An Overview of Multiple Sequence Alignment Methods Applied to Transmembrane Proteins

Cristian Zambrano-Vega[1(✉)], Byron Oviedo[1], Ronald Villamar-Torres[2], Miguel Botto-Tobar[3,4], and Marcos Barros-Rodríguez[5]

[1] Facultad de Ciencias de la Ingeniería, Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador
{czambrano,boviedo}@uteq.edu.ec
[2] Université de Montpellier, Montpellier, France
villamartorresronaldoswaldo@yahoo.es
[3] Eindhoven University of Technology, Eindhoven, The Netherlands
m.a.botto.tobar@tue.nl
[4] Universidad de Guayaquil, Guayaquil, Ecuador
miguel.bottot@ug.edu.ec
[5] Universidad Técnica de Ambato, Ambato, Ecuador
ma-barros@uta.edu.ec

Abstract. Transmembrane proteins (TMPs) have received a great deal of attention playing a fundamental role in cell biology and are considered to constitute around 30% of proteins at genomic scale. Multiple Sequence Alignment (MSA) problem has been studied for some years and researchers have proposed many heuristic and stochastic techniques tailored for sequences of soluble proteins, considering that there are a few particular differences that ought to be taken into consideration aligning TMPs sequences, these techniques are therefore not optimal to align this special class of proteins. There is a small number of MSA methods applied specifically to TMPs. In this review, we have summarized the features, implementations and performance results of three MSA methods applied to TMPs: PRALINE^{TM}, TM-Coffee and TM-Aligner. These methods have illustrated impressive advances in the accuracy and computational efforts aligning TMPs sequences.

Keywords: Multiple Sequence Alignment · Transmembrane proteins
Computational bilogy

1 Introduction

Given the biomedical of TMPs and the crevice between the number of illuminated TMPs structures and the number of TMPs groupings, arrangement examination methods are significant. Over the past years, Transmembrane Proteins (TMPs) or Integral Proteins have taken a extraordinary deal of consideration

M. Botto-Tobar et al. (Eds.): CITT 2018, CCIS 895, pp. 410–419, 2019.
https://doi.org/10.1007/978-3-030-05532-5_30

playing a fundamental role in cell biology and are among the foremost tended to targets of pharmaceutical and life science research. TMPs are non-soluble proteins secured in a cell membrane and containing one or more membrane-spanning sections isolated with intra or extra-cellular domains of variable length [3]. They carry out fundamental capacities in numerous cellular and physiological processes, such as cell-cell recognition, molecular transport and signal transduction. Around 30% of proteins encoded by the mammalian genome are transmembrane proteins [2,28]. TMPs are difficult to study [23] and are well known for their complexities in deciding their structures experimentally. Only 3227 (α: 2848, β: 366) TMP structures are available till date with Protein Data Bank of TMP with the version 2017.06.16 [20]. Given the biomedical significance of TMPs and the huge and developing gap between the solved TMPs structures and the TMPs sequences, sequence analysis methods are very significant.

The special environment of a layer protein compared to a water-soluble protein leads to particular natural weights on their groupings: it is transcendently lipophilic, needs hydrogen-bonding potential, and gives small screening of electrostatic interaction [10]. Be that as it may, to date, as it were a little number of MSA strategies have been proposed particularly for TMPs, or tried utilizing TMPs datasets

Multiple Sequence Alignment (MSA) is the process of aligning more than two biological sequences (Protein, DNA or RNA), has many applications in field of computational biology: protein structure prediction, functional genomics, genomic annotation, gene regulation networks, or homology searches. Most current MSA procedures have been built, and tested, to align homologous soluble proteins. Indeed in spite of that numerous such procedures are still applicable to transmembrane regions, yielding a very lower alignment accuracy than for soluble proteins [11]. There are a few particular differences that ought to be taken into consideration, TM domains have an adjusted amino-acid composition and different conservation patterns as compared to soluble proteins. The unique environment of a TMP compared to a water-soluble protein leads to distinct environmental pressures on their sequences: it is predominantly lipophilic, lacks hydrogen-bonding potential, and provides little screening of electrostatic interaction [10]. However, to date, only a few number of MSA techniques have been proposed expressly for TMPs, or tested using TMPs datasets.

In this paper we present an overview of these few MSA methods applied to TMPs: PRALINE^{TM} [25] and TM-Coffee [3] based on homology extension tested on datasets of TMPs from the BALiBASE v2.0 [1] benchmark, and TM-Aligner [2] based on dynamic programming and Wu-Manber algorithm, tested over the BaliBASE v3.0 reference set 7 of α-helical TMPs proteins [5], Pfam [8] and GPCRDB [16] databases that contains structure based alignment of TMPs.

The content of the paper is structured as follows: the next section presents a formal definition of the MSA problem. Section 3 details the transmembrane substitution rates proposed in literature. An overview of the state-of-the-art of MSA methods applied to TMPs is described in Sect. 4. The benchmarking for TMPs is presented in Sect. 5. Section 6 illustrates a summary of the results

presented by the methods described in Sect. 4. And Sect. 7 describes concluding comments and propose some works for our future research.

2 Multiple Sequence Alignment

In this section, we describe a definition of the MSA problem as follows (MSA score functions are maximized):

Definition 1. *Σ represents a finite alphabet set and S a set of k sequences $(s_1, s_2, \ldots, s_k)$ of different length l_1 to l_k with $s_i = s_{i1}s_{i2}, \ldots, s_{il_i} (1 \leq i \leq k)$, where for DNA sequences, Σ is composed by 4 characters of the nucleotides {A, T, G, C} and for protein sequences, Σ is composed of 20 characters of the amino acids {A, D, C, F, E, H, G, K, I, M, L, P, N, R, Q, T, S, W, V, W, Y}; to find an optimal alignment S' of S, with respects to a scoring function $f(S')$, such that:*

$$S' = (s'_{ij}),\ with\ 1 \leq i \leq k, 1 \leq j \leq l, max(l_i) \leq l \leq \sum_{i=1}^{k} l_i \tag{1}$$

satisfying:

1. *$s'_{ij} \in \Sigma \cup \{-\}$, where "-" represents the gap;*
2. *each row $s'_i = s'_{i1}s'_{i2}, \ldots, s'_{il} (1 \leq i \leq k)$ of S' is the same sequence s_i if we remove all the gap characters;*
3. *the length of the all the k sequences are equals;*
4. *S' has no fully gaps columns.*

The complexity of finding an optimal alignment is $O(k2^k L^k)$, where k is the number of sequences and L is the $max\{l_1, l_2, \ldots, l_k\}$ [29].

Figure 1 illustrates on the left a set of four unaligned sequences. Then, an aligned solution (MSA) for these sequences is illustrated on the right, with two columns totally aligned.

```
s1: AGERSLAATLVC            s1: AG------ERSLAA--TLV-C
s2: DNAILAHERLSIJ           s2: DNAILAH-ER-------LSIJ
s3: CNGYLFIEQLNA            s3: CNGYLFI-E---Q----L-NA
s4: FGLVSDVFEARHMQRLN       s4: FGLVSDVFEARH--MQRL--N
```

Fig. 1. On the left unaligned sequences and on the right an aligned solution example.

3 Substitution Rates for TMPs

PAM [6] or BLOSUM [13] are traditional score matrices communly utilized for sequence retrieval and alignment, but are consequently not ideal to align TM domains [12]. Substitution rates for TMPs, S_{ij}, are communly based on

the frequency of AA substitutions, q_{ij}, in a set of homologous sequences, as indicated by:

$$S_{ij} = \frac{1}{\lambda} \ln \left(\frac{q_{ij}}{f_i f_j} \right) \tag{2}$$

where λ represents a constant, and f_i represents the background frequencies of AA [12].

Various substitution matrix have been suggested to take the evolutionary trends particularly to transmembrane domains, such as: JTT [18], PHAT [24], the asymmetric SLIM matrices [22] and the bbTM matrix for transmembrane β-barrels [17].

4 Multiple Sequence Alignment Applied to TMPs

Very few methods have been proposed to perform MSA of TMPs. The initial proposal is presented by Cserzö *et al.* in [4], describing an algorithm which locates helical TM segments. They demonstrated that corresponding helices in another membrane related protein can be pinpointed just with the location of TM helices of a protein. Evaluating the applicability of their proposal, obtained a good starting point for homology modeling of a G-protein couple receptor (human rhodopsin and bacteriorhodopsin).

The STAM (Simple Transmembrane Alignment Method) method was proposed by Shafrir and Robert in [26] represents a second attempt to improve alignment accuracy by combining two substitution matrices since the frequencies of occurrence of the various AAs differ for TM and water-soluble regions. They identified regions likely to form TM α-helices and apply a higher penalty for insertion/deletions in the TM regions than that of a penalty in the loop region (non-TM regions). To our knowledge STAM is considered as the first software that was specifically targeted at TMPs.

Other study presented by Forrest *et al.* [10] proposed that the use of a bipartite scheme (composed by BLOSUM62 and PHAT) does not significantly improve MSA of TMPs. Introduce HOMEP, a benchmark data set of homologous membrane protein structures and assess current strategies for homology modeling of TMPs.

And recently, three new MSA software for TMPs have been proposed and represent the main topics in this work: PRALINE^{TM}, TM-Coffee and TM-Aligner. This methods are described follow:

4.1 PRALINE^{TM}

PRALINE^{TM} incorpores transmembrane specific information into the previously developed multiple alignment tool PRALINE [14,15]. The strategy includes 3 core functions:

Profile Preprocessing. In the 'preprofile' method, for every sequence a master-slave alignment is created, containing data about neighboring sequences and used in subsequent progressive alignment. These sequence pre-profiles avoid mistakes during the progressive steps and are more informative than single sequences [14,15].

Bipartite Alignment Scheme. Predicts for every input sequence its TM topology utilizing three different predictors: HMMTOP v2.1 [27], TMHMM v2.0 [21] and Phobius [19]. Theses predictors are installed locally and run independently within the PRALINE^{TM} program. Second, to reliably predicted TM positions, the profile-scoring scheme applies the TM-specific substitution score PHAT, applying the following Eq. 3 to score a pair of profile columns x and y:

$$S(x,y) = \sum_{i=1}^{20} \sum_{j=1}^{20} \alpha_i \beta_j M(i,j) \tag{3}$$

where $M(i,j)$ is the exchange weight for residues i and j provided by the selected substitution matrix M and α_i and β_j are the frequencies with which residues i and j appear in columns x and y, respectively. To ensure conflictingly predicted positions don't contrarily impact the alignment quality, two profile columns are coordinated utilizing the PHAT matrix [24] just in the event that every residue in the column is predicted to be part of a TM region, but profile columns are aligned utilizing the BLOSUM62 matrix by default.

Tree-Based Consistency Iteration. In the tree-based consistency iteration used by PRALINE-TM, every edge of a guide tree is utilized to separate the alignment in 2 subalignments, which are progressively realigned. The recent alignment is held just if an enhanced SOP score (Sum-of-Pairs) is accomplished. This score is computed by the entirety of the substitution values of both the BLOSUM62 and PHAT matrix (depending on the TM topology of the AA pair). One iterative cycle suggests that every edge of the tree is visited once.

4.2 TM-Coffee

Chang *et al.* presents in [3] the TM-Coffee software, a TM version of PSI-Coffee able to align TMPs, while utilizing a decreased reference database for homology extension, demonstrating how it can be adjusted and joined with a consistency based approach to improve the MSA of α-helical TMPs. TM-Coffee is included in the T-Coffee software, and a web version is accessible at: http://tcoffee.crg.cat/tmcoffee. With the aim of assess the performance of TM-Coffee, Chang *et al.* tested their proposal over the reference 7 of BAliBASE v2.0 benchmark [1] that contains alpha-helical transmembrane proteins, and demonstrated a relevant improvement over accurate strategies such as MSAProbs, MAFFT, PROMALS, ProbCons, PRALINE^{TM} and Kalign.

Position Specific Iterative PSI/TM-COFFEE WEB-SERVER. A web server version was developed by Floden *et al.* and presented in [9]. This version also allows a rapidly perform of homology extension, using PSI-BLAST searches against a choice of reduced complexity redundant and non-redundant database. Furthermore, using the HMMTOP algorithm outputs topological prediction of TMPs. Login procedure is not required.

4.3 TM-Aligner

TM-Aligner (Transmembrane Membrane proteins - Aligner) [2] is the most recent web-server sequence alignment tool of transmembrane proteins. Use Wu-Manber [30] and dynamic string matching algorithm [7]. The performance of TM-Aligner is assessed over Pfam database, GPCRDB and BaliBASE v3.0 reference set 7 of α-helical TMPs. Has been developed in Perl, C and PHP under a web server on Linux operative system. It is free and available at: http://lms.snu.edu.in/TM-Aligner/.

Scoring Scheme. TM-Aligner uses by default the PHAT substitution matrix [24], defines a gap insertion penalty value of eigth and a gap extension penalty value of one. The alignment process is based on dynamic programming. Aligns all regions independently.

TM-Aligner Workflow. Given a set of input unaligned sequences, the TM-Aligner workflow is described as follows:

- Predict TMs domains into the sequences using TMHMM (Transmembrane Hidden Markov Model) [21].
- Classify into different groups the input sequences, based on the TMs segments of each sequence.
- For overall alignment process, a seed alignment is defined using classes with the dominant number of TM sequences.
- Input protein sequences are separated into regions of TM, non-cytoplasmic and cytoplasmic.
- Dynamic programming technique is used to aligned all these regions independently.
- Most similar sequences are aligned using Progressive or tree-based strategy,
- Add less similar sequences to alignment until all sequences are aligned, successively.
- An initial guide-tree is created using UPGMA method, this guide-tree describes sequence relatedness.
- Wu-Manber algorithm is used to stitch the TM domains with non-cytoplasmic and cytoplasmic segments.

5 Benchmarking Transmembrane Alignments

In this section we detail two benchmarks used to assess the alignment accuracy of TMPs effectively, and used by PRALINE^{TM}, TM-Coffee and TM-Aligner.

5.1 BAliBASE

BAliBASE [5] is one of the classical benchmark from the literature. Includes a set of alignments obtained from manual alignment and/or structural databases. Contains a special reference set of TMPs, called Reference set 7 [1], with 8 accurately aligned TMPs families namely 7tm, msl, dtd, acr, photo, ion, Nat and ptga. Contains 435 sequences in total. Have from 2 to 14 TM α-helices per sequence. The core domains are authors-defined, examining the alignment of structurally equivalent residues only. The main goal of BAliBASE is assess the capacity of the strategies to recapitulate these core domains, mostly made of α-helices. Furthermore, contains a program to assess accuracy of the candidate alignments over reference alignments provided by the benchmark, called Baliscore that includes two metrics: Total Column (TC) and Sum-of-Pairs (SP) scores.

5.2 Pfam Database

The Pfam [8] is available at http://pfam.xfam.org, is a large database that contains a set of preserved protein families represented by HMMs (Hidden Markov Models) and MSA. With the aim of accurately identify the gap penalty, length parameter in profile hidden markov model and position-specific AA frequency, seed alignment are based on principal protein sequences. The last version of Pfam 31.0 contains 16712 entries and 604 clans, in this release over 36% of Pfam entries are placed within a clan. All the information for every entry as obtained from UniProt Reference Proteomes.

6 Performance Comparison

In this section, we present a performance comparison between PRALINE^{TM}, TM-Coffee and TM-Aligner, compared with themselves and other classical alignment methods. Sum-of-Pair (SP) score of BAliBASE and CPU processing time (in seconds) were considered for each protein family of BAliBASE reference set7. All these results were taken from the literature [2,3,25]. Table 1 shows the individual and average SP, bold values are the best score for each set.

In Table 1 we see that TM-Coffee achieves the highest average SP over all eight datasets and the best individual SP score for the 7TM, ACR, DTD and PTGA sets. TM-Aligner and PRALINE^{TM} achieves the best individual SP score for the MSL and ION sets, respectively. Furthermore, MAFFT is relatively robust on TM sequences, obtains the best individual SP score for the NAT and PHOTO sets. Table 2 shows the results evaluating the CPU processing time in seconds. These results were taken from [2]. Basharat *et al.* tested individually the tools using single threaded machine with two available cores, including TM-Aligner [2].

Table 1. Comparison between the PRALINE^{TM}, TM-Coffee and TM-Aligner methods and other widely-used multiple alignment tools

Set	ClustalW	Muscle	Mafft	ProbCons	Praline	Promals	Kalign	TM-Aligner	PRALINE^{TM}	TM-Coffee
7TM	0.85	0.84	0.84	0.88	0.82	0.83	0.48	0.82	0.86	**0.88**
ACR	0.91	0.95	0.94	0.94	0.93	0.91	0.92	0.92	0.94	**0.95**
DTD	0.79	0.86	0.84	0.88	0.82	0.85	0.50	0.87	0.86	**0.88**
ION	0.35	0.52	0.51	0.53	0.35	0.50	0.29	0.51	**0.54**	0.54
MSL	0.86	0.87	0.85	0.85	0.81	0.85	0.70	**0.89**	0.87	0.84
NAT	0.63	0.74	**0.77**	0.75	0.72	0.75	0.28	0.75	0.71	0.72
PHOTO	0.89	0.90	**0.93**	0.91	0.92	0.91	0.50	0.92	0.93	0.91
PTGA	0.46	0.55	0.73	0.72	0.40	0.74	0.32	0.70	0.68	**0.74**
AVG	0.72	0.78	0.80	0.81	0.72	0.79	0.50	0.80	0.80	**0.81**

Table 2. Comparison of CPU processing time in seconds of TM-Coffee and TM-Aligner methods and other widely-used multiple alignment tools, PRALINE^{TM} is not included because the standalone version is unavailable [2].

Set	PROMALS	ClustalW	Muscle	Mafft	Kalign	TM-Aligner	TM-Coffee
PTGA	17633	5	28	38	3	17	778
ACR	35622	8	28	35	6	26	1836
MSL	1055	1	3	12	1	3	17
DTD	21885	6	32	44	3	24	1443
PHOTO	3962	1	3	26	1	7	38
ION	18521	4	78	45	6	26	1385
NAT	21055	6	32	54	3	21	602
7TM	35865	19	52	117	6	56	4346
AvG	19450	6	32	46	4	23	1306

7 Conclusions

There is a lot MSA methods proposed in the literature, but to date, there is a small number of MSA methods proposed specifically for TMPs. In this work we have addressed three of the MSA methods applied to TMPs: PRALINE^{TM}, TM-Coffee and TM-Aligner. We have summarized their principal features and illustrated a performance comparison between them and other classical MSA methods evaluating SP score and the CPU processing time over the protein family of BAliBASE reference set-7. TM-Coffee achieves high accuracy results, however TM-aligner is the faster method in terms of CPU processing time (seconds).

The studied methods suggest that 2D structure prediction and dynamic programming (TM-Aligner), bipartite scheme using membrane-specific scoring matrices (PRALINE^{TM}) and homology extension (TM-Coffee) can increase alignment quality for TMPs.

Finally, considering the complexity of the problem, we suggest that the alignment process of TMPs can be tackled with stochastic methods, introducing an

alternative technique useful from a biological point of view. Furthermore, with parallel techniques to reduce the execution time.

Acknowledgement. This work has been supported by the 5th convocation of Fondo Competitivo de Investigación Científica y Tecnológica FOCICYT of the Universidad Técnica Estatal de Quevedo from Ecuador.

References

1. Bahr, A., Thompson, J.D., Thierry, J.C., Poch, O.: BAliBASE (benchmark alignment database): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res. **29**(1), 323–326 (2001). https://doi.org/10.1093/nar/29.1.323
2. Bhat, B., Ganai, N.A., Andrabi, S.M., Shah, R.A., Singh, A.: TM-Aligner: multiple sequence alignment tool for transmembrane proteins with reduced time and improved accuracy. Sci. Rep. **7**(1), 1–8 (2017). https://doi.org/10.1038/s41598-017-13083-y
3. Chang, J.M., Di Tommaso, P., Taly, J.F., Notredame, C.: Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinform. **13**(4), S1 (2012). https://doi.org/10.1186/1471-2105-13-S4-S1
4. Cserzö, M., Bernassau, J.M., Simon, I., Maigret, B.: New alignment strategy for transmembrane proteins. J. Mol. Biol. **243**(3), 388–396 (1994). https://doi.org/10.1006/jmbi.1994.1666
5. Thompson, J.D., Koehl, P., Ripp, R., Poch, O.: BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins: Struct. Funct. Bioinform. **61**(1), 127–136 (2005). https://doi.org/10.1002/prot.20527
6. Dayhoff, M., Schwartz, R., Orcutt, B.C.: A model of evolutionary change in proteins. Atlas Protein Seq. Struct. **5**, 345–352 (1978)
7. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
8. Finn, R.D., et al.: The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. **44**(D1), D279–D285 (2016). https://doi.org/10.1093/nar/gkv1344
9. Floden, E.W., Tommaso, P.D., Chatzou, M., Magis, C., Notredame, C., Chang, J.M.: PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic Acids Res. **44**(W1), W339–W343 (2016). https://doi.org/10.1093/nar/gkw300
10. Forrest, L.R., Tang, C.L., Honig, B.: On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys. J. **91**(2), 508–517 (2006). https://doi.org/10.1529/biophysj.106.082313
11. Frishman, D.: Structural Bioinformatics of Membrane Proteins (2010). https://doi.org/10.1007/978-3-7091-0045-5
12. Frishman, D.: Structural Bioinformatics of Membrane Proteins. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-7091-0045-5
13. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89**(22), 10915–10919 (1992)

14. Heringa, J.: Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput. Chem. **23**(3), 341–364 (1999). https://doi.org/10.1016/S0097-8485(99)00012-1
15. Heringa, J.: Local weighting schemes for protein multiple sequence alignment. Comput. Chem. **26**(5), 459–477 (2002). http://www.sciencedirect.com/science/article/pii/S0097848502000086
16. Isberg, V., et al.: GPCRdb: an information system for g protein-coupled receptors. Nucleic Acids Res. **44**(D1), D356–D364 (2016). https://doi.org/10.1093/nar/gkv1178
17. Jimenez-Morales, D., Adamian, L., Liang, J.: Detecting remote homologues using scoring matrices calculated from the estimation of amino acid substitution rates of beta-barrel membrane proteins. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1347–1350, August 2008. https://doi.org/10.1109/IEMBS.2008.4649414
18. Jones, D., Taylor, W., Thornton, J.: A mutation data matrix for transmembrane proteins. FEBS Lett. **339**(3), 269–275 (1994)
19. Käll, L., Krogh, A., Sonnhammer, E.L.: A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. **338**(5), 1027–1036 (2004). https://doi.org/10.1016/j.jmb.2004.03.016
20. Kozma, D., Simon, I., Tusnády, G.E.: PDBTM: protein data bank of transmembrane proteins after 8 years. Nucleic Acids Res. **41**(D1), D524–D529 (2013). https://doi.org/10.1093/nar/gks1169
21. Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L.: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes1. J. Mol. Biol. **305**(3), 567–580 (2001)
22. Müller, T., Rahmann, S., Rehmsmeier, M.: Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics **17**(suppl1), S182–S189 (2001). https://doi.org/10.1093/bioinformatics/17.suppl_1.S182
23. Newby, Z.E., et al.: A general protocol for the crystallization of membrane proteins for x-ray structural investigation. Nat. Protoc. **4**(5), 619 (2009)
24. Ng, P.C., Henikoff, J.G., Henikoff, S.: PHAT: a transmembrane-specific substitution matrix. Bioinformatics **16**(9), 760–766 (2000). https://doi.org/10.1093/bioinformatics/16.9.760
25. Pirovano, W., Feenstra, K.A., Heringa, J.: Praline^{TM}: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics **24**(4), 492–497 (2008). https://doi.org/10.1093/bioinformatics/btm636
26. Shafrir, Y., Guy, H.R.: STAM: simple transmembrane alignment method. Bioinformatics **20**(5), 758–769 (2004). https://doi.org/10.1093/bioinformatics/btg482
27. Tusnády, G.E., Simon, I.: The hmmtop transmembrane topology prediction server. Bioinformatics **17**(9), 849–850 (2001). https://doi.org/10.1093/bioinformatics/17.9.849
28. Wallin, E., Heijne, G.V.: Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. **7**(4), 1029–1038 (1998)
29. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. J. Comput. Biol. **1**, 337–348 (1994)
30. Wu, S., Manber, U.: Fast text searching: allowing errors. Commun. ACM **35**(10), 83–91 (1992)