PROTEINS: Structure, Function, and Genetics 19:55–72 (1994)

Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure

Burkhard Rost and Chris Sander
Protein Design Group, European Molecular Biology Laboratory, D-69012, Heidelberg, Germany

ABSTRACT **Using evolutionary information contained in multiple sequence alignments as input to neural networks, secondary structure can be predicted at significantly increased accuracy. Here, we extend our previous three-level system of neural networks by using additional input information derived from multiple alignments. Using a position-specific conservation weight as part of the input increases performance. Using the number of insertions and deletions reduces the tendency for overprediction and increases overall accuracy. Addition of the global amino acid content yields a further improvement, mainly in predicting structural class. The final network system has a sustained overall accuracy of 71.6% in a multiple cross-validation test on 126 unique protein chains. A test on a new set of 124 recently solved protein structures that have no significant sequence similarity to the learning set confirms the high level of accuracy. The average cross-validated accuracy for all 250 sequence-unique chains is above 72%. Using various data sets, the method is compared to alternative prediction methods, some of which also use multiple alignments: the performance advantage of the network system is at least 6 percentage points in three-state accuracy. In addition, the network estimates secondary structure content from multiple sequence alignments about as well as circular dichroism spectroscopy on a single protein and classifies 75% of the 250 proteins correctly into one of four protein structural classes. Of particular practical importance is the definition of a position-specific reliability index. For 40% of all residues the method has a sustained three-state accuracy of 88%, as high as the overall average for homology modelling. A further strength of the method is greatly increased accuracy in predicting the placement of secondary structure segments.**

Key words: secondary structure prediction, prediction of secondary structure class, prediction of secondary structure content, evolutionary information, multiple alignment profiles

INTRODUCTION

The Widening Sequence-Structure Gap

About 30,000 protein sequences are known (SWISSPROT release 25.0[1]). Only for about 1300 has the three-dimensional (3D)* structure (PDB[2,3]) been experimentally determined. Less than 300 of the known structures are unique.[4,5] Large-scale sequencing projects result in an explosive widening of the sequence–structure gap. It is well established that the formation of the 3D structure is determined by the sequence.[6,7] Can theory contribute to close the sequence–structure gap by predicting structure from sequence? The task is not a difficult one if the sequence of unknown structure (SOS) has a significant sequence identity to a protein of known 3D structure. In this case, modeling by homology allows an accurate prediction of the 3D structure for SOS.[8–16] Currently, this technique allows us to model the 3D structure for about 7,200 proteins.[17] For the remaining 21,500 known sequences the 3D structure cannot be predicted generally from the sequence. For short peptides molecular dynamics can be used to predict the structure.[18,19] For some cases, threading techniques can be used to find proteins of known 3D structure that have no significant sequence similarity to SOS but the same 3D structure.[10,20–30] However, for the majority of known proteins theory fails to close the sequence–structure gap by prediction of 3D structure. For these cases, in principle, the situation has not changed over the last 20 years: the goal has to be simplified.

*Abbreviations used: 3D, three-dimensional; SWISSPROT, data bank of protein sequences; PDB, protein data bank of known structures; DSSP, dictionary of secondary structures of proteins; HSSP, data base of homology derived structure of proteins; SOS, protein sequence of unknown three-dimensional structure; PHD, profile network system from Heidelberg (three levels of networks for the prediction of secondary structure).

Received October 4, 1993; revision accepted December 20, 1993.

Simplification to Secondary Structure Prediction

One example for simplification is the description of structure in terms of one-dimensional strings of secondary structure. The formation of secondary structure is important for the stability of a protein.[31,32] From the early 1970s, secondary structure had been predicted based on properties of amino acid stretches.[33–41] A decade ago, the accuracy in predicting secondary structure in three states—helix, strand, and loop—was below 60%.[42] In the 1980s, elaborated algorithms improved accuracy to above 60%.[43–51] One method that had been used without improving the performance was neural networks.[52,53] In a review of predictions by neural networks Hirst and Sternberg[54] summarized: "Recent reviews[55,56] suggest that 65% appears to be the maximum attainable performance of a variety of methods of secondary structure prediction." Why were the predictions not more accurate? One reason is that most methods used information local in sequence. However, the formation of secondary structure is determined not only by local interactions. Another reason was that the information used for the predictions was not sufficient. This made neural networks applied to the problem perform similarly to classical methods, although networks, in principle, can process higher order correlation in the patterns to be classified than information theory can. (Note: the prediction task can be formulated as a pattern classification: given a pentapeptide with a preference to form a helix, the task is to sort this peptide into the helix class.)

Structure Is More Conserved Than Sequence

Studying multiple alignments reveals that structure is more conserved than sequence.[8,57,58] In other words, proteins with different sequences can adopt the same 3D structure. What we see in alignments of native proteins is an evolutionary record of the unlikely: a pair of proteins residues evolved in nature is almost sure to have identical 3D structure if the two sequences have 30% identical residues.[8,13] Of course, not any two residues can be exchanged. On the contrary, the pattern of residue substitutions within one structure family contains specific information about the structure. A straightforward idea is to use this information for predictions.[59–61] Evolutionary information as present in multiple alignments has recently been used for the secondary structure prediction for single proteins.[62–73] To facilitate an evaluation of the performance of predictions based on multiple alignments, larger data sets have to be investigated. In an earlier study we used 130 protein chains[74,75]; recently Levin et al.[76] published an analysis based on some 60 proteins. With a neural network system the prediction accuracy could be improved for the first time above a sustained level of 70% three-state accuracy.[77]

Here, we describe an improvement of the network system (dubbed PHD). The following questions will be answered. How important are the details in a multiple alignment for the prediction? Can the performance be improved by preprocessing the profile? Can the pattern of insertions and deletions be used profitably? Does it pay to include global information such as the content of amino acids in the whole protein? How does the performance depend on the choice of the data set used for evaluation? Are neural networks particularly well suited for the task?

METHODS

Seven-Fold Cross-Validation on 126 Proteins

For the evaluation of the method 7-fold cross-validation was performed, as described earlier.[75,77,78] The data set used comprised 126 globular protein chains (Set 1, Table I) with no significant pairwise sequence identity (length dependent cut-off, e.g., $< 25\%$ for alignment lengths > 80 residues[13]). The networks were trained seven times on different sets of 108 protein chains and tested on the remaining 18, such that in the end each of the 126 proteins had been used for testing. The testing sets were deliberately chosen such that they did not reflect exactly the relative distribution of helix, strand, and loop in the data bank, to make the result less dependent on the current data bank. This was done to achieve a more general result, as the relative distribution of the data bank in 1993 might not exactly be the same as the one of the data bank at some future time. To investigate the influence of the particular choice of the data set on the result, four alternative sets were tested: (1) a set of 62 unique proteins as used by various authors,[42] (2) a set of 82 protein fragments used in a recent study on the improvement of a statistical method by using a multiple alignment,[76] (3) a set of 124 protein chains the experimentally determined structures of which were published recently (Set 2, Table II), and (4) a set of five proteins for which expert predictions were published (note: subset of Set 2). For testing the first two sets, (1) and (2), and same cross-validation from the experimental coordinates was performed as for the 126 proteins of Set 1 (Table I). For testing the second two sets, (3) and (4), the networks were trained on all proteins of Set 1, as no protein in this set has significant sequence identity to any of the proteins of Set 2 (Table II). The secondary structure assignment was done according to DSSP.[79] The 8 structure classes were converted to three states in the following way: DSSP "H" and "G" → here: helix (dubbed α or H), DSSP "E" → here: strand (β or E), and all others to loop (L).

TABLE I. 126 Protein Chains Used for Training and Testing the Networks (Set 1)*

256b_A	2aat	8abp	6acn	lacx	8adh	3ait	1ak3_A	2alp	9api_A
9api_B	1azu	3b5c	1bbp_A	1bds	1bmv_1	1bmv_2	3blm	4bp2	2cab
7cat_A	1cbh	1cc5	2ccy_A	1cd4	1cdt_A	3cla	3cln	4cms	4cpa_I
6cpa	6cpp	4cpv	1crn	1cse_I	6cts	2cyp	5cyt_R	1eca	6dfr
3ebx	5er2_E	1etu	1fc2_C	1fdl_H	1fdx	1fkf	2fnr	2fxb	1fxi_A
4fxn	3gap_A	2gbp	2gcr	1gd1_O	2gls_A	2gn5	1gp1_A	4gr1	1hip
6hir	3hmg_A	3hmg_B	2hmz_A	5hvp_A	2i1b	3icb	7icd	1il8_A	9ins_B
1l58	1lap	5ldh	2lh4	2lhb	1lrd_3	2ltn_A	2ltn_B	5lyz	1mcp_L
2mev_4	2or1_L	1ovo_A	2pab_A	1paz	9pap	2pcy	4pfk	3pgm	2phh
1pyp	1r09_2	2mhu	1mrt	1ppt	1rbp	1rhd	4rhv_1	4rhv_3	4rhv_4
3rnt	7rsa	2rsp_A	4rxn	1s01	1sdh_A	4sgb_I	1sh1	2sns	2sod_B
2stv	2tgp_I	1tgs_I	3tim_A	6tmn_E	2tmv_P	1tnf_A	4ts1_A	2tsc_A	1ubq
2utg_A	9wga_A	2wrp_R	1wsy_A	1wsy_B	4xia_A				

*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% α, 21% β, and 47% L, resolution ≤ 2.5Å for crystal structures. Nomenclature: the Protein Data Bank (PDB) identifier (first four characters) is followed by the chain identifier.

Three Levels of Neural Networks Using Profiles From Multiple Alignments

The system of networks is described in detail elsewhere.[77] Here, only the main idea is given. We used three levels of different networks (Fig. 1). First, a sequence-to-structure network: input is the profile of amino acid substitutions (as given in the HSSP data base[17]) for a stretch of 13 consecutive residues in a protein; output is the secondary structure state of the central residue (helix, strand, loop). Second, a structure-to-structure network: input is the output of the first level network for a window of 17 consecutive residues; the target output is again the secondary structure type of the central residue. The second level introduces a correlation between the secondary structure of adjacent residues, i.e., accounts for the fact that helices and strands span over at least two adjacent residues. Third, an arithmetic average over various networks (termed jury decision): input is the output of network architectures trained with different input information and with different training procedures; output is the combined prediction of all nets (Figs. 1 and 2). The networks each consisted of two layers (input-hidden, hidden-output) with 15 hidden units. The units were fully connected between the layers.

The whole system is independently trained seven times. The training procedure is the usual backpropagation algorithm[80] performed in two ways. (1) Unbalanced training: the patterns are chosen at each algorithmic time step of the error minimization at random, i.e., according to the relative distribution of helix (about 32%), strand (about 21%), and loop (about 47%). Thus, examples for loop are presented twice as often as those for strand. (2) Balanced training: the samples are predicted equally often. Consequently, an example for strand is presented 1.5 times more often than for the unbalanced training and an example for loop 1.5 times less often. (Note: a similar idea had been successfully used before for a statistical prediction method.[50])

In addition to the profile of amino acid substitutions from the multiple alignment, a conservation weight was used as input.[77] This weight places a higher weight on residues that are particularly well conserved throughout the multiple alignment (Fig. 1). This conservation weight is used on both levels, that of sequence-to-structure and that of structure-to-structure network. The alignment program[13] implicitly has a tendency to down weight similar sequences in generating the sequence profiles. A further explicit down weighting has not been investigated.

Adding Number of Insertions and Deletions to the Input

Insertions and deletions (termed indels) in multiple alignments occur more often in loop regions than in regular secondary structure elements such as helix and strand.[81,82] This implies that the number of insertions and deletions (indels) at a particular sequence position of the alignment carries information about secondary structure: the more insertions and/or deletions found in a region, the more likely it is a loop region (provided the alignment is correct). The number of indels used for the input of the first level networks was determined by adding two input units for each residue position. Thus, the input vector s for one pattern μ related to the secondary structure state of residue μ in the data set was (for a window of $w = 13$ consecutive residues):

s_{k*j} = frequency of amino acid k at position j, for $k = 1,...,21$

s_{22*j} = conservation weight at position j

$$s_{23*j} = \frac{N_{\text{ins}}(j)}{N_{\text{ali}}}$$

$$s_{24*j} = \frac{N_{del}(j)}{N_{\text{ali}}}$$

for the residues at positions $j = \mu - 6,...,\mu + 6$

TABLE II. 124 Newly Determined Protein Structures (Set 2)*

1ace, acetyl cholinesterase; 1atf, antifreeze polypeptide type A; 1col, antibacterial protein colicin A (C-terminal domain); 1cox, cholesterol oxidase; 1cpk_E; cAMP-dependent protein kinase; 1dfn_B, defensin HNP-3; 1end, glutathione synthase; 1f3g, phosphocarrier; 1gly, glucoamylase; 1gmf_A, granulocyte–macrophage colony-stimulating factor; 1hcc, glycoprotein; 16th complement control protein of factor h; 1hdd_C, engrailed homeodomain complex with DNA; 1hrh, ribonuclease H domain of HIV-1 reverse transcriptase; 1hsc, heat shock protein hsc70; 1ifb, intestinal fatty acid binding protein; 1msb_A, mannose binding protein A (lectin domain); 1nsb_B, neuraminidase sialidase; 1pi2, serine proteinase inhibitor; 1rop, ColE1 repressor of primer; 1sar_A, endoribonuclease SA, 1snv, sindbis virus capsid protein; 2fgf, basic fibroblast growth factor; 2gbl, protein G (bl domain); 2pk4, human plasminogen kringle 4; 2hip_B, high potential iron sulfur protein; 2scp_A, sarcoplasmic calcium binding protein; 2zta_A, GCN4 leucine zipper; 3trx, thioredoxin; 3znf, zinc finger DNA binding domain; 5enl, enolase; 5p21, CH-ras p21 protein (amino acids 1–166); anpc_macam, antifreeze glycoprotein type III; act1, actin (complex with DNase I); arc, Arc repressor DNA-binding protein; cdk2, cyclin-dependent kinase; csrc, sh3 domain of tyrosine kinase src; dp3b_ecoli, β-subunit of *E. coli* DNA polymerase III holoenzyme; hxk, yeast hexokinase b; luxf_phole, flavoprotein related to bacterial luciferase; nifk, nitrogenase molybdenum-iron; pik3, phosphatidylinositol 3-kinase; pdr, phthalate dioxygenase reductase; pou1, POU-specific domain; rrxa-su, retinoid X receptor αDNA binding domain; sh2, v-src tyrosine kinase transforming protein sh3: spectrin SH3 homologue domain; u1a, RNA-binding domain of U1 small nuclear ribonucleoprotein A

1aai, ricin; 1aak, ubiquitin conjugating enzyme; 1aap, protease inhibitor domain of Alzheimer's amyloid; 1abm, manganese superoxide dismutase; 1ads, aldose reductase with bound NADPH; 1aso, ascorbate oxidase; 1atn, deoxyribonuclease T complex with actin; 1baa, barley endochitinase; 1bbh, cytochrome c; 1bbk, methylamine dehydrogenase; 1bbl, e3-binding domain of the dihydrolipoamide; 1bbt, foot and mouth disease virus; 1bib, (1)biotin repressor-biotinylated lysine complex; 1brd, bacteriorhodopsin; 1cpc, c-phycocyanin; 1d66, gal4 (residues 1–65) complex with 19mer dna; 1eco, hemoglobin (erythrocruorin, carbonmonoxy); 1fba, fructose-1,6-bisphosphate aldolase; 1fha, ferritin (h-chain) mutant; 1fkt, fk506 and rapamycin-binding protein; 1gmp, ribonuclease from *Streptomyces aureofaciens* (RNase SA); 1grc, glycinamide ribonucleotide transformylase; 1grd, glucocorticoid receptor dna-binding domain; 1gst, isoenzyme 3-3 of glutathione S-transferase; 1hc6, arthropodan hemocyanin; 1hlh, helix–loop–helix domain; 1isu, high-potential iron–sulfur protein (hipip); 1izb, insulin mutant(e(b13)q; 1lig, ligand-binding domain of the *Salmonella tyrphimurium;* 1lpe, apolipoprotein-*e3 (LDL receptor binding domain); 1lts, heat-labile enterotoxin; 1lz3, lysozyme; 1min, nitrogenase molybdenum–iron protein; 1mrm, mandelate racemase; 1ms2, MS2 virus (bacteriophage); 1mup, major urinary protein complex with 2-(sec-butyl); 1nip, nitrogenase iron protein; 1ofv, oxidized flavodoxin; 1omf, matrix porin (ompf); 1ova, ovalbumin (egg albumin); 1paf, pokeweed antiviral protein; 1pdc, PDC-109 type II B-domain; 1pde, the pyruvate decarboxylase (elp); 1phg, cytochrome P-450cam from *Pseudomonas putida* camphor; 1ppb, thrombin in covalent complex with *d*-pheproarg; 1pya, pyruvoyl-dependent histidine decarboxylase (1-histidine); 1pyg, pyridoxal-5′-pyrophosphoryl derivative of glycogen; 1r09, rhinovirus 14 (HRV14) complex with antiviral agent; 1rnd, ribonuclease A; 1rve, eco rv endonuclease; 1tfg, transforming growth factor type β2 (tgf-b2); 1tie, erythrina trypsin inhibitor (kunitz) de-3; 1tlk, telokin; 1tmd, trimethylamine dehydrogenase; 1trb, thioredoxin reductase NADPH; 1ula, purine nucleoside phosphorylase, 1vaa, mhc class T h-2k_d and vesicular stomatitis virus; 2bpa, bacteriophage phix 174 capid proteins; 2cdv, cytochrome c_3; 2dpv, canine parvovirus; 2fcr, flavodoxin; 2had, haloalkane dehalogenase; 2hbg, hemoglobin (deoxy); 2hhm, human inositol monophosphatase dimer; 2mad, methylamine dehydrogenase MADH; 2pia, phthalate dioxygenase reductase; 2plv, poliovirus, 2pmg, phosphoglucomutase; 2sic, subtilisin BPN; 2sn3, protein neurotoxin variant-3; 2snv, sindbis virus capsid protein; 3chy, che*y; 3ink, interleukin-2; 3pgk, phosphoglycerate kinase; 3rub; ribulose-1,5-bisphosphate carboxylase/oxygenase; 3sc2, serine carboxypeptidase II; 3sgb, proteinase b from *Streptomyces griseus* (SGPB); 5fbp, fructose-1,6-bisphosphatase (*fru-1,6-*pase); 5nn9, neuraminidase n9

*The 124 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 Set 0 (Table I) used for training the networks (31,976 residues with 32% α, 22% β, and 46% L, resolution ≤3.5 Å for crystal structures). Nomenclature: where possible the Protein Data Bank (PDB) identifier (first four characters) followed by the chain identifier is given; otherwise the code of SWISSPROT is used.

with $N_{ins}(j)$ being the number of insertions at sequence position j of the alignment, $N_{del}(j)$ the number of deletions at that position, and N_{ali} the number of sequences in the alignment (only introduced to normalize the input units to 1.0). Note: the first term, s_{k*j}, describes 20 units for the amino acids plus one for a spacer allowing the extension of a window beyond the N- and C-terminal ends of a protein.

Adding the Global Amino Acid Composition to the Input

Sequence conservation in a multiple alignment is determined not only by local interactions. Whether or not a certain protein is evolutionarily conserved can depend on interactions between residues more than say 20 positions apart in sequence. Consequently, the residue substitution patterns introduce

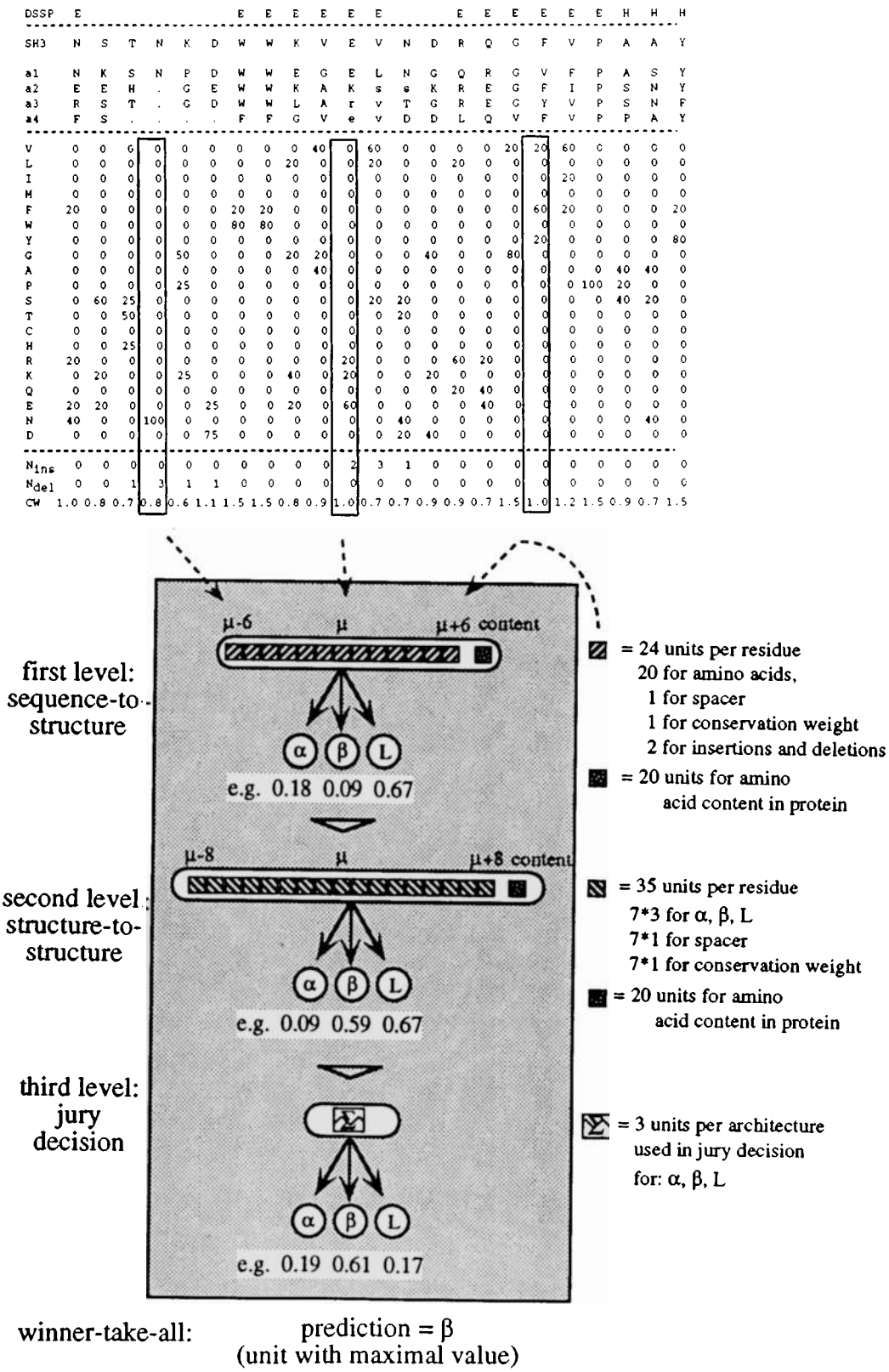

Fig. 1. Three levels of neural networks. From the multiple alignment (here guide sequence SH3 plus 4 other proteins a1–a4, note: lower case letters indicate deletions in the aligned sequence) a profile of amino acid occurrences is compiled. To the resulting 20 values at one particular position μ in the protein (one column) three values are added: the number of deletions and insertions, and the conservation weight (CW). Thirteen adjacent columns are used as input. The whole network system for secondary structure prediction consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks.

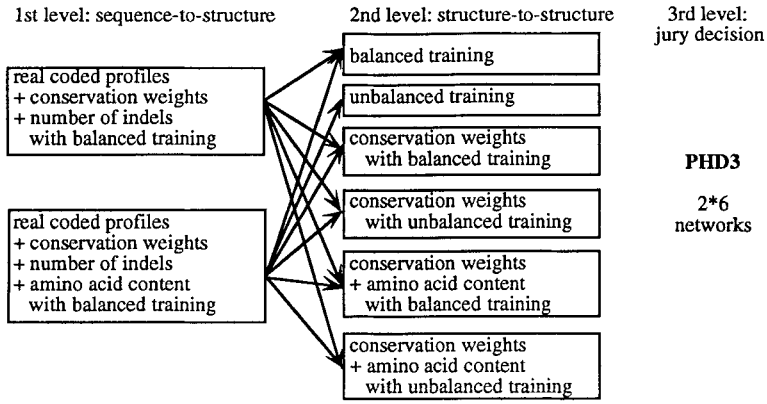

Fig. 2. PHD3: jury on 12 different neural networks. The ultimate network system mixing the information from all specific compilations of the profiles consists of 12 differently trained second-level networks. The quadrangles give the input of each of the networks used. The profiles of amino acid substitution were projected onto an interval of 0 to 1 (number of occurrences in multiple alignment divided by number of sequences). For the input to the first level network, these 20 numbers can be coded each by multiple binary (0 or 1) input units (binary coding), or each by a real number between 0 and 1.[77] For the ultimate system (labeled PHD3) only real coding was used at the first level. The input to the second level was coded in all cases by seven binary units per real number.

implicitly nonlocal information. However, explicitly only local information was included so far. The restriction to local information is one disadvantage of today's secondary structure prediction methods.[83,84] One global aspect available on the level of sequence is the amino acid composition of a protein. Such information has been used to predict the protein's structural class, like all-α, all-β, α/β (next section).[85–88] Here, we coded the amino acid composition by 20 additional units on both levels of networks:

$$s_{v+k} = \frac{N_{AA(k)}^{all}}{N} - \frac{N_{AA(k)}^{window}}{w}, \quad \text{for } k = 1,\ldots, 20$$

with N being the number of residues in a protein, $N_{AA(k)}^{all}$ the number of amino acids of type k in the whole protein, and $N_{AA(k)}^{window}$ the number of amino acids in the window of w consecutive residues used for one input vector. The additional units are added to the first and second network level (Fig. 1). The coding of amino acid content starts at the vth unit of the input. For the first level sequence-to-structure network with a window of $w = 13$, $v = 13 \times (20+1+1+2) = 312$ (20 units for the amino acids, 1 for the spacer, 1 for the conservation weight, and 2 for the number of indels). For the second level structure-to-structure network with a window of $w = 17$, $v = 17 \times 7 \times (3+1+1) = 512$ (3 units for the three output units of the first level, i.e., the three secondary structure types, 1 for the spacer, 1 for the conservation weight, and the factor 7 for coding each of the five real numbers by 7 binary units[77]). The different neural networks used for the final jury decision of the ultimate network system are sketched in Figure 2.

Prediction of Structural Class

The knowledge of secondary structure content might provide useful boundary conditions for the theoretical as well as the experimental determination of protein structure. It can directly contribute to the assessment of the folding type of a new protein. One experimental way to estimate secondary structure content is circular dichroism spectroscopy.[89] The accuracy of the secondary structure content predicted by the network system can simply be calculated as the difference between observed and predicted content averaged over all protein chains.[42,77] Levitt and Chothia[90] have pointed out that proteins fall into well-defined structural classes. Figure 3 shows that such a classification is not clear cut. Here, we used the classification according to Zhang and Chou[87] as given in the legend to Figure 3. Knowledge about function,[91,92] protein class,[93–97] or overall secondary structure content[39] has been used by others in the hope of improving secondary structure prediction methods. For proteins of unknown structure the structural class is, of course, not known. Our experience is that there is no practical advantage in training on specific structure classes, given the margin of error in identifying the structural class of a protein.[98] But how accurate is the prediction of structural class? The task is made difficult by the fact that the classification is not clear cut (Fig. 3). Small errors in prediction of secondary structure content result in false classifications. What could be achieved if homology modeling were possible? An analysis of 140 alignment pairs[90] shows that even a method allowing accurate prediction of the 3D structure reaches only a value of 90% (all-α, 96%; all-β, 70%) in classifying proteins into

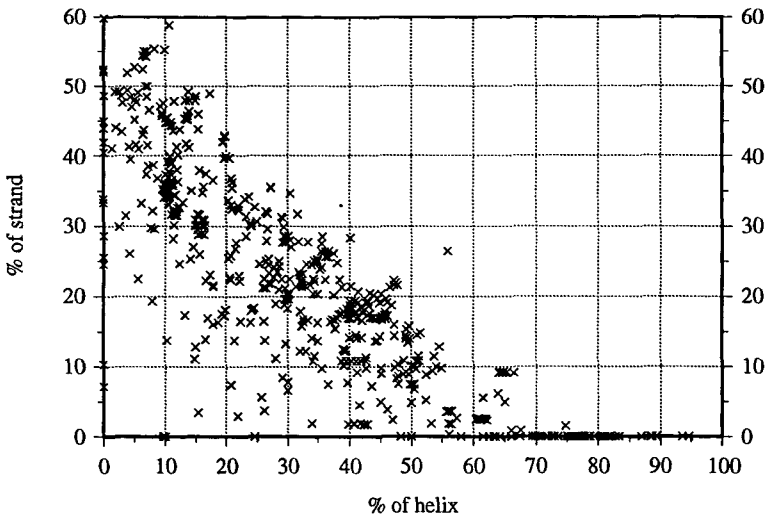

Fig. 3. Content of helix vs. content of strand for 1000 PDB proteins. For all proteins in PDB (release 24.0[2]), i.e., about 1000, the content of helix and strand is compiled according to DSSP.[79] Distinct chains are not averaged separately. The division lines are apparently not clear-cut. One possible classification scheme is[87] all-α, %α ≥ 45% and %β < 5%; all-β, %α < 5% and %β ≥ 45%; mix, %α ≥ 30% and %β ≥ 20%.

four structural classes (the measure is defined in the legend to Fig. 7).

RESULTS

The Details of the Multiple Alignment Are Important

The first question arising from the use of the information in multiple alignments is how important are the details of the alignment? We compared two different coding schemes: binary vs. real. The binary coding projects the profile entries (percentages: 0–100, Fig. 1) onto four intervals: 0–2, 3–33, 34–66, and 67–100. The result of a three level network using binary coding in terms of overall accuracy is 2 percentage points inferior to a network using real coding (Table III, "binary profiles" vs. "PHD0"). For the example in Figure 1 with 5 sequences in the alignment, this result is not surprising, since, e.g., a value of 40 or 60 results in the same input signal. But what if the number of the aligned sequences is higher, say 50. Does it really make a difference to have 50 instead of 55? Only one-third of all proteins from PDB can be aligned to less than 20 proteins (HSSP data base, Reinhard Schneider, private communication). Roughly the same figure holds for the proteins of Set 1. Thus, the gain by real coding shows that indeed the fine details of the alignments are important for the prediction.

The Addition of a Conservation Weight in PHD1 Improves Performance

The second question is of a more technical nature: is the network capable of extracting all the important information from the profiles, or does it facilitate the prediction task to additionally add information such as the conservation weight that is derived from preprocessing the profile? The result is that the addition of the conservation weight to the input of the first level sequence-to-structure and of the second level structure-to-structure network adds half a percentage point in terms of overall accuracy. The results for PHD0 (using only the profile) given in Table III are an average over the 126 proteins of Set 1 plus 4 membrane chains: the comparable average for PHD1 (using additionally the conservation weight) is Q_3 = 70.2%, i.e., 0.5 percentage points higher than the one for PHD0.

The Addition of Indel Information in PHD2 Improves the Overall Accuracy and Reduces the Tendency for Overprediction

The next question was: can the patterns of insertions and deletions be used to further improve the prediction? As mentioned above these patterns carry, in particular, information about the occurrence of loop regions. Thus, what one expects is that the prediction of loop becomes more accurate. Indeed, using the number of indels on the first level (sequence-to-structure network) improves the accuracy for loop from 72.3 to 76.9% (data not shown). The overall result is improvement in accuracy to 71.4% (Table III, PHD1 vs. PHD2) and reduction of the tendency for overprediction. The latter corresponds to a significant decrease in the accuracy of correctly predicting observed helices and strands (2–4 percentage points reduction in $Q_\alpha^{\%obs}$ and $Q_\beta^{\%obs}$, Table III). At the same time helices and strands predicted have a higher probability of being correct ($Q_\alpha^{\%pred}$ and $Q_\beta^{\%pred}$ increase by 1–5 percentage points, Table III). This means that the tendency of

TABLE III. Prediction Accuracy for Various Networks*

		Method				
	Reference net	Binary profiles†	PHD0†	PHD1	PHD2	PHD3
Q_3	62.1	67.6	69.7	70.8	71.4	71.6
I	0.13	0.17	0.25	0.25	0.27	0.27
$Q_\alpha^{\%obs}$	57	70	70	72	68	70
$Q_\alpha^{\%pred}$	60	70	72	73	78	76
C_α	0.40	0.56	0.58	0.60	0.62	0.61
$Q_\beta^{\%obs}$	41	66	64	66	64	62
$Q_\beta^{\%pred}$	53	57	57	60	61	63
C_β	0.35	0.47	0.50	0.52	0.52	0.52
Sov_3^{obs}	61.8	—	—	72.4	72.7	72.8
Sov_3^{pred}	54.5	—	—	67.6	67.7	67.9
Sov_α^{obs}	59.0	—	—	75.9	73.9	75.1
Sov_α^{pred}	48.7	—	—	70.9	74.4	73.3
Sov_β^{obs}	56.9	—	—	73.4	73.9	72.0
Sov_β^{pred}	48.5	—	—	65.4	66.6	67.8

*Given are average over 7-fold cross-validation on the 126 protein of Table I. Note: the results marked with a dagger relate to the 126 globular soluble proteins of Table I plus 4 chains of the photoreaction center. The inclusion of the four trans-membrane chains results in an average overall accuracy roughly 0.5 percentage points lower than the average over exclusively globular proteins (e.g., for PHD0 the average over 130 proteins is 70.2%).

Abbreviations for networks: reference net, only first level sequence-to-structure net with unbalanced training using single sequences instead of profiles from multiple alignments as input (similar to the networks evaluated by others on smaller data sets[52,53]). The other networks give the results for the three level system; "binary," the substitution profiles (values between 0 and 1) input to the first level are coded by 4 binary (0 or 1) units each. For the three methods labeled PHD the profiles are coded by real units (values between 0 and 1). "PHD0" uses only the profiles as input; "PHD1" uses additionally conservation weights. "PHD2" adds the number of indels on the first level. Only "PHD3" uses explicitly global information as given by the amino acid content (Fig. 2).

Per-residue measures (all values given in percent; for a more detailed definition see Rost and Sander[77]): Q_3, residues predicted correctly in 3 states (helix, strand, loop) divided by all residues; $Q_\alpha^{\%obs}$, correctly predicted residues in helix divided by observed residues in helix; $Q_\beta^{\%obs}$, same as previous for strand; $Q_\alpha^{\%pred}$, correctly predicted residues in helix/strand divided by predicted residues in helix/strand; $Q_\beta^{\%pred}$, same as previous for strand; I, information measure defined by

$$I = 1 - \frac{\sum_{i=1}^{3} a_i \ln a_i - \sum_{ij=1}^{3} A_{ij} \ln A_{ij}}{N \ln N - \sum_{i=1}^{3} b_i \ln b_i}$$

where N is the number of residues in the data bank, a_i the number of residues predicted to be in secondary structure i, b_i the number of residues observed to be in i, and A_{ij} the number of residues predicted to be in i and observed to be in j.[77] C_α is the Matthew correlation coefficient for helix and C_β that for strand.[125]

Segment based measures (given in percent; for explicit discussion see Rost et al.[99])

$$Sov = \frac{1}{N} \sum_s \frac{\text{minov}(s_1; s_2) + \delta}{\text{maxov}(s_1; s_2)} \text{len}(s_1)$$

where N is the total number of residues, s_1 and s_2 are two secondary structure segments (one from the observed string of secondary structure, the other from the predicted string), and $\text{len}(s_1)$ is the number of residues in the segment of sequence 1. The sum is taken over all segment pairs $s = \{s_1, s_2\}$. The actual overlap between the two segment is minov, i.e., the number of residues for which both segments have, e.g., en H (helix) in common; maxov is the total extent of both segments, i.e., the number of residues for which either of the two has, say, the assigned state H. The accepted variation δ assures a ratio of 1.0 when there are only minor deviations at the ends of segments; it is chosen to be smaller than minov and smaller than half the length of segment s_1. The ratio minov/maxov is constrained to a maximum value of 1.0, i.e., the allowance cannot lead to a "more than perfect" value of fractional overlap, Sov_3 gives values for 3 states, Sov_α those for helices, and Sov_{gb} those for strand. The superscript "obs" indicates that the length of the observed segments was used for weighting (likelihood that an observed segment is correctly predicted). In contrast, "pred" labels the weighting by the length of the predicted segments (likelihood that a predicted segment is correct).

the network to overpredict helices and strands is reduced, a desirable effect. A curious side effect is that the indel information has more influence on the prediction of helices than on strands (differences between PHD1 and PHD2 higher for helices, see Table III).

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. This observation cannot be explained clearly. Two potential reasons are the following. First, indel information might be more important on the sequence than on the structure level. The conservation weight allows the net-

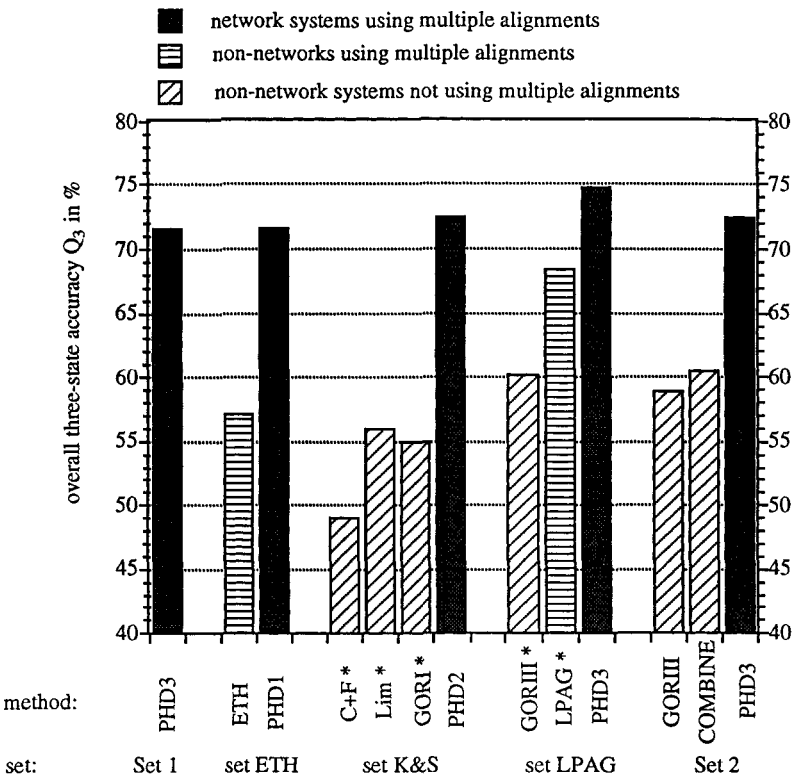

Fig. 4. Overall three-state accuracy for various predictions. The overall three state accuracy of the network systems is compared to various predictions. Set 1 and set 2 are given in Tables I and II. "ETH5" labels the five proteins for which an expert prediction from Gerloff et al.[69] is comparable, "K&S" is the set of 62 proteins as used by Kabsch and Sander,[42] and "LPAG" the one used by Levin et al.[76] The results marked with an asterisk were taken from the literature: "C + F" (Chou–Fasman[33]), "Lim,"[40] and "GORI"[39] are given in Kabsch and Sander[42]; "GORIII" and "LPAG" for set LPAG are from Levin et al.[76] (other abbreviations as in Table III).

work to focus on particular residues of the protein. Thus, the information added by using additional units for the conservation weight is not restricted to the sequence level. This might not as clearly be the case for indels. The number of indels is strongly correlated to sequence information. Thus, the inclusion of indel units pays off only on the first level of sequence-to-structure network. Second, the coding scheme used is not optimal. The second level structure-to-structure network codes one residue position by 5 units (helix, strand, loop, spacer, conservation weight). Indels contribute another 2 units. The network might fail to learn that the first 3 units are less important than the last 4. In other words, the indel information might dominate the structure information.

The Addition of the Global Amino Acid Content in PHD3 Has a Marginal Influence in Terms of Local Measures

Does it pay to explicitly include global information to the input, such as the amino acid content in the entire protein? The inclusion of global information by using the amino acid content of the protein (outside of the window) presented as additional input influences the local measures like the overall accuracy only marginally: $Q_3 = 71.6\%$ (Table III, compared to 71.4% for PHD2). The tendency to better predict loop regions (than PHD1 does) is maintained, as well as the tendency for less overprediction in helix and strand. The improvement in overall accuracy is mainly caused by a more accurate prediction of helices (Table III). However, the dominant effect of using global information is a significant improvement of prediction in terms of global measures like the correctness of the predicted content of secondary structure. This point will be discussed in detail below. (Note: including the length of the protein as additional input did not yield further improvement. This again might partly be attributed to a nonoptimal coding scheme.)

From 62.1 to 71.6%—Untangling the Contributions to the Improvement

The three level network system is already *rather* complex. Which information contributes which fraction to the improvement of prediction accuracy? Using a standard neural network consisting of one

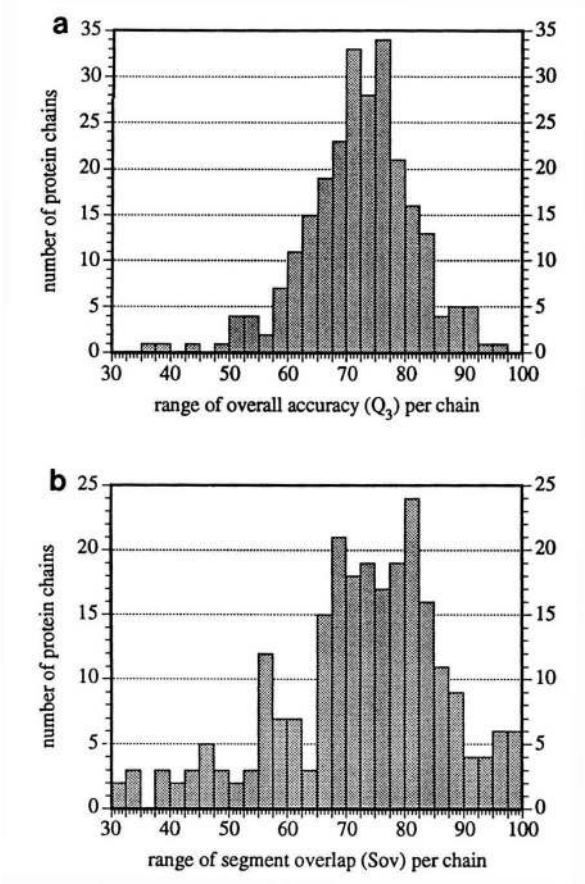

Fig. 5. Expected variation of prediction accuracy with protein chain. The distribution of the per protein chain three state accuracy can be interpreted as the expected variation of prediction accuracy for protein sequences of unknown structure. Given is the distribution over all 250 protein chains from Set 1 (Table I) and Set 2 (Table II). (**a**) The average in per-residue accuracy (Q_3) over all chains is 72.2% with a standard deviation of 9.3%. (**b**) The average in segment overlap (Sov, as defined in Table III) is 72% with a much larger standard deviation of 15.8%.

layer with single sequences[52,53,93,95,100–105] results in an accuracy of 62.1% on Set 1 (dubbed reference net in Table I). A comparable first level network using profiles, conservation weight, indels, and amino acid composition scores at an overall accuracy of 69.5%. Thus, the multiple sequence input information accounts for roughly 7.5 percentage points of the improvement. The best network of the second level reaches 70.6%. Such a network trained on both levels by presenting the patterns during the training according to their relative occurrence in the data set (unbalanced training) shares with the reference network a low accuracy in predicting strand regions ($Q_\beta^{\%obs}$ = 51%). Balanced training (presenting the three secondary structure types equally often during training) results in a lower overall accuracy (Q_3 = 68.5% on the first level and 69.1% on the second), but a substantially more accurate prediction of strand ($Q_\beta^{\%obs}$ = 65%). The final decision improves the overall accuracy from 69.1–70.6% on the second level to 71.6%. The introduction of the second level (structure-to-structure) network yields a further improvement not revealed by the per-residue scores: the predictions look by far more protein-like, i.e., the average length of the predicted secondary structure segments is similar to the predicted averages (for helix: predicted 9.2 residues per segment, observed 9.1; for strand: predicted 4.8, observed 5.1). This is reflected by the increase of prediction accuracy in terms of the segment overlap (Table III).

Overall Accuracy Above 72% Evaluated on 250 Unique Protein Chains

How much does the result depend on the choice of data set used for evaluation? One difficulty in the literature on secondary structure prediction methods has always been that the data sets used for evaluation were too small and/or allowed for significant sequence identities between the proteins used for developing the method and those used for evaluating it (insufficient cross-validation). In addition, different authors use different sets. Given the variation of prediction accuracy between different protein chains (Fig. 5), it is possible to improve the prediction accuracy—by chance or deliberately—by selecting an even larger set of sequences for which the accuracy is higher, e.g., leaving out the worst predicted proteins in Fig. 5 results in an average >80% for the best 50 proteins. How representative are the 126 proteins of Set 1? We performed 7-fold cross-validation tests on two other sets: first, a set of 62 proteins (labeled "K&S" in Table IV) used for a comparative study of the quality of secondary structure prediction a decade ago,[42] and second, a set of 82 protein fragments (labeled "LPAG" in Table IV) used in a recent study on the improvement of classical prediction methods by use of alignment information.[76] For these the network system scores 1–3 precentage points higher in overall accuracy than for the set of 126 proteins (Table I) discussed so far. This analysis allows for a direct comparison of the network system with methods of secondary structure prediction still widely used: first the Chou and Fasman[33,34] algorithm with an overall accuracy of 49%[42] compared to 72.5% of PHD2 (Table II, Fig. 4) and second, the GORIII method[47] with an overall accuracy of 60.2%[76] compared to 74.8% of PHD3 (data not shown).

Will the method score equally high for the next 100 proteins? Since we first asked this question[77] many experimentally determined structures became available. The result for Set 2 containing 124 new proteins (Table III) is surprisingly even better than that for Set 2. The overall accuracy of PHD3 becomes 72.5%. None of the 124 new proteins of Set 2 had any significant sequence identity to any of the proteins in Set 1. It is very likely that none of the proteins from Set 2 has significant sequence identity

TABLE IV. Performance for Various Prediction Methods*

	Method										
	COM-BINE	ETH	PHD1	C+F†	PHD2	LPAG†	PHD3	GORIII	COM-BINE	S83	PHD3
Set	"5"	"5"	"5"	"K&S"	"K&S"	"LPAG"	"LPAG"	Set 2	Set 2	Set 2	Set 2
N_{prot}	5	5	5	62	62	60	60	124	124	124	124
Q_3	57.6	57.2	71.6	49	72.5	68.5	74.8	58.9	60.9	61.1	72.5
I	0.09	0.14	0.29	—	0.30	—	0.34	0.10	0.12	0.13	0.28
$Q_\alpha^{\%obs}$	73	50	69	42	68	—	70	57	68	60	71
$Q_\alpha^{\%pred}$	56	80	83	45	80	—	81	57	56	60	78
C_α	0.38	0.49	0.62	—	0.64	—	0.67	0.37	0.42	0.42	0.64
$Q_\beta^{\%obs}$	17	47	55	52	71	—	78	39	30	44	62
$Q_\beta^{\%pred}$	58	48	74	35	63	—	69	48	59	50	65
C_β	0.23	0.30	0.55	—	0.57	—	0.62	0.30	0.32	0.33	0.53
Sov_3^{obs}	57.3	63.2	76.2	—	72.7	—	76.8	58.8	61.4	62.2	73.9
Sov_3^{pred}	51.3	57.3	72.2	—	69.5	—	74.7	50.2	55.5	58.0	69.4
Sov_α^{obs}	69.7	61.1	75.4	—	71.5	—	70.3	57.0	68.1	62.4	75.1
Sov_α^{pred}	48.9	81.5	80.8	—	77.8	—	76.5	44.4	50.2	57.7	75.0
Sov_β^{obs}	26.5	55.4	69.7	—	74.4	—	83.3	53.0	45.7	53.6	72.5
Sov_β^{pred}	45.7	59.4	77.0	—	69.6	—	76.5	42.3	52.0	53.2	68.5

*For different protein sets the performance of network systems is compared to alternative prediction algorithms:

Set "5": set of 5 proteins for which a comparison is possible to the expert prediction "ETH," from Gerloff et al.[62–64,69,126,127] The proteins are 1cpk chain E, cAMP-dependent protein kinase[128]; ksrc_avisr, SRC tyrosine kinase[129]; spcn_chick, SH3 domain of spectrin[130]; p85b_human, phospatidylinositol 3-OH kinase[106,107]; and nifk_azovi, molybdenum-iron nitrogenase.[131] "COMBINE" gives for comparison the performance of a classical prediction not based on alignment information.[46]

Set "K&S": set of 62 unique proteins first used by Kabsch and Sander for a comparative assessment of prediction accuracy.[42] C + F gives the result of the Chou–Fasman prediction[33,34] evaluated on this set (values taken from Kabsch and Sander[42]).

Set "LPAG": set of 82 protein fragments (from less than 20 structure families), used for prediction of secondary structure based on multiple alignments with an information theory approach.[76] LPAG gives the result the authors published for the performance of their method using the information from multiple sequence alignments (Levin et al.,[76] and Garnier, private communication).

Set 2: see Table II: "GORIII" and "COMBINE"[46] are standard information theory algorithms (both methods were published to reach Q_3>63% based on an evaluation on a smaller data set[55]). "S83" (Segment 83) is an unpublished method.[42]

The measures and the abbreviations for the networks are explained in the footnote to Table III. Note: values marked with a dagger were taken from the literature.

to any protein used to develop earlier prediction methods, thus these methods can be evaluated directly on Set 2. This we did for three methods available to us: Segment 83 (Kabsch and Sander, unpublished), GORIII,[47] and COMBINE.[46] All three classical methods have an overall accuracy more than 10 percentage points lower than that of PHD3 (Fig. 3, Table IV).

The Expected Accuracy for a New Protein Varies Between 63 and 81%, But Reliably Predicted Regions Can Be Identified

All these numbers might be less interesting for a potential user of the prediction who only wants to know: how good is the prediction on a new protein of unknown structure, say a protein called SOS? The discouraging message is, for the 250 proteins in Set 1 and Set 2 the standard deviation is as large as 9.3% (Fig. 5). But the prediction is significantly worse in isolated cases. 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. Two recent examples of failure of the prediction method are the phospatidylinositol 3-OH kinase p85_human (PIK3)[106,107] and the antifreeze protein type III anpc_macam,[108] both predicted at low accuracy of about 40%. Thus, there is a small but nonvanishing chance that the prediction for SOS is grossly wrong.

A more encouraging result, in addition to the relatively high average accuracy, is that the network prediction allows the identification of regions which are predicted with higher reliability (reliability index defined in caption of Fig. 6). An impressive 40% of all residues are predicted at an expected accuracy of 88% (Fig. 6). This accuracy is comparable to what can be expected if homology modeling were possible for SOS.[99] Thus, it is possible to estimate from the prediction whether or not the result for SOS is likely to be worse than the average over all proteins (72.3%): if the reliability index is ≥6 for more than

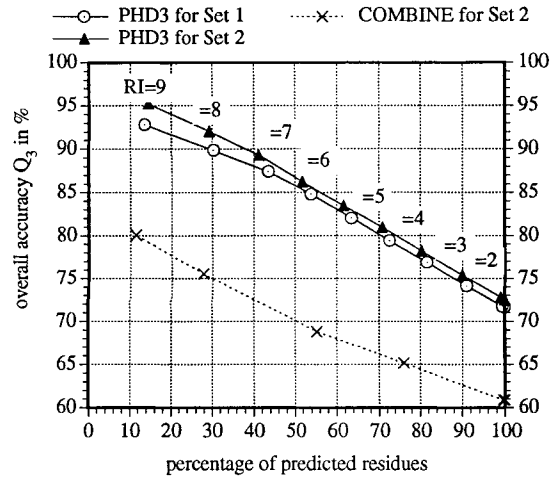

Fig. 6. Expected prediction accuracy for residues with a reliability index above a given cut-off. Plotted are averages of the three state accuracy over all those residues with reliability index RI > n, n = 0, . . . , 9. This index is simply defined by:

$$RI = \text{INTEGER}\,[10 \times (out_{max} - out_{next})]$$

where out_{max} is the output of the output unit with highest value, and out_{next} that of the unit with the next highest value. The factor 10 normalizes RI to integer values from 0 to 9. RI = 9 corresponds to a rather reliable prediction. Shown are the results for two different evaluation sets (Set 1 and Set 2), e.g., about 30% of all residues have RI > 7 and of these about 90% are correctly predicted by PHD3.

half of the residues of SOS then the accuracy can be expected to be higher than average 72% (Fig. 6).

The Neural Network Method vs. Statistical and Expert System Methods

Hirst and Sternberg[54] concluded: "The applications of neural networks to problems in protein sequence analysis, although interesting, have not yielded significant improvements over other current methodologies." They then speculated that this might change with growing databases or the incorporation of other information. This raises the question whether the time they foresaw has already come. Are neural networks particularly well suited for the task of predicting secondary structure from profiles of multiple alignments? The answer has to remain preliminary because of two reasons. First, by the end of 1993 only two analyses of nonnetwork methods using information from multiple sequence alignments have been evaluated on larger data sets (see below). The expert predictions based on multiple alignments are available for some examples, only. Second, our network system is perhaps not the optimal solution for the problem. We managed to improve the performance by some 5–7 percentage points over the last 2 years. Others might continue on this road in the future.

Which are the comparable methods for predicting secondary structure based on the information from multiple alignments? One attempt is to use the information theory as incorporated into the GOR[37–39,47,109] method. A recent publication simply deduces the prediction from an average over the predictions for each sequence in the multiple alignment.[76] An alternative method compiles an average over the information content of the multiple alignment (Altenberg and Sander, unpublished). Both methods reach overall accuracies clearly below 70%. Levin et al. published a value of 68.5% for using sequence alignments. They report as well a result of 69.6% based on the alignment of C^{α} traces, but for proteins of unknown structure the C^{α} traces are of course also unknown, consequently, C^{α} alignments cannot be obtained. If the network system (PHD3) is evaluated with 7-fold cross-validation on the same data set as used by Levin et al.,[76] the three-state overall accuracy rises to 74.8% (Table IV, Fig. 4). Thus, the nonnetwork system is some 6 percentage points inferior to the network system using the same information.

Another comparison is the one between the network system and an expert system. This comparison continues to be overemphasized in the literature in the sense that it is based on very few cases, and "one swallow does not make a summer."[110] The variation of the prediction accuracy on small data sets is substantial (Fig. 5). The overall accuracy of the expert prediction by Gerloff et al.,[69] averaged over 5 proteins, is 57.2% (that of a statistical method not using multiple alignments is 57.6%, Table IV). On the same set the network system scores at 71.6% (Table IV, Fig. 4). These comparisons show that the network system is indeed very suitable for incorporating the additional evolutionary information contained in the multiple sequence alignments.

Prediction of Structural Class Comparable to Methods Specialized on This Task

How accurately can secondary structure content be predicted? As a simple measure we used the differences between observed and predicted secondary structure content averaged over all proteins of Set 1. The results is that the additional input of the global amino acid content incorporated in PHD3 yields a more accurate prediction of the secondary structure content and consequently of the structural class (Table V). How do these result compare to other methods?

First, the prediction of secondary structure content: Muskal and Kim[85] allowed for homology between testing and training set. Consequently, their method has to compete with homology modelling (Table V), whereas the values given for the networks have to be compared to random prediction (Table V). Additionally, the test set used by Muskal and Kim contained less than 20 proteins. The performance of PHD3 on the best of the seven test sets was about the same as the one reported by Muskal and Kim,

TABLE V. Prediction of Secondary Structure Content and Structural Class*

	Method						
	PHD2	PHD3	PHD3†	GORIII†	COMBINE†	HM	RAN
Δhelix	9.0 ± 8.3	8.5 ± 8.0	7.8 ± 6.8	11.3 ± 9.4	11.2 ± 9.0	2.8 ± 3.8	32.1 ± 20.8
Δstrand	7.6 ± 8.0	7.5 ± 8.1	7.3 ± 7.9	10.6 ± 9.8	13.2 ± 10.6	2.7 ± 3.2	21.3 ± 14.5
$C_\alpha^{\text{Pearson}}$	0.86	0.87	0.91	0.78	0.83	0.97	−0.36
C_β^{Pearson}	0.74	0.74	0.73	0.46	0.51	0.97	−0.22
All-α	80.0	85.7	94.1	85.7	66.7	94.1	0.0
All-β	66.7	50.0	0.0	0.0	0.0	86.7	0.0
α/β	33.3	50.0	55.6	50.0	0.0	100	0.0
Rest	72.0	74.1	74.5	65.8	67.7	89.7	71.2
Q^{class}	70.6	74.6	75.8	66.1	66.1	90.0	44.7

*"HM" labels a prediction by homology modeling and "RAN" a random prediction[99]; the other methods are as in Tables III and IV. The results marked with a dagger refer to Set 2; the others refer to Set 1 (or comparable sets as for HM and RAN). The following measures are listed:

$$\Delta\text{content}_i = \sum_{\mu}^{\text{all chains}} |\text{content}_{i\,\mu}^{\text{obs}} - \text{content}_{i\,\mu}^{\text{pred}}|, \qquad i = \alpha,\beta$$

with $\text{content}_{i\,\mu}^{\text{obs}}$ being the observed content of secondary structure of type i in protein chain μ, and $\text{content}_{i\,\mu}^{\text{pred}}$ the predicted content. Pearson correlation coefficient:

$$C_i^{\text{Pearson}} = \frac{\langle x\,y\rangle - \langle x\rangle\langle y\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2}\sqrt{\langle y^2\rangle\langle y\rangle^2}}, \qquad i = -\alpha,\beta,\text{L}$$

where x and y are the observed and predicted content of the secondary structure of type i (here in three states: helix, strand, loop). Brackets $\langle\ \rangle$ indicate the average over all proteins under investigation. The percentages of proteins predicted correctly to be in either of the four structural classes: all-α, all-β, α/β, and rest is given by

$$\text{class} = \frac{\text{number of chains predicted correctly in class } i}{\text{number of chains predicted in class } i}$$

The total average over all four classes is defined as

$$Q^{\text{class}} = \frac{\sum_{i=1}^{4} \text{number of chains predicted correctly in class } i}{\text{number of chains predicted}}.$$

although the network is not specialized on predicting secondary structure content, and although we did not allow for any significant sequence homology. The results of a network system also compares favorably with those from circular dichroism spectroscopy.[77]

Second, comparison of the prediction of structural class: other authors allowed for homologies between test and training sets. Among various analyses[87,88,111–116] there is only one[88] that reports a thorough cross-validation on a set of 64 proteins. Again, the authors allow for homologies between training and testing set proteins. In spite of this, their results are not better than those given here (Table V).

Summing up, PHD3 classifies 75% of the 250 proteins correctly into one of the four classes all-α, all-β, α/β, and rest (Fig. 7). This is at least comparable to the results obtained by other methods that specialize on this task but do not used the information of multiple sequence alignments. One reason for the prediction of structural class is the hope to improve the prediction of secondary structure by using the information about structural class as input. We showed earlier[98] that the difference between a network trained specifically on all-α proteins and another one trained on proteins of all classes is only marginal. Given the additional error margin in finding the structural class, the balance shifts in favor of the training on proteins of all structural classes. Note, however, that prediction by homology is clearly superior to any other method, including an experimental estimate of secondary structure content by circular dichroism spectroscopy. Figure 7 illustrates that even for a prediction by homology with correlation coefficients close to 1, the error in predicting the structural class is some 10%. This error is mainly caused by the fact that the line between the four classes is not clear cut (Fig. 3).

CONCLUSIONS

The final network system PHD3 described here reaches an overall accuracy of 71.6% on Set 1 and 72.5% on Set 2. The network scores at a sustained level of higher than 72% on the combined set of 250 unique globular soluble protein chains, which in total have about 57,000 residues. The best score published for a nonnetwork method using sequence alignments is 68.5%.[76] On the same set of some 82 protein fragments PHD3 scores at 74.8% (Table IV, Fig. 4). Another comparison to a method using information from multiple alignments (based on a

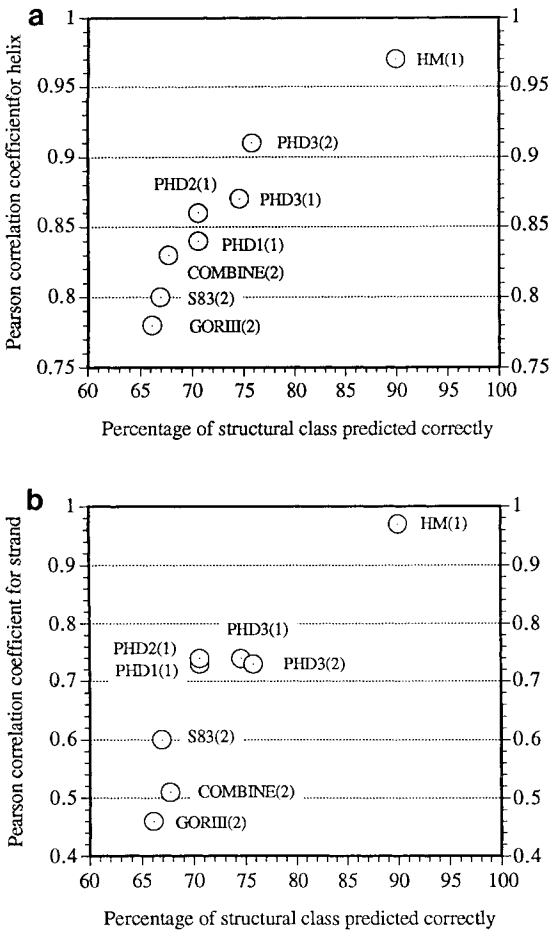

Fig. 7. Accuracy in predicting structural class vs. Pearson correlation coefficient for helix (**a**) and for strand (**b**). The Pearson correlation coefficient is often used when assessing the success of circular dichroism estimates.[89,132–135] It measures the correlation between predicted and observed content of secondary structure (defined in Table V). Typical values for circular dichroism spectroscopy are helix, 0.84; antiparallel sheet, 0.41; parallel sheet, 0.37; and loop, 0.56.[133] The values for the correctness of predicting structural class relate on the four classes as given in Figure 3 (definition of measure in Table V). "HM" labels a prediction by homology modeling. The other methods are given in the footnote to Table III (in parentheses: number indicating whether Set 1 or Set 2 was used).

very small set of 5 proteins) for the expert predictions of Gerloff et al.[69] looks equally favorable: 57.2 vs. 71.6% (Table IV, Fig. 4). Classical methods like Chou–Fasman and GORIII evaluated on the same data set yield overall accuracies 23 to 13 percentage points inferior to PHD3. Thus we conclude that the time that "the power of neural networks may be exploited in the analysis of protein sequences"[54] has come already.

Given a difference of more than 20 percentage points to the Chou–Fasman method, is this a substantial improvement? Prediction methods have to be evaluated in relation to what the worst and the best possible predictions yield. The lower limit can be given by a random prediction (about 35%) and the upper limit by a prediction based on homology building of the 3D structure, possible if a sequence homologue of known 3D structure to the new protein (SOS) exists (88.4%).[99] The span in between these two values is accessible to methods predicting secondary structure from the sequence. Normalizing the overall accuracy such that a random prediction yields 0% and a prediction by homology 100% reveals that the network system presented here is almost three times as accurate as the prediction of Chou–Fasman, and substantially better than any other prediction method (Fig. 8). Of practical interest is the definition of a reliability index. Alternative methods enabling the definition of a reliability index predict about 2–4 times fewer residues at a given accuracy (on not comparable data sets).[47,52,117,119]

The segment-based accuracy exceeding 72% (as measured in the segment overlap defined in Table III) proves that the network system presented here is relatively more successful in producing protein-like predictions than previously used neural networks: the reference network (Table III) yields a lower segment than per-residue score (Table III), indicating that the prediction is not as similar to what is observed in globular proteins.

Using amino acid content as additional input does not result in a significant increase of the overall accuracy (PHD2 vs PHD3). However, the network incorporating this global information (PHD3) is clearly superior to the one not using it (PHD2) in predicting the global content of secondary structure and consequently the structural class of a protein (Table V, Fig. 7). Of all 250 proteins 75% are correctly classified into one of the four classes: all-α, all-β, α/β, and rest.

The analysis given here is based on a data set of 250 unique proteins. Will the result hold for the next 250, the structure of which will be experimentally determined? Long-term use of PHD will provide an answer (the method is available for fully automatic use,[120] send the word *help* by electronic mail to the internet address *PredictProtein@EMBL-Heidelberg.de* for detailed instructions). Indeed, secondary structure prediction as presented here is only successful in predicting the clichés contained in the data bank. But what are these clichés? Given a novel fold, will PHD be able to correctly predict the secondary structure? In principle it should not, but what is a novel fold? An example is the recently solved flavoprotein related to the subunit of bacterial luciferase luxf_phole[121] that has been presented as a novel fold by the cystallographers. They are probably right in that such a fold is not yet in the data bank. Yet, the prediction at an overall accuracy of 78% indicates that the novel fold is based on the same local preferences as those already present in the data bank. If the universe of folds is lim-

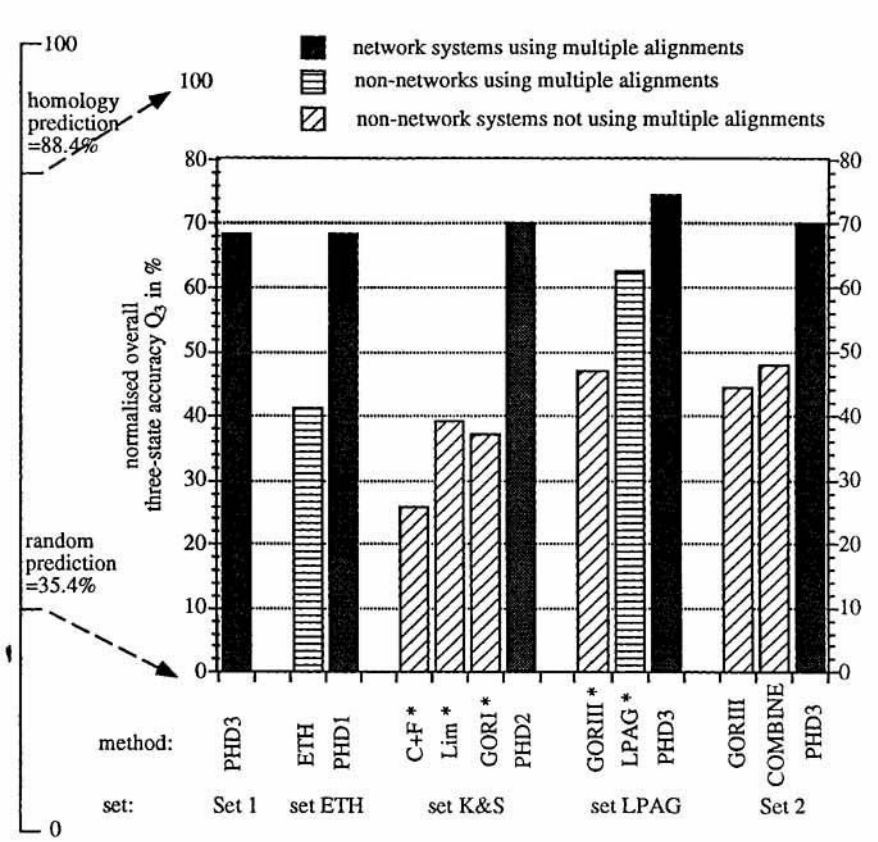

Fig. 8. Normalized overall three-state accuracy for various predictions. A lower limit for secondary structure prediction accuracy is given by a random prediction (=35.4%), an upper limit by the performance of homology modelling (=88.4% on Set 1).[99] This figure shows the same results as Figure 4, but the values are normalized such that a random prediction scores at 0% and homology modeling at 100%.

ited,[23,122–124] the prediction has a good chance to be accurate in the near future, at least, as long as the experimental techniques for the determination of structure remain restricted to the same features of proteins as today.

Of course, the real goal of predictions is to reduce the sequence–structure gap, i.e., to predict 3D structure rather than a one-dimensional abstraction of it in the form of secondary structure strings. For the time being, methods that predict more dimensions than one do not work generally reliably. In some cases, e.g., threading techniques are successful, in others, they fail; it is currently difficult in general to distinguish true positives from the background. The growth of the data bank coupled with technical improvements made it possible to substantially improve secondary structure predictions from 56% overall accuracy a decade ago[42] to now above 72%. On the way to the prediction of 3D structure, this is only a small step, but a promising one.

ACKNOWLEDGMENTS

We thank three colleagues at EMBL: Gerrit Vriend and Reinhard Schneider for assistance and helpful ideas, and Uwe Hobohm for providing the latest lists of unique proteins. We also thank Jean Garnier (INRA, Paris) for having kindly provided the software of GORIII and COMBINE. We also express our gratitude to all those who made coordinates of experimentally determined protein 3D structures available.

REFERENCES

1. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucl. Acids Res. 20:2019–2022, 1992.
2. 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. 112:535–542, 1977.
3. Abola, E.E., Bernstein, F.C., Koetzle, T.F. "The Protein Data Bank." Oxford: Oxford University Press, 1988: 69–81.
4. Hobohm, U., Sander, C. "Selection of Representative Protein Data Sets." Heidelberg, FRG: EMBL, 1993.
5. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Prot. Sci. 1:409–17, 1992.
6. Epstein, C.J., Goldberger, R.F., Anfinsen, C.B. The genetic control of tertiary protein structure: studies with model systems. Cold Spring Harbour Symp. Quant. Biol. 28:439–449, 1963.
7. Anfinsen, C.B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
8. Chothia, C., Lesk, A.M. The relation between the diver-

gence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
9. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
10. Overington, J., Johnson, M.S., Sali, A., Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. Soc. London, B 241:132–145, 1990.
11. Summers, N.L., Karplus, M. Modeling of globular proteins. J. Mol. Biol. 216:991–1016, 1990.
12. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. Proteins 11:52–58, 1991.
13. Schneider, R., Sander, C. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
14. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226:507–533, 1992.
15. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.
16. Taylor, W. New paths from dead ends. Nature (London) 356:478–480, 1992.
17. Schneider, R., Sander, C. The HSSP data base of protein structure-sequence alignment. Nucl. Acids Res. 21:3105–3109, 1993.
18. Karplus, M., Petsko, G.A. Molecular dynamics simulations in biology. Nature (London) 347:631–639, 1990.
19. Gunsteren, W.F.v. Molecular dynamics studies of proteins. Curr. Opin. Str. Biol. 3:167–174, 1993.
20. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature (London) 319:199–203, 1986.
21. Baumann, G., Frömel, C., Sander, C. Polarity as a criterion in protein design. Prot. Engin. 2:329–334, 1989.
22. Crippen, G., M. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 30:4232–4237, 1991.
23. Finkelstein, A.V., Reva, B.A. A search for the most stable folds of protein chains. Nature (London) 351:497–499, 1991.
24. Lüthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
25. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature (London) 356:1992.
26. Overington, J., Donnelly, D., Johnson, M.S., Sali, A., Blundell, T.L. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1:216–226, 1992.
27. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. J. Mol. Biol. 232:805–825, 1993.
28. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. J. Mol. Biol. 213:859–883, 1990.
29. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
30. Sippl, M.J., Jaritz, M. Predictive power of mean force pair potentials in protein folding. Proteins, submitted.
31. Pauling, L., Corey, R.B. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proc. Natl. Acad. Sci. U.S.A. 37:729–740, 1951.
32. Pauling, L., Corey, R.B., Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 37: 205–234, 1951.
33. Chou, P.Y., Fasman, U.D. Prediction of protein conformation. Biochemistry 13:211–215, 1974.
34. Chou, P.Y., Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. 47:45–148, 1978.
35. Lewis, P.N., Momany, F.A., Scheraga, H.A. Folding of polypeptide chains in proteins: A proposed mechanism for folding. Proc. Natl Acad. Sci. U.S.A. 68:2293–2297, 1971.
36. Pain, R.H., Robson, B. Analysis of the code relating sequence to secondary structure in proteins. Nature (London) 227:62–63, 1970.
37. Robson, B., Pain, R.H. Analysis of the code relating sequence to conformation in proteins: Possible implications for the mechanism of formation of helical regions. J. Mol. Biol. 58:237–259, 1971.
38. Robson, B. Conformational properties of amino acid residues in globular proteins. J. Mol. Biol. 107:327–356, 1976.
39. Garnier, J., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
40. Lim, V.I. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J. Mol. Biol. 88:857–872, 1974.
41. Finkelstein, A.V., Ptitsyn, O.B. Statistical analysis of the correlation among amino acid residues in helical, β-structural and non-regular regions of globular proteins. J. Mol. Biol. 62:613–624, 1971.
42. Kabsch, W., Sander, C. How good are predictions of protein secondary structure? FEBS Lett. 155:179–182, 1983.
43. Zhang, X., Mesirov, J.P., Waltz, D.L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
44. Salzberg, S., Cost, S. Predicting protein secondary structure with a nearest-neighbor algorithm. J. Mol. Biol. 227: 371–374, 1992.
45. Ptitsyn, O.B., Finkelstein, A.V. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 22:15–25, 1983.
46. Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J. Secondary structure prediction: Combination of three different methods. Prot. Engin. 2:185–91, 1988.
47. Gibrat, J.-F., Garnier, J., Robson, B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol. 198:425–443, 1987.
48. Levin, J.M., Robson, B., Garnier, J. An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett. 205:303–308, 1986.
49. Levin, J.M., Garnier, J. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim. Biophys. Acta 955:283–295, 1988.
50. Gascuel, O., Golmard, J.L. A simple method for predicting the secondary structure of globular proteins: Implications and accuracy. CABIOS 4:357–365, 1988.
51. Viswanadhan, V.N., Denckla, B., Weinstein, J.N. New joint prediction algorithm (Q7-JASEP) improves the prediction of protein secondary structure. Biochemistry 30: 11164–11172, 1991.
52. Holley, H.L., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86:152–156, 1989.
53. Qian, N., Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884, 1988.
54. Hirst, J.D., Sternberg, M.J.E. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry 31: 615–623, 1992.
55. Garnier, J., Levin, J.M. The protein structure code: What is its present status? CABIOS 7:133–142, 1991.
56. Heijne, G.v. Computer analysis of DNA and protein sequences. Eur. J. Biochem 199:253–256, 1991.
57. Dickerson, R.E., Timkovich, R., Almassy, R.J. The cytochrome fold and the evolution of bacterial energy metabolism. J. Mol. Biol. 100:473–491, 1976.
58. Pastore, A., Lesk, A.M. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. Proteins 8:133–155, 1990.
59. Maxfield, F.R., Scheraga, H.A. Improvements in the pre-

diction of protein topography by reduction of statistical errors. Biochemistry 18:697–704, 1979.
60. 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. 195:957–961, 1987.
61. Nishikawa, K. Assessment of secondary structure prediction of proteins: Comparison of computerized Chou-Fasman method with others. Biochim. Biophys. Acta 748: 285–299, 1983.
62. Benner, S.A., Cohen, M.A., Gerloff, D. Predicted secondary structure for the Src homology 3 domain. J. Mol. Biol. 229:295–305, 1993.
63. Benner, S.A. Predicting de novo the folded structure of proteins. Curr. Opin. Str. Biol. 2:402–412, 1992.
64. Benner, S.A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure of the catalytic domain of protein kinases. Adv. Enz. Reg. 31:121–181, 1990.
65. Barton, G.J., Newman, R.H., Freemont, P.S., Crumpton, M.J. Amino acid sequence analysis of the annexin supergene family of proteins. Eur. J. Biochem. 198:749–760, 1991.
66. Crawford, I.P., Niermann, T., Kirchner, K. Prediction of secondary structure by evolutionary comparison: Application to the α subunit of tryptophan synthase. Proteins 2:118–129, 1987.
67. Frampton, J., Leutz, A., Gibson, T.J., Graf, T. DNA-binding domain ancestry. Nature (London) 342:134, 1989.
68. Gibson, T.J., Thompson, J.D., Abagyan, R.A. Proposed structure for the DNA-binding domain of the helix-loop-helix family of eukaryotic gene regulatory proteins. Prot. Engin. 6:41–50, 1993.
69. Gerloff, D.L., Jenny, T.F., Knecht, L.J., Gonnet, G.H., Benner, S.A. The nitrogenase MoFe protein. FEBS Lett. 318:118–124, 1993.
70. Musacchio, A., Gibson, T., Lehto, V.-P., Saraste, M. SH3—an abundant protein domain in search of a function. FEBS Lett. 307:55–61, 1992.
71. Russell, R.B., Breed, J., Barton, G.J. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. FEBS Lett. 304:15–20, 1992.
72. Rost, B., Sander, C. Jury returns on structure prediction. Nature (London) 360:540, 1992.
73. Niermann, T., Kirschner, K. Improving the prediction of secondary structure of 'TIM-barrel' enzymes (Corrigendum). Prot. Engin. 4:359–370, 1991.
74. Rost, B., Sander, C. Exercising multi-layered networks on protein secondary structure. Elba, Italy: Int. J. Neural Syst. 209–220, 1992.
75. 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. 90:7558–7562, 1993.
76. Levin, J.M., Pascarella, S., Argos, P., Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. Prot. Eng, 6:849–854, 1993.
77. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
78. Rost, B., Sander, C., Schneider, R. Progress in protein structure prediction? TIBS 18:120–123, 1993.
79. Kabsch, W., Sander, C. Dictionary of Protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
80. Rumelhart, D.E., Hinton, G.E., Williams, R.J. Learning representations by back-propagating error. Nature (London) 323:533–536, 1986.
81. Lesk, A.M. "Protein Architecture—A Practical Approach." Oxford: Oxford University Press, 1991: 287.
82. Pascarella, S., Argos, P. Analysis of insertions/deletions in protein structures. J. Mol. Biol. 224:461–471, 1992.
83. Rackovsky, S. On the nature of the protein folding code. Proc. Natl. Acad. Sci. U.S.A. 90:644–648, 1993.
84. Rao, S., Zhu, Q.-L., Vajda, S., Smith, T. The local information content of the protein structural database. FEBS Lett. 322:143–146, 1993.
85. Muskal, S.M., Kim, S.-H. Predicting protein secondary structure content. A tandem neural network approach. J. Mol. Biol. 225:713–727, 1992.
86. Dubchak, I., Holbrook, S.R., Kim, S.-H. Prediction of protein folding class from amino acid composition. Proteins 16:79–91, 1993.
87. Zhang, C.-T., Chou, K.-C. An optimization approach to predicting protein structural class from amino acid composition. Prot. Sci. 1:401–408, 1992.
88. Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.S. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Prot. Sci. 2:1171–1182, 1993.
89. Johnson, C.W.J. Protein secondary structure and circular dichroism: A practical guide. Proteins 7:205–214, 1990.
90. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature (London) 261:552–558, 1976.
91. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and other characters: I. Angular distribution. J. Biochem. 94: 981–995, 1983.
92. Nishikawa, K., Ooi, T. Correlation of the amino acid composition of a protein to its structural and biological characteristics. J. Biochem. 91:1821–1824, 1982.
93. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214:171–182, 1990.
94. Muggleton, S., King, R.D., Sternberg, M.J.E. Protein secondary structure prediction using logic-based machine learning. Prot. Engin. 5:647–657, 1992.
95. Sasagawa, F., Tajima, K. Prediction of protein secondary structures by a neural network. CABIOS 9:147–152, 1993.
96. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Secondary structure assignment for α/β proteins by a combinatorial approach. Biochemistry 22:4894–4904, 1983.
97. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
98. Rost, B., Sander, C. Secondary structure prediction of all-helical proteins in two states. Prot. Engin. 6:831–836, 1993.
99. Rost, B., Schneider, R., Sander, C. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235:13–26, 1994.
100. Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nørskov, L., Olsen, O.H., Petersen, S.B. Protein secondary structure and homology by neural networks. FEBS Lett. 241:223–228, 1988.
101. Bossa, F., Pascarella, S. PRONET: A microcomputer program for predicting the secondary structure of proteins with a neural network. CABIOS 5:319–320, 1990.
102. Fariselli, P., Compiani, M., Casadio, R. Predicting secondary structures of membrane proteins with neural networks. Eur. Biophys. J. 22:41–51, 1993.
103. Fogelman-Soulié, F., Mejía, C. "Incorporating Knowledge in Multi-Layer Networks: The Example of Proteins Secondary Structure Prediction." Berlin: Springer, 1990: 185–194.
104. Hayward, S., Collins, J.F. Limits on α-helix prediction with neural network models. Proteins 14:372–381, 1992.
105. Tchoumatchenko, I., Vissotsky, F., Ganascia, J.-G. How to make explicit a neural network trained to predict proteins secondary structure. ACASA, LAFORIA-CNRS, Université Paris VI, 4 Place Jussieu, 75 252 Paris, CEDEX 05, France, 1993.
106. Kohda, D., Hatanaka, H., Odaka, M., Mandiyan, V., Ullrich, A., Schlessinger, J., Inagaki, F. Solution structure of the SH3 domain of phospholipase C-γ. Cell 72: 953–960, 1993.
107. Koyama, S., Yu, H., Dalgarno, D.C., Shin, T.B., Zydowsky, L.D., Schreiber, S.L. Structure of the PI3K SH3 domain and analysis of the SH3 family. Cell 72:945–952, 1993.
108. Sönnichsen, F.D., Sykes, B.D., Chao, H., Davies, P.L. The nonhelical structure of antifreeze protein type III. Science 259:1154–1157, 1993.
109. Robson, B. Analysis of the code relating sequence to conformation in globular proteins—Theory and application of expected information. Biochem. J. 141:853–867, 1974.

110. Robson, B., Garnier, J. Nature (London) 361:506, 1993.
111. Klein, P., DeLisi, C. Prediction of protein structural class from the amino acid sequence. Biopolymers 25:1659–1672, 1986.
112. Klein, P., Jacquez, J.A., DeLisi, C. Prediction of protein function by discriminant analysis. Math. Biosci. 81:177–189, 1986.
113. Klein, P. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874:205–215, 1986.
114. Deleage, G., Roux, B. An algorithm for protein secondary structure prediction based on class prediction. Prot. Engin. 1:289–294, 1987.
115. Deleage, G., Roux, B. "Use of Class Prediction to Improve Protein Secondary Structure Prediction." New York: Plenum Press, 1989: 587–597.
116. Sheridan, R.P., Dixon, J.S., Venkataghavan, R., Kuntz, I.D., Scott, K.P. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymers 24:1995–2023, 1985.
117. Rooman, M.J., Kocher, J.P., Wodak, S.J. Prediction of protein backbone conformation based on seven structure assignments: Influence of local interactions. J. Mol. Biol. 221:961–979, 1991.
118. Rooman, M.J., Kocher, J.-P., Wodak, S.J. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 31:10226–10238, 1992.
119. Rooman, M.J., Wodak, S.J. Extracting information on folding from the amino acid sequence: Consensus regions with preferred conformation in homologous proteins. Biochemistry 31:10239–10249, 1992.
120. Rost, B., Schneider, R., Sander, C. PHD—an automatic server for protein secondary structure prediction. CABIOS, in press.
121. Moore, S.A., James, M.N.G., O'Kane, D.J., Lee, J. Crystal structure of a flavoprotein related to the subunits of bacterial luciferase. EMBO J. 12:1767–1774, 1993.
122. Chothia, C. One thousand protein families for the molecular biologist. Nature (London) 357:543–544, 1992.
123. Finkelstein, A.V., Reva, B.A. Search for the stable state of a short chain in a molecular field. Prot. Engin. 5:617–624, 1992.
124. Finkelstein, A.V., Badretdinov, A.Y., Ptitsyn, O.B. Physical reasons for secondary structure stability: α-Helices in short peptides. Proteins 10:287–299, 1991.
125. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405:442–451, 1975.
126. Benner, S.A., Gerloff, D.L. Predicting the conformation of proteins: Man versus machine. FEBS Lett. 325:29–33, 1993.
127. Benner, S.A., Cohen, M.A., Gerloff, D. Correct structure prediction? Nature (London) 359:781, 1992.
128. Knighton, D.R., Zheng, J., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S., Sowadski, J.M. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science 253:407–414, 1991.
129. Yu, H.T., Rosen, M.R., Shin, T.B., Seidel-Dugan, C., Brugge, J.S., Schreiber, S.L. Solution structure of the SH3 domain of Src and identification of its ligand-binding site. Science 258:1665–1668, 1992.
130. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., Saraste, M. Crystal structure of a Src-homology 3 (SH3) domain. Nature (London) 359:851–855, 1992.
131. Kim, J., Rees, D.C. Crystallographic structure and functional implications of the nitrogenase molybdenum-iron protein from Azotobacter vinelandii. Nature (London) 360:553–560, 1992.
132. Perczel, A., Hollósi, M., Tusnády, G., Fasman, G.D. Convex constraint analysis: A natural deconvolution of circular dichroism curves of proteins. Prot. Engin. 4:669–679, 1991.
133. Perczel, A., Park, K., Fasman, G.D. Deconvolution of the circular dichroism spectra of proteins: The circular dichroism spectra of the antiparallel β-sheet in proteins. Proteins 13:57–69, 1992.
134. Böhm, G., Muhr, R., Jaenicke, R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. Prot. Engin. 5:191–195, 1992.
135. Andrade, M.A., Chacón, P., Merelo, J.J., Morán, F. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. Prot. Engin. 6:383–390, 1993.