

Better 1D Predictions by Experts With Machines

Burkhard Rost*

European Molecular Biology Laboratory, Heidelberg, Germany

ABSTRACT Accuracy of predicting protein secondary structure and solvent accessibility has been improved significantly by using evolutionary information contained in multiple sequence alignments. For the second Asilomar meeting, predictions were made automatically for all targets using the publicly available prediction service PredictProtein. Additionally, a semiautomatic procedure for generating more informative alignments was used in combination with the PHD prediction methods. Results confirmed the estimates for prediction accuracy. Furthermore, the more informative alignments yielded better predictions. The fairly accurate predictions of 1D structure were successfully used by various groups for the Asilomar meeting as first step toward predicting higher dimensions of protein structure. *Proteins, Suppl. 1:192–197, 1997.* © 1998 Wiley-Liss, Inc.

Key words: prediction of protein secondary structure; residue solvent accessibility; multiple alignments; neural networks

INTRODUCTION

Simplifying the Structure Prediction Problem

The second Asilomar meeting has confirmed that after 40 years of ardent research, theory still cannot predict protein three-dimensional structure (3D) from sequence, in general.¹ However, the rapidly growing sequence-structure gap (number of known protein structures vs. number of known protein sequences) has enticed theoreticians to solve simplified prediction problems.² An extreme simplification is the prediction of protein structure in one dimension (1D), as represented by strings of, for example, secondary structure, and residue solvent accessibility. Theoreticians are lucky because the 1D prediction problem is not only the task they can accomplish best, but in that even partially correct predictions of 1D structure are useful, for example, for predicting protein function, or functional sites.

Breakthrough of Third Generation Prediction Methods

The first generation of 1D prediction methods were based on physicochemical principles, expert rules, and statistics of single residues.³ The second generation incorporated the influence of residues

adjacent to the residue for which 1D structure was predicted (local information).⁴ These secondary structure prediction methods shared three major shortcomings: (1) prediction accuracy was limited to about 60% accuracy (percentage of residues predicted correctly in either of the three states helix, strand, other), (2) β strands were predicted at typically < 40% accuracy, (3) predicted secondary structure segments were, on average, only half as long as observed segments. Some methods were tailored to overcome one of these problems (long-range information^{5,6}; β strand accuracy⁷; length⁸). However, only recently have automatic methods been developed that overcome most of these shortcomings. The most important trick of the third generation prediction tools of the 90's is the use of evolutionary information contained in multiple alignments of protein families.^{2,9–21} To outsiders the superiority of the third generation tools over their predecessors (which unfortunately are still being used by major sequence analysis packages, such as GCG²²) may appear marginal. However, the usefulness of the third generation methods was demonstrated in the second Asilomar meeting in which these automated tools were routinely used by sequence analysis experts.

MATERIALS AND METHODS

From Sequence to 1D Structure

The major step in improving 1D predictions has been the use of evolutionary information contained in multiple sequence alignments. Generating the information fed into the neural network system PHD²⁰ required four steps: a data base search for homologues (method BLAST²³), (2) a refined profile-based dynamic-programming alignment of the most likely homologues (method MAXHOM²⁴), (3) a decision for which proteins will be considered as homologues (length-dependent cut-off for pairwise sequence identity^{25,26}), and (4) a final refinement, and extraction of the resulting multiple alignment. In general, prediction accuracy is better when predictions are based on better alignments. Better alignments are defined by: (i) fewer incorrectly aligned residues; (ii) greater divergence within the family of sequences. In practice, these two conditions are oppo-

*Correspondence to: Dr. Burkhard Rost, EMBL, 69 012 Heidelberg, Germany.

E-mail: rost@embl-heidelberg.de; http://www.embl-heidelberg.de/~rost/

Received 9 May 1997; Accepted 18 August 1997

nents in that less similar homologues are more likely to be misaligned.

Completely Versus Almost Automatic

The PHD prediction methods are automatically available via the internet service PredictProtein²⁰ (send the word *help* to PredictProtein@EMBL-Heidelberg.DE, or use the WWW interface²⁷). Users have the choice between the fully automatic procedure taking the query sequence through the entire cycle, or expert intervention into the generation of the alignment. For the Asilomar contest, both these modes of operation were explored. The following changes were made with respect to the usual PredictProtein service:

1. Rather than SWISS-PROT, a nonredundant data base of all known protein sequences was searched.
2. The cutoff for accepted homologues was lowered from 30% to 25% pairwise sequence identity.
3. The final list of putative homologues was inspected visually; some proteins were excluded from the list. PredictProtein users typically continue with two additional time-consuming expert interventions.
4. Visual correction of the final alignment.
5. Investigation of how the prediction of particular segments depends on the alignment.

Steps 4 and 5 were not performed for the Asilomar targets.

Prediction Targets

Secondary structure and residue solvent accessibility was predicted for all Asilomar targets. Here, results were compiled for the 15 targets for which structures were available. Secondary structure predictions comprised the predictions of secondary structure state (helix, strand, other), and a reliability index for each residue; relative solvent accessibility predictions gave the percentage of predicted solvent accessibility (additionally projected onto a two-state model (buried: $\leq 16\%$, exposed $> 16\%$), and a reliability index for each residue.

RESULTS

What Went Well?

Prediction accuracy within expected range

(1) Secondary structure prediction: the accuracy was about 74% (three-state per-residue and per-segment scores). (2) Solvent accessibility prediction: the accuracy was about 69% (two-state score); the correlation between observed and predicted relative solvent accessibility was 0.5; predictions were best for residues observed in strands, and worst for residues with no regular secondary structure; predictions were best for the charged amino acids aspartic,

glutamic, and lysine, and for methionine. Overall, secondary structure prediction accuracy using PHD on the Asilomar proteins was slightly higher than expected; solvent accessibility prediction accuracy slightly lower than expected.²⁰

Reliability of prediction correlated with accuracy

Prediction accuracy varied largely between different proteins (Fig. 1A). However, the reliability of PHD predictions enabled estimating on which side of such a distribution the prediction for a given protein was expected (Fig. 1B). Furthermore, individual residues could be correctly labelled for which the prediction was expected to be more likely accurate (Fig. 1C). For example, half of all residues predicted in a helix were predicted with the highest reliability index; 90% of these were correctly predicted (Fig. 1C).

More informative alignments yielded better predictions

By manually improving the multiple alignments used for the PredictProtein server (Fig. 1), the prediction of secondary structure improved from 70% (three-state per-residue accuracy for fully-automatic alignment selections from PredictProtein server²⁰) to 74% (semiautomatic alignment selection used for CASP2 submissions). Solvent accessibility predictions were improved from 67% (two-state per-residue accuracy for fully automatic alignment selection) to 69% (CASP2 submissions).

What Went Wrong?

Confusing helices and strands

The two examples shown in Figure 2 represented the worst cases for the secondary structure prediction in terms of per-residue and per-segment accuracy. However, even more fatal for using 1D predictions for further steps toward 3D prediction were cases for which the prediction confused helices and strands. On average, such bad predictions were made for about 8% of all residues. Particularly bad were the values for target t11 (19% of residues confused), and for target t32 (Fig. 2; 13% of residues confused). None of the confused segments was predicted with high reliability indices.

Helices too long, strands too short

Helices were predicted at a higher than average length (12.5 predicted vs. 10.3 observed); strands were predicted at a lower than average length (5.2 vs. 6.7). These values were not representative for the PHD averages, and might have originated from unusually high percentages of secondary structure in the 15 CASP2 targets (38% helix, 24% strand compared to about 32% helix and 21% strand in a representative subset of PDB,²⁸ data not given).

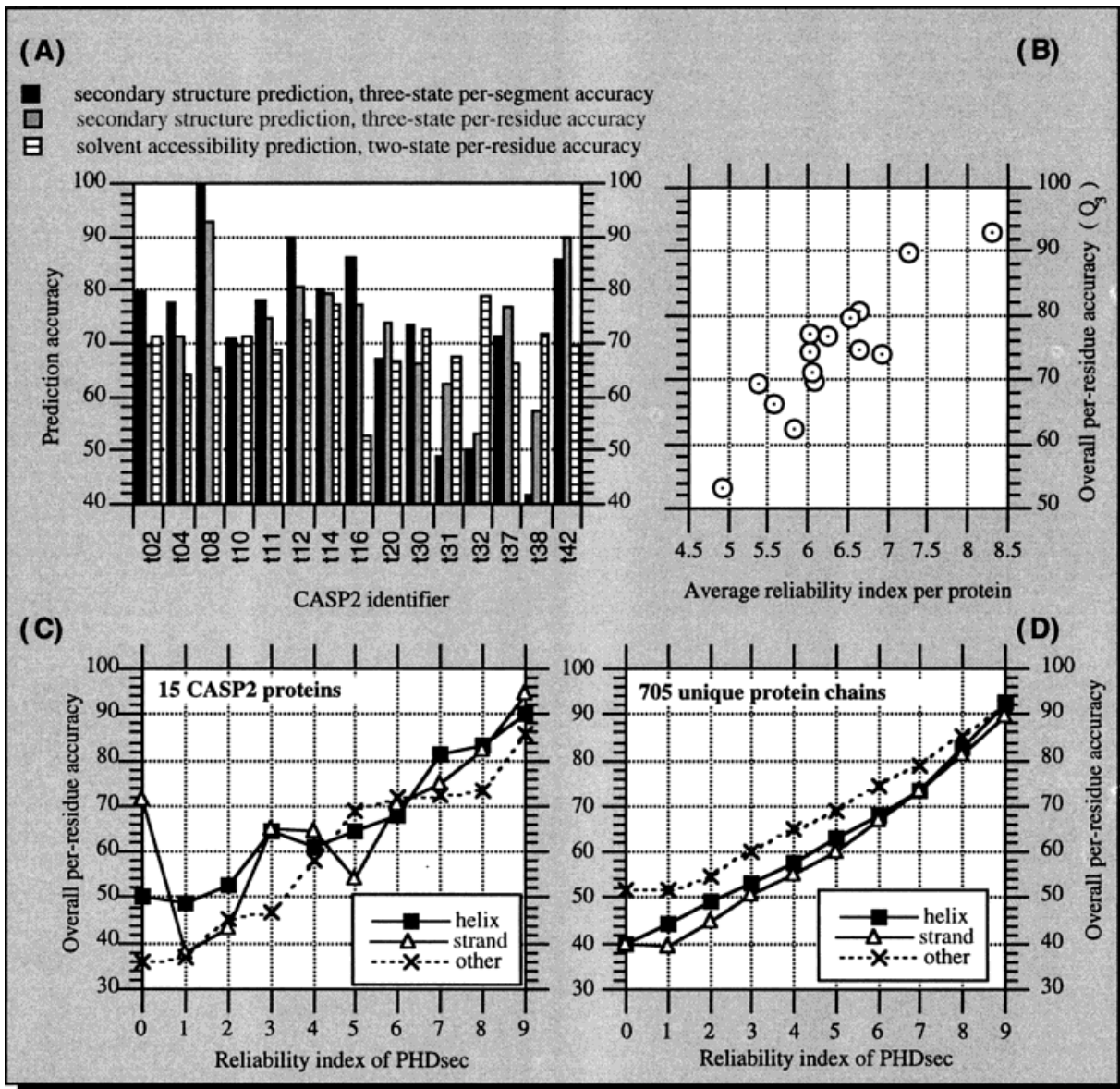


Fig. 1. Prediction accuracy for CASP2 targets. **A:** 1D prediction accuracy was described by three scores: (1) per-segment accuracy in predicting secondary structure (Sov_3 , defined in Ref. 37), (2) per-residue accuracy in predicting secondary structure (Q_3 , defined in Ref. 37), and (3) per-residue accuracy in predicting residue solvent accessibility (Q_2 , defined in Ref. 13). Exceptionally bad predictions did not coincide, i.e., the lowest value for each score did not occur for the same protein. In fact, for the three proteins for which secondary structure prediction was worst (t31, t32, t38) accessibility predictions were rather accurate; and the worst predicted accessibility for t16 coincided with an extremely good secondary structure prediction. **B:** The reliability index, scaled from 0 (low) to 9 (high), reflects the strength of the prediction for each residue. Here, the reliability index was averaged over each protein. The protein average correlated with the overall per-residue accuracy of secondary structure prediction: the worst predicted protein (t32) had the lowest average reliability index; the best predicted ones (t08, t42) had the highest average

reliability indices. **C and D:** The 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. How likely was a residue predicted in an α helix with a reliability index of n , predicted correctly? The two plots were derived for different test sets, (C) reflected the statistics on 15 Asilomar targets, (D) statistics on a 50 times larger set of 705 sequence-unique proteins. For example, prediction accuracy tended to surpass the 70% accuracy level for residues predicted at levels of $RI \geq 6$; about two-thirds of all residues were predicted at that level. To illustrate the fraction of residues predicted at highest reliability: for the set of 705 about 40% of the helical residues were predicted at $RI = 9$ (93% of these were correct); and for the Asilomar set about 50% of the helical residues were predicted at $RI = 9$ (90% of which were correctly predicted). Note that Figures C and D show the noncumulative values. The cumulative distributions answering the question "How high is the expected prediction accuracy for all residues predicted at higher reliability?" is given elsewhere.^{10,12,13,20,33}

```

t32, beta-cryptogein, 98 residues, Q3 = 53, Sov3 = 50
.....1.....2.....3.....4.....5.....6.....7.....8
AA TACTATQQTAAYKTLVLSILSDASFNQCSTDSGYSLTAKALPTTAQYKLMCASTACNTMIKKIVTLNPPNCDLTVPTSGL
Obs HHHHHHHHHHHHHH HHHHHHHHHH HHHHHHHH HHHHHHHHHHHH EE
PHD HHHHHHEEEEE EEEEE HHHHHHHHHHHHHHHHHH EEE EE
RI 9976545436645561154334410114688612445111487236754313699999999962129996254166427
.....9.....10.....11.....12.....13.....14.....15.....16
AA VLVVYSYANGFSNKCSSL
Obs EE HHHHHHHHHHHH
PHD EEEEE
RI 888551137844123579

t38, CBDN1, cellulose degradation, 152 residues, Q3 = 57, Sov3 = 42
.....1.....2.....3.....4.....5.....6.....7.....8
AA ASPIGEGTFDDGPEGWVAYGTDGPLDTSTGALCVAVPAGSAQYGVGVVINGVAIEEGTTYTLRYTATASTDVTVRALVQG
Obs EEEEEEE EEEEEEE EEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
PHD EE EEEEE EEEEE EEEEE EEEE EEEEEEEEE EEEEEEE
RI 96763254479987742448899514578759999549985213888752258852555799999851253499999836
.....9.....10.....11.....12.....13.....14.....15.....16
AA NGAPYGTVLDTSPALTSEPRQVTETFTASATYPATPAADDPEGQIAFQLGGFSADAWTLCLDDVALDSEVEL
Obs EE EEEEE EEE EEEEEEEEE EEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
PHD E EEE EEEEEEE EEEEE EEEEE EEEEEEEEEEEEE
RI 999943122333434778416999997313478878889998289998389764421899895434345419

```

Fig. 2. Examples for prediction errors. Two examples of errors in secondary structure prediction. Secondary structure prediction was worst for these two proteins (Fig. 1A, note: for t38 the prediction had to be based on a single sequence). AA, amino acid in one-letter code; Obs, secondary structure assignment based on

3D structure by DSSP³⁸; PHD, prediction by neural network system; RI, reliability of prediction (0 is low, 9 is high). Symbols for secondary structure assignments: H, α helix; E (extended), β strand; blank, other.

Structural class predicted at accuracy levels below average

The composition of secondary structure enables a rough classification of proteins into structural classes.²⁹ On average, secondary structure predictions from PHDsec predict 75% of all proteins correctly in one of the four classes: all- α , all- β , α/β , other.^{20,30} For the CASP2 targets the class classification was correct for 67% of the proteins, only. The dominant error was to predict strands for all- α proteins (and to consequently place those proteins into the class "other," rather than into the class "all- α "). However, the average content of secondary structure was predicted about as accurately as expected: differences between predicted and observed compositions were 7% for helix, and 9% for strand. Thus, the difference between CASP2 and expected classification error could be attributed to the small dataset.

Overprediction of buried residues

Most buried residues ($\leq 16\%$ relative solvent accessibility) were predicted as buried (76%). However, this was accomplished by an overprediction of buried residues, as only 60% of the residues predicted to be buried were actually observed in that state. The dominant error was a strong overprediction of completely buried (0% accessible) residues. In general,

residues were clearly overpredicted in the ranges 49–64% accessibility, and clearly underpredicted in the ranges 1–4% and 64–81%. (*Note:* the other side of the same coin was that exposed residues were underpredicted: 80% of the residues predicted to be exposed were observed in that state, however, only 64% of the residues observed in the exposed state were actually predicted.)

Why?

Correct alignment crucial for correct prediction

Alignments used for the input to the PHD neural networks should be both informative (high level of diversity; many sequences), and correct. The semiautomatic generation of multiple alignments used for the CASP2 submissions clearly improved the information content in the alignments, and thus prediction accuracy. However, including proteins from the twilight zone³¹ of 25–30% pairwise sequence identity may be fatal in two ways. First, some of the included proteins may not be structurally similar to the protein for which 1D structure is predicted. Second, the lower the level of pairwise sequence identity, the higher the likelihood of misaligning some residues. This became particularly obvious, when the alignments for the worst predicted proteins were changed (*after* the meeting). Secondary structure prediction accuracy could be increased by simply excluding

some less likely family members: for target t31, Q_3 (three-state per-residue accuracy) from 62% to 68%, Sov_3 (three-state per-segment accuracy) from 54% to 65%; for target t32, Q_3 from 53% to 56%, Sov_3 from 54% to 56%; for target t38 Q_3 from 57% to 63%, Sov_3 from 42% to 48%. Prediction accuracy was clearly below average for proteins for which no alignments were available (such as for t38, Fig. 2). The second effect of falsely aligned residues was difficult to estimate. However, the extent of the first effect illustrated that alignment errors were fatal.

Prediction accuracy lower for unusual proteins

The PHD neural networks were trained on globular water-soluble proteins; predictions tend to be wrong for other proteins. One example from the CASP2 set was t32 a small protein (98 residues; Fig. 2) which is stabilized by three cysteine-bridges. Fundamental mistakes in the secondary structure prediction were around the cysteine-bridges (Fig. 2). However, on average proteins with cysteine-bridges were not predicted less accurately (Arthur Lesk, this issue). For the prediction of solvent accessibility another effect becomes crucial: the interaction between protein chains: overall accessibility was predicted worst for t16. However, many of the residues “falsely” predicted as buried were actually observed at interfaces between the three chains of the protein (data not shown). In general, residues predicted to be buried and observed to be exposed, often indicate binding interfaces.³²

Confusing helices and strands partly due to using local information

A fatal error for prediction-based modeling is the confusion of helices and strands. Exactly this fatal error happens frequently for PHD predictions (for 7 of the 15 CASP2 targets; statistics on a larger dataset³³). Often the beginning and the ends of the confused segments are correctly identified (target t02, strand 13; target t11, strand 1; target t14, strand 4, helix 8; target t16, helix 2; target t31 strand 8). Only for two confused segments the reliability of at least one residue was above a value of $RI = 6$ (helix 8 in t14, and helix 1 in t16). Nevertheless, how can a segment be placed correctly if the type is confused? Secondary structure formation is partly determined by residue interactions non local in sequence. Such information is captured by the PHD predictions only to some extent. A region may have a higher preference for forming a helix than a strand (and vice versa), but interactions nonlocal in sequence may result in that the formation of a β sheet (α helix) is energetically more favorable. Indeed, the confusion between helices and strands can often be attributed to hydrogen bonds stabilized by nonlocal interresidue contacts.³⁴

Fifteen proteins are not representative

Some of the “errors” were specific for the CASP2 targets. The major reason for that was that 15 proteins were not enough to comprise a representative subset of all proteins (difference between Fig. 1C and D).

CONCLUSIONS: WHAT DID WE LEARN? Easy To Be Wise Afterward?

Inspecting the examples for which predictions went wrong tended to produce arguments for why that was so. However, such reasoning in some cases appeared rather premature: proteins for which secondary structure was predicted below average tended to differ from those for which solvent accessibility was predicted below average (Fig. 1A), although many of the arguments would apply to both prediction methods (such as the stabilization by cysteine bridges for target t32).

Generating More Informative Alignments Is Straightforward

The difference in prediction accuracy between the fully automatic and the semiautomatic selection of the alignment (two to four percentage points) illustrated that prediction accuracy could be improved significantly without changing the final prediction method (PHD²⁰). The procedure used for the CASP2 submission could be automated. (The major technical problem, currently, is the lack of CPU resources available at EMBL for the PredictProtein service.) Another point was illustrated for the CASP2 targets: monitoring how predictions change in response to the alignment (including more or less proteins) is an excellent means of arriving at better expert-driven predictions.

CASP: Good for Testing Methods, But Not Representative

1D structure predictions comprise excellent examples for prediction methods, in general, since we have large datasets for which we can estimate prediction accuracy. Such tests reveal that prediction accuracy differs between different proteins (with one standard deviation of about ten percentage points). How many proteins are representative? To approach the answer, consider the following experiment: first, average prediction accuracy and its standard distribution are compiled for a set of 705 unique proteins chains³³; second, from the set of 705 chains 20 are picked at random; this is repeated until average accuracy and standard distribution match that of the set of 705 proteins. How many repeats would it take? The answer: about five to ten. Thus, the following conclusions from 1D predictions in CASP2 evolve for users (and editors): don't trust too much methods that (1) were not tested in CASP, (2) revealed much lower values of accuracy than

published, and (3) that were successful in CASP, but never evaluated on larger databases.

1D Predictions Now Accurate Enough as First Step in Structure Prediction

Many of the third-generation predictions of 1D structure are accurate enough to become a first step in predicting higher dimensions of protein structure (Arthur Lesk, this issue). A prominent application of PHD predictions was threading of the CASP2 targets (e.g., Murzin, or Fischer, Eisenberg et al., this issue). Even an automatic PHD-threading procedure yielded^{35,36} relatively good results for recognizing the correct fold.

ACKNOWLEDGMENTS

I thank Sean O'Donoghue (EMBL, Heidelberg) for helpful discussions and for critically reading the manuscript, all those who contributed essentially to the CASP2 meeting by making their experimental structure determinations available before publication, the assessors Arthur Lesk (LMB, Cambridge) and Michael Levitt (Stanford University), and all those who were involved in organizing that meeting, to name a few: John Moult (CARB, Washington), Tim Hubbard (Sanger Centre, England), Stephen Bryant (NIH, Washington), Jan Pedersen (CARB, Washington), and Krzysztof Fidelis (LNL, Livermore).

REFERENCES

- Rost, B., O'Donoghue, S.I. Sisyphus and prediction of protein structure. *CABIOS* 13:345–356, 1997.
- Rost, B., Sander, C. Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* 25:113–136, 1996.
- Kabsch, W., Sander, C. How good are predictions of protein secondary structure? *FEBS Lett.* 155:179–182, 1983.
- Fasman, G.D. *Prediction of Protein Structure and the Principles of Protein Conformation*. New York: Plenum, 1989.
- Maxfield, F.R., Scheraga, H.A. Improvements in the prediction of protein topography by reduction of statistical errors. *Biochemistry* 18:697–704, 1979.
- Zvebil, 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.
- Gascuel, O., Golmard, J.L. A simple method for predicting the secondary structure of globular proteins: Implications and accuracy. *CABIOS* 4:357–365, 1988.
- Kabsch, W., Sander, C. *Segment83*. unpublished 1983.
- 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.
- Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–599, 1993.
- Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona fide prediction of aspects of protein conformation. *J. Mol. Biol.* 235:926–958, 1994.
- Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55–72, 1994.
- Rost, B., Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20:216–226, 1994.
- Wako, H., Blundell, T.L. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. Mol. Biol.* 238:682–692, 1994.
- Wako, H., Blundell, T.L. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol.* 238:693–708, 1994.
- Barton, G.J. Protein secondary structure prediction. *Curr. Opin. Struct. Biol.* 5:372–376, 1995.
- Gerloff, D.L., Chelvanayagam, G., Benner, S.A. A predicted consensus structure for the protein-kinase c2 homology (c2h) domain, the repeating unit of synaptotagmin. *Proteins* 22:299–310, 1995.
- Salamov, A.A., Solovyev, V.V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *J. Mol. Biol.* 247:11–15, 1995.
- Di Francesco, V., Garnier, J., Munson, P.J. Improving protein secondary structure prediction with aligned homologous sequences. *Protein Sci.* 5:106–113, 1996.
- Rost, B. PHD: Predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* 266:525–539, 1996.
- Thompson, M.J., Goldstein, R.A. Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* 25:38–47, 1996.
- Devereux, J., Haeblerli, P., Smithies, O. GCG package. *Nucleic Acids Res.* 12:387–395, 1984.
- Altschul, S.F., Gish, W. Local alignment statistics. *Methods Enzymol.* 266:460–480, 1996.
- Schneider, R. *Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen*. Doctoral thesis, University of Heidelberg, 1994.
- Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826, 1986.
- Sander, C., Schneider, R. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9:56–68, 1991.
- Rost, B. PredictProtein: Internet prediction service. WWW document (<http://www.embl-heidelberg.de/predictprotein/>): EMBL, 1997.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., et al. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542, 1977.
- Levitt, M., Chothia, C. Structural patterns in globular proteins. *Nature* 261:552–558, 1976.
- Rost, B. Observed secondary structure content for 721 proteins. WWW document (<http://www.embl-heidelberg.de/~rost/Res/96A-SecStrContent.html>): EMBL Heidelberg, Germany, 1996.
- Doolittle, R.F. *Of URFs and ORFs: A primer on How To Analyze Derived Amino Acid Sequences*. Mill Valley, CA: University Science Books, 1986.
- Hubbard, T., Tramontano, A., Barton, G., et al. Update on protein structure prediction: Results of the 1995 IRBM workshop. *Folding Design* 1:R55–R63, 1996.
- Rost, B. Expected prediction accuracy of PHD. WWW document (<http://www.embl-heidelberg.de/~rost/Res/96D-ExpAccuracyPHD.html>): EMBL Heidelberg, Germany, 1996.
- Rychlewski, L., Godzik, A. Secondary structure predictions: In quest of forces that shape the local protein structure. The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, USA, 1996.
- Rost, B. TOPITS: Threading one-dimensional predictions into three-dimensional structures. In: Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S. (eds.). *Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, England. Menlo Park, CA: AAAI Press, 1995:314–321.
- Rost, B., Schneider, R., Sander, C. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270:471–480, 1997.
- Rost, B., Sander, C., Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13–26, 1994.
- Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637, 1983.