BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 4 2000
Pages 404–405

The PSIPRED protein structure prediction server

Liam J. McGuffin[1, 2], Kevin Bryson[1] and David T. Jones[1, 2]

[1]Protein Bioinformatics Group, Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK and [2]Department of Biological Sciences, Brunel University, Uxbridge, UB8 3PH, UK

Received on November 8, 1999; accepted on December 14, 1999

Abstract

Summary: *The PSIPRED protein structure prediction server allows users to submit a protein sequence, perform a prediction of their choice and receive the results of the prediction both textually via e-mail and graphically via the web. The user may select one of three prediction methods to apply to their sequence: PSIPRED, a highly accurate secondary structure prediction method; MEMSAT 2, a new version of a widely used transmembrane topology prediction method; or GenTHREADER, a sequence profile based fold recognition method.*

Availability: *Freely available to non-commercial users at http://globin.bio.warwick.ac.uk/psipred/*

Contact: *david.jones@brunel.ac.uk*

Introduction

Efficient automatic methods for protein structure prediction are becoming increasingly important as a result of the influx of genomic data arising from sequencing projects. Methods for predicting topologies of both globular and membrane bound proteins have been publicly available as individual programs designed for specific platforms. However, in order to make methods more accessible, structure prediction web servers incorporating these programs are becoming more prevalent.

The PSIPRED protein structure prediction server incorporates three recently developed methods (PSIPRED, GenTHREADER and MEMSAT 2) for predicting structural information about a protein from its amino acid sequence alone. PSIPRED (Jones, 1999a) carries out a reliable secondary structure prediction on a protein and gives its name to the prediction server itself. GenTHREADER (Jones, 1999b) quickly and reliably recognizes the fold of a protein and MEMSAT 2 (Jones *et al.*, 1994; Jones, 1998) is the latest version of a robust method for inferring the structure and topology of transmembrane proteins.

The user can submit a protein sequence to the server either in single letter amino acid code format or in FastA format, and then they may select one of the following prediction methods.

Prediction of secondary structure (PSIPRED)

A new highly accurate secondary structure prediction method, PSIPRED incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated BLAST) (Altschul *et al.*, 1997).

Using a stringent cross validation procedure to evaluate performance, PSIPRED has been shown to be capable of achieving an average Q_3 score of 76.5%. This is the highest level of accuracy published for any method to date. Predictions produced by PSIPRED were also submitted to the CASP3 (Third Critical Assessment of Structure Prediction) server and assessed during the CASP3 meeting, which took place in December 1998 at Asilomar. Assessors at CASP3 ranked PSIPRED first out of all of the secondary structure prediction methods evaluated, achieving an average Q_3 score of 73.4% over the hardest category and an overall average of 77.3% across all submitted targets (Orengo *et al.*, 1999).

A Java front-end application, PSIPREDView, has also been developed to interpret the output from PSIPRED and to provide users with publication quality graphical representations of their secondary structure prediction (Figure 1). A hypertext link to the location of the images is included in the e-mail following the text representation of the prediction.

Prediction of transmembrane structure and topology (MEMSAT)

MEMSAT 2 is the latest version of the widely used all-helical membrane prediction method MEMSAT (Jones *et al.*, 1994). It predicts the secondary structure and topology of integral membrane proteins by recognizing topological models. It differs from the original method by making use of multiple sequence alignments generated by PSI-BLAST, across which the log-likelihood ratio scores are summed. As with other methods, the use of consensus information in the scoring of different topological models has produced a significant increase in prediction accuracy. From a benchmark test set of 86 transmembrane proteins of known topology, the original MEMSAT method achieved

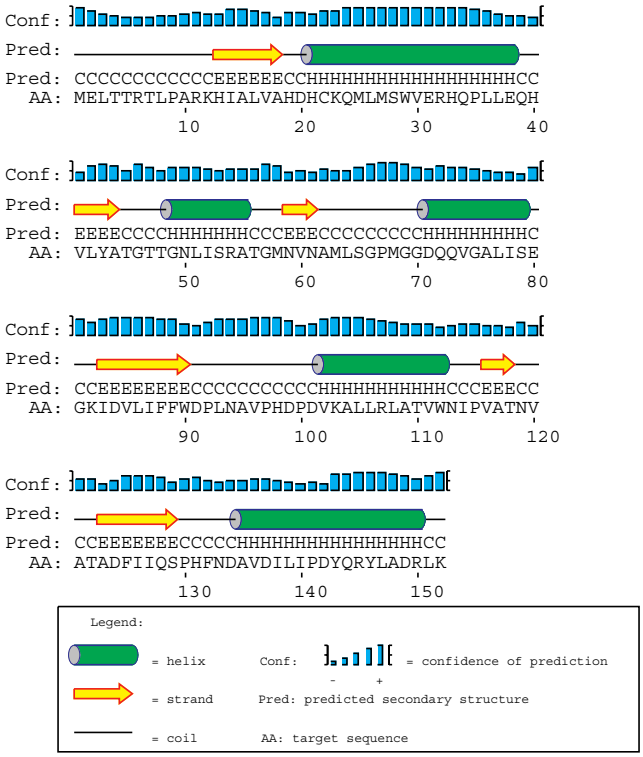

Fig. 1. PSIPRED graphical output from prediction of methylglyoxal synthase (CASP3 target 'T0081') produced by PSIPREDView—a Java visualization tool that produces two-dimensional graphical representations of PSIPRED predictions.

a success rate of only 78% (67 proteins out of 86 had a correctly predicted topology). In contrast to this, MEMSAT 2 was able to correctly predict the correct topologies of 80 out of 86, giving MEMSAT 2 an estimated accuracy of over 93% at predicting the structure and topology of all-helical transmembrane proteins and the location of their constituent helical elements within a membrane.

Fold recognition (GenTHREADER)

A new fold recognition method, GenTHREADER (Jones, 1999b), can be applied to either whole, translated genomic sequences (proteomes) or individual protein sequences, as in the case of the PSIPRED server. The method is aimed particularly at detecting superfamily relationships (i.e. fold similarities resulting from common ancestry), and exploits a conventional sequence profile-based alignment algorithm to generate sequence-structure alignments, which are then analysed by a set of statistical potentials previously used for threading (Jones *et al.*, 1992).

Of the open reading frames in the *Mycoplasma genitalium* genome, GenTHREADER has assigned some 47% (presently 51%) to at least one protein of known three-dimensional structure (Jones, 1999b). The accuracy of this method has been assessed by the first Critical Assessment of Fully Automated Structure Prediction Methods (CAFASP) where it ranked in first place for recognition of superfamily relationships (which is its intended application) and second place overall (Fischer, 1999).

A second variant of the GenTHREADER method is also provided on the PSIPRED server. This method first calculates a multiple sequence profile for the target sequence using PSI-BLAST, and uses this profile to make an alignment to sequences corresponding to each member of the fold library. This variation of the GenTHREADER method can sometimes detect relationships which are not picked up by the other method (and vice versa).

Server implementation

One significant benefit of using PSI-BLAST as the front-end to these prediction methods is that the multiple sequence alignment phase of each prediction method is far less time consuming than for other servers which typically use traditional multiple sequence alignment methods to generate sequence profiles. Indeed, rather than requiring expensive server hardware, the PSIPRED server runs very effectively on a cheap dual-processor Linux machine fitted with 512 Mb of RAM. Despite the low-cost of this system, the PSIPRED server is currently able to complete predictions in most cases in less than 2 min, and currently processes around 100 prediction requests a day. We anticipate that the current hardware will be able to cope with five times this load before an additional server will need to be installed.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.

Fischer,D., Christian,B., Bryson,K., Elofsonn,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1: Critical assessment of fully automated structure prediction methods. *PROTEINS Suppl.*, **3**, 209–217.

Jones,D.T. (1998) Do transmembrane protein superfolds exist? *FEBS Letters*, **423**, 281–285.

Jones,D.T. (1999a) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. (1999b) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Orengo,C.A., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. *PROTEINS Suppl.*, **3**, 149–170.