Bioinformatics - Applications Note

JPred: A Consensus Secondary Structure Prediction Server

James A. Cuff†‡, Michele Clamp‡

Asim S. Siddiqui†, Matt Finlay†

and Geoffrey J. Barton†‡

†Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford, OX1 3QU, UK

‡European Molecular Biology Laboratory Outstation - The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Corresponding Author: G. J. Barton, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Keywords: protein; secondary structure prediction; combination of methods; www server

Summary: *An interactive protein secondary structure prediction Inter-*

net server is presented. The server allows a single sequence or multiple alignment to be submitted, and returns predictions from six secondary structure prediction algorithms that exploit evolutionary information from multiple sequences. A consensus prediction is also returned which improves the average Q_3 *accuracy of prediction by 1% to 72.9%. The server simplifies the use of current prediction algorithms and allows conservation patterns important to structure and function to be identified.*

Availability: http://barton.ebi.ac.uk/servers/jpred.html

Contact: geoff@ebi.ac.uk

When predicting the secondary structure of a protein 'blind', without knowledge of the answer, it is useful to exploit the features of all available prediction algorithms rather than rely on one. Combination of methods has been applied successfully in a number of accurate predictions of protein secondary structure (e.g. see [Edwards & Perkins, 1996, Crawford *et al.*, 1987, Russell *et al.*, 1992, Livingstone & Barton, 1994, Livingstone & Barton, 1996, Russell & Barton, 1993]). Unfortunately, combining prediction methods on a large scale is complicated by the fact that prediction programs have very different input requirements and output formats. In order to perform a recent large-scale comparative analysis of secondary structure prediction algorithms [Cuff & Barton, 1998], we developed flexible software to standardise the input and output requirements of 6 prediction algorithms. In this Applications Note we describe a development of this work to provide the fully automatic JPred WWW server for multiple secondary structure prediction.

The server accepts two input types, a family of aligned protein sequences or a single protein sequence. If a single sequence is submitted, an automatic process creates a multiple sequence alignment, prior to prediction [Cuff & Barton, 1998].

Six different prediction methods (DSC [King & Sternberg, 1996], PHD [Rost & Sander, 1993], NNSSP [Salamov & Solovyev, 1995], PREDATOR [Frishman & Argos, 199 ZPRED [Zvelebil *et al.*, 1987] and MULPRED (Barton, 1988, unpublished) are then run, and the results from each method are combined into a simple file format.

The NNSSP, DSC, PREDATOR, MULPRED, ZPRED and PHD methods were chosen as representatives of current state of the art secondary structure prediction methods, that exploit the evolutionary information from multiple sequences. Each derives its prediction using a different heuristic, based upon nearest neighbours (NNSSP), jury decision neural networks (PHD), linear discrimination (DSC), consensus single sequence method combination (MULPRED), hydrogen bonding propensities (PREDATOR), or conservation number weighted prediction (ZPRED).

The predictions and corresponding sequence alignment are rendered in coloured HTML, Java [Clamp *et al.*, 1998] and Postscript. The predictions are coloured and aligned with their corresponding family of sequences. Physico-chemical properties, solvent accessibility, prediction reliability and conservation number values [Zvelebil *et al.*, 1987] for each amino acid are included in the output. The original ASCII text data from each of the prediction methods can also be downloaded. For example, BLAST results, MSF and HSSP format alignments, pair comparison files and so on.

The text based MULPRED output is of particular interest as it is a combination of different single sequence prediction methods (GOR [Garnier *et al.*, 1978], Chou-Fasman (1974), Lim (1974), Rose (1978), and Wilmot & Thornton (1988)). While the automatic consensus within the MULPRED program is not as accurate as current methods [Cuff & Barton, 1998], the profile based output, when combined with the modern algorithms is helpful for human interpretation of the prediction.

A consensus prediction based upon a simple majority wins combination of NNSSP, DSC, PREDATOR and PHD is provided by the JPred server. If there is a tie, the prediction from PHD is used. In our independent test this approach gave the highest accuracy compared to all other combinations [Cuff & Barton, 1998].

The consensus prediction achieved an average Q_3 score of 72.9%, where Q_3 is the percentage of residues predicted correctly for the three conformational states, strand helix and loop. This result is 1% better than PHD (71.9%) for the same data. The segment overlap score [Rost *et al.*, 1994] for the consensus method improves by 0.1 %, to 75.4%. These results were obtained on a non-redundant set of 396 protein domain sequences, that did not contain sequences similar to the proteins used to train the methods. Cuff & Barton 1998, includes a full analysis and description of the data sets, similarity cutoffs, accuracies and methods used for this test.

As well as providing a more accurate consensus prediction, the JPred server also permits the different prediction methods to be viewed concurrently with the alignment. This allows for easy interpretation and analysis of the prediction and multiple sequence alignment. Interactive analysis and re-alignment can also be carried out with the Java viewer and editor [Clamp *et al.*, 1998], where one may interactively change the colouring within the alignment to highlight important residues and conserved features.

In summary, JPred provides an automatic and simple to use tool to assist in accurate secondary structure prediction.

Acknowledgements

We thank Drs B. Rost, D. Frishman, V. Solovyev, R. King and M. Zvelebil, for permitting us to include their software in the JPred server and for helpful discussions. This work was supported in part by grants from the Medical Research Council and the Royal Society. JC is an Oxford Centre for Molecular Sciences/MRC student.

References

[Chou & Fasman, 1974] Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β–sheet, and random coil regions calculated from proteins. *Biochem.* **13**, 211–222.

[Clamp *et al.*, 1998] Clamp, M. E., Cuff, J. A. & Barton, G. J. (1998). Jalview - a java multiple sequence alignment viewer and editor. http://barton.ebi.ac.uk/.

[Crawford *et al.*, 1987] Crawford, I. P., Niermann, T. & Kirchner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins,* **2**, 118–129.

[Cuff & Barton, 1998] Cuff, J. A. & Barton, G. J. (1998). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins, submitted.

[Edwards & Perkins, 1996] Edwards, Y. J. K. & Perkins, S. J. (1996). Assessment of protein fold predictions from sequence information: The predicted α/β doubly wound fold of the von Willebrand factor type A domain is similar to its crystal structure. *J. Mol. Biol.* **260**, 277–285.

[Frishman & Argos, 1997] Frishman, D. & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins,* **27**, 329–335.

[Garnier *et al.*, 1978] Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

[King & Sternberg, 1996] King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* **5**, 2298–2310.

[Lim, 1974] Lim, V. I. (1974). Algorithms for prediction of α helices and β structural regions in globular proteins. *J. Mol. Biol.* **88**, 873–894.

[Livingstone & Barton, 1994] Livingstone, C. D. & Barton, G. J. (1994). Secondary structure prediction from multiple sequence data: blood clotting factor xiii and yersinia protein tyrosine phosphatase. *Int. J. Pept. Prot. Res.* **44**, 239–244.

[Livingstone & Barton, 1996] Livingstone, C. D. & Barton, G. J. (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. *Meth. Enz.* **266**, 497–512.

[Rose, 1978] Rose, G. D. (1978). Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature,* **272**, 586–591.

[Rost & Sander, 1993] Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

[Rost *et al.*, 1994] Rost, B., Sander, C. & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235**, 13–26.

[Russell & Barton, 1993] Russell, R. B. & Barton, G. J. (1993). An SH2–SH3 domain hybrid. *Nature,* **364**, 765.

[Russell *et al.*, 1992] Russell, R. B., Breed, J. & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters,* **304**, 15–20.

[Salamov & Solovyev, 1995] Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest- neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11–15.

[Wilmot & Thornton, 1988] Wilmot, A. C. M. & Thornton, J. M. (1988). Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.* **203**, 221–232.

[Zvelebil *et al.*, 1987] Zvelebil, M., Barton, G., Taylor, W. & Sternberg, M. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.