Poster H-9

Improved membrane protein topology prediction by domain assignments

Authors:
Andreas Bernsel (*Stockholm Bioinformatics Center, Stockholm University*)
Gunnar von Heijne (*Department of Biochemistry and Biophysics, Stockholm University*)

Short Abstract: We have identified a set of domains that, when found in soluble proteins, have conpartment-specific localization of a kind relevant for membrane protein topology prediction. Using these domains as prediction constraints, we are able to provide high-quality topology models for 11% of the membrane proteins extracted from 38 eukaryotic genomes.

Long Abstract:
Alpha-helical transmembrane proteins constitute about 20% of all proteins encoded by most genomes (Krogh et al. 2001) and are responsible for several vital processes in the cell. In addition, the medical importance of membrane bound receptors, channels, and pumps as targets for drugs is well established. Still, for the large majority of membrane proteins, the structure or even the topology, i.e., the positions and in/out-orientations of all transmembrane helices, is not known experimentally. The continuously growing amount of sequence data, in combination with the limited amount of structural data available, highlight the need for better and more accurate theoretical structure prediction methods, particularly for the annotation of membrane proteins.

 Protein domains are modular, independently evolving, and structurally similar amino acid segments, which may exist alone in single-domain proteins, or may combine to form multi-domain proteins. Although covalent combinations between transmembrane domains, i.e., domains with one or more membrane spanning regions, rarely occurs, covalent combinations between soluble domains and transmembrane domains are observed frequently (Liu et al. 2004). Moreover, domains are often compartment-specific, and information about domain occurrence can be used to predict the subcellular localization of soluble proteins (Mott et al. 2002).

 Here, we explore the possibility that the presence of compartment-specific extra-membaneous protein domains in transmembrane protein sequences might be used as a constraint in a subsequent topology prediction step, in much the same way that experimentally determined “anchor points” have been used to constrain topology predictions (Kim et al. 2003; Rapp et al. 2004; Daley et al. 2005). Unconstrained topology predictions are correct for only ~55-60% of all membrane proteins (Melén et al. 2003), while compartment-specific domains that are always located on just one side of a membrane (facing, e.g., the extracellular space or the cytosol) can be identified with high reliability. If such a domain is found in a membrane protein, that particular segment in the protein sequence can be fixed to the corresponding side of the membrane before applying a sequence-based topology prediction algorithm on the rest of the sequence. Here, we show that domains of this kind are found in at least 11% of many eukaryotic proteomes, and that a significant improvement in topology prediction can be achieved by using these domains as prediction constraints.

 Our basic approach consists of three steps:

 - Domain selection. Identify compartment-specific domains that always reside on either the in- or outside of the membrane. Each domain is represented by a profile Hidden Markov Model (HMM). In general, we considered domains annotated as “extracellular” in the

SMART 4.0 domain database to reside outside of the membrane (i.e. on the non-cytoplasmic side), and domains annotated as “signaling” to reside on the inside of the membrane (i.e. on the cytoplasmic side), which is in agreement with e.g. (Mott et al. 2002). In an attempt to assess the validity of this assumption, the domains were assigned to 297 homology reduced sequences of membrane proteins with experimentally known topologies. This resulted in 48 domain hits, contained in 29 (10%) of the sequences. Out of all domain hits, 47 (98%) were in agreement with the topology. One domain was in conflict with a known topology, and was thus removed from the domain collection. Although the test set is small, we consider our domain collection as highly reliable.

 - Domain assignment. The final domain list used for placing constraints on the topology predictions consisted of 367 domains, of which 146 were “IN-domains” (i.e. appear only on the cytoplasmic side of the membrane), and 221 were “OUT-domains” (i.e. appear only on the non-cytoplasmic side of the membrane). For each query sequence, we now try to find one or more of the domains and fix those residues to the corresponding side of the membrane.

 - Topology prediction. In the last step, we use a sequence-based method to predict the topology of the remaining part of the protein sequence, with the domain(s) found in the previous step constrained to either the in- or outside of the membrane.

 Based on the constrained predictions, the topologies of all proteins containing at least one soluble domain were analyzed. 66% of those were single-spanning proteins, compared to just 37% in the complete set of predicted membrane proteins, suggesting that our method will have particular impact on single-spanning proteins. Single-spanning proteins are often mis-predicted by the current topology prediction methods, mostly due to an inversion of the predicted topology such that the TM-segment is correctly located but the overall orientation is wrong. Large extra-membraneous domains carry little or no orientational information in the current predictors, and our domain-based method thus solves a major weakness in these methods.

 References:

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. (2001) Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes. J Mol Biol 305: 567-580.

 Liu, Y., Gerstein, M., and Engelman, D.M. (2004) Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. Proc Natl Acad Sci U S A 101: 3495-3497.

 Mott, R., Schultz, J., Bork, P., and Ponting, C.P. (2002) Predicting protein cellular localization using a domain projection method. Genome Res. 12: 1168-1174.

 Kim, H., Melén, K., and von Heijne, G. (2003) Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and prediction. J Biol Chem 278: 10208-10213.

 Rapp, M., Drew, D.E., Daley, D.O., Nilsson, J., Carvalho, T., Melén, K., de Gier, J.W., and von Heijne, G. (2004) Experimentally based topology models for E. coli inner membrane proteins. Prot Sci 13: 937-945.

 Daley, D.O., Rapp, M., Granseth, E., Melén, K., Drew, D., and von Heijne, G. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. Science: 308:1321-1323.

 Melén, K., Krogh, A., and von Heijne, G. (2003) Reliability measures for membrane protein topology prediction algorithms. J Mol Biol 327: 735-744.