Sequence analysis

**PRALINE™: a strategy for improved multiple alignment of transmembrane proteins**

Walter Pirovano, K. Anton Feenstra and Jaap Heringa*

Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received on November 20, 2007; revised on December 20, 2007; accepted on December 21, 2007

Advance Access publication January 2, 2008

Associate Editor: Burkhard Rost

**ABSTRACT**

Motivation: Membrane-bound proteins are a special class of proteins. The regions that insert into the cell-membrane have a profoundly different hydrophobicity pattern compared with soluble proteins. Multiple alignment techniques use scoring schemes tailored for sequences of soluble proteins and are therefore in principle not optimal to align membrane-bound proteins.

Results: Transmembrane (TM) regions in protein sequences can be reliably recognized using state-of-the-art sequence prediction techniques. Furthermore, membrane-specific scoring matrices are available. We have developed a new alignment method, called PRALINE™, which integrates these two features to enhance multiple sequence alignment. We tested our algorithm on the TM alignment benchmark set by Bahr et al. (2001), and showed that the quality of TM alignments can be significantly improved compared with the quality produced by a standard multiple alignment technique. The results clearly indicate that the incorporation of these new elements into current state-of-the-art alignment methods is crucial for optimizing the alignment of TM proteins.

Availability: A webserver is available at http://www.ibi.vu.nl/programs/pralinewww.

Contact: heringa@cs.vu.nl

1 INTRODUCTION

Over the past years, integral membrane proteins have received a great deal of attention. They carry out essential functions in many cellular and physiological processes, such as signal transduction, cell-cell recognition and molecular transport. Membrane proteins are likely to constitute 20–30% of all ORFs contained in genomes (Jones, 1998; Wallin and von Heijne, 1998).

Unfortunately, the number of determined transmembrane (TM) structures in the PDB is still very low: <2% of all structures solved show a membrane topology (www.pdb.org; Tusnady et al., 2005). Despite of a solid exponential growth of the number of membrane protein structures (White, 2004), their determination remains a difficult task, such that they will continue to lag behind relative to the number of elucidated soluble protein structures.

Transmembrane (TM) regions show a modified hydrophobicity and conservation pattern as compared with soluble proteins. Conventional scoring matrices such as PAM (Dayhoff et al., 1978) or BLOSUM (Henikoff and Henikoff, 1992), routinely used for sequence retrieval and alignment, are therefore in principle not suitable to align membrane-bound protein regions. Jones et al. (1994) for instance noticed that polar residues are highly conserved in these regions, whereas hydrophobic residues are more interchangeable, and developed the JTT TM substitution matrix. Ng et al. derived a new TM-specific substitution matrix called PHAT, which was shown to outperform the JTT matrix, especially on database searching (Ng et al., 2000). Meanwhile several groups focused on the development of accurate membrane topology predictors such as HMMTOP (Tusnady and Simon, 1998, 2001), TMHMM (Krogh et al., 2001; Sonnhammer et al., 1998), Phobius (Kall et al., 2004, 2005) and MEMSAT (Jones, 2007; Jones et al., 1994a). The topic has recently been reviewed by Punta et al. (2007).

Not many techniques however have been developed to improve the alignment of TM proteins. The method STAM (Shafir and Guy, 2004) represents an early attempt to improve alignment accuracy by combining different substitution matrices. A more recent study by Forrest et al. (2006) reported that the use of a bipartite scheme (consisting of BLOSUM62 and PHAT) does not significantly improve membrane protein sequence alignments. They suggest that the previously reported progress is more likely to depend on the separation of the TM blocks or on the settings of specific gap penalties.

In this study we have investigated the effects of incorporating TM specific information into the previously developed multiple alignment tool PRALINE (Heringa, 1999, 2002). This information is integrated in a ‘soft’ way, compared with for instance, the STAM approach where TM segments are first chopped and then aligned separately. In our approach the choice of the matrix depends on consistent TM predictions over a column and is determined dynamically during the alignment procedure. We also explore an additional iterative strategy to further optimize the alignments.

We have tested the algorithm on the TM benchmark alignments of BALiBASE (Bahr et al., 2001). This reference set contains more than 400 reliably aligned TM sequences, divided into eight families. The alignments are manually

*To whom correspondence should be addressed.
curated and at the moment they constitute by far the largest available benchmark. By applying the PHAT substitution matrix on accurately predicted TM regions combined with a proper gap penalty setting, we show that we are able to significantly improve the alignment quality.

2 METHODS

The ‘basic’ and ‘global profile pre-processing’ (‘pre-profile’ or ‘prepro’) PRALINE progressive alignment algorithms, underlying the strategies tested in this study, are described in detail in previously published works (Heringa, 1999, 2002). In brief, the ‘basic’ PRALINE alignment method simply follows the classic progressive alignment protocol where sequences are aligned following the order of the guide tree. In the ‘pre-profile’ method for each sequence a so-called master-slave alignment is constructed, containing information about neighboring sequences, which are then used in subsequent progressive alignment. It has been shown that these sequence pre-profiles are more informative than single sequences and help to avoid mistakes during the progressive steps (Heringa, 2002). For the PRALINE™ tool we present here, we first predict for each input sequence its TM topology using a state-of-the-art predictor. Second, the profile-scoring scheme simply applies TM-specific substitution scores from the PHAT matrix to reliably predict TM positions. Finally, we incorporated an alternative iterative scheme to enhance the alignment quality. In Figure 1 an overview of the PRALINE™ strategy is given.

2.1 Scoring scheme

The current PRALINE profile-scoring scheme uses the following equation to score a pair of profile columns $x$ and $y$:

$$S(x, y) = \sum_{i=1}^{20} \sum_{j=1}^{20} \alpha_i \beta_j M(i,j)$$

where $\alpha_i$ and $\beta_j$ are the frequencies with which residues $i$ and $j$ appear in columns $x$ and $y$, respectively, and $M(i,j)$ is the exchange weight for residues $i$ and $j$ provided by the selected substitution matrix $M$. By default profile columns are aligned using the BLOSUM62 matrix. Two profile columns will be matched using the PHAT matrix only in case each residue in the column is predicted to be member of a TM segment. This is done to guarantee that inconsistently predicted positions do not negatively influence the alignment quality. As a result, and contrary to the STAM method (Shafirir and Guy, 2004), our approach potentially allows TM segments to be aligned to non-TM segments. The BLOSUM62 and PHAT substitution matrices are normalized using their diagonal elements as described in (Abagyan and Batalov, 1997).

2.2 Transmembrane topology predictors

Transmembrane topologies are predicted using three different state-of-the-art methods: HMMTOP v2.1 (Tusnady and Simon, 2001), TMHMM v2.0 (Krogh et al., 2001) and Phobius (Kall et al., 2004). All predictors are installed locally and run independently within the PRALINE™ program.

2.3 Multiple alignment benchmark

To evaluate the performance of the two standard PRALINE alignment techniques, the PRALINE™ method and other state-of-the-art multiple alignment programs, we used the BALiBASE (v2.0) reference alignment set of TM proteins (Bahr et al., 2001). The set includes eight accurately aligned TM families. The total number of sequences is 435 with an average length of 567 residues. The number of TM helices per sequence varies from 2 to 14.

The BALiBASE ‘testing’ program was used to evaluate alignment accuracy against the benchmark. Accuracy is measured with two alternative scores: the ‘SP score’ measures the fraction of correctly aligned residue pairs while the ‘TC score’ fraction of correctly aligned columns.

2.4 Alignment methods tested

First of all we tested the performance of the original PRALINE tools compared with the PRALINE™ application where two matrices are combined. It is commonly thought that gaps within TM regions should be more penalized than gaps in soluble regions. We therefore evaluated different combinations of gap penalties, to see whether the sensitivity of the approach resides in the gap penalty settings or the specific TM matrix. In addition we compared the results obtained to other multiple alignment routines, which are designed for standard alignment purposes. These include: ClustalW v1.83 (Thompson et al., 1994), MUSCLE v3.52 (Edgar, 2004a, b), MAFFT v6 (Katoh et al., 2005) and ProbCons v1.12 (Do et al., 2005) and all of these programs were run using default parameter settings.

2.5 Tree-based consistency iteration

We also employed the potential benefits of an additional iterative strategy. At the heart of it lies the tree-dependent consistency iteration, which is similar to the tree-dependent strategy proposed by Hirosawa et al. (1995) and its implementation in the MUSCLE method (Edgar, 2004). In this scenario each edge of the phylogenetic (guide) tree is used to divide the alignments into two subalignments, which are successively realigned. The new alignment is retained only if a higher Sum-of-Pairs score is achieved. In our case this score is obtained by summing the substitution values of both the BLOSUM62 and PHAT matrix (depending on the TM topology of the amino acid pair). For the tree-based consistency strategy one iterative cycle means that each edge of the tree is visited once. The maximum number of iterations is set to 20.

3 RESULTS AND DISCUSSION

3.1 Performance of the PRALINE™ methods compared with the standard PRALINE methods

First of all we sought to understand whether a general improvement of alignment quality could be observed when including the TM-specific information. We therefore extensively tested both ‘PRALINE basic’ and ‘PRALINE prepro’ using the three selected TM topology predictors (see Methods Section) and gap—open penalties ranging from 12 to 18 (in steps of 1) for both the soluble and the TM regions. An additional parameter, the pre-profile cut-off, was varied from...
8.0 to 15.0 (in steps of 0.5) following the global pre-processing conditions defined in Heringa (2002). This parameter indicates to what extent other neighboring sequences are included into the sequence pre-profiles. The results of the ‘PRALINE basic’ strategies are summarized in Table 1, the results of the ‘PRALINE prepro’ strategies in Figures 2 and 3.

The most striking observation to be made from both Table 1 and Figure 2a and b is the positive effect on the alignment quality of the PHAT matrix applied on reliably predicted TM regions. Here the results are shown at an arbitrary gap-open penalty of 15.0 and gap-extension penalty of 1.0 for both the soluble and the TM regions; the outcomes are consistent over all combinations of gap-open penalties.

A notable increase can be observed for all three TM predictors, albeit Phobius gives the best performance overall. Phobius has shown to be one of the most accurate TM topology predictors, especially on sequences that also contain a signal peptide (Jones, 2007; Kall et al., 2004).

Concerning the pre-profile cut-off it can be noticed from Figure 2a and b that the optimal parameter settings lie between 11.0 and 12.0. In this range the highest SP and TC scores are reached and also maximum improvement relative to the standard pre-profile technique is attained. Consistency of these scores was estimated by 8-fold cross-validation, each time leaving one BAliBASE alignment out and retaining the other seven (data not shown). Standard deviation over the cross-validated SP scores was below 4%, and minimal SD of around 1.5% were reached at highest SP score. For the TC scores, SD were between 3 and 4.5%; here the differences are small enough that the optimal choice also lies at the highest TC scores.

Finally, Figure 3 shows the performance of the ‘PRALINETM prepro – Phobius’ method at different gap-open penalties: for the soluble and TM regions different penalty combinations were tested. An optimal gap-open penalty for the soluble regions is hard to define, though the optimum for TM regions lies between 15.0 and 18.0. We also varied both the soluble and TM gap-extension penalties from 1.0 to 1.5, but no significant differences were observed.

### Table 1. Performance of the PRALINE and PRALINE\textsuperscript{TM} basic strategies on reference set 7 of BAliBASE (at gap-open and gap-extension penalties of 15.0 and 1.0 for both the soluble and the transmembrane regions)

<table>
<thead>
<tr>
<th>Method</th>
<th>SP score</th>
<th>TC score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRALINE basic</td>
<td>0.646</td>
<td>0.231</td>
</tr>
<tr>
<td>PRALINE\textsuperscript{TM} basic – HMMTOP</td>
<td>0.679</td>
<td>0.264</td>
</tr>
<tr>
<td>PRALINE\textsuperscript{TM} basic – TMHMM</td>
<td>0.725</td>
<td>0.254</td>
</tr>
<tr>
<td>PRALINE\textsuperscript{TM} basic – Phobius</td>
<td>0.737</td>
<td>0.268</td>
</tr>
</tbody>
</table>

3.2 Iterative strategies further improve the PRALINE\textsuperscript{TM} method

Next, we studied the effects of the additional tree-based consistency iterative strategy (described in Section 2.5) on the PRALINE\textsuperscript{TM} method. Following our above results we now varied the pre-profile cut-off between 11.0 and 12.0, the gap-open penalties from 12.0 to 18.0 for soluble regions (in steps of 1.0) and from 15.0 to 18.0 for TM regions (in steps of 0.5).

The optimal parameter setting found at a pre-profile cut-off of 11.0 combined with a gap-open penalty combination of 15.0 for the soluble regions and 16.5 for the TM regions. The above mentioned parameter settings define the final PRALINE\textsuperscript{TM} method.

3.3 Contributions of substitution matrices and gap penalty settings on the alignment quality

We also investigated independent contributions to the alignment quality coming from the PHAT matrix and specific gap penalties. For this purpose, we tested PRALINE\textsuperscript{TM}, defined in the previous section, at different gap-open penalties for the TM
regions. In the first run the TM segments were aligned using the PHAT matrix. In the other two runs we used either the standard BLOSUM62 matrix or the PHAT matrix for the entire sequence. The results in Figure 4 clearly show that only the combination of BLOSUM62/PHAT matrices yield optimal results. The runs in which only one matrix is applied to the whole sequence, even when optimized gap penalties are used, produce much less reliable alignments. On the other hand, we noticed that applying a slightly higher gap-open penalty to the TM regions relative to that for soluble regions can have some additional benefits. These influences however are much less pronounced, implying that the results obtained are not very sensitive to the TM gap-open settings.

We further tested the BLOSUM55 matrix, since it has a comparable entropy to that of the PHAT matrix ($H = 0.5637$ and $H = 0.5605$, respectively). We optimized the alignment parameters for the BLOSUM55/PHAT combination as in Section 3.2 and observed a decreased performance of 5 percentage points on average. A likely explanation of this effect can be that the evolutionary scenarios underlying TM and soluble regions are different.

### 3.4 Comparison with other alignment methods

Finally we compared our algorithm with widely-used multiple alignment methods, which are designed for aligning soluble proteins. Results are shown in Table 2. The standard PRALINE (i.e. ‘prepro’ without TM information) method with optimized parameter settings over this dataset is included for reference (at a gap-open penalty of 15.0 the optimal pre-profile cut-off is 8.5, see Fig. 2). Notably, all methods reach SP scores that are twice as high as corresponding TC scores. The latter score is a much stricter measure, but arguably also more meaningful since evolutionary analysis is usually performed on whole alignment columns. We see that PRALINE achieves the highest SP score for two datasets and the highest TC score for four datasets. Concerning the averages over all eight datasets, ProbCons slightly outperforms MAFFT (−0.6 percentage points) and PRALINE (−0.5 percentage points) on the SP score. On the more critical TC score PRALINE clearly scores best (+1.5 and +5.1 percentage points compared with ProbCons and MAFFT, respectively). ClustalW and MUSCLE score considerably lower on almost all datasets. The standard PRALINE method achieves a SP score comparable to ClustalW, but can be placed between MAFFT and ProbCons with respect to the TC score. The inclusion of TM information in PRALINE yields +8.0 percentage points for SP scoring and +4.0 percentage points for TC scoring compared with standard PRALINE.

It should be stressed that the PRALINE and PRALINE methods were optimized on the TM dataset, whereas the other methods were run at default settings. Concerning this, both MAFFT and ProbCons are relatively robust on TM sequences. Nonetheless, the results show clearly that our TM-based strategy can significantly improve the quality of TM protein sequence alignments, and should be considered a promising avenue for other applications as well.

### 4 CONCLUSIONS

We present a new strategy designed to accurately align protein families adopting a TM topology. We conclude that the alignment quality can be improved significantly using a
TM-specific substitution matrix and proper gap penalty settings. In our view the improvement is mainly attributed by the fact that the bipartite scheme, using BLOSUM62 and PHAT, is applied in a flexible manner to undivided sequences during each step of the alignment procedure. To the best of our knowledge, the magnitude of the success accomplished has not been reported elsewhere to date. Other attempts where TM and soluble regions were aligned independently did not succeed in making significantly better alignments (Forrest et al., 2006).

In fact, in those approaches the definition of the TM segment is of crucial importance as TM segments cannot be aligned with non-TM segments, such that incorrectly delineated TM regions are likely to lead to misaligned TM and soluble segments. Even bigger problems arise when the number of TM segments varies within families. PRALINE\textsuperscript{TM} aligns undivided sequences instead and applies substitution scores from the PHAT matrix only where predictions are 100% consistent. The flexibility of the algorithm allows TM segments to be aligned with non-TM segments if other signals prevail over the TM signals.

It should be stressed that the choice of the prediction method can play an important role. Although in this article we did not explicitly test the quality of the different methods, their specific algorithms certainly affect the alignment outcomes. In general, all three methods tested enhanced the alignment accuracy, while Phobius emerged as the most valuable tool for TM alignment. Phobius is considered one of the most accurate TM topology predictors and its main advantage resides in the ability to discriminate between TM segments and signal peptides (Jones, 2007; Kall et al., 2004). The fact that most BALiBASE reference alignments contain predicted signal peptides explains at least partly the leading role of Phobius when used in our alignment strategy.

It is reassuring that the alignment quality we obtain with PRALINE\textsuperscript{TM} is correlated with the prediction quality of the TM prediction methods reported in the literature for TM sequences containing signal peptides (Jones, 2007; Kall et al., 2004). Further improvements could come from incorporating prediction confidence levels, or combining the TM topology predictions in a single consensus prediction.

Concerning gap penalties we noticed that strict gap penalty settings for TM regions improve the overall performance. However, these effects should not be overestimated: we found that the optimal gap-open penalty applied to the TM segments was only about 10% higher than the standard gap-open penalty applied to soluble regions.

None of the methods included here was able to align more than 40% of the reference alignment columns on average, so that further optimization remains a challenging task. Nonetheless this research has shed some new light on the alignment of TM protein families and shows that TM-awareness is an important concept for optimizing multiple sequence alignment quality.

**ACKNOWLEDGEMENTS**

The authors wish to thank Thomas W. Binsl for invaluable discussions and help with the implementation. We are also grateful to Gabor E. Tusnády, Anders Krogh and Lukas Käll for providing the source code of the transmembrane topology predictors and to Dmitrij Frishman and Andrei A. Mironov for early discussions. Financial support was provided by the
Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 3.2.2).

Conflict of Interest: none declared.

REFERENCES


Kall,L. et al. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics, 21(Suppl. 1), i251–i257.


