

PROCEEDINGS OF BCEC97

BIOCOMPUTING

AND

**EMERGENT
COMPUTATION**

Skövde, Sweden

1-2 September 1997

Editors

Dan Lundh

University of Skövde, Sweden

Björn Olsson

University of Skövde, Sweden

Ajit Narayanan

University of Exeter, UK



 **World Scientific**
Singapore • New Jersey • London • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

P. O. Box 128, Farrer Road, Singapore 912805

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

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

BIOCOMPUTING AND EMERGENT COMPUTATION

Copyright © 1997 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 981-02-3262-4

Printed in the United Kingdom.

Preface

The Bio-Computing and Emergent Computation conference (BCEC97) is intended not only as a general conference at the intersection between biology and computer science, but also as a guide to future developments and applications in this interdisciplinary area. The organisers and contributors to this conference share a vision: that both computer science and biology will benefit from a cross-fertilisation between two otherwise independent areas. Recent advances in biomolecular science and gene cloning have significantly increased our understanding of the role and function of human DNA. Computer tools for searching large databases of genes and enzymes are widely available and are being added to constantly, but there is still an increasing demand for more powerful computational tools for handling and storing biomolecular information. Also, computer scientists are becoming aware that traditional biologically-inspired computational paradigms, such as artificial life, genetic algorithms, cellular automata and neural networks, for instance, need to be more solidly based on the most recent advances in biomolecular research so that more plausible models are realised.

These intertwined approaches have created even further interest and raised many questions from biologists, computer scientists and others, mainly because of the scope and depth of knowledge required to make significant progress in this area. New problem solving strategies are having to be devised, and old challenges are having to be re-addressed from new angles.

The contributions to the BCEC97 conference are unique in that they span models at the levels of atoms, molecules, proteins, organelles/cells, individuals and populations. These levels are reflected in the work of the invited speakers: Stuart Hameroff, Lila Kari, Burkhardt Rost, Anders Lansner, Rolf Eckmiller and David Fogel.

In the midst of all this excitement, we should also not lose sight of increasing concern among the general public concerning the ethical dimension of future bio-computing and emergent computing research. Several important ethical issues raised are: (a) computational tools for genetic engineering and gene cloning dehumanise us even further by identifying the mechanisms and processes which give rise to human spirit and soul; (b) just as nuclear physicists were allowed to delve into the basic units of matter and indirectly released weapons which can destroy all humanity with the push of one button, so molecular biologists, computational biologists and biological computer scientists are delving into the basic units of life and similarly, and indirectly, may release weapons of unimaginably powerful proportions; and (c) life depends on diversity, and the basic process of creating life depends on mixing genetic pools so that resulting descendants are varied in their genetic make-up — diversity is threatened by cloning, which is the process of duplicating genes rather than diversifying them. These are important issues to bear in mind as we undertake research in our laboratories.

We (the organisers of this conference) take the opportunity to thank all the

Learning From Evolution To Predict Protein Structure

Burkhard Rost

European Molecular Biology Laboratory, 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.

1 Introduction: Predicting Protein Structure

Proteins are the machinery of life: The information for life is stored by a four-letter alphabet in the genes (DNA). Proteins perform all important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, recognition, and transmission of signals. Thus, genes are the blueprints or library, and proteins are the machinery of life. Proteins are formed by joining amino acids by peptide bonds into a stretched chain. This protein sequence comprises a translation of the four-letter DNA alphabet into a 20-letter alphabet of native amino acids. Proteins differ in length (from 20 to over 40,000 amino acids), and in the arrangement of the amino acids (dubbed residues, when joined in proteins). In water, the chain folds up into a unique three-dimensional (3D) structure (i.e. the co-ordinates of all atoms). The main driving force is the need to pack residues for which a contact with water is energetically unfavourable (hydrophobic residues) into the interior of the molecule. This appears possible through the formation of a macroscopic substructure called secondary structure ([1-5]; Fig. 1; for introduction into protein structure [6]; for review on principles of folding [7]).

Sequence determines structure determines function: Protein structure determines protein function. But what determines structure? The hypothesis that structure (also referred to as 'the fold') is uniquely determined by the sequence, has been verified for many proteins [8]. Particular proteins (chaperones) often play a role in the folding pathway, and in correcting misfolds [9]. However, it is still generally assumed that the final structure is at the free-energy minimum. Thus, all information about the native structure of a protein is coded in the amino acid sequence, plus its native

solution environment. Can the code be deciphered, i.e. can 3D structure be predicted from sequence? In principle, the code could be deciphered from physico-chemical principles using, e.g., molecular dynamics [10-13]. In practice, such approaches are frustrated by two principle obstacles. Firstly, energy differences between native and unfolded proteins are extremely small (order of 1 kcal/mol). Secondly, the high complexity (i.e. co-operativity) of protein folding requires several orders of magnitudes more computing time than we anticipate to have over the next decades. Thus, the inaccuracy in experimentally determining the basic parameters, and the limited computing resources become fatal for predicting protein structure from first principles [13]. The only successful structure prediction tools are knowledge-based, using a combination of statistical theory and empirical rules.

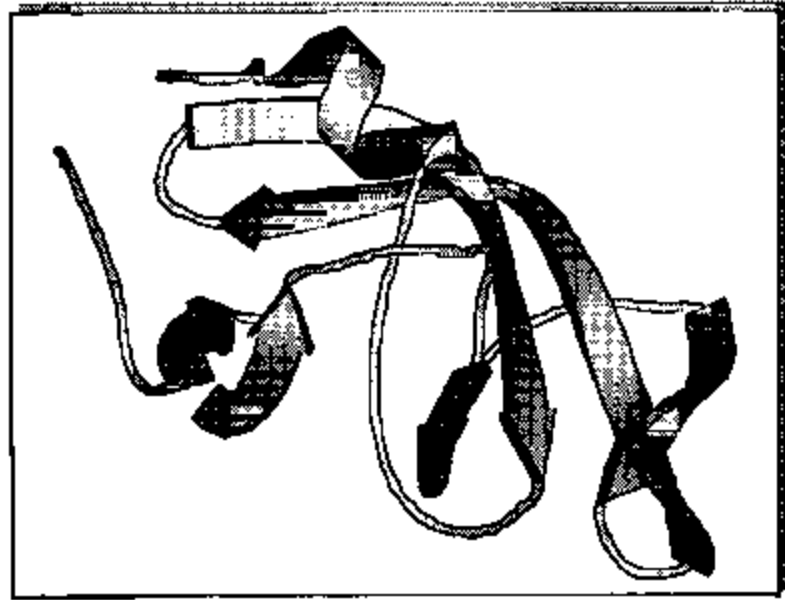


Fig. 1. Representation of HIV-1 protease monomer (Protein Data Bank code 1HHF) in 3D. The trace of the protein chain in 3D is plotted schematically as a ribbon C α -trace (alpha carbons = backbone of protein). Strands are indicated by arrows, the short helix is on the right towards the end (C-term) of the protein. Graph made with MOLSCRIPT [100].

The sequence-structure gap is rapidly increasing: Currently, databases for protein sequences (e.g. SWISS-PROT [14]) are expanding rapidly, largely due to large-scale genome sequencing projects. The first four entire genome sequences have been published; they represent all three terrestrial kingdoms: (1) prokaryotes: *haemophilus influenzae* [15], and *mycoplasma genitalium* [16]; (2) eucaryotes: yeast [17]; and (3) archaens: *methanococcus jannaschii* [18]. At least, another dozen of genomes will be completely sequenced before the end of 1997 [19, 20]; the entire human genome is likely to be known by the year 2003. This implies that the expansion of genome, and hence, protein sequences is supposedly the only field outgrowing the speed in development of computer hardware. It also implies, that despite significant improvements of structure determination techniques the gap between the number of proteins for which structure is deposited in public databases (PDB [21]), and the number of proteins for which sequences are known is increasing.

Why not simply look by microscope at the 3D structure? Unfortunately, the techniques to experimentally determine 3D structure of a protein are rather complicated. Solving a structure can take from one to several years. Consequently, 3D structure is known for less than 3% of the known protein sequences [22, 23]. The most accurate way to predict 3D structure from the sequence is by homology modelling, i.e., search for a protein with similar sequence that has a known 3D structure and then model the 3D structure of the unknown protein in analogy to the known one. Such techniques lead to a reduction of the sequence-structure gap to 10-25% [22-26].

Can the egg be unboiled? When an egg is boiled, the proteins it contains unfold. Can this procedure be reversed in theory? Can the encrypted code of protein structure be deciphered? Or, can theory help to bridge the sequence-structure gap? Indeed, for over 40 years, there has been an ardent search for methods to predict protein structure from sequence (reviews [22, 23, 27, 28]; books: [29, 30]; practical approaches and hints: [24, 31, 32]). Many methods were found which looked initially very promising - but always the hope has been dashed [5]. How well do we do in practice?

No general prediction of structure from sequence, yet: John Moult (CARB, Washington) has initiated an important, and unique experiment [33]: those who determine protein structures submitted the sequences of proteins for which they were about to solve the structure to a 'to-be-predicted' database; for each entry in that database predictors could send in their predictions before a given deadline (the public release of the structure); finally, the results were compared, and discussed during a workshop (in Asilomar, California). Two such experiments have been completed: in December 1994 (Proteins special issue, Vol. 23, 1995), and in December 1996 (to be published in Proteins, 1997). The results of both experiments demonstrated clearly that the goal to predict structure from sequence has not been reached, yet. So, no improvement despite ardent attempts, and the explosion of knowledge deposited in databases?

Here, I sketched neural network based methods (PHD series) for the prediction of 3D 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 [22]. 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 [6, 34]. 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 [6, 35-37]. The second generation incorporated the influence of residues adjacent to the residue for which 1D structure was predicted (local information) [37]. 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: [38, 39]; strand accuracy: [40]; length: [41]). 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 [42].

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) [43-45]. Input was the sequence vector composed of 13-21 residues; output the secondary structure state of the central residue. However, this simple device was not better than any other good predictor 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 [46]. (However, due to inappropriate choices of the test sets this was not revealed by the first publications [47].)

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 [34, 46, 48]. In particular, 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 [34, 46, 48]. (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.

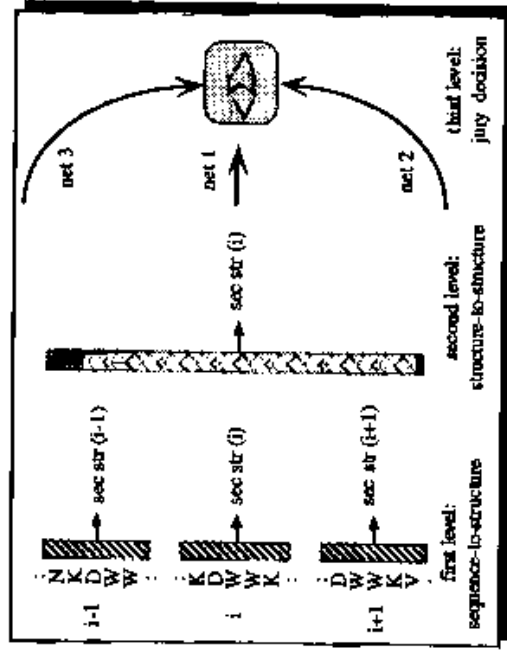


Fig. 2. Three level system for prediction of secondary structure: First level, sequence-to-structure network: a window of $a = 13$ adjacent residues was shifted through all proteins. For each window the task of the network was to predict the secondary structure state of the central residue (here: D, W, W). Second level, structure-to-structure network: a window of $a = 17$ adjacent residues was shifted through all proteins. Again the task was to predict the secondary structure for the central residue. Input were the output values, i.e., the predictions, of the 1st level net (as shown the 2nd level predicts the secondary structure for W at position i). Third level, jury decision: the output from differently trained networks for the same sequence position were summed. The secondary structure prediction for residue W at sequence position i was assigned to the unit with the maximal sum.

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 [34, 46, 48]. 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. 2). 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 [34, 46, 48].

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 [34, 46, 48].

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% [46, 48]. 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 [45, 46].

3 Profiting From The Experiment Evolution

3.1 The Wealth Of Evolutionary Information

Variation in sequence space: The exchange of a few residues can destabilise a protein [49]. This implies that the majority of the 20^N possible sequences of length N form different structures. Has evolution created such an immense variety? Random errors in the DNA 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 structural changes may destabilise the structure (due to high complexity). Thus, residue exchanges conserving structure are statistically unlikely. However, the evolutionary

pressure to conserve function has generated a record of the unlikely: structure is more conserved than sequence [50-52]. Indeed, all naturally evolved protein pairs that have 35 of 100 pairwise identical residues have similar structures [53, 54]. Even more, the majority of structurally similar protein pairs has less than 15% sequence identity [55, 56].

Long-range information in multiple sequence alignments: The residue substitution patterns observed between proteins of a particular family, i.e., changes that conserved structure, are highly specific for the structure of that family. Furthermore, the evolutionary information contained in alignments of sequence families (Fig. 3), 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 [57].

Feeding profiles of residue exchanges into the networks: The simplest way to use evolutionary information was [34]. (1) A sequence of unknown structure (U) was aligned against the database of known sequences (no information of structure required). (2) Proteins with significant sequence identity to U [53, 54] were extracted and re-aligned by the multiple alignment algorithm MaxHom [58]. (3) For each position the profile of residue exchanges in the final multiple alignment was compiled (Fig. 3), and was used as input to the 1st level sequence-to-structure network. (For training and testing the networks I used the public HSSP database containing the sequence families for all proteins of known structure [25].)

3.2 Secondary Structure Prediction (PHDsec)

Significant improvement in overall accuracy: Using evolutionary information in the simple way described, rose prediction accuracy already from about 65% to over 70%. Additional incorporation of specific information from the alignments [34] yielded a further improvement to over 72% accuracy 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% [34, 59]. The neural network system described, here, was the first to surpass the magic line of 70% accuracy [46], and proved four years after its implementation still to be the most accurate method at the Asilomar prediction contest in 1996 [60].

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 highest and the one with next highest value) provided an extremely useful index for the reliability of the prediction for each residue [34], and for the likelihood that the prediction for the entire protein was below, or above the average of 72% [34, 60]. This allows in practice to focus on regions predicted with higher reliability.

3.3 Transmembrane Helix Prediction (PHDtm)

Important class problematic for determining 3D: Even in the optimistic scenario that in the near future most structures will be experimentally determined or predicted by homology modelling, one important class of proteins will still be missing: transmembrane proteins. The major obstacle with these proteins is that they do not

crystallise, and are hardly tractable by NMR spectroscopy. Consequently, structure predictions are even more needed for this 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 for globular proteins (PHDsec) failed in predicting transmembrane helices. Hence, the networks were re-trained on transmembrane proteins. Largely, the resulting prediction system (PHDhm) was similar to the one used for predicting secondary structure for globular proteins [34]. One difference was the number of

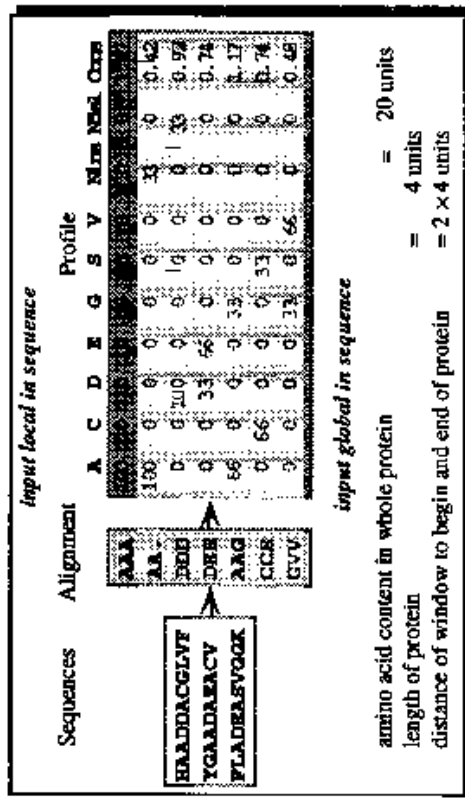


Fig. 3. Generating profiles of residue substitutions. (1) Proteins were taken from the database of known structures PDB [21]. (2) Proteins with similar sequence were searched in the SWISS-PROT database of known sequences [14]. (3) Structural homologues (here three) were selected based on the level of sequence identity [53], and were aligned by the alignment program MaxHom [58]. (4) At each residue position the occurrence (percentage) of each amino acid (given in one-letter code) was compiled along with the number of insertions (Ins) and deletions (Del) necessary to render an optimal alignment. (5) The resulting profile was input to the neural network, instead of just the sequence of the first protein. Amino acids E and D are more similar in terms of their biochemical properties than E and C. The conservation weight (Cons) reflects the degree of similarity of the residues found at a particular position of the alignment. In addition to the information locally available from 13 adjacent residues, global information was compiled: the content of each amino acid in the whole protein, the length of the protein, and the distance of the window from the begin and end of the protein.

put units. PHDhm distinguished two states: T (transmembrane helix), and non-T (i.e. a globular region). Again information from multiple alignments improved prediction accuracy significantly [34]. The final prediction system was, at least, as accurate as the best alternative prediction schemes [22, 23, 61].

3.4 Refining Neural Network Predictions (PHDtopology)

Problems of PHDhm: 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 [62, 63]. (1) The neural network (PHDhm) 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 PHDhm 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 (PHDhm) [62]: 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 PHDhm). Furthermore, the number of false positives (globular proteins predicted to contain membrane spanning regions fall from above 4% to below 2%. (Note: this number is extremely important to analyse entire genomes [62].)

3.5 Solvent Accessibility Prediction (PHDacc)

Important step towards predicting 3D? 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 [64]. One criterion for assessing each arrangement could be to use predictions of residue solvent accessibility [65, 66]. The solvent accessibility of a residue embedded in a protein structure can be described in several ways [65-67]. The simplest is a two-state description distinguishing between residues that are buried (relative solvent accessibility < 16%) and exposed (relative solvent accessibility $\geq 16\%$). The classical method to predict accessibility is to assign either of the two states, buried or exposed, according to residue hydrophobicity [68-72].

Evolutionary information improves prediction accuracy: Solvent accessibility at each position of the protein structure is evolutionarily conserved within sequence families [73, 74]. This fact was used to develop another neural network method for

predicting accessibility from multiple alignment information (PHDacc) [34, 73]. 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 $n^2 \leq A < (n+1)^2$. 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: Are Predictions Useful?

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 [5], as well as, more, and less intended cheating. The Asilomar meetings [33] 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 [22, 75]. The sustained levels of prediction accuracy published for the PHD methods were, supposedly, one of the major reasons 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 [24].

Typical applications of 1D predictions: The PHD series was the first structure prediction suite available by the internet server PredictProtein [34, 76-78]. Five years later, PredictProtein handles about 150-200 requests every day [77]. The background of users range from theoreticians who use predictions as one module for their prediction program (next paragraph) to biologists who use the predictions to investigate structure, function, and to suggest which residues to mutate in experiments [79-84]. Accurate prediction of secondary structure can also assist in X-ray diffraction (e.g., the GroEL crystal structure was derived making use of secondary structure predictions for the molecular replacement search [85]). In principle, the early stages of NMR frequency assignment could also be aided by knowledge of the secondary structure, although this has not been attempted. 1D predictions and predictions of transmembrane topology have proven to be quick and accurate enough for the analysis of entire genomes [86, 87]. The predictions of transmembrane helices provided a lower bound to approach the question of how

many proteins organisms need for, e.g., communication: the percentage of proteins with transmembrane helices has been estimated to be about 25% for yeast and haemophilus influenzae, and around 10-15% for mycoplasma genitalium and methanococcus jannaschii [62, 88]. Predictions of accessibility were used as basis for predicting functional sites [23], and to predict sub-cellular location [89].

1D predictions as input to threading techniques: Threading methods attempt to recognise similarities between protein folds in the absence of significant sequence identity [90]. The stakes are high, as most protein pairs of similar structure populate this region [55, 56], but the problem is highly non-trivial [90-92]. Recently, PHD predictions of 1D structure have been implemented successfully to develop a new generation of prediction-based threading methods [93-99]. Indeed, these methods are more successful than conventional sequence alignment techniques alone. A consequence was that most threading predictions presented at the Asilomar meeting of 1996 made use of 1D predictions from PHD (special issue of Proteins, 1997).

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 to all who contributed ideas, and motivating discussions: Søren Brunak (CBS, Copenhagen), Rita Casadio (Univ. Bologna), Sean O'Donoghue (EMBL, Heidelberg), Piero Parise (Univ. Bologna), Terry Glaserland (Univ. Chicago), Gunnar von Heijne (Univ. Stockholm), Tim Hubbard (Sanger, Hinxton), Rainer Kühnen (Univ. Heidelberg), Chris Sander (EBI, Hinxton), Michael Scharf (Take5, Heidelberg), Reinhard Schneider (LION, Heidelberg), Manfred Sipp (Univ. Salzburg), Sara Solta (Western Univ., Chicago), Anna Tramantano (IRBM, Rome), Alfonso Valencia (CNB, Madrid), Gerrit Vriend (EMBL, Heidelberg). Thanks to Chris Sander (EBI, Hinxton), and to Matti Saraste (EMBL, Heidelberg) for financial support.

References

1. Yang, A.-S., Honig, B.: Free energy determinants of secondary structure formation. 2. Antiparallel beta-sheets. *J. Mol. Biol.* 252 (1995) 366-376
2. Yang, A.-S., Honig, B.: Free energy determinants of secondary structure formation. 1. Alpha-helices. *J. Mol. Biol.* 252 (1995) 351-365
3. Sippel, M. J., Ortner, M., Jartz, M., Jäckner, P., Flöckner, H.: Helmholtz free energies of atom pair interactions in proteins. *Folding & Design* 1 (1996) 289-298
4. Sippel, M. J.: Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biol.* 260 (1996) 644-648
5. Honig, B., Cohen, F. E.: Adding backbone to protein folding: why proteins are polypeptides. *Folding & Design* 1 (1996) R17-R20
6. Brünnin, C., Toote, J.: Introduction to Protein Structure. New York, London: Garland Publ. (1991)
7. Lutman, F. E., Rose, G. D.: Protein folding-what's the question? *Proc. Natl. Acad. Sci. U.S.A.* 90 (1993) 439-441

8. Anfinsen, C. B.: Principles that govern the folding of protein chains. *Science* 181 (1973) 223-230
9. Corrales, F. J., Fersht, A. R.: Kinetic significance of GroEL14 (GroES72) complexes in molecular chaperone activity. *Folding & Design* 1 (1996) 265-273
10. Levitt, M., Warsbel, A.: Computer simulation of protein folding. *Nature* 253 (1975) 694-698
11. Shortle, D., Wang, Y., Gillespie, J., Wrabl, J. O.: Protein folding for realists: a timeless phenomenon. *Protein Sci.* 5 (1996) 991-1000
12. Karplus, M., Petsko, G. A.: Molecular dynamics simulations in biology. *Nature* 347 (1990) 631-639
13. van Gunsteren, W. F.: Molecular dynamics studies of proteins. *Curr. Opin. Str. Biol.* 3 (1993) 167-174
14. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucl. Acids Res.* 25 (1997) 31-36
15. Fleischmann, R. D., et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (1995) 496-512
16. Fraser, C. M., et al.: The minimal gene complement of *Mycoplasma genitalium*. *Science* 270 (1995) 397-403
17. Goffeau, A., et al.: Life with 6000 genes. *Science* 274 (1996) 546-567
18. Bult, C. J., et al.: Complete genome sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. *Science* 273 (1996) 1058-1073
19. Gaasterland, T.: Maggie genome sequencing project list. WWW document (<http://www.ncs.anl.gov/home/gaasterland/genomea.html>), Univ. Chicago (1997)
20. Gaasterland, T., Sengen, C.: Automated Microbial Genome Analysis. Report, Tutorial held at ISMB '96 (1996)
21. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., 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
22. Rost, B., Sander, C.: Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* 25 (1996) 113-136
23. Rost, B., O'Donoghue, S. I.: Sisyphus and prediction of protein structure. CABIOS (1997) in press
24. Rost, B., Schneider, R.: Pedestrian guide to analysing sequence databases. In Ashman, K. eds. *Core techniques in biochemistry*. Heidelberg: Springer (1997) in press
25. Schneider, R., de Darvar, A., Sander, C.: The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* 25 (1997) 226-230
26. Casari, G., De Darvar, A., Sander, C., Schneider, R.: Bioinformatics and the discovery of gene function. *Trends in Genetics* 12 (1996) 244-245
27. Rost, B., Sander, C.: Structure prediction of proteins - where are we now? *Curr. Opin. Biotech.* 5 (1994) 372-380
28. Barton, G. J.: Protein secondary structure prediction. *Curr. Opin. Str. Biol.* 5 (1995) 372-376
29. Doolittle, R. F.: Computer methods for macromolecular sequence analysis. San Diego: Academic Press (1996)
30. Sternberg, M. J. E.: Protein structure prediction. Oxford: Oxford Univ. Press (1996)
31. Rost, B., Valencia, A.: Pitfalls of protein sequence analysis. *Curr. Opin. Biotech.* 7 (1996) 457-461
32. Bork, P., Gibson, T. J.: Applying motif and profile searches. *Meth. Enzymol.* 266 (1996) 162-184
33. Moult, J., Pedersen, J. T., Hudson, R., Fidelis, K.: A large-scale experiment to assess protein structure prediction methods. *Proteins* 23 (1995) ii-iv
34. Rost, B.: PhD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* 266 (1996) 525-539
35. Kabesh, W., Sander, C.: How good are predictions of protein secondary structure? *FEBS Lett.* 155 (1983) 179-182
36. Schulz, G. E., Schirmer, R. H.: Principles of Protein Structure. Heidelberg: Springer (1979)
37. Fasman, G. D.: Prediction of protein structure and the principles of protein conformation. New York: London: Plenum (1989)
38. Maxfield, F. R., Scheraga, H. A.: Improvements in the Prediction of Protein Topography by Reduction of Statistical Errors. *Biochem.* 18 (1979) 697-704
39. 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) 957-961
40. Gascuel, O., Colman, J. L.: A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *CABIOS* 4 (1988) 357-365
41. Kabesh, W., Sander, C.: Segmen83, unpublished (1983)
42. Garnier, J., Levin, J. M.: The protein structure code: what is its present status? *CABIOS* 7 (1991) 133-142
43. Qian, N., Sejnowski, T. J.: Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.* 202 (1988) 865-884
44. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Laurup, B., Nørskov, L., Olsen, O. H., Petersen, S. B.: Protein secondary structure and topology by neural networks. *FEBS Lett.* 241 (1988) 223-228
45. Holley, H. L., Karplus, M.: Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* 86 (1989) 152-156
46. Rost, B., Sander, C.: Secondary structure prediction of all-beta proteins in two states. *Prot. Engin.* 6 (1993) 831-836
47. Rost, B., Sander, C., Schneider, R.: Progress in protein structure prediction? *TBS* 18 (1993) 120-123
48. Rost, B., Sander, C.: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 90 (1993) 7558-7562
49. Dao-pui, S., Söderlund, E., Baase, W. A., Wozniak, J. A., Sauer, U., Matthews, B. W.: Cumulative site-directed charge-change replacements in bacteriophage T4 lyozyme suggest that long-range electrostatic interactions contribute little to protein stability. *J. Mol. Biol.* 221 (1991) 873-887
50. Doolittle, R. F.: Of URFe and ORFe: a primer on how to analyze derived amino acid sequences. Mill Valley California: University Science Books (1986)
51. Chothia, C., Lesk, A. M.: The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5 (1986) 823-826
52. Lesk, A. M.: Protein Architecture - A Practical Approach. Oxford, New York, Tokyo: Oxford University Press (1991)
53. Sander, C., Schneider, R.: Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9 (1991) 56-68
54. Rost, B.: Twilight zone of protein sequence alignments. *J. Mol. Biol.* (1997)
55. Rost, B.: Protein structures sustain evolutionary drift. *Folding & Design* (1997) in press
56. Rost, B., O'Donoghue, S., Sander, C.: Protein structures evolve at random - almost. WWW document (<http://www.embl-heidelberg.de/~rost/Papers/Pre-evolution96.html>), EMBL (1996)
57. Goebel, U., Sander, C., Schneider, R., Valencia, A.: Correlated mutations and residue contacts in proteins. *Proteins* 18 (1994) 309-317
58. Schneider, R.: Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen. Ph.D. thesis, Univ. of Heidelberg (1994)

59. Rost, B.: Accuracy of predicting buried helices by PHDsc. WWW document (<https://www.embl-heidelberg.de/~rost/Res/96B-PredBuriedHelices.html>), EMBL Heidelberg, Germany (1996)
60. Rost, B.: Better ID predictions by experts with machines. *Proteins* (1997) submitted Apr 30, 1997
61. von Heijne, G.: Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23 (1994) 167-192
62. Rost, B., Casadio, R., Fariselli, P.: Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* 5 (1996) 1704-1718
63. 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
64. Colten, F. E., Frenkel, S. R.: The combinatorial approach. In Sternberg, M. J. E. eds. *Protein structure prediction*. Oxford: Oxford Univ. Press (1996) 207-228
65. Lee, B. K., Richards, F. M.: The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55 (1973) 379-400
66. Chothia, C.: The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105 (1976) 1-12
67. Connolly, M. L.: Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221 (1983) 709-713
68. Eisenberg, D., Weiss, R. M., Terwilliger, T. C.: The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* 81 (1984) 140-144
69. Kyte, J., Doolittle, R. F.: A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157 (1982) 105-132
70. Ponnuswamy, P. K., Prabhakaran, M., Manavalan, P.: Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* 623 (1980) 301-316
71. Young, L., Jernigan, R. L., Cowell, D. G.: A role for surface hydrophobicity in protein recognition. *Prot. Sci.* 3 (1994) 717-729
72. Tanford, C.: *The hydrophobic effect: formation of micelles and biological membranes*. New York: John Wiley & Sons (1980)
73. Rost, B., Sander, C.: Conservation and prediction of solvent accessibility in protein families. *Proteins* 20 (1994) 216-226
74. Rost, B.: Average conservation of 1D structure between remote homologues. WWW document (<http://www.embl-heidelberg.de/~rost/Res/96E-Conservation1D.html>), EMBL Heidelberg, Germany (1996)
75. Rost, B., Sander, C.: Progress of 1D protein structure prediction at last. *Proteins* 23 (1995) 295-300
76. Rost, B., Sander, C., Schneider, R.: PHD - an automatic server for protein secondary structure prediction. *CABIOS* 10 (1994) 53-60
77. Rost, B.: PredictProtein - internet prediction service. WWW document (<http://www.embl-heidelberg.de/predictprotein/>), EMBL (1997)
78. Rost, B., Sander, C.: Jury returns on structure prediction. *Nature* 360 (1992) 540
79. Rawlings, D. J., et al.: Mutation of unique region of Bruton's tyrosine kinase in immunodeficient XID Mice. *Science* 261 (1993) 358-361
80. Lupas, A., Kosler, A. J., Walz, J., Baummeister, W.: Predicted secondary structure of the 20S proteasome and model structure of the putative peptide channel. *FEBS Lett.* 354 (1994) 45-49
81. Hubbard, T. J. P., Park, J.: Fold recognition and ab initio structure predictions using Hidden Markov models and β -strand pair potentials. *Proteins* 23 (1995) 398-402
82. Springer, T. A.: Folding of the N-terminal, ligand-binding region of integrin α -subunits into a β -propeller domain. *Proc. Natl. Acad. Sci. U.S.A.* 94 (1997) 65-72
83. Valencia, A., Hubbard, T. J., Muga, A., Bafios, S., Llorca, O., Carrascano, J., Valpuesta, J. M.: Prediction of the structure of GroES and its interaction with GroEL. *Proteins* 22 (1995) 199-209
84. Hansen, J. E., Lund, O., Nielsen, J. O., Brunak, S., Hansen, J.-E. S.: Prediction of the secondary structure of HIV-1 gp120. *Proteins* 25 (1996) 1-11
85. Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D. C., Josciniak, A., Horwich, A. L., Sigler, P. B.: The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* 371 (1994) 578-586
86. Koonin, E. V., Tatusov, R. L., Rudd, K. E.: Protein sequence comparison at genome scale. *Meth. Enzymol.* 266 (1996) 295-322
87. Odgren, P. R., Harvie, L. W. J., Fey, E. G.: Phylogenetic occurrence of coiled coil proteins: implications for tissue structure in metazoans via a coiled coil tissue matrix. *Proteins* 24 (1996) 467-484
88. Rost, T. B.: Sneaking in genomes for helical transmembrane proteins. Talk presented at 'Distance based approaches to protein structure prediction III', Copenhagen, Denmark (manuscript in preparation), EMBL-PDG-12/96, EMBL, Heidelberg, Germany (1996)
89. Andrade, M. A., O'Donoghue, S. I., Rost, B.: Evolution carved sub-cellular location into protein surfaces. (1997) in submission
90. Sippl, M. J.: Knowledge-based potentials for proteins. *Curr. Opin. Str. Biol.* 5 (1995) 229-235
91. Sippl, M. J., Jaritz, M., Hendlich, M., Ortner, M., Lachner, P.: Applications of Knowledge Based Mean Fields in the Determination of Protein Structures. In Doniach, S. ed. *Statistical Mechanics, protein structure and protein-substrate interactions*. New York, London: Plenum Press (1994) in press
92. Lachrop, R. H.: The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Prot. Engin.* 7 (1994) 1059-1068
93. Russell, R. B., Copley, R. R., Barton, G. J.: Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* 259 (1996) 349-365
94. Rost, B.: TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures. In Rawlings, C., et al. eds. *Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, England: Menlo Park, CA: AAAI Press (1995) 314-321
95. Rost, B.: Fitting 1-D predictions into 3-D structures. In Bohr, H., Brunak, S. eds. *Protein folds: a distance based approach*. Boca Raton, Florida: CRC Press (1995) 132-151
96. Rost, B., Schneider, R., Sander, C.: Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270 (1997) 1-10
97. Fischer, D., Rice, D. W., Bowse, J. U., Eisenberg, D.: Assigning amino acid sequences to 3D protein folds. *FASEB J.* 10 (1996) 126-136
98. Fischer, D., Eisenberg, D.: Fold recognition using sequence-derived properties. *Prot. Sci.* 5 (1996) 947-955
99. Fischer, D., Elofsson, A., Rice, D., Eisenberg, D.: Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In Hunter, L., Klein, T. eds. *Pacific Symposium on Biocomputing, Hawaii, 1996*. World Scientific Publishing Co., Singapore (1996) 300-318
100. Kraulis, P.: MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24 (1991) 946-950