Protein Engineering vol.15 no.9 pp.745–752, 2002

On filtering false positive transmembrane protein predictions

Miklos Cserzö[1,2], Frank Eisenhaber[3], Birgit Eisenhaber[3] and Istvan Simon[4]

[1]University of Birmingham, School of Biosciences, Edgbaston, Birmingham B15 2TT, UK, [3]IMP Bioinformatics, Dr Bohr-Gasse 7, A-1030 Vienna, Austria and [4]Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, P.O. Box 7, H-1518 Budapest, Hungary

[2]To whom correspondence should be addressed.
E-mail: miklos@bip.bham.ac.uk

While helical transmembrane (TM) region prediction tools achieve high (>90%) success rates for real integral membrane proteins, they produce a considerable number of false positive hits in sequences of known non-transmembrane queries. We propose a modification of the dense alignment surface (DAS) method that achieves a substantial decrease in the false positive error rate. Essentially, a sequence that includes possible transmembrane regions is compared in a second step with TM segments in a sequence library of documented transmembrane proteins. If the performance of the query sequence against the library of documented TM segment-containing sequences in this test is lower than an empirical threshold, it is classified as a non-transmembrane protein. The probability of false positive prediction for trusted TM region hits is expressed in terms of *E*-values. The modified DAS method, the DAS-TMfilter algorithm, has an unchanged high sensitivity for TM segments (~95% detected in a learning set of 128 documented transmembrane proteins). At the same time, the selectivity measured over a non-redundant set of 526 soluble proteins with known 3D structure is ~99%, mainly because a large number of falsely predicted single membrane-pass proteins are eliminated by the DAS-TMfilter algorithm.

Keywords: automated sequence database screening/DAS-TMfilter/genome sequence annotation/transmembrane region prediction

Introduction

As the gap between the number of available sequences and the existing experimental data on characterized biomolecules continues to grow, the intellectual challenge of structure and function prediction from biomacromolecular sequences becomes of increasing practical importance. Transmembrane proteins are targets of primary interest since many of them are surface receptors or enzymes that are easily accessible to pharmaceutical interference. At the same time, their membrane-bound nature causes considerable difficulties for experimental structure determination, so limiting access to important data about their functions.

In most known cases, membrane-spanning parts of transmembrane proteins consist of helices perpendicular to the membrane plane. However, helices can also be tilted, as in light-harvesting complex (Kuhlbrandt *et al.*, 1994). Helices have even been found that lie parallel to the membrane plane (Picot *et al.*, 1994). Another group of membrane proteins, the porins (Weiss and Schulz, 1992), are constructed of 16 β-sheets arranged as a barrel, giving rise to a large central hole. There are indications that the membrane-spanning segments can consist of single β-strands, which can span the membrane with fewer residues than an α-helix (Hucho *et al.*, 1994).

In this paper, we concentrate on integral membrane proteins with transmembrane helices, which will be referred to as TM proteins; all other proteins will be termed non-TM proteins. Note that porins, which lack TM helices are, thus, considered as non-TM proteins in this paper: the DAS-TMfilter prediction method described in this paper generally cannot detect the β-strand-like transmembrane protein segments in porins. The problem of detecting non-helical TM segments is outside the scope of this paper.

Helical transmembrane regions are generally characterized simply as continuous stretches of, mainly, hydrophobic residues and were, therefore, early targets for bioinformatics approaches (Engelman *et al.*, 1986; Eisenhaber *et al.*, 1995). A variety of techniques have been applied to locate TM segments and new methods continue to be published. Available methods include (i) sliding window averaging with amino acid hydrophobicity scales (Kyte and Doolittle, 1982; Engelman *et al.*, 1986; Hirokawa *et al.*, 1998; Juretic *et al.*, 1998; Pasquier *et al.*, 1999; Jayasinghe *et al.*, 2001), (ii) amino acid residue distribution criteria (Jones *et al.*, 1994; Persson and Argos, 1996; McGuffin *et al.*, 2000), (iii) sequence profile analysis (von Heijne, 1992; Cserzö *et al.*, 1997), (iv) neural network analysis (Rost *et al.*, 1995), (v) hidden Markov models (Sonnhammer *et al.*, 1998; Tusnady and Simon, 1998, 2001a; Pasquier and Hamodrakas, 1999; Krogh *et al.*, 2001), (vi) molecular mechanics modeling (Nikiforovich, 1998) and (vii) combinations of these methods (Nilsson *et al.*, 2000; Tompa *et al.*, 2001). It should be noted that none of these methods are trained for the prediction of non-helical TM regions.

A few of the most advanced prediction tools perform with a success rate close to 95% for known transmembrane sequences (Moller *et al.*, 2001; Simon *et al.*, 2001). They are effective in locating transmembrane segments in real TM proteins, but they tend also incorrectly to identify other hydrophobic clusters in globular proteins as helical transmembrane segments. As a result, as many as 20–40% of non-TM query sequences may give false positive hits in such prediction processes (Jayasinghe *et al.*, 2001; Tompa *et al.*, 2001). Strictly, feeding non-TM queries into these tools is inappropriate, as the methods are neither designed nor optimized for this role. However, the mass production of genomic sequence data continues to put great pressure on the bioinformatics community to supply a reliable TM annotation tool.

The issue of reducing the false positive error rates of existing prediction tools urgently needs to be addressed. In this paper, we propose DAS-TMfilter, a modification of the DAS method

('dense alignment surface' algorithm) (Cserzö *et al.*, 1994, 1997), that achieves a substantial decrease in the false positive error rate. In this procedure, a sequence with initially predicted transmembrane region(s) is re-tested in a second step that compares it with transmembrane segments in a sequence library of documented transmembrane proteins. If the performance of the query sequence in this second test is below an empirically determined threshold, the query is finally classified as non-transmembrane sequence. Further, we evaluate the probability of false positive prediction for trusted TM region hits in terms of ***E***-values. At the same time, the modified method does not fall below the ~95% threshold in recognizing genuine TM regions. That rate is typical for advanced TM segment recognition techniques. To simplify discussion, we omit comparisons with the many previously published methods and compare the fidelity of the new method with the results of recent comparative surveys (Moller *et al.*, 2001; Simon *et al.*, 2001).

The paper is organized as follows. First, we describe the approach in general terms. The Methods section enunciates the exact mathematical formulation of the algorithm. Then we describe results of validation tests and interpretation of the output results.

Theoretical considerations

The issue of the TM learning set

Most of the known TM prediction techniques are essentially knowledge-based, so the quality of a 'gold standard' learning and test set is critical for their parametrization and performance evaluation. A close look at the available 'experimentally determined' database of 128 protein examples provided by Moller *et al.* (Moller *et al.*, 2000) shows that it is not as 'golden' as one would expect for a rigorous standard. Less than 10% of the commonly used reference database entries are results of atomic resolution X-ray diffraction. The remainders are derived from indirect measurements: a very limited number of residues are identified as being inside or outside the cell or in the cell membrane by chemical methods. Typically, a standard TM region length restriction is imposed onto the sequence and the most hydrophobic stretch between the labeled key residues on each side of the membrane is declared as transmembrane helix. As a rule, a length between 20 and 25 residues is assumed but the database includes reported examples of TM segments longer than 40 residues. There are more than 100 different hydrophobicity scales in the literature (Nakai *et al.*, 1988; Palliser and Parry, 2001). Depending on which scale the authors prefer, the same experimental data can result in different transmembrane segment position assignments. Even in the case of X-ray structure determination, the termini of the TM regions are not unambiguously determined: The termini of the helices are defined by the hydrogen bond between residues i and $i + 3$ but X-ray structural data show that many helices extend into the aqueous phase to some degree (Tusnady and Simon, 2001b).

As a result, the available experimental data on TM proteins can be used to compile a good collection of hydrophobic segments representing cores of transmembrane regions, but the stated margins of the TM regions are not very reliable. For example, we found some preferences for amino acid residues with flexible backbone and small side chains at the margins of TM regions (necessity to form loop structures) but the noise in the database did not allow us to assign statistical significance to this finding.

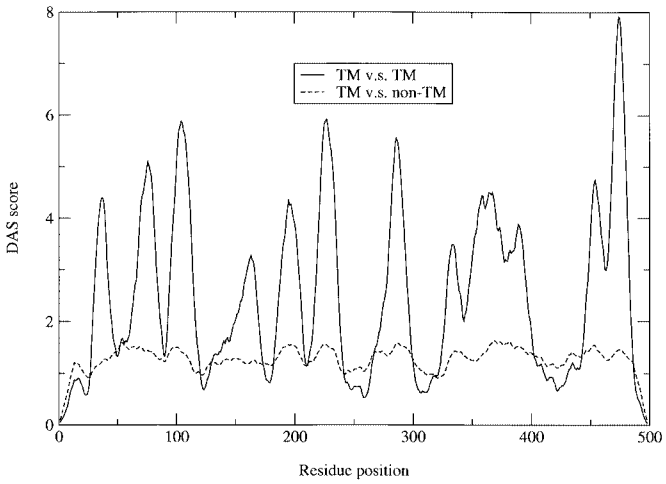

Fig. 1. DAS profiles of a TM protein as function of residue number. The library DAS profile Λ(TLCA_RICPR, ***Q***) for the SWISS-PROT sequence with the accession number P19568, a member of the TM library, has been averaged (i) over all library sequences as query ***Q*** (full line) and (ii) over all sequences of the non-TM set as query ***Q*** (dashed line). This example shows that non-TM sequences have a clear tendency to produce low library profiles whereas the reported TM regions can be recognized as peaks of the profile computed with true TM region proteins.

The unreliability of experimentally reported margins of TM regions also sets limits on the comparison of predicted and documented TM segments. In our accuracy tests, we consider a reported region as predicted if there is any overlap with the predicted segment. In the following, we use one set of documented transmembrane proteins (called 'TM set' or 'TM library') as positive examples and contrast it with another set of known non-transmembrane proteins (a non-redundant set of soluble proteins with known 3D structure called 'non-TM set').

Modification of the DAS method: the DAS-TMfilter algorithm

Generally, other properties in addition to window-averaged hydrophobicity are required to distinguish between TM regions and hydrophobic stretches in globular proteins, but the current status of the learning database suggests that it appears unlikely that such properties might be formulated as explicit condition as in the case of hydrophobicity. Moreover, not all transmembrane helices are equally hydrophobic, for example those surrounded by other TM regions. Thus, hydrophobicity thresholds derived as averages over learning sets might be too low to recognize single hydrophobic helices in non-TM proteins as false positives.

At this point, we thought that a prediction technique such as the 'dense alignment surface' (DAS) method (Cserzö *et al.*, 1994, 1997), which relies on direct comparisons of a query sequence with learning set sequences at all stages of the prediction process, might have the potential to define implicitly the additional conditions. Originally, DAS was a low-stringency dot-plot method for comparing a query sequence against a collection of library sequences consisting of non-homologous membrane proteins. TM regions in the query can be recognized by characteristic black/white patterns in the dot plot (see Figure 1 in Cserzö *et al.*, 1994). If a special scoring matrix RReM (previously derived from neighbor relationships of residues and found to assign high scores to exchanges that maintain residue polarity) is applied, the resulting hydrophobicity profiles for the query sequence predict the location

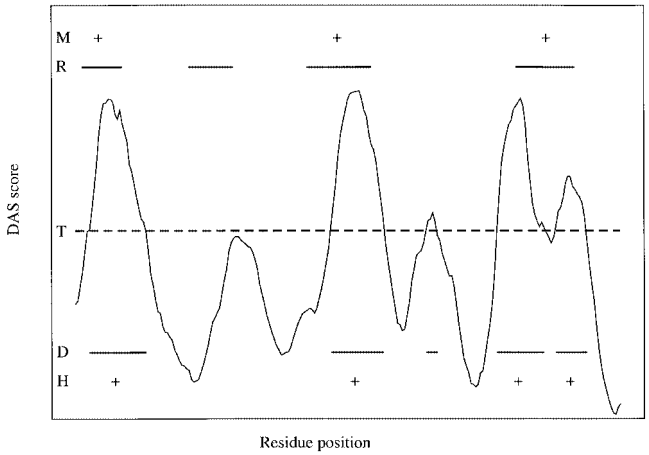

Fig. 2. Guideline for the calculation of the quality score Φ. The schematic DAS profile Λ(***i***,***Q***) for a query ***Q*** as function of the sequence positions of a hypothetical TM library sequence *i* is shown together with the score threshold (***T***, dashed line). The annotated TM segments (***R***) as well as the predicted TM core regions (***D***) are shown as horizontal bars. Overlaps of annotated with predicted TM regions as well as overlaps of predicted with annotated regions are denoted with + signs (***M*** and ***H***, respectively). The values of the variables in this example are $\boldsymbol{R} = 4$, $\boldsymbol{D} = 5$, $\boldsymbol{M} = 3$ and $\boldsymbol{H} = 4$.

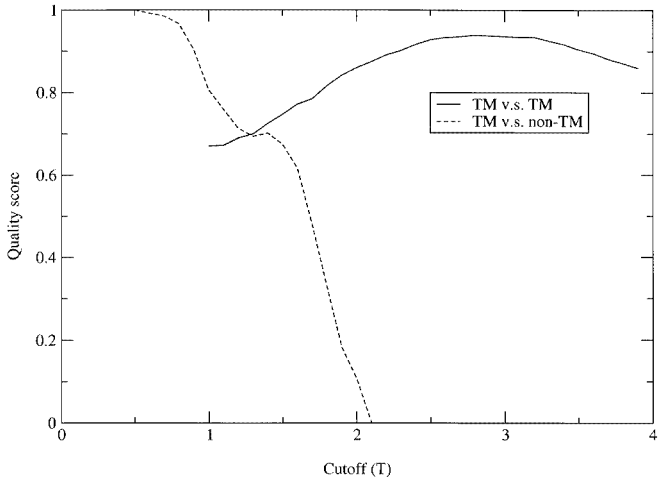

Fig. 3. Relationship between quality score Φ and DAS score threshold ***T*** for TM proteins and non-TM proteins averaged over large sequence sets. DAS profiles for all of the 128 TM set proteins averaged separately (i) against the TM set sequences (full line) and (ii) against all non-TM set sequences used as library in this calculation (dashed line) have been computed. It should be emphasized that our definition of the quality score Φ overestimates the prediction efficiency at low cutoffs $\boldsymbol{T} << 2$.

of the potential transmembrane core segments with high precision (Cserzö *et al.*, 1997).

The DAS method and the window-averaged hydrophobicity profile methods are different in principle. DAS describes the hydrophobic segments at three levels. First, a TM fragment is similar to any other as they are all made up of hydrophobic residues. Second, if two TM fragments are aligned, then the similarity remains high even if the two fragments are shifted relative to each other (condition of even distribution of hydrophobicity). Finally, if there are several TM fragments in either sequence, we expect to observe alternative alignment matches for all combinations of TM fragments. All three conditions are included in the new mathematical formulation of the DAS method (Equations 1–10). With its current algorithmic modification and parametrization, DAS profiles of query sequence are calculated and regions above a threshold ($\boldsymbol{T} = 2.5$ for window size $\boldsymbol{W} = 13$ was finally found optimal) are considered likely to be cores of transmembrane segments.

In the second step, the query sequence is used in a 'reverse' prediction cycle. At this stage, the query sequence is used to 'predict' TM segments in the sequences of the TM library. The results of the predictions are compared with the location of the known TM segments. The quality of this prediction distinguishes between TM or non-TM query type. Our experience shows that high-value library profiles with high quality scores are obtained when the query is a real TM protein. Weak profiles and low quality scores indicate non-TM queries (Figure 1). The error rate, i.e. the frequency of the wrong assignment, is significantly lower than in a direct application of any TM prediction method alone.

The quality score Φ evaluating the overlap of prediction and annotation can be calculated as presented in Equation 11 and Figure 2. Possible values of Φ are real numbers between 0 and 1; $\Phi = 1$ in the ideal case. We computed library profiles for all members of the learning set averaged separately: (i) over all TM set sequences and (ii) over all non-TM set sequences. The quality scores of predictions were then determined as a function of the threshold ***T*** (Figure 3). For non-TM input sequences, the quality Φ rapidly decreases with increasing threshold ***T***. With genuine TM proteins as input, the quality Φ is close to a maximum for a score threshold $\boldsymbol{T} \sim 25$. On the basis of the DAS curves of the library profiles, it seems that the TM core region prediction algorithm apparently best distinguishes the TM queries from the non-TM queries at around this cutoff value for the profile score.

We conclude that the quality score Φ is a reliable indicator for distinguishing whether a query with predicted TM core regions is a genuine TM protein or non-TM protein. For predicted single-pass transmembrane proteins, a quality score of $\Phi \leqslant 0.8$ provided a good criterion by which to reject many false positive predictions. We would suggest that this should allow more reliable automated screening of non-annotated genomic for TM proteins than has previously been possible. The details of this process are explained in the Methods section.

Methods

Databases

We selected 128 protein sequences with experimentally determined transmembrane topology for test purposes (Moller *et al.*, 2000); these are referred to as the 'TM set'. As contrasting examples of TM-free proteins, 527 sequences with known 3D structure were chosen. These 527 proteins constitute a representative subset of the PDB: a snapshot of the 'PDB-select' (Hobohm and Sander, 1994; Berman *et al.*, 2000). One entry, the structure 1PRC, which is a photosynthetic reaction center with many true TM helices, was excluded. The remaining 526 proteins are referred to as the 'non-TM set'. This provides two datasets of comparable sizes, which approximately reflect the estimated proportions of these two types of proteins in complete genomes.

RReM scoring matrix

The residue replacement matrix (RReM scoring matrix) used for generating the alignment surface of protein pairs is based on the neighborhood selectivity of amino acid pairs (up to 10 residues distant from each other in the sequence) and characterizes whether a certain amino acid is disfavored in

terms of its observed frequency versus its frequency expected by chance (Tudos *et al.*, 1990). The matrix has been recalculated from the combined SWISS-PROT/TREMBL database snapshot of March 2001 containing over 120 million residues (Bairoch and Apweiler, 2000). The calculations are described elsewhere (Cserzö and Simon, 1989; Tudos *et al.*, 1990).

The DAS-TMfilter algorithm: step 1

The DAS algorithm has undergone substantial modifications since its first publication (Cserzö *et al.*, 1994), so the current version of the algorithm is formally described in detail here. We denote the query protein sequence $\boldsymbol{Q}$ with sequence length $\boldsymbol{q}$ as $\boldsymbol{q}$-tuple of amino acid residues $\boldsymbol{a_j}$ ($1 \leqslant \boldsymbol{j} \leqslant \boldsymbol{q}$)

$$\boldsymbol{Q} = (\boldsymbol{a}_1,\boldsymbol{a}_2,...\boldsymbol{a}_q) \tag{1}$$

and the $\boldsymbol{i}$th TM library sequence $\boldsymbol{L}_i$ with length $\boldsymbol{l}(\boldsymbol{i})$ as $\boldsymbol{l}(\boldsymbol{i})$-tuple of amino acid residues $\boldsymbol{b_k}$ ($1 \leqslant \boldsymbol{k} \leqslant \boldsymbol{l}(\boldsymbol{i})$)

$$\boldsymbol{L}(\boldsymbol{i}) = (\boldsymbol{b}_1,\boldsymbol{b}_2,...\boldsymbol{b}_{l(i)}) \tag{2}$$

The square 20×20 RReM scoring matrix $\boldsymbol{Z}(\boldsymbol{a},\boldsymbol{b})$ evaluates the exchange from amino acid type $\boldsymbol{a}$ to amino acid type $\boldsymbol{b}$. The raw alignment surface $\boldsymbol{A}_0(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i}))$ of the query sequence $\boldsymbol{Q}$ and the library sequence $\boldsymbol{L}(\boldsymbol{i})$ is the matrix

$$\boldsymbol{A}_0(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i})) = [\boldsymbol{Z}(\boldsymbol{a}_j,\boldsymbol{b}_k)]_{1 \leqslant j \leqslant q, 1 \leqslant k \leqslant l(i)} \tag{3}$$

In the next step, alignment surface values from $\boldsymbol{A}_0$ are averaged along diagonal segments and a new alignment surface $\boldsymbol{A}_1$ is obtained. Here, we take advantage of the sequence similarity between any two TM regions and between the same two TM regions after small alignment shifts. Hydrophobic clusters in the query sequence will produce high values in $\boldsymbol{A}_1$ but not polar stretches since there are none in the TM library.

$$\boldsymbol{A}_1(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i})) = \left\{ \sum_{n=-w}^{w} \boldsymbol{Z}(\boldsymbol{a}_{j+n},\boldsymbol{b}_{k+n}) \right\}_{1 \leqslant j \leqslant q, 1 \leqslant k \leqslant l(i)} \tag{4}$$

In this equation, terms $\boldsymbol{Z}(\boldsymbol{a}_{j+n},\boldsymbol{b}_{k+n})$ with $\boldsymbol{j}+\boldsymbol{n} \leqslant 0$, $\boldsymbol{j}+\boldsymbol{n} > \boldsymbol{q}$, $\boldsymbol{k}+\boldsymbol{n} \leqslant 0$ or $\boldsymbol{k}+\boldsymbol{n} > \boldsymbol{l}(\boldsymbol{i})$ are set to zero. The window size $\boldsymbol{W} = 2\boldsymbol{w}+1$ is a pre-selected parameter. Further, we calculate the average matrix element value $\boldsymbol{e}$ and its standard deviation $\boldsymbol{d}$ of a matrix obtained after window averaging along diagonals from a randomized version of $\boldsymbol{A}_0$. These two values are used to normalize the $\boldsymbol{A}_1$ matrix and to screen all matrix values below a constant $\boldsymbol{c} = 1$ (one standard deviation):

$$\boldsymbol{A}_2(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i})) = \left\{ \boldsymbol{x} = \frac{1}{\boldsymbol{d}} \left[\sum_{n=-w}^{w} \boldsymbol{z}(\boldsymbol{a}_{j+n},\boldsymbol{b}_{k+n}) - \boldsymbol{e} \right] - \boldsymbol{c} \; \boldsymbol{x} > 0 \right\}_{\boldsymbol{x} \leqslant 0} \quad 1 \leqslant j \leqslant q, 1 \leqslant k \leqslant l(i) \tag{5}$$

This matrix will now be used for the computation of profiles for the library and the query sequences. The library profiles need to be stored for later use in the second step of the DAS-TMfilter algorithm for the exclusion of false positive predictions. Thus, the master profile $\Lambda(\boldsymbol{i},\boldsymbol{Q})$ for the $\boldsymbol{i}$th library sequence from the query $\boldsymbol{Q}$, a vector with $\boldsymbol{l}(\boldsymbol{i})$ components, is calculated as

$$\Lambda(\boldsymbol{i},\boldsymbol{Q}) = \frac{1}{2\boldsymbol{w}+1} \left(..., \sum_{n=-w}^{w} \Lambda_0(\boldsymbol{i},\boldsymbol{k}+\boldsymbol{n},\boldsymbol{Q}), ... \right)_{1 \leqslant k \leqslant l(i)} \tag{6}$$

with

$$\Lambda_0(\boldsymbol{i},\boldsymbol{k},\boldsymbol{Q}) = \frac{1}{\boldsymbol{q}} \sum_{j=1}^{q} \boldsymbol{A}_2(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i}))_{j,k}. \tag{7}$$

Again, terms with $\boldsymbol{k}+\boldsymbol{n} \leqslant 0$ or $\boldsymbol{k}+\boldsymbol{n} > \boldsymbol{l}(\boldsymbol{i})$ are set to zero. Finally, the profile $\mathrm{K}_0(\boldsymbol{Q},\boldsymbol{i})$ for the query sequence weighted with the $\boldsymbol{i}$th library profile is calculated as

$$\mathrm{K}_0(\boldsymbol{Q},\boldsymbol{i}) = \frac{1}{2\boldsymbol{w}+1} \left(..., \sum_{n=-\mathrm{w}}^{w} \mathrm{K}_1(\boldsymbol{Q},\boldsymbol{j}+\boldsymbol{n},\boldsymbol{i}), ... \right)_{1 \leqslant j \leqslant q} \tag{8}$$

with

$$\mathrm{K}_1(\boldsymbol{Q},\boldsymbol{j}+\boldsymbol{n},\boldsymbol{i}) = \frac{1}{\boldsymbol{l}(\boldsymbol{i})} \sum_{k=1}^{l(i)} \boldsymbol{A}_2(\boldsymbol{Q},\boldsymbol{L}(\boldsymbol{i}))_{j,k} \cdot \Lambda(\boldsymbol{i},\boldsymbol{Q})_{\mathrm{k}} \tag{9}$$

It is thought that this weighting increases discrimination if the library protein is a transmembrane protein with two or more TM regions. The DAS profile $\mathrm{K}(\boldsymbol{Q})$ for the query $\boldsymbol{Q}$ is obtained as average over all library sequences:

$$\mathrm{K}(\boldsymbol{Q}) = \frac{1}{\boldsymbol{N}} \left(..., \sum_{i=1}^{N} \mathrm{K}(\boldsymbol{Q},\boldsymbol{i})_j, ... \right)_{1 \leqslant j \leqslant n} \tag{10}$$

N is the total number of library sequences. Regions above the threshold $\boldsymbol{T}$ (selected as $\boldsymbol{T} = 2.5$) in $\mathrm{K}(\boldsymbol{Q})$ are considered putative core regions of transmembrane segments. The minimum distance between putative TM core region peaks is required to be larger than 33 residues; when such conflicts occur, the smaller peak is suppressed. This threshold was derived from less than 10 counter-examples found in the learning set after application of the DAS-TMfilter algorithm. Note that this rule is purely empirical and the small number of examples suggests that it is not critical. This completes the first step of the DAS-TMfilter algorithm.

The DAS-TMfilter algorithm: step 2

The second step of the DAS-TMfilter algorithm is designed to flag likely false predictions among all TM helix hits. The profiles $\Lambda(\boldsymbol{i},\boldsymbol{Q})$ for the TM library (defined with Equations 6 and 7) are used in the quality check back-end filter and treated separately. They are also searched for above-threshold regions and their coincidence with transmembrane segments annotated in the description of the library proteins is checked. We calculate a quality score Φ:

$$\Phi(\boldsymbol{Q}) = \sqrt{\frac{\boldsymbol{H}}{\boldsymbol{D}} \cdot \frac{\boldsymbol{M}}{\boldsymbol{R}}} \tag{11}$$

where $\boldsymbol{R}$ is the number of all annotated TM regions in the database of TM proteins used for post-processing, $\boldsymbol{D}$ is the number of all core regions predicted in the library with the query $\boldsymbol{Q}$ using the library profiles $\Lambda(\boldsymbol{i},\boldsymbol{Q})$, $\boldsymbol{M}$ is the number of annotated TM regions in any library sequence that overlap with a predicted one and $\boldsymbol{H}$ is the number of predicted core regions in any library sequence that overlap with an annotated TM region (see Figure 2 for a graphic explanation). If $\boldsymbol{D} = 0$, Φ is set to 0. We found out that a quality score $\Phi \leqslant 0.8$ is a good criterion for the rejection of a predicted helical transmembrane region in a putative single-TM segment protein. To derive useful criteria for multi-TM segment proteins, larger datasets need to be investigated to obtain

more than just singular examples of false positive predictions. This will be one of our future tasks.

In this work, the window size $W = 2w + 1$ for the alignment surface scan was fixed to 13 residues. Calculations with other possible values suggest that the algorithm is not very sensitive to the value of this variable but $W = 13$ seems optimal. This value represents the core region of a minimal length TM helix of 19 residues, omitting three residues at each end.

Computation of probabilities of false positive prediction

The DAS-TMfilter profile scores are not easily interpreted, since they are not directly comparable to results from other prediction methods that might hit into the same query sequence region. Any good prediction method attempts to introduce a probabilistic measure that estimates the reliability of predictions. In this method, we derive an E-value for each predicted TM core region.

Lets assume that the value of a DAS-TMfilter profile for a given query sequence position is a random variable with normal distribution. Then, the local profile maxima over a given sequence stretch can be considered extreme-value distributed. We derived a set of profile values corresponding to local maxima with sequential distance of at least 13 residues (one window length) from the non-TM set (total 525 proteins after exclusion of 1PRC, a true transmembrane protein, and 1COL, a protein with facultative transmembrane regions). The search resulted in 5425 data points for a total sequence length of 139 624 residues in the non-TM set, i.e. one peak per about 26 residues. The empirical distribution was fitted to with an extreme value distribution function where $P(\mathbf{score} \geqslant S)$ is the probability of a finding a profile score larger or equal to S by chance:

$$P(score \geqslant S) = 1-\exp\{-\exp[-\lambda(S-u)]\}. \quad (12)$$

The correlation coefficient between $1n\{-1n\ (P_{observed}(score \leqslant S))\}$ and $-\lambda(S-u)$ is 0.99956 for the coefficients $\lambda = 3.529828$ and $\lambda u = 3.624375$. The regression is validated by the t-tests for the regression coefficients (Student's t-test values for slope λ and intercept λu are –2485.7 and 2051.7, respectively) and Fisher's test for comparison with the function average ($F = 6178714$); and all significances are clearly below 0.001. For the computation of the E-values (probability of false positive TM prediction), the sequence length q of the query enters the equation

$$E(score \geqslant S) = \frac{q}{26}(1-\exp\{-\exp[-\lambda(s-u)]\}. \quad (13)$$

Obviously, the chance of finding a transmembrane region-like segment in a random sequence increases with sequence length. For example, the score threshold of 2.5 for TM core region prediction corresponds to an expected false positive prediction of 0.0055 for this individual region in a short sequence stretch. For a 260-residue protein, the probability is higher and estimated as 5.5%. For example, TM core scores 3.0, 4.0 and 5.0 yield P-values of 9.4×10^{-4}, 2.8×10^{-5} and 8.1×10^{-7}, respectively, for a small sequence span and a 10-times higher E-value for a medium-sized protein. Thus, scores higher than $S = 2.5$ will result in dramatically lower rates of false positive prediction and, consequently, in more reliable assignments.

Results and discussion

Computation of prediction accuracy: self-consistency test

In the first cycle, the DAS-TMfilter algorithm with window size $W = 13$ was applied for each of the 128 sequences of the TM library against the total library but the actual sequence itself (self-consistency test). Since the algorithm is symmetrical for any two sequences, a total number of 8128 individual pairs of profiles was obtained in the calculations. The new program recognizes 94% of the documented TM regions. There are 12 signal peptides in proteins of the TM library; 10 of them are matched by a strong DAS peak. Apparently, signal peptides can be considered true TM segments, but not permanent ones as they are cleaved off from the matured protein. Therefore, the peaks for the signal peptides were counted as true positives in our work. The two signal peptides missed were considered as false negatives.

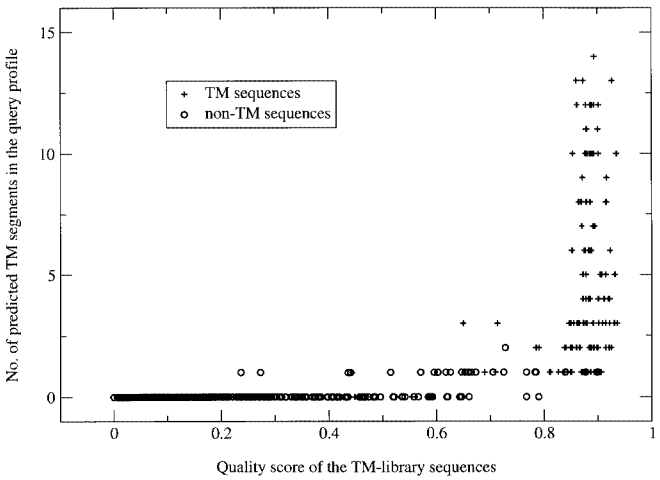

Fig. 4. Relationship between the number of predicted TM segments and the quality score for queries from the TM set and the non-TM set. The number of predicted TM segments in a query plotted against the quality score Φ of each query for the individual sequences of the TM set (marked +) and for the sequences of the non-TM set (marked ○). For the first part, the TM core prediction run, the top eight sequences in Table I (first column) have been taken. All 128 sequences of the TM set have been used for quality score calculations. Note that the quality score is a discrete value and, therefore, it may happen that some markers in the plot are on top of each other.

The original DAS code was fine-tuned on a TM-database of 44 prokaryotic sequences and achieved 96% overall predictive power. The learning set used in the present study is almost three times that size and also includes eukaryotic sequences. These results demonstrate the stability of the DAS predictions as the database size increases.

The efficiency of the DAS-TMfilter with respect to the number of predicted TM regions at the whole protein level is as follows: in 88 of the 128 sequences the correct number of TM segments was detected. The method finds less than the reported number of TM segments for 20 proteins: one segment is missed in 11, two in five and three segments were missed in four. False positive segments were predicted for 20 TM proteins: one false positive TM region was found in 14 learning set sequences, two false positive TM segments were detected in three, three false positives in two and four false positives in one protein.

Computation of prediction accuracy: false positive predictions of proteins with membrane-spanning regions

The second cycle of calculation compared the set of profiles derived from the non-TM set sequences against the profiles from the TM set. This analyzed 67 328 pairs of individual profiles. In Figure 4, the number of predicted TM segments detected in the DAS curves of the query sequence is shown as a function of the quality score Φ of the TM library sequences against that query. Each protein of the non-TM set (○) and of the TM set (+) is shown. This map establishes a link between the hydrophobicity

information contained in the query sequence and the effect of the query on the DAS profiles of the library sequences.

The two sets of sequences are reasonably separated. For most of the non-TM sequences, no TM segment is detected. In two sequences of the 526 proteins, two TM segments are assigned to a non-TM protein (oxidoreductase 1LCI with quality Φ = 0.67; colicin A C-terminal pore-forming domain 1COL with quality Φ = 0.96). The two predicted TM segments for colicin A (PDB-code 1COL), which occur as a twin peak over a core segment between residues 532 and 575, coincide with those helices that are predicted to enter the membrane (residues 528–548 and 555–575 in the annotation of SWISS-PROT entry P04480). This bacterial toxin is stable both in soluble form and, after conformational changes, in membrane-inserted form. Owing to this amphipatic behavior, we do not consider this as a false positive hit. In the case of the TM set, only three sequences of the 128 (with SWISS-PROT accessions P32839, P07371, P03655) were missed. The most critical region of the comparison in Figure 4 is the line corresponding to predicted single-pass proteins. There are 29 TM and 29 non-TM proteins in this zone and these are fairly well discriminated by setting an appropriate quality score Φ limit. For example, there are four TM proteins below the Φ = 0.8 value (P25060, P06008, P32175, P32897) and only five non-TM sequences (PDB accessions 1CPO, 1LTS, 1NOX, 1OXA, 1PVC) above that value. With this setting of parameters, we miss 7/128 TM proteins (5.5%, rate of positive prediction = 94.5%) at the expense of including 7/526 non-TM proteins (1.3%). We think that the quality score Φ limit applied should ultimately be related to the number of predicted TM regions in a target protein. The current data allow us to define this parameter fairly accurately for single-pass TM proteins. Further tests with larger sets of negative queries will be required to find the optimal value for TM/non-TM discrimination in proteins with multiple membrane-spanning segments.

We examined the sequences of limited numbers of false negative and false positive results that persist with this improved algorithm. All of the non-predicted TM regions in genuine TM proteins were relatively rich in glycine, serine and other small residues and/or they included multiple polar residues. It is possible that these annotated TM regions may only be able to function as transmembrane segments when in a complex that includes other TM proteins. The predicted TM regions that were incorrectly assigned in non-TM proteins are largely α-helical and typically occur in proteins that include a many-layered packing of secondary structure elements that comprise 1–2 long, very hydrophobic helices that are packed against other secondary structural elements within the core of a globular structure. It is not a surprise that, as exemplified by 1COL, such helices may sometimes function as transmembrane regions after a conformational change has been triggered by insertion of the protein into a membrane and/or a multi-protein complex (Lakey and Slatin, 2001).

Computation of prediction accuracy: false positive prediction of single TM segments

We wanted to have an estimate for the false positive prediction rate of single TM regions within a given query protein. A predicted transmembrane core region can be characterized by, for example, its peak height in the profile, the sequence length of the core or the area between the profile and a horizontal line corresponding to the threshold $\boldsymbol{T}$ = 2.5. We found that these values were not correlated with one another. Since the peak height is the major parameter for core selection, we used this value for assessing the probability of false positive prediction.

The distribution function of peak height of local maxima in the DAS profiles of sequences in the non-TM set matches an extreme-value distribution very well (see Methods for details), suggesting that the DAS profile value is normally distributed and the peak size of its local maxima is extreme-value distributed. With these assumptions, we could calculate a probability of false positive prediction ($\boldsymbol{E}$-value) for comparison with other prediction methods (Altschul *et al.*, 1997; Eisenhaber *et al.*, 1999, 2001). Obviously, the $\boldsymbol{E}$-value depends also on sequence length since the TM core region can be anywhere in the sequence. For example, a score of 2.5 for a predicted TM core region in a medium-sized protein corresponds to an expected false positive prediction of ~5% for this individual region.

Reduction of the learning set

Generally, knowledge-based prediction methods are expected to be cross-validated with statistical procedures such as the jack-knife test to monitor the stability of parameters relative to the learning set. In our case, the number of parameters is small (window size $\boldsymbol{W}$ = 13, score threshold $\boldsymbol{T}$ = 2.5, quality score threshold $\Phi \leqslant 0.8$). Since the method relies on multiply shifted alignments between putative and documented TM regions and all TM segment sequences are similar to themselves and to each other in this respect, traditional jack-knife procedures are not very sensitive. To emphasize, the concept of statistically significant sequence similarity between distantly related sequences is generally considered not applicable to sequence segments rich in TM regions because of their compositional bias. Indeed, we measure unchanged positive prediction rates over the TM set in a jack-knife test. A reduction of the learning set would be a more serious criterion.

A smaller learning set has also practical advantages since, in the presented implementation of DAS-TMfilter, the final prediction is based on several pairwise DAS runs of the query with each library sequence. This procedure aims the reduction of the noise of the individual DAS curves of the query through averaging. On the other hand, the computational time of the calculation is proportional to the number of sequences in the TM library. Using a large TM library is therefore useful only if the cost imposed by more runs gains us accurate curves. We explored the effect of the number of library sequences on the accuracy of the query curves. A series of runs were carried out where each sequence of the TM set was in turn selected as the library and the rest of the TM set were submitted against that as queries. The quality scores of the TM queries varied considerably (data not shown), suggesting that individual library sequences made different contributions to the prediction accuracy. By using only the top-scoring eight proteins (Table I, first column) as the TM library in the first computation step, we achieve an overall predictive power of 95% (recognition of transmembrane regions in the TM set). At the same time, the computation is speeded up 16-fold.

The effect of the library size in terms of discrimination between TM or non-TM query proteins were tested in a similar manner. Runs on TM set and the non-TM set against single sequence libraries suggest that again the potential library sequences have different discriminatory values. We can reduce the number of library sequences down to 16, eight or even four top-scoring proteins (Table I, second column) without risking the discriminative power of the method on the current test database.

Interestingly, there is no overlap between the top-scoring subsets of proteins from the two lists. Apparently, a TM protein can provide an accurate profile for the query or it can be very

Table I. The list of the SWISS-PROT accession numbers for the 16 top-scoring TM library sequences in terms of producing accurate query profiles (first column) and in terms of discriminative power of the type of the query sequence (second column): the final version of DAS-TMfilter uses the upper eight entries from each column

P15877	P02945
P32174	P19568
P11026	P03617
P18537	P05701
Q53068	P77921
P18582	P23215
P35523	P26790
P12691	P17448
P52205	P02699
P07038	P27125
P02916	P31602
P19673	P03805
P18783	P23889
Q54397	P23978
P08194	P36574
P03844	P11551

sensitive for calculations of prediction quality of the query, but it cannot serve the two tasks at the same time.

Application of the DAS-TMfilter program and interpretation of its output

The DAS-TMfilter prediction method operates along the following lines: individual DAS runs are performed using the query sequence of an unknown protein against the selected first set of TM library sequences. The resulting individual DAS curves of the query are averaged over the library and evaluated. If there is no peak above the empirical cutoff limit of 2.5, the query is classified as a non-TM protein. If there are two or more peaks above the cutoff, the query is recognized as a true TM protein. If only one peak is detected, the back-end filter of the program is invoked. The query is compared with the second half of the TM library and the quality of the resulting DAS curves of the library sequences is again evaluated. If the quality score Φ is higher than the empirical value, the query is classified as a TM protein. Otherwise, it is assigned as a non-TM protein. In its current implementation, the program can process more than 1000 protein sequences per hour on a standard workstation.

The prediction itself consists of a list of peaks and regions above the cutoff limit that are assigned as TM core segments. Here, however, we stress that there is a basic difference between TM cores and TM helices. The cores are the detectable parts of the TM helices. They can be very narrow if the DAS signal is weak or they can be very wide in case of strong DAS signals. The bundle-forming tendency of TM helices is well known. In such transmembrane proteins, the outer members of the helix bundle are exposed to the lipid phase and are very hydrophobic. These will yield strong DAS signals. The inner members of a bundle are buried and are often less hydrophobic and may give weaker DAS signals. A few signals are so weak that DAS-TMfilter can detect only one residue long core. Even these weak signals should be taken seriously; however, most of the false positive detections are also weak signals. The quality score Φ separates most of them.

A peak on the list of a DAS-TMfilter prediction therefore means only that it is within a TM segment: it is not informative about the start and end of the relevant TM helix. At present, the relatively small size of the TM database and the high error rate of the cited helix end-points within it prevent any serious development in this respect. Further progress will require more learning set sequences and more accurate annotation of these sequences.

Although DAS-TMfilter is a significant step in the development of TM prediction, the new algorithm still encounters problems with a few of the largely α-helical soluble proteins (these might be recognized by running fold recognition programs in parallel) and it misses some weak TM regions in true single-pass TM proteins. Considering the future of TM region prediction, it may be impossible to improve prediction rates further (in terms of sensitivity and/or selectivity) with a local sequence segment approach that does not consider the whole structure or even the potential formation of complexes.

More than a dozen efficient TM prediction tools have been published, but only two of them discriminate between real TM and non-TM queries. Both of these claim 98% efficiency in terms of TM segment recognition and 99% selectivity for the correct query type (Hirokawa *et al.*, 1998; Krogh *et al.*, 2001). In the case of the SOSUI tool, it is difficult to comment on the reported results because the method is not described in sufficient detail to judge the real merit of their approach. The TMHMM tool implements a hidden-Markov model to locate TM segments and to identify the query type on the essentially same learning sets as we used as our TM and non-TM sets, so the results can be directly compared. The efficiency of the two methods is similar in the query type identification step. The small (3%) difference in the TM detection step might be a result of the different approaches taken by the methods, for example, by the different number of model parameters (there is a handful of parameters for DAS but at least an order of magnitude more for TMHMM). Here we again emphasize that the experimental TM database is small and is very likely to include a few errors. Application of an 'unsupervised learning' approach, such as a hidden-Markov model, to such a database tends to overestimate its real efficiency and prediction accuracy may reduce if it is applied to a more comprehensive set. Moreover, the DAS-TMfilter approach is more closely related to physical principles; it contains only one sensitive parameter, the empirical cutoff, that affects efficiency. Therefore, we consider the small apparent difference in prediction accuracy between the two methods as the fluctuation of two independent estimations of the real efficiency on a database of ultimate size with a real value somewhere between the two quoted success rates. However, it is not possible to provide the exact statistical significance of this 3% difference at present.

Acknowledgements

M.Cs. acknowledges the support of an MRC Bioinformatics Infrastructure grant for the University of Birmingham and IMP for travel support. The research of B.E. and F.E. was financially supported by the Austrian agency FWF (grant P15037), by the Austrian National Bank (OeNB, Österreichische Nationalbank) and by Boehringer Ingelheim. I.S. acknowledges support from the Hungarian Scientific Research Fund (OTKA T 30566 and T 34131). M.Cs. and I.S. also acknowledge the support of the British–Hungarian Inter-Government Science and Technology Fund. We thank Bob Michell for his helpful comments during the preparation of this paper.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 5–48.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Cserzö,M. and Simon,I. (1989) *Int. J. Pept. Protein Res.*, **34**, 184–195.

Cserzö,M., Bernassau,J.M., Simon,I. and Maigret,B. (1994) *J. Mol. Biol.*, **243**, 388–396.

Cserzö,M., Wallin,E., Simon,I., von Heijne,G. and Elofsson,A. (1997) *Protein Eng.*, **10**, 673–676.
Eisenhaber,B., Bork,P. and Eisenhaber,F. (1999) *J. Mol. Biol.*, **292**, 741–758.
Eisenhaber,B., Bork,P. and Eisenhaber,F. (2001) *Protein Eng.*, **14**, 17–25.
Eisenhaber,F., Persson,B. and Argos,P. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.
Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) *Bioinformatics*, **14**, 378–379.
Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.
Hucho,F., Gorne-Tschelnokow,U. and Strecker,A. (1994) *Trends Biochem. Sci.*, **19**, 383–387.
Jayasinghe,S., Hristova,K. and White,S.H. (2001) *J. Mol. Biol.*, **312**, 927–934.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *Biochemistry*, **33**, 3038–3049.
Juretic,D., Zucic,D., Lucic,B. and Trinajstic,N. (1998) *Comput. Chem.*, **22**, 279–294.
Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) *J. Mol. Biol.*, **305**, 567–580.
Kuhlbrandt,W., Wang,D.N. and Fujiyoshi,Y. (1994) *Nature*, **367**, 614–621.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lakey,J.H. and Slatin,S.L. (2001) *Curr. Top. Microbiol. Immunol.*, **257**, 131–161.
McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) *Bioinformatics*, **16**, 404–405.
Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) *Bioinformatics*, **16**, 1159–1160.
Moller,S., Croning,M.D. and Apweiler,R. (2001) *Bioinformatics*, **17**, 646–653.
Nakai,K., Kidera,A. and Kanehisa,M. (1988) *Protein Eng.*, **2**, 93–100.
Nikiforovich,G.V. (1998) *Protein Eng.*, **11**, 279–283.
Nilsson,J., Persson,B. and von Heijne,G. (2000) *FEBS Lett.*, **486**, 267–269.
Palliser,C.C. and Parry,D.A. (2001) *Proteins*, **42**, 243–255.
Pasquier,C. and Hamodrakas,S.J. (1999) *Protein Eng.*, **12**, 631–634.
Pasquier,C., Promponas,V.J., Palaios,G.A., Hamodrakas,J.S. and Hamodrakas,S.J. (1999) *Protein Eng.*, **12**, 381–385.
Persson,B. and Argos,P. (1996) *Protein Sci.*, **5**, 363–371.
Picot,D., Loll,P.J. and Garavito,R.M. (1994) *Nature*, **367**, 243–249.
Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) *Protein Sci.*, **4**, 521–533.
Simon,I., Fiser,A. and Tusnady,G.E. (2001) *Biochim. Biophys. Acta*, **1549**, 123–136.
Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
Tompa,P., Tusnady,G.E., Cserzö,M. and Simon,I. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 4431–4436.
Tudos,E., Cserzö,M. and Simon,I. (1990) *Int. J. Pept. Protein Res.*, **36**, 236–239.
Tusnady,G.E. and Simon,I. (1998) *J. Mol. Biol.*, **283**, 489–506.
Tusnady,G.E. and Simon,I. (2001a) *Bioinformatics*, **17**, 849–850.
Tusnady,G.E. and Simon,I. (2001b) *J. Chem. Inf. Comput. Sci.*, **41**, 364–368.
von Heijne,G. (1992) *J. Mol. Biol.*, **225**, 487–494.
Weiss,M.S. and Schulz,G.E. (1992) *J. Mol. Biol.*, **227**, 493–509.

Received November 29, 2001; revised April 26, 2002; accepted May 21, 2002