

Higher order HMMs for Localization Prediction of Transmembrane Proteins

Stefan Maetschke

Mikael Bodén

Marcus Gallagher

School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, Queensland 4072, Australia
Email: stefan@itee.uq.edu.au

Abstract

Utilizing the recently published LOCATE database, we construct Hidden Markov Models (HMMs) of first, second and third order for subcellular localization prediction of transmembrane proteins. In comparison with linear Support Vector Machines (SVMs), based on overall amino acid and di-peptide composition, higher order HMMs show a significant increase in prediction performance. The best performance was achieved by a second order HMM with a correlation coefficient of 0.46. A web-service for localization prediction of transmembrane proteins has been made available at <http://pprowler.itee.uq.edu.au/TMPHMLoc>.

Keywords: HMM, SVM, subcellular localization, transmembrane protein

1 Introduction

Transmembrane proteins are inserted into the membranes of organelles and perform a variety of essential functions, such as channels, pumps, receptors and energy transducers. Current predictors for subcellular localization however, primarily target soluble proteins and ignore the characteristic topological domains of transmembrane proteins. On the other hand, topology predictors such as TMHMM (Sonnhammer, von Heijne & Krogh 1998, Krogh, Larsson, von Heijne & Sonnhammer 2001), Phobius (Käll, Krogh & Sonnhammer 2004) or HMMTOP (Tusnády & Simon 2001) are not designed for subcellular localization prediction.

Inspired by topology prediction methods, we construct a novel Hidden Markov Model (HMM) architecture for subcellular localization prediction of transmembrane proteins and compare it against two standard approaches for localization prediction of soluble proteins. More specifically, we 1) introduce the architecture and parameter estimation of the HMM, 2) measure the prediction accuracy and computation times of first, second and third order HMMs, 3) and compare the HMMs with linear Support Vector Machines (SVMs) that exploit overall amino acid and di-peptide composition as input. We utilize the recently published LOCATE database (Fink, Aturaliya, Davis, Zhang, Hanson, Teasdale, Kai, Kawai, Carninci, Hayashizaki & Teasdale 2006) and focus our comparison on five locations along the secretory pathway in mouse.

We thank Melissa Davis for many helpful discussions concerning transmembrane protein localization. This work was supported by the Australian Research Council Centre for Complex Systems.

2 Transmembrane proteins

Transmembrane proteins contain α -helical domains of hydrophobic residues that anchor the protein in the membrane. The *transmembrane domains* are usually flanked by *cap regions* that show a preference for charged residues and influence the orientation of the α -helix relative to the membrane (see Fig. 1). The more positively charged cap region of the transmembrane domain tends to reside on the cytosolic side (positive inside rule (von Heijne 1986)).

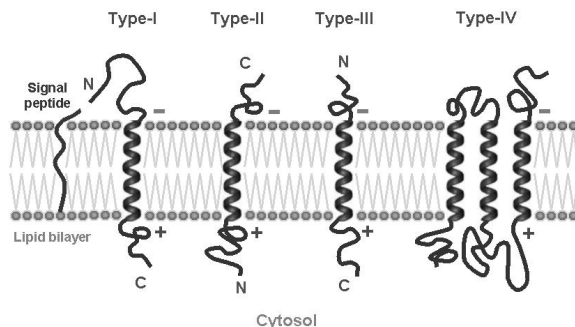


Figure 1: Transmembrane protein inserted into the lipid bilayer. The transmembrane domains form α -helices and the cap regions display a preference for charged residues (marked with plus and minus signs).

Four different types of transmembrane proteins can be distinguished¹ (Higy, Junne & Spiess 2004). Type-I proteins carry an N-terminal signal peptide which is cleaved when the protein is inserted into the membrane (von Heijne 1990). The N-terminus of the mature protein is at the luminal or extracellular side and the C-terminus is at the cytoplasmic side. The orientation of Type-III is the same as Type-I proteins, whereas Type-II proteins are reversed. Multi-spanning proteins (Type-IV) pass the lipid bilayer several times with their termini on either side of the membrane (Rapoport, Goder, Heinrich & Matlack 2004).

Transmembrane proteins are localized to almost all compartments in the cell. We focus our study on organelles along the secretory pathway (see Fig. 2). The secretory pathway is especially complex due to its dynamic localization process that requires transmembrane proteins to travel through several stations until they reach their final destination (van Vliet, Thomas, Merino-Trigo, Teasdale & Gleeson 2003).

Entry station to the secretory pathway is the endoplasmic reticulum (ER). Transmembrane proteins are cotranslationally inserted into the ER membrane

¹Note that this is a simple classification scheme that ignores important, but less frequent subtypes, such as reentrant regions in α -helical transmembrane proteins (Viklund, Granseth & Elofsson 2006).

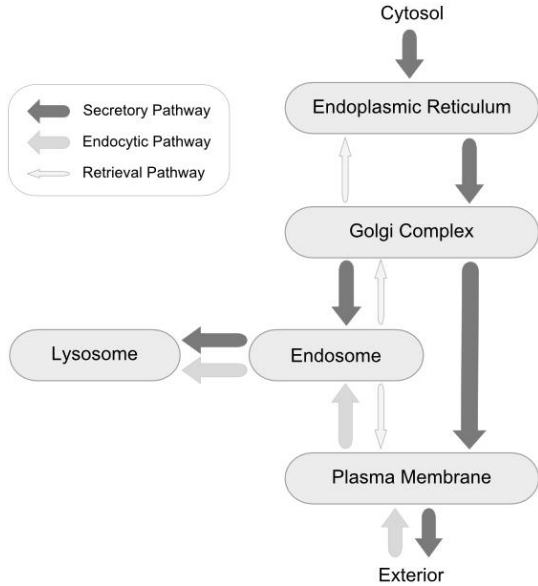


Figure 2: Schema of the secretory and endocytic pathways. The secretory pathway is directed from the interior of the cell to the exterior. The direction of the endocytic pathway is reversed.

and N-terminal signal peptides sequences are cleaved at this stage. Further transport occurs in vesicles that bud from the ER membrane and fuse with the Golgi complex (GO). At the Golgi complex, proteins are packed into coated vesicles and transported to the plasma membrane (PM) or the endosome (EN). From the endosome vesicles move proteins to the lysosome (LY). Also an indirect route exists, where proteins are exocytosed first and then internalized again, following the endocytic pathway. Additional retrieval pathways transport escaped proteins back to their original target location (van Vliet et al. 2003).

3 Related work

A multitude of prediction algorithms for protein subcellular localization have been developed. The vast majority of them is limited to soluble proteins however. We will discuss only a subset of the more recent algorithms that are related to our work.

Apart from methods that search for homologous or similarly annotated proteins in databases, the majority of current predictors exploit the amino acid or di-peptide composition and utilize SVMs to derive subcellular localization (Hua & Sun 2001, Park & Kanehisa 2003, Cui, Jiang, Liu & Ma 2004, Yu, Mendrola, Audhya, Singh, Keleti, DeWald, Murray, Emr & Lemmon 2004).

Composition based algorithms basically neglect the residue order of the sequence. To alleviate this weakness, autocorrelation functions (Feng & Zhang 2001), the pseudo amino acid composition (Chou 2001, Zhou & Doctor 2003) and the residue-coupling model (Guo, Lin & Sun 2005) have been applied.

A related approach is the partitioning of the protein sequence into sections (e.g. N-terminal, middle section, C-terminal) and the evaluation of section specific features such as amino acid composition and physicochemical properties (Small, Peeters, Legeal & Lurin 2004, Cui et al. 2004, Matsuda, Vert, Saigo, Ueda, Toh & Akutsu 2006). Yuan (1999) modeled the amino acid sequence directly with Markov chain models.

None of the aforementioned algorithms however, model the characteristic membrane spanning regions

or consider the orientation of transmembrane proteins as topology predictors such as TMHMM (Krogh et al. 2001), Phobius (Käll et al. 2004) or HMMTOP (Tusnady & Simon 2001) do. The latter utilize detailed first order HMMs to describe the transmembrane, cap and loop regions but are not designed for subcellular localization prediction. The differences between topology prediction methods and our approach will be discussed in more detail in Section 6.

The only predictor for eukaryotic membrane proteins that we are aware of is based on amino acid composition and employs a least Mahalanobis distance classifier (Chou & Elrod 1999). A data set with 2105 membrane proteins extracted from Swiss-Prot (Release 35.0) with nine different locations was used and an overall jackknife accuracy of 65.9% was reported.

Since the data set was only weakly redundancy-reduced and contained different types of membrane proteins, these results are not comparable with ours. We compiled a strictly redundancy reduced, more recent data set, that contains transmembrane proteins only.

4 Data set

All predictors were trained and tested on protein data extracted from the LOCATE² database (Fink et al. 2006). LOCATE is based on the mouse transcriptome of the FANTOM3 Isoform Protein Sequence set (IPS7), enriched by membrane organization and subcellular localization annotation.

Membrane organization is determined by *MemO* (Davis, Zhang, Yuan & Teasdale 2006), a consensus method that employs SignalP (Bendtsen, Nielsen, von Heijne & Brunak 2004) and five transmembrane topology predictors (HMMTOP, TMHMM, SVM-TM, MEMSAT, DAS) to predict signal peptides, transmembrane domains, protein orientation and subsequently protein type. Subcellular localization annotation in LOCATE is inferred from sources of varying quality (experimental, literature, predicted) but carefully reviewed.

We downloaded the XML version (`LOCATE_whole_db_v3-060810.xml`) of the database and extracted all transmembrane proteins with a unique subcellular localization annotation. The dataset was then filtered for proteins targeted to locations along the secretory pathway. Redundancy reduction was performed with BlastClust (Altschul, Gish, Miller, Myers & Lipman 1990), which removed all entries with a sequence similarity greater than 25%. The final data set contained 1351 transmembrane proteins with the following distribution: 873 plasma membrane (PM), 261 endoplasmic reticulum (ER), 141 Golgi apparatus (GO), 45 lysosome (LY), 31 endosome (EN).

5 Hidden Markov Models

A HMM is composed of a set of states $\{S_1, S_2, \dots, S_N\}$ with transition probabilities a_{ij} . In the discrete case each state emits symbols v_k from a finite symbol set $V = \{v_1, v_2, \dots, v_M\}$. The state transition probability distribution $\mathbf{A} = \{a_{ij}\}$ with $1 \leq i, j \leq N$, is defined as the probability that the model changes to state S_j given that it was in state S_i ,

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad (1)$$

where q_t describes the state the model occupies at time step t . The symbol emission probabilities $\mathbf{B} =$

²<http://locate.imb.uq.edu.au>

$\{b_j(v_k)\}$ are the probabilities that symbol v_k is emitted (or observed) when the model is in state S_j with

$$b_j(v_k) = P(o_t = v_k | q_t = S_j), \quad (2)$$

and o_t is the observed symbol at time step t (Durbin, Eddy, Krogh & Mitchison 1998).

Maximum likelihood estimates for state and emission probabilities can be directly calculated from labeled observation sequences or, for unlabeled data, gained in an unsupervised fashion utilizing a variant of the EM-algorithm (Baum-Welch). The most probable state sequence through the model is usually determined with a dynamic programming approach using the Viterbi-algorithm (Durbin et al. 1998).

In our domain, states describe sections of the protein sequence. For a first order HMM, V becomes the amino acid alphabet and t is a specific position within the sequence. Higher order HMMs are readily created by redefining V as an alphabet over pairs (second order) or n -tuples (n -th order) of amino acids (Durbin et al. 1998), and a protein is then processed as a sequence of overlapping, consecutive pairs or tuples of amino acids.

6 Localization predictor

The construction of the localization predictor can be divided into three phases. The first phase is the sequence labeling phase. The second phase is the construction of transmembrane protein models for each subcellular location based on the labeled sequences. In the third phase the protein models are aggregated in a localization model. In the following the three phases will be described in more detail.

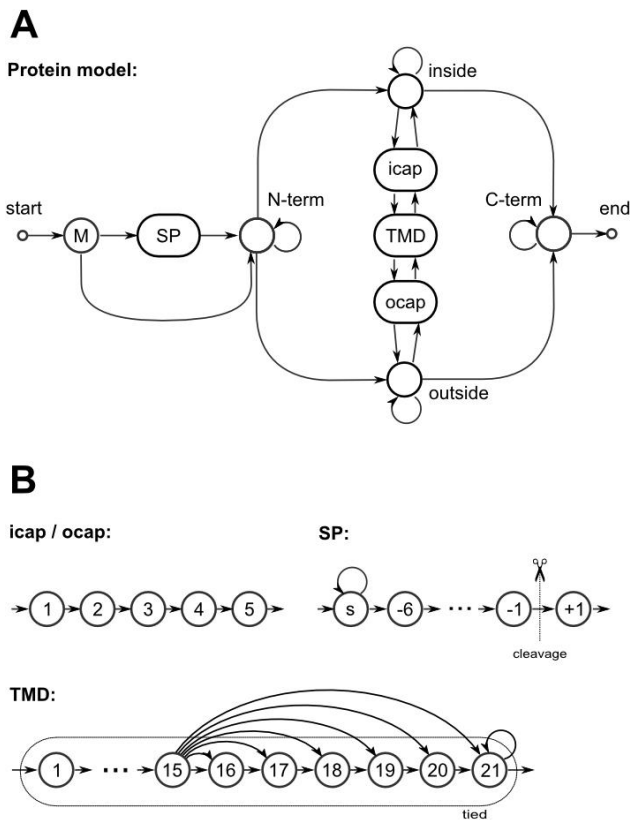


Figure 3: Prediction system. A) Transmembrane protein model for a single location. B) Details of the components of protein model A.

During the first phase every residue of the sequences in the training set is labeled with a state label

of the protein model to construct (see Fig. 3). Labels are derived from existing sequence annotations such as transmembrane domains or signal peptides.

The first residue of the sequence is always labeled as Methionine (M). In the presence of a signal peptide annotation the following residues are labeled as signal peptide states. The position downstream of the annotated cleavage site is labeled +1 and the adjacent six residues upstream are labeled -1 to -6. The remaining upstream residues are all labeled as signal peptide residues s (See SP model in Fig. 3).

The ten amino acids following the signal peptide or the Methionine state are labeled as N-terminal (N-term). The transmembrane region is labeled with 15 up to 21 distinctive state labels (according to the length of the annotated region). The emission probabilities of these states are *tied* (each state uses the same emission probability distribution, See TMD model in Fig. 3). The five residues upstream and downstream of the transmembrane domain are labeled as inside (icap) or outside (ocap) regions, represented by five states (See icap/ocap model in Fig. 3).

The last ten residues of the sequence are labeled as C-Terminal (C-term) and all remaining amino acids are marked as *inside* or *outside* residues. The membrane orientation (N-terminus inside or outside) of the protein, which is required to label inside and outside residues and cap regions, is determined according to the presence or absence of an annotated signal peptide (a signal peptide indicates a non-cytosolic N-terminal). We also used the orientation annotation provided by the topology predictors in LOCATE but found it to result in lower prediction performance (data not shown). Likewise a fixed orientation (e.g. N-terminal always outside) was found to be inferior.

In phase two the labeled sequences are grouped according to the annotated subcellular localization. For each group a HMM is constructed. The model states are directly given by the used label set. Maximum likelihood estimates for emission and transition probabilities are derived from the frequencies of state residues and state transitions in the labeled sequences in the same way Profile-HMMs are built (Durbin et al. 1998). We also calculated the model parameters utilizing the unsupervised Baum-Welch algorithm (Durbin et al. 1998) but found the resulting prediction performance inferior to the supervised approach (data not shown).

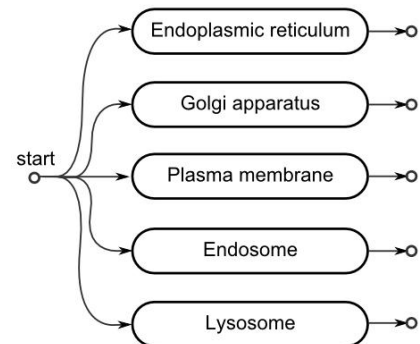


Figure 4: Aggregation of protein models within an overall HMM.

In phase three the transmembrane protein models, constructed in phase two, are aggregated in a single HMM with a unique start state but multiple end states (See part C of Fig. 4). Classification is performed by determining the Viterbi-path of the query sequence through the model and predicting the subcellular localization according to the end state of the

most probable path.

Note that the protein models are smaller (less states and parameters) than similar models employed for topology prediction (Krogh et al. 2001, Käll et al. 2004, Tusnády & Simon 2001). There are three motivations for this: 1) The annotation of transmembrane regions and signal peptides in the training data is predicted, not experimentally confirmed. An overly refined model would only predict predicted data with high accuracy. 2) The objective is to predict subcellular localization, not topology. The exact domain borders are therefore of secondary interest. 3) The data sets for some locations (e.g. endosome, lysosome) are very small and parameters for more complex models cannot be estimated reliably.

7 Results

Many algorithms for subcellular localization prediction of *soluble proteins* are based on SVMs that exploit the overall amino acid or di-peptide composition of a protein as input (Hua & Sun 2001, Park & Kanehisa 2003). We therefore compare the prediction accuracy, and training and query time, for two composition based SVMs with HMMs of varying order. In the following, SVM1 denotes a linear SVM that exploits the amino acid composition and SVM2 is a linear SVM that utilizes the di-peptide composition. HMM1, HMM2 and HMM3 refer to first, second and third order HMMs, respectively.³

The results in Table 1 show a significant increase in prediction accuracy of higher order HMMs compared to first order HMMs or SVMs. Notably the correlation coefficient of SVM1 is a magnitude smaller than that of SVM2. This suggests that the di-peptides composition is a much better representation of the typical sorting signals (e.g. ER retrieval signal K(X)KXX or lysosomal/endosomal di-leucine targeting signal) than the mono amino acid composition.

Concerning training and query time, the HMMs are fast to train but slow to query while the situation for the SVMs is reversed. The training time for the third order HMM is surprisingly high. We believe that the physical memory (1GB) was not sufficient and memory swapping took place in this case.

Classes with small numbers of training samples, such as the endosomal (EN) and lysosomal (LY) classes, cause a clear drop in prediction performance for higher order HMMs. SVM2, as a maximum margin classifier, is less effected by this difficulty, while the performance of SVM1 is poor in general. There is no significant difference in prediction performance between second and third order HMMs but the second order model features lower query times and memory requirements.

	PM	ER	GO	EN	LY	
834	25	11	3	0	PM	
125	126	8	1	1	ER	
63	22	54	0	2	GO	
21	0	1	9	0	EN	
28	4	1	0	12	LY	

Table 2: Ten-fold cross-validation confusion matrix for second order model (HMM2). Rows represent observed locations and columns represent predicted locations.

³Note that second and third order HMMs utilize the same architecture as described above but observe amino acid pairs or triples instead of single amino acids.

To gain a deeper insight into the prediction performance of the second order HMM2, we calculated the ten-fold cross-validation confusion matrix (see Table 2). The confusion matrix shows that most of the misclassified proteins are predicted as targeted to the plasma membrane (left most column). This is not surprising, since the plasma membrane class is the majority class. Also the plasma membrane is known to serve as a default location for proteins that lack specific sorting signals (Pedrazzini, Villa & Borgese 1996, Brandizzi, Frangne, Marc-Martin, Hawes, Neuhaus & Paris 2002). Interestingly, there is no confusion between endosomal and lysosomal targeted proteins and in general little confusion between proteins targeted to non-plasma membrane locations. This indicates that the current location models seem to miss some specific targeting signal, and that more sensitive models can increase the prediction accuracy without severing the discrimination between locations.

8 Conclusion

We presented a novel architecture of an HMM based localization predictor for transmembrane proteins. In contrast to topology predictors, the new architecture has less states but models the terminal regions and is of second order. The latter is in agreement with the observation that location predictors based on di-peptide composition typically achieve higher performance than classifiers that exploit the mono amino acid composition only.

By modeling the characteristic topology of transmembrane proteins, the new predictor achieves a significant increase in prediction accuracy (correlation coefficient 0.46), compared to predictors based on overall di-peptide composition. To our knowledge, it is the only localization predictor specifically for transmembrane proteins, that is currently available online (<http://pprowler.itee.uq.edu.au/TMPHMLoc>).

We took advantage of the recently published LOCATE database and concentrated our efforts on locations along the secretory pathway, which are especially difficult to distinguish between.

Further work will focus on extending the range of predicted locations, utilizing additional data sources and comparing the new predictor against a more comprehensive set of alternative methods.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990), 'Basic local alignment search tool.', *J Mol Biol* **215**(3), 403–410.
*<http://dx.doi.org/10.1006/jmbi.1990.9999>
- Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004), 'Improved prediction of signal peptides: SignalP 3.0.', *J Mol Biol* **340**(4), 783–795.
*<http://dx.doi.org/10.1016/j.jmb.2004.05.028>
- Brandizzi, F., Frangne, N., Marc-Martin, S., Hawes, C., Neuhaus, J. & Paris, N. (2002), 'The destination for single-pass membrane proteins is influenced markedly by the length of the hydrophobic domain', *Plant Cel* **14**, 1077–1092.
- Chou, K. C. (2001), 'Prediction of protein cellular attributes using pseudo-amino acid composition.', *Proteins* **43**(3), 246–255.
- Chou, K. C. & Elrod, D. W. (1999), 'Prediction of membrane protein types and subcellular locations', *Proteins* **35**, 137–153.

Method	ER	GO	PM	EN	LY	Overall	TT	QT
SVM1	0.072	0.000	0.049	0.000	0.000	0.024 (± 0.007)	0.003	0.001
SVM2	0.318	0.304	0.320	0.369	0.235	0.309 (± 0.017)	0.347	0.001
HMM1	0.172	0.232	0.240	0.160	0.034	0.168 (± 0.012)	0.028	0.660
HMM2	0.492	0.479	0.506	0.403	0.433	0.462 (± 0.020)	0.025	0.782
HMM3	0.532	0.455	0.469	0.385	0.392	0.447 (± 0.019)	0.157	0.934

Table 1: Prediction accuracy and training and query times for methods split by location. Results are 10 fold cross-validated, 10 times repeated. Method = prediction method, ER = Endoplasmic Reticulum, GO = Golgi Complex, PM = Plasma Membrane, EN = Endosome, LY = Lysosome. Overall = overall mean correlation coefficient with 95% confidence interval in brackets. A correlation coefficient of 1.0 is ideal. TT = training time in msec and QT = query time in msec per sample on a Pentium 4, 2.8 GHz with 1 GB main memory.

- Cui, Q., Jiang, T., Liu, B. & Ma, S. (2004), 'Esub8: A novel tool to predict subcellular localizations in eukaryotic organisms', *BMC Bioinformatics* **5**(66).
- Davis, M. J., Zhang, F., Yuan, Z. & Teasdale, R. D. (2006), 'MemO: A consensus approach to the annotation of a protein's membrane organization', *In Silico Biol.* **6**(0037).
- Durbin, R. M., Eddy, S. R., Krogh, A. & Mitchison, G. (1998), *Biological sequence analysis*, Cambridge University Press, Cambridge, UK.
- Feng, Z. P. & Zhang, C. T. (2001), 'Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids.', *Int. J. Biol. Macromol.* **28**(3), 255–261.
- Fink, L., Aturaliya, R., Davis, M., Zhang, F., Hanson, K., Teasdale, M., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. & Teasdale, R. (2006), 'LOCATE: a mouse protein subcellular localization database.', *Nucleic Acids Research* **34** (Database issue), D213–D217.
*<http://dx.doi.org/10.1093/nar/gkj069>
- Guo, J., Lin, Y. & Sun, Z. (2005), A novel method for protein subcellular localization: Combining residue-couple model and SVM, in Y.-P. P. Chen & L. Wong, eds, 'Proceedings of 3rd Asia-Pacific Bioinformatics Conference', Imperial College Press.
- Higy, M., Junne, T. & Spiess, M. (2004), 'Topogenesis of membrane proteins at the endoplasmic reticulum', *Biochemistry* **43**, 12716–12722.
- Hua, S. & Sun, Z. (2001), 'Support vector machine approach for protein subcellular localization prediction', *Bioinformatics* **17**(8), 721–728.
- Käll, L., Krogh, A. & Sonnhammer, E. (2004), 'A combined transmembrane topology and signal peptide prediction method', *Journal of Molecular Biology* **338**(5), 1027–1036.
- Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.', *J Mol Biol* **305**(3), 567–580.
*<http://dx.doi.org/10.1006/jmbi.2000.4315>
- Matsuda, S., Vert, J.-P., Saigo, H., Ueda, N., Toh, H. & Akutsu, T. (2006), 'A novel representation of protein subsequences for prediction of subcellular location using support vector machines', *Protein Science* **14**, 2804–2813.
- Park, K.-J. & Kanehisa, M. (2003), 'Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs', *Bioinformatics* **19**(13), 1656–1663.
- Pedrazzini, E., Villa, A. & Borgese, N. (1996), 'A mutant cytochrome b(5) with a lengthened membrane anchor escapes from the endoplasmic reticulum and reaches the plasma membrane', *Proc. Natl. Acad. Sci. USA* **93**, 4207–4212.
- Rapoport, T. A., Goder, V., Heinrich, S. U. & Matlack, K. E. S. (2004), 'Membrane-protein integration and the role of the translocation channel.', *Trends Cell Biol* **14**(10), 568–575.
*<http://dx.doi.org/10.1016/j.tcb.2004.09.002>
- Small, I., Peeters, N., Legeal, F. & Lurin, C. (2004), 'Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences', *Proteomics* **4**, 1581–1590.
- Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998), 'A hidden Markov model for predicting transmembrane helices in protein sequences.', *Proc Int Conf Intell Syst Mol Biol* **6**, 175–182.
- Tusnády, G. E. & Simon, I. (2001), 'The HMMTOP transmembrane topology prediction server', *Bioinformatics* **17**(9), 849–850.
- van Vliet, C., Thomas, E., Merino-Trigo, A., Teasdale, R. & Gleeson, P. (2003), 'Intracellular sorting and transport of proteins', *Progress in Biophysics & Molecular Biology* **83**, 1–45.
- Viklund, H., Granseth, E. & Elofsson, A. (2006), 'Structural classification and prediction of reentrant regions in α -helical transmembrane proteins: Application to complete genomes.', *J Mol Biol* doi:10.1016/j.jmb.2006.06.037.
- von Heijne, G. (1986), 'The distribution of positively charged residues in bacterial inner membranes correlates with the trans-membrane topology.', *EMBO J* **5**, 3021–3027.
- von Heijne, G. (1990), 'The signal peptide', *Journal of Membrane Biology* **115**, 195–201.
- Yu, J. W., Mendrola, J. M., Audhya, A., Singh, S., Keleti, D., DeWald, D. B., Murray, D., Emr, S. D. & Lemmon, M. A. (2004), 'Genome-wide analysis of membrane targeting by *S. cerevisiae* pleckstrin homology domains.', *Molecular Biology of the Cell* **13**(5), 677–688.
- Yuan, Z. (1999), 'Prediction of protein subcellular locations using Markov chain models.', *FEBS Lett.* **451**(1), 23–26.
- Zhou, G.-P. & Doctor, K. (2003), 'Subcellular location prediction of apoptosis proteins.', *Proteins* **50**(1), 44–48.
*<http://dx.doi.org/10.1002/prot.10251>