. Rost, B.: Protein structures sustain evolutionary drift. *Folding & Design* 2 (1997) S19-S24
39. Rost, B.: Marrying structure and genomics. *Structure* 6 (1998) 259-263
40. Goebel, U., Sander, C., Schneider, R., Valencia, A.: Correlated mutations and residue contacts in proteins. *Proteins* 18 (1994) 309-317
41. Schneider, R.: *Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen.* Ph.D. thesis, Univ. of Heidelberg (1994)
42. Rost, B.: Better 1D predictions by experts with machines. *Proteins Suppl.* 1 (1997) 192-197
43. von Heijne, G.: Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23 (1994) 167-192
44. Rost, B., Casadio, R., Fariselli, P.: Topology prediction for helical transmembrane proteins at 86
45. Rost, B., Casadio, R., Fariselli, P.: Refining neural network predictions for helical transmembrane proteins by dynamic programming. In States, D., et al. eds. *Fourth International Conference on Intelligent Systems for Molecular Biology.* St. Louis, M.O., U.S.A.; Menlo Park, CA: AAAI Press (1996) 192-200
46. Cohen, F. E., Presnell, S. R.: The combinatorial approach. In Sternberg, M. J. E. eds. *Protein structure prediction.* Oxford: Oxford Univ. Press (1996) 207-228
47. Lee, B. K., Richards, F. M.: The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55 (1971) 379-400
48. Chothia, C.: The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105 (1976) 1-12

49. Connolly, M. L.: Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221 (1983) 709-713
50. Tanford, C.: *The hydrophobic effect: formation of micelles and biological membranes*. New York: John Wiley & Sons (1980)
51. Kyte, J., Doolittle, R. F.: A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157 (1982) 105-132
52. Eisenberg, D., Weiss, R. M., Terwilliger, T. C.: The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sc. U.S.A.* 81 (1984) 140-144
53. Rost, B., Sander, C.: Conservation and prediction of solvent accessibility in protein families. *Proteins* 20 (1994) 216-226
54. Rost, B.: Average conservation of 1D structure between remote homologues. WWW document (<http://www.embl-heidelberg.de/~rost/Res/98F-ConservationOf1D.html>), EMBL Heidelberg, Germany (1996)
55. Rost, B., Sander, C.: Progress of 1D protein structure prediction at last. *Proteins* 23 (1995) 295-300
56. Rost, B.: PredictProtein - internet prediction service. WWW document (<http://www.embl-heidelberg.de/predictprotein>), EMBL (1997)
57. Rost, B., Schneider, R.: Pedestrian guide to analysing sequence databases. In Ashman, K. eds. *Core techniques in biochemistry*. Heidelberg: Springer (1998) in press
58. Bernstein, F. C., Koetzle, T. F., Williams, G. J. R., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M.: The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112 (1977) 535-542
59. von Heijne, G.: Membrane protein structure prediction. *J. Mol. Biol.* 225 (1992) 487-494
60. Kraulis, P. J.: *J. Appl. Crystallography*, 24 (1991), 940-950.