Learning from Evolution to Predict Protein Structure

Burkhard Rost

European Molecular Biology Laboratory

EMBL, 69 012 Heidelberg, Germany; rost@embl-heidelberg.de; http://www.embl-heidelberg.de/~rost/

Invited talk

Gøteborg, Sweden, Jun. 22-26, 1998 (ECMI98)

Learning from Evolution to Predict Protein Structure

Burkhard Rost

European Molecular Biology Laboratory

EMBL, 69 012 Heidelberg, Germany; rost@embl-heidelberg.de; http://www.embl-heidelberg.de/~rost/

Abstract. 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.

Introduction

The sequence-structure gap is rapidly increasing: Currently, databases for protein sequences (e.g. SWISS-PROT (1)) are expanding rapidly, largely due to large-scale genome sequencing projects. Despite significant improvements of structure determination techniques the gap between the number of proteins for which structure is deposited in public databases (PDB (2)), and the number of proteins for which sequences are known is increasing: 3D structure is known for less than 3% of the known protein sequences (3, 4). The most accurate way to predict 3D structure from the sequence is by homology modelling, i.e., search for a protein with similar sequence that has a known 3D structure and then model the 3D structure of the unknown protein in analogy to the known one. Such techniques lead to a reduction of the sequence-structure gap to 10-35% (3-7).

No general prediction of structure from sequence, yet: John Moult (CARB, Washington) has initiated an important, and unique experiment (8): those who determine protein structures submitted the sequences of proteins for which they were about to solve the structure to a 'to-be-predicted' database; for each entry in that database predictors could send in their predictions before a given deadline (the public release of the structure); finally, the results were compared, and discussed during a workshop (in Asilomar, California). Two such experiments have been completed: in December 1994 (Proteins special issue, Vol. 23, 1995), and in December 1996 (to be published in Proteins, 1997). The results of both experiments demonstrated clearly that the goal to predict structure from sequence has not been reached, yet. So, no improvement despite ardent attempts, and the explosion of knowledge deposited in databases?

Here, I sketch neural network based methods (PHD series) for the prediction of 1D aspects (secondary structure, transmembrane helices, solvent accessibility) of protein structure. The methods illustrate that (1) neural networks as black-boxes fail to improve prediction accuracy, (2) neural networks are sufficiently flexible to carve expertise from biology into the tool, (3) the quantum leap in prediction accuracy achieved in the 90's has unearthed from

implementing evolutionary information into neural networks, (4) and that the new generation of prediction methods is extremely useful in assisting, facilitating, and speeding-up experiments in molecular biology.

Conclusions: Are Predictions Useful?

Structure prediction: work in progress... Native 3D structures of proteins are encoded by a linear sequence of amino acid residues. To predict 3D structure from sequence is a task challenging enough to have occupied a generation of researchers. Have we finally succeeded? The bad news is: no, we still cannot predict structure for any sequence. The good news are: we have come closer, and growing databases facilitate the task.

Predictions in 1D: significant improvement by larger databases: The rich information contained in the growing sequence and structure databases enables improving the accuracy of 1D predictions. Here I sketched, how evolutionary information input to neural network systems yielded better predictions of secondary structure, solvent accessibility, and transmembrane helices. These predictions of protein structure in 1D are significantly more accurate, and more useful than five years ago.

Conditions to become useful: In the field of structure prediction we have witnessed blooming over-optimism (9), as well as, more, and less intended cheating. The Asilomar meetings (8) to some extent are succeeding in separating the chaff from the wheat. The sustained levels of prediction accuracy published for the PHD methods were, supposedly, one of the major reason for their success. Another important issue is that of making the method available. Molecular biologists do NOT have the time to become experts in running programs. Thus, methods should be easy-to-use, and available via the internet (5).

Typical applications of 1D predictions: The PHD series was the first structure prediction suite available by the internet server PredictProtein (10-12). Five years later, PredictProtein handles about 150-200 request every day (11). The background of users range from theoreticians who use predictions as one module for their prediction program (next paragraph) to biologists who use the predictions to investigate structure, function, and to suggest which residues to mutate in experiments (13-18). Accurate prediction of secondary structure can also assist in X-ray diffraction (e.g., the GroEL crystal structure was derived making use of secondary structure predictions for the molecular replacement search (19)). In principle, the early stages of NMR frequency assignment could also be aided by knowledge of the secondary structure, although this has not been attempted. 1D predictions and predictions of transmembrane topology have proven to be quick and accurate enough for the analysis of entire genomes (20, 21). The predictions of transmembrane helices provided a lower bound to approach the question of how many proteins organisms need for, e.g., communication: the percentage of proteins with transmembrane helices has been estimated to be about 25% for yeast and haemophilus influenzae, and around 10-15% for mycoplasma genitalium and methanococcus jannaschii (22, 23). Predictions of accessibility were used as basis for predicting functional sites (4), and to predict sub-cellular location (24).

1D predictions as input to threading techniques: Threading methods attempt to recognise similarities between protein folds in the absence of significant sequence identity (25). The stakes are high, as most protein pairs of similar structure populate this region (26), but the problem is highly non-trivial (25, 27, 28). Recently, PHD predictions of 1D structure have been implemented successfully to develop a new generation of prediction-based threading methods (29-33). Indeed, these methods are more successful than conventional sequence alignment techniques alone. A consequence was that most threading predictions presented at the Asilomar meeting of 1996 made use of 1D predictions from PHD.

What next? Most breakthroughs in protein structure prediction were achieved over the last six years. Thus, although we still cannot solve the general prediction problem, progress has been made. In general, however, we could ask the question - is it worth persevering with structure prediction, given that it is clearly such a difficult task? The answer is: yes. The methods which have spun off from structure prediction have already given us considerable insight into the first four complete genomes. Perseverance with structure prediction will yield fruit in about five years time when the human genome will be known.

References

1. Bairoch, A. & Apweiler, R. (1997) Nucl. Acids Res. 25, 31-36.2. Bernstein, F. C., et al. (1977) J. Mol. Biol. 112, 535-542.3. Rost, B. & Sander, C. (1996) Annu. Rev. Biophys. Biomol. Struct. 25, 113-136.4. Rost, B. & O'Donoghue, S. I. (1997) CABIOS in press.5. Rost, B. & Schneider, R. (1997) Pedestrian guide to analysing sequence databases (Springer, Heidelberg).6. Schneider, R., et

al. (1997) Nucl. Acids Res. 25, 226-230.7. Casari, G., et al. (1996) TiGs 12, 244-245.8. Moult, J., et al. (1995) Proteins 23, ii-iv.9. Honig, B. & Cohen, F. E. (1996) Folding & Design 1, R17-R20.10. Rost, B. (1996) Meth. Enzymol. 266, 525-539.11. Rost, B. (1997) PredictProtein - prediction service (http://www.embl-heidelberg.de/predictprotein).12. Rost, B. & Sander, C. (1992) Nature 360, 540.13. Rawlings, D. J., et al. (1993) Science 261, 358-361.14. Lupas, A., et al. (1994) FEBS Lett. 354, 45-49.15. Hubbard, T. & Park, J. (1995) Proteins 23, 398-402.16. Springer, T. A. (1997) Proc. Natl. Acad. Sc. 94, 65-72.17. Valencia, A., et al. (1995) Proteins 22, 199-209.18. Hansen, J. E., et al. (1996) Proteins 25, 1-11.19. Braig, K., et al. (1994) Nature 371, 578-586.20. Koonin, E. v., et al. (1996) Meth. Enzymol. 266, 295-322.21. Odgren, P. R., et al. (1996) Proteins 24, 467-484.22. Rost T, B. (1996) Sneaking in genomes for helical transmembrane proteins (EMBL, Heidelberg, Germany).23. Rost, B., et al. (1996) Prot. Sci. 5, 1704-1718.24. Andrade, M. A., et al. (1997) in submission.25. Sippl, M. J. (1995) Curr. Opin. Str. Biol. 5, 229-235.26. Rost, B. (1997) Folding & Design 2, S19-S24.27. Sippl, M. J., et al. (1994) Applications of Knowledge Based Mean Fields in the Determination of Protein Structures (Plenum Press, New York, London).28. Lathrop, R. H. (1994) Prot. Engin. 7, 1059-1068.29. Russell, R. B., et al. (1996) J. Mol. Biol. 259, 349-365.30. Rost, B. (1995) TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures (Menlo Park, CA: AAAI Press, Cambridge, England).31. Rost, B. (1995) Fitting 1-D predictions into 3-D structures (CRC Press, Boca Raton, Florida).32. Rost, B., et al. (1997) J. Mol. Biol. 270, 471-480.33. Fischer, D. & Eisenberg, D. (1996) Prot. Sci. 5, 947-955.