

WAVPRED: A WAVELET-BASED ALGORITHM FOR THE PREDICTION OF TRANSMEMBRANE PROTEINS*

PATRICK FISCHER[†], GUY BAUDOUX[‡], AND JOHAN WOUTERS[§]

Abstract. Mathematical filtering using wavelet transform was applied to analyse hydrophathy signals of membrane proteins. The accuracy of our localization of transmembrane approaches that of well-established methods. The analysis of hydrophobicity plots using wavelets presents advantages with respect to other “filtering” methods based on fixed windows or Fourier transforms and compared to “training” (neural networks) techniques. Although the method embodies principles that have long been appreciated, its simplicity makes it a very useful tool for the evaluation of protein membrane-spanning segments. Generalization of use of wavelets should be encouraged in other aspects of bioinformatics.

1. Review on Wavelets

Since all the results presented in the sequel are one dimensional, only 1D wavelets theory is introduced in this part. The generalization to higher dimension is relatively easy and is based on tensor products of basis functions.

Any signal, which can be seen as a 1D mathematical function, can be represented by a sum of fundamental or simple functions called basis functions. The most famous example, the Fourier series

$$s(t) = \sum_{k=-\infty}^{+\infty} c_k e^{ikt} \quad (1.1)$$

is valid for any 2π -periodic function sufficiently smooth. Each basis function e^{ikt} is indexed by a parameter k which is related to a frequency. In (1.1), $s(t)$ is written as a superposition of harmonic modes with frequencies k . The coefficients c_n are given by the integral

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} s(t) e^{-ikt} dt. \quad (1.2)$$

Each coefficient c_k can be viewed as the average harmonic content of $s(t)$ at frequency k . Thus the Fourier decomposition gives a frequency representation of any signal. The computation of c_k is called the decomposition of s and the series on the right-hand side of (1.1) is called the reconstruction of s .

Although this decomposition leads to good results in many cases, some disadvantages are inherent to the method. One of them is the fact that all the information concerning the space behavior of the signal is completely lost in the Fourier description. For instance, a discontinuity or a localized high variation of the frequency will not be well described by the Fourier representation. The underlying reason lies in the nature of complex exponential functions used as basis functions. They all cover the entire real line and differ only with respect to frequency. They are not suitable for representing the behaviour of a discontinuous function or a signal with high localized oscillations. Like the complex exponential functions of the Fourier decomposition, wavelets can be

*Received: January 25, 2002; Accepted (in revised version): April 23, 2002.

[†]MAB, Université de Bordeaux 1, 351 Cours de la Libération, 33405 Talence, France (fischer@math.u-bordeaux.fr).

[‡]SmithKline Beecham Biologicals, Rixensart, Belgium.

[§]Facultés Universitaires Notre Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Belgium.

used as basis functions for the representation of a signal. But, unlike the complex exponential functions, they are able to restore the positional information as well as the frequency information.

1.1. Continuous Wavelets. Functions depending on two variables a and b respectively linked to frequency and position are used to define the mathematical transformation:

$$d_b^a = \int dx f(x) \psi_{a,b}(x), \quad (1.3)$$

where $\psi_{a,b}(x)$ plays the same part as the exponential functions in the Fourier transform. A possibility is to construct $\{\psi_{a,b}(x)\}_{a,b}$ from a function $g(x)$ by translating and modulating it:

$$\psi_{a,b}(x) = g(x - b) e^{iax}, \quad a, b \in \mathbb{R} \quad (1.4)$$

where $g(x)$ is a window function. In spite of the improvement brought by this ‘‘pseudospectral’’ representation, this transformation is not perfect, and in particular it is not adapted to describe accurately functions which exhibit high variations. This kind of phenomenon is generally very localized in space whereas low variations often spread over a large area. To overcome this disadvantage (a fixed-size window function), analyzing functions with position support widths adapted to their frequency need obviously to be defined.

The idea is to apply dilations on top of translations previously introduced. Starting with a function ψ well localized in position and frequency spaces, a family of analyzing functions can be constructed:

$$\psi_{a,b}(x) = |a|^{1/2} \psi\left(\frac{x-b}{a}\right), \quad a \in \mathbb{R}^*, b \in \mathbb{R}. \quad (1.5)$$

The initial function is called the mother wavelet. Here, b is a position parameter and $1/a$ is homogeneous to a frequency.

The continuous wavelet transform is an isometry from $L^2(\mathbb{R})$ into $L^2(\mathbb{R} \times \mathbb{R}, a^{-2} dadb)$.

Similar to the definition of the inverse Fourier transform, it is possible to define a reconstruction formula that allows to rewrite $f(x)$ as an expansion. The following theorem specifies few characteristics of continuous wavelet theory:

Let ψ be a normalized function belonging to $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, of which the Fourier transform $\hat{\psi}$ verifies the following equality:

$$\int d\xi \frac{|\hat{\psi}|^2}{|\xi|} = K < \infty. \quad (1.6)$$

Then, the conservation of the norm defined by,

$$\frac{1}{K} \int \int \frac{dad b}{a^2} |d_b^a|^2 = \int dx |f(x)|^2, \quad (1.7)$$

and the possibility to recover the function $f(x)$ using the reconstruction formula defined as follows:

$$f(x) = \frac{1}{K} \int \int \frac{dad b}{a^2} d_b^a \psi_{a,b}(x), \quad (1.8)$$

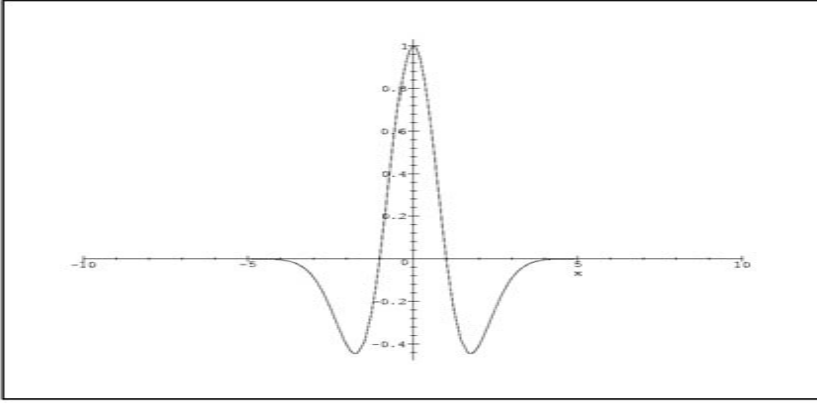


FIG. 1.1. The “Mexican hat” mother wavelet.

are ensured. The condition (1.6) means that any oscillating function localized in both spaces and whose integral over the whole space \mathbb{R} is null can be used as a mother wavelet. A typical choice for ψ is,

$$\psi(x) = (1 - x^2)e^{-\frac{x^2}{2}} \quad (1.9)$$

the second derivative of the Gaussian function, sometimes called the “Mexican hat” function (Figure 1.1).

1.2. Discrete and Orthonormal Wavelets. Discrete wavelets correspond to the choices $a = a_0^m, b = nb_0a_0^m$, indicating that the translation parameter b depends on the chosen dilation rate. The family of wavelets becomes, then, for $m, n \in \mathbb{Z}$,

$$\psi_{m,n} = a_0^{-m/2} \psi(a_0^{-m}x - nb_0). \quad (1.10)$$

The dilation step a_0 is generally taken greater than one and the translation step b_0 different from zero. It is also possible to define wavelets that constitute an orthonormal basis. They are defined as the collection,

$$\psi_{j,k} = 2^{-j/2} \psi(2^{-j}x - k), j, k \in \mathbb{Z}. \quad (1.11)$$

The simplest and most famous example of orthonormal wavelet basis is the Haar system already known at the beginning of the century [6]:

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 1/2 \\ -1 & 1/2 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The construction of orthonormal wavelet bases is presented in the following section.

1.3. Multiresolution Analysis. The theoretical construction of orthogonal wavelet families is intimately related to the notion of multiresolution analysis:

DEFINITION: A multiresolution analysis is a decomposition of the Hilbert space $L^2(\mathbb{R})$ of physically admissible functions (i.e square integrable functions) into a chain of closed subspaces,

$$\dots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \dots$$

such that

- $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$ and $\bigcup_{j \in \mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$
- $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j-1}$
- $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$
- There is a function $\varphi \in V_0$ such that $\{\varphi(x-k)\}_{k \in \mathbb{Z}}$ is an orthonormal basis of V_0

Let W_j be the orthogonal complementary subspace of V_j in V_{j-1} :

$$V_j \oplus W_j = V_{j-1}. \quad (1.13)$$

This space contains the difference in information between V_j and V_{j-1} , and allows the decomposition of $L^2(\mathbb{R})$ as a direct form:

$$L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j. \quad (1.14)$$

Then, there exists a function $\psi \in W_0$, called the mother wavelet, such that $\{\psi(x-k)\}_{k \in \mathbb{Z}}$ is an orthonormal basis of W_0 . The corresponding wavelet bases are then characterized by:

$$\varphi_{j,k}(x) = 2^{-j/2} \varphi(2^{-j}x - k), \quad k, j \in \mathbb{Z}, \quad (1.15)$$

$$\psi_{j,k}(x) = 2^{-j/2} \psi(2^{-j}x - k), \quad k, j \in \mathbb{Z}. \quad (1.16)$$

The mother wavelet corresponding to the chosen wavelet basis verifies:

$$\int_{\mathbb{R}} dx \psi(x) x^m = 0, \quad m = 0, \dots, M-1, \quad (1.17)$$

which means that it has M vanishing moments.

Since the scaling function $\varphi(x)$, and the mother wavelet $\psi(x)$ belong to V_{-1} , they admit the following expansions:

$$\varphi(x) = \sqrt{2} \sum_{k=0}^{L-1} h_k \varphi(2x - k), \quad h_k = \langle \varphi, \varphi_{-1,k} \rangle, \quad (1.18)$$

$$\psi(x) = \sqrt{2} \sum_{k=0}^{L-1} g_k \varphi(2x - k), \quad g_k = (-1)^k h_{L-k-1}, \quad (1.19)$$

where the number L of coefficients is connected to the number M of vanishing moments and is also connected to other properties that can be imposed to $\varphi(x)$. Functions verifying (1.18) or (1.19) have their support included in $[0, \dots, L-1]$. Furthermore, if there exists a coarsest scale, $j = n$, and a finest one, $j = 0$, the bases can be rewritten as:

$$\varphi_{j,k}(x) = \sum_{l=0}^{L-1} h_l \varphi_{j-1, 2k+l}(x), \quad j = 1, \dots, n, \quad (1.20)$$

and

$$\psi_{j,k}(x) = \sum_{l=0}^{L-1} g_l \varphi_{j-1,2k+l}(x), \quad j = 1, \dots, n. \quad (1.21)$$

The wavelet transform of a function $f(x)$ is then given by two sets of coefficients defined as

$$d_k^j = \int_{\mathbb{R}} dx f(x) \psi_{j,k}(x), \quad (1.22)$$

and

$$s_k^j = \int_{\mathbb{R}} dx f(x) \varphi_{j,k}(x). \quad (1.23)$$

Starting with an initial set of coefficients s_k^0 , and using (1.20) and (1.21), coefficients d_k^j and s_k^j can be computed by means of the following recursive relations:

$$d_k^j = \sum_{l=0}^{L-1} g_l s_{2k+l}^{j-1}, \quad (1.24)$$

and

$$s_k^j = \sum_{l=0}^{L-1} h_l s_{2k+l}^{j-1}. \quad (1.25)$$

Coefficients d_k^j and s_k^j are considered in (1.24) and (1.25) as periodic sequences with the period 2^{n-j} . The set d_k^j is composed by coefficients corresponding to the decomposition of $f(x)$ on the basis $\psi_{j,k}$, and s_k^j may be interpreted as the set of averages at various scales. The principle of applying (1.24) and (1.25) is usually symbolized by the pyramid scheme (Figure 1.2).

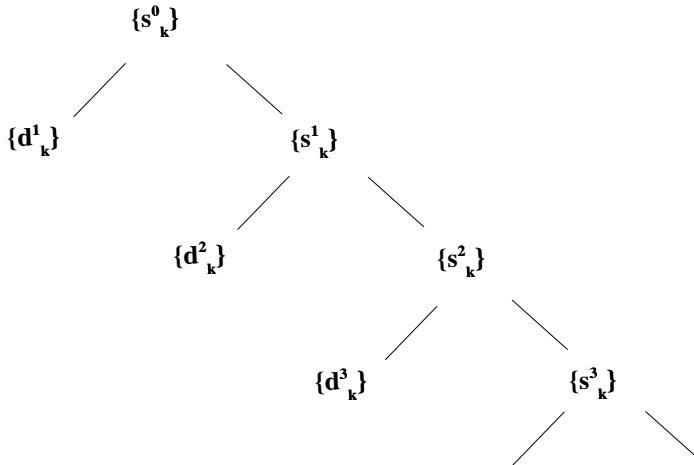
In practice, the wavelet expansion (like the Fourier one) must be truncated at some finest and coarsest scales. A Fast Wavelet Transform based on the pyramid scheme is used to compute this expansion with a very low computational cost (in the same way as the FFT is used for a Fourier expansion) [11].

The mathematical construction of the wavelet bases (particular choice for the mother wavelet) is not described in this paper, all the details can be found in the literature [5, 14, 1].

Here we want to present an original application of wavelet analysis to the prediction of transmembrane helices using hydrophobicity scales.

2. HTMs Prediction

A well-known problem in protein modelling is the prediction of the position of transmembrane helices (HTMs) in protein sequences. One of the most commonly used methods in this field is the interpretation of hydropathy profiles. This method was first introduced by Kyte and Doolittle [10] who used a window of 19 residues to smooth the hydropathy data, to enable the detection of potential transmembrane helices as peaks in a two dimensional plot. Rao and Argos [16], using several physico-chemical properties of amino acids and repeated window smoothing further refined the method.

FIG. 1.2. *The pyramid scheme.*

Following the same concept, Von Heijne [7] proposed the use of a trapezoidal window to smooth the hydropathy data, and Edelman [2] presented an optimal predictor (a window with position weights and amino acids weight scale) for the prediction of HTMs. Those methods perform quite well (according to Sonnhammer et al. [19]). They were, however, outmoded by more recent programs including TMAP [15], MEMSAT [8], PHDhtm [17, 18] or TMHMM [19].

Despite these recent developments, predictions based on hydropathy analysis remain interesting tools for theoretical and practical reasons. Theoretically, approaches based on hydrophobicity scales can “provide insight into the principles governing the formation of secondary structures” (quoting King and Sternberg [9] when they compared their secondary structure method, DSC, to neural network techniques).

Practically, hydrophobicity analysis enables the rapid scanning of entire genomes, in searches for transmembrane helices, which are indicators of potential antigens or drug receptors.

3. Methods

The method is schematically presented on Fig. 3.1. The protein sequence is first converted into a “hydropathy signal” using the scales of Kyte and Doolittle [10], Eisenberg [3], or GES [4], and the mean hydropathy of the whole protein $\langle H \rangle$ is computed (Step (1)). The numerical sequence is then subjected to a mathematical filtering (grey square I), followed by a “biochemical” filtering (grey square II).

In the mathematical filtering, the wavelet transform is applied using a mother wavelet with 10 vanishing moments, the so-called Daubechies 20 [1] (Step (2)). Then, in the filtering step (3), the wavelet coefficients corresponding to high frequencies (for frequency parameter $j = 0$ to $j = 4$), are turned to zero, and finally, the signal is reconstructed and around zero by subtraction of its mean value (Step (4)), producing a zero mean filtered hydropathy signal whose maximum, m , is used in the subsequent

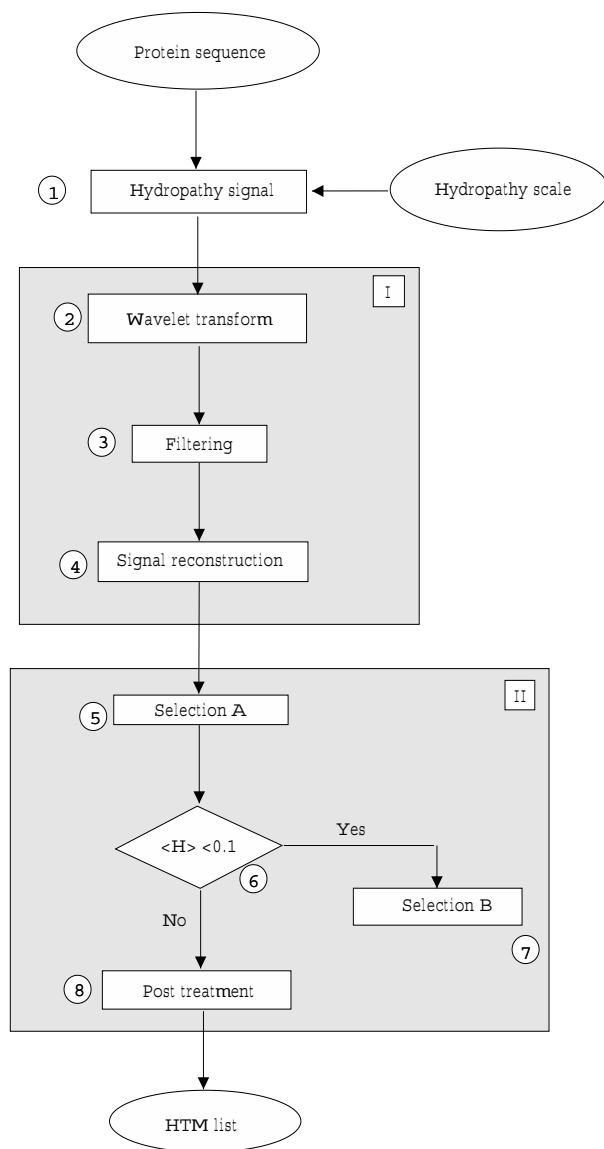


FIG. 3.1. General flowchart presenting the different steps of both the mathematical filtering (grey square I) and the “biochemical” filtering (grey square II) implemented in our filtering method based on wavelets.

treatment. The case of the CYDB-ECOLI protein sequence is shown to illustrate the action of the mathematical filtering. The linear (lettered) sequence (Figure 3.2) is first numerized using a hydrophathy scale, here the so-called Kyte-Doolittle scale (Figure 3.3). Then the wavelet filtering is applied leading to a smoother signal (Figure 3.4).

```

MIDYEVLRFIWWLLVGVLLIGFAVTDGFDMGVGMLTRFLGRNDTERRI
MINSIAPHWDGNQVWLITAGGALFAAWPMVYAAAFSGFYVAMILVLA
SLFFRPVGFDRYSKIEETRWNRNMWDWGIFIGSFVPLVIGVAFGNLLQG
VPFNVDEYLRLYYTGNFFQLNPFGLLAGVVSVGMIIQTGATYLMQRT
VGELHLRTRATAQVAALVTLVCFALAGVWVMYGIDGYVVKSTMDHYA
ASNPLNKEVVREAGAWLVNFNNTPIWAIWALGVVLPPLLITLARMMDKA
AWAFVFSSLTACIILTAGIAMFPFVMPSTMMNASLTMWDATSSQLT
LNVMTWVAVVLPVILLYTAWCYWKMFGKITKEDIERNTHSLY

```

FIG. 3.2. Linear sequence of the CYDB-ECOLI protein.

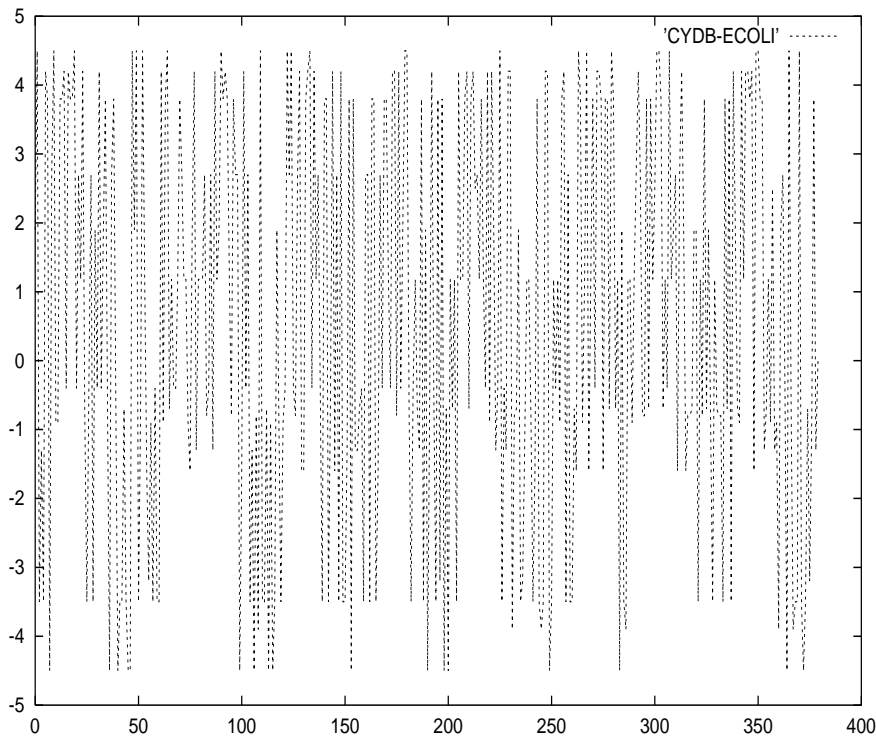


FIG. 3.3. Hydrophathy plot of the CYDB-ECOLI sequence.

In the “biochemical” filtering step, a tentative list of HTMs is created first, by selection of all amino acids having a positive filtered hydrophathy (Step (5)), and for each transmembrane segment s so defined, the number of amino acid is computed N_s .

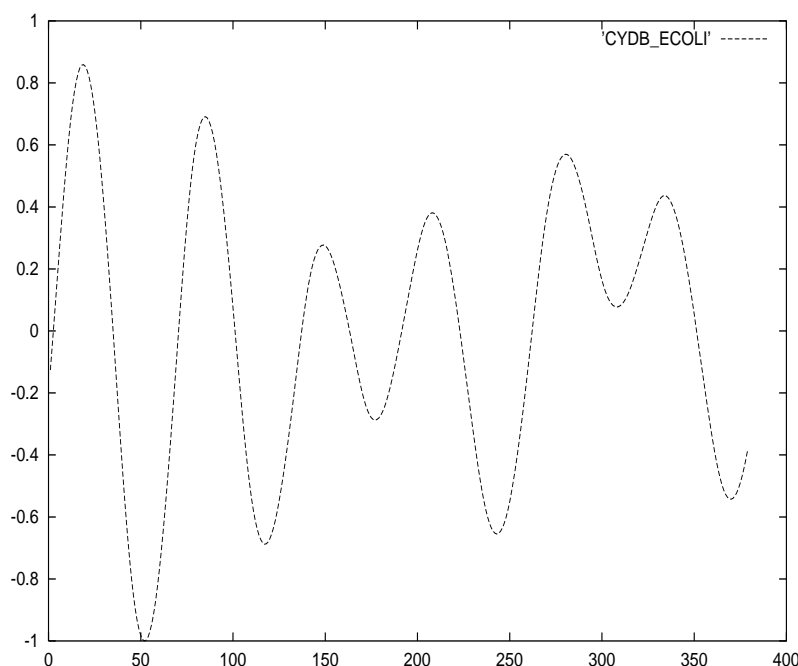


FIG. 3.4. *Filtered hydropathy plot of the CYDB-ECOLI sequence.*

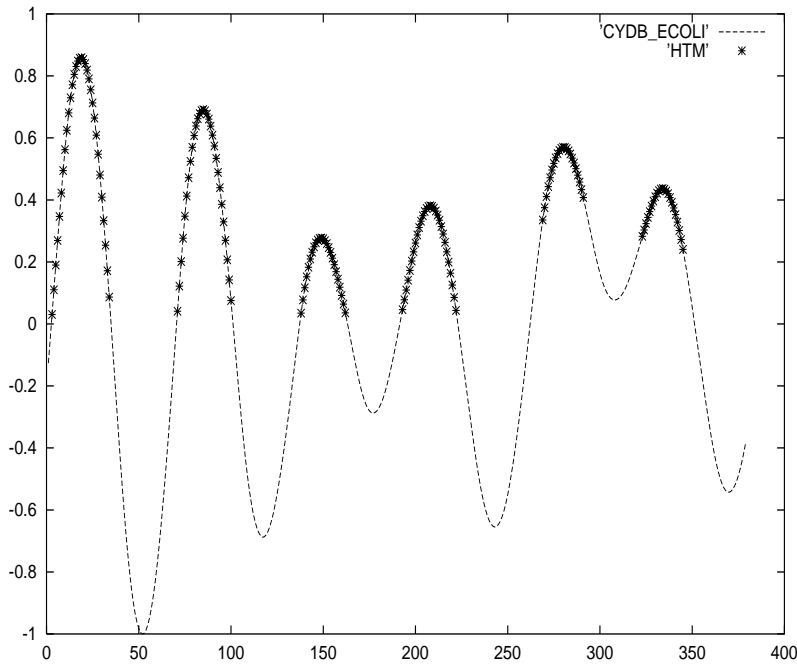
This tentative list is re-examined in two steps:

- in a cutting procedure inspired from Rost et al. [17]:
 - if N_s is smaller than 11 residues, s is discarded.
 - if N_s is between 38 and 72 residues, s is cut into 2 equal parts, declaring the 3 central amino acids as non-transmembranar.
 - if N_s is greater than 72 residues, s is cut into 3 or more parts of 24 residues, declaring the 3 amino acids at the limits as non transmembranar.
- in an overall hydrophobicity filtering:
 - if $\langle H \rangle$ is less than 0.1 (Step (6)), a second selection (Step (7)) of segments is made using the following criteria :
 - if $\langle H \rangle$ is less than -0.2, only predicted segments for which the maximum of filtered hydropathy is greater than $m.0.7$ are considered as correct;
 - if $\langle H \rangle$ is between -0.2 and 0.1, only predicted segments for which the maximum filtered hydropathy is greater than $m.0.5$ are considered as correct.

The “biochemical” filtering applied to the CYDB-ECOLI protein sequence gives a prediction of 8 well localized HTMs (Figure 3.5). This prediction can be compared to the experimental datas (Table 3.1).

4. Material

The 83 membrane protein sequences test set, and the complementary 48 sequences used by Rost et al. [18] were obtained from the authors. Together, they form the

FIG. 3.5. *HTMs prediction for the CYDB-ECOLI sequence.*TABLE 3.1. *Predicted and observed HTMs for the CYDB-ECOLI sequence.*

HTMs	1	2	3	4	5	6	7	8
Observed	9-28	80-99	123-142	165-184	206-225	263-282	293-312	337-356
Predicted	1-30	73-98	126-144	164-182	207-227	259-286	290-318	337-357

131-test sequences set. The 160-membrane sequences test set was retrieved from the World Wide Web site of the TMHMM: (<http://www.cbs.dtu.dk/krogh/TMHMM/>) [19]. The programs were first written for the MATLAB software, then an independant C version called WavPred has been created. Codes can be obtained from the authors upon request.

5. Results and Discussion

The analysis of hydrophobicity plots using wavelets presents advantages with respect to other filtering methods based on fixed windows or Fourier transforms. The display of the hydrophobic character of a protein using the Kyte and Doolittle technique [10] was the first successful empirical algorithm for transmembrane helix prediction. It is based on a fixed window filtering procedure. Results are, however, dependent on the size of the window and can lead to different localization of HTMs. This procedure is implemented (as black box) in a variety of software. Non specialized users may not be aware of the influence of the size of the window and produce useless

results. Beside the need of an *a priori* determination of the window size, this method is also unable to render the signal at the borders (typically there is a blind region at the beginning and the end of the sequence).

The Fourier transform reconstruction (obtained by using only the low frequencies) is also unable to produce an accurate hydrophathy graph. Indeed, the peaks in a graph obtained after Fourier analysis may not correspond to the position of HTMs. This is inherent to the Fourier analysis, whose basis functions spread all over the real line. In their construction process (Step (4) in Fig. 3.1), the low-frequency components are distributed all along the sequence, even in the parts where they are not necessary.

In the wavelet reconstruction (using only the low-frequency coefficients) the basis functions are localized and thus lead to a more accurate (locally) filtered hydrophathy signal. The plots obtained in this way are easy to interpret and indicate the precise localization of the HTMs of a protein.

We have applied our filtering procedure to the three sequences sets described in the materials section. The results, characterised by commonly used performance indices, are presented in Table 6.1. In this table, data compiled from the literature were also collected.

At first inspection, our method compares favorably with highly accurate ones. The sensitivity of 94.4% for the 83-sequences set is near the 96.2% of TMHMM and the 98.8 of PHDhtm. This is also the case for the specificity. The Q2, not given for all methods, is good for the 160-sequences set (89.7%), but much weaker (76.2) for the 83-sequences set compared to PHDhtm. Our method is so able to distinguish accurately between HTM and non-HTM regions in protein sequences. The good values of the sensitivity and selectivity, obtained on isolated sequences, show that a totally general property (hydrophathy), if filtered correctly, can have important predictive power. This is an advantage over methods based on a training procedure with known protein sequences and whose results with totally new sequences can be doubtful.

However, the QM is rather low: 60.0% compared to the 80–90% of the best methods. That means that the method is unable, at this stage of development, to predict the whole set of HTMs of a protein. As a consequence, an HTM topology hypothesis cannot be based on the predictions. A detailed examination of the results can, however, explain this fact. Indeed, in the 160-sequences set, 49 observed HTMs were not detected. Examined individually, those segments are generally less hydrophobic than the others and so fall below our cut off criteria. In further developments, additional criteria should be included.

6. Conclusions and Perspectives

The increasing number of DNA and protein sequences entering databases makes necessary the use of rapid and faithful algorithms to predict protein structures (and functions). Among proteins, membrane proteins are usually involved in important functions, but prediction of their structure represents a real challenge, because their three-dimensional structures are often very difficult to obtain experimentally. In this perspective, accurate localization of transmembrane hydrophobic stretches along the amino acid sequence yields valuable information.

In this work, we show that the use of wavelet analysis is an interesting alternative to existing methods. This procedure allows the rapid generation of clear hydrophathy plots useful to locate and identify segments rich in apolar residues. Due to its simplicity and its graphical nature, we think that the wavelet filtering of hydrophathy plots can have its place in general protein analysis packages.

To reach a good level of accuracy in the detection of HTMs, our method is depen-

dent upon the post-treatment of the signal obtained after wavelet reconstruction. To increase the performances and so make HTM topology predictions feasible, the use of complementary physico-chemical properties is needed. Alternatively, more sophisticated wavelet analysis techniques (wavelet packet methods for instance) could also be used in order to extract additional features of the hydrophathy plots (similar wavelet techniques were already used by Mandell et al. [12, 13] in mode matches between the hydrophobicity of ligands and receptors).

Moreover, our results are obtained on single-sequence analysis. It is an advantage because the sequences obtained from large genome sequence projects are in some cases unrelated. But we expect better results from the usage of aligned sequences, as was observed by Rost et al. [17]. Generalization of the usage of wavelets should also be encouraged in other domains of bioinformatics, in particular for the prediction of globular helices, transmembranar strands, denoising of dot plots, etc.

TABLE 6.1. *Numerical results; (N_{prot} : number of proteins in the test set; N_{obs} , number of observed HTMs; Q_2 , percentage of residues predicted correctly in either of the two states, HTM or not-HTM; Q_M , percentage of proteins for which all HTMs were predicted correctly; N_{pred} , number of HTM predicted; N_{corr} , number of HTM correctly predicted (a predicted HTM is counted as correct if there is an overlap of at least 5 residues with a true helix); Sens (sensitivity), percentage of observed HTMs correctly predicted; Spec (Specificity), percentage of predicted HTMs that are correct. nd: not determined).*

Method	N_{prot}	N_{obs}	Q_2	Q_M	N_{pred}	N_{corr}	Sens	Spec	Ref
OP	37	195	88.0%	43.2%	221	193	99.0%	87.3%	4
TMAP	28	126	nd	82.1%	125	123	97.6%	98.4%	6
TopPred	26	134	nd	73.1%	142	134	100%	94.4%	3
PHDhtm	83	341	93.6%	88.0%	354	337	98.8%	95.2%	8,9
PHDhtm	131	539	93.8%	89.3%	552	533	98.9%	96.6%	8,9
MEMSAT	83	341	nd	80.7%	nd	nd	96.8%	94.6%	7
MEMSAT	160	696	nd	73.8%	nd	nd	93.3%	95.6%	7
TMHMM	83	341	nd	83.1%	nd	nd	96.2%	97.6%	5
TMHMM	160	696	nd	83.8%	nd	nd	97.1%	97.7%	5
WavPred	83	341	76.2%	63.9%	352	322	94.4%	91.5%	
WavPred	129	535	75.5%	55.1%	536	487	91.1%	90.9%	
WavPred	160	696	89.7%	60.0%	696	647	93.0%	93.0%	

REFERENCES

- [1] I. Daubechies, *Orthogonal bases of compactly supported wavelets*. Comm. Pure Appl. Math. 41:909–996, 1988.
- [2] J. Edelman, *Quadratic minimization of predictors for protein secondary structure: Application to transmembrane alpha-helices*. J. Mol. Biol. 232:165–191, 1993.

- [3] D. Eisenberg, R. Terwillinger, and W. Wilcox, *Hydrophobic moments and protein structure*. Faraday Symp. Chem. Soc. 17:109–120, 1982.
- [4] D. Engelman, T. Steitz, and A. Goldman, *Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins*. Anni. Rev. Biophys. Chem. 15:321–353, 1986.
- [5] A. Grossmann and J. Morlet, *Decomposition of hardy function into square integrable wavelets of constant shape*. SIAM J. Math. Anal. 15:723–736, 1984
- [6] A. Haar, *Zur theorie der orthogonalen functionensysteme*. Math. Ann., 69:331+, 1910.
- [7] G. Von Heijne, *Membrane protein structure prediction hydrophobicity analysis and the positive-inside rule*. J. Mol. Biol. 225:487–494, 1992.
- [8] D. Jones, W. Taylor, and J. Thornton, *A model recognition approach to the prediction of all-helical membrane protein structure and topology*. Biochemistry 235:1501–1531, 1994.
- [9] R.D. King and M. Sternberg, *Identification and application of the concepts important for accurate and reliable protein secondary structure prediction*. Protein Science. 5:2298–2310, 1996.
- [10] J. Kyte and R.F. Doolittle, *A simple method for displaying the hydropathic character of proteins*. J. Mol. Biol. 157:105–132, 1982.
- [11] S. Mallat, *Multiresolution approximation and wavelets orthogonal bases of L^2* . Trans. Am. Math. Soc. 315:69–87, 1989.
- [12] A. Mandell, K. Selz, and M. Shlesinger, *Mode matches and their locations in the hydrophobic free energy sequences of peptide ligands and their receptor eigenfunctions*. Proc. Natl. Acad. Sci. USA. 94:13576–13581, 1997.
- [13] A. Mandell and M. Owens, et al, *Mode Matches in Hydrophobic free energy eigenfunctions predict peptide-protein interactions*. Biopolymers. 46:89–101, 1998.
- [14] Y. Meyer, *Ondelettes et opérateurs I. ondelettes*. Actualités Mathématiques, 1990
- [15] B. Persson and P. Argos, *Prediction of transmembrane segments in protein utilising multiple sequence alignments*. J. Mol. Biol. 237:527–535, 1994.
- [16] J. Rao and P. Argos, *A conformational preference parameter to predict helices in integral membrane proteins*. Biochem. Biophys. Acta. 869:197–214, 1986.
- [17] B. Rost, R. Castadio, P. Fariselli, and C. Sander, *Transmembrane helices predicted at 95% accuracy*. Protein Sci. 4:521–533, 1995.
- [18] B. Rost, P. Fariselli, and R. Castadio, *Topology prediction for helical transmembrane proteins at 86% accuracy*. Prot. Sci. 5:1704–1718, 1996.
- [19] E. Sonnhammer, G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences*. Intelligent Systems in Molecular Biology. Menlo Park, CA: AAAI Press, 175–182, 1998.