TM-MSAligner: a tool for multiple sequence alignment of transmembrane proteins

Joel Cedeño-Muñoz[1], Cristian Zambrano-Vega[2], and Antonio J. Nebro[3],[4⋆]

[1] Facultad de Ciencias Pecuarias y Biológicas, State Technical University of Quevedo
Quevedo, Los Ríos, Ecuador
jacedeno@uteq.edu.ec
[2] Facultad de Ciencias la Ingeniería, State Technical University of Quevedo
Quevedo, Los Ríos, Ecuador
czambrano@uteq.edu.ec
[3] ITIS Software, Edificio de Investigación Ada Byron, University of Málaga,
Málaga, 29071, Spain
[4] Dept. de Lenguajes y Ciencias de la Computación, University of Málaga,
ajnebro@uma.es
Málaga, 29071, Spain

Abstract. Transmembrane proteins (TMPs) are crucial to cell biology, making up about 30% of all proteins based on genomic data. Despite their importance, most of the available software for aligning protein sequences focuses on soluble proteins, leaving a gap in tools specifically designed for TMPs. Only a few methods target TMP alignment, with just a couple of the available to researchers. Considering that there are a few particular differences that ought to be taken into consideration aligning TMPs sequences, standard MSA methods are ineffective to align TMPs. In this paper, we present TM-MSAligner, a software tool designed to deal with the multiple sequence alignment of TMPs by using a multi-objective evolutionary algorithm. Our software include features such as transmembrane substitution matrix dynamically used according to the topology region, a high penalty to gap opening and extending, and two MSA quality scores, Sum-Of-Pairs with Topology Prediction and Aligned Segments, that can be optimized at the same time. This approach reduce the number of Transmembrane (TM) and non-Transmembrane (non-TM) broken regions and improve the TMP quality score. TM-MSAligner outputs the results in an HTML format, providing an interactive way for users to visualize and analyze the alignment. This feature allows for the easy identification of each topological region within the alignment, facilitating a quicker and more effective analysis process for researchers.

Keywords: Multiple sequence alignment, transmembrane proteins, multi-objective optimization, evolutionary algorithms, software framework

⋆ Corresponding author: ajnebro@uma.es

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

1 Introduction

The study of Transmembrane Proteins (TMPs) sequences has taken increasing attention in recent years due to their fundamental roles in various biological processes and their significance as potential drug targets and life science research [14, 9, 15]. Transmembrane proteins are involved in vital cellular functions, such as signal transduction, ion transport, and cell adhesion, making them key players in maintaining cellular homeostasis and energy production [11]. Sequence analysis methods for TMPs are of great interest in the biomedical and bioinformatics domains and understanding the structural and functional aspects of these proteins is crucial for unraveling the complexities of cellular mechanisms.

Multiple Sequence Alignment (MSA) remains one of the most powerful tools for assessing evolutionary sequence relationships and for identifying structurally and functionally important protein regions [11]. It serves as a foundational step for a range of further analyses in protein family studies, including homology modeling, predicting secondary structures, and understanding phylogenetic relationships. Transmembrane regions exhibit unique amino acid compositions and conservation patterns, differing significantly from soluble proteins. Traditional MSA methods fail to consider these distinctions when aligning TMPs, resulting in reduced accuracy of the alignments. Furthermore, there are few techniques available that can align TMPs while also optimizing for more than one MSA quality score, highlighting a gap that we address in this paper.

We introduce TM-MSAligner, a novel software tool aimed at finding multiple sequence alignments of TMPs using a multi-objective evolutionary algorithm [2]. These are stochastic nature-inspired search algorithms belonging to the family of metaheuristics [1] that do not guarantee to find optimal solutions but they usually provide accurate solutions in a reasonable amount of time. The alignment of TMPs is formulated as a bi-objective optimization where the Sum-Of-Pairs with Topology Prediction and Segments Aligned are defined as scores to be maximized, so its optimum is a set of trade-off solutions between the two objectives known. In the field of multi-objective optimization, this set is known as the Pareto set and their correspondence in the objective space is referred as to Pareto front. Due to the stochastic feature of multi-objective evolutionary algorithms, they provide as a result an approximation to the Pareto front.

The results obtained when TM-MSAligner is executed are generated in both CSV and HTML format, allowing the latter to plot the found alignments, so that the user can choose the solution that best meet defined criteria and the HTML page will show the alignment using the MSABrowser[5] viewer. In this way, the topology of the amino acids is shown in different colours, which facilitates the process of identifying the topological regions included in the alignments.

The rest of the paper is structured as follows. The package is described in Section 2 and an usage example is detailed in Section 3. The next section includes a discussion about our tool and, finally, Section 5 provides the conclusions and future works.

[5] MSABrowser: `https://thekaplanlab.github.io/`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

Parameter/Component	Type	Domain	Depedency
algorithmResult	c	{externalArchive, population}	
populationSizeWithArchive	i	[10, 200]	algorithmResult == ExtArch
externalArchive	c	{CDA, unbounded, hypevolume}	algorithmResult == ExtArch
offspringPopulationSize	i	[1, 400]	
selection	c	{tournament, random}	
selectionTournamentSize	i	[2, 10]	selection == tournament
replacement	c	{rankingAndDensityEstimator}	
ranking	c	{dominance, strength}	
densityEstimator	c	{crowdingDistance, knn}	
kValueForKNN	i	[1, 3]	densityEstimator == knn
variation	c	{crossoverAndMutation}	
crossover	c	{SPX}	
crossoverProbability	r	[0.0, 1.0]	
mutation	c	{IRG, MAGG, SCG, SANGG}	
mutationProbabilityFactor	r	[0.0, 2.0]	

Table 1. Parameter space of TM-MSAligner. Types: (c)ategorical, (i)nteger, (r)eal. (CDA: crowdingDistanceArchive, IRG: insertRandomGap, MAGG: mergeAdjuntedGapsGroups, SCG: shiftClosedGaps, SANGG: splitANonGapsGroup, ExtArch: external archive)

2 Software description

The core of TM-MSAligner is a multi-objective evolutionary algorithm that combines features of algorithms such as NSGA-II [3] and SPEA2 [16]. The package is implemented in Java as a Maven project and uses the jMetal framework for multi-objective optimization using metaheuristics [4][8] as a dependence. TM-MSAligner is an open source project under MIT license[6].

2.1 Software architecture

The evolutionary algorithm of TM-MSAligner architecture is based on a workflow of components where each component is a step of the algorithm. Some of these components have more than one implementation and they can also have control parameters. The parameter space of TM-MSAligner is shown in Table 1.

The approach adopted to create the initial population is to pre-compute alignments by using existing tools (e.g., ClustalW2, T_Coffee, Muscle, Kalign, Mafft, Probcons, etc.) and recombining them to get the desired number of individuals for the population. This way the algorithm is able of providing accurate solutions faster than initializing the population with alignments obtained by filling the gaps randomly.

The evaluation of the population can be carried out sequentially or in parallel. The parallel model can be synchronous, where the behavior of the evolutionary algorithm does not change, or asynchronous [5], which can be more efficient than the synchronous one when using a large number of cores.

[6] TM-MSAligner: `https://github.com/jMetal/TM-MSAligner`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

2.2 Transmembrane Proteins features

TM-MSAligner includes some features to lead with TMPs sequences. The most relevant are:

- **Topology Prediction**: Topology prediction is used to identify transmembrane regions within the protein sequences. We have used DeepTMHMM [6], a deep learning protein language model-based algorithm to predict the topology of both alpha-helical and beta-barrels proteins. This software exports the results in a *.3line* format file, where each line represents the name of the sequence, the sequence of amino-acids and the TM topology, respectively. This information must be included in the input data file.
- **Transmembrane substitution matrix**: The substitution matrix is determined dynamically in our algorithm. It depends on consistent TM predictions over a column. The Sum-Of-Pairs MSA quality score applies the PHAT [10](Predicted Hydrophobic and Transmembrane) substitution matrix on consistently predicted TM regions and BLOSUM [7] substitution matrix on non-TM regions.
- **Regions based Gap Penalty**: A TM-regions based Gap Penalty is incorporated in our proposal. The TM regions and non-TM regions are respectively denoted with different *Open Gap Penalties* and *Extended Gap Penalties*, so, with the aim of make TM regions harder to be broken during the aligning, the gap penalties in TM regions is higher than gap penalties in non-TM regions.
- **Aligned Segments**: The second fitness score to optimize is to generate alignments with the highest number of aligned regions with the same topology inside the MSA.

2.3 Execution modes

TM-MSAligner allows to use a number of choices to configure and run the evolutionary algorithm:

- `TM_MSAligner`: All parameters can be set manually, allowing advanced users to fine-tune the settings. This execution mode is provided through a class that is typically executed in an integrated development environment (IDE), such as Eclipse or IntelliJ Idea.
- `ConfigurableTM_MSAligner`: The algorithm accepts as input a string with a particular combination of parameter values, what it is very convenient if we intend to run TM-MSAligner from the command line.
- `BAliBASETest`: This mode provides an adapted version of TM-MSAligner to solve the instances of the BAliBASE-ref7 benchmarking [12]. The unaligned sequences, the topology information of the proteins and the pre-computed alignments are saved in the `resources` folder of the project.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

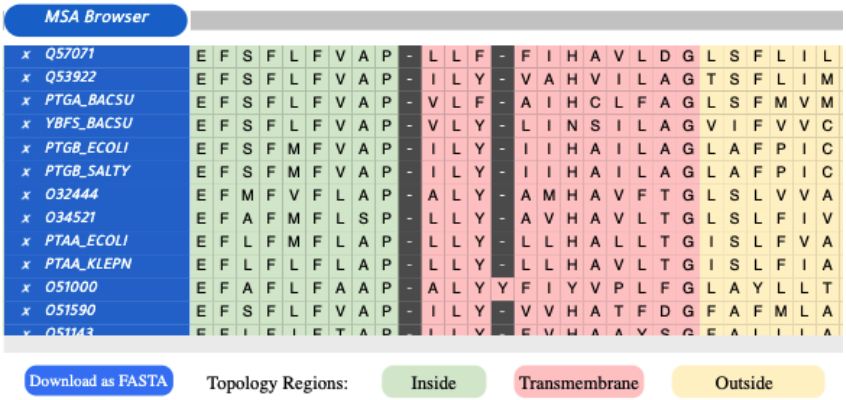

Fig. 1. Example of how MSAViewer displays the alignments, giving a different color to each topology for an easy identification of the regions.

2.4 Output Results

TM-MSAligner generates the following results:

- A Web page with the plot of the Pareto Front approximation and the visualization of the alignments selected by user, who can click the preferred point over the plot figure.
- List of alignments which represents the Pareto Front approximation. Each MSA solution is illustrated in HTML format using the viewer MSABrowser[13]. With the aim of identify the regions into the MSA, the topology of the aminoacids is colored with different values, as can be seen in the example of a MSA visualization shown in Figure. 1.
- The output of the *Observer* selected. It can be a plot figure or the best scores reached by the algorithm during its execution.

3 Illustrative example

To illustrate the working of TM-MSAligner, we use it to solve the TMPs dataset of BAliBASE called reference 7 [12]. The original sequences, the transmembrane topology information, and the pre-computed alignments are saved in the *ref7* sub-folder of the *resources* directory.

Figure 2 show the Pareto front approximation obtained with the highlighted solution having the highest Aligned Segment score and Sum-Of-Pairs with Topology Prediction scores.

Figure 3 depicts the visualization of the alignments assigned to the solution with the highest and lowest Sum-Of-Pair with Topology Prediction score values. We can observe that this score penalizes the insertion of Opening Gaps, even more inside transmembrane regions, and that the alignment with the highest score has fewer TM regions broken than the lowest one.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

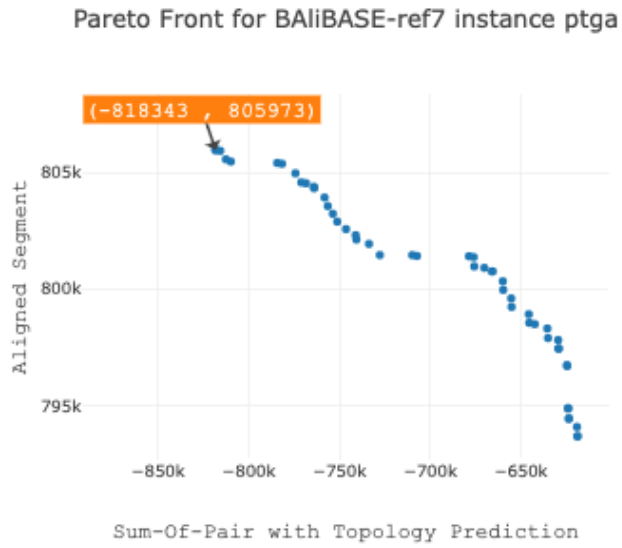

Fig. 2. Solution selected with the highest Segment Aligned score.

4 Discussion

Nowadays, there are many MSA procedures that have been built and tested to align homologous soluble proteins, but only few have been adapted to lead with TMPs and only a pair are currently available. Given the biomedical importance of TMps and the large and growing gap between the number of solved TMP structures and the number of TMP sequences, sequence analysis techniques are crucial.

To address this challenge, TM-MSAligner is a software tool that allows users with a bioinformatics background to find sets of accurate alignments for TMPs representing trade-offs solutions according to the Sum-Of-Pairs with Topology Prediction and Aligned Segment objectives. The alignments performed by our software will have more conserved TM and non-TM regions. The parallelism features of TM-MSAligner can help to accelerate the optimization process by making use of the available cores of modern CPUs.

5 Conclusions

We have presented TM-MSAligner, a software tool developed to find multiple sequence alignment of transmembrane proteins by using a multi-objective evolutionary algorithm. For the purpose to improve the alignment accuracy for TMPs, some specific features have be taken, adapting our MSA method to lead with TMPs.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

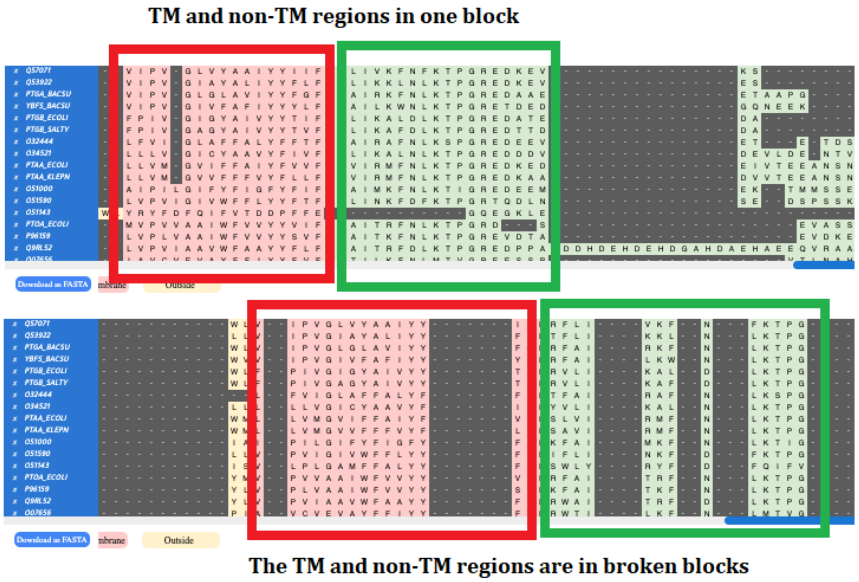

Fig. 3. Visualization of the Alignments with the highest (image above) and lowest (image below) Sum-Of-Pair with Topology Prediction scores.

As TM-MSAligner can be executed from the command line, only a minimum knowledge of the Java development tools is required, while experienced users have access to the source code, so they have the have the chance to extend the package with new components (e.g., mutation and crossover operators) and new algorithms.

TM-MSAligner has an open source license and it is hosted in a GitHub repository containing the source code and the documentation. The software tool can be downloaded and executed on Windows, Linux and macOS computers.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Innovation via Grant PID2020-112540RB-C41 (AEI/FEDER, UE) and by the Junta de Andalucía, Spain, under contract QUAL21 010UMA.

References

1. Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys 35(3), 268–308 (2003)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10

2. Coello Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A.: Evolutionary Algorithms for Solving Multi-Objective Problems. Springer, New York, second edn. (September 2007), iSBN 978-0-387-33254-3
3. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
4. Durillo, J.J., Nebro, A.J.: jMetal: A java framework for multi-objective optimization. Advances in Engineering Software 42(10), 760–771 (2011)
5. Durillo, J.J., Nebro, A.J., Luna, F., Alba, E.: A study of master-slave approaches to parallelize nsga-ii. In: 2008 IEEE International Symposium on Parallel and Distributed Processing. pp. 1–8 (2008)
6. Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Armenteros, J.J.A., Marcatili, P., Nielsen, H., Krogh, A., Winther, O.: Deeptmhmm predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv (2022)
7. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89(22), 10915–10919 (1992)
8. Nebro, A.J., Durillo, J.J., Vergne, M.: Redesigning the jMetal multi-objective optimization framework. Genetic and Evolutionary Computation Conference pp. 1093–1100 (7 2015)
9. Ng, D.P., Poulsen, B.E., Deber, C.M.: Membrane protein misassembly in disease. Biochimica et Biophysica Acta (BBA) - Biomembranes 1818(4), 1115–1122 (2012), protein Folding in Membranes
10. Ng, P.C., Henikoff, J.G., Henikoff, S.: Phat: a transmembrane-specific substitution matrix. BIOINFORMATICS 16, 760–766 (2000)
11. Pirovano, W., Abeln, S., Feenstra, K.A., Heringa, J.: Multiple alignment of transmembrane protein sequences, pp. 103–122. Springer Vienna, Vienna (2010)
12. Thompson, J.D., Koehl, P., Ripp, R., Poch, O.: Balibase 3.0: Latest developments of the multiple sequence alignment benchmark. Proteins: Structure, Function, and Bioinformatics 61(1), 127–136 (2005)
13. Torun, F.M., Bilgin, H.I., Kaplan, O.I.: MSABrowser: dynamic and fast visualization of sequence alignments, variations and annotations. Bioinformatics Advances 1(1), vbab009 (08 2021)
14. Wallin, E., Heijne, G.V.: Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Science 7(4), 1029–1038 (1998)
15. Yin, H., Flynn, A.D.: Drugging membrane protein interactions. Annual Review of Biomedical Engineering 18(1), 51–76 (2016), pMID: 26863923
16. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm. Tech. Rep. 103, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland (2001)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_10