Signal Peptide Prediction in Single Transmembrane Proteins Using the Continuous Wavelet Transform

I.A. Avramidou[1], I.K. Kitsas[1] and L.J. Hadjileontiadis[1]

[1]Aristotle University of Thessaloniki/ Dept. of Electrical & Computer Engineering, GR-54124 Thessaloniki, Greece

Abstract—**The signal peptide (SP) is directly associated with a protein's translocation across biological membranes and consequently with the expression of its functional role. A scheme of predicting the exact position of the SP within a protein is proposed in this work, by applying the continuous wavelet transform (CWT) to the hydrophobic sequence of the protein. The scheme was developed with regard to proteins of known structure extracted from public available databases. The results have verified the effectiveness of the method, which is comparable to existing methods, thus revealing a novel and fast approach to the prediction of SP in single transmembrane proteins, with a prospect of a generalized application.**

Keywords—**Signal peptide, prediction, transmembrane, proteins, continuous wavelet transform.**

I. Introduction

In both prokaryotic and eukaryotic cells, proteins are allowed entry into the secretory pathway only if they are endowed with a specific targeting signal: a signal peptide (SP). The SP is in most cases a transient extension to the amino terminus of the protein and is removed once its targeting function has been carried out [1].

A number of computational methods aiming at the prediction of the exact position of the SP in the primary amino acid sequence have been developed. The majority of them is based on neural networks [2], [3] trained and tested on a set of experimentally derived SPs from eukaryotes and prokaryotes, or on Hidden Markov Models (HMMs) [4]-[6], which model the different sequence regions of a SP in a series of interconnected, by transition probabilities, states. Likewise, other schemes have been proposed either implementing a position weight matrix approach [7], [8], or being based on sequence alignment techniques [9], or using support vector machines [10], [11].

However, most of the methods mentioned above incorporate some kind of data dependences or complexity, thus leaving room for further research on the prediction of SPs by applying novel detection schemes. Neural network methods [2], [3], for example, have to sacrifice the computational cost of training on the altar of better accuracy. Similarly, weight matrices become more precise as the amount of data on which they are based increases. Moreover, the general architecture of learning systems, such as neural networks and HMMs, makes it difficult to trace the cause of false predictions. Finally, sequence alignment techniques call for the maintenance of large sets of reference data.

In this paper, we have developed a method, namely Continuous Wavelet Transform – SP Detector (CWT-SPD), for the prediction of SPs, which is devoid of the drawbacks described above. Specifically, the method focuses on the identification of the last amino acid of the primary amino acid sequence that belongs to a SP, which in most cases is referred to as cleavage site, on the grounds that the SP is usually a transient extension to the amino terminus of the protein as mentioned before. The proposed algorithm is based on the CWT and it is applied to the arithmetic sequence produced by the conversion of the amino acid sequence to an arithmetic sequence, by means of the Kyte and Doolittle [12] hydrophobicity scale. The isolation of the area of CWT coefficients that represents the SP precedes the prediction, which is made according to the sum of coefficients across all scales. The method has been applied to a dataset of human proteins with a single transmembrane segment that are endowed with a reported SP. The results have indicated that CWT-SPD is a fast and effective approach to the problem of SP prediction.

II. Materials and Methods

A. Dataset Characteristics

The dataset used in this work consists of documented transmembrane proteins extracted from SWISS-PROT Release 46.0 [13]. From the initial set of 12108 human protein sequences, automatically selected based on the presence in the feature table of the 'TRANSMEM' keyword, a subset of 1390 sequences was extracted, containing all the proteins with a single-membrane segment. This subset was further refined by excluding similar sequences, so that every pair of sequences had less than 30% of sequence identity, thus resulting to 499 single transmembrane proteins. These proteins were further divided in two groups, with respect to the existence of a SP, by the presence of the 'SIGNAL' keyword in the feature table. The process described above resulted in the following subsets: one with 327 single transmembrane proteins with a SP and another, with 172 single

P.D. Bamidis and N. Pallikarakis (Eds.): MEDICON 2010, IFMBE Proceedings 29, pp. 65–68, 2010.
www.springerlink.com

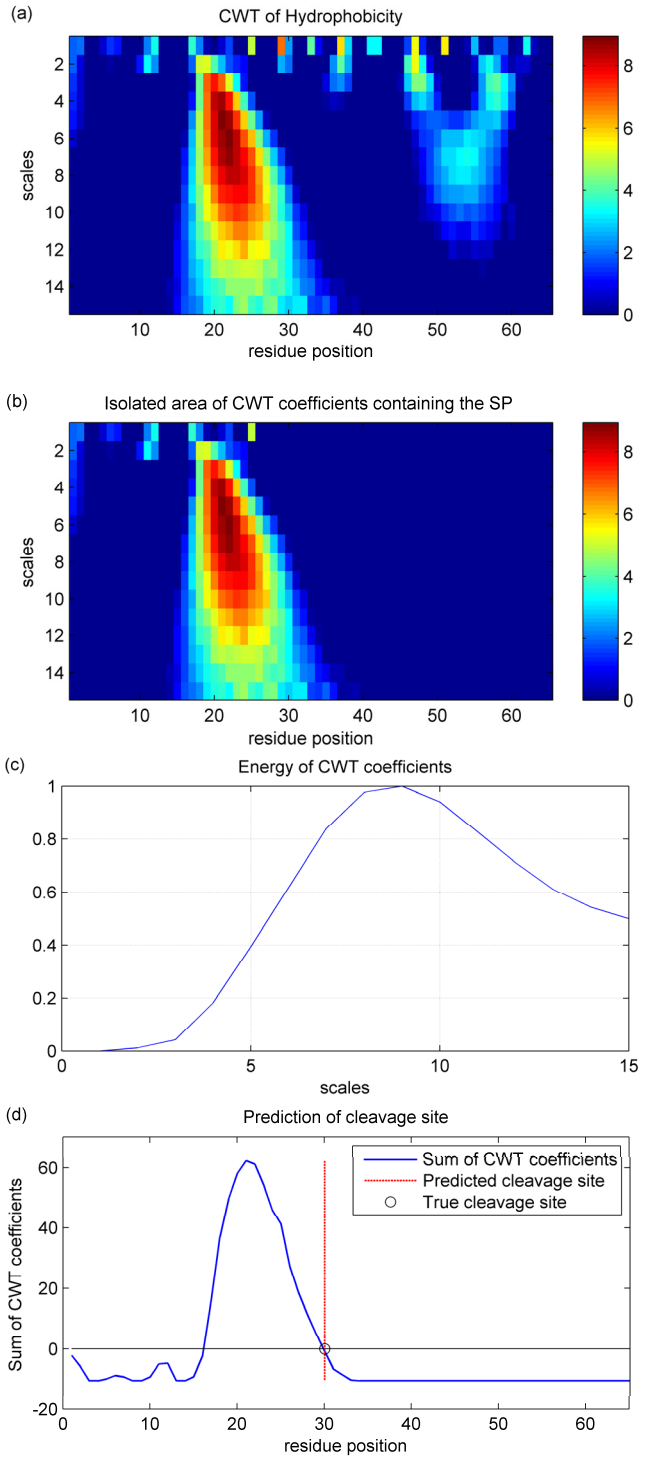

Fig. 1 (a) The magnitude of the CWT of hydrophobicity sequence derived from the transmembrane protein Q9Y5G1 (SWISS-PROT Release 46.0), (b) the isolated section of (a) corresponding to the SP, (c) the energy of the signal resulting by summation of the isolated coefficients across all residues, (d) prediction (red vertical line) of the cleavage site (circle) of the SP by summation of CWT coefficients across all scales.

transmembrane proteins without a SP in the feature table. The method was tested on data derived from the first of these subsets. The SPs taken as reference had an average length of 23 amino acids.

B. Continuous Wavelet Transform (CWT)

The continuous wavelet transform of a series of hydrophobic values, x(n), is defined as [14]

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_0^n x(n) \psi^* \left(\frac{n-b}{a} \right) dn \,, \qquad (1)$$

where α is a scaling parameter and b a dilation parameter; $(a,b \in \Re,\ a>0)$. n is the amino acid sequence length of the protein containing the SP, while $\psi(n)$ is the analyzing mother wavelet scaled by the factor a, and dilated by a factor b. The Mexican Hat Wavelet [15] was chosen for the realization of the CWT. This symmetrical wavelet, which is defined as the second derivative of the Gaussian probability density function, was selected in order to ensure common reference with other SP prediction methods.

C. Prediction of the Signal Peptide

The CWT coefficients are first thresholded thus keeping only positive values. Next the area of coefficients that represents the SP is identified and isolated, taking into account three parameters: a) the distance from the amino terminus of the vertical axis used to indicate it, b) the location of the coefficient with the greatest value, and c) the location and amplitude of the peaks resulting from the sum of the coefficients across all scales. A range of effective scales is dynamically selected in the following step, comprising at least 70% of the energy of the signal resulting by summation of the isolated coefficients across all residues. Finally, the prediction is made according to the zero-crossing point of the signal resulting from the sum of the coefficients across the range of effective scales. A characteristic example of the steps of SP prediction along with the modifications that take place at the CWT domain is depicted in Fig.1. In particular, Figs. 1(a) and 1(b) correspond to the CWT coefficients (before and after isolating the SP area), whereas Fig. 1(c) depicts the energy of the signal resulting by summation of the isolated coefficients across all residues. The cleavage site of the SP is determined by the zero-crossing, as illustrated in Fig. 1(d).

D. Evaluation and Performance Indices

The method proposed in this paper is evaluated according to the methodology described by Cuthbertson [16], which

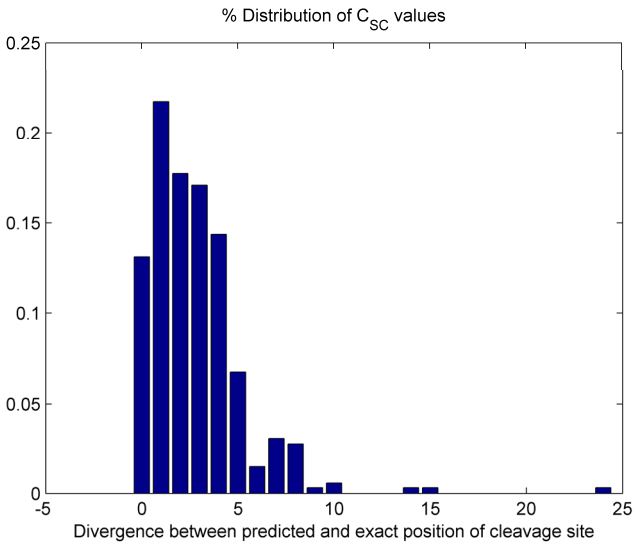

Fig. 2 The distribution of C_{sc} values for the CWT-SPD method applied to the entire dataset.

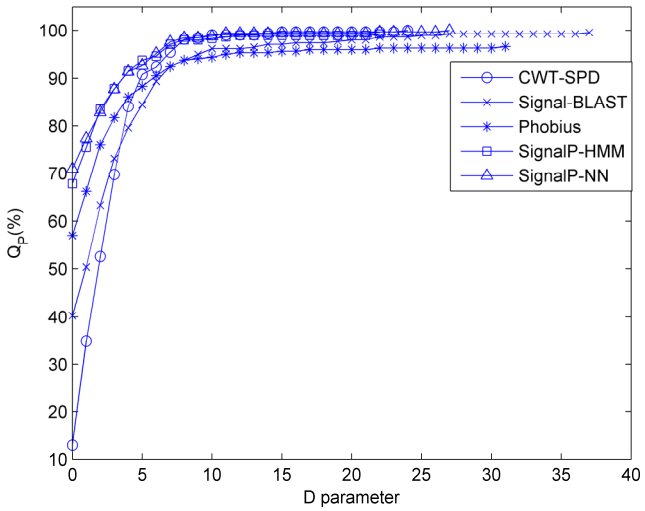

Fig. 3 Comparative analysis of the prediction power Q_P for the CWT-SPD, Signal-BLAST, Phobius, SignalP-HMM and SignalP-NN methods as a function of D parameter.

introduces the index C_{sc} defining it as the absolute deviation between the actual cleavage site and the predicted value.

As far as the performance of the method is concerned, it is measured by the index Q_P, defined by Tusnády and Simon [17], as follows:

$$Q_P = 100\sqrt{\frac{N_C}{N_O}\cdot\frac{N_C}{N_P}}, \qquad (2)$$

where N_C, N_P and N_O are the number of SPs that have been correctly predicted, that have been located and that actually exist, respectively.

Moreover, an index D, corresponding to the maximum acceptable value of C_{SC} in order to consider a SP prediction to be correct, was introduced and taken into account during the evaluation process and comparison of CWT-SPD with previous works.

III. Results

The application of CWT-SPD to the transmembrane protein Q9Y5G1 (SWISS-PROT Release 46.0) as already described is illustrated in Fig. 1. As shown in Fig.1(d) the zero-crossing of the signal resulting by summation of the CWT coefficients across all scales coincides with the true cleavage site (30th residue). The CWT-SPD was further applied to the entire set of single transmembrane proteins with a SP and initially evaluated by estimating the index C_{sc}, as shown in Fig. 2. In particular, the latter illustrates the distribution of the C_{SC} score across the proteins included in the set. From this figure, it is apparent that, for the most (90.8%) of the sequences of the dataset, the prediction of the cleavage site was within a distance of five residues away from the true position, thus limiting the estimation error within acceptable ranges. From the same figure, it is also clear that the distribution exhibits a peak (21.7%) corresponding to a deviation of just a single residue between the prediction and the actual cleavage site, whereas a percentage of 13.1% results in an exact match between the true and predicted position of the cleavage site.

The performance of the CWT-SPD regarding the examined dataset was also measured by means of the index Q_P, as shown in Fig. 3. This figure depicts the resulted Q_P values for the CWT-SPD method (line with circle markers), for the parameter D ranging from 0 to 24 for the proposed method. In the same figure, the resulting Q_P values for the methods Signal-Blast [18], Phobius [19], SignalP-NN and SignalP-HMM [20] are also shown, with the parameter D taking values up to 37 (Signal-BLAST).

IV. Discussion

The results of the proposed CWT-SPD method were compared to those derived from four efficient SP prediction methods reported in the literature. The four methods were applied on the aforementioned set of proteins using their Web-interface. An effort was made to include algorithms with a variety of methodologies, such as sequence alignment techniques (Signal-BLAST), HMMs (Phobius,

Table 1 Comparison of performance by Q_P index

Method	D=4	D=5	D=8
CWT-SPD	84.10	90.83	98.17
Signal-BLAST	79.57	84.49	93.70
Phobius	86.13	88.34	93.73
SignalP-NN	91.44	92.66	98.47
SignalP-HMM	91.53	93.69	98.00

D: maximum acceptable value of deviation between the actual cleavage site and the predicted value, Q_P: Prediction Power index for the entire dataset of single transmembrane proteins with signal peptide for three indicative values of D.

SignalP-HMM), and neural networks (SignalP-NN). As illustrated in Fig. 3, the CWT-SPD method superpasses the Signal-BLAST and Phobius methods for D taking values greater than three and four, respectively. Furthermore, a comparable performance is observed between the CWT-SPD method and the SignalP-HMM and SignalP-NN methods for more flexible values of the D parameter. Moreover, as shown in Fig. 3, it is clear that assigning values greater than eight to the parameter D has little effect on the ranking of the methods.

A comparative exhibition of the index Q_P regarding all the aforementioned methods, for indicative values of D is presented in Table 1. From this table it is apparent that Signal-Blast exhibits a rather poor performance compared to the rest of the methods with values of Q_P at least 3% lower than the three methods with best performance for all cases. As far as the Phobius method is concerned, the prediction power of 86.13%, when D equals four, is clearly comparable to the 84.1% of the CWT-SPD. However, the superiority of the latter is apparent for values of D greater than four. Regarding the SignalP-NN and SignalP-HMM methods it is obvious that they exhibit a similar behavior, better than the rest of the methods (Q_P>91%), however CWT-SPD, outperforms the latter for values of D greater than seven and favorably compares with the first (Q_P>98.17%). This is even more significant, considering the independence of the CWT-SPD method from training procedures and datasets.

V. Conclusions

An effective scheme of predicting the position of the SP in the primary amino acid sequence of single transmembrane proteins has been proposed. The effectiveness of the CWT in detecting special features of a numerical sequence justifies its selection for the detection of the SPs in an amino acid sequence. Ongoing work is on the way towards the expansion of the method to different datasets, the improvement of the overall performance with respect to the distribution of the CWT coefficients across high and low scales and the characterization of a protein as to whether it contains a SP or not.

References

1. Von Heijne G (1990) The signal peptide. J Membr. Biol 115:195-201
2. Nielsen H, Engelbrecht J, Brunak S, Von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1-6
3. Fariselli P, Finocchiaro G, Casadio R (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics 19:2498-2499 DOI 10.1093/bioinformatics/btg360
4. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model, ISMB Proc., Int. Conf. Intell. Syst. Mol. Biol., Montreal, Canada, 1998, pp 122-130
5. Zhang Z, Wood W (2003) A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics 19:307-308
6. Käll L, Krogh A, Sonnhammer L (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027-1036 DOI 10.1016/j.jmb. 2004. 03. 016
7. Von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683-4690
8. Hiller K, Grote A, Scheer M, Münch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32:375-379 DOI 10.1093/nar/gkh378
9. Frank K, Sippl M (2008) High-performance signal peptide prediction based on sequence alignment techniques. Bioinformatics 24:2172-2176 DOI 10.1093/bioinformatics/btn422
10. Chou KC (2001) Prediction of protein signal sequences and their cleavage sites. Proteins 42:136-139
11. Vert JP (2002) Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings, PSB Proc., Pac. Symp. Biocomput., Lihue, Hawaii, USA, 2002, pp 649-660
12. Kyte J, Doolittle R (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132
13. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin M, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365-370
14. Addison P (2002) The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine and finance. Institute of Physics (IOP) Publishing, Bristol
15. Daubechies I (1992) Ten lectures on wavelets. SIAM
16. Cuthbertson A, Doyle D, Sansom M (2005) Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 18:295-308
17. Tusnády G, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283:489-506
18. Signal-BLAST at http://sigpep.services.came.sbg.ac.at/signalblast.html
19. Phobius at http://phobius.sbc.su.se/
20. SignalP 3.0 Server at http://www.cbs.dtu.dk/services/SignalP

Author: I. A. Avramidou, I. K. Kitsas and L. J. Hadjileontiadis
Institute: Dept. of Electrical & Computer Engineering, Aristotle University of Thessaloniki
Street: University Campus, GR-54124
City: Thessaloniki
Country: Greece
Email: io.avramidou@gmail.com, {ikitsas, leontios}@auth.gr