A Local Search Appproach for Transmembrane Segment and Signal Peptide Discrimination

Sami Laroum[1,2], Dominique Tessier[1], Béatrice Duval[2], and Jin-Kao Hao[2]

[1] UR1268 Biopolymères Interactions Assemblages,
INRA, 44300 Nantes, France
{slaroum,tessier}@nantes.inra.fr
[2] LERIA, 2 Boulevard Lavoisier, 49045 Angers, France
{bd,hao}@info.univ-angers.fr

Abstract. Discriminating between secreted and membrane proteins is a challenging task. This is particularly true for discriminating between transmembrane segments and signal peptides because they have common biochemical properties. In this paper, we introduce a new predictive method called LSTranslocon (Local Search Translocon) based on a Local Search methodology. The method takes advantage of the latest knowledge in the field to model the biological behaviors of proteins with the aim of ensuring good prediction. The LS Prediction approach is assessed on a constructed data set from Swiss-Prot database and compared with one of the best methods from the literature.

Keywords: Subcellular localization, amino acid position, transmembrane segment insertion, local search.

1 Introduction

Subcellular localization of proteins is important for the understanding of gene/protein function. Most eukaryotic protein cells are synthesized in the cytosol and are translocated to various subcellular compartments. Recent studies have led to a better understanding of the transport mechanisms of protein entering the secretory pathway and a complete review can be found in [1]. During biosynthesis, newly synthesized proteins that contain a targeting signal are directed towards the endoplasmic reticulum (ER). The targeting signals are either N-terminal signal sequences called signal peptides (SP) or, in the case of many membrane proteins that lack discrete signal peptides, the first transmembrane sequence called a signal anchor (SA). Then, proteins are translocated across the ER through the translocon channel. If the segment of amino acids inside the channel contains the "right key", the translocon opens sideways and the protein fits in the membrane. Otherwise, the protein is fully translocated across the ER membrane and released into the ER lumen. This phenomena induces an additional difficulty for discrimination between SP and SA and requires special techniques in order to obtain reliable predictive results.

Many prediction methods for this difficult protein subcellular localization problem have been developed over the years to localize proteins with signal

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2010, LNCS 6023, pp. 134–145, 2010.

peptides ([2,3]), or to localize protein with transmembrane segments ([4,5]). However, in spite of their high prediction performance, in certain cases they still yield false predictions. Discrimination between secreted and transmembrane proteins thus remains a challenging task [6]. The frequent false classifications are mainly a consequence of the fact that these proteins contain both a hydrophobic stretch that can be interpreted either as the hydrophobic core region of a signal peptide - the targeting N-terminal signal of secreted proteins - or as a transmembrane segment (TM segment).

Based on an analysis of known soluble and transmembrane sequences, the purpose of the present study was to design a predictive method to define the rules for membrane insertion during the crossing of the translocon. We looked for the code that enables the opening of the translocon knowing that it has to distinguish peptides that share certain biochemical properties: signal peptides, signal anchors and helical transmembrane segments. By sliding a window of about 19 amino acids along the sequence of a protein - the width of the ER membrane -, if we know the code for opening the translocon, we will be able to predict if the protein is a soluble protein or if it is integrated in the membrane.

The method presented in this paper is based on a local search approach and takes advantage of the latest biological knowledge. For this purpose, we consider subcellular localization prediction as a combinatorial search problem and devise a local search based procedure to determine near-optimal solutions. One notices that this is the first time that such an approach is applied to this difficult prediction problem.

The paper is organized as follows: in Section 2, we present some state-of-the-art prediction methods. In Section 3, we describe our approach for membrane protein recognition. In Section 4, we show computational results and comparisons with one best performing method. Finally, in Section 5, we conclude the paper by giving a summary and the future work.

2 Prediction Methods

Many predictive algorithms dedicated to signal peptide or transmembrane segments are available. The first prediction methods were based on the evaluation of the hydrophobicity of each amino acid. The method used a "sliding window" with fixed width (19 residues) to identify transmembrane segments along the protein sequence and the hydrophobicity average was calculated for amino acids within the window [10]. At the moment, the most popular predictors implemented as web servers for both signal peptide or transmembrane predictions, are usually built upon learning systems such as neural networks (NN) [11], Hidden Markov Model (HMM) [16] and Support Vector Machines (SVM) ([17,18]).

Programs dedicated to signal peptide prediction try to localize precisely the N-terminal signal sequence, the signal sequence cleavage site, or a combination of both features. The current most widely used methods are PrediSi [12] and SignalP3.0 [13].

Membrane topology[1] prediction programs evaluate if a protein is likely to be a membrane protein or not and how many membrane protein domains it has. Sometimes, more precise information is added with the orientation and the boundaries of the TM domains. Some popular transmembrane predictors are TOPPred [15] and HMMTOP [14]. Description and evaluation of the main methods can be found in ([19,20]).

However, few methods focus on the difficulty of a correct discrimination between transmembrane segments and signal peptides. One of the best methods dedicated to this difficult prediction problem is Phobius [7] published in 2004. The Phobius method is based on a HMM model which combines a TM helix submodel with a SP submodel. Transition between states are arranged such that the position of SP is located at the N-terminal of the sequence -the hydrophobic region of a SP is seldom located after the first 30 amino acids- whereas TM segments can be found at any position in the sequence. An evolution of the Phobius method is proposed with Philius [8] which implements a topology prediction by dynamic Bayesian networks. The state transition topology of Philius exactly mimics that of Phobius, and performances of Philius are close to those obtained with Phobius. The last method, SPOCTOPUS [9] combines a neural network method with a hidden Markov model. SPOCTOPUS first performs a homology search to create a sequence profile. This method reports a very high accuracy, but it is difficult to link these results with a biological interpretation.

3 A New Approach for Recognition of TM Proteins

3.1 New Biological Knowledge

Recent studies expand understanding of the recognition and the insertion of TM segments by the translocon ([21,22]). Briefly, the experiments describe the insertion of the membrane proteins into the ER membrane by the evaluation of the amino acid contribution during the insertion process. These studies suggest that insertion or not of helical transmembrane segment depends mainly on the local contribution of each amino acid.

First, in their experiments Hessa *et al.* assess the contribution of each amino acid in different positions along the membrane. The key observation is that the amino acid position plays a determining role during targeting by the translocon. The insertion would be mainly related to an interaction energy between the sequence of amino acids committed in the translocon and the membrane. To calculate this energy of interaction, Hessa *et al.* suggest a hydrophobic scale called "biological hydrophobicity scale". For each amino acid a curve determines its influence according to its position in the translocon.

The experimental studies show that (1) each amino acid has a different hydrophobic index for different sequence positions [21] and (2) the scales are symmetric across sequence positions [23]. For example, Figure 1 displays the curves for two amino acids.

[1] Membrane topology describes which regions of the polypeptide chain span the membrane.

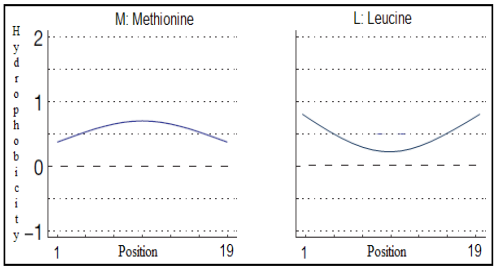

Fig. 1. Contribution scales of two amino acids Methionine and Leucine. The curves describe the contribution according to the position on a 19-residue segment.

Unfortunately, the experiments leading to these curves are very complex to realize, expensive and time-consuming and predictive systems issued from this work such as SCAMPI [24] do not allow a good distinction between SP and TM segments.

3.2 Our Proposal: *In Silico* Fine-Tuning of the Curves

Considering the difficulty of biological experiments that determine the scale of position-specific amino-acid contributions, we propose in this work, to determine *in silico* the scale curves. Our goal is to determine curves that enable a good discrimination between SP and TM segments.

Prediction System. We shall denote in the following by a one of the 20 amino acids that will be represented by a one-letter abbreviation, i.e. $a \in \{$A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y$\}$.

For each amino acid a, we have to determine a curve noted $C[a]$, defined on the interval $[1, 19]$ that represents the 19 positions in a segment that we consider as relevant for membrane insertion. Let us note that this position parameter may vary and for more generality we denote it by $l = 19$ in the following. Therefore, we use $C[a, j] = C[a](j)$ to denote the value of curve $C[a]$ for an integer j where j represents a position, $j \in [1, 19]$.

When we consider a sequence Seq of amino acids of length $l = 19$, we use the notation $Seq = < a^{(1)}a^{(2)} \dots a^{(l)} >$, where $a^{(j)}$ is the index of the amino acid in position j in the segment $a^{(j)}$ is therefore a letter in $\{$A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y$\}$). The hydrophobicity of sequence Seq is defined as follows as the average of $C[a^{(j)}, j]$ for j varying from 1 to l:

$$E(Seq) = \frac{\sum_{j=1}^{l} C[a^{(j)}, j]}{l}$$

Now given a longer sequence $Seq = < a^{(1)}a^{(2)} \dots a^{(n)} >$ of size $n > l$, a sliding window of fixed length l is scanned on the sequence and we define the hydrophobicity of the sequence as the maximum hydrophobicity value of a sub-sequence of length l denoted by:

$$E(Seq) = \max_{1 \leq k \leq n-l+1} \{E(Seq_k)\}$$

where $Seq_k = < a^{(k)} a^{(k+1)} \dots a^{(k+l-1)} >$.

The distinction between a SP and a TM segment is given by the value of $E(Seq)$ and a threshold τ. Seq is classified as a signal peptide when $E(Seq) < \tau$, and Seq is classified as a transmembrane segment when $E(Seq) > \tau$. So a set of curves $\left(C[a^i]\right)_{i=A}^{i=Y}$ determines a classification system for discrimination of SP and TM segments. Our goal is to optimize the 20 curves in order to obtain a good classification accuracy of SP and TM segments.

The quality of such a classification system can be evaluated by the *Area Under the ROC Curve* (AUC) [27]. A ROC[2] curve is obtained by selecting a series of thresholds τ and plotting sensitivity on the Y axis versus specificity on the X axis. The AUC gives the probability that a classifier will rank a randomly selected positive example higher than a randomly selected negative example.

Our problem is therefore an original and difficult optimization problem that can be solved by local search.

3.3 Local Search for Determination of the Curves $C[a]$

Local search approach is a metaheuristic method which is known to be an effective technique for solving computationally hard optimization problems [25]. Our local search algorithm moves in the space of candidate solutions to optimize the AUC of the associated classification system until a solution deemed optimal is found.

Representation of a Solution. In our problem, a candidate solution S is a set of 20 curves $\left(C[a^i]\right)_{i=A}^{i=Y}$ where each curve $C[a]$ is defined on $[1, 19]$. We suppose that each curve is symmetric, so for each amino acid a, $C[a, j] = C[a, 20-j]$ and we decide to approximate each curve by an amino acid contribution function of the following polynomial form:

$$Y = \alpha x^2 + \beta \tag{1}$$

This contribution function is defined by two parameters: H_{middle}, the Y extremum of the function obtained for $j = 10$, and $H_{extremity}$ the value of the function for $j = 1$ and $j = 19$. So a curve is entirely fixed by a pair of values $(H_{extremity}, H_{middle})$. Figure 2 displays an example of a curve of the amino acid contribution function.

Initialization. To start the local search process, we consider curves based on the hydrophobic values given by the Kyte & Doolittle scales [10], which are known to be one of the best hydrophobic indexes [26]. Each scale considers that the index is independent of the position of the amino acid in the sequence and so in our formalization a constant curve h_i is associated to each amino acid a, so our initial solution is $S_0 = (C[a^i])_{i=A}^{i=Y}$ such that $\forall j \in [1, 19]$, $C[a^i, j] = h_i$.

[2] The term ROC comes from signal detection theory and means *Receiver Operating Characteristic*.

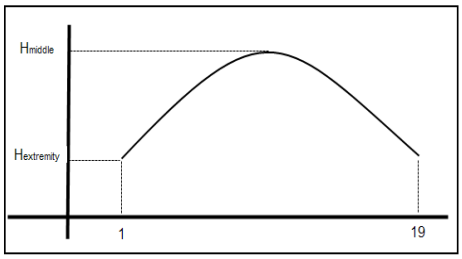

Fig. 2. Curve of contribution function defined by two parameters

Neighborhood. The local search process moves from solution S to an improving neighboring solution S' of the current solution S. To obtain a neighboring solution, a curve from S is randomly selected and then modified to generate a new curve. More precisely, let C denote our selected curve. As explained previously, C is entirely defined by a couple $(H_{extremity}, H_{middle})$. A new curve C' is generated by a new couple $(H_{extremity} + \epsilon, H_{middle} + \delta)$, with the constraints that $(\epsilon, \delta) \in \{-0.5, 0, 0.5\}^2$ or $(\epsilon, \delta) \in \{-0.3, 0, 0.3\}^2$. This leads to a total of 18 combinations. Removing the two trivial cases with $(\epsilon, \delta) = (0, 0)$, one can generate 16 new curves from an existing curve. In other words, each solution S has 16 neighboring solutions (for a chosen curve). These neighbor solutions are examined by the local search procedure at each iteration.

Evaluation Function. At each local search iteration, all candidate neighbors of the current solution S are evaluated. According to the principle of *steepest descent*, the best solution is chosen to replace the current solution S and the local search process is iterated from this new solution. The quality of a neighbor solution S' is assessed by the evaluation function AUC(S') that estimates the ability of the solution to obtain a suitable discrimination between SP and TM segments, using the classification system based on the curves of S'. Let us note that it is sufficient to compute the AUC of the classification system associated to a set of curves to evaluate a solution. The determination of a specific threshold τ is not necessary at this moment.

LSTranslocon Procedure. The general LSTranslocon procedure is composed of two main phases: the modeling phase and τ-calibration phase. The modeling phase consists in the determination of the best curves for the discrimination between signal peptide and transmembrane segments, as just explained. The stopping condition of this phase is obtained by evaluation of the proposed solutions on a validation dataset (see next Section). When the classification between SP and TM segments becomes stable - the AUC value remains unchanged - the search process ends and returns the best solution.

After the modeling phase, the τ-calibration phase is performed to determine the most appropriate threshold τ that permits to classify the protein sequences

as transmembrane sequences or signal peptides. This threshold corresponds to the best classifier of the ROC curve.

4 Experimentations and Discussion

4.1 Benchmark

The literature does not offer suitable ready-to-use dataset for our protein localization problem. Consequently, to assess the performance of our proposed method, we built a high quality benchmark database where the desired constituent sequences were extracted from the most recent version of Swiss-Prot database 57.8 (released on 22 September 2009) according to the following four steps: (1) The selected proteins are only those that are marked in the OC (organism classification) line as "eukaryota or eukaryotic", the eukaryotic proteins differ from prokaryotic proteins in particular in the addressing in the cell. (2) For the proteins obtained from the above step, we extract those which were marked as "signal peptide" and "transmen" in the FT (Feature Table) line, (3) We removed those which were annotated with uncertain terms for their signal peptide or transmem, such as "potential", "probable", or "by similarity" (4) For the resulting data set, the sequence identity is checked and analyzed by using the program CD-HIT [28], which produces a non-redundant dataset at the 50% sequence identity level.

By strictly following the above steps, we finally obtained a benchmark database for eukaryotic proteins. The database contains 5469 sequences with signal peptide and 798 transmembrane protein segments.

In order to equalize the sizes of the database between SP and TM segments, we randomly selected 6% from the dataset with SP. Thus, the final benchmark database contains 900 SP and 798 TM segments.

Now, from the selected data with signal peptide, we extracted the first 55 amino acids (because the maximum length of SP is 55 amino acids). Note that in our study we consider a SA as TM segment, for this reason we selected only the first TM or SA segment according to its annotations in Swiss-Prot. In the case where the selected segment has a length inferior to 19, we expanded the selected window to obtain a segment superior or equal to 19 amino acids.

The statistical distribution of amino acids in the benchmark database is presented in Figure 3. The figure shows that some amino acids are less represented than others in the data sequences.

This benchmark database is used to evaluate our approach by a process of cross-validation that involves 10 experiments. For each experiment, we build from the initial benchmark database three types of datasets: a *training* set, a *validation* set and a *test* set. The training set is used for the modeling phase (optimization of the curves) and the determination of the threshold τ. The validation set is used to determine the stopping condition and to avoid overfitting. And the test set is used to evaluate the classification accuracy of our prediction model. Each of these sets is constituted by randomly selecting from our initial

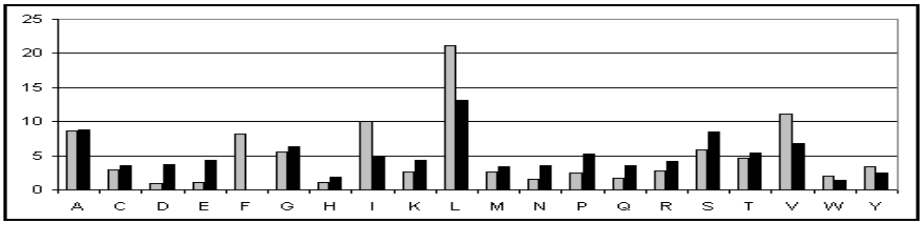

Fig. 3. Statistical distribution of each amino-acid in the dataset. The Y axis gives the frequency of each amino acid. The dark bars indicate the distribution for the data with SP segments while the grey bars show the distribution with TM segments (or SA).

database according to the following proportions: 60% for training data, 10% for validation data and 30% for test data.

The classification accuracy reported in the next Section is the averaged value calculated only on the *test* sets of the ten experiments.

4.2 Results and Discussion

This section reports the experimental results that assess our approach. The main focus of our approach is to determine a new hydrophobic scale that enables biological understanding of the insertion phenomena inside the translocon. Therefore, we first present a set of curves obtained by LSTranslocon and discuss them. Then, we demonstrate that our approach can lead to an effective predictor for the discrimination between SP and TM segments.

Figure 4 shows the 20 curves (one curve per amino acid) obtained when we apply our method to the benchmark database described in 4.1. We can remark that the shapes of the curves are quite different. These shapes suggest that some amino acids like Proline (P) or Methionine (M) facilitate more the insertion when they are embedded inside the membrane in the middle of the curve whereas other amino acids like Leucine (L) or Serine (S) prefer the interface positions at the extremities of the curves. The experimental curves given by Hessa et al [23] were obtained *in vitro*. The procedure is complex and time consuming. Our computational approach allows us to obtain similar curves much more rapidly. However, we observe that for some amino acids and especially those which are less frequent in our database, our curves have a different shape with respect to those suggested before. One possible explanation of this observation is that these amino acids do not have a great influence in the insertion process, which explains their lack of representation in the benchmark database and the difficulty to properly adjust their insertion curve. Now we turn our attention to the predictor system that results from our local search approach. We compare our prediction rates with those obtained by Phobius, one of the best performing prediction tools (see section 2). As these two programs do not completely share the same objectives, a fair comparison is difficult. For all that, we performed two experiments to evaluate the capabilities of these programs.

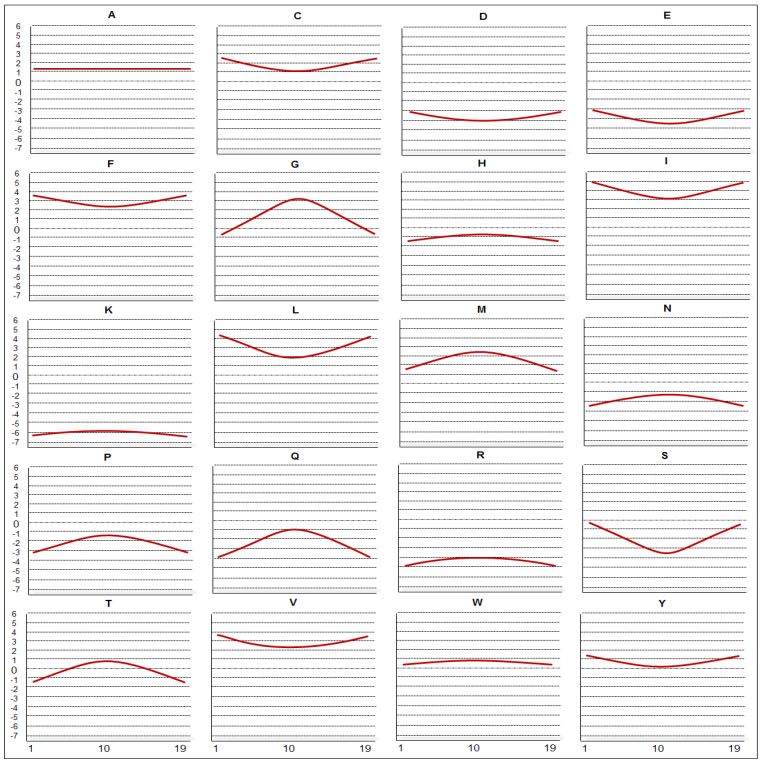

Fig. 4. Curves of insertion for each amino acid arbitrarily selected from one experiment. The curves describe the contribution according to the position on a 19-residue segment. The X axis shows the positions and the Y axis indicates the insertion index.

In the two experiments, LSTranslocon is trained according to the experimental protocol described in Section 4.1 and the Phobius predictor available on the web is used. The two programs are evaluated 10 times on the same test sets and the results are summarized in Table 1.

In the first experiment, the 10 test sets are generated as explained in Section 4.1. On these datasets, Phobius achieves about 84% average accuracy whereas LSTranslocon is limited to 80% (see results "Experiment 1" in Table 1). The Phobius web server accepts the whole protein sequence as input data whereas LSTranslocon requires only the TM or SP segments (see Section 4.1). The Phobius model takes advantage of the supplementary information. In fact, in its model, the transition between states are arranged such that the position of SP are located at the N-terminal region whereas TM segments are more often found later in the sequence.

To confirm this observation, we conduct a second experiment. This time, Phobius and LSTranslocon are tested on 10 restricted test sets : we only keep the TM segments located in the N-terminal regions from the previous 10 test

Table 1. Results of two experiments that compare our method LSTranslocon and Phobius. Each cell of the table indicates the *best classification accuracy* achieved by the corresponding classifier on the given test set.

Experiment 1			Experiment 2		
Data set	LSTranslocon	Phobius	Data set	LSTranslocon	Phobius
Test 1	0.79	0.85	Test 1'	0.80	0.82
Test 2	0.81	0.86	Test 2'	0.83	0.81
Test 3	0.82	0.84	Test 3'	0.81	0.81
Test 4	0.81	0.87	Test 4'	0.81	0.83
Test 5	0.79	0.85	Test 5'	0.81	0.81
Test 6	0.81	0.84	Test 6'	0.80	0.81
Test 7	0.81	0.87	Test 7'	0.82	0.84
Test 8	0.80	0.85	Test 8'	0.82	0.82
Test 9	0.80	0.84	Test 9'	0.81	0.81
Test 10	0.79	0.82	Test 10'	0.80	0.79
Average	0.80	0.84	Average	0.81	0.81

sets. We consider that a TM segment is not in the N-terminal region, if the start position of the TM segment is beyond the 30^{th} amino acid. The results "Experiment 2" show that the accuracy of Phobius decreases and is equal to the accuracy obtained with LSTranslocon. As we look for the curves which explain the insertion phenomena inside the translocon, we do not want to exploit the statistical bias concerning the TM and SP positions inside the protein sequence. Under this strong constraint, we observe that the results reported in this paper are as good as those obtained by Phobius. This demonstrates that our approach, which is quite new, has an interesting potential that we hope to improve in the near future.

5 Conclusion

In this paper, we have presented a new method based on the Local Search Approach for the discrimination of transmembrane segments and signal peptides. The method integrates the latest knowledge acquired in the biological field and presents the insertion curves in the membrane for each amino acid. LSTranslocon is evaluated with the ability to maximize the distinction between TM segments and SPs. These two characteristics ensure that our method constitutes a complete approach and gives a good explanation of insertion machinery.

The main advantage of such a predictive method is the chemically interpretable rules that will enable experts to understand biological phenomena. Furthermore, the proposed method can be applied to very large data, even whole proteome datasets.

Several improvements to the proposed method can be envisaged. One immediate possibility would be to study alternative functions to optimize the curves and to introduce more biological knowledge to provide more effective guidance

of the local search process. Another natural extension would be to reinforce the basic local search procedure by more powerful metaheuristics. Moreover, a next step will be to apply our approach on full sequences with the aim to localize the TM segment positions.

Acknowledgements

This research was partially supported by the region Pays de la Loire (France) with its "Bioinformatics Program" (2007-2009). The authors are grateful to the reviewers for their useful comments.

References

1. Mandon, E.C., Trueman, S.F., Gilmore, R.: Translocation of proteins through the Sec61 and SecYEG channels. Current Opinion in Cell Biology 21, 501–507 (2009)
2. Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G.: Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Bio. 300, 1005–1016 (2000)
3. Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai, K.: WoLF PSORT: protein localization predictor. Nucleic Acids Research 35, W585–W587 (2007)
4. Viklund, H., Granseth, E., Elofsson, A.: Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: Application to complete genomes. J. Mol. Biol. 361, 591–603 (2006)
5. Nugent, T., Jones, D.T.: Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 10 (2009)
6. Kaell, L., Krogh, A., Sonnhammer, E.L.L.: Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. Nucleic Acids Research 35, W429–W432(2007)
7. Kall, L., Krogh, A., Sonnhammer, E.L.L.: A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338, 1027–1036 (2004)
8. Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A., Noble, W.S.: Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput. Biol. 4 (2008)
9. Viklund, H., Bernsel, A., Skwark, M., Elofsson, A.: SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 2, 2928–2929 (2008)
10. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132 (1982)
11. Small, I., Peeters, N., Legeai, F., Lurin, C.: Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4, 1581–1590 (2004)
12. Hiller, K., Grote, A., Scheer, M., Munch, R., Jahn, D.: PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Research 32, W375–W379 (2004)
13. Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S.: Improved prediction of signal peptides: SignalP 3.0. J. Mol. Bio. 340, 783–795 (2004)
14. Tusnady, G.E., Simon, I.: The HMMTOP transmembrane topology prediction server. Bioinformatics 17 (2001)

15. Von Heijne, G.: Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule. J. Mol. Bio. 225, 487–494 (1992)
16. Bagos, P.G., Liakopoulos, T.D., Hamodrakas, S.J.: Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. BMC Bioinformatics 7 (2006)
17. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728 (2001)
18. Garg, A., Bhasin, M., Raghava, G.P.S.: Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J. Biol. Chem. 280, 14427–14432 (2005)
19. Klee, E.W., Sosa, C.P.: Computational classification of classically secreted proteins. Drug Discovery Today 12, 234–240 (2007)
20. Klee, E.W., Ellis, L.B.M.: Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics 6 (2005)
21. Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., von Heijne, G.: Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 433, 377–381 (2005)
22. Hessa, T., White, S., von Heijne, G.: Membrane insertion of a potassium-channel voltage sensor. Science 307, 1427 (2005)
23. Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H., von Heijne, G.: Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature 450, 1026-1030 (2007)
24. Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., Elofsson, A.: Prediction of membrane-protein topology from first principles. Proc. Natl. Acad. Sci. USA 105, 7177–7181 (2008)
25. Hoos, H., Stützle, T.: Stochastic Local Search: Foundations & Applications. Morgan kaufmam Publishers Inc., San Francisco (2004)
26. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S.: Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18, 298–305 (2002)
27. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Labs (2004)
28. Weizhong, L., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006)