Improved detection of homologous membrane proteins by the use of topology predictions

Maria Hedman *

September 21, 2001

Supervisor: Arne Elofsson Stockholm Bioinformatics Center, Stockholm University

*Stockholm Bioinformatics Center, SCFAB, Stockholm University, SE-106 91 Stockholm, Sweden E-mail:maria@sbc.su.se

Abstract

25 % of the sequences in all genomes code for membrane proteins. These proteins are involved in fundamental reactions in the cell and are of particular importance in pharmaceutical industry. Membrane proteins have vastly different properties than globular proteins. However search and alignment methods that are available today are developed and optimized primarily for globular sequences and have a sub optimal performance for membrane sequences. Here we present a novel prediction based alignment method for membrane sequences that perform significantly better than standard sequence alignment methods. We also show that predicted secondary structure information can increase the specificity of PSI-BLAST for membrane sequences.

Contents

1 Introduction

To search for homolugous sequences in databases of known sequences has been shown to be one of the best methods to gain information about an unknown protein. In contrast to experimental studies, database searching is fast and suitable for handling large sets of sequences. As the genome projects proceed the sequence databases continue to grow and we are every day presented with sequences for which a lot remains to be known about their structure, function and evolutionary origin. The demand for computer algorithms which are able to make use of the sequence databases as tools in biological research is huge.

Most sequence database search programs are based on sequence alignments in order to compare and measure the similarity between sequences. The alignments can be calculated in various ways modeling different biological perspectives. The first sequence alignment algorithm described in biological literature, the Needleman-Wunsch algorithm [SC70], considers similarities over the full length of the sequences compared. Shortly after a slight variant, the Smith-Waterman algorithm [TM81], was presented for finding optimal local alignments of two sequences. The idea of finding local regions of similarity agrees with the observation that distant homolugous sequences often share isolated regions of similarity even though the overall sequence similarity is low. The Smith-Waterman algorithm has formed the bases for many subsequent alignment algorithms and database search programs, such as BLAST [AGM+90] and FastA [DW85].

The introduction of PSI-BLAST [AMS+97] significantly increased the ability to find weak but biologically interesting similarities. Distantly related sequences may not be possible to detect from pairwise sequence alignments only. The innovation of PSI-BLAST was that, following an initial BLAST search it creates a multiple sequence alignment and a position specific profile from hits that match the query above a certain threshold. The profile is used as scoring scheme in the next iteration of the program. The use of multiple sequence alignments aids the detection of distant similarities by using information on conserved regions within families of related sequences.

Almost all currently available alignment methods are developed and optimized for globular sequences. However recent studies show that 20-30 % of the genes in all genomes code for membrane proteins [KLvHS01].

Alignment programs such as BLAST and PSI-BLAST are sub optimal methods to identify relationships between membrane sequences. Since membrane sequences are partly located in the membrane they consist of long stretches of hydrophobic residues with low compositional complexity and non related membrane sequences have higher degree of sequence similarity than unrelated non membrane sequences. Consequently membrane sequences display significant differences from globular proteins but scoring matrices, parameter settings and statistics used in

BLAST and PSI-BLAST are based on properties observed for average proteins.

The widespread use of database searching and the great interest in membrane proteins in medical research underscores the importance of developing accurate alignment techniques for membrane proteins. The issue of calculating scoring matrices for membrane proteins have been addressed previously [NHH00, TSR01]. The results show a slight improvement compared to ordinary scoring matrices (BLOSUM62) but no comparison against multiple sequence search methods such as PSI-BLAST is done in any of these studies.

In this study we present a novel prediction based alignment method for membrane proteins. The technique is similar to what has been used in threading methods for globular proteins [FE96] where addition of predicted secondary structure information in the alignment process has proved to significantly aid fold recognition. The fact that secondary structures for membrane sequences can be predicted with very high accuracy [BRPC95] suggests that the same technique has potential to significantly aid the alignments of membrane proteins. Further we evaluate the performance of PSI-BLAST for the ability to detect homolugous membrane proteins.

2 Background and Theory

Modern biology has more and more turned into a data rich science. Novel biological experiments create massive amounts of data and the need for storing and communicating large data sets grow continuously. The most obvious example is the recently finished nucleotide sequence determination of the human genome. A new field of science dealing with issues, challenges and new possibilities created by the biological databases has emerged: bioinformatics.

Improving database search algorithms is one of the most important challenges in bioinformatics today. The methods to find related sequences have evolved from fairly simple pairwise sequences comparisons to methods which are able to reveal relationships even when sequence similarity is very low.

In this section some of the basic techniques which are of particular importance for this work are presented. The special features of membrane sequences and methods for predicting transmembrane helices will be treated separately at the end of the section.

2.1 Sequence alignment

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the protein databases continue to grow the chance of finding such homologies in-

creases and the databases become more and more useful tools. Algorithms which are able to reveal biologically significant information for large sets of sequences in reasonable time are therefore of great importance.

Alignment of sequences is a fundamental process in the algorithms underpinning most sequence database search programs. The alignments are based on scoring schemes which state the probabilities of all possible amino acid substitutions. In the simplest case identical amino acids are scored one and every pair of non identical amino acids is scored zero. Search algorithms based on this scoring system will be of limited usage, since only sequences with very high sequence identity will be found. In order to detect biologically interesting sequences, which are more distantly related, the scoring system has to model the process of evolution.

The point accepted mutation (PAM) and blocks substitution matrices (BLOSUM) are two of the most popular matrix series [MRB78, SJ92]. The PAM matrix is computed by counting mutations between closely related sequences to obtain PAM 1 target frequencies. The PAM 1 matrix is multiplied by itself to get series which correspond to larger evolutionary distances. The PAM 1 matrix estimates scores for accepting one mutation per 100 positions; the PAM 170 is constructed by multiplying PAM 1 by itself 170 times and estimates 170 accepted point mutations per 100 positions.

Whereas the PAM model is based on closely related sequences, the BLOSUM model is based on substitution rates observed over large evolutionary distances. BLOSUM 62, for example, is derived from clusters of related sequence segments that are less than 62% identical.

To address the occurrence of substitutions and deletions in protein evolution the alignment algorithm may allow gaps to be introduced in the alignment. The number and length of the gaps are restricted by the use of scoring penalties. The final alignment score is then a function of the identity between aligned residues and the number and length of the gaps introduced.

Different alignment algorithms perform alignments in rather different ways. The first alignment algorithm described in biological literature was the Needleman-Wunsch algorithm [SC70], which aims at optimizing the alignments over the full length of the sequences (global alignment). Subsequently a slight variant was proposed, the Smith-Waterman algorithm [TM81], which searches for isolated regions of similarity between the sequences (local alignment). Local alignments may produce more sensitive and biological interesting alignments since functional sites are localized in relatively short regions, which are conserved irrespective of insertions and deletions in the sequence. Several popular sequence database search programs are based on the Smith-Waterman algorithm.

2.1.1 BLAST

The Needleman-Wunsch and Smith-Waterman algorithms are too time consuming to be realistic tools for comparisons against large sets of sequences, which is the case in database searching. To address the issue of speed database search programs use heuristic techniques.

The BLAST (Basic Local Alignment Search Tool) [AGM+90] searches sequence databases for optimal local alignments to a query sequence. The original BLAST program can only produce ungapped alignments. A modification of the algorithm has been introduced to generate gapped alignments [AMS+97]. The algorithm seeks one initial pair of sub sequences which form an ungapped alignment. Dynamic programming is used to extend the central pair of aligned residues in both directions to yield the final gapped alignment.

Protein sequences may contain regions which have low compositional complexity (i.e., regions within a sequence that have high densities of particular residues, e.g. GAPGAPGAPGAPGAP... such as occur in repetitive often tightly structured sequences such as collagen). These regions will result in a high number of spurious high-scoring matches biasing the result. To overcome this problem SEG may be used as a part of the BLAST routine. SEG uses the method of Wootton and Federhan [JS96] to divide a sequence into regions of high and low complexity. The output is a sequence just like the input sequence except that if low-complexity regions are found, the amino acid characters in these regions are replaced by X's. A BLAST search ignores these X regions.

To know whether a given alignment constitutes evidence for homology it helps to know how strong an alignment can be expected from chance alone. BLAST reports the raw score of the calculated alignments as well as assessments of their statistical significance. These assessments take the form of E-values. The E-value E for a given alignment depends upon its score S, as well as the length m and n of the sequences. The parameters K and λ can be thought of simply as natural scales for the search space size and the scoring system used.

$$E = Kmne^{-\lambda S}$$

The E-value represents the number of distinct alignments with equivalent or superior score that might have been expected to have occurred purely by chance. An E-value close to zero suggests a significant hit.

2.1.2 PSI-BLAST

More distant protein similarities may be impossible to detect from pairwise sequence comparisons only. Information identified from multiple sequence alignments of sequences in the same family may be used to increase the sensitivity.

PSI-BLAST (Position-Specific Iterative BLAST) [AMS+97] allows the creation of position specific profiles from sequences which match the query sequence above a certain threshold. The basic steps of the algorithm are described below.

1. PSI-BLAST takes a single sequence as input. The sequence is compared to the database using the gapped BLAST program.

2. Hits above a certain threshold are used to create a multiple sequence alignment and then a profile. The profile takes the form of a $L \times 20$ matrix, where L equals the length of the query sequence. The profile holds information on how conserved a particular residue is.

3. The profile is compared to the protein database, again seeking local alignments. The BLAST algorithm is used slightly modified, the query of length L and the 20×20 scoring matrix used originally are replaced by a position specific scoring matrix of dimension $L \times 20$.

4. PSI-BLAST estimates the statistical significance of the local alignments found.

5. Finally PSI-BLAST iterates by returning to step (2) an arbitrary number of times or until no further hits above the threshold are detected.

PSI-BLAST uses information on conserved regions, derived from multiple sequence alignments, and the profile is refined in each round of the program. The iterative process increases the search sensitivity but may cause problems as well. Once a false hit is incorporated in the profile the search will hereafter be biased to accept many more unrelated sequences. The threshold for inclusion of new sequences in the profile is therefore a parameter of crucial importance.

2.2 Threading techniques

Sequence alignments can be used to gain information about the structure of a unknown sequence, provided that homolugous sequences of known structure exist. However, there are many proteins with similar structure where no obvious homology has been detected.

Fold recognition methods, threading methods, present an alternative method to predict the structure of a unknown amino acid sequence. The idea is to fit the unknown probe sequence to a database of known target folds and evaluate which fold is most compatible with the sequence. Threading methods can roughly be divided in two categories: structural methods [JRD91, TRM92, MS92, MD93] and prediction based methods [FE96, RSS97]. Structural methods create a compatibility function [DDJD96] that describes how well a probe sequence matches

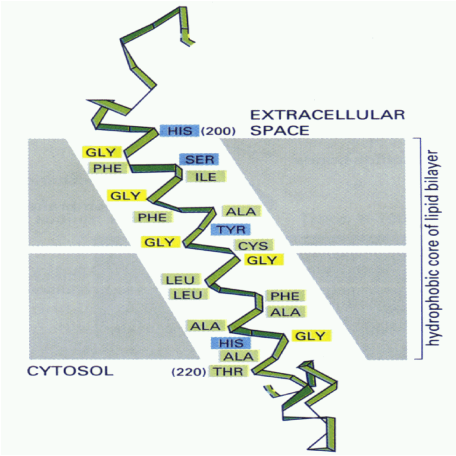

Figure 1. Illustration of membrane spanning helix

the target fold. The compatibility function may refer to the energy of each residue [JRD91] or structural properties of neighboring residues [MS92].

In contrast to the structural methods, prediction methods also take into account sequence derived predicted properties such as the predicted secondary structure of the probe sequence. These methods add a positive score if the predicted secondary structure for a certain residue agrees with the secondary structure state of the residue. Inclusion of predicted secondary structure information has proved to improve fold recognition by about 25% [FE96].

In this study we present a prediction based alignment method for membrane sequences that uses a similar technique to what has been done in prediction based threading studies. The fact that secondary structures of membrane proteins can be predicted with very high accuracy suggests that the same approach should be especially suitable for membrane sequences.

2.3 Membrane proteins

Cells and organelles within cells are bounded by membranes, a lipid bilayer that is hydorphilic on its two outer sides and hydrophobic in between. Proteins embedded in the membrane serve as mediators between the cell and its environment. The protein molecule may be bound to the membrane in different ways. It may traverse the membrane as one alpha helix or form several transmembrane alpha helices, which are connected by loops and traverse the membrane several times. Membrane proteins may also have several beta strands which form a channel through the membrane. The residues buried in the membrane have to perform

hydrophobic interactions with the membrane lipids and therefore tend to have hydrophobic properties.

Figure 1 shows an illustration of a single spanning transmembrane alpha helix. The membrane is about 30 Å wide and transmembrane helices are typically about 20-30 amino acids long to span the membrane. Due to the strictly limited number of residue types, transmembrane regions can be seen as stretches of low compositional complexity.

Experimental studies of membrane proteins are complicated by their special features and the characteristics of the environment. Membrane proteins are difficult to express in large quantities [RC95], hard to purify and especially difficult to crystallize. Extracting transmembrane proteins from the membrane will very easily disrupt them from their native structure. While there are x-ray structures representing hundreds of different folds of globular proteins, less than 10 different integral membrane protein structures are known to high resolution. The membrane structures that are known so far indicates that the architectural principles for proteins embedded in the membrane are far less diverse than those for globular proteins. The structure and assembly of membrane proteins are carefully described in [vHG97] and [vHG99].

2.3.1 Predicting transmembrane helices

Knowledge of the presence and location of transmembrane helices in a protein sequence is important for functional annotation an function analysis. Since residues in transmembrane regions need to perform interactions with the lipid molecules in the interior of the membrane transmembrane alpha-helices consist of unusually long stretches of hydrophobic residues which generally are easy to detect.

Some transmembrane helices can be located with high reliability by simply using hydrophobicity plots. In these plots the residue numbers are plotted versus the corresponding hydrophobicity indices. The hydrophobicity index states the hydrophobic character of a certain amino acid residue. Peaks above the zero line indicate possible transmembrane regions. Figure 2 shows hydrophobicity plots for glycophorin and bacteriorhodopsin. Compared to methods for secondary structure predictions of globular proteins transmembrane alpha helix prediction methods have proved to be significantly more accurate. It is not unusual that methods for predicting the location of transmembrane regions do 95 % correct predictions [MEI+97] [BRPC95].

2.3.2 TMHMM

In this study we used TMHMM [KLvHS01] to predict transmembrane alpha helices in the sequences. TMHMM does not only predict the locations of trans-

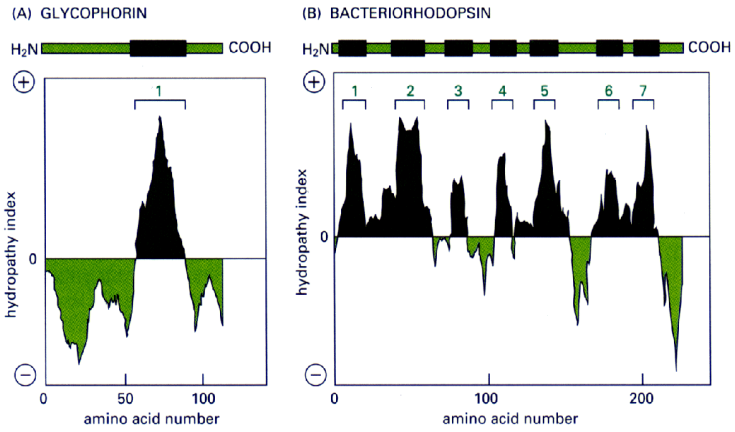

Figure 2. Hydrophobicity plots for a single membrane spanning proteins, glycophorin (A) and a multiple spanning membrane protein, bacteriorhodopsin (B). The residue numbers from N-terminal to C-terminal are plotted on the x-axis versus the corresponding hydrophobicity indices on the y-axis. The corresponding transmembrane regions are shown on top of the plots

membrane regions but also the topology of the protein including information on the orientation of the helices. The method is based on a Hidden Markov Model (HMM) and predicts correctly the entire topology for around 80% of the sequences in a standard dataset.

HMM's were initially used in speech recognition but have been successfully used in computational biology, e.g. to model the statistical structure of genomes [G.A92] and as a way to statistically describe a multiple sequence alignment [KKCea97]. The model is probabilistic and consists of a number of interconnecting states. In the case of TMHMM each state corresponds to a region or position in the protein being modeled. In the simplest case a model for a transmembrane sequences consists of three states: one for inside loops, one for transmembrane regions and one for outside loops. Each state has an associated probability distribution over the 20 amino acids.

The states are connected to each other in a way that corresponds to actual situation in the cell. The state for inside loop is connected to itself and to the transmembrane helix state. The fact that inside loop state is connected to itself reflects that the loop can be longer than one residue. The inside loop state is also connected to the transmembrane state since after the inside loop begins the transmembrane helix. In this way the sates form a cyclic network which describes the architecture of a transmembrane protein. The aim is to model the biological situation in the cell as closely as possible. The path of a protein sequence through

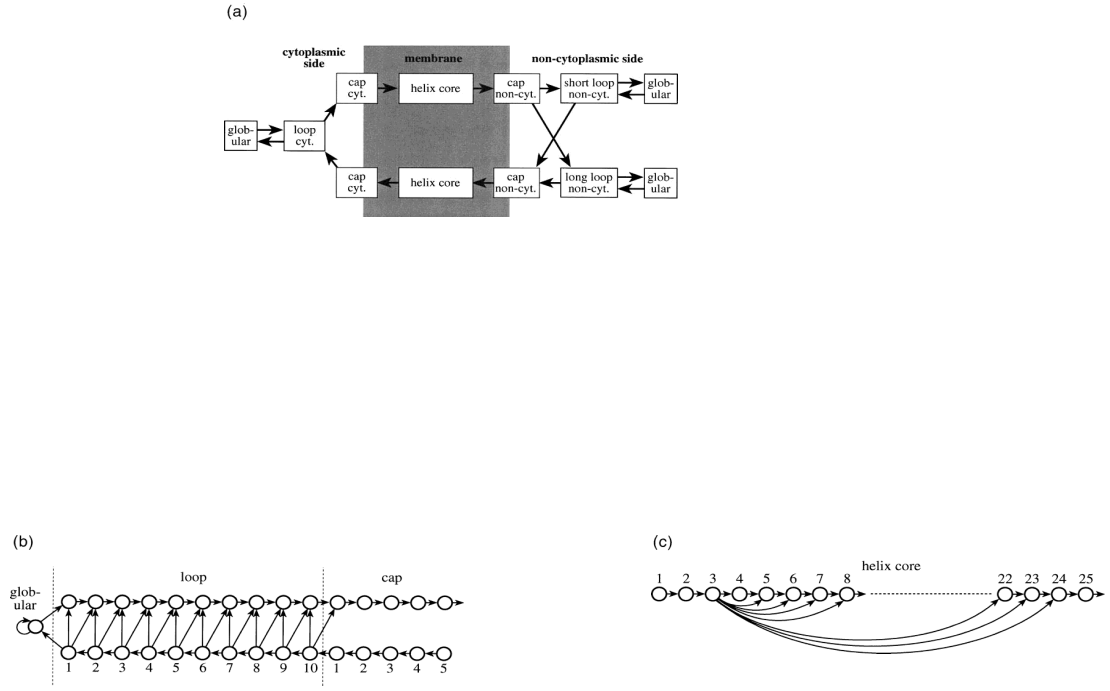

Figure 3. The layout of the hidden Markov model used in TMHMM. (a) The overall layout. Each box corresponds to one or more states in the HMM. Cyt. represents the cytoplasmic side of the membrane and non-cyt the other side. (b) The detailed structure of the inside and outside loop models and helix cap models. (c) The structure of the model for the helix core modeling lengths between five and 25, which translates to helices between 15 and 35 when the caps are included. Picture from [KLvHS01]

the states with the highest probability is likely to describe the actual topology of the protein.

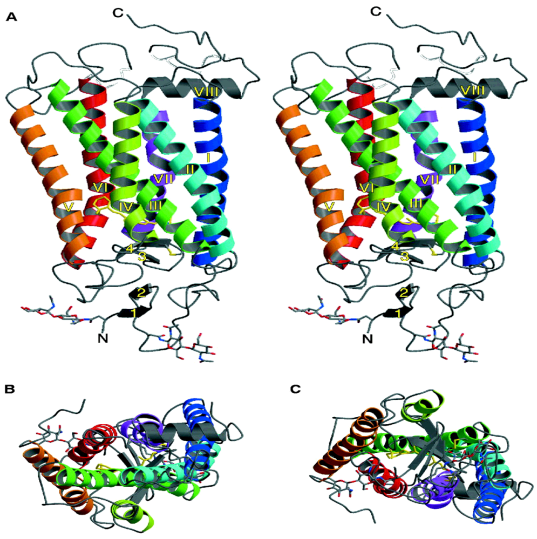

Figure 4. Ribbon drawings of rhodopsin. (A) Parallel to the plane of the membrane (stereo view). A view into the membrane plane is seen from the cytoplasm ic (B) and intradiscal side (C) of the membrane. Picture from [KTT+00]

2.3.3 GPCRDB

G protein coupled receptors (GPCRs) form a large superfamiliy of proteins that transduce signals across the cell membrane. At the external side of the membrane a signal is created by the binding of a photon or ligand. The signal is transmitted through the membrane and mediates a G-protein interaction at the cystolic side. The signal pathway in the cell is continued through the activation of a second messenger. GPCRs are involved in many fundamental cell response reactions and cell-cell signaling processes and play a crucial role in medical science and pharmaceutical industry. More than 50 % of existing medicines act on a GPCR [J96].

The structure of bovine Rhodopsin is known through diffraction data [KTT+00]. Rhodopsin is a GPCR which is activated by light and induces a chain of reactions that leads to vision. The structure of rhodopsin shows seven transmembrane helices connected by six loops of varying length. Rhodopsin provides information of the general structure of GPCRs, which all share the same overall structure, see figure 4.

The GPCRDB [FJM+98], aims at collecting as much GPCR related data as possible and making the information publicly available. The data comprise sequences and multiple sequence alignments but also other information such as ligands and ligand binding data, 3D models, literature references etc. The se-

quences are ordered in classes and sub classes based on their sequence similarity and functionality.

3 Methods

3.1 Test set

One of the most fundamental issues when evaluating the performance of methods for detecting relationships between sequences is to have a good way of stating which hits are correct and which hits are false. Several studies [HE99, LE99] use the Scop [ASTC95] classification to create benchmarks used for evaluating the performance of different recognition methods. The Scop database is hierarchically ordered. Each protein domain is classified into a family where the sequences have a clear evolutionary relationship. Families are ordered into superfamilies that is a sub classification of the fold category. Proteins in the same superfamily are of probable evolutionary origin, while the fold level is characterized by major structural similarity. Scop is from several aspects the ideal choice for evaluating search and prediction methods. The manual classification makes it independent of any specific sequence or structure comparison method and it is generally considered to be of very high quality.

In the case of membrane proteins there is no correspondence to the Scop database, simply because there are not enough structure information available.

In this study we use a selected set of GPCRs from GPCRDB[1] (December 2000 release) to create a test set of membrane sequences used in evaluating the performance of the alignment methods tested. A similar benchmark has been used in a recent study of Rehmsmeier [TSR01].

The sequences from the five major classes in GPCRDB were downloaded. The set was reduced so that no two sequences had more than 50 % sequence identity according to FastA. In order to get a more even distribution between the classes 50 sequences from class A were randomly selected and all other sequences in class A were removed from the set. The final test set contained 100 sequences from five families, see table 3.1.

3.2 Database

The tests were run against SWISS-PROT (release 39). In order to easily detect hits to related sequences all sequences with more than 95% sequence identity to any sequence in GPCRDB were excluded from SWISS-PROT and replaced by the sequences in GPCRDB before the comparisons were made. By doing this we

[1] GPCRDB is available at http://www.gpcr.org/7tm

Class	GPCRDB	Test set
Class A rhodopsin like	1207	50
Class B secretin like	86	19
Class C metabotropic glutamate/phermone	62	18
Class D fungal phermone	16	12
Class E cAMP receptors	4	1

Table 1. The five major classes in GPCRDB and the number of sequences in each class in the entire database and the test set respectively

assume that all sequences in GPCRDB are GPCRs and that all sequences with less than 95% sequence identity to any of these sequences are not GPCRs. This may not be entirely correct but the errors will effect the different methods equally and we believe that it is the best we can do.

3.3 Brief outline of the prediction based algorithm

Here we explore the ideas of prediction based threading in the alignment of membrane sequences. In contrast to prediction based threading for fold recognition [FE96], which match the predicted secondary structure of the query against the known structure of the template, we match a prediction against a prediction. Further we use a transmembrane specific predictor, TMHMM [KLvHS01], for the secondary structure information.

TMHMM is used to predict the location and orientation of possible transmembrane alpha helices in the query and database sequences. Each residue becomes associated with one of three topology states, *inside*, *membrane* or *outside*, according to the TMHMM prediction. The topology state is used as well as the residue type in the subsequent alignment procedure. If the residues in an aligned pair represent a topology match, e.g. *membrane* - *membrane* or *inside* - *inside*, an additional score is added to the substitution score. The value of the additional score for the different kinds of matches determines the influence of the predicted information on the total alignment score.

3.4 Single sequence searches

The first tests were done as single sequences searches. First the sequences in the test set were compared against the database using an ordinary Smith-Waterman type algorithm (SW), where BLOSUM 62 was used as scoring matrix and a gap opening penalty of -10 and a gap extension penalty of -4.

Further predicted secondary structure information was included in the alignment algorithm (SW&PRED). The membrane topology was predicted with TMHMM for the query and database sequences and an additional score of one was added for *membrane* - *membrane* matches.

Due to computer and time limitations no extensive search for the optimal weights for secondary structure matches and/or gap penalties was possible.

3.5 Multiple sequence searches

It is known that information from multiple sequence alignments can be used to increase the sensitivity in database searching. In addition to the single sequence searches described above we evaluated the performance when profile information was included in the alignments (PSI&PRED). The profiles were created by running PSI-BLAST against the database for each sequence in the test set. PSI-BLAST was run for five iterations with an E-value threshold for inclusion of new sequences in each iteration of 1e-15. An ordinary sequence-profile alignment algorithm was used and an additional score of one was added for *membrane* - *membrane* matches. The gap penalties were the same as for the single sequence searches.

3.6 PSI-BLAST

PSI-BLAST is one of the most sensitive database search methods available. In order to evaluate the performance of prediction based alignments compared to the performance of currently available search methods we ran PSI-BLAST for the same set of sequences. We have noticed that the E-value threshold for inclusion of new sequence in each iteration is of great importance for the performance of PSI-BLAST for membrane sequences. We evaluated the performance of PSI-BLAST for three different threshold values, 1e-3, 1e-5 and 1e-15. No filtering of the query sequence was used and a maximum of five iterations were allowed in each search.

3.7 Score normalization

For the PSI-BLAST searches the E-value was used for scoring. In the SW type algorithms the raw alignment score S was normalized by the length m and n of the compared sequences, $\frac{S}{log(m \times n)}$.

3.8 Measure the performance

The comparisons were ordered in decreasing normalized score order and analyzed on two levels of similarity. In the first analysis we considered the ability to recog-

	Class test	GPCR test
ClassA - ClassA	correct	ignored
ClassA - GPCR	ignored	correct
ClassA - !GPCR	false	false

Table 2. Summary of how correct and false matches are stated in the two tests performed in this study. ClassA - ClassA represents a match between sequences in the same class, ClassA - GPCR a match between GPCRs in different classes and ClassA - !GPCR represents a match between a GPCR and a non GPCR.

nize GPCRs in the same class (Class test). All matches between sequences in the same class (according to the classification in GPCRDB) were considered correct while matches between GPCRs in different families were ignored. Secondly we evaluated the ability to recognize more distantly related sequences (GPCR test). In this case all hits to GPCRs outside the own class were considered correct while hits to sequences in the same class were ignored, see table 2.

The results from the various searches were analyzed by calculating spec-sens plots and Mathews correlation coefficient (MCC) as measurements of the performance.

The spec-sens plots describe the fraction possible correct hits found as a function of the fraction found hits being correct. The main advantage is that it measures the ability of a method to reliable find all pairwise matches in the databases. The fraction possible correct hits found, the sensitivity, is defined as:

$$sens(S) = \frac{P_t(S)}{P_t(S) + N_f(S)}$$

where $P_t(S)$ is the number of true positives, i.e. the number of correct hits having a score above S and $N_f(S)$ being the number of false negatives, i.e. the number of correct hits with a score less than S. The specificity measures the probability that a pair of sequences with a score greater than a S really is a true hit, defined as:

$$spec(S) = \frac{P_t(S)}{P_t(S) + P_f(S)}$$

where $P_f(S)$ is the number of false positives, i.e. the number of false hits that have a score above S and $P_t(S)$ is defined as above. The sensitivity is plotted as a function of specificity, each point corresponding to a certain score S.

Matthews correlation coefficient is a related way of measuring the performance, defined as:

$$MCC(S) = \frac{P_t(S)N_T(S) - P_f(S)N_f(S)}{\sqrt{(N_t(S) + N_f(S))(N_t(S) + P_f(S))(P_t(S) + N_f(S))(P_t(S) + P_f(S))}}$$

Method	Description
SW	Smith-Waterman, BLOSUM 62
SW&PRED	Smith-Waterman BLOSUM 62 & secondary structure predictions
PSI-3	PSI-BLAST, 5 iterations, threshold 1e-3
PSI-5	PSI-BLAST, 5 iterations, threshold 1e-5
PSI-15	PSI-BLAST, 5 iterations, threshold 1e-15
PSI&PRED	Profile alignments PSI-15 profiles & secondary structure predictions

Table 3. Summary of the alignment methods in the study

where P_t is the number of true positives, N_t the number of true negatives, N_f the number of false negatives and P_f the number of false positives. A MCC value of 1 indicates that there is a score S such that all matches with a score higher than S are correct matches and all hits with a score below S are false matches. A MCC value of 0, on the other hand, indicates a very low information content and only random ability to discriminate between false and correct matches.

4 Result and Discussion

In this study we have examined six alignments methods and evaluated their ability to reliable recognize homologous membrane sequences. The methods are presented in table 3. Two methods, SW and SW&PRED, are single sequence search methods where SW is ordinary Smith-Waterman alignments and SW&PRED is Smith-Waterman alignments with additional secondary structure information. The other methods are multiple sequences search methods, where PSI-3, PSI-5 and PSI-15 are PSI-BLAST searches and PSI&PRED is sequence-profile alignments with additional secondary structure information. The various PSI-BLAST searches are labeled according to the threshold values used, e.g. PSI-3 was run with a threshold value of 1e-3 for inclusion of new sequences in each iteration.

It is known that relationships between proteins span a very broad range, from almost identical sequences to apparently unrelated sequences sharing only rough 3D fold. Finding homologous sequences on the various levels of similarity poses different problems to the search algorithms. Methods that perform well at finding

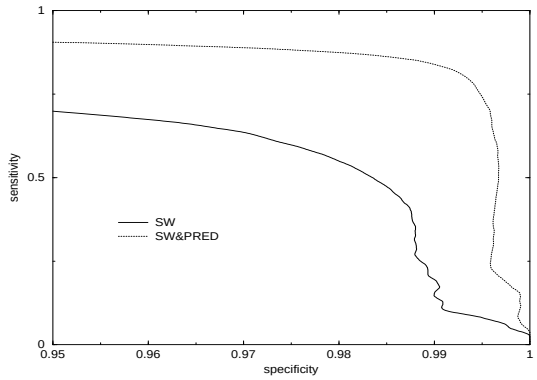

Figure 5. Comparison of Smith-Waterman (SW) and the prediction based algorithm (SW&PRED) for the ability to detect GPCRs in the same class

sequence similarities may be of limited usefulness in the case of finding sequences that have similar fold but low degree of sequence similarity.

In this study we have evaluated the performance of the various methods on two levels of similarity. Fist we considered the ability to recognize sequences in the same class of GPCRs and secondly we considered the ability to recognize GPCRs outside the own class. In the first case matches to GPCRs outside the own class were ignored and in the second case matches to sequences in the same class were ignored. By separating the two levels of similarity we believe that we get a clearer picture of the performance. If hits to sequences in the same class would not have been ignored in the second case these hits would dominate the result and make it hard to detect the differences in performance at the more distant level of similarity.

4.1 Secondary structure information improves alignments of membrane sequences

The result of the comparison of the sequences in the test set against the database for prediction based alignments is compared to the performance of ordinary SW alignments in figure 5. It is clear that the inclusion of additional secondary structure information significantly improves the result. At all specificities the prediction based method has higher sensitivity. The improvement is also clear by comparing the MCC values which increased from 0.83 to 0.93, see table 4.

Method	MCC-class	MCC-GPCR
SW	0.83	0.01
SW-PRED	0.93	0.24
PSI-3	0.95	0.60
PSI-5	0.96	0.46
PSI-15	0.98	0.16
PSI-PRED	0.98	0.22

Table 4. Summary of the best obtained Matthews correlation coefficients (MCC) at class and GPCR level respectively

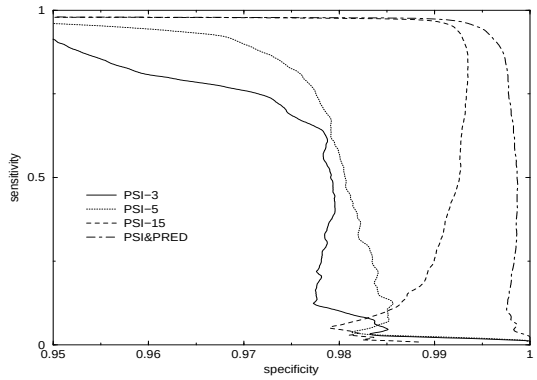

Figure 6. Comparison of PSI-BLAST for various threshold values (PSI-xx) and prediction based profile search (PSI&PRED) for the ability to detect GPCRs i n the same class

4.2 Performance of PSI-BLAST

PSI-BLAST is one of the most sensitive sequence based database search methods available. Due to the iterative approach and the position specific profiles PSI-BLAST is able to detect more distantly related sequences. The statistics used in BLAST, however, are based on information derived from globular sequences. Since unrelated membrane sequences have higher degree of sequence similarity than unrelated non membrane sequences a BLAST alignment between membrane sequences may have a significant E-value even if the sequences are unrelated.

In PSI-BLAST all matches above a certain E-value threshold are incorporated in the profile which is used in the next iteration. The easiest way to avoid false hits to be incorporated is to use a more strict threshold value when PSI-BLAST is used to search for membrane sequences.

The performance of PSI-BLAST for three different threshold values is shown in figure 6. The default threshold value, 1e-3, results in a lower specificity compared to more strict threshold values. Our assumption that a more strict threshold value improves the result when PSI-BLAST is used for membrane sequences seems to be correct.

It is obvious from figure 6 that even at a very strict E-value threshold PSI-BLAST results in a number of very high scoring false hits (the specificity in figure 6 never reaches 1). High scoring false hits is a common problem when PSI-BLAST is used without manual checking of the sequences incorporated in the profiles. Once a false hit is incorporated every subsequent iteration will result in more false hits and the final result will be biased by a high number of high scoring incorrect matches.

Among the top 100 false hits for PSI-15 57 hits have no transmembrane regions according to TMHMM, 35 have one and two of the most high scoring false hits have eight transmembrane regions. It seems as if PSI-BLAST might find false hits to membrane as well as to globular proteins and that we did not suffer from miss-classified GPCRs in the database.

The left column in table 4 shows a summary of the performance of the different methods at family level. The addition of secondary structure information (SW&PRED) significantly improves the performance compared to ordinary SW alignments but PSI-BLAST still is the best method, especially at strict threshold values.

The fact that PSI-BLAST profiles as well as predicted secondary structure information improve the search results suggests that the combination of profiles and predicted secondary structure information would improve the performance even further.

The multiple search method PSI&PRED in figure 6 explores profile as well as secondary structure information. PSI-BLAST was run for five iterations with an

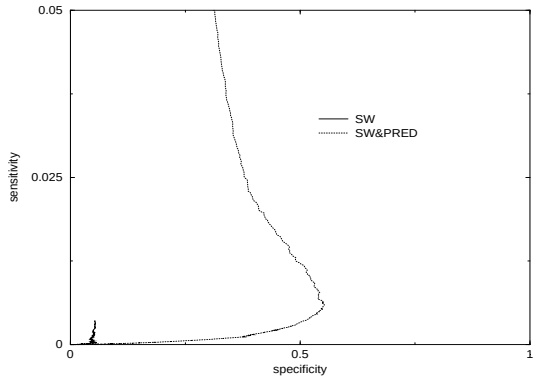

Figure 7. Comparison of Smith-Waterman (SW) and the prediction based alignment method (SW&PRED) for the ability to detect GPCRs outside the own class

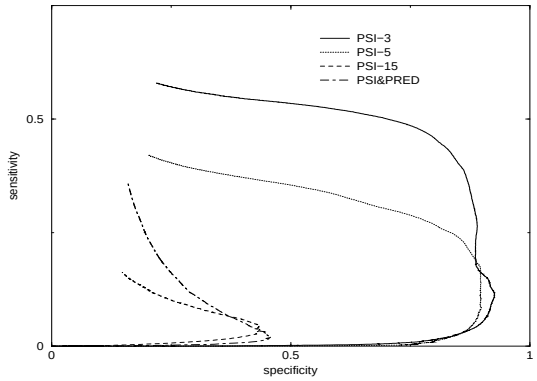

Figure 8. Comparison of PSI-BLAST for various threshold values (PSI-xx) and the prediction based profile alignments (PSI&PRED) for the ability to detect GPCRs outside the own class

E-value threshold of 1e-15 and the profile used in the last iteration was used in PSI&PRED. The inclusion of predicted secondary structure information into a profile search increases the specificity compared to PSI-15 but at an error rate of more than 1 % no improvment over PSI-15 can be seen. The result indicates that the inclusion of predicted membrane regions into profiles reduces the problem of high scoring false matches but the sensitivity still seems to be limited by the profile.

4.3 Ability to find more distantly related sequences

Single sequence alignments in general are weak methods to identify relationships on distant levels of similarity. Figure 7 shows the result of SW alignments compared to prediction based alignments for the ability to recognize sequences outside the own class of GPCRs. It is clear that the introduction of secondary structure information significantly increases the ability to find sequences outside the own class. While ordinary SW alignments never reaches more than 10 % specificity SW&PRED finds about 1% of the correct matches at 50% specificity. By scoring matches between membrane regions it is possible to detect relationships based on similarities in secondary structure even if sequence similarity is very low.

In contrast to single sequence alignments, the main advantage of PSI-BLAST is the ability to find distantly related sequences. PSI-BLAST with a less restrictive threshold value performs significantly better than any other method figure 8, figure 7 and table 4 right column. PSI-3 detects about 50% of the non class related GPCRs at 80% specificity.

5 Conclusions and future steps

In this study we present a prediction based alignment method for membrane sequences. The approach is based on a similar technique that has previously been explored in fold recognition studies for globular proteins [FE96]. TMHMM is used to predict presence and location of transmembrane helices in the sequences being aligned. The alignment is based on a Smith-Waterman type algorithm and uses an extra score in addition to the substitution score. Aligned residues that are predicted to be located in transmembrane helices are given an additional positive score.

We show that the additional secondary structure information significantly improves alignments of membrane sequences. For single sequence searches this method perform significantly better than standard Smith-Waterman alignments (figure 5). The performance can probably be improved further by optimizing the weighting for topology matches. Here we have scored *membrane* - *membrane* matches without considering the direction of the helices. It would be interesting to study the effect of scoring *inside* - *inside* and *outside* - *outside* matches as well. Other parameters such as the gap penalties can probably also be optimized further.

Further we show that the combination of predicted topology and profile information increases the specificity compared to PSI-BLAST (figure 6). We used profiles created in the last iteration of PSI-BLAST with a threshold value of 1e-15 for inclusion of new sequences in each iteration. Since a less strict threshold

value makes it possible to find many more sequences, especially on distant levels of similarity, it would be interesting to evaluate the effect of topology information in combination with profiles created with a threshold of 1e-5 or 1e-3. The result presented in figure 6 suggests that it is possible to reach a high sensitivity by the use of profiles and a high specificity by the use of predicted topologies.

It should be noticed that we used the profiles created by PSI-BLAST, i.e. no topology information were used in the creation of the profiles. To include secondary structure information in PSI-BLAST is a non-trivial task but based on the results presented here it seems to be a possibility for the future to drastically improve detection of homolugous membrane proteins.

Aknowledgements

I would like to thank my supervisor Arne Elofsson at Stockholm Bioinformatics Center for help and guidance throughout the project. I would also like to thank Gunnar von Heijne and Karin Melén for assistance and valuable ideas during the work.

References

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugen W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol*, 215:403–410, 1990.

[AMS+97] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[ASTC95] Murzin AG, Breener SE, Hubbard T, and Chothia C. A structure classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, 1995.

[BRPC95] Rost B, Casadio R, Fariselli P, and Sander C. Transmembrane helices predicted at 95% accuracy. *Protein Sci*, 4(3):521–533, 1995.

[DDJD96] Fischer D, Rice DW, Bowie JU, and Eisenberg D. Assigning amino acid sequences to 3d protein folds. *FASEB J*, 10:126–136, 1996.

[DW85] Lipman D.J. and Pearson W.R. Rapid and sensitive protein similarity searches. *Scince*, 227:1435–1441, 1985.

[FE96] Daniel Fischer and David Eisenberg. Protein fold recognition using sequence-derived prediction. *Protein Science*, 5:947–955, 1996.

[FJM+98] Horn F., Weare J, Beukers M.W., Hörsch S., Bairoch A., Chen W., Edvardsen Ø, Champagne F., and Vriend G. Gpcrdb: an information system for g protein-coupled receptors. *Nucleic Acids Research*, 26:275–279, 1998.

[G.A92] Churchill G.A. Hidden markov chain and the analysis of genome structure. *Computers and Chemistry*, 16(2):107–115, 1992.

[HE99] Jeanette Hargbo and Arne Elofsson. Hidden markov models that use predicted secondary structures for fold recognition. *Proteins: Structure Function, and Genetics*, 36:68-76, 1999.

[J96] Drews J. Genomic sciences and the medicine of tomorrow. *Nature Biotechnol*, 14:1516–1518, 1996.

[JRD91] Bowie JU, Luthy R, and Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[JS96] Wootton JC and Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*, 266:513–525, 1996.

[KKCea97] Karplus K, Sjölander K, Barrett C, and et. al. Predicting structures using hidden markov models. *Proteins Suppl*, 1:134–139, 1997.

[KLvHS01] Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *J. Mol. Biol*, 305:567–580, 2001.

[KTT+00] Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, and Miyano M. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science*, 289(5480):739–745, 2000.

[LE99] Erik Lindahl and Arne Elofsson. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol*, 295:613–625, 1999.

[MD93] Wilmanns M and Eisenberg D. Three-dimensional profiles from residue-pair preferences: Identification of sequences with beta/alpha-barrel fold. *Proc Natl Acad Sci USA*, 90:1379–1383, 1993.

[MEI+97] Cserzo M., Wallin E., Simon I, von Heijne G., and Elofsson A. Prediction of transmembrane alpha-helices ni procaryotic membrane proteins: the dense alignment surface method. *Protein Eng*, 10(6):673–676, 1997.

[MRB78] Dayhoff M.O., Schwartz R.M., and Orcutt B.C. Matrices for detecting distant relationship. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[MS92] Sippl MJ and Weitckus S. Detection of native like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins Struct Funct Genet*, 13:258–271, 1992.

[NHH00] Pauline C. Ng, Jorja G. Henikoff, and Steven Henikoff. Phat: a transmembrane-specific substitution matrix. *Bioinformatics*, 16(9):760–766, 2000.

[RC95] Grisshammer R and Tate CG. Overexpression of integral membrane proteins for structural studies. *Q Rev Biophys*, Aug 28(3):315–422, 1995.

[RSS97] Burkhard Rost, Reinhard Schneider, and Chris Sander. Protein fold recognition by prediction-based threading. *J. Mol. Biol*, 270:471–480, 1997.

[SC70] Neddleman S.B. and Wunsch C.D. A general method apllicable to the search for similarities in the amino aicd sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[SJ92] Henikoff S. and Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science USA*, 89(22):10915–10919, 1992.

[TM81] Smith TF and Waterman MS. Identificaton of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[TRM92] Jones D. T., Taylor W. R., and Thornton J. M. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[TSR01] Müller T, Rahmann S, and Rehmsmeier. Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics*, 17 Suppl 1:S182–S189, 2001.

[vHG97] von Heijne G. Principles of membrane protein assembly and structure. *Progr.Biophys.Mol.Biol*, 66(2):113–139, 1997.

[vHG99] von Heijne G. Recent advances in the understanding of membrane protein assembly and structure. *Quart Rev.Biophys*, 32(4):285–307, 1999.