BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1438–1439
doi:10.1093/bioinformatics/btr096

Structural bioinformatics

Advance Access publication February 23, 2011

The MEMPACK alpha-helical transmembrane protein structure prediction server

Timothy Nugent*, Sean Ward and David T. Jones*

Bioinformatics Group, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The experimental difficulties of alpha-helical transmembrane protein structure determination make this class of protein an important target for sequence-based structure prediction tools. The MEMPACK prediction server allows users to submit a transmembrane protein sequence and returns transmembrane topology, lipid exposure, residue contacts, helix–helix interactions and helical packing arrangement predictions in both plain text and graphical formats using a number of novel machine learning-based algorithms.

Availability: The server can be accessed as a new component of the PSIPRED portal by at http://bioinf.cs.ucl.ac.uk/psipred/.

Contact: d.jones@cs.ucl.ac.uk; t.nugent@cs.ucl.ac.uk

Received on November 25, 2010; revised on January 27, 2011; accepted on February 17, 2011

1 INTRODUCTION

Given the biological and pharmacological importance of transmembrane (TM) proteins and the difficulties associated with obtaining their crystal structures, the use of bioinformatics approaches to direct experimental work while furthering our understanding of their structure and function is essential. The MEMPACK prediction server applies a selection of machine learning-based tools to predict TM topology—the total number of TM helices, their boundaries and in/out orientation relative to the membrane—with the addition of lipid exposure, residue contacts, helix–helix interactions, culminating in prediction of the optimal helical packing arrangement using a force-directed algorithm. Figure 1 provides an example of some of the server output. The underlying tools have recently been shown to provide significant improvements in prediction accuracy compared with existing methods. It is hoped that this service will be of benefit to the broader scientific community.

2 METHODS

In order to predict TM protein topology, the server employs the MEMSAT3 (Jones, 2007) and MEMSAT-SVM (Nugent and Jones, 2009a) methods which are based on neural network and SVM classifiers, respectively. Both methods use a dynamic programming algorithm to return a list of the most likely topologies returned by overall likelihood and are also capable of predicting the presence of signal peptides and, in the case of MEMSAT-SVM, reentrant helices—membrane penetrating helices that enter and exit the membrane on the same side, common in many ion channel families. The methods were trained using PSI-BLAST (Altschul *et al.*, 1997) profile data generated from the Möller dataset (Möller *et al.*, 2000), in the case of MEMSAT3, or a crystal structure-based training set, in the case of MEMSAT-SVM, and achieved maximum topology prediction accuracies of 78% (Möller set) and 89% (crystal structure set) when fully cross-validated using a jack knife test. The higher fraction of eukaryotic sequences in the Möller set compared with the relative bias toward prokaryotic sequences in the crystal structure set suggest that the strong performance of these two methods makes their combination ideally suited to whole-genome annotation of alpha-helical TM proteins.

3 PREDICTION OF THE OPTIMAL HELICAL PACKING ARRANGEMENT

Despite significant efforts to predict TM protein topology, comparatively little attention has been directed toward developing a method to help users determine possible 3D packing arrangements for helices. Our novel tool MEMPACK (Nugent and Jones, 2009b) uses a range of features to predict residue contacts and helix–helix interactions before using this information to predict the optimal helical packing arrangement. First, an SVM classifier, trained using lipid exposed residue profiles labelled according to molecular dynamics simulation data (Sansom *et al.*, 2008), is used to predict per residue lipid exposure. This information is then combined with PSI-BLAST profile data for each interacting residue and additional sequence-based features as input data for an SVM to predict residue contacts. Combining these results with predicted topology information, helix–helix interactions can then be predicted and used to optimally arrange the helices using a graph-based approach. By employing a force-directed algorithm, the method attempts to minimize edge crossing while maintaining uniform edge length, attributes common in native structures. Finally, a genetic algorithm is used to rotate helices in order to prevent residue contacts occurring across the longitudinal helix axis. Under stringent cross-validation on a non-redundant test set of 74 protein chains, the method achieved 70% lipid exposure and 67% helix–helix interaction prediction accuracy—both significant improvements over existing methods—and was able to produce a helical packing arrangement which closely resembled a 2D slice taken from the crystal structure approximately normal to the likely plane of the lipid bilayer in 14

*To whom correspondence should be addressed.

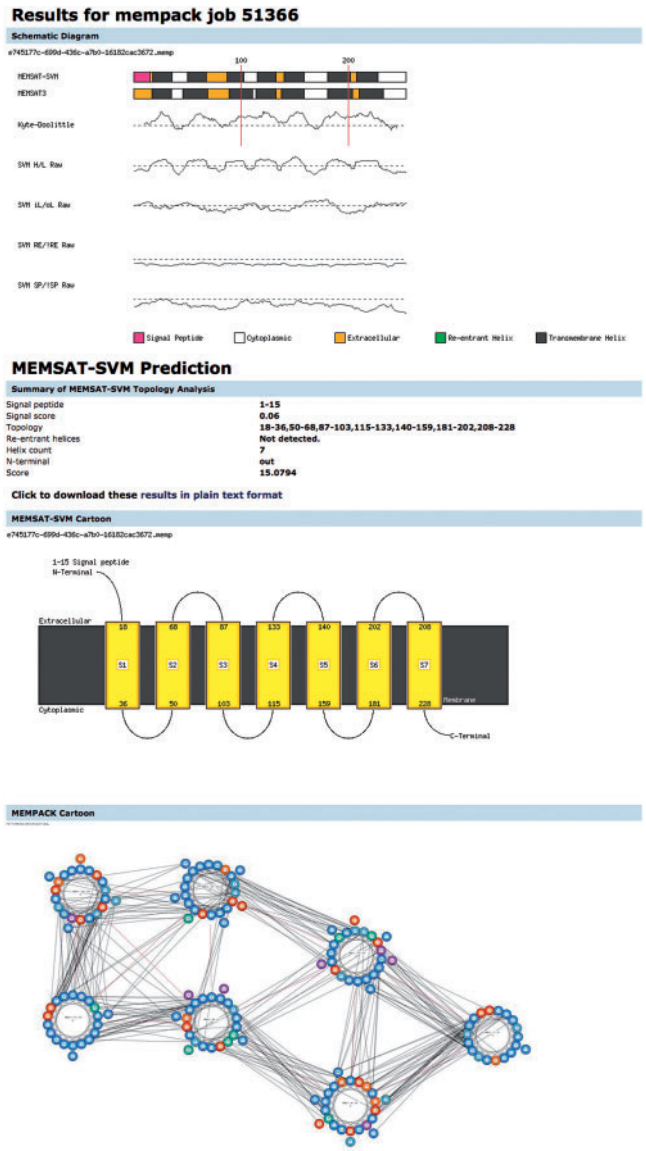

Fig. 1. Sample output for Archaerhodopsin-1, showing predicted transmembrane regions via MEMSAT and MEMSAT-SVM, the MEMSAT-SVM helix orientation cartoon and the predicted helical packing arrangement from MEMPACK. The plots underneath the schematic topology diagram show the raw scores generated by the SVMs that distinguish between TM helices and loop regions (H/L), inside loops and outside loops (iL/oL), reentrant loops or non-reentrant loops (RE/!RE) and signal peptides or non-signal peptides (SP/!SP). Colors in the MEMPACK cartoon indicate hydrophobic residues (blue), polar residues (red) and charged residues (green for negative, purple for positive). Lines between residues indicate a predicted interaction.

out of 23 cases, where all helix–helix interactions were successfully predicted. Of the remaining 51 cases, 34 were partially predicted while 17 had no predicted interactions, highlighting the challenges that remain for helix–helix interaction prediction in TM proteins.

Funding: Part of this work was supported by the BioSapiens project, which is funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health' (contract number LSHG-CT-2003-503265). Funding was also provided by the Biotechnology and Biological Sciences Research Council and the Wellcome Trust (grant number GR066745MA). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the article.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.

Möller,S. *et al.* (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Nugent,T. and Jones,D.T. (2009a) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Nugent,T. and Jones,D.T. (2009b) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput. Biol.*, **6**, e1000714.

Sansom,M.S. *et al.* (2008) Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem Soc. Trans.*, **36**, 27–32.