Multi-neighborhood search for discrimination of signal peptides and transmembrane segments

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

[1] LERIA, 2 Boulevard Lavoisier, 49045 Angers, France
[2] UR 1268 Biopolymères Interactions Assemblages,
INRA, 44300 Nantes, France
{laroum,bd,hao}@info.univ-angers.fr
{tessier}@nantes.inra.fr

Abstract. A key step in study of biosynthesis of membrane proteins is to look for the code that could be used to explain and predict which proteins would eventually be inserted in the membrane and which proteins would be secreted into the ER lumen when they cross the translocon channel. The aim of this work is to present an improvement of a previous method based on a local search approach. The proposed method relies on new in-depth biological observations to design a new search space for the local search algorithm. Experiments conducted on a dedicated dataset show that our new approach leads to improved outcomes in terms of prediction rates.
Keywords: Amino Acid Position, Transmembrane Segment Insertion, Signal peptide, Local Search, Multi-Neighborhood Search

1 Introduction

Proteins transported across the endoplasmic reticulum (ER) membrane include soluble proteins and membrane proteins. Recent studies have led to a better understanding of the transport mechanisms of these proteins ([1], [2]). A targeting signal localized in the N-terminal sequence and called signal peptide (SP) guides the nascent protein to the ER membrane. Next, the nascent protein gets into the sec61 translocon, a protein complex located in the ER membrane. The translocon discriminates between the proteins which cross the ER membrane and are released in the ER lumen and the proteins which get inserted in the ER membrane. When membrane proteins lack discrete signal peptides, the first transmembrane sequence directs the nascent protein to the membrane like a signal peptide. In this case, the first transmembrane segment is called a signal anchor (SA).

The recognition inside the translocon channel is based on identifying the "right key". If the segment of amino acids contains the code, 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.

There exist several methods using both experimental and statistical data that are optimized to predict the insertion of membrane proteins. Some of them

are based on experimental works that try to elucidate precisely how membrane proteins get inserted or secreted through the ER membrane. Scampi [3] is a prediction method using recently published experimental results of the energetics of insertion of a single transmembrane (TM) segment into the ER membrane [4]. MINS [5] and MINS2 [6] use computational methods for predicting the membrane insertion free energies of protein sequences. Following the same principle, we proposed in a previous work a prediction method based on a local search algorithm [7]. The idea was to mimic the insertion phenomena as closely as possible by modeling the likelihood of each amino acid residing in the membrane. This work assumed that the insertion efficiency of a TM segment depends on its amino acid composition and on the position of the amino acids within the TM segment. It led to a new in silico scale composed of 20 curves where each curve represents the insertion profile of one amino acid. These curves are also used to discriminate between SP and TM segments, a problem that is still not fixed.

In this paper, we present a multi-neighborhood local search (MN-LS) approach which is based on new biological knowledge. This approach explores two spaces which are composed of straight lines and symmetric curves respectively and employs different neighborhoods to explore these spaces. The basic idea of MN-LS is to optimize separately the straight lines and symmetric curves by adjusting a straight line or a symmetric curve each time. Tested on a dedicated data set, the proposed approach proves to be able to provide good prediction accuracy as well as more interesting results for the curves.

The remainder of this paper is organized as follows. In Section 2, we present some biological knowledge to understand the problem and a summary of our previous method. In section 3, we describe the construction of a new database and we discuss the representation of the data. Section 4 and 5 present the improvements of the approach and give the experimental results on three datasets. The conclusion and perspectives are given in section 6.

2 Local search for modeling amino acid insertion curves

2.1 Biological knowledge

This work deals with the recognition of two types of proteins: those secreted in the ER lumen and those inserted in the ER membrane [8] and tries to identify the 'code' recognized by the translocon.
Recently, several experiments were designed to read the 'sequence code'. Hessa *et al.* [9] carried out a series of in vitro experiments which assess the contribution of each amino acid in different positions along the membrane. The experiments revealed that the amino acid position plays a determining role during targeting by the translocon. The hydrophobicity of an amino acid is related to its transfer energy from a polar medium such as the cytoplasm to an apolar medium such as the membrane. So, Hessa *et al.* suggest a biological hydrophobicity scale derived from their experiments. Even if most of the hydrophobicity scales have been derived experimentally, we assume that we can elaborate an in silico scale based on the study of the insertion phenomena from two sets of protein segments which

cross the translocon and share the same chemical hydrophobic profile : SP and TM segments. This scale could benefit from a larger quantity of data stored in the protein databases and consequently could be much more precise. In this scale, each amino acid has different hydrophobic indexes for different positions and the scale is represented by 20 symmetric curves across sequence positions [3]. The length of the curve is 19 amino acids which corresponds to the thickness of the membrane.

2.2 Local search for in silico determining the curves

In a previous system called LSTranslocon [7], we used a local search approach to determine *in silico* the hydrophobic indexes of the amino acids. Following the hypotheses assessed by biological experiments, we search indexes defined by symmetric curves over $l = 19$ positions, and a solution is therefore a set of 20 curves, one for each amino acid.

Given a solution, the insertion index of a sequence of amino acids of length l is the average of its indexes. In the case of a longer sequence, a sliding window of fixed length l is scanned on the sequence and we define the insertion index of the sequence as the maximum index calculated on a sub-sequence of length l. The distinction between a SP and a TM segment is decided according to the following principle: if the insertion index is lower than a threshold τ then the sequence is a SP, otherwise the sequence is a TM segment. The quality of this classifier is evaluated by the Area Under the ROC Curve (AUC) [10] that estimates the ability of the solution to obtain a suitable discrimination between SP and TM segments.

LSTranslocon tries to maximize the AUC and solves this optimization problem by a local search algorithm that has the following characteristics.

Search space. A configuration s is a set of 20 curves defined on the interval [1..19] and each curve is defined by an equation $Y = \alpha(x - X_0)^2 + \beta$. In fact we represent each curve by Y_{ext} the value at the extremities of the interval for $x = 1$ and $x = 19$ and Y_{mid} the value for $x = 10$.

Initialization. The initial configuration s_0 is a set of 20 constant values given by the hydrophobic scale of Kyte and Doolittle [11].

Neighborhood. A neighbor of a configuration s is defined by randomly selecting a curve C from s and by computing a new curve C', by slight modifications of C. In this implementation, each configuration has a neighborhood of size 16 which is visited by a descent algorithm. The stopping condition is reached when the AUC becomes stable on a validation set.

Dataset. To assess LSTranslocon, we used a database, called SWP-v1, of 900 SP and 798 TM segments extracted from the Swiss-Prot database 57-8 (released on 22 September 2009).

Training, validation and test. This database is used to train and evaluate LSTranslocon according to the following cross-validation process that involves 10 experiments. For each experiment, the initial dataset is split into three parts by randomly drawing 60% of the data for the training set, 10% for the validation set and 30% for the test set. The training set is used for the optimization of the

insertion curves of the amino acids and for the determination of the threshold τ. The validation set is used to determine the stopping condition and to avoid overfitting. When the curves and the threshold are computed, the test set is used to evaluate the classification accuracy of the predictive model.

For each experiment, a 10-fold cross-validation provides an average value for the classification accuracy achieved by the system.

Results The results of LSTransLocon were encouraging since we obtained a predictive accuracy of 80% on our benchmark dataset, which is quite close to the results of Phobius [12], when we consider TM segments located in the N-terminal region of the protein.

However, we observed that different runs of LSTranLocon may lead to different solutions, that means different insertion curves for the amino acids. For an amino acid that has few occurences in the dataset, it is not surprising that we cannot correctly adjust its curve, but this phenomenon is also observed for some frequent amino acids, like Leucine.

Therefore we tried to analyze the reasons for this instability. The following sections propose different modifications of our approach in order to obtain more reliable results and to improve the discrimination of TM segments and SP.

3 Benchmark dataset and representation of the data

Since the publication of our previous results with LSTranslocon, new releases of the database Swiss-Prot are available. Each new version contains more proteins and up-to-date information. It is very important for our approach to deal with numerous and reliable data. Therefore we constructed a new benchmark dataset from the Swiss-Prot database 57.15 (released on 02-march-2010).

This section describes precisely how this dataset, named SWP-v2, is extracted. It also presents the first improvement of our approach, namely a change of representation for TM segments.

3.1 Construction of a benchmark dataset: SWP-v2

All datasets generated during the experiments are extracted according to the following steps: (1) The selected proteins are only those that are marked in the OC (organism classification) line as "eukaryota", 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 "transmem" in the FT (Feature Table) line. (3) For the proteins having a signal peptide, we only select those marked in the CC (subcellular localization) line by "secreted". For transmembrane proteins, we select those marked at line CC (subcellular localization) by "membrane" or "endoplasmic reticulum". (4) We remove the proteins having a SP and annotated as "potential", "probable", or "by similarity". However, for transmembrane proteins, we only remove proteins annotated "probable", or "by similarity". (5) For the resulting dataset, the sequence identity is checked and analyzed by using

the program CD-HIT [13], 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 1050 sequences with signal peptide and 734 transmembrane proteins.

The signal peptide is located in N-terminal region, and the length varies between different proteins. For eucaryotic proteins the average length of signal peptide ranges from 22 to 32 amino acid residues [14]. So, we represent the SP with the first 32 amino acids.

Note that in our study we consider a SA as a TM segment. Furthermore, the polytopic membrane proteins are different from the bitopic membrane proteins because they do not have the same ability to be inserted in the membrane [6]. 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 represent a TM segment by a sequence of 19 amino acids.

3.2 Influence of TM segment representation

The annotation of TM segments in UniProtKB/Swissprot is based on the information given by the published papers [14]. Nevertheless, experimentally proven transmembrane regions are generally annotated with the qualifier Potential due to the difficulties in determining their precise boundaries. In our dataset, several of the 734 TM proteins are annotated with the qualifier Potential and in this case, the transmembrane regions are predicted by the application of predictive tools TMHMM, Memsat, Phobius and the hydrophobic moment plot method of Eisenberg and coworkers. These prediction tools introduce a bias in our learning data. We want to determine whether the annotation of TM segments has an influence on our prediction method.

Therefore, we carried out several experiments to study the influence of TM segment representation on the performance of discrimination between SP and TM segments. We consider the chain of amino acids that represents the annotation of the TM segment in the dataset SWP-v2 and we widen this chain by adding a certain number of amino acids before and after this annotation position.

Table 1 reports the accuracy obtained with LSTranslocon method on SWP-v2. Δ represents the number of amino acids added before and after the extraction window of the TM segments. Note that $\Delta 0$ means that the TM segments are represented with their original length (annotation in Swiss-prot). Δk represents a chain of amino acids enlarged of $2k$ amino acids, k before and k after the annotated position of the TM segment.

We observe that widening the extraction window of the TM improves the predictive performance. Therefore, for all the following experiments, a TM segment is represented by a chain of amino acids which is widened by 10 amino acids before and after the given position of the TM segment.

Widening	$\Delta 0$	$\Delta 2$	$\Delta 4$	$\Delta 6$	$\Delta 8$	$\Delta 10$
Average accuracy	0.792	0.808	0.816	0.825	0.825	0.827
Standard deviation	0.0204	0.0069	0.0144	0.0063	0.0131	0.0087

Table 1. The table reports an evaluation of the effect of TM segment length for discrimination between SP and TM segments. Each cell of table indicates the average accuracy and the standard deviation achieved by a 10-fold cross-validation on SWP-v2.

3.3 Statistical data analysis

Our method adjusts the insertion curve of each amino acid in order to obtain a good classification of the TM segments and SP of the training dataset. So it is interesting to observe the statistical distribution of the amino acids in our benchmark database. This information is presented in Figure 1. Note that our dataset SWP-v2 represents a small portion of the total database Swiss-Prot.

We can observe that some amino acids, like Histidine, have very few occurences in SWP-v2 and the same observation is true in SwissProt. Other amino acids like Leucine (L), Alanine (A), Valine (V), Isoleucine (I), Phenilalanine (F) and Serine (S) are very frequent in SWP-v2 as well as in SwissProt. However, in our data the amino acid Phenilalanine (F) is quite frequent but is not very present in the complete database. Besides, most of these frequent amino acids have a high hydrophobic value.

We shall exploit this information in section 4 to propose different training models for the amino acids.

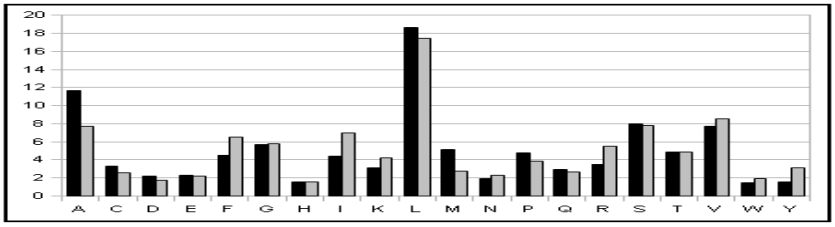

Fig. 1. Statistical distribution of each amino-acid in the SWP-v2. The Y axis gives the percentage of occurrences of each amino acid in the SP segments (dark bars) and in the TM segments (grey bars) in the SWP-v2 dataset.

3.4 Evaluation on other test datasets

As explained before, we use a cross-validation process to evaluate the new method described in this paper. The training set and the validation set are used to learn the curves and to propose a classification threshold; then the test set is used to evaluate the accuracy of the resulting classifier. A series of 10

experiments gives an average accuracy calculated on SWP-v2, that is extracted from the Swiss-prot database.

We also propose to test our method on additional sets of membrane proteins. These sets are extracted from the SCAMPI dataset [3]. This database is interesting because it contains proteins with known 3D structure and the position of the TM segment is more reliable. So this is a good benchmark for a predictive method.

SCAMPI dataset [3] is divided into two collections of proteins. The first collection is a "high-resolution" set of 123 transmembrane proteins. The maximum homology among these 123 proteins is 40%. The second collection is a "low-resolution" set of 146 proteins with homology reducing at 40% sequence identity.

These two collections provide examples that are different from the TM segments of SWP-v2 because SWP-v2 contains proteins with a unique TM segment while the proteins from Scampi contain several TM segments. On the contrary, SWP-v2 and Scampi datasets contain the same SP. So to follow an unbiased methodology, we apply a cross-validation process where the test set is a collection of TM segments of Scampi completed by a set of SP drawn from SWP-v2. By dividing SWP-v2 in 10 parts, we can achieve 10 experiments and we call "ScampiHigh" the case where the test TM segments are obtained from the Scampi collection of high resolution, and "ScampiLow" the other case.

4 New search space of insertion indexes for amino acids

4.1 Principle

As explained in section 2.2, our previous method [7] represents the insertion index of each amino acid by a symmetric curve defined on $l = 19$ positions. We observed that the curves of some amino acids were unstable in the sense that different runs of our local search algorithm may provide different shapes of curves for the same amino acid. One explanation for this observation is that these amino acids are not very frequent in our database and the algorithm has some difficulty to properly adjust their insertion curve.

Therefore, in our new algorithm, we propose to consider that an insertion index may be defined by a constant straight line or by a symmetric curve. A symmetric curve is defined by two parameters (Y_{ext}, Y_{mid}) whereas a straight line is defined by a sole parameter Y_{mid}.

According to the statistical distribution of amino acids in our benchmark dataset (see section 3.3) and to the hydrophobicity values of amino acids, we propose to consider two clusters of amino acids.

On the one hand, the amino acids Alanine (A), Phenylalanine (F), Isoleucine (I), Leucine (L) and Valine (V) are highly frequent in our data. We also notice that they are known to have high hydrophobic values. So they form a group noted $\mathcal{C}$ and for each element of this group we search a symmetric curve to define the insertion index. The second group $\mathcal{D}$ contains all the fifteen other amino acids

and for each element of this group we search a straight line to define the insertion index.

These hypotheses define a new search space for our new algorithm, where a solution (s) is a set of 15 straight lines and 5 curves, each being defined on the interval [1..19]. As a constant line is defined by a unique parameter, this reduces the number of parameters that characterize a solution.

To explore this search space, we adopt the following two-stage strategy. We start with an initial solution (s_0) defined by 20 constant hydrophobicity values. Our algorithm first optimizes the values of the 15 straight lines. Then when an optimum is reached for the lines, we optimize the 5 symmetric curves. In each case, we explore the whole neighborhood of the current solution in order to choose the solution that provides the best AUC improvement.

In table 2, we compare the results obtained with this search space to the results obtained with LSTranslocon. In the two cases, we use the scale of Kyte and Doolittle [11] as a initial solution.

We notice that the new search space enables a slight improvement of the discrimination performance on the three test sets.

Moreover, this new search space guarantees a stability of the insertion curves, since we reduced the problem dimensionality and the algorithm tries to adjust symmetric curves only for amino acids that are very frequent in the data.

Method	New search space			LSTranslocon		
Data	SWP-v2	ScampiHigh	ScampiLow	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.829	0.807	0.824	0.827	0.795	0.804
Standard deviation	0.0131	2.4e-05	0.0077	0.0087	0.0002	0.0125

Table 2. Results of the algorithm with the new search space compared to LSTranslocon: average accuracy and standard deviation achieved by a 10-fold cross-validation on 3 datasets.

4.2 Influence of the initial scale

In a local search process, the initial solution may be randomly chosen or defined by relevant knowledge to provide good conditions for the algorithm. In all our previous experiments, we chose the Kyte and Doolittle scale [11] to fix the constant values of the initial solution. The literature provides other hydrophobicity scales. This section studies whether a particular scale is better suited to initiate our search process.

We consider the three following hydrophobicity scales: Kyte and Doolittle [11], Eisenberg [16] and Engelman [17]. We study what results can be obtained with the new search space described in the previous section when the initial solution (s_0) is defined by each of these scales. These scales are not normalized and their range are quite different. So we give different variation steps and different running times to the search process to lead this experiment in a fair manner.

Table 3 shows that the results obtained with the Kyte and Doolittle scale and with the Eisenberg scale are comparable with a slight advantage to Eisenberg

scale. The Engelman scale provides less interesting results, especially on the Scampi data where the accuracy is below 0.80.

So, in the rest of the paper, the experiments use the Eisenberg scale to define the initial solution of the local search process.

Scale	Kyte and Doolittle		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.829	0.807	0.824
Standard deviation	0.0131	2.4e-05	0.0077
Scale	Eisenberg		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.837	0.817	0.826
Standard deviation	0.0134	3.9e-05	0.0063
Scale	Engelman		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.806	0.777	0.791
Standard deviation	0.0001	0.0001	0.0078

Table 3. Comparison of three hydrophobic scales. Each cell reports the average accuracy and the standard deviation achieved by a 10-fold cross-validation on 3 datasets.

5 Multi-Neighborhood Search

To carry out efficiently the search task of determining in silico the hydrophobic indexes of the amino acids, we introduce in this section a multi-neighborhood local search (MN-LS) algorithm which is based on the observations made in Sections 3 and 4. This algorithm is designed to explore two different search spaces in two sequential phases in order to optimize 15 straight lines associated to the amino acids of the group $\mathcal{D}$ (see Section 4.1) and the set of 5 symmetric curves of the group $\mathcal{C}$ (see Section 4.1).

For this purpose, our MN-LS algorithm employs different neighborhoods and uses steepest descent strategies to explore these neighborhoods in a sequential manner.

5.1 Neighborhoods

Neighborhood of straight lines (group $\mathcal{D}$): Since each straight line is defined by the value of the Y_{mid} parameter, a solution of the first search space (group $\mathcal{D}$ of the 15 amino acids) is identified by a vector of 15 values. Given such a solution (s), we define a neighboring solution by adding a shift value ϵ to or subtracting ϵ from one single component of (s). Since (s) contains 15 components, each solution has exactly 30 neighboring solutions. Large values for ϵ lead to important changes of a straight line while small values for ϵ give only slight changes. In this paper, ϵ is experimentally set at 0.7.

Neighborhood of symmetric curves (group $\mathcal{C}$): Recall that each symmetric curve is defined by a couple (Y_{ext}, Y_{mid}). A solution of the second search space (group $\mathcal{C}$ of the 5 amino acids) represents thus 5 symmetric curves which can be considered as a vector of 5 couples (Y_{ext}, Y_{mid}). Given such a solution (s), we generate a neighboring solution by adding a shift value ϵ to or subtracting ϵ from one couple (Y_{ext}, Y_{mid}) of (s) by excluding (Y_{ext}+ϵ, Y_{mid}+ϵ) and (Y_{ext}-ϵ, Y_{mid}-ϵ). Since for each couple (Y_{ext}, Y_{mid}) of (s), we have six neighbors {(Y_{ext}, Y_{mid}+ϵ), (Y_{ext}, Y_{mid}-ϵ), (Y_{ext}+ϵ, Y_{mid}), (Y_{ext}-ϵ, Y_{mid}) (Y_{ext}+ϵ, Y_{mid}-ϵ), (Y_{ext}-ϵ, Y_{mid}+ϵ)}, each solution has exactly 30 neighboring solutions.

In this paper, ϵ is experimentally set at 0.7, 0.5 and 0.3.

5.2 Move strategy and multi-neighborhood exploration

As explained in Section 4.1, each candidate solution is assessed according to the associated AUC score. To make a move from the current solution, the search algorithm examines all the neighboring solutions and picks the best improving neighbor (according to the AUC score) to replace the current solution. The search stops if no such an improving neighbor exists in the neighborhood.

The algorithm starts its exploration by examining the first search space composed of 15 straight lines. This is simply realized by applying the steepest descent strategy to the given neighborhood. When this phase is finished, the algorithm proceeds to the next phase for the optimization of symmetric curves.

To explore the neighborhoods of symmetric curves, the algorithm operates successively by examining the second search space with the neighborhood defined by the largest ϵ value 0.7. Upon reaching a local optimum, the algorithm switches to the next neighborhood defined by $\epsilon = 0.5$ to find another local optimum solution. The search then continues with the neighborhood defined by $\epsilon = 0.3$. We justify these successive explorations by the fact that it is preferable to make important changes to ensure a large exploration of the search space at the beginning of the search and limit the changes for finer examination toward the end of the search.

For both phases of the search, the initial solution is generated by using the values given by the Eisenberg scale (See Section 4.2).

5.3 Experiments

Table 4 (higher part) shows the experimental results that assess our approach MN-LS (a multi-neighborhood search combined with the new search space). For comparison, we recall in the table (middle part) the best results shown in the preceding section with the new search space and a simple neighborhood as well as the results achieved by LSTranslocon (lower part). The three methods begin with an initial solution generated from the Eisenberg scale (See Section 4.2). Each method is evaluated 10 times on the three test datasets.

We observe that the results of MN-LS and the algorithm with the new search space are very close, with a slight improvement by MN-LS on SWP-v2. When we

Method	MN-LS		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.866	0.811	0.826
Standard deviation	0.006	0.0001	0.0037
Method	New search space		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.837	0.817	0.826
Standard deviation	0.0134	3.9e-05	0.0063
Method	LSTranslocon		
Data	SWP-v2	ScampiHigh	ScampiLow
Average accuracy	0.834	0.797	0.809
Standard deviation	0.0285	0.0002	0.0145

Table 4. Comparison between MN-LS using the new search spaces and LSTranslocon.

compare MN-LS with LSTranslocon, we observe that MN-LS achieves improved classification accuracy on the three datasets. These results highlight the importance of separating the search space into straight lines and symmetric curves and the interest of using multi-neighborhoods for searching good solutions.

6 Conclusion

In this paper, we have introduced two features to improve a previous local search approach for membrane protein prediction. Based on biological observations, the new method optimizes first a set of 15 straight lines corresponding to the set of amino acids with a low hydrophobic value, followed by learning symmetric curves for the 5 remaining amino acids which are highly hydrophobic.

To explore the two spaces (straight lines and symmetric curves), the proposed method investigates different neighborhoods and examines them in a sequential and exhaustive manner. The experimental results show that the method gives better results compared with LSTranslocon in terms of classification accuracy and stability of insertion curves.

Concerning the recognition of TM segments in the full sequence of a protein, we observe that the proposed approach is able to predict the position of the first TM segment. For the other TM segments of a protein, it seems that the proposed threshold is not adequate. This is not a surprising result since our training phase is devoted to the discrimination between SP and the first TM segments of proteins.

We are currently studying alternative curves to get closer to the real phenomenon of membrane protein insertion. In [5, 18], the authors use asymmetric curves because it appears that some amino acids are better in the insertion at the N-terminal region than at the C-terminal region. Moreover, we are investigating other search approaches like genetic algorithm in order to improve further the classification accuracy.

7 Acknowledgments

This research was partially supported by the region Pays de la Loire (France) with its "Bioinformatics Program" (2007-2010). 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. Rapoport, Tom A.: Protein transport across the endoplasmic reticulum membrane. Febs Journal. 275, 4471-4478 (2008)
3. 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)
4. 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-U2 (2007)
5. Park, Y., Helms, V.: Prediction of the translocon-mediated membrane insertion free energies of protein sequences. Bioinformatics. 24, 1271-1277 (2008)
6. Park, Y. and Helms, V.: MINS2: Revisiting the molecular code for transmembrane-helix recognition by the Sec61 translocon. Bioinformatics. 24, 1819-1820 (2008)
7. Laroum, S., Tessier, D., Duval, B., Hao, J. K.: A Local Search Appproach for Transmembrane Segment and Signal Peptide Discrimination. Lecture Notes in Computer Science. 6023, 134-145 (2010)
8. Cheng, Z.:Protein translocation through the Sec61/SecY channel. Bioscience Reports. 30, 201-207 (2010)
9. Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, SH., von Heijne, G.: Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature. 433, 377-381 (2005)
10. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Labs (2004)
11. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol .Biol. 157, 105-132 (1982)
12. Kall, L., Krogh, A., Sonnhammer, E.L.L.: A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338, 1027-1036 (2004)
13. 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)
14. Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S.: Improved prediction of signal peptides: SignalP 3.0. J. Mol .Biol. 340, 783-795 (2004)
15. Junker, V.L, Apweiler, R., Bairoch, A.: Representation of functional information in the SWISS-PROT data bank. Bioinformatics. 15, 1066-1067 (1999)
16. Eisenberg, D., Weiss, R.M., Terwillger, T.C: The helical hydrophobic moment - A measure of the amphiphilicity of a helix. Nature. 299-371 (1982)
17. Engelman, D.M., Steitz, T.A, Golman, A.: Identifing nonpolar transbilyer helices in amino acid sequences of membrane proteins. Annual Review of Biophysics and Biophysical Chemistry. 321-353 (1986)
18. Chamberlain, A.K., Lee, Y., Kim, S., Bowie, J.U.: Snorkeling preferences foster an amino acid composition bias in transmembrane helices. J. Mol .Biol. 339, 471-479 (2004)