The future of protein secondary structure prediction accuracy

Dmitrij Frishman[1] and Patrick Argos[2]

Background: The accuracy of secondary structure prediction for a protein from knowledge of its sequence has been significantly improved by about 7% to the 70–75% range by inclusion of information residing in sequences similar to the query sequence. The scientific literature has been inconsistent, if not negative, regarding chances for further improvement from the vast knowledge to be provided by genome sequencing efforts.

Results: By applying a prediction technique that is particularly sensitive to added sequence information to a standard set of query sequences with related primary structures taken from chronologically successive releases of the SWISS-PROT database, it is shown that prediction accuracy can be expected to reach 80–85% with a large 10-fold increase in present sequence knowledge.

Conclusions: Even with present prediction approaches, improvement in prediction accuracy can still be expected, albeit limited to no more than 10%.

Addresses: [1]Martinsried Institute for Protein Sequences, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18a, 82152 Martinsried, Germany. [2]European Molecular Biology Laboratory, Postfach 102209, Meyerhofstraβe 1, 69012 Heidelberg, Germany; e-mail: argos@embl-heidelberg.de.

Correspondence: Dmitrij Frishman
e-mail: frishman@mips.biochem.mpg.de

Key words: information content, protein structure, secondary structure, sequence alignment, sequence databases

Received: **10 Feb 1997**
Accepted: **13 Feb 1997**

Published: **11 Mar 1997**
Electronic identifier: **1359-0278-002-00159**

Folding & Design 11 Mar 1997, **2**:159–162

Introduction

Since the classic works of Chou and Fasman [1] and Lim [2] were published more than two decades ago, the accuracy of protein secondary structure prediction in three states (α-helix, β-strand, and coil) from sequence information has been steadily increasing at an average rate of a little less than 1% a year. Until recently, the major source of improvement in prediction from single sequence information has been the application of more sophisticated recognition algorithms, such as neural networks [3,4] and the nearest neighbor approach [5,6], along with the growth of available protein tertiary structures used for training [7]. In recent years, more remarkable improvement has been achieved by utilization of multiply aligned sequence homologs with the best reported accuracies exceeding 70% [8–13]. It has been demonstrated that the additional information contained in a set of related primary structures yields an accuracy gain of 5–7% relative to prediction from only a single sequence [14].

The achievable secondary structure prediction accuracy has been a major topic of discussion [15,16]. It has been argued that further significant improvement in accuracy is unlikely. However, a theoretical study [17] has noted that the information potential of current sequence/structure databases has not been exhausted and has suggested that significantly higher prediction accuracies, up to 85%, are possible from consideration of higher order, rather than purely local, information as well as from the extended knowledge of sequence homologs.

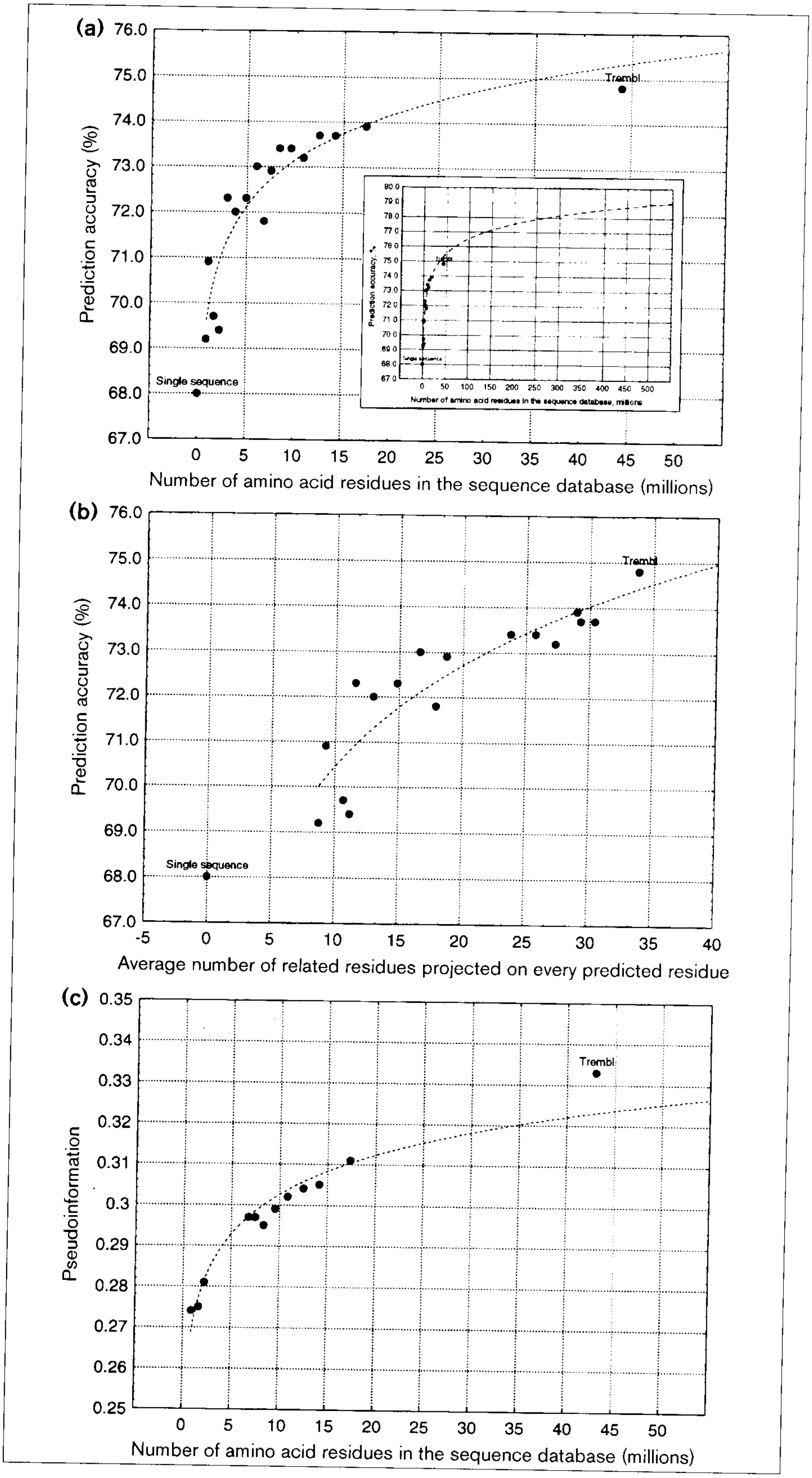
We have investigated the role of homologous sequence information in secondary structure prediction by conducting a large-scale computational experiment in which the growth of the available sequence data was artificially simulated by considering chronologically successive, but existing, releases of the SWISS-PROT database and, as the asymptotical case, the TREMBL database created by translating all coding frames in a very recent EMBL nucleotide sequence database [18,19]. It is shown that the amount and quality of sequence data available crucially influence the prediction accuracy; however, the improvement expected can be no more than 10% with present approaches, and probably less.

Results and discussion

Prediction accuracy versus sequence database size

Figure 1a shows the dependence of the prediction accuracy achieved by our secondary structure prediction program PREDATOR on the number of amino acid residues in a given sequence database release. One immediate observation is that the availability of even a very small sequence database (release 2 with 3939 amino acid sequences and 900 163 residues) improved the prediction compared to that from the single query sequence (an improvement from 68% to 69.2%). Further, the curve appears steep in the region corresponding to the SWISS-PROT releases 2–16 (years 1986–1990; accuracy increase from 69.2% to 73%) and then begins to flatten despite the explosive growth in the number of available protein sequences. Although the TREMBL database contains three times more sequence data than the largest SWISS-PROT release considered in

Figure 1



(a) Dependence of secondary structure prediction accuracy on the size of the prot sequence database used to extract homologous information. The insert shows logarithmic extrapolation of the data for a 1 fold increase in the sequence data availabl **(b)** Dependence of the secondary structur prediction accuracy on the average numb amino acid residues extracted from the protein sequence database through datab searches and alignments over each predic residue in our sample of 125 protein chain **(c)** Dependence of the average pairwise alignment information content, or pseudoinformation, per aligned residue on database size.

Each dot corresponds to one release of the SWISS-PROT database; only even release numbers (from 2 to 32) were considered to reduce computational requirements. Values for the single-sequence case (no homologo information available) and for the TREMBL database simultaneous to release 32 of SWISS-PROT (November 1995) are also shown. Logarithmic regression is shown as dashed line.

this work, the gain in accuracy achieved by its use was a mere 0.9%. Logarithmic extrapolation of the available data to the case of a 10-fold increase of the database size (corresponding roughly to sequence knowledge in two human genomes) shows that the potential for prediction improvement from multiple sequences is not exhausted and that accuracy close to 80% is feasible (see insert, Fig. 1a).

Prediction accuracy versus protein family size

The quality of the prediction does not depend directly on the total volume of the sequence database, but rather on the number of sequences related to the query sequence. Furthermore, only significant and nontrivial similarities in the range of 25–90% residue identity with the query sequence contribute to the prediction. Figure 1b illustrates the dependence of the prediction accuracy on the average number of individual database residues reliably related to each of the residues in the query sequence through careful subsequence pairwise alignment. After the relatively steep growth corresponding to early database releases, the plot acquires a nearly linear character, with approximately 0.5–1.0% accuracy improvement per every additional five related residues. These relationships suggest that the slowing growth in prediction accuracy is a result of decreasing addition of new sequences related to the particular set of 125 protein chains tested.

Prediction accuracy versus data quality

Another crucial factor in prediction accuracy is the quality of the related sequence sets available for the prediction. Addition of subsequences trivially related to the query sequence with very high percent residue identity after alignment does not add substantially new information. On the other hand, using sequences questionably related to the query sequence (identity of 20–25%) is counterproductive, as the relationship may not imply structural similarity. For both these extreme cases, the information content will be low (e.g. for 15% and 80% or $\Omega^{0,m}_q = 0.15$ and 0.80, the pseudoinformation values will be $I^{0,m}_q = 0.28$ and 0.17, respectively; see Materials and methods) and the contribution of the corresponding pairwise alignments downweighted. Availability of sequences related to the query sequence in the range of 36% identity has the strongest influence on the prediction quality. As seen in Figure 1c, the average information content per each aligned residue used for prediction is steadily growing with each sequence database release, but is unlikely to reach its optimal value of 0.37.

Pairwise versus multiple alignments

Reliance on rigorous pairwise alignment between the sequence to be predicted and other related sequences or sequence fragments avoids many difficulties characteristic of hierarchical multiple and global sequence alignments, where unreliably related sequence regions are more likely [20,21]. It must also be stressed that the predictions in this work are not consensus predictions for an entire protein family and are made for one protein sequence considering related subsequences through pairwise comparisons. This process avoids the limitations imposed on the achievable prediction accuracy by the variation of observed secondary structures amongst different family members [22].

Conclusions

The limiting factor in secondary structure prediction accuracy from multiple and related sequences is not the principal inability of machine intelligence methods to make use of additional information in ever larger sequence families, but the natural limitations on the amount and diversity of available sequence information resulting from sequencing efforts. With an increase in sequencing speed and target species, the average achievable accuracy of prediction still has a potential for improvement. Nevertheless, 80–85% correctness would appear to be the upper limit without a breakthrough in prediction approaches.

Materials and methods

Secondary structure prediction algorithm

Secondary structure predictions were effected with the program PREDATOR [13,23]. The average prediction accuracy of the method is 68% from a single sequence and 75% from multiple sequence sets. The two most novel features of the algorithm are utilization of secondary structure propensities based on both local and long-range effects, and utilization of similar sequence information in the form of carefully selected sequence fragments, taken from available databases and significantly related to those of the query sequence through pairwise local alignment, rather than global multiple alignments of entire sequences. The secondary structure propensities of the related subsequences (1 – *m*) are combined with (projected onto) those of the query sequence 0 and weighted according to their information content *I* (or pseudoinformation) taken from the corresponding pairwise alignments; namely, $I^{0,m}_q = -\Omega^{0,m}_q \ln\Omega^{0,m}_q$ where $\Omega^{0,m}_q$ is the fraction of identical residues in the local alignment *q* for sequence fragment *m*. $I^{0,m}_q$ reaches its maximum value (0.367) when $\Omega^{0,m}_q = 0.36$ or 36%.

The source code, documentation and executables of our secondary structure prediction program PREDATOR are freely available for academic users via anonymous ftp from ftp.ebi.ac.uk (directories /pub/software/unix/predator and /pub/software/dos/predator). Protein sequences can be submitted for secondary structure prediction either via the internet to http://www.embl-heidelberg.de/predator/predator_info.html or through electronic mail to predator@embl-heidelberg.de. A mail message containing HELP in the first line will be appropriately answered.

Training and testing

Predictions were generated for a list of 125 nonhomologous proteins that was published by Rost and Sander [9] and now constitutes a comparative standard. Related sequences were extracted through FASTA (version 2.0) database searches [24] using a uniform cutoff threshold of 0.0001 for statistical significance of subsequence relationships. Each sequence set was made nonredundant such that no two sequence members shared more than 95% identical residues after alignment. A full jackknife procedure was performed to test achievable accuracy by excluding one of the 125 protein structures and the corresponding sequence set, deriving database statistics from the remaining 124 structures and recognition parameters from the remaining 124 sequence sets, and finally using this information to predict secondary structure for the excluded protein. The final prediction accuracy resulted from averaging over the 125 proteins, each under jackknife conditions.

Acknowledgements

We would like to thank Hans-Werner Mewes for careful reading of the manuscript and David Starks-Browning for his friendly assistance in using the EMBL computer system.

References

1. Chou, P.Y. & Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry* **13,** 222–245.
2. Lim, V.I. (1974). Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J. Mol. Biol.* **88,** 873–894.
3. Qian, N. & Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202,** 865–884.
4. Holley, L.H. & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
5. Levin, J.M., Robson, B. & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* **205,** 303–308.
6. Zhang, X., Mesirov, J.P. & Waltz, D.L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225,** 1049–1063.
7. Bernstein, *et al.*, & Tasumi, M. (1977*).* The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542.
8. Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195,** 957–961.
9. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–599.
10. Mehta, P.K., Heringa, J. & Argos, P. (1995). A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.* **4,** 2517–2525.
11. Salamov, A.A. & Solovyev, V.V. (1995). Prediction of protein secondary structure by combining nearest-neighbour algorithms and multiple sequence alignments. *J. Mol. Biol.* **247,** 11–15.
12. Geourjon, C. & Deléage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple sequences. *Comput. Appl. Biosci.* **11,** 681–684.
13. Frishman, D. & Argos, P. (1997). 75% accuracy in protein secondary structure prediction. *Proteins* **27,** 329–335.
14. Levin, J., Pascarella, S., Argos, P. & Garnier, J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6,** 849–854.
15. Kabsch, W. & Sander, C. (1983). How good are predictions of protein secondary structure? *FEBS Lett.* **155,** 179–182.
16. Russell, R.B. & Sternberg, M.J.E. (1995). Structure prediction: how good are we? *Curr. Biol.* **5,** 488–490.
17. Rao, S., Zhu, Q.-L., Vaida, S. & Smith, T. (1993). The local information content of the protein structural database. *FEBS Lett.* **2,** 143–146.
18. Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* **24,** 21–25.
19. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. & Cameron, G.N. (1993). The EMBL data library. *Nucleic Acids Res.* **21,** 2967– 2971.
20. Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences. The twilight zone revisited. *J. Mol. Biol.* **249,** 816–831.
21. Di Francesco, V., Garnier, J. & Munson, P.J. (1996). Improving protein secondary structure prediction with aligned homologous sequences. *Protein Sci.* **5,** 106–113.
22. Russell, R.B. & Barton, G.J. (1993). The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* **234,** 951–957.
23.. Frishman, D. & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9,** 133–142.
24. Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444-2448.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad.htm – for further information, see the explanation on the contents pages.