PROTEINS: Structure, Function, and Genetics 23:295–300 (1995)

Progress of 1D Protein Structure Prediction at Last

Burkhard Rost and Chris Sander
EMBL, 69012 Heidelberg, Germany

ABSTRACT **Accuracy of predicting protein secondary structure and solvent accessibility from sequence information has been improved significantly by using information contained in multiple sequence alignments as input to a neural network system. For the Asilomar meeting, predictions for 13 proteins were generated automatically using the publicly available prediction method PHD. The results confirm the estimate of 72% three-state prediction accuracy. The fairly accurate predictions of secondary structure segments made the tool useful as a starting point for modeling of higher dimensional aspects of protein structure.**

Key words: automatic prediction of protein secondary structure and solvent accessibility, neural networks

SPREADING OPTIMISM BY PUBLISHING HIGH SCORES

Protein secondary structure[1,2] had been predicted from sequence[3] before the first three-dimensional (3D) structures were determined by crystallography.[4,5] Two decades (and dozens of methods[6–8]) later, the accuracy of secondary structure prediction was still not better than 50–55%[9] (percentage of residues predicted correctly in either of the three states helix, strand, rest). In the hunt for higher scores, a growing data base of known structures[10] and refined methods pushed the accuracy to above 60–65%.[11–22] Occasionally, even higher values were reported, but tests on representative data sets revealed that prediction accuracy was about 60% by 1992.[23] Overoptimistic claims by predictors nourished scepticism of potential users. One major point about prediction methods became clear at the Asilomar prediction contest: exaggerated claims are more damaging than genuine errors. Even a prediction method of limited accuracy can be useful if the user knows what to expect. For the editors of scientific journals this implies that no prediction method should be published that has not been sufficiently cross-validated.

HOW TO EVALUATE PREDICTION METHODS

A proper evaluation of prediction methods, in our view, needs to meet four requirements. (1) No significant pairwise sequence identity: the proteins used for setting up a method (training set) and those used for evaluating it should have a *pairwise* sequence identity of less than 25% (length-dependent cut-off[24]), otherwise homology modeling could be applied which would be much more accurate than ab initio predictions.[25,26] (2) Sufficiently large data set: all available unique proteins should be used for testing (currently more than 400[27]), evaluations based on too small numbers are not representative, e.g., prediction accuracy varied about eight percentage points between seven randomly chosen sets, each containing 20 protein chains (and about 4,000 residues).[28] (3) Avoid comparing apples with oranges: no matter which data sets are used for a particular evaluation, results should always be reported additionally on standard sets.[28] (4) No optimization with respect to test set: a seemingly trivial—and often violated—rule is that methods should never be optimized with respect to the data set chosen for final evaluation. For example, most methods are evaluated in n-fold cross-validation experiments (splitting the data set into n different training and test sets). The exact number of n is not important provided the test set is representative and the cross-validation results are NOT misused to again change parameters.

HOW TO IMPROVE PREDICTIONS OF SECONDARY STRUCTURE AND SOLVENT ACCESSIBILITY

One key to more accurate predictions of 1D structure, such as secondary structure or solvent accessibility, has been the use of evolutionary information. This idea has long been in the literature.[29–32] Most often, experts base single-case predictions on multiple alignments.[33–49] The reason for the relevance of evolutionary information is that profiles of residue exchanges in naturally evolved protein families are highly specific for details of a particular protein

Abbreviations: 3D, three-dimensional; 1D, one-dimensional; DSSP, data base containing the secondary structure and solvent accessibility derived from the experimentally determined coordinates of proteins of known 3D structure; PHD, profile-based neural network prediction of secondary structure (PHDsec) and solvent accessibility (PHDacc).

Received April 3, 1995; revision accepted July 3, 1995.
Address reprint requests to Burkhard Rost, EMBL, 69012 Heidelberg, Germany.

TABLE I. Accuracy for Blind Predictions,*,†,‡

		Secondary structure						Accuracy for solvent accessibility	
		Per-residue accuracy			Per-segment accuracy				
Proteins	N_{res}	Q_3(%)	I	BAD (%)	Sov_H (%)	Sov_E (%)	Sov_3 (%)	Q_2 (%)	Cor
chmu	188	94.1	0.54	0.0	100.0	—	99.1	78.7	0.61
kau	769	71.7	0.30	1.4	72.3	75.4	67.4	45.0	0.14
L14	122	68.0	0.22	4.1	44.4	91.5	83.8	74.6	0.57
ppdk	869	70.3	0.24	4.0	80.6	47.7	71.0	70.1	0.44
prosub	72	55.6	0.22	13.6	61.7	57.1	66.2	81.9	0.56
RTP	122	64.8	0.11	10.7	80.8	14.3	71.1	67.2	0.35
synapto	134	68.7	0.22	6.7	0.0	84.0	75.7	79.9	0.61
staufen3	67	46.3	0.10	2.9	52.2	57.7	52.8	73.1	0.37
xyla	345	75.7	0.36	1.7	84.0	65.0	75.8	79.7	0.61
all 9	2688	71.6	0.28	3.4	80.0	66.5	72.8	63.0	0.38
set 94a	55325	72.1	0.27	3.9	75.1	72.3	73.4		
set 94b	67811							74.6	0.54

*Abbreviations for symbols: Q_3, overall three-state per-residue accuracy, i.e., the percentage of residues predicted correctly in either of the three states helix, strand, rest[28]; I, information content of prediction matrix: this entropy-based measure of per-residue accuracy is less intuitive than the three-state percentage but captures well in a single value the balance between correct, false, over, and under prediction, as given by a 3×3 matrix A where A_{ij} gives the number of residues predicted in state i and observed in state j[28,26,59]; *BAD*, percentage of residues predicted in helix and observed in strand, or predicted in strand and observed in helix; *Sov*, overlap between observed and predicted secondary structure segments: this fuzzy score yields 100% even when the two segments (observed and predicted) differ slightly in length, or are slightly displaced (it is found to be the best among several alternatives of defining per-segment accuracy[26]; Sov_H, overlap between observed and predicted helical segments; Sov_E, overlap between observed and predicted strand segments; Sov_3, overlap between all segments (helix, strand, rest); Q_2, two-state overall per-residue accuracy, i.e., percentage of residues predicted correctly in either of the two states buried or exposed (buried $<16\%$ solvent accessible, exposed $\geq 16\%$[50]); *Cor*, correlation coefficient between observed and predicted relative solvent accessibility[50]; N_{res}, is the number of residues.
†Abbreviations for proteins: *chmut*, N-terminal of P-protein of *E. coli*; *kau*, urease α-subunit; *L14*, 50 S ribosomal protein L14; *ppdk*, pyrovate orthophosphate dikinase; *prosub*, propeptide of subtilisin BPN; *RTP*, replication termination protein; *staufen3*, domain 3 of staufen; *synapto*, C2 domain of synaptotagmin; *xyla*, xylanase; *set 94a*, 250 unique proteins used to evaluate PHDsec[23]; *set 94b*, 238 unique proteins used to evaluate PHDacc.[50]
‡Proteins predicted but not evaluated (coordinates not available to us): carboxymuconate lactonizing enzyme (cmle); glycerol-3-phosphate cytidylyltransferase (gpct); isopenicilin N-synthase (ipns); 6-phospho-β-D-galactosidase (pbdg).

structure. One tool to capture the richness of such information is a neural network. For one neural network method the following levels of performance accuracy have been achieved: secondary structure prediction, per-residue accuracy of 72 ± 9%[23]; solvent accessibility prediction: correlation coefficient between observed and predicted accessibility of 0.54 ± 0.12[50] (Table I). Other programmable methods have been shown to benefit from evolutionary information to about the same extent.[32,51,52]

WOMAN WITH MACHINE

Predictions of 1D structure by a system of profile-based neural networks (PHD) are publicly available (for information, send the word *help* to the internet address *PredictProtein@EMBL-Heidelberg.DE*, or use the World Wide Web site *http://www.embl-heidelberg.de/predictprotein/predictprotein.html*). In the Asilomar contest, we submitted blind predictions for 13 proteins (Table I, for four proteins no results are given as the experimental coordinates were not available). The accuracy of PHD predictions depends partly on parameters that can be influenced by the user, such as the sequences taken for a multiple alignment, or the details of the alignment. Furthermore, experts can "filter" predictions according to additional knowledge. Such fine-tuning can improve prediction accuracy in a few cases markedly (as exemplified by comparing the results of PHD based on MaxHom[24] alignments and those refined by Tim Hubbard, see Tom Defay and Fred Cohen, this issue). For the Asilomar contest, we deliberately contributed "unfiltered" automatic predictions to illustrate both the accuracy reachable in laboratories without experts and the starting point for experts to possibly improve predictions. Of course, there is no direct competition between woman and machine; intelligent experts use accurate automated methods and attempt to refine the prediction by their expertise.

WHAT WENT RIGHT?

Prediction accuracy within expected range. First, the secondary structure prediction accuracy using PHD on the Asilomar proteins was within the range of what had been published as expected accuracy (Table I). The correlation between observed and predicted relative solvent accessibility was on the low side of the expected range. The deviation from expected values of accuracy indicates that a set of nine proteins is too small to derive generally valid estimates (sets of 20 or even 60 proteins—evaluation

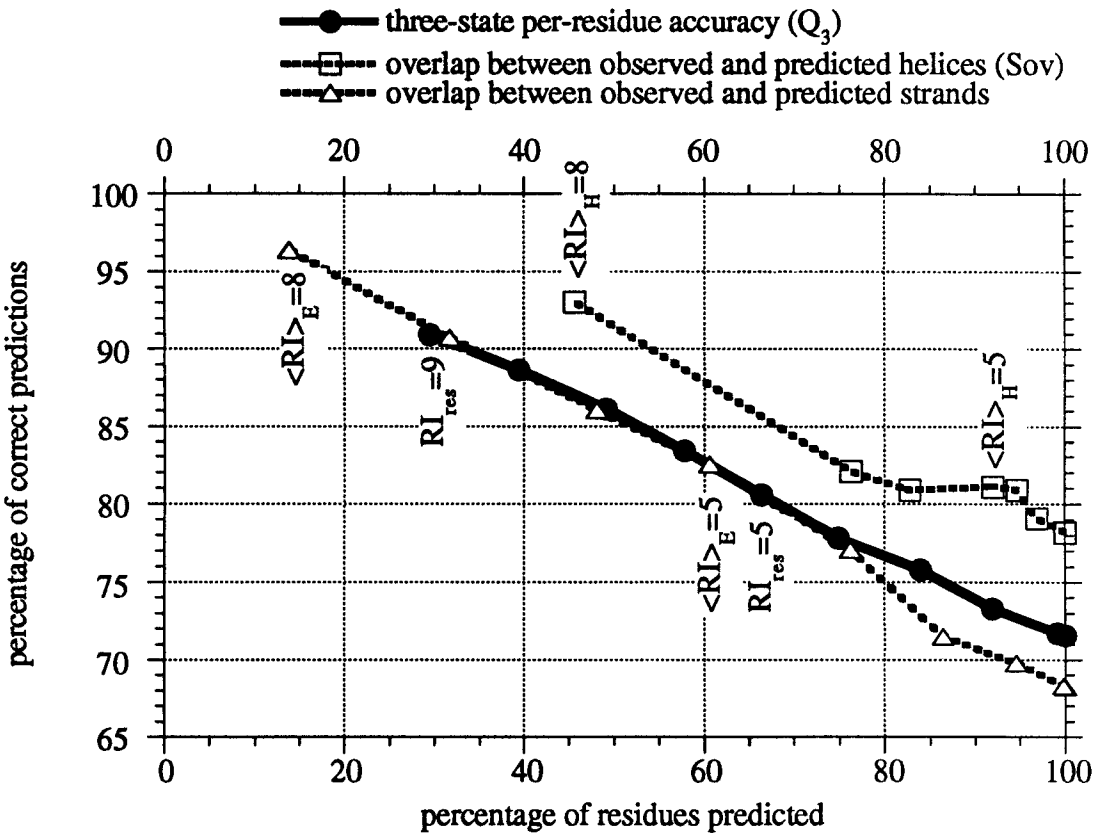

Fig. 1. Estimating reliability of prediction for residue subsets. Expected prediction accuracy can be raised above the 90% level at the expense of not predicting secondary structure for regions with a low reliability index. The reliability index, scaled from 0 (low) to 9 (high), reflects the strength of the prediction. How accurate is the prediction on the subset of residues predicted with indices greater or equal to a certain value (accuracy on vertical axis)? And what fraction of residues is predicted at a given level of minimal reliability (horizontal axis)? The accuracy is given in terms of per-residue (solid line with full circles) and per-segment values (dotted lines, with open triangles for strands and open squares for helices). For points, the values for the respective index are explicitly given in the figure: per-residue accuracy (Q_3), $RI_{res} \geq n$ expected accuracy for the subset of all residues with reliability index above n; per-segment values (Sov), $\langle RI \rangle_{H,E=n}$ expected accuracy for all segments for which the reliability index over all residues in that segment was n. For example, about 65% of all residues are predicted at a reliability of at least five ($RI_{res} = 5$); the three-state per-residue accuracy for these is 81%; 59% of all strands are predicted at an average reliability of at least five ($RI_E = 5$); the segment overlap for these is 86%; more than 90% of all helices are predicted at an average of at least five ($RI_H = 5$); the segment overlap for these is 81%.

set published a decade ago[9]—have been shown to be too small to derive valid estimates for prediction accuracy[23,28]). For example, helices were predicted better than expected; strands on the other hand were predicted worse than expected (Table I).

Reliability of prediction correlates with accuracy. Prediction accuracy varies between different proteins. Fortunately, the reliability of PHD predictions enables one to estimate on which side of such a distribution the prediction for a given protein is to be expected. For example, the 30% of residues predicted with highest reliability are predicted at an average accuracy of over 90% (Fig. 1); and the most reliable one-third of the strands and helices were predicted at segment-based accuracy values of >90% (Fig. 1).

Using 1D predictions to guide modeling of 3D structure. For xylanase the relative solvent accessibility was predicted more accurately than the expected average (Table I). Liisa Holm used this prediction of accessibility and the corresponding prediction of secondary structure to infer a (correct) ab initio model of the 3D structure. However, the topography of the model corrected by one of us (C.S.) was incorrect in part. For some proteins in the contest Tim Hubbard used PHD predictions as a baseline for correct predictions of topography (Hubbard and Park and Defay and Cohen, this issue).

WHAT WENT WRONG?

Helices predicted as too long. Helices were predicted at a higher than average length (13.5 residues instead of 10). This may partly be explained by the unusual content of secondary structure in the nine proteins evaluated here: helix, 39% vs. about 32% in a set of representative proteins[27,23]; strand 18 vs. 21%. These deviations basically indicate that a data set of nine proteins is too small to be representative.

```
   prosub: propeptide of subtilisin BPN,  71 residues,    Q3  =  55%
  ....,....1....,....2....,....3....,....4....,....5....,....6....,....7.
AAEKKYIVGFKQTMSTMSAAKKKDVISEKGGKVQKQFKYVDAASATLNEKAVKELKKDPSVAYVEEDHVAHAY
Obs    EEEEEE          HHHHHHHHH                 EEEEE   HHHHH      EEEEEE
PHD    EEEEEE    HHHHHHHHH    EE    HHHHHHHHHHHHHHHH HHHHHHHHHH    EEEEE
RI733777541105566574312222434524456643343352140199999981699923665311112 79

   RTP: replication termination protein, 116 residues,    Q3  = 64%
  ....,....1....,....2....,....3....,....4....,....5....,....6....,....7
AASSTGFLVKQRAFLKLYMITMTEQERLYGLKLLEVLRSEFKEIGFKPNHTEVYRSLHELLDDGILKQIKVK
Obs          HHHHHHHHHHHHHH          HHHHHHHHH          HHHHHHHHHHHHH   EEEEEEE
PHD    HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHH    HHHHHEHHH
RI52159999999999999877999999726999999999999259972689863053447522533202 33
   ....,....8....,....9....,....10...,....11...,.
AA KEGAKLQEVVLYQFKDYEAAKLYKKQLKVELDRCKKLIEKALSDNF
Obs         EEEEEEE  HHHHHHHHHHHHHHHHHHHHHHHHHHHHH
PHD HH     HHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHHHH
RI 3189653778999721544557544569999999999999713699

   staufen3: domain 3 of staufen,          67 residues,    Q3  =  46%
  ....,....1....,....2....,....3....,....4....,....5....,....6....,..
AAPISQVHEIGIKRNMTVHFKVLREEGPAHMKNFITACIVGSIVTEGEGNGKKVSKKRAAEKMLVELQK
Obs  HHHHHHHHHHH    EEEEEEEE       EEEEEEEE  EEEEEEEE     HHHHHHHHHHHHH
PHD                EEEEE             EEEEE     EE              HHHHHHEE
RI977642344444035811213579985553148713555255159899973345623432113136 8
```

Fig. 2. Three examples of errors in secondary structure prediction. Abbreviations used: AA, amino acid in one-letter code; Obs, secondary structure assignment based on 3D structure by DSSP[58]; PHD, prediction by neural network system; RI, reliability of prediction, 0 is low, 9 is high; this index reflects the strength of the prediction, i.e., the difference between the output unit (note, the PHD networks have three output units coding for helix, strand, and rest) with the highest value (winner unit) and the output unit with the next highest value. Symbols for secondary structure assignments: H, α-helix; E (extended), β-strand; blank, rest.

Fatal for 3D modeling: confusing helices and strands. The three examples for secondary structure prediction shown (Fig. 2) are predicted at accuracy levels below average; *prosub* and *staufen3* were actually the worst predictions (Table I). For those two the dominant regular structure was detected. However, for the replication termination protein *RTP*—although predicted at a better overall accuracy than *prosub* and *staufen3* (Table I)—both strands were missed near the N-terminus and one (much too long) helix was predicted instead of two shorter ones. Reassuringly, the long helix has a region of low reliability in the central region (at YGL, Fig. 2). Both for *prosub* and *RTP* the wrong number of predicted secondary structure segments would have prevented any further correct model building, but this is less true for *staufen3,* although it is predicted with less than 50% single-residue accuracy.

Overprediction of buried residues. Prediction of relative solvent accessibility went wrong particularly for multimeric proteins. The dominant error was a strong overprediction of completely buried (0% accessible) residues (data not shown).

WHY?

Diversity of multiple alignments crucial for prediction success. For secondary structure, twice as many residues were predicted at the highest reliability index as was expected from testing PHDsec[23]; and for most values of the reliability index of the accessibility prediction, the observed values of accuracy were below the expected levels. Note that prediction strength depends on the diversity of the sequence family aligned. For some of the predictions, the alignments did not contain many sequences, e.g., only two sequences were aligned to the replication termination protein (*RTP*). This resulted in unusually high levels of prediction strength (reliability index). In general, the correctness and diversity of the multiple alignment used for prediction are crucial factors influencing prediction accuracy.

Unpredictable accuracy for nonsoluble proteins. The neural networks used for 1D predictions were trained on globular water-soluble proteins; predictions tend to be wrong for other proteins.[53] The replication termination protein illustrated this point as an interaction between dimers is crucial.[54]

Generic prediction of secondary structure segments. A fatal error for prediction-based modeling is the confusion of helices and strands (which is above average for two of the proteins in Fig. 2). An observation from many PHD predictions is that exactly this fatal error happens in some cases. For the nine contest proteins, in three out of ten cases the ends of

the secondary structure segments were correctly predicted, but the segment is a strand rather than a helix (two strands predicted as helix in *RTP* and the strand at position 24–25 of *prosub* in Fig. 2). How can a segment be placed correctly if the type is confused? Stretches of some 9–17 adjacent residues (input to neural networks) have preferences for forming regular arrays of backbone hydrogen bonds. However, some short sequence motifs can occur in both strands and helices.[55] A region may have a higher preference for forming a helix than a strand, but interactions nonlocal in sequence may result in that the formation of a β-sheet is energetically more favorable. Thus, the confusion between helices and strands indicates that the prediction is focused on preferences for formation of regular secondary structure, rather than on preferences for forming certain secondary structure types. A hypothesis that would have to be verified is that exactly those segments are confused which are formed due to nonlocal interactions.

WHAT DID WE LEARN?

First, prediction of 1D protein structure is now sufficiently accurate to be useful as a starting point for, e.g., threading techniques (single examples: Tim Hubbard, Craig Livingstone et al., Geoffrey Barton, Stephen Benner and Dietlind Gerloff in this issue; for an automated threading method[56,57]). Second, in some cases accuracy is sufficient to base correct modeling of 3D structure on 1D predictions. Third, even false predictions can contribute useful information. For example, for the least accurately predicted replication termination protein (Fig. 2) the 3D structure for one region was modeled accurately by mean-force potential-based threading (Manfred Sippl, this issue). However, the threading-based modeling excluded the N-terminal helix. This helix was predicted correctly by PHD. Thus, reliable segment predictions are useful as a starting point for 3D modeling, and PHD constitutes a tool for focusing on the most reliably predicted segments.

ACKNOWLEDGMENTS

We should like to express our gratitude to John Moult (CARB, Washington, D.C.) and to his colleagues for having organized an extremely interesting meeting that emphasized the importance of objective evaluation in a research field of increasing importance. Thanks to Tim Hubbard (MRC, Cambridge, U.K.) for encouragement that paved the way to California. Furthermore, thanks to colleagues from EMBL Heidelberg: Gert Vriend, for valuable ideas in developing PHD; Reinhard Schneider, for MaxHom and important help and discussions; Antoine de Daruvar, for managing the painful struggle with wrong formats and CPU crashes when running the PredictProtein server; and all other members of the Protein Design Group for many discussions. Thanks also to the referees for very helpful comments. Last, but not least, thanks to all those who contributed protein structure coordinates from NMR and X-ray crystallography.

REFERENCES

1. 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.
2. 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.
3. Szent-Györgyi, A. G., Cohen, C. Role of proline in polypeptide chain configuration of proteins. Science 126:697, 1957.
4. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. J., Davies, D. R., Phillips, D. C. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London) 185:422–427, 1960.
5. Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, G., Will, G., North, A. T. Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. Nature (London) 185:416–422, 1960.
6. Sternberg, M. J. E., Thornton, J. M. Prediction of protein structure from amino acid sequence. Nature (London) 271: 15–20, 1978.
7. Schulz, G. E., Schirmer, R. H. "Principles of Protein Structure." New York: Springer, 1979.
8. Fasman, G. D. "Prediction of Protein Structure and the Principles of Protein Conformation." New York: Plenum, 1989.
9. Kabsch, W., Sander, C. How good are predictions of protein secondary structure? FEBS Lett. 155:179–182, 1983.
10. 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.
11. 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.
12. Ptitsyn, O. B., Finkelstein, A. V. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 22:15–25, 1983.
13. Taylor, W. R., Thornton, J. M. Prediction of super-secondary structure in proteins. Nature (London) 301:540–542, 1983.
14. 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.
15. Biou, V., Gibrat, J. F., Levin, J. M., Robson, B., Garnier, J. Secondary structure prediction: Combination of three different methods. Prot. Eng. 2:185–191, 1988.
16. 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.
17. Gascuel, O., Golmard, J. L. A simple method for predicting the secondary structure of globular proteins: implications and accuracy. CABIOS 4:357–365, 1988.
18. Qian, N., Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884, 1988.
19. Holley, H. L., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86:152–156, 1989.
20. Zhang, X., Mesirov, J. P., Waltz, D. L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
21. Maclin, R., Shavlik, J. W. Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. Machine Learn. 11: 195–215, 1993.

22. Munson, P. J., Di Francesco, V., Porrelli, R. Prediction of protein secondary structure using linear and quadratic logistic models with penalized maximum likelihood estimation. In: "27th Hawaii International Conference on System Sciences." Hunter, L., ed. Wailea, HI: IEEE Computer Society Press, 1994: 375–384.
23. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
24. Sander, C., Schneider, R. Database of homology-derived structures and the structurally meaning of sequence alignment. Proteins 9:56–68, 1991.
25. *Russell*, R. B., Barton, G. J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234:951–957, 1993.
26. Rost, B., Sander, C., Schneider, R. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235: 13–26, 1994.
27. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Prot. Sci. 3:522–524, 1994.
28. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
29. Zuckerkandl, E., Pauling, L. Evolutionary divergence and convergence in proteins. In: "Evolving Genes and Proteins." Bryson, V., Vogel, H. J., eds. New York: Academic Press, 1965: 97–166.
30. Maxfield, F. R., Scheraga, H. A. Improvements in the prediction of protein topography by reduction of statistical errors. Biochemistry 18:697–704, 1979.
31. Sweet, R. M. Evolutionary similarity among peptide segments is a basis for prediction of protein folding. Biopolymers 25:1565–1577, 1986.
32. 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.
33. 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.
34. Frampton, J., Leutz, A., Gibson, T. J., Graf, T. DNA-binding domain ancestry. Nature (London) 342:134, 1989.
35. 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.
36. 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.
37. Niermann, T., Kirschner, K. Improving the prediction of secondary structure of 'TIM-barrel' enzymes (Corrigendum). Prot. Eng. 4:359–370, 1991.
38. Benner, S. A., Cohen, M. A., Gerloff, D. Correct structure prediction? Nature (London) 359:781, 1992.
39. 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.
40. Rost, B., Sander, C. Jury returns on structure prediction. Nature (London) 360:540, 1992.
41. Barton, G. J., Russell, R. B. Protein structure prediction. Nature (London) 361:505–506, 1993.
42. Benner, S. A., Cohen, M. A., Gerloff, D. Predicted secondary structure for the Src homology 3 domain. J. Mol. Biol. 229:295–305, 1993.
43. Boscott, P. E., Barton, G. J., Richards, W. G. Secondary structure prediction for modelling by homology. Prot. Eng. 6:261–266, 1993.
44. 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.
45. Gibson, T. J., Thompson, J. D., Heringa, J. The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding nucleic acid. FEBS Lett. 324:361–366, 1993.
46. Musacchio, A., Gibson, T., Rice, P., Thompson, J., Saraste, M. The PH domain: A common piece in the structural patchwork of signalling proteins. TIBS 18:343–348, 1993.
47. Robson, B., Garnier, J. Nature (London) 361:506, 1993.
48. Livingstone, C. D., Barton, G. J. Secondary structure prediction from multiple sequence data: Blood clotting factor XIII and Yersinia protein-tyrosine phosphatase. Int. J. Peptide Protein Res. 44:239–244, 1994.
49. Perkins, S. J., Smith, K. F., Williams, S. C., Haris, P. I., Chapman, D., Sim, R. B. The secondary structure of the von Willebrand factor type A domain in factor B of human complement by Fourier transform infrared spectroscopy. J. Mol. Biol. 238:104–119, 1994.
50. Rost, B., Sander, C. Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226, 1994.
51. Levin, J. M., Pascarella, S., Argos, P., Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. Prot. Eng. 6:849–854, 1993.
52. Salamov, A. A., Solovyev, V. V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247:11–15, 1995.
53. Rost, B., Casadio, R., Fariselli, P., Sander, C. Prediction of helical transmembrane segments at 95% accuracy. Prot. Sci. 4:521–533, 1995.
54. Langley, D. B., Smith, M. T., Lewis, P. J., Wake, R. G. Protein-nucleoside contacts in the interaction between the replication terminator protein of Bacillus subtilis and the DNA terminator. Mol. Microbiol. 10:771–779, 1993.
55. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
56. Rost, B. Fitting 1D predictions into 3D structures. In: "Protein Folds: A Distance Based Approach." Bohr, H., Brunak, S., eds. Boca Raton, FL: CRC Press (in press).
57. Rost, B. TOPITS: Threading one-dimensional predictions into three-dimensional structures. In: "The Third International Conference on Intelligent Systems for Molecular Biology." Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengaver, T., Wodak, S., eds. Cambridge, U. K., July 16–19, 1995: Menlo Park, CA: AAAI Press, 1995: 314–321.
58. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
59. Wang, Z.-X. Assessing the accuracy of protein secondary structure. Nature Struct. Biol. 1:145–146, 1994.