

Proteomics: Predicting Proteins Structure ¹

Lisete Sousa ²

Dept. of Statistics and Operational Research and Center of Statistics and Applications,
Lisbon University, e-mail: lmsousa@fc.ul.pt

Abstract: As the Genomic era comes to maturity with the full sequencing of entire genomes, the challenge ahead consists in understanding how genes originate the many proteins with diverse structures and functions present in the make up of living organisms. This new focus ushers in the era of Proteomics, i.e. the systematic study of the diverse properties of proteins and ultimately the characterization of an organisms Proteome. The complexity of this new field has provided mathematicians with the opportunity to develop statistical methods to predict the structure and possible function of proteins. This work focuses essentially on the prediction of transmembrane protein topology. State of the art is reviewed and a novel method is presented.

Keywords: Proteomics, transmembrane protein topology, hidden Markov models, Gibbs sampling

1. State of the Art

Transmembrane proteins play a central role in cellular biology and are responsible for a great variety of human diseases. Currently, there is a need to produce and publish novel methods for predicting the topology of transmembrane proteins, so that biological databanks exhibit a high-quality biochemical annotation and that investigators have access to alternative methods.

Early methods for prediction of transmembrane helices used hydrophobicity analysis alone. Indeed some helices can be located with high reliability from a hydrophobicity plot, but others cannot. Another signal shown to be associated with transmembrane helices is the abundance of positive charged residues in the part of the sequence on the cytoplasmic side of the membrane, the positive-inside rule. By combining charge bias with hydrophobicity analysis, better predictions can be obtained. Examples are Toppred2, MEMSAT and TMAP. Despite this relatively consistent topogenic signal, correct prediction of the location and orientation of all transmembrane segments has been proved to be a difficult problem. Local methods like TMPred, DAS and PHD have as main weakness a lack of specificity for membrane proteins.

HMMTOP and TMHMM are global approaches, which determine statistically the most probable topology for the whole topology according the underlying (hidden Markov) model. Möller et al. (2001) published the first independent evaluation of the performance of some existing tools, and showed that TMHMM was, at the time, the best performing transmembrane prediction program. Generally these methods are benchmarked on a set

¹The present paper is financially supported by the projects PTDC/MAT/64353/2006 and FCT/POCI2010.

²Address of correspondence: FCUL, Bloco C6 - Piso 4, Campo-Grande 1749-016 Lisboa, Portugal

of proteins with experimentally verified topology. Käll and Sonnhammer (2002) investigated if the accuracy measured on such data sets can be expected in an unbiased genomic analysis or if there is a bias towards 'easily predictable' proteins in the benchmark data sets. The benchmark data set showed significantly higher levels of agreement between different methods than in 10 tested genomes. All these methods are not accurate enough, as they have at most 50-60% accuracy in terms of the prediction of whole TM topology. In order to improve this accuracy, more recently, Arai et al. (2004) developed a prediction server of transmembrane topology based on a consensus approach by combining the results of several existing methods (TMpred, TMAP, MEMSAT, HMMTOP, TMHMM, DAS, and other more) in two steps: ConPred_elite and ConPred_all. The authors just refer that the prediction accuracy of ConPred_elite is almost 100%, although is achieved by sacrificing the prediction coverage (20-30%), and that ConPred_all improves the prediction accuracy up to 11%.

Topology prediction of transport proteins located in the plasma membrane is also a matter of study. These membrane transporters have a defined membrane topology with transmembrane helices and loop segments that usually are placed outside the membrane. There is, however, increasing evidence that some of the loops have "intramembrane segments", i.e., reenter the membranes as "reentrant loops" or even cross the membrane as a nonhelical intramembrane loop segment. Currently the most used methods to predict transmembrane topology of transport proteins systems do not allow nonhelical intramembrane segments. More recently, Viklund et al. (2006) constructed TOP-MOD, a simple method based on hidden Markov models that focuses on predicting reentrant regions. Lasso et al. (2006) implemented TMLOOP to predict membrane dipping loops in polytopic membrane proteins. TMLOOP applies discovered patterns as weighted predictive rules in a collective motif method. BETH is a flexible model which can be adapted to predict intramembrane loop segments, specifically the topology of SGLT1 (Raja et al., 2003; Sousa et al., 2004).

2. New Approach Arising

An HMM-based method, named BETH, is presented here (Sousa, 2004). BETH is ambitious in the sense that it does not incorporate aminoacids properties and does not map closely to the biological system as for TMHMM. However, BETH incorporates informative prior knowledge about meaningful quantities such as the length of helix, in and out segments, which is a big advantage compared to other existing methods. For comparison with other methods, BETH was applied on the test set of 160 transmembrane proteins used by Sonnhammer et al. (1998), revealing an accuracy in predicting the correct topology of 71.9%. This percentage increases up to 84% if only small proteins (less than 200 aminoacids) are analysed. Concerning the total of 696 transmembrane helices (in the test set) BETH accuracy is even better, 96.1%. On a set of 40 soluble proteins BETH only predicted wrongly one as being a transmembrane protein. No transmembrane proteins, in the test set, were predicted as soluble proteins. This shows that BETH discriminates well between transmembrane and non-transmembrane proteins.

BETH has essentially two handicaps. First, the high percentage of helices falsely merged. Approximately 25% of the transmembrane helices were merged in pairs. The second handicap is the orientation of the helices, frequently predicted as running to the same side from where it is coming. The orientation problem motivated the development of

a more sophisticated model which consisted of four states instead of three to characterize the segments location (two states for transmembrane segments). Improvements in the original model have been made, namely, considering a non-homogeneous Markov chain and incorporating prior knowledge about some parts of the protein (Dias, 2006). At the moment, a web-based tool is being developed (BETHWeb) to make BETH algorithm available online (Dos Santos, 2008).

3. Statistical Methodology

Hidden Markov models have been used to predict transmembrane protein topology (Sonhammer et al., 1998; Krogh et al., 2001). Although transmembrane and intracellular/extracellular domains cannot usually be modeled as a string of independent letters, in many cases taking a first-order Markov dependence between aminoacids is a sufficiently complex initial model for capturing much of the structure that is believed to be present in transmembrane proteins. The role of the hidden Markov model is to provide a mechanism for modeling the type of region (transmembrane, inside, outside) at each location in the sequence. Crucially, the model for base transitions depends on the hidden label process and the analysis must allow for the fact that these data are missing. For much of their history, hidden Markov models have been implemented using recursive algorithms developed for parameter estimation and for restoring the hidden Markov chain. However, Bayesian methods can be applied to these subjects using a hidden Markov model implemented by Markov chain Monte Carlo methods (Boys et al., 2000)

BETH adopts a Bayesian perspective and prior information is incorporated into a procedure for identifying transmembrane regions. Inferences are made by simulating from the posterior distribution using Markov chain Monte Carlo methods and, in particular, Gibbs sampling with data augmentation. The advantages of this approach are that it provides a relatively straightforward framework for incorporating prior information and permits the detection of segment types allowing for posterior parameter uncertainty.

The original version of BETH (Sousa, 2004), considers only three states: one for inside loops, one for transmembrane regions and one for outside loops. In fact, the goal here is to detect transmembrane regions and the overall topology of the protein, therefore it is just taken into account if an aminoacid is in the membrane or inside/outside the cell. Obviously, chemical factors (hydrophobicity, charge, etc.) are not explicitly included in a model containing only three states. Nevertheless, the model captures all of this information, which is implicitly contained in the aminoacids.

Dias (2006) improved BETH through the adoption of a non homogeneous Markov model, sensible to the hydrophobicity and the hydrophobic region variables. These conditioning variables allowed the determination of a prior distribution for the parameters representing the transitions between topologies based upon the characteristics of the protein under study. He also proposed three distinct strategies of restricting the dimension of transmembrane helices: (1) using the hydrophobicity variable; (2) considering a transition scheme with 72 topologies; (3) adopting a counting variable. It was also shown how the knowledge of part of the protein topology can be incorporated in the model.

We transformed BETH, considering four states, and a non-homogeneous Markov chain, to model this specific intramembrane segments. Again, a Bayesian perspective is adopted through the hidden Markov model and prior information is incorporated into the procedure for identifying both kinds of transmembrane domains. The model is used for

predictions on the sodium-D-glucose cotransporter SGLT1 where preliminary experimental evidence for an intramembrane segment (of the C-terminal loop 13) exists (Raja et al., 2003; Sousa et al., 2004).

References

- Arai M., Mitsuke H., Ikeda M., Xia J.-X., Kikuchi T., Satake M. and Shimizu T. (2004) ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Research* 32, W390–W393.
- Boys R.J., Henderson D.A. and Wilkinson D.J. (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov Models. *Applied Statistics* 49, 269–285.
- Dias P. (2006) *Predição da Topologia de Proteínas Transmembranares: Modelos de Markov Escondidos No Homogêneo sob uma Perspectiva Bayesiana*. MSc. Thesis, Lisbon University.
- Dos Santos S. (2008) *Desenvolvimento de Sistema online para a Predição da Topologia de Proteínas Transmembranares: Algoritmo BETH*. MSc. Thesis, Lisbon University.
- Käll L. and Sonnhammer E. (2002) Reliability of transmembrane predictions in whole-genome data. *Federation of European Biochemical Societies Letters* 532, 415–418.
- Krogh A., Larsson B., von Heijne G. and Sonnhammer E. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* 305, 567–580.
- Lasso G., Antoniw F.J. and Mullins G.L.J. A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics* 22, e290297.
- Möller S., Croning M.D.R. and Apweiler R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653.
- Raja M.M., Tyagi N.K. and Kinne R.K.H. (2003) Phlorizin recognition in a C-terminal fragment of SGLT1 studied by tryptophan scanning and affinity labeling. *The Journal of Biology and Chemistry* 278 (49), 49154–49163.
- Sonnhammer E.L.L., von Heijne G. and Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow et al., eds., AAAI Press, Calif., 175–182.
- Sousa L. (2004) *Development of statistical methods for the analysis of molecular genetic data*. PhD Thesis, Lisbon University.
- Sousa L., Althof T., Raja M.M., Kinne R. and Urfer W. (2004) Prediction of Nonhelical Intramembrane Segments in Membrane Proteins by a Bayesian Methodology. Technical Report 4/2004, Dortmund University.
- Viklund H., Granseth E. and Elofsson A. (2006) Structural classification and prediction of reentrant regions in non-helical transmembrane proteins: application to complete genomes. *Journal of Molecular Biology* 361, 591–603.