BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 8 2009, pages 1086–1088
doi:10.1093/bioinformatics/btp102

Structural bioinformatics

iMembrane: homology-based membrane-insertion of proteins

Sebastian Kelm[1,*], Jiye Shi[2] and Charlotte M. Deane[1]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG and [2]UCB Celltech, Branch of UCB Pharma S.A., 208 Bath Road, Slough, SL1 3WE, UK

Received on August 29, 2008; revised on January 7, 2009; accepted on February 17, 2009

Advance Access publication February 23, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: iMembrane is a homology-based method, which predicts a membrane protein's position within a lipid bilayer. It projects the results of coarse-grained molecular dynamics simulations onto any membrane protein structure or sequence provided by the user. iMembrane is simple to use and is currently the only computational method allowing the rapid prediction of a membrane protein's lipid bilayer insertion. Bilayer insertion data are essential in the accurate structural modelling of membrane proteins or the design of drugs that target them.

Availability: http://imembrane.info. iMembrane is available under a non-commercial open-source licence, upon request.

Contact: kelm@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online and at http://www.stats.ox.ac.uk/proteins/resources.

1 INTRODUCTION

Membrane proteins constitute ∼30% of all known proteins and are one of the largest classes of drug targets. They have roles in a multitude of biological processes such as cell recognition and neurotransmitter transport (Müller *et al*., 2008). Unfortunately, they are extremely hard to purify and crystallize, making experimentally determined structures rare. Current computational structure prediction methods are also not ideal, as they are designed to work on globular, soluble proteins.

However, even if a membrane protein's structure is obtained, whether experimentally or computationally, we still do not hold the whole solution to the problem: the protein's position within the lipid bilayer remains unknown. Natural ligands or drugs must be able to access the part of the protein to which they bind. Therefore, it is important to be able to distinguish the parts of the protein that are within the lipid bilayer from those that are solvent-accessible. This information is not currently available from experiments. Structures obtained by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy do not reflect the protein's native lipid bilayer environment.

There are several sequence-based methods to predict the position of transmembrane (TM) helices (e.g. TMHMM, Krogh *et al*., 2001) and β-barrels (e.g. HMM-B2TMR, Martelli *et al*., 2002). For reviews see Cuthbertson *et al*. (2005) and Bagos *et al*. (2005). The boundaries of putative TM helices or sheets tend to be predicted inaccurately and vary between different prediction methods. Half-helices, which span only a part of the membrane, are also hard to predict with existing tools. More importantly, all the above methods use a simple two-state membrane model (in membrane/not in membrane), occasionally with the addition of an uncertainty margin around the prediction. None of the available tools provides a detailed prediction of each residue's position within the lipid bilayer, or its contacts with the different regions of the membrane lipids.

There are some structure-based methods, which predict a protein's position within the membrane. These usually model the membrane as a hydrophobic slab, delimited by parallel planes (e.g. OPM, Lomize *et al*., 2006). The position of these planes is determined by using an energy function, which takes physical and/or statistical properties of amino acid residues as arguments.

In contrast to these largely simplified models, a recently developed method (Scott *et al*., 2008) uses coarse-grained molecular dynamics (MD) simulations in order to better account for the complexity of the lipid bilayer. Protein X-ray structures are simulated in the presence of membrane lipids, which self-assemble into a lipid bilayer. Simulation results include a summary listing the fraction of time each residue spent in contact with the different parts of the membrane lipids (polar head groups or hydrophobic lipid tails). A growing number of these simulation results are being made available online, in the Coarse-Grained database (CGDB, http://sbcb.bioch.ox.ac.uk/cgdb/). CGDB currently contains over 228 lipid bilayer self-assembly simulations for 138 PDB proteins covering 101 SCOP families, 90 superfamilies and 58 folds.

Performing MD simulations—even coarse-grained ones—requires large amounts of time and processing power. In this article, we present iMembrane, a simple method allowing the projection of the existing simulation results onto proteins of homologous structure or sequence. We show that these projected results do not vary greatly from those obtained in original coarse-grained simulations. Where performing an original simulation would take days on a compute server, our method takes mere seconds on a modern desktop computer. In addition, we are able to apply our method to proteins where only sequence information is available.

Here we use CGDB as our dataset. However, our method could theoretically be applied to any database of MD simulation results. Additional datasets will be included in future releases of iMembrane.

2 ALGORITHM

iMembrane accepts either a sequence, in FASTA format (Pearson, 1990), or a structure, in PDB format (Berman *et al*., 2000), as input.

*To whom correspondence should be addressed.

In the case of a structure, its sequence is first extracted from the ATOM coordinates of the structure file. Typically, a BLAST (Altschul *et al*., 1990) sequence search is now carried out against the CGDB of membrane proteins. Matches are then re-aligned to the query using either MUSCLE (Edgar, 2004) sequence alignment or MAMMOTH (Ortiz *et al*., 2002) structure superposition. These alignments are then annotated using the CGDB protein's simulation results. A flow diagram of the iMembrane algorithm, including alternative search methods, is available in the Supplementary Material.

A residue's annotation is provided as a single letter per residue: N (not in contact with the membrane), H (in contact with the polar head groups of the membrane lipids) or T (in contact with the lipid hydrophobic tails). In the first instance, these letters simply represent an interpretation of the raw simulation results, as provided in the CGDB.

We also provide a simplified model, which abstracts the membrane as a three-layered slab, with an inner region around the membrane lipids' hydrophobic tails, and two peripheral regions surrounding the membrane lipids' polar head groups. The boundaries of these layers are calculated by fitting parallel planes onto the membrane contact data.

This model allows us to then use each residue's 3D coordinates to determine in which layer of the membrane it resides, or whether it is outside the membrane. iMembrane does this automatically for the CGDB proteins and then uses this information to annotate any homologous proteins aligned to them.

In the case where the input to our method is a structure, we can use the same procedure to assign every residue in the query protein to one of the membrane (or non-membrane) layers defined by the aligned CGDB protein. This step is performed in a Pymol environment (DeLano, 2002).

In the case of a sequence-only input, the query's 3D information is missing. Therefore, we can only annotate those residues that are aligned to a CGDB protein's residue. In the future, an additional structure prediction step will be implemented, such that we will be able to annotate every residue of a sequence-only input, as well as give back its proposed structure.

3 VISUALIZATION

We visualize the predicted membrane insertion of the input protein using (i) a colour-annotated sequence alignment and/or (ii) a coloured 3D structure as shown in Figure 1. The sequence-based visualization is always provided, whereas the coloured structure output is currently restricted to the case where the input itself was a structure.

4 ACCURACY

iMembrane's accuracy was tested using a leave-one-out cross-validation on the CGDB data. The prediction results for each hit were compared to the original annotation generated directly from the corresponding MD simulation result in the CGDB. A Q3 score was calculated, representing the fraction of annotated residues assigned to the correct annotation (T, H or N; see Fig. 1). In addition, a Q2 score was calculated by merging the two types of membrane layers (T and H).

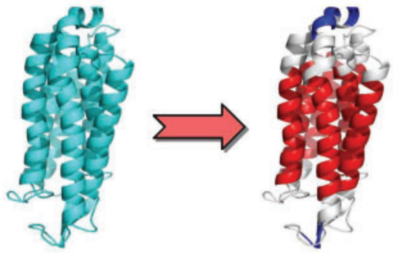

Fig. 1. The structure of PDB entry 2JAF before (left) and after (right) annotation with iMembrane. Shades show the membrane layers. Top to bottom: non-membrane (dark blue), polar head group layer (white), lipid tail layer (dark red), polar head group layer (white) and non-membrane (dark blue).

Independent of the input type (structure or sequence), a sequence identity of >35% tends to result in a Q3 accuracy >70% and a Q2 accuracy of ∼90% and above in the membrane layer prediction. A slight upwards trend can be observed with increasing sequence identity. Below 35% sequence identity, homolog detection and sequence alignment quality is known to decline (Rost, 1999). As our method depends entirely on the alignment between the query and database proteins, its accuracy varies greatly below ∼35% sequence identity, in the case where the input is a sequence. For structure input, this boundary is pushed down to 20% sequence identity. The use of improved alignment methods more suitable for distant homologs will benefit the accuracy of iMembrane in future releases.

A range of accuracy plots can be found in the Supplementary Material.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the Oxford Protein Informatics Group, H. Saadi and M. Sansom for useful discussion and feedback, and also K. Scott and A. Chetwynd for help with the CGDB.

Funding: Biotechnology and Biological Sciences Research Council (to S.K.); University of Oxford Systems Biology Doctoral Training Centre (to C.M.D.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bagos,P.G. *et al*. (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, **6**:7.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–42.

Cuthbertson,J.M. *et al*. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel*., **18**, 295–308.

DeLano,W.L. (2002) The PyMOL molecular graphics system. Available at http://www.pymol.org (last accessed date 15 February, 2008).

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.

Lomize,M.A. *et al*. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Martelli,P.L. *et al*. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18** (Suppl. 1), S46–S53.

Müller,D.J. *et al*. (2008) Vertebrate membrane proteins: structure, function, and insights from biophysical approaches. *Pharmacol. Rev.*, **60**, 43–78.

Ortiz,A.R. *et al*. (2002) MAMMOTH (Matching Molecular Models Obtained from Theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Scott,K.A. *et al*. (2008) Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure*, **16**, 621–630.