Proceedings of the
Twenty-Seventh Hawaii International Conference on System Sciences

Volume V:
Biotechnology Computing

Edited by Lawrence Hunter

Sponsored by:
The University of Hawaii
The University of Hawaii College of Business Administration

In cooperation with:
The IEEE Computer Society
Association for Computing Machinery

IEEE Computer Society Press
Los Alamitos, California
Washington - Brussels - Tokyo
The papers in this book comprise the proceedings of the meeting mentioned on
the cover and title page. They reflect the authors' opinions and, in the interests
of timely dissemination, are published as presented and without change. Their
inclusion in this publication does not necessarily constitute endorsement by the
editors, the IEEE Computer Society Press, or the Institute of Electrical and
Electronics Engineers, Inc.

Published by
IEEE Computer Society Press
10882 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1284

© 1994 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source.
Libraries are permitted to photocopy beyond the limits of US copyright law, for private
use of patrons, those articles in this volume that carry a code at the bottom of the first
page, provided that the per-copy fee indicated in the code is paid through the
Copyright Clearance Center, 27 Congress Street, Salem, MA 01970. For other copying,
reprint, or republication permission, write to IEEE Copyrights Manager, IEEE Service
Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

IEEE Computer Society Press Order Number:
Volume V only: 5090-02
Five Volume Set: 5030-02
Library of Congress Number 72-180444
IEEE Catalog Number 84TH0807-2
ISBN 0-8186-5090-7 (paper)
ISSN 1060-3425

Additional copies can be ordered from:

IEEE Computer Society Press
Customer Service Center
10882 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1284
Tel: (714) 821-8380
Fax: (714) 821-4641
Email: cs.orders@computer.org

IEEE Computer Society
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: (908) 981-1385
Fax: (908) 981-8667

IEEE Computer Society
13, avenue de l'Aquilon
B-1200 Bruxelles
BELGIUM
Tel: +32-2-770-2185
Fax: +32-2-770-8505

IEEE Computer Society
Doshisha Building
2-18-1 Minami-Okayama
Minato-ku, Tokyo 107
JAPAN
Tel: +81-3-3408-3118
Fax: +81-3-3408-3553

Production Editor: Penny Siloma and Mary E. Kavanaugh
Printed in the United States of America by Braun-Brumfield, Inc.

The Institute of Electrical and Electronics Engineers, Inc.
Evolution and Neural Networks: 
Protein Secondary Structure Prediction Above 71% Accuracy

Burkhard Rost, Chris Sander, and Reinhard Schneider
EMBL Heidelberg, Meyerhofstr 1, D-69117 Heidelberg, Europe

Abstract
Some 30,000 protein sequences are known. For 1,000 of these structures are experimentally known. Another 4,000 can be modeled by homology. For the remaining 25,000 sequences, the tertiary structure cannot be predicted generally from the sequence. A reduction of the problem is the projection of 3D structure onto a one-dimensional string of secondary structure assignments. Predictions in three states rate between 35% (random) and 85% (homology matching) accuracy. Here, we present an improvement of a neural network system using information about evolutionary conservation. The method achieves a sustained overall accuracy of 71.4%. A test on 45 new proteins confirms the estimated accuracy. Of practical importance is the definition of a reliability index for each residue position; e.g., about 40% of the predicted residues have an expected accuracy of 85%. The method has been made publicly available by an automatic e-mail server.

Introduction
The number of known protein sequences (30,000 SwissProt, release 25.0 [1]) is growing much faster than that of known protein structures (1,000 PDB [2]). Less than 200 of the known structures are unique [3]. This situation underscores the increasing need for theoretical predictions of structural features of proteins. Suppose, one has a sequence and wants to know as much as possible about the structure. How can theory help? Say the sequence of unknown structure is SOS. If there is a protein with a sequence similar to SOS in the database of known structures, model building by homology allows prediction of the structure of SOS with reasonable accuracy [4-10]. If not, i.e., if SOS belongs to the majority of the 80% of known sequences which do not have homologous known structure [17], there still might be a chance to model the fold. If SOS is very short, molecular dynamics could perhaps help to fold it up [18-24]. If SOS

1 Abbreviations used: 3D: three-dimensional; PDB: Protein Data Bank of known three-dimensional structures; Swissprot: database of known sequences; DSSP: database of Homology-derived Structures; PSSM: Dictionary of Secondary Structure of Protein; PSSM: Profile of Secondary Structure of Protein; PSSM: Profits network from Heidelberg; three levels of severing for the prediction of secondary structure; SOS: Sequence of unknown structure.
Table 1: Comparison of single residue scores

<table>
<thead>
<tr>
<th>method</th>
<th>N</th>
<th>N_{prot}</th>
<th>Qin</th>
<th>Qin</th>
<th>Qin</th>
<th>corr</th>
<th>corr</th>
<th>corr</th>
<th>I</th>
<th>Sov</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>24291</td>
<td>140</td>
<td>88.4</td>
<td>88</td>
<td>88</td>
<td>99</td>
<td>0.84</td>
<td>0.85</td>
<td>0.77</td>
<td>0.616</td>
</tr>
<tr>
<td>RAN</td>
<td>2162</td>
<td>94</td>
<td>77.1</td>
<td>94</td>
<td>94</td>
<td>97</td>
<td>0.54</td>
<td>0.55</td>
<td>0.54</td>
<td>0.286</td>
</tr>
<tr>
<td>PHD</td>
<td>23191</td>
<td>126</td>
<td>71.6</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.286</td>
</tr>
<tr>
<td>PHD on 5K</td>
<td>9006</td>
<td>5</td>
<td>57.2</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.286</td>
</tr>
<tr>
<td>ETH on 5K</td>
<td>9006</td>
<td>5</td>
<td>71.6</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.286</td>
</tr>
<tr>
<td>PHD on 5K</td>
<td>9006</td>
<td>5</td>
<td>71.6</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.286</td>
</tr>
</tbody>
</table>

Abbreviations for methods: PDB: A set of 140 proteins with similar 3D structure [54]. RAN: 94 pairs of largely dissimilar 3D structure [54]. PHD: Neural network system including models cross-validated on the 126 proteins of Table 2. PHD on 5K: Same network tested on the 45 proteins of Table 3. ETH on 5K: Results for 5 predictions of experts [55-58]. The proteins are: Igla; cAMP-dependent protein kinase [59]; sact; SRC tyrosine kinase [60]; sh3: SH3 domain of spectrin [61]; pkb; p85b_huma, phosphatidylinositol 3-OH kinase [62,63]; and nfe2_MOUSE, molybdenum-iron nitrogensen [64]. PHD on 5: Results for network system on the same 5 proteins.

structure between sequence pairs of similar 3D structure. A comparison of 140 protein pairs of known structure shows that these have about 88% of their residues in identical secondary structure states beta, strand, or loop [54]. Such a comparison raises interesting questions regarding the measures used to evaluate predictions. The overall three state accuracy (percentage of correctly predicted residues, usually termed Q3) has been used by most researchers in the field. More information about the quality of a prediction is contained in correlation coefficients such as the one defined by Matthews [65], or in an entropy like measure I (defined in the captions of Table 1).

1.2: Evaluating the accuracy of predicting secondary structure segments

All these numbers comparing the correctness of single residues ignore that secondary structure elements are extended objects. In practice, it is more important to predict roughly the correct placement of all elements than to predict some elements completely correctly and others completely incorrectly. A simple way to measure the correctly predicted helices and strands is to count all those which overlap at least for half of their length between prediction and observation [47, 58, 66]. For random predictions such a measure scores at 50%, which is permisibly high [54]. More informative is the fractional segment overlap Sov. For the comparison between an observed and a predicted secondary structure string, Sov is a sum over the fractions between the region for which both strings have, e.g., a H (or beta) and the region in which either of the two (or both) has a H. For example:

\[
\text{Sov} = \sum_i \frac{\text{min} \text{ov}(i) + S}{\text{max} \text{ov}(i)}
\]

where the sum is over all segments i, with \(\text{min} \text{ov}(i)\) being the length of the observed segment i, \(\text{min} \text{ov}(i)\) the overlap (upper bar), \(\text{max} \text{ov}(i)\) the spanned region (lower bar) and \(S\) an allowed deviation chosen to be \(\text{min} \text{ov}(i)\) and \(\text{len}(i)/2\). \(S\) assures that segments are counted as equal which differ only at the edges (detailed definition in [54]).

Sov is about 33% for a random prediction and about 90% for the comparison of the secondary structure for 3D similar protein pairs (Table 1). A reasonable definition of the goal of secondary structure prediction is to predict the segments as well as could be done by homology modelling, i.e. to approach a value of Sov=90%, and to additionally reach a per-residue accuracy in the order of 85% [54].

2: Improving secondary structure prediction by incorporation of evolutionary information

2.1: Setting up evolutionary records of neutral sequence variation

The mutation of a single residue typically causes an approximate reduction of the free energy difference
between native and unfolded state of about 1 kcal/mol [67]. Thus, the exchange of a few residues can already destabilise a protein of more than 100 residues [68, 69]. Does this imply that two proteins with some different residues have a different 3D structure? Random errors in the DNA lead to the wrong translation of the information coded in the genes into sequences of amino acids. These errors are the basis for evolution [70, 71]. The function of a protein is mainly determined by its 3D structure. Mutations resulting in a structural change are not likely, since the protein cannot perform its task. Thus, only those errors are likely to be accepted which do not alter the structure [72]. Consequently, the known proteins are a record of exploration for variation of sequence with no effect to structure.

How can the maximal variation be measured? A practical way is to compare the variation in sequence for proteins with the same structure. This has been done using some 500-1,300 proteins of known structure [11]. The result is that a cut-off for significant sequence similarity can be defined, such that it is very likely that two protein sequences with a mutual sequence identity above this value have the same structure. The cut-off depends on the length of the fragments for which the two sequences can be aligned. E.g., for alignment length > 80 residues a pair of proteins with only 25% identical residues has the same 3D structure. Of course not any two residues can be exchanged anywhere in the sequence. Instead, the possible exchanges depend on the details of the structure and on the chemical properties of the amino acids involved. Thus, the pattern of residue substitution carries information rather specific for a particular protein structure.

For the generation of alignments we used the MaxHammer/SSSP algorithm that builds up the alignment in essentially two steps. In sweep 1, the sequences are aligned consecutively to the guide sequence (SOS) by a standard dynamic programming method [73]. After each sequence has been added to the alignment an alignment profile is compiled. This profile contains the occurrence of each amino acid at each position in the alignment. The profile is used to align the next sequence. In sweep 2, after all sequences with significant homology have been picked from Swickpopt, the profile is recomputed, and the dynamic programming algorithm starts once again to align consecutively the sequences, this time using the conservation profile as derived after completion of sweep 1. In addition, a conservation weight is calculated at each position of the alignment [17].

2.2: Using evolutionary information for predictions

Recently, the use of evolutionary information was shown to improve the prediction accuracy both for individual cases [55, 56, 58, 74–81] and for sets of proteins [82–84]. The first method that broke the 70% barrier in Q3 when tested on more than 100 unique proteins was a system of three levels of multi-layered feed-forward networks (' neural networks'). In brief, the method is the following (more detailed descriptions are given in [53] and [82]).

The profiles from the multiple alignments are used as input to a first level two-layered feed-forward network ('sequence-to-structure net'). This is done by shifting a window of 13 residues successively through the sequence. The output of the network consists of three real numbers between 0 and 1 which give the probability for the residue at the centre of a particular window to be in a helix, strand, or loop. The first level sequence-to-structure net outputs the prediction for one single residue for each window. Thus, there is no explicit correlation between the secondary structure of adjacent residues, as observed in real protein structures. This shortcoming is corrected by feeding the output of the first level sequence-to-structure net into a second level structure-to-structure network (with the input window extending over 17 residues).

| Table 2: 126 protein chains used for training and testing the networks |

Representative set of 126 globular protein chains with less than 25% pairwise similarity for lengths > 80 used for training and testing the method C4,395 residues with 32% a, 21% b, and 47% l, resolution ≤ 2.5 Å for crystal structures). Nomenclature: The Protein Data Bank (PDB) identifier (first four characters) is followed by the chain identifier.

| 256b_A | 2aa_2 | 8bp_2 | 6acn | 1aa2 |
| 256b_B | 3jk | 1e93_A | 2a4p | 9mpl_A |
| 1phem_B | 1e93_A | 3ibc | 1bbp_A | 1bol |
| 2t5t_A | 1e92 | 1go3 | 2as4 | 1a4d |
| 1chx_A | 3cna | 4cna | 4cpa | 1d3y |
| 6gpp | 6gpp | 4epv | 1cm | 1asw |
| 6a8 | 4e43 | 3cna | 1e9d | 3f3f |
| 5c5a | 5e9e | 1e9d | 1laa | 1d3y |
| 1tifs | 1tifs | 1enr | 2a4p | 1a4d |
| 2gh_A | 2gh_A | 2as4 | 1g1l_A | 4giz |
| 6hu | 6hu | 1g1l_B | 2as4 | 3vrd |
| 2ih_b | 2ih_b | 3d9d | 1e93 | 9mpl |
| 1l89 | 1l89 | 3ibc | 1bbp_A | 1bol |
| 3nie_b | 2as4 | 1asw | 1l89 | 9mpl |
| 2mev_A | 2or.L | 1ca3 | 2apb_A | 1paw |
| 9pap | 9pap | 4epf | 3gpm | 2fhh |
| 11hp | 11hp | 2a4p | 1asw | 1paw |
| 11hp | 11hp | 4epf | 4epf | 1paw |
| 3nit | 3nit | 2apb_A | 4giz | 1d3y |
| 1ieb_A | 1ieb_A | 1g1l_A | 2as4 | 2cod |
| 1ieb_A | 1ieb_A | 1g1l_A | 2as4 | 2cod |
| 1ieb_A | 1ieb_A | 1g1l_A | 2as4 | 2cod |
| 1ieb_A | 1ieb_A | 1g1l_A | 2as4 | 2cod |

387
3. Improvement to 71.4% by use of indel information

3.1: Coding insertions and deletions from alignments as additional input units

Sequence alignments typically allow for insertions and deletions. Given the following two sequence sketches:
- Aligning without insertions gives:
  - LNITFTCMW
  - LEERGEMW

- i.e., three residues identical in both

sequences instead of two for the comparison shown above. Allowing gaps, the optimal alignment has 4 residues in both sequences:
  - LNITFTCMW
  - LEERGEMW

Deleting the E between T and G in the first sequence would have had the same effect of 4 identical residues. Insertions and deletions can more often occur in loop regions than in regular secondary structure elements like helix and strand [85, 86]. This implies that the number of insertions and deletions at a particular sequence position of the alignment carries information about secondary structure; the more insertions or deletions, the more likely it is a loop region (provided the alignment is sufficiently diverse).

The information about insertions and deletions (indels) was used for the input of the networks by adding two input units per basic cell. The input vector for the first 13 residues of a protein is:

\[
\begin{aligned}
0 & \quad N_{al}(i) \\
0 & \quad \frac{N_{al}(i)}{N_{al}}, \quad \text{and} \quad \frac{N_{al}(i)}{N_{al}} \quad \text{for} \ i = 1, \ldots, w
\end{aligned}
\]

where \(N_{al}(i)\) is the number of insertions at sequence position \(i\) of the alignment, \(N_{al}(i)\) the number of deletions at that position, and \(N_{al}\) the number of sequences in the alignment (only introduced to normalise the input units to 1). \(w\) is the window size (number of consecutive residues used for one input vector). The choice of 23 (and consequently 24 for the next unit) is because the 20 first units are used for coding the 20 different amino acids, 1 for the solvent, and the 22nd for the conservation weight.

3.2: Effect of the explicit use of indel information for the input

Using the indel information for the first level sequence-to-structure networks increases the accuracy, using it for the second level structure-to-structure decreases the accuracy. The number of correctly predicted residues observed to be in helix or strand is inferior to the system not using indels. This is explained by the fact that the number of insertions and deletions is in particular informative for the existence of loop regions [85].

The improvement obtained by indels is a further increase of almost one percentage point over the network ignoring insertions and deletions. An increase of some 5-6 percentage points over the best results published previously which is not entirely comparable due to allowing for significant sequence identity in the data [87].

Studying evolution obviously helps tremendously in efforts to predict secondary structure. Two questions arise: (1) How much of the improvement stems from the improvements on the side of the neural networks (3 levels, balanced learning scheme [53])? (2) How does the result compare to non-network methods using evolutionary information?

(1) There are three main components which improve the performance on the network side. First, the jury decision (3rd level) which is about one percentage point superior to the best second level network. Second, the balanced training procedure (presenting helices, strands and loop examples equally often during training), which increased the accuracy for strand by more than 10 percentage points. And third, the use of a second level which does accurately influence the overall accuracy (from e.g. 68.1 on the 1st level to 68.9 on the 2nd level), but the length distribution of the predicted segments is more protein like and the improvement in terms of the segment measure Sow is from 69.2 to 71.9%. Thus, in terms of overall accuracy about 2 percentage points stem from enhancing the network. Without using multiple alignments a one level network results in about 61% overall accuracy when cross-validated on the 126 proteins. Consequently, about 9 percentage points of the increase stems from the use of multiple alignment information. About 2 of these from the additional usage of conservation weights and indels.
(ii) A comparable increase to the one of the network system from 57.5 to 66.1% for using the Robson method 
[40, 41, 87] with multiple sequence alignment information 
was reported earlier on the basis of 11 proteins [83]. 
A comparison using the same proteins on a statistically 
significant number of samples is missing. One possible 
comparison is that with the expert predictions of the ETH 
Zürich, which can currently be based on five proteins 
(Table 1). On this small set of examples 'man with machine' does some 10 percentage points better than 'man 
without machine' [83].

4: The reliability index provides an accurate 
measure for the accuracy of the prediction

4.1: The expected variation of prediction accuracy with protein chain is some 20% 

The numbers might be less interesting and largely 
confusing if the only thing one wants to know is: how 
good is the prediction on the test protein SOS? The 
discovering message to the potential user is: for the 126 
proteins used in the cross-validation test the standard 
development was 9.5%, i.e. the prediction of SOS is likely to 
be 71.4±19% accurate (the interval of ±2 standard 
developments covers more than 95% of the samples). 
But errors could be significantly worse. Secondary structure 
prediction in proteins is not as reliable as the helices in 
motifs are. So, all the unusual SOS is 
compared to known structures, the less likely is a good 
prediction.

4.2: 36% of the residues predicted at the level of homology modelling

A more encouraging message for the user is that the 
network prediction allows the identification of regions 
which are predicted with higher reliability. About 36% of 
all residues are predicted at an accuracy of 90%, i.e. 
comparable to what can be expected if homology 
modelling were possible for SOS (Fig. 1).

A different question is: how reliable is the correctness 
of the prediction of a helix or a strand? We compute the 
average reliability index of all predicted segments and 
evaluate the correctness as a function of the average 
reliability (Fig. 2). The result is that about three quarters 
of all predicted helices score at a fractional overlap 
Sov>80%, but only a few have Sov>90%. This relatively 
poor performance stems from the fact that the fractional 
overlap is higher if computed as the percentage of 
observed segments (Sov=72.7%) than if computed as the 
percentage of predicted segments (Sov=57.7%). In other 
words, the probability to predict all observed segments 
especially correctly is higher than the probability that all 
predicted segments are correct. For SOS this implies that 
a residue predicted to be in a helix with a reliability of 9 
has a chance of >95% to be correctly predicted, but the 
helix predicted around that residue might have a different 
placement and/or extension than the prediction suggests. 
The segments predicted with higher reliability could be 
used for a molecular dynamics minimisation using the 
secondary structure segments as rigid bodies [90].

![Figure 1: Expected prediction accuracy for residues with a reliability index above a given cut-off](image1)

Plot are averages of the three state accuracy over all 
those residues with reliability index >0, m=0, ..., 9. E.g. 
about 22% of all residues have Sov>7 and of those 92% are 
correctly predicted by PDB.

![Figure 2: Fractional overlap for helices and strands vs. average reliability per segment](image2)

The average reliability is computed for all predicted 
helices and strands. Plotted are the fractional overlaps for 
all segments predicted at an average reliability >0, m=0, ..., 8. E.g., about 25% of all helices yield a value for Sov 
(caption of Table 1) of about 85%. By definition the 
numbers given here are the probabilities for a predicted 
segment to be correct, whereas the numbers in Table 1 
summarize the probabilities that an observed segment is 
correctly predicted.
5: The automatic secondary structure prediction from Heidelberg

If theoretical tools are designed to contribute to reducing the sequence structure gap, then the improvement of secondary structure prediction methods has a reason only in case the method is made available to potential users. If one wants to predict the secondary structure of a protein one might feel inclined to use the best available method to do so. But which one is that? The comparison of different methods is made difficult by the fact that most methods are evaluated on different data sets. One way out of this dilemma is to set up an automatic prediction service via the internet (for instructions send the word 'help' to the Internet address: PredictProtein@EMBL-Heidelberg.DE). The PHD method can now publicly be tested on new proteins. Thus, the method might be used as a reference point for the development of better tools.

The output of the server consists of a multiple alignment generated by the program MaxHom [11] and written in the format of the HSSP files (available via anonymous ftp: ftp.embl-heidelberg.de) and of the prediction of secondary structure in three states. The accuracy of the prediction depends crucially on the information contained in the multiple alignment. In the first nine months of 1993 some 6,000 predictions were requested.

6: Evaluation of PHD on 45 recently solved proteins

The PHD method has been shown to yield on average $Q_3 = 71.4\%$ and $Sov = 72.7\%$ in a multiple cross-validation test on 125 proteins. Two questions remain. (i) How accurate will the prediction turn out to be on the next 125 proteins the structure of which will be experimentally solved? (ii) Has the testing set influenced the development of the method, i.e., was the result implicitly optimised on a given data set? After we had performed all analyses we tested 45 proteins which had no sequence similarity to those used for training (Table 3). The result confirmed the values obtained by the cross-validation: $Q_3 = 71.6\%$ and $Sov = 72.8\%$ (Fig. 3, Table 1).
indicates that (i) there is a good chance for the network method to score equally high for the next hundreds of proteins and that (ii) the attempt to not optimise the development of the method with respect to a particular data set was successful.

In Fig. 4, explicit examples for some cases are given. The zinc finger DNA binding domain (PDB code: 3amn) is predicted to have no regular secondary structure at all, when only the sequence itself is used. An alignment with 7 sequences improves the results significantly (although inferior to the average). The anti-freeze protein type III (Swissprot: bapc_macaec (89)) is predicted almost as badly as a random prediction would do. The prediction of the anti-freeze type I protein (1fas) confirmed the model.

Table 3: 45 proteins with recently solved structure

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetylcholinesterase</td>
<td>acetylcholinesterase (complex with Dnase I)</td>
</tr>
<tr>
<td>Bapc_macaec</td>
<td>anti-freeze zinc finger protein (type III)</td>
</tr>
<tr>
<td>Arc</td>
<td>Arc repressor DNA-binding protein</td>
</tr>
<tr>
<td>B. amm</td>
<td>antibacterial protein</td>
</tr>
<tr>
<td>colae</td>
<td>colae A (C-terminal domain)</td>
</tr>
<tr>
<td>Coq</td>
<td>coq-6 (A-amyloid protein)</td>
</tr>
<tr>
<td>Domain</td>
<td>Domain of tyrosine kinase src</td>
</tr>
<tr>
<td>Ep3</td>
<td>Ep3 (sensory beta subunit of E.coli DNA polymerase II helicosome)</td>
</tr>
<tr>
<td>Epp</td>
<td>Epp (Guanine nucleotide exchange factor)</td>
</tr>
<tr>
<td>Fvtr</td>
<td>Fvtr (basic fibroblast growth factor)</td>
</tr>
<tr>
<td>G11</td>
<td>G11 (a factor)</td>
</tr>
</tbody>
</table>
| Ig | Ig |}

with almost 100% per-residue accuracy. The prediction of the ATPase fragment of heat shock protein hsp70 (1hsp) is about average. An interesting detail is that the helix predicted around residues 64-72 is almost correct; the 3D structure has a helix-like turn.

Conclusion and outlook

Secondary structure prediction methods operate in a range between 55 and 85% three-state accuracy. Until the early 90s methods have hovered about 60-66% overall accuracy. Due to lacking rigour in the evaluation of the results, it is likely that 50% is closer to where prediction methods score in practice. It is not sufficient to evaluate prediction methods based on per-residue measures; instead, the accuracy in terms of segments should be taken into account. The evolutionary information contained in multiple alignments can be used to pass prediction accuracy over 70% by using a three level system of neural networks. The performance of the network system can be improved by explicitly using the information about insertions and deletions contained in the multiple alignment. The final system scores at 71.4% and 77.2% in a multiple cross-validation test on 126 proteins. A test on 45 proteins with recently solved structure shows that the high level of accuracy is likely to be a reasonable estimate for future predictions. Prediction accuracy is not equally distributed over all sample proteins. Instead, there is a considerable variation with the protein chain. However, the network method permits an assessment of the reliability of the predictions: 56% of all sites are predicted at a level of 88% accuracy which is comparable to the prediction by homology modelling. Methods for the prediction of secondary structure have an impact on the research in molecular biology only if they are made available to potential users. The network predictions can be obtained by electronic mail.

In the wake of large DNA-sequencing projects, one requirement for prediction of structural features is speed: the automatic prediction of secondary structure can easily keep track with sequencing. Another requirement is quality: the prediction accuracy has been improved significantly by the profile network method. But is this worth while the effort? A practical answer is given by the community of researchers repeatedly using the prediction service to assist their research in molecular biology. However, there are two severe restrictions of the method:

(i) Most contemporary theoretical predictions are successful at most for the exons in the data bank. The number of domain realise in nature might be limited [30, 91]. This makes it promising to learn from already existing cases. But, new motifs typical compared to what has been found so far, certainly cannot be determined other than by experiment. (ii) The description of a protein structure as one-dimensional
Figure 4: Some explicit examples for neural network predictions

<table>
<thead>
<tr>
<th>Protein 3aaf (zinc finger DNA binding domain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
</tr>
<tr>
<td>Obs</td>
</tr>
<tr>
<td>PDB</td>
</tr>
<tr>
<td>PID</td>
</tr>
<tr>
<td>PHD</td>
</tr>
</tbody>
</table>

without alignment
with alignment

<table>
<thead>
<tr>
<th>Protein 1n0c: (AAT3ase fragment of heat shock protein hsp70, only the first 80 of 382 residues shown here)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
</tr>
<tr>
<td>Obs</td>
</tr>
<tr>
<td>PDB</td>
</tr>
<tr>
<td>PID</td>
</tr>
</tbody>
</table>

Abbreviations used: AA: amino acid sequence; Obs: observed secondary structure; PDB: predicted secondary structure. The secondary structure is assigned by DSSP [38] with the abbreviations: H: helix; E: extended strand; G: 310 helix; B: β-bulge; and S: spaces for loop regions. Note: for training the network and for evaluating the accuracy, 310 helices were converted to H, β-bulges to loop.

string of secondary structure segments is one of the most drastic and simple reductions of the underlying 3D reality. This reduction can be helpful only as long as more advanced methods are missing.

Secondary structure predictions can certainly be improved further. The goal for such improvements should be to increase the segment accuracy (SOV) in order to obtain predictions more on the real goal: the prediction of 3D structure from the sequence. What should be the next step on the road to practically solving the protein folding problem? To reach practical contributions to the reduction of the growing gap between proteins of known sequences and those of known structure, theoretical tools will have to be improved by increasing the dimension of the features that are predicted. The goal is a prediction of 3D structure: this should not be forgotten while attempting to improve predictions in one dimension.

References


