Improving Prediction of Hydrophobic Segments along a Transmembrane Protein Sequence using Adaptive Multiscale Lifting

Marina I. Knight[*], Guy P. Nason

September 9, 2005

Abstract

Abstract: Established methods for transmembrane protein segment prediction are often based upon hydrophobicity analysis. Classical wavelet multiscale methods have proved successful in the prediction task. However, they implicitly model protein chain residues as being equally spaced. Our main motivation is to challenge this assumption by developing a new multiscale 'lifting' technique that utilizes irregularly spaced residues, where the spacing is derived from resolved 3D information obtained from similar aligned proteins. For different protein families we calculate asymmetrical dissimilarity matrices of order 20 that estimate the 'distance' between residue types. We use our new adaptive lifting technique to regress the Kyte and Doolittle hydrophobicity index upon residues (now irregularly spaced using information from the distance matrices) and use the regression to predict transmembranar segments. We compare the results obtained through our method with the ones obtained through the usage of classical wavelets, and show that incorporating 3D resolved structure improves overall prediction (both in terms of the existence of predicted segments compared to experimentally determined ones and also the proportion of correctly predicted segments).
The software is available from http://www.maths.bris.ac.uk/∼maxmp/proteomics.html

1 Introduction

Membrane proteins are an important class of protein structures, but the experimental determination of their three-dimensional (3D) structure can often be very difficult. For this type of protein, predicting various structural aspects using only the information contained in the residue sequence is a problem of interest, see for example Lio and Vannucci (2000).

Since transmembrane proteins are found spanning the plasma membrane, the segments embedded in the lipid bilayer primarily consist of hydrophobic amino acids, and this feature can be used in order to identify them. Typical methodology for predicting transmembrane segments includes *hydrophobicity analysis* focused on helical transmembrane proteins— for example Kyte and Doolittle (1982), Engelman *et al*. (1986), Lio and Vannucci (2000). However, hydrophobicity analysis as a tool for prediction is not limited at transmembrane segments only, but has also been used for hydrophobic cores of globular proteins (see Hirakawa *et al*. (1999) for instance).

Wavelet based smoothing methods (Lio and Vannucci (2000), Fisher *et al*. (2003)) have been used and shown to perform well in the task of transmembrane segment prediction. So far, classical wavelet methods have been used. This means that the residues within the protein chain are modelled as being equally spaced. If each residue is thought of as a 3D structure determined by its atoms, then plausibly one should not automatically consider the distances between any two residues to be equal.

If additionally one was presented with supplementary secondary and tertiary structure information, then precise local information (which typically we do not have) would be gained on the residue positions. Then a 3D parametric function could be fitted in order to accurately obtain the inter-residue distances.

Our work is motivated by the intuition that we might improve transmembrane segment prediction if we were somehow able to take into account the resolved 3D information contained in proteins that are similar to our proteins of interest. Making use of this additional information would help estimate the (true) inter-residue distances and improve upon the estimation of the function that 'models' the hydrophobicity level along the protein.

[*] School of Mathematics, University of Bristol, University Walk, BS8 1TW, Bristol

Since the discrete wavelet transform cannot be directly used on irregularly spaced grids, we will *use an adaptive lifting scheme* (see Nunes *et al*. (2004)), which constructs wavelets that adjust to the protein features and are able to work on irregularly spaced observations. Also, we will *construct a measure for the distances between consecutive residues of a protein* by using the information contained in a corresponding set of sequentially aligned proteins with determined 3D structure.

We later show on the proteins from Rost *et al*. (1995) that transmembrane segment prediction improves by incorporating the inter-residue distances. All the proteins in the study are helical, we discuss in the final section the possibility of other structures such as beta-barrels.

We will now briefly introduce the steps we have taken in our analysis, while the next section provides a detailed description of the methodology. The whole approach is relying on analysing the hydropathy profile associated to each protein, and we will thoroughly discuss its construction, which also involves estimating the inter-residue distances. We will base our transmembrane segments prediction on a denoised version of the hydropathy signal. We address the statistical problem of denoising by using wavelet methodology, hence we will briefly introduce basic concepts on wavelets. Since we will use second generation wavelets, we then concentrate on the description of our algorithm, which produces adaptively constructed wavelet functions to decompose the signal at each step. The wavelet coefficients will then be subjected to a thresholding technique, discussed in the denoising section. Once the denoised profile is obtained, we class as transmembranar the segments that are longer than 11 residues (Rost *et al*. (1995)) and correspond to residues with hydrophobicities higher than the smoothed average.

2 Method

2.1 The distance matrix and the hydropathy plot

Various measures for the hydrophobicity of each amino acid have been constructed (for example the scale of Kyte and Doolittle, or the Eisenberg scale), and also combined measures of hydrophobicity and helicity to be used in the context of helical transmembrane proteins (see for instance the Lio and Vannucci scale). By means of these scales, *the primary structure of the protein can be converted into a hydropathy profile*, i.e. we obtain a signal which on the horizontal axis has the residues in their order of appearance in the chain, and on the vertical axis their corresponding values from the hydrophobicity index. After investigating the compatibility of our method with the previously mentioned scales, we decided to use in our study the Kyte and Doolittle measure of hydrophobicity.

In previous studies, the residues were processed assuming that they were equally spaced. As explained in the introduction, we will challenge this assumption and construct a coordinate for each residue in the chain. The coordinate corresponding to each residue will indicate its estimated distance to the previous and next residues.

We now turn to the way we construct the coordinates for each residue. First we determine which protein sequences with resolved 3D structure are similar to the protein we study, through a fast alignment method, FastA, using the scoring matrix BLOSUM62.

Our aim is to use the known 3D structure of the aligned protein sequences in order to estimate the distance between each pair of consecutive residues in the primary structure of the protein of interest. This is done by identifying all the appearances of each specific residue pair in the primary structures of the aligned chains, and then computing all the corresponding Euclidean distances; their average will give us the measure we need. In computing the Euclidean distance between two residues, the x, y, z coordinates (as given by their corresponding PDB file) of all their atoms are used. The result is a 20×20 asymmetrical matrix D, where D_{ij} contains the average of the Euclidean distances computed between the residues i and j, from all the aligned chains where they appear in this order. We should emphasize here that D is not symmetric, hence the distance from Arg to Lys, say, is different to that from Lys to Arg.

At this point one might like to refer to Figure 1, which gives an indication of the range of estimated distances between different pairs of amino acids, as well as their variation. This distance matrix has been computed using 402 matching proteins, each with various sequence lengths.

The distance matrix in Figure 1 is less variable than those obtained from specific protein families. The amino acids are clustered according to their R–group nature, and separated in Figure 1 by white lines. So, GLY–ILE form the first group of amino acids with aliphatic R–groups, SER and THR are non-aromatic amino acids with hydroxyl R–groups, CYS and MET have sulphur containing R–groups, ASP–GLN are

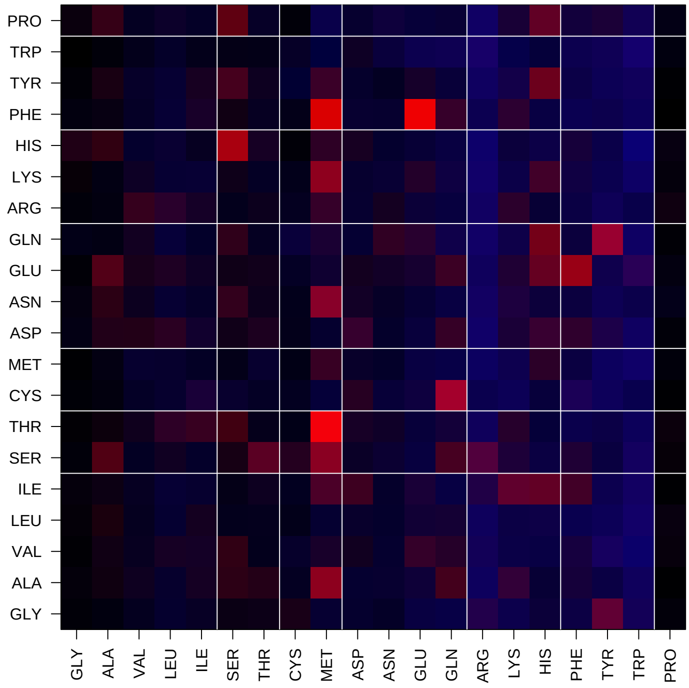

Figure 1: Overall ‘average’ distance matrix (in Ångström units). The intensity of a pixel (dark/light) corresponds to the mean distance for that residue pair. The colour of a pixel (blue/red) corresponds to the standard deviation for that amino acid pair. The brightest pixel in the figure occurs at MET–THR, with a distance of 10.2. It is also one of the most variable, with a standard deviation of 16.7– MET–PHE is the most variable (most red) with a standard deviation of 18.7. The darkest combination is GLY–TRP, with a distance of 4.5, while the least variable is MET–TRP, having a standard deviation of 0.28. The lower and upper quartiles of the mean distance (standard deviation) are 5.2 and 6.1 (0.74 and 2.52).

acidic amino acids and their amides, ARG–HIS are basic amino acids, PHE–TRP are amino acids with aromatic rings and PRO is the sole imino acid. It is notable, for example, that pairs consisting of amino acids with aromatic rings typically have low standard deviations, but middling mean distances.

A shortcoming of this matrix approach is that it only takes into consideration the distances between consecutive residues, and hence models the protein as being a straight chain, rather than modelling its 3D shape. Overcoming this restriction and trying to estimate the 3D function behind the protein shape is an interesting point for future research, and we suspect that it would bring us even closer to correctly estimating the hydrophobicity level as a function of the protein's amino acid composition and shape.

Note that some residue pairs will only appear in the primary structure of the protein being investigated, and not also in the primary structures of the chains aligned to it. In this case, we use the distances supplied by an appropriate distance matrix, computed as follows. In the dataset we are going to investigate (Rost *et al*. (1995)), 15 proteins belong to the tetraspanin family (TM4SF), 22 belong to the ligand-gated ionic channel family (TC 1.A.9), and the rest belong to different families. The last group consists of only 9 proteins, hence we added another 10 proteins randomly selected from the set of 83 cross-validation proteins used in the same study by Rost *et al*. (1995). This way the size of this group was boosted to 19 proteins. The natural division of the dataset into families has led us to construct average distance matrices corresponding to each of the two main families. In the calculation of each matrix, the chains aligned to the sequences belonging to each family have been used. Hence for each of the proteins in one of these families, the missing distances will be imputed from its corresponding overall distance matrix. Along with these two matrices, we have also computed another two family-specific distance matrices, based this time on the structure of the entire proteins (rather than only on the chains) that were aligned to sequences belonging to each family. If the family-specific matrix computed based on the aligned chains contains no information on a particular residue combination, the missing value is taken from the family-specific matrix which uses the entire protein structure. When we analyse a protein that does not belong to one of these two families, we use the distances supplied by an overall ‘average’ distance matrix (see Figure 1), computed from a database comprising 402 proteins with determined 3D structure— the ones aligned to all the proteins investigated.

Having estimated the distances between each pair of consecutive residues in the protein of interest, we compute a coordinate value for each residue in the chain, based on its distance to the previous residue. Using these coordinates, the residues will be plotted on the horizontal axis, hence rather than considering them to be equally spaced, they will have an uneven distribution.

Figure 2 shows an example of hydrophobicity signal, which is very wiggly and a visual assesment is virtually impossible, hence proper statistical tools are needed to denoise it.

2.2 Wavelets and the hydrophobicity profile

Classically constructed wavelets are families of functions based on dilations and translations of a single function, called the mother wavelet. They have the ability of providing representations for square integrable functions, either by continuous linear superpositions of wavelets, or by discrete series expansions of wavelets (Daubechies (1992)). By their construction, classical wavelet decompositions work only on equally spaced grids, with lengths of the form 2^N, and modifications are required in order to overcome these limitations (Cohen *et al*. (1993)).

For this reason, we will construct second generation wavelets, capable of working on irregularly spaced grids of any length. Stemming from the lifting scheme idea introduced by Sweldens (see for instance Sweldens (1997)), we have constructed an adaptive lifting scheme, which we are going to employ in our study.

2.3 An adaptive lifting scheme

The lifting algorithm

We can think of the hydropathy signal as being a function f sampled at n irregularly-spaced points on the real line, x (n is the number of residues in the chain of the protein of interest, x gives their associated coordinates and f is the chosen hydropathy scale). Our aim is to transform the sampled function values by means of lifting into a set of detail and scaling coefficients (representing the high and low 'frequencies', respectively), where each coefficient relates to a certain scale.

Our lifting transform will not follow the classical idea of splitting the sequence into odds and evens, but following Jansen *et al*. (2001, 2004) we concentrate on removing one point at each iteration. Briefly described, the algorithm has the following steps:

- First **choose a point to be lifted**. Say the initial stage is n, when we collected the n sampled points, and the next stage is $n-1$. Denote the point to be removed by j_n.
- **Predict** the function value at j_n by using regression over the cloud of points determined by a neighbourhood (denote it by J_n) of j_n. Generate a detail coefficient $d_{j_n} := c_{n,j_n} - \sum_{i \in J_n} a_i^n c_{n,i}$, where $c_{n,i} := f(x_{n,i})$ and $(a_i^n)_{i \in J_n}$ are the weights obtained through regression.
- **Update** the function values at the neighbours $c_{n-1,i} := c_{n,i} + b_i^n d_{j_n}$, $\forall i \in J_n, i \neq j_n$. The aim of this stage is to preserve constant the average signal, and the weights $(b_i^n)_{i \in J_n}$ will be obtained from this condition.
- After obtaining the detail and updating the values of the remaining points, **remove the point** j_n.
- **Reiterate the lifting transform**: decompose the signal down to preserving only two low frequency coefficients, the rest of them being detail coefficients.

As an observation, this construction induces a parallel construction of scaling and wavelet functions (Jansen *et al*. (2004)).

Adaptivity in the Lifting Algorithm

Since we want a *transform which adjusts itself to suit the signal structure*, we have introduced the option of adaptive lifting prediction steps. In the lifting procedure, there are two sources of adaptiveness we can use— the order of regression and the configuration of neighbours (including their number). This gives rise to two adaptive methods:

- The first method is *adaptive over the order of regression used in the prediction scheme*. The algorithm chooses at each step the type of regression (linear, quadratic or cubic, with or without intercept) which generates the smallest detail in absolute value. The wavelet bases constructed like this adapt themselves to the smoothness of the signal, investigated within a user-specified configuration of neighbours. We will refer to this procedure as **AdaptPred**.

- The second adaptive method *minimises the detail coefficients not only over the regression schemes, but also over the neighbourhood structure*. In other words, several configurations of neighbours are tested with the first adaptive transform, and the one yielding the smallest detail coefficient will be chosen. Hence the wavelet bases constructed through this procedure adapt themselves to the smoothness of the signal within the best predictive window at each step. The name of this procedure is **AdaptNeigh**.

For details regarding the constructions above and their implications, the reader is directed to Nunes *et al.* (2004).

2.4 Denoising the hydrophobicity profile

Wavelets constructed following the above procedure (hence able to work on irregularly spaced data) are going to be employed for detecting the transmembrane segments of helical transmembrane proteins.

Since the transmembrane segments are sequences of predominantly hydrophobic residues, we want to detect the points at which sharp changes occur in the signal. This amounts to modelling the profile as noise-contaminated, and estimating the underlying signal.

Mathematically, we write each of our (independent) observations $(f_{n,i})_{i\in\overline{1,n}}$ as $f_{n,i} = g_{n,i} + \varepsilon_{n,i}$, where $g_{n,i}$ is the population value to be estimated and $\varepsilon_{n,i}$ is an identically distributed, independent noise, assumed to follow a $N(0, \sigma^2)$ distribution. In other words, based on just one observation, $f_{n,i}$, at each sampled point i of the grid, we want to estimate the true value of the signal at i, $g_{n,i}$. The assumption of independent observations is a necessary mathematical requirement, which we are aware that is likely to hold only approximately for our transmembrane prediction.

In practice, most of the time the true signal is not sparse, but transformed through a discrete wavelet transform (DWT) or through a lifting algorithm, the resulting sequence of wavelet coefficients has the property of being sparse. Hence the observed signal will first be decomposed into coarse scale coefficients and wavelet coefficients (details). Intuitively, the coarse scale coefficients are capturing the ‘big’ features of the signal, while the noise mostly contaminates the details. Some of these details of course, are going to represent true features of the signal (and are also contaminated by noise), while others will be due only to the noise.

When the noise corrupted signal f is transformed through the lifting algorithm into a set of scaling and wavelet coefficients, it means that the above model will be converted into $d_j = d_j^* + e_j$, with $(d_j)_j$ being the observed wavelet coefficients, $(d_j^*)_j$ the ‘true’ wavelet coefficients and e_j the transform of the noise ε_j. We only note here that the lifting transform is not an orthogonal transform (while the DWT is), and hence care must be taken in assessing the distributional properties of the true and observed wavelet coefficients. For an in depth discussion refer to Nunes *et al.* (2004).

In order to establish which of the observed wavelet coefficients represent true non-zero population wavelet coefficients, a threshold needs to be estimated for each detail. In our approach we will use an adapted version of the empirical Bayes procedure (for details see Johnstone and Silverman (2005), Nunes *et al.* (2004)). Briefly, the empirical Bayes approach relies on the property of the ‘true’ wavelet coefficients of having a sparse structure which allows us to place independent prior distributions describing each of them as being zero with a probability π (to be estimated) or to have come from a quasi-Cauchy distribution, with probability $1 - \pi$. In shrinking the observed wavelet coefficients, we will use the posterior means of the developed posterior distributions of the ‘true’ details.

Once the wavelet coefficients have been thresholded, the transform is inverted, yielding a estimated version of the initial signal.

2.5 Predicting the Transmembrane Segments

The estimated hydropathy profile will be used to predict the transmembranar segments. All the residues corresponding to smoothed hydrophobicities larger than the estimated average will be considered to be transmembranar, provided that they form segments which are longer than 11 residues.

3 Implementation

We tested our method using 46 of the 48 proteins from Rost *et al.* (1995) (the double–blind set, available from http://www.embl-heidelberg.de/∼rost/Papers/1996_phdtop/Blind.html). The search on the AD1 antigen retrieves the entry 'cd63-rat', which subsequently appears in the database, and the glutamate receptor A precursor contains a much longer sequence than the rest, causing memory difficulties. As mentioned in section 2.1, we added a set of 10 proteins to the 46 proteins dataset, in order to boost the the number of proteins that do not belong to either of the two families. As a consequence, we report the overall results obtained on all 56 proteins. We compared our results against those obtained by using the least asymmetric Daubechies wavelets with 8 vanishing moments (usually denoted Daubechies 's8') for decomposing the signal down to 4 levels, combined with the empirical Bayes procedure for shrinking the wavelet coefficients, using the posterior means (*Daub mean*). We based our wavelet choice on a comparative study between several Daubechies wavelets, with different vanishing moments. The same choice has been previously reported in the literature (Lio and Vannucci (2000)). As a remark, since Daubechies 's8' wavelets were used, the estimated distances between residues have been ignored, and considered to be equal.

In Nunes *et al.* (2004), AdaptPred and AdaptNeigh were tested in an extensive simulation study and they proved to be very powerful in the task of shrinkage. For denoising smooth signals or signals with a small number of discontinuities, AdaptPred with 2 neighbours (*AP2*) performs best, while for denoising non-smooth signals, AdaptNeigh using up to two neighbours at each stage (*AN1*) gives the best results. Hence when denoising our hydrophobicity data, we have focused on these two methods.

Note that in the decomposition using adaptive wavelets, we kept the same number of scaling coefficients as in the decomposition using Daubechies 's8'.

4 Results

4.1 Prediction accuracy measurements

Both methods produce their corresponding predicted transmembranar segments which we have to compare against the experimental data and assess which is the better prediction. We believe that there is no obvious measure that would give a concise answer as to which of the predictions is better, and hence we used several measures for the accuracy of prediction:

- **Measures referring to the residue accuracy** (see for example Rost and Sander (1993)): the percentage of residues predicted correctly in either of the two states (transmembranar or not), Q_2; the percentage of residues which are correctly predicted to be transmembranar, relative to the number of residues observed to be transmembranar (Q_{obs}) and relative to the number of residues predicted to be transmembranar (Q_{pred}).

- **Measures referring to the segment accuracy** (see for example Rost *et al.* (1996)): the number of correctly predicted transmembrane segments, N_{corr}, where a segment is considered to be correctly predicted if there is an overlap of at least 5 residues with a true one; sensitivity, i.e. the percentage of observed transmembrane segments that were correctly predicted, *Sens*; specificity, i.e. the percentage of predicted transmembrane segments that are correct, *Spec*; segment overlap Sov_{obs}, Sov_{pred}, which are more sophisticated measures for evaluating (on a scale from 0% to 100%) respectively the correctness of segment prediction versus the true segments and the fraction of the predicted segments that is correct (for more details see Zemla *et al.* (1999)).

4.2 Discussion of results

Using the above measures, we evaluated the performance of our method versus the performance of the method employing Daubechies 's8' on equally spaced grids.

After investigating the AdaptPred with two closest neighbours and AdaptNeigh with at most two neighbours methods, both with posterior median and with posterior mean thresholding, we concluded that AdaptNeigh method using posterior mean shrinkage (*AN1 mean*), gives the best results throughout the study, hence this is the method we recommend, followed by AdaptNeigh with posterior median thresholding (*AN1 median*). Occasionally (even though very rare), it happens for the AdaptNeigh technique to produce predicted segments that are too short (average length under 14 residues) or too long (average length over 34 residues), situation when AdaptPred using two closest neighbours and posterior mean shrinkage (*AP2 mean*) would be chosen.

We found out that on the proteins belonging to the tetraspanin TM4SF family, the classical method mostly gives very good prediction, with only a few exceptions. On the leukocyte antigen CD37 (UniProt entry 'cd37-human') Daubechies 's8' fails almost completely to recognize the true segments, giving Sov_{obs} and Sov_{pred} values of 0.5 and 0.33 respectively.

As said before, we tested AdaptPred with 2 closest neighbours and AdaptNeigh using at most 2 neighbours, both using posterior median thresholding and posterior mean shrinkage. The results show that AdaptNeigh with either type of shrinkage and AdaptPred with posterior mean shrinkage give the best predictions. Overall, our segment prediction accuracy is very similar to the one obtained through the classical method, as showed by the results in Table 1. We obtain higher sensitivity (i.e. the percentage of observed transmembranar segments that were correctly predicted), and very similar specificity, as well as very similar *Sov* values. These values indicate an accurate segment prediction, judged not only by the simple criterion of considering a segment correctly predicted if there is an overlap of at least 5 residues with a true one, but also by the better measure provided by *Sov*, which takes into account the change points as well. The per-residue measures indicate a better behaviour for the classical method, but we should keep in mind that this measure should be considered with care, since we are primarily interested in sequences of residues and their positions within the chain.

At a close examination of the results based on which we obtained Table 1, we notice that our method provides more homogenous estimations, and there is no failure of prediction for any of the proteins, unlike the classical method.

On the ligand-gated ionic channel (TC 1.A.9) family, the classical based wavelet methods give quite poor predictions most of the time, with *Sov* values ranging from (0.51,0.3) to at most (1,0.51). For four proteins, values around (0.5,0.3) are obtained, hence the classical method fails to make a good prediction for them. Most of the *Sov* values are concentrated around (0.8,0.45), indicating that there is a tendency of overpredicting segments (predicting segments that are not truly transmembranar), and also of not being able to correctly detect the boundaries of the true segments. This generally translates in predicting a segment as being the merging of 2 or, in a few cases, even 3 true segments.

Our methods give an improved prediction for most of the proteins. AdaptNeigh method gives better predictions than AdaptPred, and this time AdaptNeigh using posterior mean shrinkage is superior to the same method, but employing posterior median thresholding. Most of the *Sov* values for AdaptNeigh using posterior mean shrinkage are within the range of (0.8-1,0.5-0.8), considerably higher than the results obtained using the classical wavelets.

Also for this family, the prediction performance given by our method is more homogenous than in the classical case. By examining Table 2, we notice that while improving the sensitivity (the boundaries of the true segments are correctly identified, and segments are seldom merged), we do not seem to be able to significantly improve upon the specificity of our prediction (some segments are falsely predicted as transmembrane).

We now examine a protein belonging to this family: we chose the gamma-aminobutyric-acid receptor gamma-3 subunit precursor (UniProt entry 'gac3-mouse'), which displays the typical behaviour of both methods. It has a chain of length 467 residues, to which chains coming from eight proteins with determined 3D structure have been aligned. The inter-residue distances were computed based on these proteins, and the values of the missing pairs were imputed from the overall distance matrix corresponding to this family.

The experimentally determined transmembrane segments are believed to be:
255-277, 281-303, 315-337, 444-467.

Our method predicts the following segments:
5-15, 77-88, 116-133, 232-249, 253-277, 288-303, 317-332, 447-467, while by the usage of Daubechies wavelets, we obtain
1-15, 72-92, 118-134, 157-173, 230-296, 306-337, 443-467.

The hydropathy profile obtained is given in Figure 2:

The observed and predicted segments correspond to Figure 3:

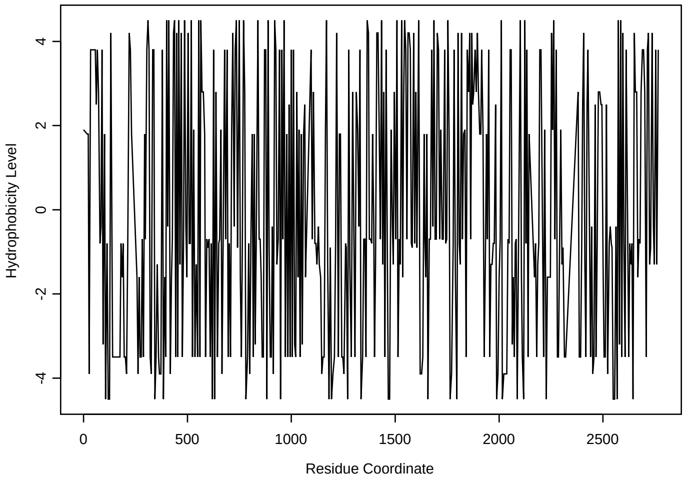

Figure 2: Hydropathy Profile of ‘gac3-mouse’

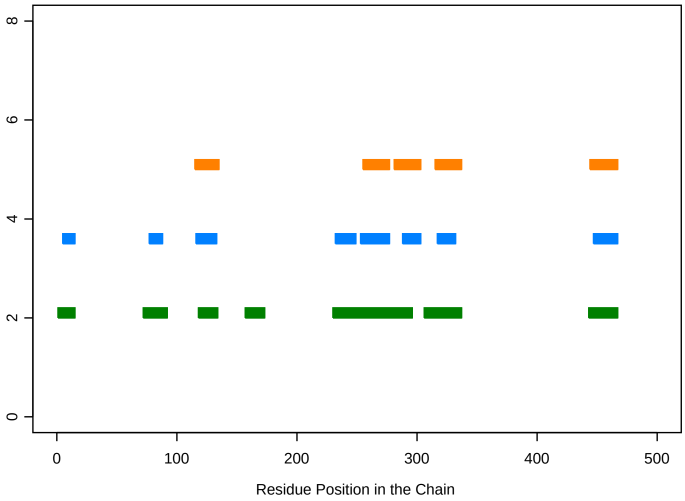

Figure 3: Predicted segments for ‘gac3-mouse’: red=True, blue=AdaptNeigh1, green=Daub ‘s8’

Figure 4 shows the corresponding coarse versions of the hydropathy profile of ‘gac3-mouse’:

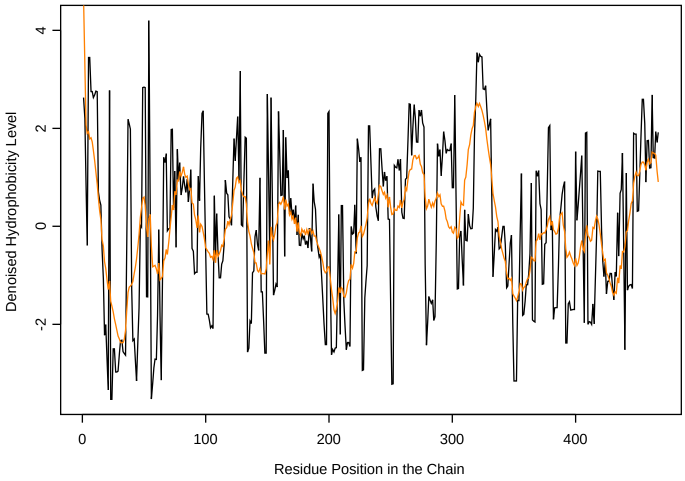

Figure 4: Centred denoised hydropathy profile of ‘gac3-mouse’ using AdaptNeigh1, Daub ‘s8’, in black, red respectively

When measuring the performance of these segments, we obtain the following results:

our method:
$Q_2 = 0.83$, $Q_{obs} = 0.82$, $Q_{pred} = 0.55$, $< L >_{obs} = 23.25$, $< L >_{pred} = 17.12$, $N_{obs} = 4$, $N_{pred} = 8$, $N_{corr} = 4$, $Sens = 1$, $Spec = 0.5$, $Sov_{obs} = 1$, $Sov_{pred} = 0.57$
Daubechies 's8':
$Q_2 = 0.75$, $Q_{obs} = 0.92$, $Q_{pred} = 0.44$, $< L >_{obs} = 23.25$, $< L >_{pred} = 27.71$, $N_{obs} = 4$, $N_{pred} = 7$, $N_{corr} = 3$, $Sens = 0.75$, $Spec = 0.43$, $Sov_{obs} = 0.72$, $Sov_{pred} = 0.44$

We notice in this example the behaviour described earlier, in that both methods overpredict the transmembrane segments, and the classical wavelets are also merging some of the true segments.

All of the methods mentioned in the beginning of this paper (see for example Rost *et al*. (1995, 1996) or Fisher *et al*. (2003)), besides the mathematical filtering, employ a bio-chemical filtering as well, which we keep minimal (we only cut the helices containing at most 10 residues). Such further filtering and inspection of the already predicted segments will considerably improve the prediction specificity and sensitivity (and will also improve Sov_{pred}, Sov_{obs}), by eliminating some of the unlikely segments, or splitting the segments considered to be too large into two or more segments. In our study, a closer examination of the obtained predicted segments in the ligand-gated ionic channel (TC 1.A.9) family shows that a lot of the segments wrongly predicted as transmembranar are very short (11-15 residues), and hence unlikely to 'survive' a bio-chemical filtering procedure. It may also be that some of them are too long, and splitting them into more segments might be a solution. In our approach, we kept exclusively a mathematical filtering procedure and investigated its behaviour with and without the information given by multiple aligned sequences with known 3D structure. Having improved upon the basic mathematical prediction, various other procedures (such as the bio-chemical filtering discussed above) could then be joined, and contribute to an improved final prediction.

Finally, for the rest of proteins, the ones belonging to different families, the predictions of both our method and of the one employing classical wavelets are quite good, with the exception of three proteins which have only one (true) transmembranar segment. For these proteins, our methods and the classical one have very similar performances, in that the *Sov* values are around (0.9-1, 0.2-0.6), indicating that the methods correctly identify the true segment, but additionally predict false ones. For the remaining 16 proteins, none of the methods fails and the range of *Sov* values is (0.64-1,0.67-1) for AdaptNeigh with posterior mean shrinkage method, and (0.67-1, 0.51-1) for the classical method. Comparing the results obtained on the whole set of 19 proteins, we see that our method either outperforms the results obtained through the Daubechies wavelets or gives similar results, and in only three cases we obtain worse results than by using the Daubechies 's8'. For this group of proteins, the best results are obtained by predicting through AdaptNeigh with at most 2 neighbours using posterior mean shrinkage, too. Examining Table 3 we notice that we obtain improved specificity values for the AdaptNeigh technique using posterior mean shrinkage as compared to the results obtained through the usage of classical wavelets, and a similar sensitivity value. This is reflected also by examining the *Sov* values.

To conclude, examine Table 4, which combines all the previous data to show the overall tendency. We compared the *Sov* values (since these are the most complete measures for the segment prediction accuracy) obtained through our methods versus the ones obtained by using classical wavelets. For performing the comparisons we used paired t-tests, since the sample size is large enough so that the tests should be robust against non-normality. For each of our methods, and both for Sov_{obs} and Sov_{pred}, we tested the null hypothesis of no difference between the mean *Sov* value of our method and the mean *Sov* value of the classical method, versus the alternative that our method provides a higher *Sov* value than the one obtained through the classical wavelets. We indicated the highly significant differences in Table 4. Based on a careful examination of the data and on the results of the significance tests, we conclude that we improve the quality of prediction by using resolved 3D structure of proteins that are similar to the proteins to be analysed— both in terms of the correctness of the segments with respect to the true segments and the proportion of predicted segments that are correct.

Wavelet methods using a second filtering step based on the chemical properties of the residues, report final sensitivity and specificity values of 0.93 and over. With no such further filtering, we obtain a value of 0.90, indicating that if we additionally use such a procedure, we should obtain even higher sensitivity values. We refer to the sensitivity and specificity values since they are the measures usually reported in the literature, but we stress again that a much better measure, indicating more accurately the behaviour of the method, is *Sov*. Its observed value also confirms an improvement of the prediction accuracy. Due to the ligand-gated ionic channel family (TC 1.A.9), the specificity value drops to 0.70, a higher value than the one corresponding to the classical method — 0.62, but yet a smaller value than the ones reported by the previous studies (Lio and Vannucci (2000), Fisher *et al*. (2003)), in which further to the mathematical filtering, a step of biochemical filtering is employed. The value of Sov_{pred} (0.76) is higher than the one provided by the specificity index, and it also points towards the existence of an improvement with respect to the classical method (which has Sov_{pred} of 0.67).

For the initial dataset consisting of 46 proteins, we have also tested our methodology using two different types of matrices for imputing the missing values when computing the coordinate of each residue. We remind the reader that so far we primarily used two matrices, one for each of the families (see section 2.1). In the calculation of these matrices we used the chains with determined 3D structure that were aligned to the sequences belonging to each family, respectively. Solely for estimating the missing values in these matrices, we used another two matrices computed based on the structure of the entire proteins aligned to sequences belonging to each family. Now, we have also tested our methods using imputed values straight from the distances provided by these matrices. And finally, regardless the family to which the protein belongs, we have used the overall matrix computed based on 376 proteins, all the proteins aligned to the 46 proteins being analysed. Our intuition was that the prediction should slightly decrease in accuracy by using less specific information. The tests proved that the specificity has slightly decreased, from the overall 0.76 to 0.74 (this difference being mainly due to the specificity decrease in the ligand-gated ionic channel family, from 0.59 to 0.55), while the sensitivity was not influenced.

As a note, if the sequence of interest has no aligned sequences with resolved 3D structure, then the corresponding overall matrix can be used for computing all the inter-residue distances.

5 Conclusions and further work

This article has developed a new multiscale technique for transmembrane protein segment prediction. The new technique improves on earlier wavelet methods by utilising resolved 3D structure information from similar proteins to provide irregularly spaced residues. The irregular spacing is generated by order-20 distance matrices which calculate inter-residue distances over families of similar proteins (and also a generic ‘all-protein’ matrix for use when a family matrix cannot produce a distance for a particular combination). This construction aims at obtaining a better estimate of the true function that models the level of hydrophobicity along the protein. We tested our method on helical transmembrane proteins, and consequently we generated distance matrices that reflect the helicity property. An interesting direction would be to further extend the study to beta-barrel transmembrane proteins. For the future, the ‘paradigm’ provides a way of generalising multiscale algorithms for irregularly spaced objects (such as proteins) and hence lifting shows great promise for directly utilising 3D resolved information in a mathematical multiscale manner which is informed by the biochemical reality.

6 Acknowledgements

Marina Knight gratefully acknowledges the support of an ORS award and Bristol University Research Scholarship. Guy Nason was partially supported by EPSRC Advanced Research Fellowship AF/001664.

References

Daubechies, I. (1992). *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA.

Cohen, A., Daubechies, I., Vial, P. (1993). Wavelets on the Interval and Fast Wavelet Transforms. *Appl. Comput. Harmon. Anal.*, **1**, 54–81.

Engelman, D.M., Steitz, T.A. and Goldman, A. (1986). Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins. *Annu. Rev. Biophys. Biophysical Chem.*, **15**, 321–353.

Fisher, P., Baudoux, G. and Wouters, J. (2003). WAVPRED: a Wavelet-Based Algorithm for the Prediction of Transmembrane Proteins. *Comm. Math. Sci.* **1**, 44–56.

Hirakawa, H., Muta, S. and Kuhara, S. (1999). The Hydrophobic Cores of Proteins Predicted by Wavelet Analysis. *Bioinformatics*, **15**, 141–148.

Jansen, M., Nason, G.P. and Silverman, B.W. (2001). Scattered data smoothing by empirical Bayesian shrinkage of second generation wavelet coefficients. In Unser, M. and Aldroubi, A. (eds)*Wavelet Applications in Signal and Image processing*, Proceedings of SPIE, **4478**, 87–97.

Jansen, M., Nason, G.P. and Silverman, B.W. (2004). Multivariate Nonparametric Regression using Lifting. *Technical Report*, **04:17**, Department of Mathematics, University of Bristol, (submitted for publication).

Johnstone, I.M. and Silverman, B.W. (2005). Empirical Bayes Selection of Wavelet Thresholds. *Ann. Statist.*, **33**, (to appear).

Johnstone, I.M. and Silverman, B.W. (2002). EbayesThresh: R and S-PLUS software for Empirical Bayes Thresholding. Unpublished manuscript (available from the CRAN archive).

Kyte, J. and Doolittle, R.F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.*, **157**, 105–132.

Lio, P. and Vannucci, M. (2000). Wavelet Change-Point Prediction of Transmembrane Proteins. *Bioinformatics*, **16**, 376–382.

Nunes, M.A., Knight, M.I. and Nason, G.P. (2004). Adaptive Lifting for Nonparametric Regression. *Technical Report*, **04:20**, Department of Mathematics, University of Bristol, (submitted for publication).

Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995). Transmembrane Helices Predicted at 95% Accuracy. *Protein Science* **4**, 521–533.

Rost, B., Fariselli, P. and Casadio, R. (1996). Topology Prediction for Helical Transmembrane Proteins at 86% Accuracy. *Protein Science* **5**, 1704–1718.

Rost, B. and Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.*, **232**, 584–599.

Sweldens, W. (1997). The Lifting Scheme: a Construction of Second Generation Wavelets. *SIAM J. Math. Anal.*, **29**, 511–546.

Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999). A Modified Definition of *Sov*, a Segment Based Measure for Protein Secondary Structure Assesment. *PROTEINS: Structure, Function and Genetics* **34**, 220–223.

	N_{pred}	N_{corr}	*Sens*	*Spec*	$< L >_{obs}$	$< L >_{pred}$	Sov_{obs}	Sov_{pred}	Q_2	Q_{obs}	Q_{pred}
AP2 mean	62	60	1.00	0.97	22.08	25.45	0.96	0.95	0.88	0.94	0.78
AN1 median	63	60	1.00	0.95	22.08	22.05	0.95	0.94	0.88	0.87	0.81
AN1 mean	62	60	1.00	0.97	22.08	20.36	0.93	0.93	0.88	0.82	0.84
Daub mean	58	57	0.95	0.98	22.08	28.90	0.93	0.92	0.90	0.96	0.80

Table 1: Results obtained on the TM4SF family (15 proteins, 60 experimentally determined transmembrane segments). N_{pred}, N_{corr} give the number of predicted, respectively correctly predicted transmembrane segments; *Sens*, *Spec* give the sensitivity, specificity of prediction; $< L >_{obs}$, $< L >_{pred}$ are the average length of the observed, predicted segments; Sov_{obs}, Sov_{pred} evaluate the correctness of prediction versus the true segments, and the fraction of the predicted segments that is correct; Q_2 is the percentage of correctly predicted residues, Q_{obs}, Q_{pred} measure the percentage of correctly predicted residues relative to the number of observed, respectively predicted transmembranar residues

	N_{pred}	N_{corr}	*Sens*	*Spec*	$< L >_{obs}$	$< L >_{pred}$	Sov_{obs}	Sov_{pred}	Q_2	Q_{obs}	Q_{pred}
AP2 mean	168	72	0.82	0.43	22.34	25.21	0.85	0.49	0.77	0.95	0.45
AN1 median	173	73	0.83	0.42	22.34	21.71	0.84	0.50	0.76	0.84	0.44
AN1 mean	148	73	0.83	0.49	22.34	19.77	0.84	0.59	0.82	0.79	0.52
Daub mean	165	63	0.72	0.38	22.34	27.07	0.75	0.44	0.75	0.96	0.43

Table 2: Results obtained on the TC 1.A.9 family (22 proteins, 88 experimentally determined transmembrane segments). N_{pred}, N_{corr} give the number of predicted, respectively correctly predicted transmembrane segments; *Sens*, *Spec* give the sensitivity, specificity of prediction; $< L >_{obs}$, $< L >_{pred}$ are the average length of the observed, predicted segments; Sov_{obs}, Sov_{pred} evaluate the correctness of prediction versus the true segments, and the fraction of the predicted segments that is correct; Q_2 is the percentage of correctly predicted residues, Q_{obs}, Q_{pred} measure the percentage of correctly predicted residues relative to the number of observed, respectively predicted transmembranar residues

	N_{pred}	N_{corr}	*Sens*	*Spec*	$< L >_{obs}$	$< L >_{pred}$	Sov_{obs}	Sov_{pred}	Q_2	Q_{obs}	Q_{pred}
AP2 mean	98	82	0.90	0.84	23.89	23.59	0.90	0.77	0.82	0.79	0.75
AN1 median	106	83	0.91	0.78	23.89	19.13	0.91	0.78	0.80	0.71	0.76
AN1 mean	95	82	0.90	0.86	23.89	17.99	0.89	0.82	0.83	0.66	0.84
Daub mean	98	80	0.88	0.82	23.89	24.02	0.92	0.75	0.81	0.80	0.73

Table 3: Results obtained on the rest of the proteins (19 proteins, 91 experimentally determined transmembrane segments). N_{pred}, N_{corr} give the number of predicted, respectively correctly predicted transmembrane segments; *Sens*, *Spec* give the sensitivity, specificity of prediction; $< L >_{obs}$, $< L >_{pred}$ are the average length of the observed, predicted segments; Sov_{obs}, Sov_{pred} evaluate the correctness of prediction versus the true segments, and the fraction of the predicted segments that is correct; Q_2 is the percentage of correctly predicted residues, Q_{obs}, Q_{pred} measure the percentage of correctly predicted residues relative to the number of observed, respectively predicted transmembranar residues

	N_{pred}	N_{corr}	*Sens*	*Spec*	$< L >_{obs}$	$< L >_{pred}$	Sov_{obs}	Sov_{pred}	Q_2	Q_{obs}	Q_{pred}
AP2 mean	328	214	0.90	0.65	22.80	24.73	0.90^{β}	0.71^{β}	0.80	0.88	0.60
AN1 median	342	216	0.90	0.63	22.80	20.93	0.89^{γ}	0.71^{β}	0.80	0.79	0.60
AN1 mean	305	215	0.90	0.70	22.80	19.32	0.88	0.76^{α}	0.83	0.75	0.68
Daub mean	321	200	0.84	0.62	22.80	26.52	0.86	0.67	0.80	0.89	0.59

Table 4: Overall results (56 proteins, 239 experimentally determined transmembrane segments). N_{pred}, N_{corr} give the number of predicted, respectively correctly predicted transmembrane segments; *Sens, Spec* give the sensitivity, specificity of prediction; $< L >_{obs}$, $< L >_{pred}$ are the average length of the observed, predicted segments; Sov_{obs}, Sov_{pred} evaluate the correctness of prediction versus the true segments, and the fraction of the predicted segments that is correct; Q_2 is the percentage of correctly predicted residues, Q_{obs}, Q_{pred} measure the percentage of correctly predicted residues relative to the number of observed, respectively predicted transmembranar residues; α indicates a significantly higher *Sov* value for the corresponding method than for *Daub mean* at 99% confidence level, while β corresponds to a significantly higher result for our (corresponding) method at 95% confidence level and γ indicates a significantly higher result for our (corresponding) method at 90% confidence level