

Rost papers Bioinformatics 1992-1999

Columbia University
Department of Biochemistry and Molecular Biophysics
650 West 168th Str. BB217
New York, NY 10032, USA

Email: rost@columbia.edu
WWW: <http://cubic.bioc.columbia.edu/>
Tel: +1-212-305-3773
Fax: +1-212-305-7932

Bibliography

1. B Rost and C Sander (1992) Jury returns on structure prediction. **Nature** 360:540.
2. B Rost and C Sander (1992) Exercising multi-layered networks on protein secondary structure. In: O Benhar, S Brunak, P DelGiudice and M Grandolfo (eds.). *Neural Networks: From Biology to High Energy Physics*. Elba, Italy: International Journal of Neural Systems:209-220.
3. B Rost and G Vriend (1993) Neural networks in chemistry. **CDA News** 8:24-27.
4. B Rost and C Sander (1993) Prediction of protein secondary structure at better than 70% accuracy. **J. Mol. Biol.** 232:584-599.
5. B Rost, C Sander and R Schneider (1993) Progress in protein structure prediction? **TIBS** 18:120-123.
6. B Rost and C Sander (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. **Proc. Natl. Acad. Sci. U.S.A.** 90:7558-7562.
7. B Rost and C Sander (1993) Secondary structure prediction of all-helical proteins in two states. **Prot. Engin.** 6:831-836.
8. T Meitinger, A Meindl, P Bork, B Rost, C Sander, M Haasemann and J Murken (1993) Molecular modelling of the norrie disease protein predicts a cysteine knot growth factor tertiary structure. **Nat. Gen.** 5:376-380.
9. B Rost, C Sander and R Schneider (1994) Phd - an automatic server for protein secondary structure prediction. **CABIOS** 10:53-60.
10. B Rost, C Sander and R Schneider (1994) Redefining the goals of protein secondary structure prediction. **J. Mol. Biol.** 235:13-26.
11. B Rost and C Sander (1994) Combining evolutionary information and neural networks to predict protein secondary structure. **Proteins** 19:55-72.
12. B Rost, C Sander and R Schneider (1994) Evolution and neural networks - protein secondary structure prediction above 71% accuracy. In: L Hunter (eds.). *27th Hawaii International Conference on System Sciences*. Wailea, Hawaii, U.S.A.: Los Alamitos, CA: IEEE Society Press:385-394.
13. B Rost and C Sander (1994) 1d secondary structure prediction through evolutionary profiles. In: H Bohr and S Brunak (eds.). *Protein structure by distance analysis*. Amsterdam, Oxford, Washington: IOS Press:257-276.
14. B Rost and C Sander (1994) Conservation and prediction of solvent accessibility in protein families. **Proteins** 20:216-226.
15. B Rost and C Sander (1994) Structure prediction of proteins - where are we now? **Curr. Opin. Biotech.** 5:372-380.
16. B Rost, R Casadio, P Fariselli and C Sander (1995) Prediction of helical transmembrane segments at 95% accuracy. **Prot. Sci.** 4:521-533.
17. L Holm, B Rost, C Sander, R Schneider and G Vriend (1994) Data based modeling of proteins. In: S Doniach (eds.). *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions*. New York: Plenum Press:277-296.
18. B Rost and C Sander (1995) Protein structure prediction by neural networks. In: M Arbib (eds.). *The handbook of brain theory and neural networks*. Cambridge, MA: Bradford Books/The MIT Press:772-775.
19. B Rost (1995) Fitting 1-d predictions into 3-d structures. In: H Bohr and S Brunak (eds.). *Protein folds: A distance based approach*. Boca Raton, Florida: CRC Press:132-151.
20. B Rost (1995) Topits: Threading one-dimensional predictions into three-dimensional structures. In: C Rawlings, D Clark, R Altman, L Hunter, T Lengauer and S Wodak (eds.). *Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, England: Menlo Park, CA: AAAI Press:314-321.
21. B Rost and C Sander (1995) Progress of 1d protein structure prediction at last. **Proteins** 23:295-300.
22. B Rost (1996) Nn which predicts protein secondary structure. In: E Fiesler and R Beale (eds.). *Handbook of neural computation*. New York: Oxford Univ. Press:G4.1.
23. B Rost (1996) Phd: Predicting one-dimensional protein structure by profile based neural networks. **Meth. Enzymol.** 266:525-539.
24. B Rost, R Casadio and P Fariselli (1996) Refining neural network predictions for helical transmembrane proteins by dynamic programming. In: D States, P Agarwal, T Gaasterland, L Hunter and RF Smith (eds.). *Fourth International Conference on Intelligent Systems for Molecular Biology*. St. Louis, M.O., U.S.A.: Menlo Park, CA: AAAI Press:192-200.
25. B Rost, R Casadio and P Fariselli (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. **Prot. Sci.** 5:1704-1718.
26. B Rost and A Valencia (1996) Pitfalls of protein sequence analysis. **Curr. Opin. Biotech.** 7:457-461.
27. B Rost, R Schneider and C Sander (1997) Protein fold recognition by prediction-based threading. **J. Mol. Biol.** 270:471-480.
28. B Rost (1997) Protein structures sustain evolutionary drift. **Folding & Design** 2:S19-S24.

29. B Rost and SI O'Donoghue (1997) Sisyphus and prediction of protein structure. **CABIOS** 13:345-356.
30. B Rost (1997) Learning from evolution to predict protein structure. In: B Olsson, D Lundh and A Narayanan (eds.). BCEC97: Bio-Computing and Emergent Computation. Skövde, Sweden: World Scientific:87-101.
31. B Rost, S O'Donoghue and C Sander (1998) Midnight zone of protein structure evolution. EMBL Heidelberg <http://www.columbia.edu/~rost/Papers/PreEvolution96.html>.
32. MA Andrade, SI O'Donoghue and B Rost (1998) Adaptation of protein surfaces to subcellular location. **J. Mol. Biol.** 276:517-525.
33. B Rost (1997) Better 1d predictions by experts with machines. **Proteins** Suppl. 1:192-197.
34. B Rost (1998) Protein structure prediction in 1d, 2d, and 3d. In: PvR Schleyer, NL Allinger, T Clark, J Gasteiger, PA Kollman, HF Schaefer III and PR Schreiner (eds.). The encyclopaedia of computational chemistry. Chichester: John Wiley & Sons:2242-2255.
35. B Rost and R Schneider (1999) Pedestrian guide to analysing sequence databases. In: K Ashman (eds.). Core techniques in biochemistry. Heidelberg: Springer:in press.
36. B Rost (1999) Twilight zone of protein sequence alignments. **Prot. Engin.** 12:85-94.
37. B Rost (1999) Neural networks for protein structure prediction: Hype or hit? CUBIC, Columbia University, Dept. Biochemistry & Molecular Biophysics <http://www.columbia.edu/~rost/Papers/>.
38. B Rost (1998) Marrying structure and genomics. **Structure** 6:259-263.
39. B Rost (1999) Short yeast orfs: Expressed protein or not? CUBIC preprint: CUBIC, Columbia University, Dept. of Biochemistry & Mol. Biophysics http://cubic.bioc.columbia.edu/papers/1999_globe.

Abstracts

Jury returns on structure prediction

Burkhard Rost & Chris Sander

Quote: 1992 Nature 360, 540

Exercising Multi-layered Networks on Protein Secondary Structure

Burkhard Rost & Chris Sander

Quote: 1992 Neural Networks: From Biology to High Energy Physics 3, 209-220

The quality of a multi-layered network predicting the secondary structure of proteins is improved substantially by: (i) using information about evolutionarily conserved amino acids (increase of overall accuracy by six percentage points), (ii) balancing the training dynamics (increase of accuracy for strand), and (iii) combining uncorrelated networks in a jury (increase two percentage points). In addition, appending a second level structure-to-structure network results in better reproduction of the length of secondary structure segments.

Neural Networks in Chemistry

Burkhard Rost & Gerrit Vriend

Quote: 1993 CDA News 8, 24-27

The attempts to understand the functioning of the brain and to improve computers have profited from one another since the early days of electronic calculating. Over the last decade the application of artificial neural networks - implemented on computers - has become popular for various pattern recognition tasks. The basic procedure is that patterns are presented to a network that learns to extract intrinsic features and to group the patterns into classes. The networks do not only perform arbitrarily complicated distinction tasks but are able, as well, to generalize, i.e., to perform the classification for new patterns. This means that a network for e.g. zip code recognition does not only learn to distinguish between the hand-written zip codes on those envelopes presented to it for learning, but it learns as well to distinguish the codes for any future incoming envelope. This can be achieved, because the network learns to extract certain features from the letters and learns to relate these features to the required decision like "this is 6900 for Heidelberg" or not. Typical applications are the recognition of faces, speech, handwriting, QSAR analysis, particle detection in high energy physics, the prediction of developments in stock exchange markets, and the prediction of protein and gene structure. (For a survey of the properties of neural networks see e.g. the books of Hertz, 91; or Müller, 90, or the article by Cowan, 90.

Prediction of protein secondary structure at better than 70% accuracy

Burkhard Rost & Chris Sander

Quote: 1993 J. Mol. Biol. 232, 584-599

We have trained a two layered feed-forward neural network on a non-redundant database of 130 protein chains to predict the secondary structure of water-soluble proteins. A new key aspect is the use of evolutionary information in the form of multiple sequence alignments that are used as input in place of single sequences. The inclusion of protein family information in this form increases the prediction accuracy by 6-8 percentage points. A combination of three levels of networks results in an overall three state accuracy of 70.8% for globular proteins (sustained performance). If four membrane protein chains are included in the evaluation, the overall accuracy drops to 70.2%. The prediction is well balanced between α -helix, β -strand and loop. 65% of the observed strand residues are predicted correctly. The accuracy in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy. The performance accuracy is verified by a seven-fold cross-validation test, and an additional test on 26 recently solved proteins. Of particular practical importance is the definition of a position-specific reliability index. For the half of the residues predicted with high reliability the overall accuracy increases to better than 82%. A further strength of the method is the more realistic prediction of segment length. The protein family prediction method is available for testing by academic researchers via an electronic mail server.

Progress in protein structure prediction?

Burkhard Rost,

Reinhard Schneider & Chris Sander

Quote: 1993 TIBS 18, 120-123

Prediction of protein secondary structure is an old problem and progress has been slow over the years. Recently, spectacular success has been claimed in the blind prediction of the catalytic subunit of the cAMP dependent protein kinase. When predictions in this and other test cases are assessed critically, some claims of prediction success turn out to be exaggerated, but a kernel of real progress remains: protein structure prediction can be improved substantially when a family of related sequences is available. Enough so that molecular biologists equipped with a new amino acid sequence and a multiple sequence alignment in hand may be tempted to test the new prediction methods.

Improved prediction of protein secondary structure by use of sequence profiles and neural networks

Burkhard Rost & Chris Sander

Quote: 1993 Proc. Natl. Acad. Sci. U.S.A. 90, 7558-7562

The explosive accumulation of protein sequences in the wake of large scale sequencing projects is in stark contrast to the much slower experimental determination of protein structures. Improved methods of structure prediction from the gene sequence alone are therefore needed. Here, we report a substantial increase in both the accuracy and quality of secondary structure predictions, using a neural network algorithm. The main improvements come from the use of multiple sequence alignments (better overall accuracy), from 'balanced training' (better prediction of β strands) and from 'structure context training' (better prediction of helix and strand lengths). The best method, cross-validated on seven different test sets purged of sequence similarity to learning sets, achieves a three-state prediction accuracy of 69.7%, significantly better than any previous method. The improved distribution and accuracy of helices and strands makes the predictions well suitable for use in practice as a first estimate of structural type of newly sequenced proteins.

Secondary structure prediction of all-helical proteins in two states

Burkhard Rost & Chris Sander

Quote: 1993 Prot. Engin. 6, 831-836

Can secondary structure prediction be improved by prediction rules that focus on a particular structural class of proteins? To help answer this question, we have assessed the accuracy of prediction for all-helical proteins, using two conceptually different methods and two levels of description. An overall two-state single residue accuracy of about 80% can be obtained by a neural network, no matter whether it is trained on two states (helix, non-helix), or first trained on three states (helix, strand, loop) and then evaluated on two states. For four test proteins, this is similar to the accuracy obtained with inductive logic programming. We conclude that on the level of secondary structure, there is no practical advantage in training on two states, especially given the added margin of error in identifying the structural class of a protein. In the further development of these methods, it is increasingly important to focus on aspects of secondary structure that aid in the construction of a correct three-dimensional model, such as the correct placement of segments.

Molecular modelling of the Norrie disease protein predicts a cysteine knot growth factor tertiary structure

Thomas Meitingner, Alfons Meindl, Peer Bork, Burkhard Rost, Chris Sander, Martina Haasemann & Jan Murken

Quote: 1993 Nat. Gen. 5, 376-380

PHD - an automatic server for protein secondary structure prediction

Burkhard Rost,

Reinhard Schneider & Chris Sander

Quote: 1994 CABIOS 10, 53-60

In the middle of 1993, more than 30,000 protein sequences are known. For 1000 of these the three-dimensional (tertiary) structure is experimentally solved. Another 7000 can be modelled by homology. For the remaining 21,000 sequences secondary structure prediction provides a rough estimate of structural features. Predictions in three states rate between 36% (random) and 88% (homology modelling) overall accuracy. Using information about evolutionary conservation as contained in multiple sequence alignments, the secondary structure of 4700 protein sequences was predicted by the automatic e-mail server PHD. For proteins with at least one known homologue, the method has an expected overall three-state accuracy of 71.4% for proteins with at least one known homologue (evaluated on 126 unique protein chains).

Redefining the goals of protein secondary structure prediction

Burkhard Rost,

Reinhard Schneider & Chris Sander

Quote: 1994 J. Mol. Biol. 235, 13-26

Secondary structure prediction recently has surpassed the 70% level of average accuracy, evaluated on the single residue states helix, strand and loop (Q_3). But the ultimate goal is reliable prediction of tertiary (3D) structure, not 100% single residue accuracy for secondary structure. A comparison of pairs of structurally homologous proteins with divergent sequences reveals that considerable variation in the position and length of secondary structure segments can be accommodated within the same 3D fold. It is therefore sufficient to predict the approximate location of helix, strand, turn and loop segments, provided they are compatible with the formation of 3D structure. Accordingly, we define here a measure of segment overlap (Sov) that is somewhat insensitive to small variations in secondary structure assignments. The new segment overlap measure ranges from an ignorance level of 37% (random protein pairs) via a current level of 72% for a prediction method based on sequence profile input to neural networks (PHD) to an average 90% level for homologous protein pairs. We conclude that the highest scores one can reasonably expect for secondary structure prediction are a single residue accuracy of $Q_3 > 85\%$ and a fractional segment overlap of $Sov > 90\%$.

Combining evolutionary information and neural networks to predict protein secondary structure

Burkhard Rost & Chris Sander

Quote: 1994 Proteins 19, 55-72

Using evolutionary information as contained in multiple sequence alignments as input to neural networks, secondary structure can be predicted at significantly increased accuracy. Here, we extend our previous three-level system of neural

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

Evolution and Neural Networks - Protein secondary structure prediction above 71% accuracy

Burkhard Rost,
Reinhard Schneider & Chris Sander

Quote: 1994 27th Hawaii International Conference on System Sciences 5, 385-394

Some 30,000 protein sequences are known. For 1,000 the structure is experimentally solved. Another 4,000 can be modeled by homology. For the remaining 25,000 sequences, the tertiary structure (3D) cannot be predicted generally from the sequence. A reduction of the problem is the projection of 3D structure onto a one-dimensional string of secondary structure assignments. Predictions in three states rate between 36% (random) and 88% (homology modelling) accuracy. Here, we present an improvement of a neural network system using information about evolutionary conservation. The method achieves a sustained overall accuracy of 71.4%. A test on 45 new proteins confirms the estimated accuracy. Of practical importance is the definition of a reliability index at each residue position: e.g. about 40% of the predicted residues have an expected accuracy of 88%. The method has been made publicly available by an automatic e-mail server.

1D secondary structure prediction through evolutionary profiles

Burkhard Rost & Chris Sander

Quote: 1994 Protein structure by distance analysis 257-276

For only about one third of the new proteins, three-dimensional (3D) structure can be predicted. For the remaining two thirds, a compromise has to be made. An extreme

simplification is the projection of 3D structure onto a string of 1D secondary structure assignments. Here, we report how neural networks can be configured such that strand is predicted significantly better, and that the prediction looks like native proteins in terms of the length of predicted segments. Using evolutionary information contained in multiple sequence alignments as input to neural networks, secondary structure can be predicted at significantly increased accuracy. Pre-processing the alignment information by using a position-specific conservation weight and the number of insertions and deletions in each alignment position is found to be advantageous. Addition of the global amino acid content yields a further improvement, mainly in predicting structural class. The final network system has a sustained overall accuracy of more than 72% evaluated on 250 sequence-unique chains. Of particular practical importance is the definition of a position-specific reliability index. For 40% of all residues the method has a sustained three-state accuracy of 88%, as high as the overall average for homology modelling.

Conservation and prediction of solvent accessibility in protein families

Burkhard Rost & Chris Sander

Quote: 1994 Proteins 20, 216-226

Predicting three-dimensional (3D) protein structure alone from sequence in general is currently an insurmountable task. As intermediate step, a much simpler task has been pursued extensively: prediction of a projection of 3D structure onto 1D strings of secondary structure. Here, we present an analysis of another 1D projection of 3D structure: the relative solvent accessibility of a residue. We show that solvent accessibility is less conserved in 3D families than secondary structure: the average correlation of relative solvent accessibility between 3D homologues is only 0.66. This value provides an effective practical upper limit for the accuracy of predicting accessibility from sequence. We introduce a neural network system that predicts relative solvent accessibility (projected onto 10 discrete states) using evolutionary profiles of amino acid substitutions derived from multiple sequence alignments. Evaluated in a cross-validation test on 126 unique proteins, the correlation between predicted and observed relative accessibility is 0.54. For a ternary (buried, intermediate, exposed) description of relative accessibility the fraction of correctly predicted residue states is about 58%. In absolute terms, this accuracy appears poor, but given the relatively low conservation of accessibility in 3D families (correlation 0.66), the network system is not far from optimal performance. Prediction is best for buried residues, e.g. 86% of the completely buried sites are correctly predicted as having 0% relative accessibility.

Structure prediction of proteins - where are we now?

Burkhard Rost & Chris Sander

Quote: 1994 Curr. Opin. Biotech. 5, 372-380

Although the structure-from-sequence prediction problem remains fundamentally unsolved, new and promising methods in 3D, 2D, and 1D have reopened the field. Pseudopotentials or information values derived from the databases can distinguish between correct and incorrect models (3D).

Interresidue contacts (2D) can be detected by the analysis of correlated mutations, albeit with low accuracy. Significantly improved prediction of secondary structure (1D) from multiple sequence alignments is now available in daily practice.

Prediction of helical transmembrane segments at 95% accuracy

Burkhard Rost,

Rita Casadio, Piero Fariselli & Chris Sander

Quote: 1995 Prot. Sci. 4, 521-533

We describe a neural network system that predicts the locations of transmembrane helices in integral membrane proteins. By using evolutionary information as input to the network system, the method significantly improved on a previously published neural network prediction method that had been based on single sequence information. The input data was derived from multiple alignments for each position in a window of 13 adjacent residues: amino acid frequency, conservation weights, number of insertions and deletions, and position of the window with respect to the ends of the protein chain. Additional input was the amino acid composition and length of the whole protein. A rigorous cross-validation test on 69 proteins with experimentally determined locations of transmembrane segments yielded an overall two-state per-residue accuracy of 95%. About 94% of all segments were predicted correctly. When applied to known globular proteins as a negative control, the network system incorrectly predicted fewer than 5% of globular proteins as having transmembrane helices. The method was applied to all 269 open reading frames from the complete yeast VIII chromosome. For 59 of these at least two transmembrane helices were predicted. Thus, the prediction is that about one fourth of all proteins from yeast VIII contain one transmembrane helix, and some 20% more than one.

Data based modeling of proteins

Liisa Holm, Burkhard Rost, Chris Sander & Gert Vriend

Quote: 1994 Statistical Mechanics, Protein Structure, and Protein Substrate Interactions 277-296

Protein Structure Prediction by Neural Networks

Burkhard Rost & Chris Sander

Quote: 1995 The handbook of brain theory and neural networks 772-775

No abstract

Fitting 1D predictions into 3D structures

Burkhard Rost

Quote: 1995 Protein folds: A distance based approach 132-151

The experimental determination of protein structure cannot keep track with the rapid generation of new sequence information. Can theory contribute? The most successful prediction method - and the only one for prediction of 3D

structure - is homology modelling. It is applicable for about one quarter of the proteins. For the rest, the prediction task has to be simplified. An extreme simplification is to project 3D structure onto 1D strings of secondary structure or solvent accessibility. For these 1D aspects of 3D structure, prediction accuracy has been improved significantly by using evolutionary information as input to neural network systems. The gain in accuracy bases on the conservation of secondary structure and relative solvent accessibility within sequence families. Secondary structure and accessibility are conserved, as well, between remote homologues. This fact can be used by fitting 1D predictions into 3D structures to detect such remote homologues. In comparison to other threading approaches, 1D threading is rather flexible. However, two factors decrease detection accuracy. First, the loss of information by projecting 3D structure onto 1D strings (in particular the loss of distances between secondary structure segments). And second, the inaccuracy of predicting 1D structure. A preliminary result is that every fifth remote homologue is detected correctly.

TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures

Burkhard Rost

Quote: 1995 Third International Conference on Intelligent Systems for Molecular Biology 314-321

Homology modelling, currently, is the only theoretical tool which can successfully predict protein 3D structure. As 3D structure is well conserved within sequence families, homology modelling allows to predict 3D structure for 20% of the SWISSPROT proteins. 20% of the proteins in are remote homologues to another PDB protein, i.e. the structures are homologous but pairwise sequence identity is not significant. Threading techniques attempt to predict such remote homologues based on sequence information to thus increase the scope of homology modelling. Here, a new threading method is presented. First, for a list of PDB proteins, 3D structure was projected onto 1D strings of secondary structure and relative solvent accessibility. Then, secondary structure and solvent accessibility were predicted by neural network systems (PHD) for a search sequence. Finally, the predicted and observed 1D strings were aligned by dynamic programming. The resulting alignment was used to detect remote 3D homologues. Four results stand out. First, even for an optimal prediction of 1D strings (taken from PDB), only about half the hits that ranked above a given threshold were correctly identified as remote homologues; only about 25% of the first hits were correct. Second, real predictions (PHD) were not much worse: about 20% of the first hits were correct. Third, a simple filtering procedure improved prediction performance to about 30% correct first hits. With such a filter, the correct hit ranked among the first three for more than 23 out of 46 cases. Fourth, the combination of the 1D threading and sequence alignments markedly improved the performance of the threading method TOPITS for some selected cases.

Progress of 1D protein structure prediction at last

Burkhard Rost & Chris Sander

Quote: 1995 Proteins 23, 295-300

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 modelling of higher dimensional aspects of protein structure.

NN which predicts protein secondary structure

Burkhard Rost

Quote: 1996 Handbook of neural computation G4.1

Currently, the prediction of three-dimensional (3D) protein structure from sequence alone poses insurmountable difficulties. As an intermediate step, a much simpler task has been pursued extensively: predicting 1D strings of secondary structure. Here, a composite neural network is described which predicts three secondary structure states (helix, strand, loop). The network system comprises two levels of feed-forward networks (one hidden layer each) and a final jury decision over differently trained networks. Training is done by an adaptive-like back-propagation. An important key features of the system is that the input is not only the sequence of one protein but the profile of a whole bunch of sequences of proteins which have the same 3D structure. The combination of the problem specific topology and the pre-processing of the input improve prediction accuracy from some 62% to 72%. Furthermore, the specific topology and training procedure successfully corrects for shortcomings of both simpler NN and classical methods. Over the last years, the system has been the best automatic predictor in a very competitive area of research

PHD: predicting 1D protein structure by profile based neural networks

Burkhard Rost

Quote: 1996 Meth. Enzymol. 266, 525-539

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 (SWISSPROT) by homology modelling based on significant sequence identity (>25%) to known 3D structures (PDB). 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, e.g., 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. 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 review I describe three prediction methods that use evolutionary information as input to neural network systems to predict secondary structure (PHDsec), relative solvent accessibility (PHDacc), and transmembrane helices (PHDhtm). I shall also illustrate 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 email prediction service (see Availability). The baseline conclusion after some 30,000 requests to the service is that 1D predictions have become accurate enough to be used as a starting point for expert-driven modelling of protein structure.

Refining neural network predictions for helical transmembrane proteins by dynamic programming

Burkhard Rost,

Piero Fariselli & Rita Casadio

Quote: 1996 Fourth International Conference on Intelligent Systems for Molecular Biology 192-200

For transmembrane proteins experimental determination of three-dimensional structure is problematic. However, membrane proteins have important impact for molecular biology in general, and for drug design in particular. Thus, prediction method are needed. Here we introduce a method that started from the output of a profile-based neural network system (PHDhtm). Instead of choosing the neural network output unit with maximal value as prediction, we implemented a dynamic programming-like refinement procedure that aimed at producing the best model for all transmembrane helices compatible with the neural network output. Preliminary results suggest that the refinement was clearly superior to the initial neural network system; and that, in terms of correctly predicting all transmembrane helices of a protein correctly, the method was more accurate than a previously applied empirical filter. The refined prediction was used successfully to predict transmembrane topology based on an empirical rule for the charge difference between extra- and intra-cellular regions (positive-inside rule). The resulting accuracy in predicting topology was better than 80%. Although a more thorough evaluation of the method on a larger data set will be required, the results compared favourably with alternative methods for the prediction of transmembrane helices and topology.

Topology prediction for helical transmembrane proteins at 86% accuracy

Burkhard Rost,

Piero Fariselli & Rita Casadio

Quote: 1996 Prot. Sci. 5, 1704-1718

Previously, we introduced a neural network system predicting the locations of transmembrane helices in integral membrane proteins based on evolutionary profiles (PHDhtm). Here, we describe an improvement and an extension of that system. The improvement is achieved by a dynamic programming-like algorithm that optimises helices compatible with the neural

network output. The extension is the prediction of topology (orientation of first loop region with respect to membrane) by applying the observation that positively charged residues are more abundant in extra-cytoplasmic regions to the refined prediction of all transmembrane helices. Furthermore, we introduce a method to reduce the number of false positives, i.e., proteins falsely predicted with membrane helices. The evaluation of prediction accuracy is based on a cross-validation and a double-blind test set (in total 131 proteins). The final method appears to be more accurate than other methods published. (1) For almost 89% ($\pm 3\%$) of the test proteins all transmembrane helices are predicted correctly. (2) For more than 86% ($\pm 3\%$) of the proteins topology is predicted correctly. (3) We define reliability indices which correlate with prediction accuracy: for the most strongly predicted half of the proteins the likelihood of predicting all transmembrane helices correctly raises to 98%; and for two-thirds of the proteins the accuracy of topology prediction was 95%. (4) The rate of proteins for which transmembrane helices are predicted falsely is below 2% ($\pm 1\%$). Finally, the method is applied to 1616 sequences of *Haemophilus influenzae*. We predict 19% of the genome sequences to contain one or more transmembrane helices. This appears to be lower than what we predicted previously for the yeast VIII chromosome (about 25%).

Pitfalls of protein sequence analysis

Burkhard Rost & Alfonso Valencia

Quote: 1996 *Curr. Opin. Biotech.* 7, 457-461

No abstract

Update on protein structure prediction: Results of the 1995 IRBM workshop

Tim Hubbard, Anna Tramontano, Geoff Barton, David Jones, Manfred Sippl, Alfonso Valencia, Arthur Lesk, John Moul, Burkhard Rost, Chris Sander, Reinhard Schneider, Armin Lahm, Raphael Leplae, Christiane Buta, Miriam Eisenstein, Ola Fjellström, Hannes Floeckner, J Guenter Grossmann, Jan Hansen, Manuela Helmer-Citterich, Flemming Steen Joergensen, Aron Marchler-Bauer, Joel Osuna, Jong Park, Astrid Reinhardt, Luis Ribas de Pouplana, Arturo Rojo-Dominguez, Vladimir Saudek, John Sinclair, Shane Sturrock, Ceslovas Venclovas and Carla Vinals

Quote: 1995 *Protein folds: A distance based approach* 132-151

Protein fold recognition by prediction-based threading

Burkhard Rost,

Reinhard Schneider & Chris Sander

Quote: 1997 *J. Mol. Biol.* 270, 471-480

In fold recognition by threading one takes the amino acid sequence of a protein and evaluates how well it fits into one of the known three-dimensional (3D) protein structures. The quality of sequence-structure fit is typically evaluated using

inter-residue potentials of mean force or other statistical parameters. Here, we present a new approach to evaluating sequence-structure fitness. Starting from the amino acid sequence we first predict secondary structure and solvent accessibility for each residue. We then thread the resulting one-dimensional (1D) profile of predicted structure assignments into each of the known 3D structures. The agreement between predicted and observed structure profile is evaluated using statistical parameters. The optimal threading for each sequence-structure pair is obtained using dynamic programming. The overall best sequence-structure pair constitutes the predicted 3D structure for the input sequence. The method is fine-tuned by adding information from direct sequence-sequence comparison and applying a series of empirical filters. Although the method relies on reduction of 3D information into 1D structure profiles, its accuracy is, surprisingly, not clearly inferior to methods based on evaluation of residue interactions in 3D. We therefore hypothesise that existing 1D-3D threading methods essentially capture not more than the fitness of an amino acid sequence for a particular 1D succession of secondary structure segments and residue solvent accessibility. The prediction-based threading method on average finds any structurally homologous region at first rank in 30% of the cases. For the 17% first hits detected at highest scores, the expected accuracy raised to 70%. However, the task to detect entire folds rather than homologous fragments, was managed much better: depending on the cut-off for what was regarded as an 'entire fold' the first hit was correct in 60-80% of all cases. The quality of the resulting 3D models depends crucially on the details of the sequence-structure alignments which can be inaccurate in detail even in cases in which the correct fold is detected.

Protein structures sustain evolutionary drift

Burkhard Rost

Quote: 1997 *Folding & Design* 2, S19-S24

A protein sequence folds into a unique three-dimensional protein structure. Different sequences, though, can fold into similar structures. How stable is a protein structure with respect to sequence changes? What percentage of the sequence are 'anchor' residues, i.e., are crucial for protein structure and function? Here, these questions are pursued by analysing large numbers of structurally homologous protein pairs. Most pairs of similar structures have sequence identity as low as expected from randomly related sequences. On average only three to four percent of all residues are 'anchor' residues (residues crucial for maintaining the structure). The symmetric shape of the distribution at low sequence identity suggests that for most structures, four billion years of evolution was sufficient to reach an equilibrium. The mean identities for convergent (different ancestor) and divergent evolution (same ancestor) of proteins to similar structures are quite close, and hence, in most cases it is difficult to distinguish between the two effects. In particular, low levels of sequence identity appear not to be indicative of convergent evolution.

Sisyphus and protein structure prediction

Burkhard Rost & Sean I. O'Donoghue

Quote: 1997 CABIOS 13, 345-356

The problem of predicting protein structure from sequence remains fundamentally unsolved despite more than three decades of intensive research effort. However, new and promising methods in 3D, 2D, and 1D prediction have reopened the field. Mean-force-potentials derived from the protein databases can distinguish between correct and incorrect models (3D). Inter-residue contacts (2D) can be detected by analysis of correlated mutations, albeit with low accuracy. Secondary structure, solvent accessibility, and transmembrane helices (1D) can be predicted with significantly improved accuracy using multiple sequence alignments. Some of these new prediction methods have proven accurate and reliable enough to be useful in genome analysis, and in experimental structure determination. Moreover, the new generation of theoretical methods is increasingly influencing experiments in molecular biology.

Learning from evolution to predict protein structure

Burkhard Rost

Quote: 1997 BCEC97: Bio-Computing and Emergent Computation 87-101

In the wake of the genome data flow, we need - more urgently than ever - accurate tools to predict protein structure. The problem of predicting protein structure from sequence remains fundamentally unsolved despite more than three decades of intensive research effort. However, the wealth of evolutionary information deposited in current databases enabled a significant improvement for methods predicting protein structure in 1D: secondary structure, transmembrane helices, and solvent accessibility. In particular, the combination of evolutionary information with neural networks proved extremely successful. The new generation of prediction methods proved to be accurate and reliable enough to be useful in genome analysis, and in experimental structure determination. Moreover, the new generation of theoretical methods is increasingly influencing experiments in molecular biology.

Midnight zone of protein structure evolution

Burkhard Rost, Sean I. O'Donoghue & Chris Sander

Quote: 1998

Today, we have a detailed and ever-widening knowledge of the evolution of DNA sequences, but what do we really know about the evolution of protein structure? Until recently, the answer was: not much. The first detailed structures were determined 26 years ago; 13 years ago, the database of atomic-resolution protein structures contained just 312 structures (PDB). Since then, due to advances in determination methods, the PDB has grown exponentially; presently it holds over 4000 entries. With this size, we can just begin to analyse the evolution of protein structure. Here, we report an analysis of all pairs of proteins in the PDB which have similar three-

dimensional (3D) structures. For each pair, we aligned the 3D structures, and measured the sequence identity (pairwise identical residues) in the aligned regions. The resulting distribution of pair identity scores shows one prominent and unexpected feature: most pairs cluster in an approximately Gaussian peak centred at 8-9% sequence identity. The distribution is surprisingly similar to that expected for 'random' pairs of completely unrelated sequences. This result has implications for our understanding of protein folding, and of the effect of convergent (different ancestor) and divergent (same ancestor) evolution on protein structure.

Surface residue composition of proteins is correlated with sub-cellular location

Miguel Andrade, Sean I. O'Donoghue & Burkhard Rost

Quote: 1998 J. Mol. Biol. 276, 517-525

One complexity in studying the general principles that determine globular protein structure is that different proteins experience different physio-chemical environments and that environment can greatly affect protein structure. Theoretical studies have tended to ignore this complexity. In this paper, we have approached this problem by grouping proteins by their sub-cellular location and looking of structural properties which are characteristic to each location. We hypothesised that, throughout evolution, each location has maintained a characteristic physio-chemical environment and hence the proteins in each location have become adapted to their environment. Hence we would expect to see differences in protein structures from different locations, particular at the surface. To test this hypothesis, we have examined all eucaryotic proteins with known three-dimensional structure and for which the sub-cellular location is known to be either nuclear, cytoplasmic, or extra-cellular. In agreement with previous studies, we find that the total amino acid composition carries a signal which identifies the location. This signal is due almost entirely to the surface residues. The surface residue signal was so strong we were able to accurately predict the location of proteins, given only a knowledge of which residues are at the surface. The result supports our hypothesis that protein structures show characteristic adaptation to their environment. The results suggest how the accuracy of prediction of location from sequence can be improved. These results also suggest several principles that proteins may use in adapting to particular physio-chemical environments; these may be useful for protein design.

Better 1D predictions by experts with machines

Burkhard Rost

Quote: 1997 Proteins Suppl. 1, 192-197

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 semi-automatic 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 towards predicting higher dimensions of protein structure.

Protein structure prediction in 1D, 2D, and 3D

Burkhard Rost

Quote: 1998 The encyclopaedia of computational chemistry 3, 2242-2255

No Abstract.

Pedestrian guide to analysing sequence database

Burkhard Rost & Reinhard Schneider

Quote: 1999 Core techniques in biochemistry in press

Over the past few years our means of communication have changed rapidly due to the growth of the World Wide Web (WWW). The Web enables molecular biologists to immediately access databases, scan literature, find information about related research and researchers, and to trace cell cultures. Wet-lab biologists can uncover information about the protein of interest without having to become experts in sequence analysis. Here, we present a variety of tools; provide an overview of the state-of-the art in sequence analysis; and described some of the principles of the methods.

Twilight zone of protein sequence alignments

Burkhard Rost

Quote: 1999 Prot. Engin. 12, 85-94

Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20-35% sequence identity. Here, I analysed more than a million sequence alignments between protein pairs of known structures to re-define a line distinguishing between true and false positives for low levels of similarity. Four results stood out. (1) The transition from the safe zone of sequence alignment into the twilight zone is described by an explosion of false negatives. More than 95% of all pairs detected in the twilight zone had different structures. More precisely, above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were. (2) Whether or not sequence homology implied structural identity depended crucially on the alignment length. For example, if ten residues were similar in an alignment of length 16 (> 60%), structural similarity could not be inferred. (3) The 'more similar than identical' rule (discarding all pairs for which percentage similarity was lower than percentage identity) reduced false positives significantly. (4) Using intermediate sequences for finding links between more distant families was almost as successful: pairs were predicted to be homologous when the respective sequence families had proteins in common. All findings are applicable to automatic database searches.

Neural networks for protein structure prediction: hype or hit?

Burkhard Rost

Quote: 1999 TICS

Neural networks have been applied to many pattern classification problems. Here, I review applications to the problem of predicting protein structure from protein sequence. Initially, many methods were apparently designed by researchers who just wanted a real-life application for their gadget. However, the competitiveness of the field separated the wheat from the chaff. Meanwhile, several neural network-based methods have contributed significantly to advancing the field of bio-informatics, and some are clearly influencing molecular biology.

Marrying structure and genomics

Burkhard Rost

Quote: 1998 Structure 6, 259-263

Today. Large-scale genome sequencing is filling up the catalogue of natural proteins at a breath-taking speed. Today, we have available not just a large number of sequences, but also glimpses of the inventory of entire organisms. This will soon improve our understanding of cells, in particular, and of life, in general. Three means will contribute: (1) sequencing genomes (genomics), (2) determining protein structures, and (3) determining protein function. Protein structure is interwoven with function. Sequencing and determining function are also routinely combined. However, what about the relation between structure determination and genomics?

Tomorrow. Structural genomics, the marriage between protein structure determination and genomics, is already beginning. Here, I attempted to illustrate the likely direction this marriage will take. Structure determination will be pushed by, and profit from genomics. Basing research and technical developments (such as drug design) on all three pillars (sequence, structure, function) will be a big step toward understanding of life.

Objectives. Structure determination will benefit from genomics in two ways (Fig. 1). (1) The mass of available sequences will facilitate quick determination of structure for most existing folds. (2) Sequences for entire organisms will help to unravel missing links in functional pathways, to explore alternative pathways, and to widen our understanding of principle mechanisms and of evolutionary cross-links.

Short yeast ORFs: expressed protein or not?

Burkhard Rost

Quote: 1999

Sequencing the entire genome of *Saccharomyces cerevisiae* (yeast) revealed about 2500 ORFs with less than 100 residues. Most of these supposedly do not correspond to expressed proteins. However, some do. How could theory help separating the wheat from the chaff? Here, I introduced a simple measure for the 'globularity' of a protein. I used this measure to

develop a novel method that firstly predicted the globularity, and secondly compared the predicted globularity to the database background. The difference between these two values provided an indication for how likely a sequence would adopt a typical globular protein structure. On average, globular domains differed from randomly chosen fragments of these domains and - to some extent - from native protein chains extending over the core domain. Thus, the method might be useful for predicting domains from sequence. Analysing a set

of 2427 short yeast ORFs, I first predicted membrane helices. The results indicated that most short ORFs with putative membrane helices would not correspond to expressed proteins. Assuming that ORFs are more likely expressed if similar to typical globular proteins, I sorted all short yeast ORFs according to their predicted globularity. About half of the non-membrane ORFs resembled globular domains.