

*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 the  
IEEE Computer Society Press  
10662 Los Vaqueros Circle  
P.O. Box 3014  
Los Alamitos, CA 90720-1264

© 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 94TH0607-2

ISBN 0-8186-5090-7 (paper)

ISBN 0-8186-5091-5 (microfiche)

ISSN 1060-3425

Additional copies can be ordered from:

IEEE Computer Society Press  
Customer Service Center  
10662 Los Vaqueros Circle  
P.O. Box 3014  
Los Alamitos, CA 90720-1264  
Tel: (714) 821-8380  
Fax: (714) 821-4641  
Email: [cs.books@computer.org](mailto:cs.books@computer.org)

IEEE Service Center  
445 Hoes Lane  
P.O. Box 1331  
Piscataway, NJ 08855-1331  
Tel: (908) 981-1383  
Fax: (908) 981-8667

IEEE Computer Society  
13, avenue de l'Aquilon  
B-1200 Brussels  
BELGIUM  
Tel: +32-2-770-2198  
Fax: +32-2-770-8505

IEEE Computer Society  
Ooshima Building  
2-19-1 Minami-Aoyama  
Minato-ku, Tokyo 107  
JAPAN  
Tel: +81-3-3408-3118  
Fax: +81-3-3408-3553

Production Editor: Penny Storme 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 the structure is experimentally solved. Another 4,000 can be modeled by homology. For the remaining 25,000 sequences, the tertiary structure (3D) 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 36% (random) and 88% (homology modelling) 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 at each residue position: e.g. about 40% of the predicted residues have an expected accuracy of 88%. The method has been made publicly available by an automatic e-mail server.

## Introduction

The number of known protein sequences (30,000 Swissprot<sup>†</sup>, 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 data bank of known structures, model building by homology allows prediction of the structure of SOS with reasonable accuracy [4-16]. If not, i.e. if SOS belongs to the majority of the 80% of known sequences which do not have homologues of 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

is too long, one can try to thread the SOS sequence into a known structure, i.e. to find a protein of known structure that has no significant sequence similarity to SOS but is likely to have the same fold [14, 25-37]. If this attempt also fails prediction in 3D is no longer possible in general. Thus, one has to simplify the problem. One extreme simplification is the prediction of one dimensional strings of secondary structure assignment.

Here, we shall introduce a novel criterion for the evaluation of secondary structure prediction methods based on segments, and shall present a neural network method that not only reaches a three-state accuracy of 71.4% by the use of evolutionary information but also improves the prediction in terms of the segment score. The high level of accuracy is confirmed by an additional test on 45 recently solved structures. The prediction method is available via electronic mail. We attempt to provide a potential user of the method with a realistic estimate of the accuracy to be expected for SOS.

## 1: Exploring the limits of secondary structure prediction

### 1.1: Secondary structure prediction between 35 and 88% in per-residue measures

A random prediction of secondary structure in three states (helix, strand, and rest, here termed loop) yields an overall per-residue accuracy of about 35% (Table 1, numbers given refer to the DSSP assignment of secondary structure [38]). This value provides a lower limit for the evaluation of predictions. Early methods like the ones of Chou and Fasman [39], Robson et al. [40, 41] and Lim [42] scored 14-19% above the random level [43]. In the 80's the value has been set up to about 63-66%, i.e. 27-30% above random [44-50]. How good can secondary structure prediction become? For quite some time it was claimed that 65-70% was the ultimate accuracy obtainable by methods which are restricted to the explicit use of local information [51, 52]. There is evidence that this value has to be readjusted. Predictions can be better than 70%, and not just on favourable subsets of proteins [53]. Another question is whether the 100% mark is a reasonable goal? An upper limit for a very accurate prediction can be deduced by comparing the similarity in secondary

<sup>†</sup> Abbreviations used: 3D: three-dimensional; PDB: Protein Data Bank of known three-dimensional structures; Swissprot: data bank of known sequences; HSSP: database of Homology-derived Structures of Proteins; DSSP: Dictionary of Secondary Structures of Proteins; PHD: Profile network from Heidelberg (three levels of networks for the prediction of secondary structure); SOS: Sequence of unknown Structure.

Table 1: Comparison of single residue scores

method	N	N <sub>prot</sub>	Q <sub>3</sub>	Q <sub>α</sub>	Q <sub>β</sub>	Q <sub>l</sub>	corr <sub>α</sub>	corr <sub>β</sub>	corr <sub>l</sub>	I	Sov
PDB	24291	140	88.4	88	86	99	0.84	0.85	0.77	0.616	89.7
RAN	12162	94	35.2	23	26	47	-0.09	0.01	0.04	0.006	38.7
PHD	23191	126	71.4	69	64	77	0.62	0.52	0.51	0.267	72.7
PHD on45	9338	45	71.6	71	56	81	0.65	0.51	0.52	0.278	72.8
ETH on 5	906	5	57.2	50	47	72	0.49	0.30	0.31	0.139	63.2
PHD on 5	906	5	71.6	69	55	86	0.62	0.55	0.54	0.286	76.2

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 indels cross-validated on the 126 proteins of Table 2. PHD on45: Same network tested on the 45 proteins of Table 3. ETH on5: Results for 5 predictions of experts [55-58]. The proteins are: lcpk: cAMP dependent protein kinase [59]; csrc: SRC tyrosin kinase [60]; sh3: SH3 domain of spectrin [61]; pik3: p85b\_human, phosphatidylinositol 3-OH kinase [62,63]; and nifk\_azovi: molybdenum-iron nitrogenase [64]. PHD on5: Results for network system on the same 5 proteins.

Abbreviations for accuracy measures: N: number of residues in the data set; N<sub>prot</sub>: number of proteins in the data set. Q<sub>3</sub>, Q<sub>α</sub>, Q<sub>β</sub>, Q<sub>l</sub>: percentages of correctly predicted residues: for all three states, and for helix, strand and loop. corr<sub>α</sub>, corr<sub>β</sub>, corr<sub>l</sub>: give the Matthews correlation coefficients [65]. I: measures the information defined by [53]. Sov: fractional overlap in detail defined by [54].

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 helix, 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 Q<sub>3</sub>) 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 permissively 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 (for helix) and the region in which either of the two (or both) has a H. For example:

observed HHHH yields the fraction:  $\frac{3}{6}$   
 predicted HHHHH

$$Sov = \sum_i \frac{\min_{ov}(i) + \delta}{\max_{ov}(i)} * \text{len}(i)$$

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

Sov is about 38% 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 1kcal/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 physico-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 MaxHom/HSSP 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 Swissprot, the profile is recompiled, 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 sequence 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 Q<sub>3</sub> 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

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 (24,395 residues with 32%  $\alpha$ , 21%  $\beta$ , and 47% L, resolution  $\leq 2.5\text{\AA}$  for crystal structures). Nomenclature: the Protein Data Bank (PDB) identifier (first four characters) is followed by the chain identifier.

256b_A	2aat	8abp	6acn	1acx
8adh	3ait	1ak3_A	2aip	9api_A
9api_B	1azu	3b5c	1bbp_A	1bda
1bmv_1	1bmv_2	3blm	4bp2	2cab
7cut_A	1cbh	1cc5	2ccy_A	1cd4
1cdt_A	3cla	3cin	4cms	4cpa_I
6cpa	6cpp	4cpv	1crn	1cse_I
6ctc	2cyp	5cyr_R	1cca	6cfr
3ebx	5er2_E	1etu	1fc2_C	1fdi_H
1fdx	1fcl	2fnr	2fxb	1fxi_A
4fxn	3gap_A	2gbp	2ger	1gdl_O
2gla_A	2gn5	1gpl_A	4gr1	1hip
6hir	3hmg_A	3hmg_B	2hmg_A	5hvp_A
2ilb	3icb	7icd	1iis_A	9ins_B
1i58	1iap	5kdh	2lh4	2lhb
1iud_3	2ltn_A	2ltn_B	5lyz	1ncp_L
2mev_4	2orl_L	1ovo_A	2pab_A	1pas
9pap	2pcy	4pfl	3pgm	2phh
1pyp	1r09_2	2mhu	1nrt	1ppi
1rbp	1rhd	4rhv_1	4rhv_3	4rhv_4
3rnt	7rua	2rap_A	4rxn	1s0l
1sdb_A	4sgb_I	1shl	2sns	2sod_B
2stv	2tgp_I	1tga_I	3tim_A	6tmm_E
2umv_P	1urf_A	4tst_A	2tuc_A	1ubq
2utg_A	9wga_A	2wrp_R	1way_A	1way_B
4xia_A				

consecutive residues). The third level is the computation of an arithmetic average over the outputs of several independently trained second level nets (jury decision). The networks used for the 3rd level (jury) are trained on the same training set, the differences stem mainly from a different order of presenting the samples during the training procedure [53].

Such a network system increases the prediction accuracy from about 61% to almost 70% by using the profiles derived from multiple alignments [82] and above 70% by also using the conservation weight computed from the sequence alignment [53]. Can the performance of the network system be improved further by using additional input information?

### 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 stretches:

LNNTEGDWW  
LEEHGEWW     Aligning without insertions gives:  
LNNTEGDWW  
LEEHGEWW     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:

LNNTEGDWW  
LEEH .GEWW     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:

$$s_{23+j} = \frac{N_{ins}(j)}{N_{all}}, \text{ and } s_{24+j} = \frac{N_{del}(j)}{N_{all}}, \text{ for } j = 1, \dots, w$$

where  $N_{ins}(j)$  is the number of insertions at sequence position  $j$  of the alignment,  $N_{del}(j)$  the number of deletions at that position, and  $N_{all}$  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 indels is strongly correlated to the sequence information. Consequently, the inclusion of indel units pays off only on the first level of sequence-to-structure network.

Adding the number of insertions and deletions in the multiple alignment increases the overall accuracy by another half percentage point to  $Q_3=71.4\%$  (sevenfold cross-validation on the 126 globular protein chains in Table 2). The increase in overall accuracy mainly stems from a more accurate prediction for loop regions (from 74% to 77%). The percentage of correctly predicted residues observed to be in helix or strand is inferior to the system not using indels. This is explained by 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. And an increase of some 5-6 percentage points over the best results published previously which is not strictly comparable due to allowing for significant sequence identity in the data [47].

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

(i) 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 increases the accuracy for strand by more than 10 percentage points. And third, the use of a second level which does scarcely 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  $Sov$  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 stem 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' [88].

#### 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 discouraging message to the potential user is: for the 126 proteins used in the cross-validation test the standard deviation was 9.5%, i.e. the prediction of SOS is likely to be  $71.4 \pm 19\%$  accurate (the interval of  $\pm 2$  standard deviations covers more than 95% of the samples). But matters could be significantly worse. Secondary structure predictions are successful in capturing the clichés contained in the data bank. So, the more unusual SOS is compared to known structures, the less likely is a good prediction.

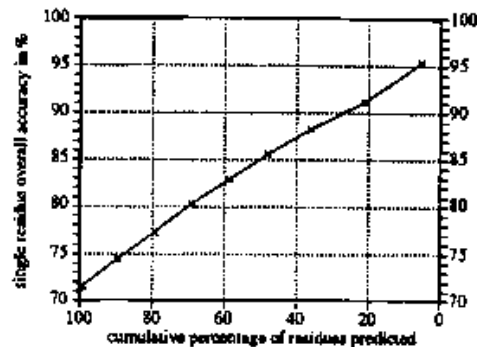
##### 4.2: 36% of the residues predicted at the level of homology modelling

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 88%, 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 ( $Sov = 67.7\%$ ). In other words, the probability to predict all observed segments essentially 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

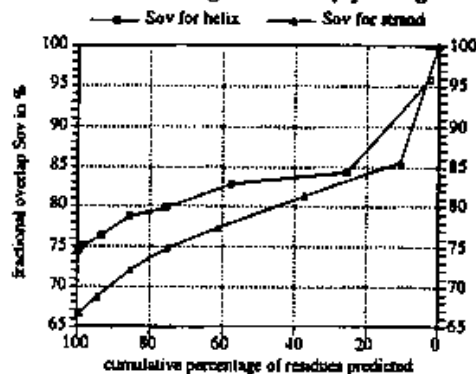
placement and/or extension than the prediction suggests. The segments predicted with higher reliability could be used for a molecular dynamics minimisation using the secondary structure segments as rigid bodies [90].

Figure 1: Expected prediction accuracy for residues with a reliability index above a given cut-off



Plotted are averages of the three state accuracy over all those residues with reliability index  $> n$ ,  $n=0, \dots, 9$ . E.g. about 22% of all residues have  $RI > 7$  and of these 92% are correctly predicted by PHD.

Figure 2: Fractional overlap for helices and strands vs. average reliability per segment



The average reliability is computed for all predicted helices and strands. Plotted are the fractional overlaps for all segments predicted at an average reliability  $> n$ ,  $n=0, \dots, 8$ . E.g., about 25% of all helices yield a value for  $Fov$  (captions 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.



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: 3znf) is predicted to have no regular secondary structure at all, when only the sequence itself is used. An alignment with 7 sequences improves the result significantly (although inferior to the average). The anti-freeze protein type III (Swissprot: anpc\_macam [89]) is predicted almost as badly as a random prediction would do. The prediction of the anti-freeze type I protein (1aif) confirmed the model

**Table 3: 45 proteins with recently solved structure**

The 45 protein chains were chosen from a much larger Protein Data Bank 'prerelease' set such that they all have less than 25% (for length > 80) similarity to any of the proteins in Table 2 used for training the networks. Nomenclature: where possible the Protein Data Bank (PDB) identifier (first four characters) followed by the chain identifier is given; else the code of Swissprot is used.

lace: acetyl cholinesterase; act1: actin (complex with DNase I); anpc\_macam: antifreeze glycoprotein type III; arc: Arc repressor DNA-binding protein; lool: antibacterial protein colicin A (c-terminal domain); lcox: cholesterol oxidase; lcpk\_E: cAMP-dependent protein kinase; csrc: sh3 domain of tyrosine kinase arc; ldfn\_B: defensin HNP-3; dp3b\_scoli:  $\beta$ -subunit of E.coli DNA polymerase III holoenzyme; lend: glutathione synthase; 5eni: enolase; lfbg: phosphocarrier; 2fgf: basic fibroblast growth factor; 2gbl: protein G (b1 domain); lgly: glucoamylase; lgmf\_A: granulocyte-macrophage colony-stimulating factor; lhcc: glycoprotein; l6th: complement control protein of factor b; lbdd\_C: engrailed homeodomain complex with DNA; 2hip\_B: high potential iron sulfur protein; lhrb: ribonuclease H domain of HIV-1 reverse transcriptase; lhsc: heat shock protein hsc70; hsk: yeast hexokinase b; lifb: intestinal fatty acid binding protein; lmsb\_A: mannose binding protein A (lectin domain); luxf\_phol: flavoprotein related to bacterial luciferase; nifk: nitrogenase molybdenum-iron; lnh\_B: neuraminidase sialidase; 5p21: CH-ras p21 protein (amino acids 1 - 166); pdr: phthalate dioxygenase reductase; lpi2: serine proteinase inhibitor; pik3: phosphatidylinositol 3-kinase; 2pk4: human plasminogen kringle 4; pou1: POU-specific domain; lrop: ColE1 repressor of primer; rxa-su: retinoid X receptor  $\alpha$  DNA binding domain; lser\_A: endoribonuclease SA; 2scp\_A: sarcoplasmic calcium binding protein; sh2: v-src tyrosine kinase transforming protein sh3: spectrin SH3 homologue domain; lsnv: sindbis virus capsid protein; ulr: RNA-binding domain of U1 small nuclear ribonucleoprotein A; 3trx: thioredoxin; 3znf: zinc finger DNA binding domain; zzia\_A: GCN4 leucine zipper.

with almost 100% per-residue accuracy. The prediction of the ATPase fragment of heat shock protein hsp70 (1hsc) 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 35 and 88% three-state accuracy. Until the early 90's methods have hovered about 60-66% overall accuracy. Due to lacking rigour in the evaluation of the results, it is likely that 60% 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 push 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  $Q_3=71.4\%$  and  $Sov=72.7\%$  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 prediction: 36% 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 people 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 clichés in the data bank. The number of folds realised in nature might be limited [30, 91]. This makes it promising to learn from already existing cases. But, new motifs atypical compared to what has been found so far, currently cannot be determined other than by experiment. (ii) The description of a protein structure as a one-dimensional

Figure 4: Some explicit examples for neural network predictions

protein 3znf (zinc finger DNA binding domain)

	.....1.....2.....3.....	
AA	RPYHCSYCNFSFKTKGNLTKHMKSKAHSKK	
Obs	B B HHHHHHHH	
PHD		without alignment
PHD	EEEEEEE HHHHHH	with alignment

protein anpc\_macam (anti freeze protein type III)

	.....1.....2.....3.....4.....5.....6.....
AA	NQASVVANQLIPINTALTLVMMRSEVVTPVGIPARDIPRLVSMQVNRVPLGTTLMPPDMVKGYPPA
Obs	EEEEE EEE EEEE EEEEE EEEE EEE EEE EEE
PHD	EEEEE HHHHH HHHHHHHH HHH

protein lhsc (ATPase fragment of heat shock protein hsp70, only the first 80 of 382 residues shown here)

	.....1.....2.....3.....4.....5.....6.....7.....
AA	KQPAVGIDLGTTSYSCVGVFQHGKVEIANDQGNRTTPSYVAFTDTERLIGDAARNOVAMNPTNTVFDKRLIGRRF
Obs	EEEEE EEEEE EEEE EE EEE EEE HHHH B
PHD	EEEEE EEEEEEE EEEE EEEE HHHHHHHHHHH HHHHHHHH

Abbreviations used: AA: amino acid sequence; Obs: observed secondary structure; PHD: predicted secondary structure. The secondary structure is assigned by DSSP [38] with the abbreviations: H: helix; E: extended strand;

G:  $3_{10}$  helix; B:  $\beta$ -bulge; and spaces for loop regions. Note: for training the network and for evaluating the accuracy,  $3_{10}$  helices were converted to H,  $\beta$ -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 orient the prediction more on the real goal: the prediction of 3D structure from the sequence. What should be the next steps on the road to practically solving the protein folding problem? To render 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

1. A. Bairoch & B. Boeckmann. "The SWISS-PROT protein sequence data bank", *Nucl. Acids Res.*, Vol. 20, pp. 2019-2022, 1992.
2. F.C. Bernstein, et al., "The Protein Data Bank: a computer based archival file for macromolecular structures", *J. Mol. Biol.*, Vol. 112, pp. 535-542, 1977.

3. U. Hobohm, M. Scharf, R. Schneider & C. Sander. "Selection of representative protein data sets", *Prot. Sci.*, Vol. 1, pp. 409-17, 1992.
4. J. Greer, "Comparative Modeling of Homologous Proteins", *Meth. Enzymol.*, Vol. 202, pp. 239-252, 1991.
5. T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg & J.M. Thornton. "Knowledge-based prediction of protein structures and the design of novel molecules", *Nature*, Vol. 326, pp. 347-352, 1987.
6. J. Greer, "Model for haptoglobin heavy chain based upon structural homology", *Proc. Natl. Acad. Sc. U.S.A.*, Vol. 77, pp. 3393-3397, 1980.
7. J. Greer, "Comparative Model-building of the Mammalian Serine Proteases", *J. Mol. Biol.*, Vol. 153, pp. 1027-1042, 1981.
8. J. Greer, "Comparative Modeling Methods: Application to the Family of the Mammalian Serine Proteases", *Proteins*, Vol. 7, pp. 317-334, 1990.
9. W.R. Taylor & C.A. Orengo, "A holistic approach to protein structure alignment", *Prot. Engin.*, Vol. 2, pp. 505-19, 1989.
10. G. Vriend & C. Sander, "Detection of Common Three-Dimensional Substructures in Proteins", *Proteins*, Vol. 11, pp. 52-58, 1991.
11. R. Schneider & C. Sander, "Database of Homology-Derived Structures and the Structural Meaning of Sequence Alignment", *Proteins*, Vol. 9, pp. 56-68, 1991.
12. W. Taylor, "New paths from dead ends", *Nature*, Vol. 356, pp. 478-480, 1992.

13. N.L. Summers & M. Karplus, "Modeling of Globular Proteins", *J. Mol. Biol.*, Vol. 216, pp. 991-1016, 1990.
14. J. Overington, M.S. Johnson, A. Sali & T.L. Blundell, "Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction", *Proceedings of the Royal Society, London*, Vol. B 241, pp. 132-145, 1990.
15. M. Levitt, "Accurate Modeling of Protein Conformation by Automatic Segment Matching", *J. Mol. Biol.*, Vol. 226, pp. 507-533, 1992.
16. L. Holm & C. Sander, "Fast and Simple Monte Carlo Algorithm for Side Chain Optimization in Proteins: Application to Model Building by Homology", *Proteins*, Vol. 14, pp. 213-223, 1992.
17. R. Schneider & C. Sander, "The HSSP data base of protein structure-sequence alignment", *Nucl. Acids Res.*, Vol. 21, pp. 3105-3109, 1993.
18. M. Karplus & G.A. Petsko, "Molecular dynamics simulations in biology", *Nature*, Vol. 347, pp. 631-639, 1990.
19. R.A. Abagyan, "Towards protein folding by global energy optimization", Vol. 325, pp. 17-22, 1993.
20. R. Abagyan & M. Totrov, "Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins", *J. Mol. Biol.*, Vol. 232, pp. in press, 1993.
21. W.F.v. Gunsteren, "Molecular dynamics studies of proteins", *Curr. Opin. Str. Biol.*, Vol. 3, pp. 167-174, 1993.
22. W.F.v. Gunsteren, "The role of computer simulation techniques in protein engineering", Vol. 2, pp. 5-13, 1988.
23. K.A. Dill, "Folding proteins: finding a needle in a haystack", Vol. 3, pp. 99-103, 1993.
24. R.L. Jernigan, "Protein folds", *Curr. Opin. Str. Biol.*, Vol. 2, pp. 248-256, 1992.
25. D. Eisenberg & A.D. McLachlan, "Solvation energy in protein folding and binding", *Nature*, Vol. 319, pp. 199-203, 1986.
26. G. Baumann, C. Frömmel & C. Sander, "Polarity as a criterion in protein design", *Prot. Engin.*, Vol. 2, pp. 329-334, 1989.
27. M.J. Sippl, "Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-based Prediction of Local Structures of Globular Proteins", Vol. 213, pp. 859-883, 1990.
28. M.J. Sippl & S. Weitckus, "Detection of Native-Like Models for Amino Acid Sequences of Unknown Three-Dimensional Structure in a Data Base of Known Protein Conformations", *Proteins*, Vol. 13, pp. 258-271, 1992.
29. G. Crippen M., "Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces", *Biochem.*, Vol. 30, pp. 4232-4237, 1991.
30. A.V. Finkelstein & B.A. Reva, "A search for the most stable folds of protein chains", Vol. 351, pp. 497-499, 1991.
31. R.A. Goldstein, Z.A. Luthey-Schulten & P.G. Wolynes, "Protein tertiary structure recognition using optimized Hamiltonians with local interactions", *Proceedings of the National Academy of Science, U.S.A.*, Vol. 89, pp. 9029-9033, 1992.
32. R. Lüthy, J.U. Bowie & D. Eisenberg, "Assessment of protein models with three-dimensional profiles", *Nature*, Vol. 356, pp. 1992.
33. R. Lüthy, A.D. McLachlan & D. Eisenberg, "Secondary Structure-Based Profiles: Use of Structure-Conserving Scoring Tables in Searching Protein Sequence Databases for Structural Similarities", *Proteins*, Vol. 10, pp. 229-239, 1991.
34. J. Overington, D. Donnelly, M.S. Johnson, A. Sali & T.L. Blundell, "Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds", *Prot. Sci.*, Vol. 1, pp. 216-226, 1992.
35. C.M. Stultz, J.V. White & T.F. Smith, "Structural analysis based on state-space modeling", *Prot. Sci.*, Vol. 2, pp. 305-314, 1993.
36. S.H. Bryant & C.E. Lawrence, "An empirical energy function for threading protein sequence through folding motif", Vol. pp. in press, 1993.
37. C. Ouzounis, C. Sander, M. Scharf & R. Schneider, "Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures", *J. Mol. Biol.*, Vol. 232, pp. 805-825, 1993.
38. W. Kabsch & C. Sander, "Dictionary of Protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features", *Biopolymers*, Vol. 22, pp. 2577-2637, 1983.
39. P.Y. Cbou & U.D. Fasman, "Prediction of protein conformation", *Biochem.*, Vol. 13, pp. 211-215, 1974.
40. J. Garnier, D.J. Osguthorpe & B. Robson, "Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins", *J. Mol. Biol.*, Vol. 120, pp. 97-120, 1978.
41. B. Robson & R.H. Pain, "Analysis of the Code Relating Sequence to Conformation in Proteins: Possible Implications for the Mechanism of Formation of Helical Regions", *J. Mol. Biol.*, Vol. 58, pp. 237-259, 1971.
42. V.I. Lim, "Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure", Vol. 88, pp. 857-872, 1974.
43. W. Kabsch & C. Sander, "How good are predictions of protein secondary structure?", *FEBS Lett.*, Vol. 155, pp. 179-182, 1983.
44. O.B. Ptitsyn & A.V. Finkelstein, "Theory of protein secondary structure and algorithm of its prediction", *Biopolymers*, Vol. 22, pp. 15-25, 1983.
45. J.M. Levin & J. Garnier, "Improvements in a secondary structure prediction method based on a search for local sequence homologues and its use as a model building tool", *Biochim. Biophys. Ac.*, Vol. 955, pp. 283-295, 1988.
46. J.M. Levin, B. Robson & J. Garnier, "An algorithm for secondary structure determination in proteins based on sequence similarity", *FEBS Lett.*, Vol. 205, pp. 303-308, 1986.
47. X. Zhang, J.P. Mesirov & D.L. Waltz, "Hybrid System for Protein Secondary Structure Prediction", *J. Mol. Biol.*, Vol. 225, pp. 1049-63, 1992.
48. S. Salzberg & S. Cost, "Predicting Protein Secondary Structure with a Nearest-neighbor Algorithm", *J. Mol. Biol.*, Vol. 227, pp. 371-374, 1992.
49. V. Biou, J.F. Gibrat, J.M. Levin, B. Robson & J. Garnier, "Secondary structure prediction: combination of three different methods", *Prot. Engin.*, Vol. 2, pp. 185-91, 1988.
50. J.-F. Gibrat, J. Garnier & B. Robson, "Further Developments of Protein Secondary Structure Prediction Using Information Theory. New Parameters and Consideration of Residue Pairs", *J. Mol. Biol.*, Vol. 198, pp. 425-443, 1987.
51. S. Rackovsky, "On The Nature of The Protein Folding Code", *Proc. Natl. Acad. Sc. U.S.A.*, Vol. 90, pp. 644-648, 1993.
52. M.J.E. Sternberg, "Secondary structure prediction.", *Curr. Opin. Str. Biol.*, Vol. 2, pp. 237-241, 1992.

53. B. Rost & C. Sander, "Prediction of protein secondary structure at better than 70% accuracy", *J. Mol. Biol.*, Vol. 232, pp. 584-599, 1993.
54. B. Rost, R. Schneider & C. Sander, "Redefining the goals of protein secondary structure prediction", *J. Mol. Biol.*, in the press, 1993.
55. S.A. Benner, M.A. Cohen & D. Gerloff, "Predicted Secondary Structure for the Src Homology 3 Domain", *J. Mol. Biol.*, Vol. 229, pp. 295-305, 1993.
56. S.A. Benner, M.A. Cohen & D. Gerloff, "Correct structure prediction?", *Nature*, Vol. 359, pp. 781, 1992.
57. S.A. Benner & D. Gerloff, "Patterns of Divergence in Homologous Proteins as Indicators of Secondary and Tertiary Structure of the Catalytic Domain of Protein Kinases", *Adv. Enz. Reg.*, Vol. 31, pp. 121-181, 1990.
58. D.L. Gerloff, T.F. Jeany, L.J. Knecht, G.H. Gonnet & S.A. Benner, "The nitrogenase MoFe protein", *FEBS Lett.*, Vol. 318, pp. 118-124, 1993.
59. D.R. Knighton, et al., "Crystal Structure of the Catalytic Subunit of Cyclic Adenosine Monophosphate-dependent Protein Kinase", *J. Mol. Biol.*, Vol. 253, pp. 407-414, 1991.
60. H.T. Yu, et al., "Solution Structure of the SH3 Domain of Src and Identification of Its Ligand-Binding Site", *Science*, Vol. 258, pp. 1665-1668, 1992.
61. A. Musacchio, M. Noble, R. Pauptit, R. Wierenga & M. Saraste, "Crystal structure of a Src-homology 3 (SH3) domain", *Nature*, Vol. 359, pp. 851-855, 1992.
62. D. Kohda, et al., "Solution Structure of the SH3 Domain of Phospholipase C-γ", *Cell*, Vol. 72, pp. 953-960, 1993.
63. S. Koyama, et al., "Structure of the PEK SH3 Domain and Analysis of the SH3 Family", *Cell*, Vol. 72, pp. 945-952, 1993.
64. J. Kim & D.C. Rees, "Crystallographic structure and functional implications of the nitrogenase molybdenum-iron protein from *Azotobacter vinelandii*", *Nature*, Vol. 360, pp. 353-360, 1992.
65. B.W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, Vol. 405, pp. 442-451, 1975.
66. B. Rost, C. Sander & R. Schneider, "Progress in protein structure prediction?", *TIBS*, Vol. 18, pp. 120-123, 1993.
67. E.E. Lattman & G.D. Rose, "Protein folding-what's the question?", *Proceedings National Academy of Sciences USA*, Vol. 90, pp. 439-441, 1993.
68. E.I. Shakhovich & A.M. Gutin, "Influence of Point Mutations on Protein Structure: Probability of a Neutral Mutation", *Journal of theoretical Biol.*, Vol. 149, pp. 537-546, 1991.
69. H.B. Zabin, M.P. Horvath & T.C. Terwilliger, "Approaches to Predicting Effects of Single Amino Acid Substitutions on the Function of a Protein", *Biochem.*, Vol. 30, pp. 6230-6240, 1991.
70. C. Darwin, *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray, 1859.
71. J. Monod, *Le hasard et la nécessité*, Paris: Seuil, 1970.
72. C. Chothia & A.M. Lesk, "The relation between the divergence of sequence and structure in proteins", *EMBO J.*, Vol. 5, pp. 823-826, 1986.
73. T.F. Smith & M.S. Waterman, "Comparison of biosequences", *J. Mol. Biol.*, Vol. 2, pp. 482-489, 1981.
74. J.F. Bazan, "Structural design and molecular evolution of a cytokine receptor superfamily", *Proc. Natl. Acad. Sc. U.S.A.*, Vol. 87, pp. 6934-6938, 1990.
75. G.J. Barton, R.H. Newman, P.S. Freemont & M.J. Crumpton, "Amino Acid sequence analysis of the annexin supergene family of proteins", *Eur. J. Biochem.*, Vol. 198, pp. 749-760, 1991.
76. T.J. Gibson, J.D. Thompson & R.A. Abagyan, "Proposed structure for the DNA-binding domain of the Helix-Loop-Helix family of eukaryotic gene regulatory proteins", *Prot. Engin.*, Vol. 6, pp. 41-50, 1993.
77. R.B. Russell, I. Breed & G.J. Barton, "Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains", *FEBS Lett.*, Vol. 304, pp. 15-20, 1992.
78. A. Musacchio, T. Gibson, V.-P. Lehto & M. Saraste, "SH3 - an abundant protein domain in search of a function", *FEBS Lett.*, Vol. 307, pp. 55-61, 1992.
79. J. Franpton, A. Lautz, T.J. Gibson & T. Graf, "DNA-binding domain ancestry", *Nature*, Vol. 342, pp. 134, 1989.
80. T. Niermann & K. Kirchner, "Improving the prediction of secondary structure of TIM-barrel enzymes (Corrigendum)", *Prot. Engin.*, Vol. 4, pp. 359-370, 1991.
81. B. Rost & C. Sander, "Jury returns on structure prediction", *Nature*, Vol. 360, pp. 540, 1992.
82. B. Rost & C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks", *Proc. Natl. Acad. Sc. U.S.A.*, Vol. 90, pp. 7558-7562, 1993.
83. M.J. Zvelebil, G.J. Barton, W.R. Taylor & M.J.E. Sternberg, "Prediction of protein secondary structure and active sites using alignment of homologous sequences", *J. Mol. Biol.*, Vol. 195, pp. 957-961, 1987.
84. G.J. Barton & R.B. Russell, "Protein structure prediction", *Nature*, Vol. 361, pp. 505-506, 1993.
85. A.M. Lesk, *Protein Architecture - A Practical Approach*. Oxford, New York, Tokyo: Oxford University Press, 1991.
86. S. Pascarella & P. Argos, "Analysis of Insertions/Deletions in Protein Structures", *J. Mol. Biol.*, Vol. 224, pp. 461-471, 1992.
87. B. Robson, "Conformational Properties of Amino Acid Residues in Globular Proteins", *J. Mol. Biol.*, Vol. 107, pp. 327-56, 1976.
88. S.A. Benner & D.L. Gerloff, "Predicting the Conformation of Proteins: Man versus Machine", *FEBS Lett.*, Vol. 325, pp. 29-33, 1993.
89. F.D. Sönnichsen, B.D. Sykes, H. Chao & P.L. Davies, "The Nonhelical Structure of Antifreeze Protein Type III", *Science*, Vol. 259, pp. 1154-1157, 1993.
90. D. Rojewski & R. Elber, "Molecular Dynamics Study of Secondary Structure Motions in Proteins: Application to Myohemerin", *Proteins*, Vol. 7, pp. 265-279, 1990.
91. A.V. Finkelstein, A.Y. Badretdinov & O.B. Ptitsyn, "Physical Reasons for Secondary Structure Stability: α-Helices in Short Peptides", *Proteins*, Vol. 10, pp. 287-299, 1991.