

Structure prediction of proteins—where are we now?

Burkhard Rost and Chris Sander

European Molecular Biology Laboratory, Heidelberg, Germany

Although the 'structure from sequence' prediction problem remains fundamentally unsolved, new and promising methods in one, two and three dimensions have reopened the field. Significantly improved one-dimensional prediction of secondary structure from multiple sequence alignments is now in routine use. In the two-dimensional approach, inter-residue contacts can be detected by analysis of correlated mutations, albeit with low accuracy. Finally, three-dimensional methods, in which pseudopotentials or information values are derived from the databases, are proving their value for distinguishing between correct and incorrect models.

Current Opinoin in Biotechnology 1994, 5:372–380

Introduction

Suppose one has a protein sequence of unknown structure termed 'SOS'. What can be learned about SOS before beginning an experiment? Data banks of protein sequences and structures are growing rapidly [1,2] as a result of large-scale sequencing projects [3] and improvements in experimental determination of three-dimensional structure [4]. Can we profit from the information flood? Does the data bank teach us how to predict the three-dimensional structure of SOS?

The most successful tool for prediction of three-dimensional structure is homology modelling. An approximate three-dimensional model (which has a correct fold, but inaccurate loop regions) can be constructed if SOS has significant similarity to a protein of known structure, evaluated in terms of sequence similarity (i.e. alignment) or sequence-structure fitness (i.e. threading) (for reviews, see May and Blundell, this issue (pp 355–360) and [5]). Homology modelling effectively raises the number of 'known' three-dimensional structures from about 1500 to 7500 [6,7]. But what if SOS has no homologue of known three-dimensional structure? Can three-dimensional structure be predicted directly from sequence?

Without detectable homology, we are still forced to resort to simplifications of the prediction problem. In the process, we can make use of the rich diversity of information in current data banks. For this review, we have selected generic methods for prediction at three different levels of simplification (see Fig. 1), namely one, two and three dimensions. Prediction in one dimension (i.e. secondary structure) can be improved significantly through the use of evolutionary information. Prediction in two dimensions (i.e. inter-residue contacts) can also, to a certain extent, profit from evolutionary information, but so far, is of only limited accuracy. Lastly, incorrect three-dimensional structures can now be detected with remarkable accuracy. In the following sections, we consider each of these prediction methods in turn.

One-dimensional approaches: predictions are successful, but of limited use

Single sequences are the dead end of structure prediction

An extreme simplification of the prediction problem is to project three-dimensional structure onto one-dimensional strings of secondary structure (see Fig. 1). If a sequence codes for the entire complexity of a three-dimensional structure [8,9], some simple sequence motifs might determine secondary structure. For many years, however, little improvement has been made to prediction accuracy [10–12,13*]. Not even the most sophisticated algorithms [14–23] have been able to overcome the principal problem: it is exceedingly difficult to extract from single sequences enough information for accurate prediction. How can we extract more information from sequence databases?

Evolution distinguishes signal from noise

At the level of protein molecules, selective pressure results from the need to maintain function, which in turn requires maintenance of the specific three-dimensional structure consistent with that function [24,25,26*,27,28]. Accordingly, conservation and mutation patterns observed in alignments contain very specific information about three-dimensional structure. How much variation is tolerated? Two naturally evolved proteins with more than 25% identical residues (length >80 residues) are extremely likely to be similar in three-dimensional structure [6]. Even so, structure may be conserved in spite of much higher divergence [28,29,30*–32*,33–35,36*,37–39]. Do we have enough data to detect structure-specific sequence motifs [40] and to correctly align very remote homologues?

Multiple alignments improve as data banks grow

When sequence similarity is sufficient, alignment procedures are (more or less) straightforward [6,41–43]. For less similar protein sequences, however, alignments may

Evolution is the key to improvement of secondary structure predictions

That evolutionary information can improve predictions was established long ago [44,45]. Recently, isolated reports of improvements in prediction have revived interest in this area [46–52]. Even so, prediction accuracy varies significantly between proteins (Fig. 2). So, does the improvement hold up when evaluated on a large data set? Indeed, simple methods based on multiple alignment information yield better predictions [53,54,55*]. By stepwise incorporation of more evolutionary information, prediction accuracy can be pushed above 72% accuracy (percentage of residues correctly predicted in either helix, strand or other conformation) [56*–58*].

Is secondary structure prediction more useful now?

How can we assess whether a level of 72% prediction accuracy is good? It is certainly reasonably good compared with the prediction of secondary structure by homology modelling [57*,59*,60*]. In addition, it should be borne in mind that some residues within a structure are predicted at higher levels of accuracy than the mean value (Fig. 2).

Is prediction accurate enough to be of practical use? A number of reasons argue that it is. First, a considerable amount of interest is currently being shown in predictions (about 1000 requests per month to an electronic mail server [61,62]). Second, some biologists find predictions very helpful for checking hypotheses about, for example, relations between proteins or putative sites for site-directed mutations. Third, in exceptional cases, predictions might initiate reasonable first guesses of three-dimensional structure [63]. Fourth, predictions can be used as a seed for methods predicting, for example, contacts between secondary structure segments [64]. At this

stage, it is useful to assess whether successful application of evolutionary information to prediction in one dimension can be generalized to two dimensions.

Two-dimensional approaches: predictions useful, but of limited accuracy

Predicting contacts is a difficult task

From the knowledge of all inter-residue contacts or distances (Fig. 1b) one can, in principle, model a three-dimensional structure using distance geometry methods [65–69]. Two questions surround such methods: first, can contacts be predicted accurately enough; and second, are all important contacts predicted? A trade-off occurs between the Scylla of predicting enough contacts and the Charibdis of predicting only correct ones (see Fig. 3). Can evolutionary information help out once again?

Correlated mutations might imply spatial proximity

In sequence alignments, some pairs of positions appear to co-vary in a physico-chemically plausible manner (i.e. a 'loss of function' point mutation is often rescued by an additional mutation that compensates for the change [70,71]). One hypothesis is that compensation would be most effective in maintaining a structural motif if the mutated residues were spatial neighbours. Recently, attempts have been made to quantify such a hypothesis [72*,73*] and to use it for contact predictions [74*,75*].

In myoglobins, a correlation is detectable between fluctuations in neighbouring charges, but no significant correlation for side-chain volume is evident [72*] (the latter can largely be explained by the density of packing [76*,77,78*]). Is the signal from correlated mutations

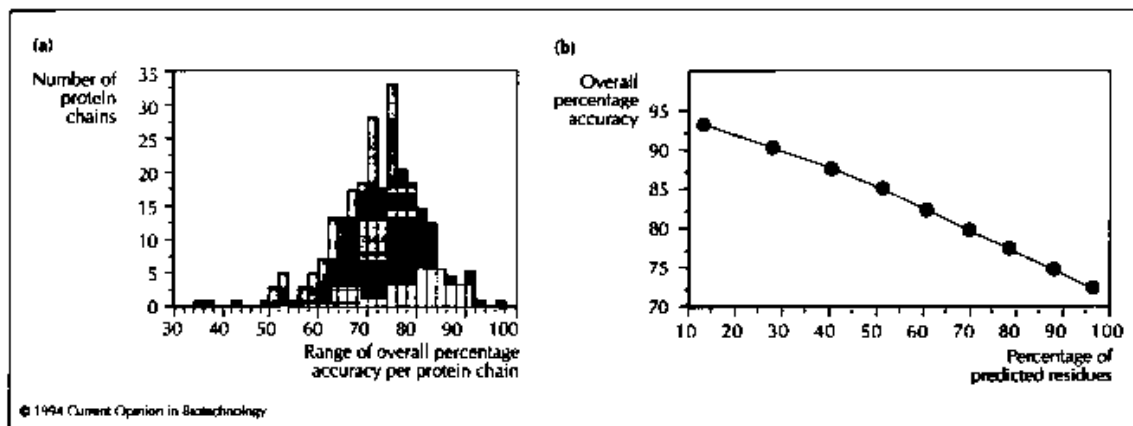


Fig. 2. Secondary structure prediction accuracy for a system of neural networks using evolutionary information evaluated on 250 protein families. (a) Prediction accuracy varies considerably among different protein families. One standard deviation is nine percentage points, so prediction accuracy for most sequences is 63–81% and the average accuracy is 72%. Because of this significant variation, prediction methods have to be evaluated on a sufficiently large set of unique proteins. (b) Residues with a higher reliability index are predicted with higher accuracy. For example, for some 40% of all residues, prediction accuracy is, on average, 88%. In practice, it is recommended that attention be focused on the most reliably predicted residues.

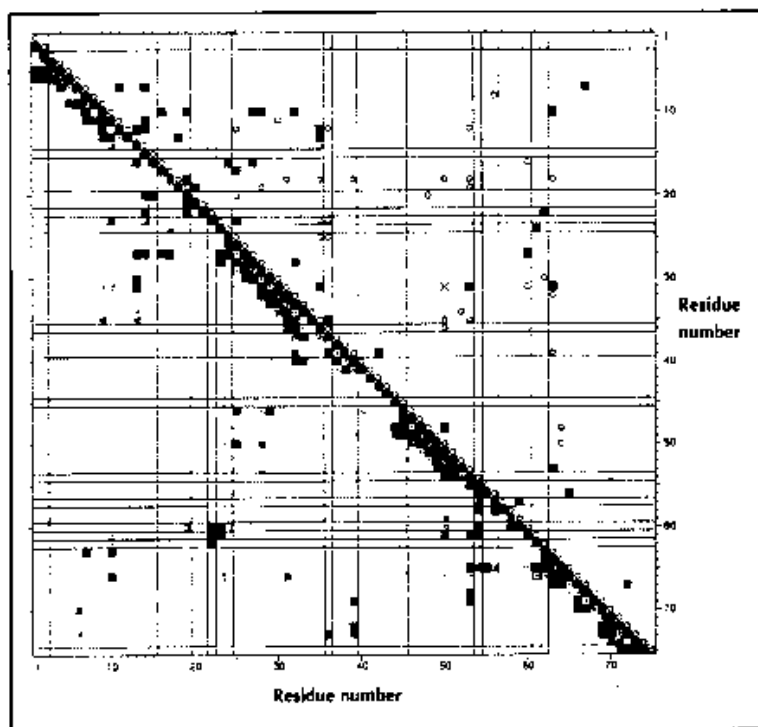


Fig. 3. Contact map for the intestine calcium-binding protein (3ICB). The map predicted on the basis of correlated mutations is shown in the upper-right portion of the figure [75*]. The map based on the observed crystal structure is shown in the lower left portion for comparison. Thirty five correctly predicted contacts (filled squares) and 34 incorrectly predicted contacts (open circles) are shown, which is a ratio of ~1:1 for the first 70 predicted contacts. The ratio depends on the threshold used to assign a mutation: higher thresholds imply a greater number of correctly predicted contacts at the expense of a smaller total number of predicted contacts. Figure provided by courtesy of Alfonso Valencia using CONAN [109].

stronger than the signal from a simple measure of conservation for single residues? One claim is that samples of residue pairs selected by evaluation of compensated changes are not as informative as those selected by a simple measure of conservation alone [73*]. So, how far can correlated mutations be used to predict contacts?

Applying a stringent significance cut-off in the prediction of contacts, a small number of residue contacts can be predicted with reasonable accuracy [74*]. Predictions of contacts based on correlated mutations are between 1.4 and 5.1 times better than random predictions [75*]. The current consensus appears to be that correlated mutations might provide sufficient information to distinguish between alternative models of three-dimensional structure, but not enough information to predict conformations *ab initio* (Fig. 3), unless additional information is acquired.

Can predictions based on correlated mutations be improved?

Why is the incorporation of evolutionary information not more successful for contact (two-dimensional) prediction? One reason is that functional and structural constraints differ in their effect on sequences [75*]. Furthermore, the dynamics of protein folding and certain aspects of protein stability may result in a correlated mutation for a residue pair separated in the final three-dimensional structure [74*]. Improvements will depend on how these effects are treated. For some purposes, however,

it might be quite useful to predict just a few contacts correctly [79,80]. One ambitious proposal is to generate much more information about structurally constrained correlated mutations by experiments that combine sequence randomization with a selection system, with the explicit goal of determining three-dimensional structure [74*].

Given that we cannot yet, in general, predict three-dimensional structure from sequence, can we at least ascertain whether or not any particular three-dimensional model for a given sequence represents the (correct) native structure?

Three-dimensional approaches: identification of native-like structures

Potentials of mean force pass the recognize-self test

We are still far from simulating protein folding, despite all the recent improvements [81*,82–84]. Even the recognize-self test [85] is difficult to pass: deliberately misfold a structure, and compare the misfolded with the native structure; if the native structure stands out as being clearly better, the recognize-self test has been successfully passed. A new boom has occurred in prediction methods (reviewed in [86*]) that have as their basis pseudo-energy functions. For example, the approach of Sippl [87] is based on the notion that physically meaningful potentials (potentials of mean force) can be derived from the set of known three-dimensional structures [88**]. The recognize-self test is successfully passed by this and sim-

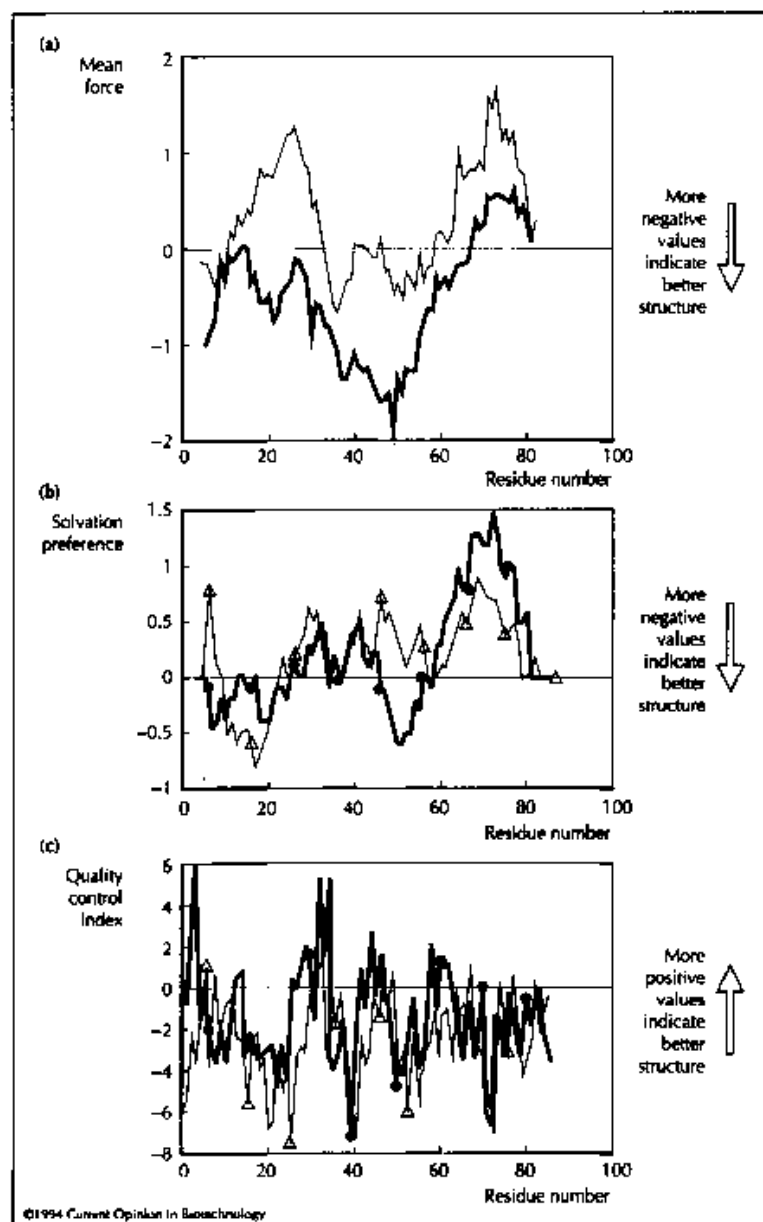


Fig. 4. Discrimination between native-like three-dimensional structures and incorrect model structures. Three methods plotting sequence against potential/pseudopotential are compared for the two structures of the DNA-binding gene V protein (Protein Data Bank datasets 2GNS and 1BGH). (a) Distance-based potential of mean force (where more negative values indicate a better structure). For 2GNS and 1BGH, the average over all residues was -2.53 and -4.95 , respectively [87,94**]. (b) Atomic solvation preferences (where smaller values indicate a better structure). For 2GNS and 1BGH, the average over all residues was 22.2 and 16.3 , respectively [91]. (c) Contact-based quality control index (where more positive values indicate a better structure). For 2GNS and 1BGH, the average over all residues was -2.75 and -1.43 , respectively [93*]. All three methods identify 2GNS as containing many errors. 1BGH may be inaccurate towards the carboxyl terminus. A similar qualitative result is found using a method [92] that relies on various structural features (backbone dihedral angles, bond lengths, planarity of rings and hydrogen-bonding patterns). On the basis of such differences, structures deposited in the data bank were predicted to contain errors [92,93*,94**].

ilar methods [89]. Can the method detect the better of two X-ray structures?

Errors in three-dimensional structure can be recognized

A major problem in determining three-dimensional structure is the quality of the atomic model derived from the interpretation of experimental data. Some methods can aid in modelling by evaluating the quality of a given three-dimensional structure [90–92,93*,94**]. The common idea in these approaches is to use the data bank to derive a criterion (or pseudo-energy) that describes a native fold. The methods differ in the choice of pseudo-

energy: distances [87,92], hydrophobicity [95], contacts [93*], or solvation [91]. The pseudo-energy of a native structure has a typical functional shape along the sequence that is characteristically different from that of a wrong model (Fig. 4).

An obvious idea is to use such pseudo-energies for threading (i.e. to fit a sequence into a three-dimensional structure) [96]. But, despite considerable improvements [26*,35,88**,97*,98,99,100*,101,102*,103*,104–106], the main problem of distinguishing between incorrect and correct remote homologues in a database search remains unsolved.

Is discrimination of native-like structure of practical use?

A number of reasons attest to the utility of pseudopotentials or information values in protein structural research. First, their ability to identify specific residues that may have an incorrect NMR or X-ray refinement model can facilitate experimental determination of three-dimensional structure. Second, the detection of a correct model amongst various alternatives has an important impact on structure prediction. In some cases, three-dimensional prediction can be simplified by external constraints (e.g. for membrane proteins and coiled-coil proteins) such that three-dimensional models can be generated. Mean-force potentials can be used, in these instances, as an additional filter for selecting the correct model from among the remaining alternatives ([79,80,107*]; O'Donoghue, unpublished data).

Conclusions

The rich information contained in the database of known protein structures can be used to ascertain whether a structural model is correct or incorrect and also, in favorable cases, whether a newly determined protein sequence fits one of the commonly occurring folds. If the detection of structural homology by such a threading procedure is successful, one can build a fairly accurate three-dimensional model. Even so, a method for the direct and general prediction of three-dimensional structure from sequence, without homology considerations, is not yet available. Current claims to the contrary, even for simple folds, should be viewed with some scepticism.

Nevertheless, some aspects of protein structure can currently be predicted with increasing accuracy, particularly when databases are tapped for the rich information they contain (sequences have been adapted over eons of evolutionary history). For instance, evolutionary information significantly improves the accuracy of secondary-structure prediction, which is now at 72±8% three-state accuracy for sequence families. In contrast, the prediction of inter-residue contacts based on correlated mutations is, so far, of very limited accuracy.

In short, even if three-dimensional structure cannot be predicted correctly from scratch, some attributes of structure can be predicted with increasing accuracy, especially if multiple-sequence information is available.

Acknowledgements

We thank the following: Séan O'Donoghue, Reinhard Schneider and Manfred Sippl for discussions and help; Christos Ousounis, Alfonso Valencia, Liisa Holm, Gerrit Vriend, Manfred Sippl and Maria Ortner for contributing data and figures; Janet Thornton and Roman Laskowski for helpful remarks.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- * of special interest
 - ** of outstanding interest
1. Bairoch A, Boeckmann B: The SWISS-PROT Protein Sequence Data Bank. *Nucleic Acids Res* 1992, 20:2019-2022.
 2. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: The Protein Data Bank: a Computer Based Archival File for Macromolecular Structures. *J Mol Biol* 1977, 112:535-542.
 3. Oliver S, van der Aart QJ, Agostini-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine C, Anwar R, Ballesta JP, Benit P, et al.: The Complete DNA Sequence of Yeast Chromosome III. *Nature* 1992, 357:38-46.
 4. Lattman EE: Protein Crystallography for All. *Proteins* 1994, 18:103-106.
 5. Sali A, Overington JP, Johnson MS, Blundell TL: From Comparisons of Protein Sequences and Structures to Protein Modelling and Design. *Trends Biochem Sci* 1990, 15:235-240.
 6. Schneider R, Sander C: Database of Homology-Derived Structures and the Structural Meaning of Sequence Alignment. *Proteins* 1991, 9:56-66.
 7. Sander C, Schneider R: The HSSP Data Base of Protein Structure-Sequence Alignment. *Nucleic Acids Res* 1993, 21:3105-3109.
 8. Anfinsen CB: Principles that Govern the Folding of Protein Chains. *Science* 1973, 181:223-230.
 9. Hartl F-U, Hlodan R, Langer T: Molecular Chaperones in Protein Folding: The Art of Avoiding Sticky Situations. *Trends Biochem Sci* 1994, 19:20-25.
 10. Sternberg MJE: Secondary Structure Prediction. *Curr Opin Struct Biol* 1992, 2:237-241.
 11. Rackovsky S: On the Nature of the Protein Folding Code. *Proc Natl Acad Sci USA* 1993, 90:644-648.
 12. Rao S, Zhu Q-L, Vajda S, Smith T: The Local Information Content of the Protein Structural Database. *FEBS Lett* 1993, 322:143-146.
 13. Rost B, Sander C, Schneider R: Progress in Protein Structure Prediction? *Trends Biochem Sci* 1993, 18:120-123.
 - * Some standards for testing secondary structure prediction methods are reviewed and used to evaluate current prediction methods.
 14. Geourjon C, Deléage G: SOPM: A Self-Optimized Method for Protein Secondary Structure Prediction. *Protein Eng* 1994, 7:157-164.
 15. Metfessel BA, Saurugger PN, Connelly DP, Rich SS: Cross-Validation of Protein Structural Class Prediction using Statistical Clustering and Neural Networks. *Protein Sci* 1993, 2:1171-1182.
 16. Muggleton S, King RD, Sternberg MJE: Protein Secondary Structure Prediction using Logic-Based Machine Learning. *Protein Eng* 1992, 5:647-657.
 17. Presnell SR, Cohen FE: Artificial Neural Networks for Pattern Recognition in Biochemical Sequences. *Annu Rev Biophys Biomol Struct* 1993, 22:283-298.
 18. Yi T-M, Lander ES: Protein Secondary Structure Prediction using Nearest-Neighbor Methods. *J Mol Biol* 1993, 232:1117-1129.
 19. Asai K, Hayamizu S, Handa K: Prediction of Protein Secondary Structure by the Hidden Markov Model. *Comput Appl Biosci* 1993, 9:141-146.
 20. Fariselli P, Compiani M, Casadio R: Predicting Secondary Structures of Membrane Proteins with Neural Networks. *Eur Biophys J* 1993, 22:41-51.
 21. Maclin R, Shavlik JW: Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou-Fasman Algorithm for Protein Folding. *Machine Learn* 1993, 11:195-215.
 22. Sasagawa F, Tajima K: Prediction of Protein Secondary Structures by a Neural Network. *Comput Appl Biosci* 1993, 9:147-152.

23. Munson PJ, 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*, Edited by Hunter L. Wailea, Hawaii, USA: IEEE Computer Society Press; 1994:375-384.
24. Farber GK, Petsko GA: The Evolution of α/β Barrel Enzymes. *Trends Biochem Sci* 1990, 15:228-234.
25. Pastore A, Lesk AM: Comparison of the Structure of Globins and Phycocyanins: Evidence for Evolutionary Relationship. *Proteins* 1990, 4:133-155.
26. Ouzounis C, Sander C, Scharf M, Schneider R: Prediction of Protein Structure by Evaluation of Sequence-Structure Fitness: Aligning Sequences to Contact Profiles Derived from 3D Structures. *J Mol Biol* 1993, 232:805-825.
- Single-residue contact environments are used to describe three-dimensional structure. Statistical preference parameters of residue types for these environments are then used to evaluate three-dimensional models generated either by cyclic permutation or by a database alignment search. Although the recognize-self test works well, remote homologues cannot be reliably detected. Results can be improved through the use of conservation weights derived from multiple sequence alignments.
27. Musacchio A, Gibson T, Rice P, Thompson J, Saraste M: The PH Domain: A Common Piece in the Structural Patchwork of Signalling Proteins. *Trends Biochem Sci* 1993, 18:343-348.
28. Doolittle RF: Convergent Evolution: The Need to be Explicit. *Trends Biochem Sci* 1994, 19:15-18.
29. Bordo D: ENVIRON: A Software Package to Compare Protein Three-Dimensional Structures with Homologous Sequences using Local Structural Motifs. *Comput Appl Biosci* 1993, 9:639-645.
30. Vingron M, Waterman MS: Sequence Alignment and Penalty Choice. *J Mol Biol* 1994, 235:1-12.
- Alignment algorithms depend on the setting of various parameters, most noticeably gap penalties. Two strategies are presented that allow the detection of biologically good alignments: first, a method to delineate efficiently all optimal alignments arising under all choices of parameters; and second, a method to study the statistical behaviour of optimal alignment scores. The analysis is illustrated using two immunoglobulin sequences.
31. Henikoff S, Henikoff JG: Performance Evaluation of Amino Acid Substitution Matrices. *Proteins* 1993, 17:49-61.
- Several types of amino acid substitution matrix are evaluated. Matrices derived directly from sequence-based or structure-based alignments of distantly related proteins perform much better than extrapolated matrices based on the Dayhoff evolutionary model. Appropriate selection of the substitution matrix is a simple and effective way of improving alignment methods.
32. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: Hidden Markov Models in Computational Biology: Applications to Protein Modelling. *J Mol Biol* 1994, 235:1501-1531.
- A fresh approach to the problem of multiple sequence alignment that represents a meeting point between machine learning and molecular biology. The method relies on bootstrapping 'profile' information as the alignment is refined.
33. Haussler D, Krogh A, Mian IS, Sjolander K: Protein Modeling using Hidden Markov Models: Analysis of Globins. In *Proceedings for the 26th Hawaii International Conference on Systems Sciences*, Edited by Hunter L. Wailea, Hawaii, USA: IEEE Computer Society Press; 1993:792-802.
34. Altschul SF: A Protein Alignment Scoring System Sensitive at All Evolutionary Distances. *J Mol Biol* 1993, 232:290-300.
35. Johnson MS, Overington JP, Blundell TL: Alignment and Searching for Common Protein Folds Using a Data Bank of Structural Templates. *J Mol Biol* 1993, 231:735-752.
36. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 1993, 262:208-214.
- Local residue patterns are used to improve multiple sequence alignments. An optimized local alignment model is determined by an iterative procedure.
37. Livingstone CD, Barton GJ: Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation. *Comput Appl Biosci* 1993, 9:745-756.
38. Heringa J, Argos P: A Method to Recognize Distant Repeats in Protein Sequences. *Proteins* 1993, 17:391-411.
39. Thompson JD, Higgins DC, Gibson TJ: Improved Sensitivity of Profile Searches through the use of Sequence Weights and Gap Excision. *Comput Appl Biosci* 1994, 10:19-29.
40. Rooman M, Wodak SJ: Identification of Predictive Sequence Motifs Limited by Protein Structure Data Base Size. *Nature* 1988, 335:45-49.
41. Schneider G, Wrede P: Development of Artificial Neural Filters for Pattern Recognition in Protein Sequences. *J Mol Biol* 1993, 236:586-595.
42. Jones DT, Taylor WR, Thornton JM: The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Comput Appl Biosci* 1992, 8:275-282.
43. Flores TP, Orengo CA, Moss DS, Thornton JM: Comparison of Conformational Characteristics in Structurally Similar Protein Pairs. *Protein Sci* 1993, 2:1811-1826.
44. Maxfield FR, Scheraga HA: Improvements in the Prediction of Protein Topography by Reduction of Statistical Errors. *Biochemistry* 1979, 18:697-704.
45. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE: Prediction of Protein Secondary Structure and Active Sites using Alignment of Homologous Sequences. *J Mol Biol* 1987, 195:957-961.
46. Benner SA, Cohen MA, Gerloff D: Predicted Secondary Structure for the Src Homology 3 Domain. *J Mol Biol* 1993, 229:295-305.
47. Benner SA, Badcoe I, Cohen MA, Gerloff DL: *Bona Fide* Prediction of Aspects of Protein Conformation. *J Mol Biol* 1994, 235:926-958.
48. Barton GJ, Russell RB: Protein Structure Prediction. *Nature* 1993, 361:505-506.
49. Bozcon PE, Barton GJ, Richards WG: Secondary Structure Prediction for Modelling by Homology. *Protein Eng* 1993, 6:261-266.
50. Gerloff DL, Jenny TF, Knecht LJ, Gonnet GH, Benner SA: The Nitrogenase Mofe Protein. *FEBS Lett* 1993, 318:118-124.
51. Gibson TJ, Thompson JD, Abagyan RA: Proposed Structure for the DNA-Binding Domain of the Helix-Loop-Helix Family of Eukaryotic Gene Regulatory Proteins. *Protein Eng* 1993, 6:41-50.
52. Livingstone CD, Barton GJ: Secondary Structure Prediction from Multiple Sequence Data: Blood Clotting Factor XIII and *Yersinia* Protein-Tyrosine Phosphatase. *Int J Pept Protein Res* 1994, in press.
53. Rost B, Sander C: Exercising Multi-Layered Networks on Protein Secondary Structure. In *Neural Networks: From Biology to High Energy Physics*, Edited by Benhar O, Brunak S, DeGiudice P, Grandolfo M, Elba, Italy: International Journal of Neural Systems; 1992:209-220.
54. Levin JM, Pascarella S, Argos P, Garnier J: Quantification of Secondary Structure Prediction Improvement using Multiple Alignments. *Protein Eng* 1993, 6:849-854.
55. Rost B, Sander C: Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks. *Proc Natl Acad Sci USA* 1993, 90:7558-7562.
- A system of neural networks is used to improve the prediction of β -strands, native-like lengths of secondary structure segments and overall accuracy. Prediction accuracy is significantly improved by using residue-substitution patterns from multiple alignments as input to the neural network system.
56. Rost B, Sander C: Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J Mol Biol* 1993, 232:584-599.
- The system of neural networks in [55] is described in detail. Further improvement of the method stem is achieved by using conservation weights from multiple alignments for input. Various measures for prediction accuracy are defined.
57. Rost B, Sander C, Schneider R: Redefining the Goals of Protein Secondary Structure Prediction. *J Mol Biol* 1994, 235:13-26.
- For a sequence family, the goal of secondary structure prediction is not 100% accuracy on a per-residue level. The comparison of (three-dimensional homologues carried out in this study suggests, instead, that perfect (in three-dimensional topography) predictions for a sequence family aver-

age at 88% three-state accuracy. An additional measure for the quality of secondary structure predictions is based on secondary structure segments rather than on single residues. For such a segment-based measure, the optimum attainable in a prediction for a sequence family is about 90%.

58. Rost B, Sander C: **Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure.** *Proteins* 1994, 19:55-72.

The system of neural networks described in [56*] is improved by using additional evolutionary information (number of insertions and deletions) and the global information of amino acid composition as input. Evaluated on a data set of 250 unique protein chains, the method yields an overall cross-validated three-state accuracy of more than 72%. The system is compared with various alternative prediction methods.

59. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Morion J-P: **Comparison of Three Algorithms for the Assignment of Secondary Structure in Proteins: The Advantages of a Consensus Assignment.** *Protein Eng* 1993, 6:377-382.

Automatic assignments of secondary structure on the basis of three-dimensional structure are not unique. The assignment depends not only on the crystallographic resolution, but also on the particular method used for assignment. In this study, three methods are compared. They assign the same secondary structure for only 63% of all residues in a data set of 154 unique proteins. A consensus assignment is also proposed.

60. Russell RB, Barton GJ: **The Limits of Protein Secondary Structure Prediction Accuracy from Multiple Sequence Alignment.** *J Mol Biol* 1993, 234:951-957.

Secondary structure assignments vary within three-dimensional families. The variation differs for different families. A measure is proposed that estimates the expected variation for a specific family on the basis of conservation of sequence within the family. Some recent predictions of secondary structure fall within the expected range of perfect prediction.

61. Rost B, Sander C, Schneider R: **PHD—An Automatic Server for Protein Secondary Structure Prediction.** *Comput Appl Biosci* 1994, 10:53-60.

62. Rost B, Sander C, Schneider R: **Evolution and Neural Networks—Protein Secondary Structure Prediction Above 71% Accuracy.** In *27th Hawaii International Conference on System Sciences*, Edited by Hunter L, Wailea, Hawaii, USA: IEEE Society Press; 1994:385-394.

63. Meitinger T, Meindl A, Bork P, Rost B, Sander C, Haasemann M, Mueken I: **Molecular Modelling of the Norrie Disease Protein Predicts a Cysteine Knot Growth Factor Tertiary Structure.** *Nature Genet* 1993, 5:376-380.

64. Hubbard TJP: **Use of β -Strand Interaction Pseudo-Potential in Protein Structure Prediction and Modelling.** In *27th Hawaii International Conference on System Sciences*, Edited by Hunter L, Wailea, Hawaii, USA: IEEE Society Press; 1994:336-344.

65. Brünger AT, Nilges M: **Computational Challenges for Macromolecular Structure Determination by X-Ray Crystallography and Solution NMR-Spectroscopy.** *Q Rev Biophys* 1993, 26:49-125.

66. Bohr J, Bohr H, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B, Petersen SB: **Protein Structures from Distance Inequalities.** *J Mol Biol* 1993, 231:861-869.

67. Saitoh S, Nakai T, Nishikawa K: **A Geometrical Constraint Approach for Reproducing the Native Backbone Conformation of a Protein.** *Proteins* 1993, 15:191-204.

68. Galaktionov SC, Marshall GR: **Properties of Intra globular Contacts in Proteins: An Approach to Prediction of Tertiary Structure.** In *27th Hawaii International Conference on System Sciences*, Edited by Hunter L, Wailea, Hawaii, USA: IEEE Society Press; 1994:326-335.

69. Taylor WR: **Protein Fold Refinement: Building Models from Idealized Folds using Motif Constraints and Multiple Sequence Data.** *Protein Eng* 1993, 6:593-604.

70. Aitschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of Co-ordinated Amino Acid Substitutions with Function in Viruses Related to Tobacco Mosaic Virus.** *J Mol Biol* 1987, 193:693-707.

71. Aitschuh D, Vernet T, Moras D, Nagai K: **Coordinated Amino Acid Changes in Homologous Protein Families.** *Protein Eng* 1988, 2:193-199.

72. Neher E: **How Frequent are Correlated Changes in Families of Protein Sequences?** *Proc Natl Acad Sci USA* 1994, 91:98-102.

A statistical theory is presented that allows evaluation of correlations in a family of aligned protein sequences. One protein family, the myoglobins, is analyzed. This author finds a high correlation between fluctuations in neighbouring charges and sequence, but finds no such correlation for side-chain volume.

73. Taylor WR, Hatrick K: **Compensating Changes in Protein Multiple Sequence Alignments.** *Protein Eng* 1994, 7:341-348.

A method is described that identifies compensating changes between residues at positions in a multiple alignment. The correlation of compensating changes among sequences is compared with the correlation of sequence conservation among sequences. The two correlations are only marginally different, leading to the selection of closer pairs by a compensation measure. The compensation measure is found to be not as good as a simpler measure based on conservation alone.

74. Shindyalov IN, Kolchanov NA, Sander C: **Can Three-Dimensional Contacts in Protein Structures be Predicted by Analysis of Correlated Mutations?** *Protein Eng* 1994, 7:349-358.

Correlated mutations are detected on the basis of reconstruction of a phylogenetic tree and analysis of mutations in the branches of this tree. A significant, but weak, tendency for residues with correlated mutations to be spatially close is observed.

75. Goebel U, Sander C, Schneider R, Valencia A: **Correlated Mutations and Residue Contacts in Proteins.** *Proteins* 1994, 18:309-317.

Correlated mutations in multiple alignments are used to predict inter-residue contacts. The result is an improvement over a random prediction of 1.4-5.1 when evaluated on 13 protein families.

76. Finkelstein AV, Gutun AM, Badretdinov AY: **Why are the Same Protein Folds Used to Perform Different Functions?** *FEBS Lett* 1993, 325:23-28.

Proteins of homologous three-dimensional structure perform various functions. These authors propound the hypothesis that some frequently occurring folding patterns can be stabilized thermodynamically by many random sequences. In contrast, the folds that are rarely, or never, observed can be stabilized only by a tiny number of random sequences. The advantageous folds are few, they tolerate various primary structures, and therefore, they can perform different functions.

77. Finkelstein AV, Nakamura H: **Weak Points of Antiparallel β -Sheets. How Are They Filled Up in Globular Proteins?** *Protein Eng* 1993, 6:367-372.

78. Gerstein M, Sonnhammer ELL, Chothia C: **Volume Changes in Protein Evolution.** *J Mol Biol* 1994, 236:1067-1078.

The variation in volume that occurs during evolution in the buried core of three different families of proteins is not significant compared with the variation to be expected for random sequences. Thus, the volume variation does not reflect compensating changes in, or global constraints upon, protein sequences. Some individual sites in the core do, however, have a significant tendency to conserve their volume.

79. Nilges M, Brünger AT: **Successful Prediction of Colled Coil Geometry of the GCN4 Leucine Zipper Domain by Simulated Annealing: Comparison to the X-Ray Structure.** *Proteins* 1993, 15:133-146.

80. O'Donoghue SL, Junius FK, King GF: **Determination of the Structure of Symmetric Colled-Coil Proteins from NMR Data: Application of the Leucine Zipper Proteins Iaa and GCN4.** *Protein Eng* 1993, 6:557-564.

81. Yun-Yu S, Mark AE, Cun-Xin W, Fuhua H, Berendsen HJ, van Gunsteren WF: **Can the Stability of Protein Mutants be Predicted by Free Energy Calculations?** *Protein Eng* 1993, 6:289-295.

Among protein mutants, components of the free energy change are shown to be highly sensitive to the computational details of the simulation. The conclusion is that free energy calculations cannot currently be used to reliably predict protein stability.

82. Van Gunsteren WF: **Molecular Dynamics Studies of Proteins.** *Curr Opin Struct Biol* 1993, 3:167-174.

83. Littman EE, Rose GD: **Protein Folding—What's the Question?** *Proc Natl Acad Sci USA* 1993, 90:439-441.

84. Abagyan RA: **Towards Protein Folding by Global Energy Optimization.** *FEBS Lett* 1993, 325:17-22.

85. Novotny J, Rashin AA, Bruccoleri RE: **Criteria that Discriminate between Native Proteins and Incorrectly Folded Models.** *Proteins* 1988, 4:19-30.

86. Wodak SJ, Rooman MJ: **Generating and Testing Protein Folds.** *Curr Opin Struct Biol* 1993, 3:247-259.

An optimistic and comprehensive review covering most of the attempts to detect remote homologues by threading techniques.

87. Sippl MJ: **The Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-Based Prediction of Local Structure of Globular Proteins.** *J Mol Biol* 1990, 213:859-883.
88. Sippl MJ: **Boltzmann's Principle, Knowledge Based Mean Fields and Protein Folding. An Approach to the Computational Determination of Protein Structures.** *J Comput Aided Mol Des* 1993, 7:473-501.
- The principles of the underlying concept of potentials of mean force are described in a very detailed and understandable manner. Both the strengths and the weaknesses of the method are explained, and examples of its application are given.
89. Hendlich M, Lackner P, Weitckus S, Flockner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ: **Identification of Native Protein Folds Amongst a Large Number of Incorrect Models. The Calculation of Low Energy Conformations from Potentials of Mean Force.** *J Mol Biol* 1990, 216:167-180.
90. Sippl MJ, Weitckus S: **Detection of Native-Like Models for Amino Acid Sequences of Unknown Three-Dimensional Structure in a Data Base of Known Protein Conformations.** *Proteins* 1992, 13:258-271.
91. Holm L, Sander C: **Evaluation of Protein Models by Atomic Solvation Preference.** *J Mol Biol* 1992, 225:93-105.
92. Laskowski RA, Moss DS, Thornton JM: **Main-Chain Bond Lengths and Bond Angles in Protein Structures.** *J Mol Biol* 1993, 231:1049-1067.
93. Vriend G, Sander C: **Quality of Protein Models: Directional Atomic Contact Analysis.** *J Appl Crystallogr* 1993, 26:47-60.
- From an initial hypothesis that atom-atom interactions are the primary determinant of protein folding, protein models are tested for proper packing by calculation of a directional contact quality index. The power of the index is demonstrated by its application to a series of successively refined crystal structures. In all cases, the model known to be more valid has a better contact quality index. This is the first attempt at deriving an atomic pseudo-potential.
94. Sippl MJ: **Recognition of Errors in Three-Dimensional Structures of Proteins.** *Proteins* 1993, 17:355-362.
- When the potentials of mean force are applied to the whole data bank of known three-dimensional structures, four cases are shown to stand out, in terms of the shape of their pseudo-energy functions. These are predicted to be incorrect structures. The detection of non-native models is possible, even in cases where only the α -carbon trace of the protein conformation is available.
95. Casari G, Sippl MJ: **Structure-Derived Hydrophobic Potential.** *J Mol Biol* 1992, 224:725-732.
96. Bowie JU, Lüthy R, Eisenberg D: **A Method to Identify Protein Sequences that Fold into a Known Three-Dimensional Structure.** *Science* 1991, 253:164-169.
97. Sippl MJ, Jantz M: **Predictive Power of Mean Force Pair Potentials in Protein Folding.** In *Distance Based Approaches to Protein Structure Determination*. Edited by Bohr H, Brunak S. Amsterdam: IOS Press; 1994:113-134.
- Some 200 proteins of known three-dimensional structure are combined in a 'polyprotein'. The polyprotein is used to generate a physically meaningful background for threading approaches.
98. Sippl MJ, Weitckus S, Flockner H: **In Search of Protein Folds. In The Protein Folding Problem and Tertiary Structure Prediction.** Edited by Merz KH, LeGrand S. Boston, Massachusetts, USA: Birkhäuser Boston Inc; 1994 in press.
99. Sippl MJ, Jaritz M, Hendlich M, Ortner M, Lackner P: **Applications of Knowledge Based Mean Fields in the Determination of Protein Structures.** In *Statistical Mechanics, Protein Structure and Protein-Substrate Interactions*. Edited by Dornisch S. New York: Plenum Press; 1994 in press.
100. Bryant SH, Lawrence CE: **An Empirical Energy Function for Threading Protein Sequence through the Folding Motif.** *Proteins* 1993, 16:92-112.
- The distance-dependent potentials of mean force introduced in [87] are refined for threading sequences into known three-dimensional structures using a new optimization algorithm. The conclusion of this sophisticated refinement procedure is that, in some cases, threading can successfully detect correct models.
101. Wilmanns M, Eisenberg D: **Three-Dimensional Profiles from Residue-Pair Preferences: Identification of Sequences with β -Barrel Fold.** *Proc Natl Acad Sci USA* 1993, 90:1379-1383.
102. Miyazawa S, Jernigan RL: **A New Substitution Matrix for Protein Sequence Searches Based on Contact Frequencies in Protein Structures.** *Protein Eng* 1993, 6:267-278.
- New edition of a classical (circa 1985) contact potential by the same authors.
103. Nishikawa K, Matsuo Y: **Development of Pseudoenergy Potentials for Assessing Protein 3D-1D Compatibility and Detecting Weak Homologies.** *Protein Eng* 1993, 6:811-820.
- The distance-dependent potentials of mean force introduced in [81*] are split into four different terms to enable a better distinction between positive and false-positive hits for threading a sequence into a known three-dimensional structure. The method is thoroughly tested on a large database. These authors conclude that they are on 'the right track', but that false positives cannot be excluded.
104. Goldstein RA, Luthey-Schulten ZA, Wolynes P: **A Bayesian Approach to Sequence Alignment Algorithms for Protein Structure Recognition.** In *27th Hawaii International Conference on System Sciences*. Edited by Hunter L. Wailea, Hawaii, USA: IEEE Society Press; 1994:306-315.
105. Stultz CM, White JV, Smith TF: **Structural Analysis Based on State-Space Modeling.** *Protein Sci* 1993, 2:305-314.
106. Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL: **Fragment Ranking in Modelling of Protein Structure: Conformationally-Constrained Environmental Amino Acid Substitution Tables.** *J Mol Biol* 1993, 229:194-220.
107. Taylor WR, Jones DT, Green NM: **A Method for α -Helical Integral Membrane Protein Fold Prediction.** *Proteins* 1994, 18:281-294.
- For seven-helix membrane proteins, the structure prediction problem is simplified, as only one secondary structure exists. Thus, all possible structures can be represented on a simple lattice. The result is a successful automatic prediction of the correct fold. The limited number of examples of comparable structures makes it difficult to say whether or not the method will work as well for any membrane protein.
108. Kabsch W, Sander C: **Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features.** *Biopolymers* 1983, 22:2577-2637.
109. Scharf M: **Analysis of Residue Pair Interactions in Proteins (in German).** [Msc thesis]. Heidelberg: University of Heidelberg; 1989.
110. Kraulis P: **MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures.** *J Appl Crystallogr* 1991, 24:946-950.

B Rost and C Sander, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012 Heidelberg, Germany.