Gr78, and Dnak entries from SWISS-PROT, eliminated the divergent Hsp110 subfamily, and kept only one member of each group of sequences that were more than 90% identical, ending with a list of 68 sequences. These we aligned in PileUp[19] using an increased gap penalty of 5. Analysis of the alignment with ALIGNED80 showed that, as the scanning window was decreased from 28 to 14, the initially detected first peak was joined by a second and then by a third, indicating the presence of a helical bundle (Fig. 4). Note that if these peaks were part of a segmented coiled coil, such as in intermediate filaments, the probabilities would be in the 90% range and the peaks would be longer. The location of the three peaks is matched by three helices predicted using the PHD server.[17] In conclusion, the analysis indicates the presence of a three-helix bundle at the C-terminal end of Hsp70 proteins.

Input File Formats

COILS accepts files in GCG (Genetics Computer Group) format and in Pearson (FASTA) format. In addition, users can adapt any sequence to be read by COILS by marking its beginning (by ">" or "[space][space]..") and end (by "*" or "//"). An input file may contain multiple sequences as long as they are delimited by markers.

ALIGNED accepts files created by PileUp and CLUSTAL V from SWISS-PROT entries (this limitation is connected to the space allocated in the alignment for the sequence names). An expanded range of input formats is planned.

Program Availability

All programs, source codes, and documentation can be downloaded from the Coils/vms folder of the anonymous ftp server FTP.BIOCHEM.MPG.DE. The programs are written in VAX Pascal and operate equally under VAX/VMS and OpenVMS. In addition, the coils folder contains C and c++ source codes for COILS that can be compiled under UNIX, as well as a compiled version of the c++ code for PC/DOS. Macstripe, a Macintosh adaptation of COILS by Alex Knight (knight@wi.mit.edu), is available on the World Wide Web at http://www.wi.mit.edu/matsudaira/coilcoil.html.

A World Wide Web (WWW) server for COILS has become available at the Swiss Institute for Experimental Cancer Research (http://ulrec3.unil.ch/software/COILS_form.html) courtesy of Kay Hofmann (khofmann@isrec-sun1.unil.ch).

[19] Genetics Computer Group, Madison, WI.

Acknowledgments

I thank Janice Lupas for programming the algorithms described in this chapter and Jeff Stock for critically reading the manuscript.

[31] PHD: Predicting One-Dimensional Protein Structure by Profile-Based Neural Networks

By BURKHARD ROST

Introduction

We still cannot predict protein three-dimensional (3D) structure from sequence alone, but we can predict 3D structure for one-fourth of the known protein sequences (SWISS-PROT[1]) by homology modeling based on significant sequence identity (>25%) to known 3D structures (Protein Data Bank, PDB[2]).[3] For the remaining, about 30,000 known sequences, the prediction problem has to be simplified. An extreme simplification is to try to predict projections of 3D structure, for example, one-dimensional (1D) secondary structure, solvent accessibility, or transmembrane location assignments for each residue. Despite the extreme simplification, the success of 1D predictions has been limited as segments from single sequences (used as input) do not contain sufficient global information about 3D structures.[4,5] Patterns of amino acid substitutions within sequence families are highly specific for the 3D structure of that family. Using such evolutionary information is the key to a significant improvement of 1D predictions.

In this chapter we describe three prediction methods that use evolutionary information as input to neural network systems to predict secondary

[1] A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **22,** 3578 (1994).
[2] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).
[3] C. Sander and R. Schneider, *Nucleic Acids Res.* **22,** 3597 (1994).
[4] W. Kabsch and C. Sander, *FEBS Lett.* **155,** 179 (1983).
[5] B. Rost, C. Sander, and R. Schneider, *Trends Biochem. Sci.* **18,** 120 (1993).

software/COILS_form.html) courtesy of Kay Hofmann (khofmann@isrec-sun1.unil.ch).

Acknowledgments

I thank Janice Lupas for programming the algorithms described in this chapter and Jeff Stock for critically reading the manuscript.

[31] PHD: Predicting One-Dimensional Protein Structure by Profile-Based Neural Networks

By BURKHARD ROST

Introduction

We still cannot predict protein three-dimensional (3D) structure from sequence alone, but we can predict 3D structure for one-fourth of the known protein sequences (SWISS-PROT[1]) by homology modeling based on significant sequence identity (>25%) to known 3D structures (Protein Data Bank, PDB[2]).[3] For the remaining, about 30,000 known sequences, the prediction problem has to be simplified. An extreme simplification is to try to predict projections of 3D structure, for example, one-dimensional (1D) secondary structure, solvent accessibility, or transmembrane location assignments for each residue. Despite the extreme simplification, the success of 1D predictions has been limited as segments from single sequences (used as input) do not contain sufficient global information about 3D structures.[4,5] Patterns of amino acid substitutions within sequence families are highly specific for the 3D structure of that family. Using such evolutionary information is the key to a significant improvement of 1D predictions.

In this chapter we describe three prediction methods that use evolutionary information as input to neural network systems to predict secondary

[1] A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **22**, 3578 (1994).
[2] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112**, 535 (1977).
[3] C. Sander and R. Schneider, *Nucleic Acids Res.* **22**, 3597 (1994).
[4] W. Kabsch and C. Sander, *FEBS Lett.* **155**, 179 (1983).
[5] B. Rost, C. Sander, and R. Schneider, *Trends Biochem. Sci.* **18**, 120 (1993).

Gr78, and Dnak entries from SWISS-PROT, eliminated the divergent Hsp110 subfamily, and kept only one member of each group of sequences that were more than 90% identical, ending with a list of 68 sequences. These we aligned in PileUp[19] using an increased gap penalty of 5. Analysis of the alignment with ALIGNED80 showed that, as the scanning window was decreased from 28 to 14, the initially detected first peak was joined by a second and then by a third, indicating the presence of a helical bundle (Fig. 4). Note that if these peaks were part of a segmented coiled coil, such as in intermediate filaments, the probabilities would be in the 90% range and the peaks would be longer. The location of the three peaks is matched by three helices predicted using the PHD server.[17] In conclusion, the analysis indicates the presence of a three-helix bundle at the C-terminal end of Hsp70 proteins.

Input File Formats

COILS accepts files in GCG (Genetics Computer Group) format and in Pearson (FASTA) format. In addition, users can adapt any sequence to be read by COILS by marking its beginning (by ">" or "[space][space]..") and end (by "*" or "//"). An input file may contain multiple sequences as long as they are delimited by markers.

ALIGNED accepts files created by PileUp and CLUSTAL V from SWISS-PROT entries (this limitation is connected to the space allocated in the alignment for the sequence names). An expanded range of input formats is planned.

Program Availability

All programs, source codes, and documentation can be downloaded from the Coils/vms folder of the anonymous ftp server FTP.BIOCHEM.MPG.DE. The programs are written in VAX Pascal and operate equally under VAX/VMS and OpenVMS. In addition, the coils folder contains C and c++ source codes for COILS that can be compiled under UNIX, as well as a compiled version of the c++ code for PC/DOS. Macstripe, a Macintosh adaptation of COILS by Alex Knight (knight@wi.mit.edu), is available on the World Wide Web at http://www.wi.mit.edu/matsudaira/coilcoil.html.

A World Wide Web (WWW) server for COILS has become available at the Swiss Institute for Experimental Cancer Research (http://ulrec3.unil.ch/

[19] Genetics Computer Group, Madison, WI.

structure (PHDsec[6–8]), relative solvent accessibility (PHDacc[9]), and transmembrane helices (PHDhtm[10]) are described. Also illustrated are the possibilities and limitations in practical applications of these methods with results from careful cross-validation experiments on large sets of unique protein structures. All predictions are made available by an automatic E-mail prediction service (see section on availability). The baseline conclusion after some 30,000 requests to the service[11] is that 1D predictions have become accurate enough to be used as a starting point for expert-driven modeling of protein structure.[12–14]

Methods

Generating Multiple Sequence Alignment

The first step in a PHD prediction is generating a multiple sequence alignment. The second step involves feeding the alignment into a neural network system. Correctness of the multiple sequence alignment is as crucial for prediction accuracy as is the fact that the alignment contains a broad spectrum of homologous sequences. By default, PHD uses the program MaxHom (Fig. 1) that generates a pairwise profile-based multiple alignment.[15] A key feature of MaxHom is the compilation of a length-dependent cutoff for significant pairwise sequence identity (Fig. 1).[15]

Multiple Levels of Computations

The PHD methods process the input information on multiple levels (Fig. 2). The first level is a feed-forward neural network with three layers of units (input, hidden, and output). Input to this first level sequence-to-structure network consists of two contributions: one from the local sequence, that is, taken from a window of 13 adjacent residues, and another from the global sequence (Fig. 2). Output of the first level network is the 1D structural state of the residue at the center of the input window. For

[6] B. Rost and C. Sander, *J. Mol. Biol.* **232,** 584 (1993).
[7] B. Rost and C. Sander, *Proc. Natl. Acad. Sci. U.S.A.* **90,** 7558 (1993).
[8] B. Rost and C. Sander, *Proteins* **19,** 55 (1994).
[9] B. Rost and C. Sander, *Proteins* **20,** 216 (1994).
[10] B. Rost, R. Casadio, P. Fariselli, and C. Sander, *Protein Sci.* **4,** 521 (1995).
[11] B. Rost, C. Sander, and R. Schneider, *CABIOS* **10,** 53 (1994).
[12] T. J. P. Hubbard and J. Park, *Proteins* **23,** 398 (1995).
[13] B. Rost, "TOPITS: Threading One-Dimensional Predictions into Three-Dimensional Structures." AAAI Press, Cambridge, July 16–19, 1995.
[14] B. Rost and C. Sander, *Proteins* **23,** 295 (1995).
[15] C. Sander and R. Schneider, *Proteins* **9,** 56 (1991).

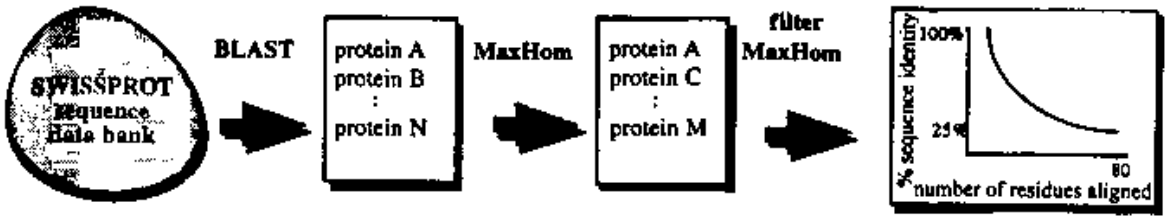

FIG. 1. First, for each protein, the SWISS-PROT database is searched for sequence homologs with a fast alignment method [BLAST, S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215,** 403 (1990)]. Second, the list of putative homologs found is reexamined with a more sensitive profile-based multiple alignment method [MaxHom, C. Sander and R. Schneider, *Proteins* **9,** 56 (1991)]. Third, a length-dependent cutoff for significant pairwise sequence identity is applied [25% + 5%, where +5% reflects a safety margin in the twilight zone (R. F. Doolittle, "Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences." University Science Books, Mill Valley, California, 1986)].

PHDsec and PHDhtm the second level is a structure-to-structure network (see below). The next level consists of an arithmetic average over independently trained networks (jury decision). The final level is a simple filter.

Number of Output Units Determined by Task

Secondary structure is coded by three units: helix, H (*H*, *G*, and *I* in DSSP, the database containing the secondary structure and solvent accessibility for proteins of known 3D structure[16]); strand, E (*E* and *B* in DSSP[16]); and none of the above, denoted loop, L. Transmembrane locations are coded by two units, one for residues being in a transmembrane helix, the other for non-membrane-bound residues (assignments from SWISS-PROT[1]). For solvent accessibility the output coding is not so straightforward. First, the value for accessibility is normalized to a relative accessibility (observed accessibility taken from DSSP[16] divided by maximal accessibility of a given residue type[9,17]) to enable a comparison between residues of different sizes. Second, the relative accessibility is projected onto ten states (for technical reasons; Fig. 2).[9]

Better Segment Prediction by Structure-to-Structure Networks

The output coding for the second level network is identical to the one for the first (Fig. 2). The dominant input contribution to the second level structure-to-structure network is the output of the first level sequence-to-structure network. The reason for introducing a second level is the follow-

[16] W. Kabsch and C. Sander, *Biopolymers* **22,** 2577 (1983).
[17] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, *Science* **229,** 834 (1985).

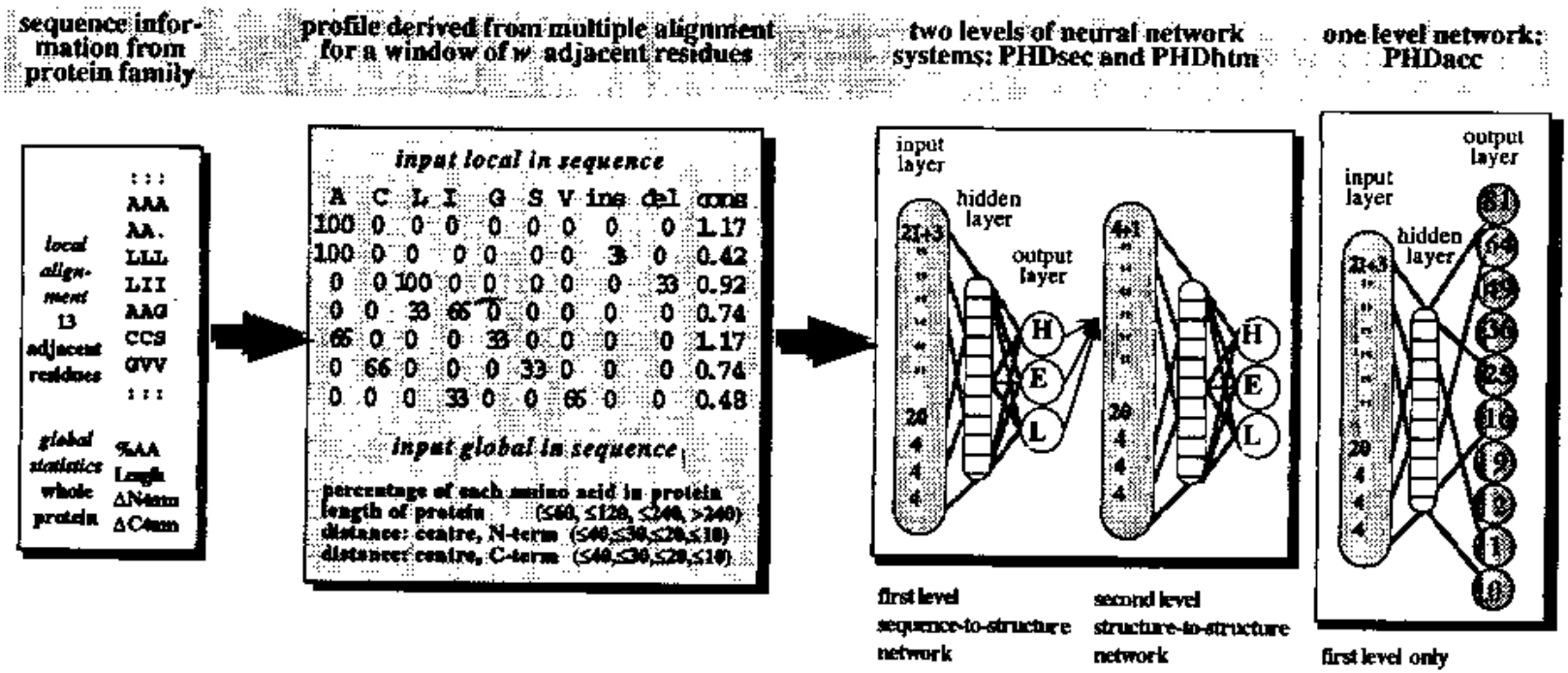

FIG. 2. First, a window of $w = 13$ adjacent residues is chosen from the alignment (here we show only $w = 7$ for clarity). Second, for each residue the profile and global information are complied from the protein. Third, the local and global information is fed into neural network systems. PHDsec and PHDhtm consist of two network levels. In the first level, involving sequence-to-structure networks, for each residue position 24 units are used, 20 for the amino acid types, 1 for a spacer allowing the window to extend over the protein ends (so that the first and last residues in a protein can be at the center of one input window), 2 for the numbers of insertions (ins) and deletions (del) in the alignment at that position, and 1 for the conservation weight (cons); the global information is coded by 20 units for the amino acid composition, 4 for the protein length, and 8 for the distances of the window with respect to the protein ends. The output units code for the 1D structural state of the central residue. For PHDsec, three output units code for helix, strand, and rest; for PHDacc, ten units code for ten levels of relative solvent accessibility (e.g., if the fourth unit has the maximal value, then the prediction is a relative solvent accessibility ≥9% and <16%); and for PHDhtm, two units code for transmembrane or not transmembrane helix. In the second level, involving structure-to-structure networks, the output of the first level is fed into a second level of structure-to-structure network, which additionally uses global information and the conservation weight as input; for example, for PHDsec, first level output = 3 units → local input to second level = 3 + 1 (spacer) + 1 (cons). The output of the second level is the same as that of the first level.

ing. Networks are trained by changing the connections between the units such that the error is reduced for each of the examples successively presented to the network during training. The examples are chosen at random. Therefore, the examples taken at time step t and at time step $t + 1$ are usually not adjacent in sequence. This implies that the network cannot learn that, for example, helices contain at least three residues. The second level structure-to-structure network introduces a correlation between adjacent residues with the effect that predicted secondary structure segments or transmembrane helices have length distributions similar to the ones observed.[6,7]

Balanced Predictions by Balanced Training

For the prediction of secondary structure and transmembrane helices, the distribution of the examples is rather uneven: about 32% of the residues are observed in helix, 21% in strand, and 47% in loop; about 18% of the residues in integral transmembrane proteins are located in transmembrane helices. Choosing the training examples proportional to the occurrence in the data set (unbalanced training) results in a prediction accuracy that mirrors this distribution; for example, strands are predicted inferior to helix or loop.[18-20] A simple way around the database bias is a balanced training: at each time step one example is chosen from each class, that is, one window with the central residue in a helix, one with the central residue in a strand, and one representing the loop class. This training results is a prediction accuracy well balanced between the output states.[6,7]

Compromise between Overprediction and Underprediction by Jury Decision

Balanced training results in improved predictions for the less populated output states (e.g., strand). However, this is associated with less accurate predictions for more populated states (loop). Consequently, the overall accuracy is lower for the balanced than for the unbalanced prediction. To find a compromise between networks with balanced and those with unbalanced training, a final jury decision is performed (effectively a compromise between over- and underprediction). The jury decision is a simple arithmetic average over, typically, four differently trained networks: all

[18] O. Gascuel and J. L. Golmard, *CABIOS* **4**, 357 (1988).
[19] B. Rost and C. Sander, *in* "1D Secondary Structure Prediction through Evolutionary Profiles" (H. Bohr and S. Brunak, eds.), p. 257. IOS Press, Amsterdam, Oxford, and Washington, D.C., 1994.
[20] A. A. Salamov and V. V. Solovyev, *J. Mol. Biol.* **247**, 11 (1995).

combinations of first level networks with balanced or unbalanced training, and with balanced or unbalanced training of second level networks (2 × 2). The final prediction is assigned to the unit with maximal output value (winner takes all).

Correcting Obvious Errors by Final Filter

For secondary structure prediction (PHDsec), the filter affects only drastic, unrealistic predictions (e.g., HEH → HHH; EHE → EEE; and LHL → LLL). For accessibility prediction (PHDacc), the filter performs an average over neighboring output units (i.e., not over adjacent residues). Only the filter used for predicting transmembrane helices (PHDhtm) is crucial for the performance. The currently implemented filter has been guided by previous experience.[21-24] Predicted transmembrane helices which are too long are either split or shortened. Predicted transmembrane helices which are too short are either elongated or deleted. All these decisions (split or shorten; elongate or delete) are based on the strength of the prediction and on the length of the transmembrane helix predicted.[10]

Avoiding Overestimating Prediction Accuracy

The three necessary conditions for an appropriate evaluation of prediction accuracy are first, that training and testing set are distinct; second, that the testing set is representative; and, third, that free parameters are not optimized on the test set which is used for the final evaluation. In more detail, first, the criterion for distinct sets is that no protein in one set has more than 25% pairwise sequence identity to any protein in the other.[15] Second, the test set has to be representative for the database (ideally for all existing proteins), that is, all known sequence families should be included, and they should be included only once. Third, no free parameter should be optimized with respect to the test set. A simple protocol for correct testing would be the following. (1) Choose a small test set (pretest, some 10 proteins) and adjust free parameters; (2) keeping the network fixed, compile the accuracy for all test proteins (real test, >100 proteins by cross-validation experiments; note that the number of splits between test and training sets for cross-validation is of no interest for the user); (3) apply the same network to another test set never used before (prerelease test, e.g., protein structures experimentally determined after the project

[21] G. von Heijne, *Nucleic Acids Res.* **14,** 4683 (1986).
[22] G. von Heijne and Y. Gavel, *Eur. J. Biochem.* **174,** 671 (1988).
[23] G. von Heijne, *J. Mol. Biol.* **225,** 487 (1992).
[24] L. Sipos and G. von Heijne, *Eur. J. Biochem.* **213,** 1333 (1993).

had started). A lower level of accuracy for the pre-release test than for the real test indicates an overfitting of free parameters. Step three should be reapplied whenever a considerable number of new structures have been added to the database (Table I).

Results

Values for Expected Prediction Accuracy Are Distributions

Statements such as secondary structure is about 90% conserved within sequence families,[25] or solvent accessibility is about 85% conserved within sequence families,[9] refer to averages of distributions. The same holds for the expected prediction accuracy (Fig. 3). Such distributions explain why some developers have overestimated the performance of their tools using data sets of only tens of proteins (or even fewer).[5] For the user interested in a certain protein, the distributions imply a rather unfortunate message: for that protein, the accuracy could be lower than 40%, or it could be higher than 90% (Fig. 3). For some of the worst predicted proteins, the low level of accuracy could be anticipated from their unusual features, for example, for crambin or the antifreeze glycoprotein type III. However, for others the reasons for the failure of PHDsec are not obvious; for example, both the phosphatidylinositol 3-kinase[26] and the Src homology domain of cytoskeletal spectrin have homologous structure,[27] but prediction accuracy varies between less than 40% (kinase) and more than 70% (spectrin). Another possible reason for a bad prediction is a bad alignment. In general, single sequences yield accuracy values about ten percentage points lower than multiple alignments.[6] Indeed, the worst case for a prediction so far is pheromone (1erp), a short protein structurally dominated by a disulfide bridge, for which there is no sequence alignment available: only 32% of the residues are predicted correctly.

Reliability of Prediction Correlating with Accuracy

An estimate where in the distributions (Fig. 3) a given prediction is to be expected is given by the prediction strength, that is, the difference between the output unit with highest value (winner unit) and the output unit with the next highest value. This difference is used to define a reliability

[25] B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235,** 13 (1994).
[26] S. Koyama, H. Yu, D. C. Dalgarno, T. B. Shin, L. D. Zydowsky, and S. L. Schreiber, *Cell (Cambridge, Mass.)* **72,** 945 (1993).
[27] A. Musacchio, M. Noble, R. Pauptit, R. Wierenga, and M. Saraste, *Nature (London)* **359,** 851 (1992).

TABLE I

ACCURACY OF SECONDARY STRUCTURE AND ACCESSIBILITY PREDICTION

Method[a]	Set[b]	N[c]	Secondary structure			Solvent accessibility			Date[j]
			Q_3[d]	I[e]	Sov_3[f]	Q_1[g]	Q_2[h]	Corr[i]	
HM: SeqAli	1	80	88.4	0.62	89.7	71.6	83.8	0.68	
HM: StrAli	1	80				73.6	84.8	0.77	
RAN	1	80	35.2	0.01	30.6	33.9	52.0	0.01	
PHD	2	126	71.6	0.27	72.8	57.9	75.0	0.54	06 92
PHD	3	124	72.5	0.28	75.6				07 93
SIMPA	3	124	60.7	0.12	61.7				
PHD	3a	112				57.9	74.7	0.54	03 94
PHD	7	60	74.8	0.34	76.8				
LPAG	7	60	68.5[k]	—	—				
PHD	8	13				60.8	79.2	0.61	
Wako & Blundell	8	13				—	76.5[k]	—	
PHD	4	27	72.0	0.28	72.4	57.6	73.4	0.55	05 94
PHD	5	59	73.0	0.30	75.7	57.0	74.0	0.54	11 94
PHD	6	9	72.1	0.27	72.8		63.0	0.38	12 94
PHD	2–6	**337**	**72.3**	**0.28**	**73.8**				03 95
		318				**57.3**	**74.2**	**0.54**	

[a] HM: SeqAli, homology modeling based on sequence alignments within sequence families [B. C. Sander, and R. Schneider, *J. Mol. Biol.* **235,** 13 (1994); B. Rost and C. Sander, *Proteins* **20,** 216 (1994)]; HM: StrAli, homology modeling based on structural alignments [B. Rost and C. Sander, *Proteins* **20,** 216 (1994)]; *RAN*, random alignments, that is, worst prediction; PHD, neural network predictions; SIMPA, statistical prediction method [J. M. Levin, *et al.*, *FEBS Lett.* **205,** 303 (1986); note that SIMPA is not reported as the best method but scored better than others (GORIII, COMBINE) on set 7]; Wako & Blundell, statistical prediction method based on alignments [Wako and Blundell, *J. Mol. Biol.* **238,** 682 (1994)]; *LPAG*, statistical prediction method based on alignments [J. M. Levin, S. P. A. Pascarella, P. Argos, and J. Garnier, *Protein Eng.* **6,** 849 (1993)].

[b] Different tests sets are numbered to indicate identical sets. 1, B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235,** 13 (1994); 2, 3, B. Rost and C. Sander, *Proteins* **19,** 55 (1994); 3a, subset of set 3, B. Rost and C. Sander, *Proteins* **20,** 216 (1994); 4 and 5, recently determined structures; 6, proteins from Asilomar prediction context, B. Rost and C. Sander, *Proteins* **23,** 295 (1995); 7, J. M. Levin, S. P. A. Pascarella, P. Argos, and J. Garnier, *Protein Eng.* **6,** 849 (1993); 8, H. Wako and T. L. Blundell, *J. Mol. Biol.* **238,** 682 (1994); 2–6, results for proteins from sets 2–6, 337 unique protein chains with a total of 74,901 residues for PHDsec, and 318 unique proteins with a total of 79,588 residues for PHDacc.

[c] N, Number of proteins used for testing (all results for test proteins with less than 25% sequence identity to proteins used for training).

[d] Q_3, Three-state overall pre-residue accuracy for secondary structure, that is, number of residues predicted correctly in helix, strand, or rest.

[e] I, Information, entropy measure for accuracy [B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235,** 13 (1994); B. Rost and C. Sander, *J. Mol. Biol.* **232,** 584 (1993)].

index for the prediction of each residue [normalized to a scale from 0 (low) to 9 (high)]. Residues with higher reliability index are predicted with higher accuracy (Fig. 4). In practice, the reliability index offers an excellent tool to focus on some key regions predicted at high levels of expected accuracy. (Note however, that the reliability indices tend to be unusually high for poor alignments.)

Prediction of Secondary Structure at Better than 72% Accuracy

PHDsec was the first secondary structure prediction method to surpass a level of 70% overall three-state per-residue accuracy.[6] The last test set with more than 300 unique protein chains and a total of more than 70,000 residues compiled for this chapter yielded a three-state per-residue accuracy better than 72% (Table I). Besides the high level of overall accuracy, predictions are well balanced (high value for I in Table I). Furthermore, PHDsec meets the demands for a reasonable prediction tool in that the accuracy measured in segment-based scores[25] is higher than per-residue scores: about 74% of the segments are correctly predicted (Sov_3 in Table I).

Structural Class Prediction Comparable to Experimental Accuracy

Proteins can be sorted roughly into four structural classes based on secondary structure content: all-α (helix $\geq$ 45%, strand $<$ 5%), all-β (strand $\geq$ 45%, helix $<$ 5%), α/β (helix $\geq$ 30%, strand $\geq$ 20%), and all others.[28,29] An experimental way to measure secondary structure content is circular dichroism spectroscopy.[30,31] A simple alternative is to use the

[28] M. Levitt and C. Chothia, *Nature (London)* **261,** 552 (1976).
[29] C.-T. Zhang and K.-C. Chou, *Protein Sci.* **1,** 401 (1992).
[30] C. W. J. Johnson, *Proteins* **7,** 205 (1990).
[31] A. Perczel, K. Park, and G. D. Fasman, *Proteins* **13,** 57 (1992).

[f] Sov_3, Three-state overall per-segment accuracy, that is, overlap of predicted and observed secondary structure segments [B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235,** 13 (1994)].
[g] Q_3, Three-state overall per-residue accuracy for accessibility, that is, percentage of residues correctly predicted as buried, intermediate, or exposed [B. Rost and C. Sander, *Proteins* **20,** 216 (1994)].
[h] Q_2, Two-state overall accuracy for accessibility, that is, percentage of residues correctly predicted as buried or exposed.
[i] Corr, Correlation between predicted and observed relative solvent accessibility [B. Rost and C. Sander, *Proteins* **20,** 216 (1994)].
[j] Date of collecting the data set.
[k] Result taken from literature.

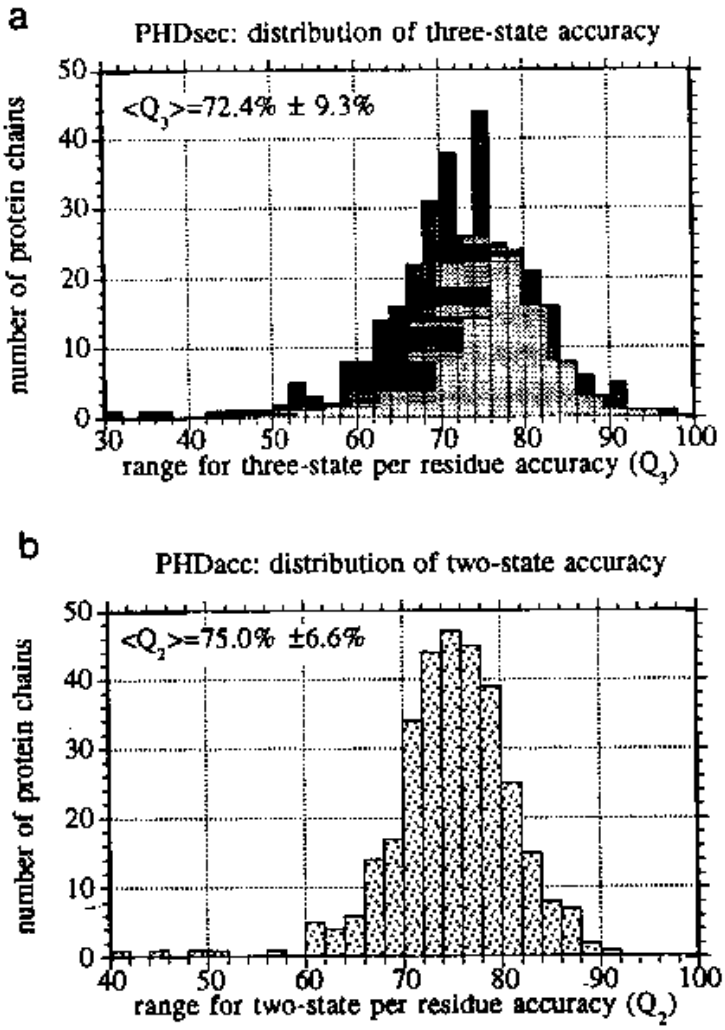

FIG. 3. Expected variation of prediction accuracy with protein chain. (a) Three-state per-residue overall accuracy for PHDsec (total of 337 chains). (b) Two-state per-residue overall accuracy for PHDacc (total of 318 chains). Given are the distributions, averages, and one standard deviation. (c) Cumulative percentage of protein chains predicted at an error level lower than the value given (error = 100 − accuracy). Error values are percentages of falsely predicted residues (PHDsec, PHDacc) and falsely predicted segments (PHDhtm). For example, for one-half of all chains, PHDhtm predicts all segments correctly (note that the total set for evaluating PHDhtm comprises only 69 chains), whereas PHDsec and PHDacc rate at about 25% falsely predicted residues.

predictions of PHDsec to compile the overall prediction of secondary structure content. Based on the predicted content, proteins are sorted into either of the four structural classes. The result is that for about 74% of all protein chains, the class is correctly predicted (Table II). The correlation between observed and predicted content is 0.88 for helix and 0.75 for strand. These values are comparable to results from circular dichroism spectroscopy (he-

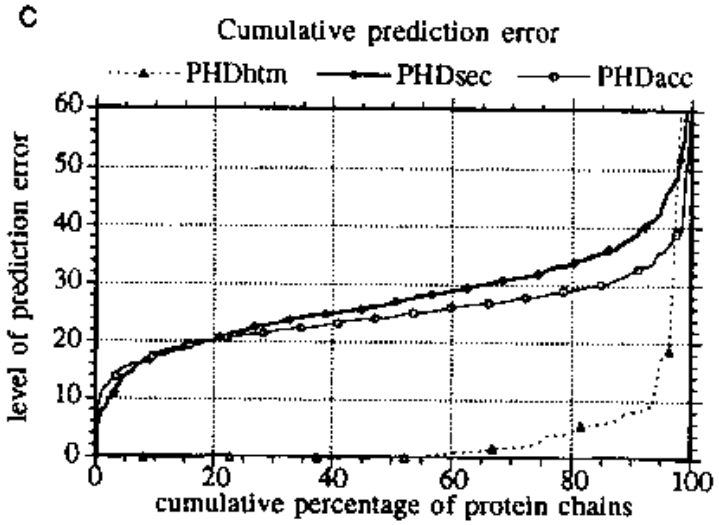

FIG. 3. (*continued*)

lix; 0.84; strand, 0.37–0.41[31]).[6] Of course, this does not imply that PHDsec can replace experiments. However, the high level of accuracy suggests using PHDsec prediction as a complement to experiments.

Prediction of Buried or Exposed Residues at 74% Accuracy

Comparing the conservation of secondary structure with that of solvent accessibility (measured in three states), we find that solvent accessibility is less conserved (Table I). Consequently, PHDacc is less accurate than PHDsec. However, the accessibility prediction is relatively close to the optimum given by homology modeling: the correlation between predicted and observed relative accessibility is 0.54 for PHDacc and would be 0.68 for sequence alignments if homology modeling were possible (Table I). More than 74% of the residues are predicted correctly in either of the two states, buried or exposed. Entirely buried residues (<4% accessible) are predicted best (data not shown).[9] PHDacc is, so far, superior to other methods (Table I). (*Note:* When a subset of 99 monomers was tested, the two-state accuracy rose to over 77%.[9])

Transmembrane Helices Predicted at 95% Accuracy

The problem in evaluating the performance of PHDhtm is the small set of proteins for which the locations of transmembrane helices have been determined reliably. Consequently, the results ought to be viewed with caution. The overall two-state per-residue accuracy of PHDhtm is 95% (Fig. 3), and the per-segment accuracy is about 96% (only 15 of 380 trans-

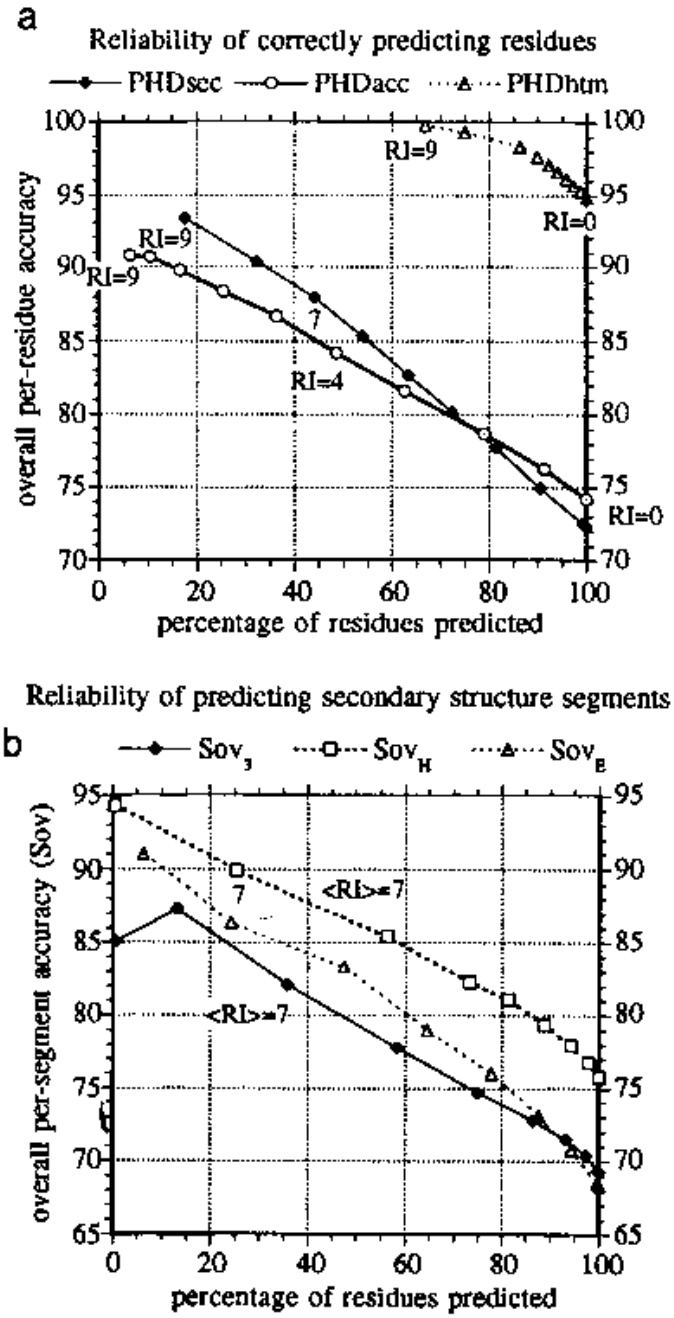

FIG. 4. (a) Expected per-residue accuracy for residues with a reliability index (RI) above a given cutoff; A level of accuracy comparable to homology modeling is reached for 49% of all residues by PHDacc ($RI \geq 4$) and for 44% of all residues by PHDsec ($RI \geq 7$). The small region covered by the reliability index of PHDhtm is dominated by strong predictions for nontransmembrane residues; the most accurately predicted residues in transmembrane helices reach a level of only 96% accuracy. (b) Expected per-segment accuracy for secondary structure segments with an average reliability index ($\langle RI \rangle$) above a given cutoff. For example, an average reliability of $\langle RI \rangle \geq 7$ is reached for (i) 36% of all segments ($Sov_3 > 82\%$), (ii) 56% of all helices ($Sov_E > 85\%$), and (iii) 24 of all strands ($Sov_H > 86\%$). [Definitions of segment overlap are given in B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235**, 13 (1994).].

TABLE II
PREDICTING SECONDARY STRUCTURE CONTENT AND STRUCTURAL CLASS

Method[a]	Set[a]	N_{prot}[a]	ΔHelix[b]	ΔStrand[b]	All-α[c]	All-β[c]	α/β[c]	Rest[c]	Q_{class}[c]
HM:SeqAli	Set 1	80	2.8 ± 3.8	2.7 ± 3.2	94.1	86.7	100.0	89.7	90.0
RAN	Set 1	80	32.1 ± 20.8	21.3 ± 14.5	0.0	0.0	0.0	71.2	44.7
PHD	Set 2	126	8.5 ± 8.0	7.5 ± 8.1	85.7	50.0	50.0	74.1	74.6
PHD	Set 3	124	7.8 ± 6.8	7.3 ± 7.9	94.1	0.0	55.6	74.5	75.8
PHD	Set 2–6	337	8.1 ± 7.9	7.1 ± 7.6	85.0	55.6	45.5	75.6	74.2

[a] See Table I, footnotes *a–c*.

[b] Error in predicting the content of helix or strand averaged over all protein chains in the data set. The error is computed as the difference between the percentage of helix (Δhelix) or strand (Δstrand) between observed and predicted. (Values are given ± one standard deviation.)

[c] Percentage of protein chains correctly predicted in either of the four classes: all-α, all-β, α/β, and all others. Q_{class} gives the percentage of protein chains correctly predicted in any of the four classes.

membrane helices in a set of 69 proteins were wrongly predicted).[10] Of further practical importance is the low level of false positives for PHDhtm. Of a set of 278 globular water-soluble proteins with unique sequences, PHDhtm predicts only 14 incorrect transmembrane helices; these errors occur mostly for proteins with highly hydrophobic β strands in the core.[10]

Availability

PHD predictions (and MaxHom alignments) are available on request by the automatic prediction service PredictProtein.[11] For detailed information send the word help as subject to the Internet address PredictProtein@EMBL-Heidelberg.DE or ventured through the World Wide Web (WWW) site: http://www.embl-heidelberg.de/predictprotein/predictprotein.html (Fig. 5). Because we sometimes have over 100 requests per day, returning a prediction may take a day or more. If no answer is received after two days, something has gone wrong (typical reasons: corrupted E-mail connection of sender or hardware problems at EMBL). In such a case, simply resubmit the request. Should the answer not appear after another two days, send a note to Predict-Help@EMBL-Heidelberg.DE. For further services (e.g., database) provided by the EMBL Protein Design Group, see http://www.sander.embl-heidelberg.de/descr/ or connect by anonymous ftp to ftp.embl-heidelberg.de.

Comments

Accuracy of Predictions. The expected levels of accuracy [PHDsec 72 ± 9% (three states), PHDacc 75 ± 7% (two states), and PHDhtm 94 ±

File Header **(optional statements in italics)**

```
Joe Sequencer, Department of Advanced Protein Research,
National Univeristy, Timbuktu
joe@amino.churn.edu
   predict secondary structure, predict accessibility;
   predict transmembrane
   return MSF; return no alignment; return HSSP profiles;
   return graph
```

File Body - request = single sequence
(anything in line(s) after '#' is interpreted as one-letter amino acid sequence)

```
# incredulase from paracoccus dementiae, translated from cDNA
KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKD
WWKVEVNDRQGFVPAAYVKKLD
```

FIG. 5. Example of a request to the automatic protein structure prediction server PredictProtein. All network methods (PHDsec, PHDacc, PHDhtm) are available. The file to be submitted consists of two parts: a head (optional key words are shown in italics) and the main body starting with a hash (#) in the first line and the one-letter code amino acid sequence in the following lines. Alternative options include submission of a list of sequences or a complete alignment. Details are given in the PredictProtein help file (see text).

6%] are valid for typical globular, water-soluble (PHDsec, PHDacc), or helical transmembrane proteins (PHDhtm) when the multiple alignment contains many and diverse sequences. High values for the reliability indices indicate more accurate predictions. (*Note:* For alignments with little variation in the sequences, the reliability indices adopt misleadingly high values.)

Usefulness of Predictions. The prediction of secondary structure can be accurate enough to assist chain tracing. Furthermore, predictions can be used as a starting point for modeling 3D structure and predicting function.[13,32–34]

Confusion between Strand and Helix. PHDsec focuses on predicting hydrogen bonds. Consequently, occasionally strongly predicted (high reliability index) helices are observed as strands and vice versa.

Strong Signal from Secondary Structure Caps. The ends of helices and strands contain a strong signal. However, on average PHDsec predicts the core of helices and strands more accurately than the caps.[19]

Accessibility Useful to Provide Upper Limits for Contacts. The predicted solvent accessibility (PHDacc) can be translated into a prediction of the number of water atoms around a given residue. Consequently, PHDacc

[32] T. J. P. Hubbard, "Use of β-Strand Interaction Pseudo-Potential in Protein Structure Prediction and Modeling." IEEE Society Press, Los Alamitos, CA, 1994.

[33] B. Rost, *in* "Fitting 1D Predictions into 3D Structures" (H. Bohr and S. Brunak, eds.), in press. CRC Press, Boca Raton, Florida, 1995.

[34] T. Meitinger, A. Meindl, P. Bork, B. Rost, C. Sander, M. Haasemann, and J. Murken, *Nat. Genet.* **5**, 376 (1993).

can be used to derive upper and lower limits for the number of interresidue contacts of a certain residue (such an estimate could improve predictions of interresidue contacts[35]).

Protein Design and Synthesized Peptides. The PHD networks are trained on naturally evolved proteins. However, the predictions have proved to be useful in some cases to investigate the influence of single mutations. For short polypeptides, the following should be taken into account: the network input consists of 17 adjacent residues, and thus shorter sequences may be dominated by the ends (which are treated as solvent).

Prediction of Porins. PHDhtm predicts only transmembrane helices, and PHDsec has been trained on globular, water-soluble proteins. How does one predict the 1D structure for porins then? As porins are partly accessible to solvent, the prediction accuracy of PHDsec was relatively high (70%) for the known structures. Thus, PHDsec appears to be applicable.

Using Prediction of Transmembrane Helices. One possible application of PHDhtm is to scan, for example, entire chromosomes for possible transmembrane proteins. The classification as transmembrane protein is not sufficient to have knowledge about function, but it may shed some light on the puzzle on genome analyses. When using PHDhtm for this purpose, the user should keep in mind that on average about 5% of the globular proteins are falsely predicted to have transmembrane helices.

Acknowledgments

I express my gratitude to the colleagues from the European Molecular Biology Laboratories who help(ed) in developing PHD. First of all, thanks to Chris Sander for intellectual, emotional, and financial support. Second, thanks to Reinhard Schneider for valuable ideas, important discussions, and for help in setting up the prediction server. Third, thanks to Antoine de Daruvar for having rewritten the server software and for now maintaining the server. Fourth, thanks to Gerrit Vriend whose ideas paved the way for the first prediction above 70% accuracy. Fifth, thanks to Séan O'Donoghue for a thorough correction of the manuscript. Finally, thanks to all those who deposit data about protein structure in public databases and thus enable the development of tools such as PHD.

[35] U. Goebel, C. Sander, R. Schneider, and A. Valencia, *Proteins* **18**, 309 (1994).