Multiple Sequence Alignment of Transmembrane Beta-Barrel Proteins

Akhil Jobby

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (MCompSc) at

Concordia University

Montréal, Québec, Canada

September 2019

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Akhil Jobby**

Entitled: **Multiple Sequence Alignment of Transmembrane Beta-Barrel Proteins**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (MCompSc)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Tse-Hsun Chen

_____ Examiner
Dr. Tristan Glatard

_____ Examiner
Dr. Aiman Hanna

_____ Supervisor
Dr. Gregory Butler

Approved by _____
Dr. Lata Narayanan, Chair
Department of Computer Science and Software Engineering

_____ 2019

_____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Multiple Sequence Alignment of Transmembrane Beta-Barrel Proteins

Akhil Jobby

The two main types of transmembrane proteins are transmembrane alpha helices (TMAH) and transmembrane beta-barrels (TMBB). From literature, we know that both are responsible for diverse biologically important functions. Since there is plenty of sequence and structural data available on various TMAH proteins, many techniques have been developed to analyze their sequence. On the contrary, not many TMBB proteins have been identified or studied. One of the most powerful sequence analysis techniques used for identifying and annotating the biological sequences is "Multiple Sequence Alignment" (MSA). It is often used for phylogenetic analysis, identification of conserved regions in the sequences, prediction of the topology of proteins, etc.

High-throughput sequencing methods generate huge volume of sequence data, but they remain largely unannotated. Hence an MSA method for TMBB would be important for sequence-based studies and identifying more of such proteins. In this thesis, we apply a method called homology extension to the MSA and adjust the strategy applied by TM-Coffee, a state of the art MSA method tested for TMAH, to make it suitable for TMBB proteins. We focus on extensively evaluating this method and comparing it with popular MSA tools.

Acknowledgments

I would like to sincerely express my gratitude towards my supervisor Dr. Gregory Butler for the guidance and support towards me throughout my studies and research. I thank him for his efforts in improving my thesis. I could not have imagined having a better supervisor than him.

Besides my supervisor, I would like to thank my parents and my brother from the bottom of my heart for not losing faith in me and for providing continuous support and love. A very special thanks to Emma, who stood by me during worse times and for being a pillar of motivation. Finally, I would like to thank my lab colleagues; Munira, Stephanie, Stuart, Qing and my friends Tejas, Gurpreet and others for their help during my studies and research.

Contents

List of Figures

List of Tables

Glossary

Amino Acid: A basic building block for proteins. Amino acids are made up of amino group (-NH2), an acidic carboxyl group (-COOH), and an organic R group (or side chain). This R group or side chain distinguishes amino acids from each other. There are 20 different general amino acid molecules represented by 20 letters.

Blast (Basic Local Alignment Search Tool): A program that allows comparison of input/query sequence with a database of sequences in a fast and efficient manner. Quality of search hits that are returned by the program can also be determined.

Blastp: Blast for protein sequences.

Blastp+: A standalone Blastp.

BLOSUM (Blocks Substitution Matrix): A substitution matrix used for protein sequence alignment.

Clustal Omega: A multiple sequence alignment program.

E-value (Expect Value): The number of hits (from Blast) one can expect to see by chance while searching a database of fixed size.

HHalign: A method to find pairwise alignments. It is part of the HH suite of programs. The name comes from the fact that it performs HMM-HMM alignments.

Hidden Markov Model: A statistical model for a Markov process with hidden states.

Hits: Similar sequences returned by Blast program.

HSP (High-scoring Segment Pairs): Subsegments of a pair of sequences that are highly similar.

MAFFT (Multiple Alignment using Fast Fourier Transform): A multiple sequence alignment program.

MBed: This method embeds sequences into n dimensions to create an n element vector. The vector is used to determine Euclidean distance to obtain guide trees.

Membrane Protein: Proteins that are embedded in, or attached to, the cell membrane.

MSA (Multiple Sequence Alignment): Sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.

MUSCLE (Multiple Sequence Comparison by Log-Expectation): A multiple sequence alignment program.

OMPDB: A database of beta-barrel outer membrane protein from gram negative bacteria.

Outer Membrane Protein: Proteins in bacteria from the outer membrane of the cell.

PAM (Point Accepted Mutation): A substitution matrix used for protein sequence alignment.

PDB (Protein Data Bank): A database of protein structures.

PDBTM (Protein Data Bank of Transmembrane Proteins): Protein structure database for transmembrane proteins.

PFam: A database of protein families.

PHAT (Predicted Hydrophobic and Transmembrane): A substitution matrix specific to transmembrane proteins.

Protein: Proteins are bio-molecules made up of a long chain of amino acids. They can be represented by letters corresponding to each amino acid.

Protein Secondary Structure: Segments of proteins representing 3-Dimensional forms such as alpha helix, beta sheet, coils and turns.

PSSM (Position Specific Substitution Matrix): A matrix used to represent patterns in sequences.

Query sequence: Input sequence to Blast program.

Residue: Amino acids in a protein can be called residues.

Substitution Matrix: In the context of amino acids, it is a matrix that describes the rate of change of residues by a model of evolution.

T-Coffee (Tree-based Consistency Objective Function for Alignment Evaluation): A multiple sequence alignment program.

TM-Coffee: A multiple sequence alignment tool for transmembrane alpha helices.

TopDB (Topology Data Bank): A database of experimentally derived topology of transmembrane proteins.

Topology: Refers to the combination and orientation of protein secondary structures.

Transmembrane Protein: Proteins embedded in the cell membrane and pass through them in single or multiple passes.

TMBB (Transmembrane Beta-Barrel Protein): Specific type of transmembrane protein that are found in the outer membrane of gram negative bacteria, mitochondria and chloroplast. The beta sheets of the protein secondary structure take on the shape of a barrel in the 3-Dimensional structure.

UPGMA (Unweighted Pair Group Method with Arithmetic Mean): A hierarchical clustering method and is often used for determining phenograms. It assumes a constant rate of evolution.

WSP (Weighted Sum of Pairs): Addition of all pairwise alignment scores but down weighting contributions from homologous sequences.

Chapter 1

Introduction

1.1 Motivation

Transmembrane beta-barrel (TMBB) proteins are one of the important classes of proteins which are involved in various function in cellular and physiological processes like substrate transport and signaling [49]. This has immense medical interest especially because, these proteins are mainly localized in the outer membrane of gram-negative bacteria.

Multiple sequence alignment (MSA) [15] has become the most robust method used to identify protein secondary structure. It is necessary to identify such protein regions that are structurally and functionally important. MSA is the prerequisite to several studies like secondary structure prediction, homology modeling, phylogenetic analysis, substrate specificity etc. MSAs can also be used to create profiles or HMMs and identify distantly related sequences. Thus, it is a very effective tool for evolutionary and molecular biology studies [45].

There are a few MSA tools that are transmembrane aware; PralineTM [47], TM-Coffee [10] and STAM [50]. However, all of them are modeled and tested on the abundantly available transmembrane alpha-helical proteins and not on transmembrane beta-barrel proteins. Since there is no method that is designed for TMBB, it is important to have such a method that can align multiple TMBB sequences. In this work, we focus on creating a procedure that will compute MSA specifically for transmembrane beta-barrel proteins. Since there is

a lack of benchmark dataset for transmembrane beta-barrel (TMBB) proteins, we construct a dataset from various sources such as research papers and protein databases.

1.2 Biological Background

The cell wall of gram-negative bacteria is composed of three layers; the exterior outer-membrane, the middle peptidoglycan layer and inner plasma membrane (also known as a cytoplasmic membrane or inner membrane) [31]. The outer membrane is made up of phospholipids, lipoproteins, surface and integral proteins. Integral proteins that run through the cross-section of these membranes are called *transmembrane proteins*. They are divided into two types main types; transmembrane alpha-helical proteins and transmembrane beta-barrel proteins, shown in Figure 1.1. This classification is based on the secondary structures that make up these proteins. The main secondary structure of these proteins are alpha helix and beta sheet, respectively, shown in Figure 1.2. Other types of secondary structures include turns and loops. Transmembrane beta-barrel (TMBB) proteins are not only found in the outer membrane of gram-negative bacteria, but also in the outer membrane of cell organelles such as mitochondria and chloroplasts. All of these proteins are generally classified as Outer Membrane Proteins (OMP)[7]. An example of a typical beta-barrel protein is OmpG [70] illustrated in Figure 1.3 [31].

After the translation process by which proteins are synthesized, anti-parallel beta-strands form beta sheets and they fold and turn to form the tertiary structure of the beta-barrel protein. When these proteins are transported and embedded in the membrane, they are called transmembrane beta-barrel proteins. They get their name from the barrel-like structure it forms with the help of the secondary structure beta sheet, illustrated in Figure 1.3. In Figure 1.3, (A) shows side view of the protein and (B) shows the top view. S1-S14 represents the 14 beta strands that join to form the barrel-like shape. L1-L7 are loops and T1-T6 are turns [70]. Beta strands in this barrel are held together by hydrogen bonds and contain hydrophilic and hydrophobic regions. Some beta-barrel proteins that occur in the plasma region of the cell have a hydrophobic core and hydrophilic exterior [56],

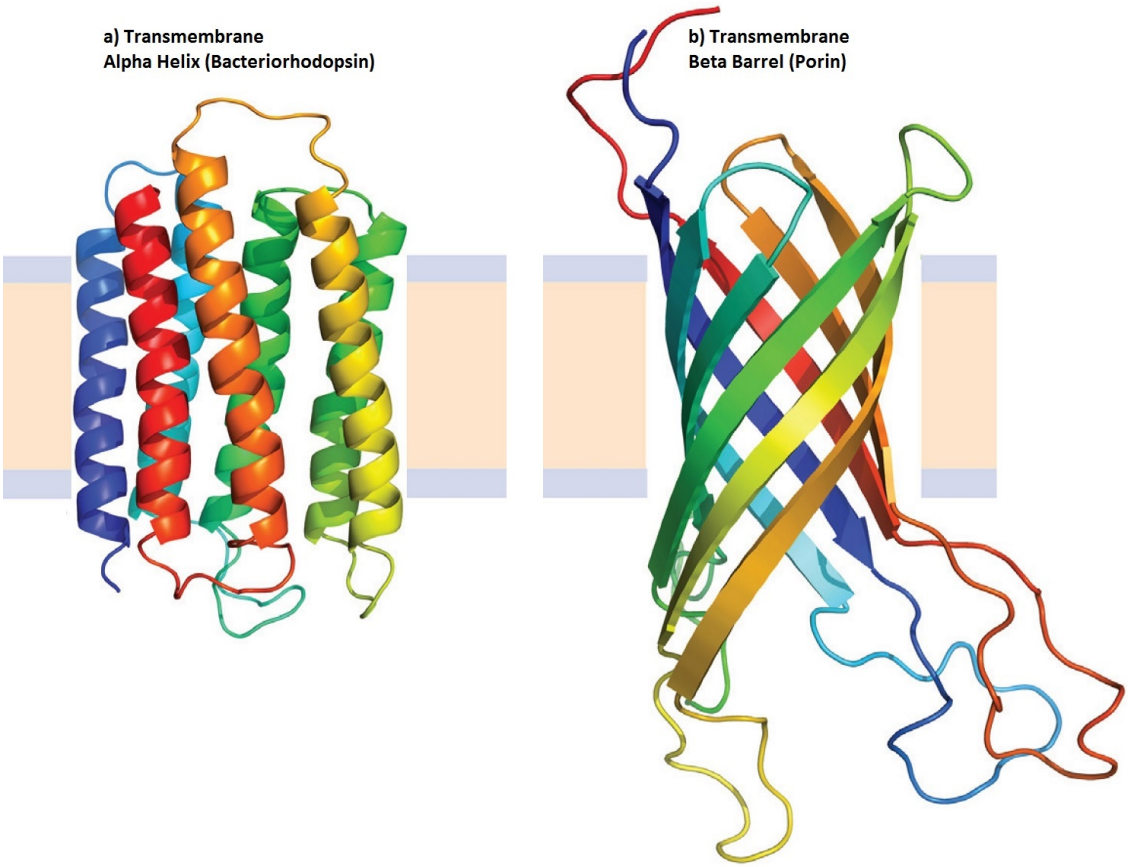

Figure 1.1: Classification of Transmembrane Proteins. This figure shows the two main types of transmembrane proteins. a) Transmembrane alpha helices. The image shows a proteins called bacteriorhodopsin and b) Transmembrane beta-barrel. TMBB protein shown here is called Porin. Original picture by Roman, E.A.; González Flecha, F.L; CC BY 3.0.

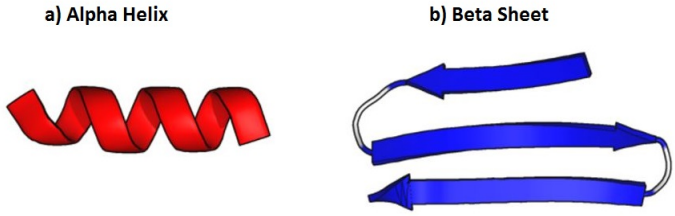

Figure 1.2: Two main types of secondary structure or proteins. Image shows tow main types of secondary structure of proteins. a) Alpha helix and b) Beta sheet. Several beta strands are joined together to form a beta sheet. Original picture by Thomas Shafee; CC BY 4.0.

whereas others [27], such as porins, have a reverse construction with a polar hydrophilic core and non-polar hydrophobic exterior to help them anchor into the outer membrane of the cell wall that is composed of hydrophobic phospholipids [69]. Transmembrane beta-barrel proteins are interchangeably called beta-barrel transmembrane, transmembrane beta-barrel or TMBB proteins.

TMBBs are crucial for pore formation, transport of molecules and voltage gating, and they are molecular targets for antimicrobial drugs [20]. Some of these proteins are also known to be bacterial toxins. TMBBs are extremely difficult to be determined experimentally [62]. As a result, not many TMBB structures have been reviewed and produced with high resolution. Today high throughput sequencing methods produce a large number of protein sequences that remain unreviewed or unannotated. Hence, there is a need for computational methods that can identify, analyze and annotate these proteins. Multiple sequence alignment strategy is a key precursor to this analysis of novel protein sequences.

1.3 Research Contribution

The contributions in this thesis are as follows:

1) Our work is to develop a multiple sequence alignment tool specific to transmembrane beta-barrel proteins. This requires the collection of a gold standard dataset of TMBBs, and the evaluation of the tool. It is crucial to construct a dataset of transmembrane beta-barrel proteins that can be used for further studies. For the purpose of this research, we collect all available TMBB protein sequences and clean them to construct a dataset of TMBB proteins. This dataset is curated from different protein databases and research papers.

2) For the first time, a TMBB specific multiple sequence alignment tool, called TMB-Coffee, is developed by adapting the T-Coffee [43] and PSI/TM-Coffee [2] programs. Our method uses a reduced database of TMBBs and constructs a library of TMBBs.

3) Evaluation of TMB-Coffee uses the dataset above. We use the NorMD [60] score to evaluate the quality of a multiple sequence alignment. We assess the reliability of the alignments produced by TMB-Coffee. We compare its performance with several general

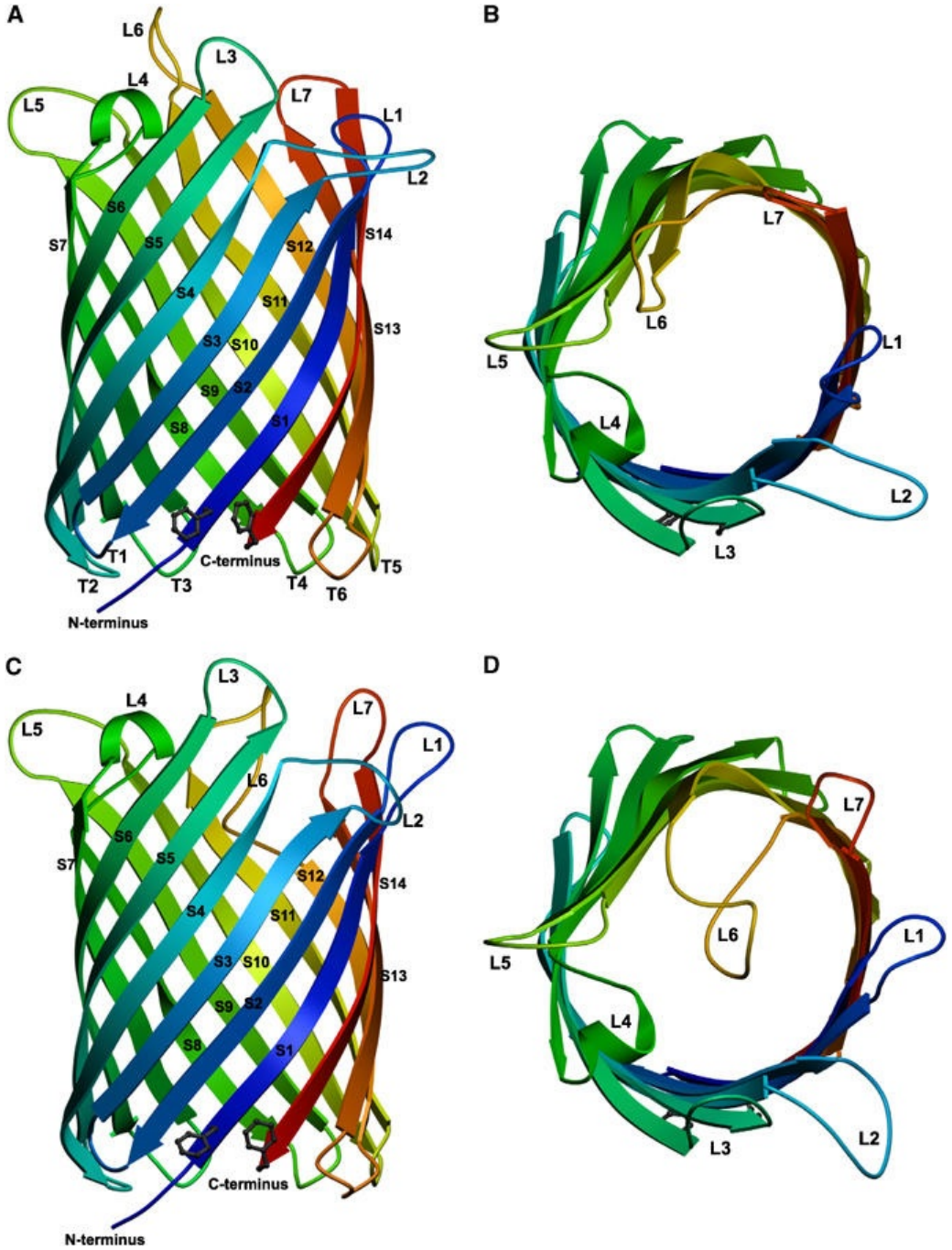

Figure 1.3: Beta Barrel Transmembrane Protein (OmpG). Overall structure of OmpG, showing the 14-stranded, antiparallel beta-barrel in the open (A, B) and closed (C, D) conformation viewed along the membrane (A, C) and from the extracellular side (B, D). The open and closed conformation of this protein may be used to allow and block substrate transport across the membrane. The beta-strands S1–S14 are rainbow-coloured starting from the N terminus (blue) to the C terminus (red) on the periplasmic side, where adjacent strands are connected by short turns T1–T6. Loops L1–L7 extend outward into the extracellular space in the open conformation, but L6 folds across the barrel entrance in the closed conformation. Reproduced from [70].

purpose MSA methods. We consider how well they align the beta-barrel regions.

1.4 Layout of Thesis

The layout of the remaining chapters in the thesis is as follows: Chapter 2 provides the necessary background information needed to understand the work done in the thesis. This includes the major tools and methods use to compare or construct MSAs and to evaluate them. Chapter 3 presents our work. Section 3.1 presents the construction of a beta barrel dataset. Section 3.2 presents the construction of local blast database used in TMB-Coffee and its evaluation. Section 3.3 presents the method used in TMB-Coffee and highlights its difference between T-Coffee and PSI/TM-Coffee. Section 3.4 presents the evaluation of TMB-Coffee. Section 3.6 and Section 3.7 presents the results and discussion, respectively. Chapter 4 concludes with contributions and limitations of the research. Appendix A contains supplementary information. It contains the dataset, and the scores of individual MSA. It also shows the version of tools used in this research and files uploaded to GitHub.

Chapter 2

Background

In this chapter we give a brief background about how proteins sequences are analysed and compared. We look into protein sequence alignment techniques, especially Multiple Sequence Alignment (MSA). We also look into different protein sequence alignment programs.

2.1 Protein sequence

Amino acids are the building blocks of proteins. Hence, proteins can be represented by a string of amino acid letter codes that follow the addition order of amino acids during protein formation. There are 20 different amino acids that appear in the genetic code. They can be represented by English letters. Table 2.1 shows list of amino acids. The abbreviations are based on IUPAC-IUB Joint Commission on Biochemical Nomenclature [34, 36].

For computational purposes, there are many file formats that can store protein information such as its sequence, structural topology, lineage etc. Fasta [33] is one such text based format that uses the letter codes to represent protein sequences. This format can also be used to represent DNA, RNA sequences and sequence alignments. Fasta format begins with a single-line description, followed by lines of sequence or alignment data. The description line is distinguished from the sequence data by a greater-than (>) symbol. For example, the protein (Uniprot ID- P0A910 [39]) can be represented in fasta format as shown in Figure 2.1. Here, "M" in the second line stands for start of the sequence, Methionine (amino acid),

Amino Acids	Three Letter Code	Single Letter Code
Glycine	GLY	G
Alanine	ALA	A
Valine	VAL	V
Leucine	LEU	L
IsoLeucine	ILE	I
Threonine	THR	T
Serine	SER	S
Methionine	MET	M
Cystein	CYS	C
Proline	PRO	P
Phenylalanine	PHE	F
Tyrosine	TYR	Y
Tryptophane	TRP	W
Histidine	HIS	H
Lysine	LYS	K
Argenine	ARG	R
Aspartate	ASP	D
Glutamate	GLU	E
Asparagine	ASN	N
Glutamine	GLN	Q

Table 2.1: Amino Acid Letter Code. Letter code follows IUPAC-IUB Commission on Biochemical Nomenclature. The nomenclature is based on the names of the amino acids.[34]

"I" stands for Isoleucine and so on.

```
>sp|P05695|PORP_PSEAE Porin P OS=Pseudomonas aeruginosa
MIRRHSCKGVGSSVAWSLLGLAISAQSLAGTVTTDGADIVIKTKGGLEVATTDKEFSFKL
GGRLQADYGRFDGYYTNNGNTADAAYFRRAYLEFGGTAYRDWKYQINYDLSRNVGNDSAG
YFDEASVTYTGFNPVNLKFGRFYTDFGLEKATSSKWVTALERNLTYDIADWVNDNVGTGI
QASSVVGGMAFLSGSVFSENNNDTDGDSVKRYNLRGVFAPLHEPGNVVHLGLQYAYRDLE
DSAVDTRIRPRMGMRGVSTNGGNDAGSNGNRGLFGGSSAVEGLWKDDSVWGLEGAWALGA
FSAQAEYLRRTVKAERDREDLKASGYYAQLAYTLTGEPRLYKLDGAKFDTIKPENKEIGA
WELFYRYDSIKVEDDNIVVDSATREVGDAKGKTHTLGVNWYANEAVKVSANYVKAKTDKI
SNANGDDSGDGLVMRLQYVF
```

Figure 2.1: Protein Sequence in Fasta format. Fasta is a text based format used to represent biological sequences like nucleotide and amino acid sequences. The description line is distinguished from the sequence data by a greater-than (>) symbol in the first column. This is followed by the sequence starting in the next line [33].

2.2 Protein Sequence Analysis

Protein sequence analysis [14] are experimental and computational methods to understand and annotate the function, structure and evolution of protein. Here, we focus on

computational methods of sequence alignment. High-throughput production of protein sequences is contributes to an ever increasing collection of sequences. These sequences before being annotated does not provide an understanding of the proteins. Hence, it is necessary to analyse and understand these sequences. There are many methods for sequence analysis; sequence alignment, profile comparison, sequence assembly, etc.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Figure 2.2: Sequence Alignment. Alignment of two protein sequences "AAB24882" and "AAB24881" using ClustalW [58]. The protein IDs are from the OMP database [63]. This figure displays only a section of the alignment that shows conserved regions in the sequences. This section shows how gaps are introduced in the 98 residues from "AAB24881" to align it with 116 residues from "AAB24882".

2.3 Sequence Alignment

Sequence alignment is a method of arranging the residues in different sequences with respect to each other. This method is used for understanding and comparing the relationship between the sequences in consideration to finally identify function, structure etc. Figure 2.2 shows how a sequence alignment looks like in general. It shows the accession id and the aligned sequences. This sequence alignment is performed by ClustalW [58] program. Sequences can be aligned either by performing *local alignment* [53] or *global alignment* [42]. Global alignment tries to align every residue of every sequence. It is an end to end alignment, whereas, local alignment aligns sub-string of the sequences. Figure 2.3 shows global and local alignment. When alignment of two sequences is performed such as in global and local alignment, it is called *pairwise alignment*.

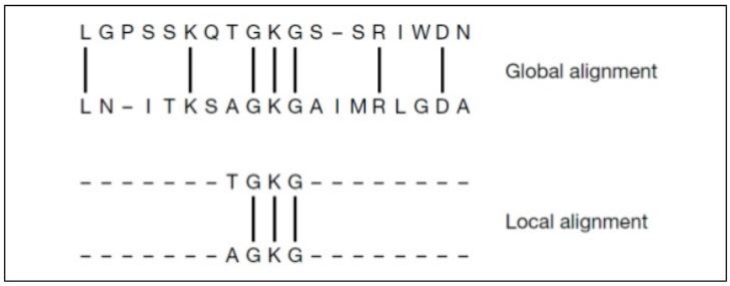

Figure 2.3: Global and Local Alignment. Global alignment shows how all residues in first sequence is aligned to all residues in second sequence. Whereas, local alignment only aligns a portion of the second sequence to the first [53, 42].

2.4 Multiple Sequence Alignment

Unlike pairwise alignment, the alignment of three or more biological sequences is called *multiple sequence alignment* (MSA) [15, 9]. MSA can be obtained for nucleotide as well as protein sequences. However in this work, we use only protein sequences. MSA is the most robust methods for identifying evolutionary information from these sequences. It is the prerequisite to several studies like secondary structure prediction, homology modelling, phylogenetic analysis etc. For illustration, MSA of three beta-barrel proteins is constructed using Clustal Omega [51]. A section of sequences in the MSA of three transmembrane beta barrel proteins is shown in Figure 2.4. Alpha-numeric key is used to distinguish the sequences, " $-$ " represents gap, " $*$ " represents fully conserved residues, " : " represents highly similar residues and "." represents low similarity.

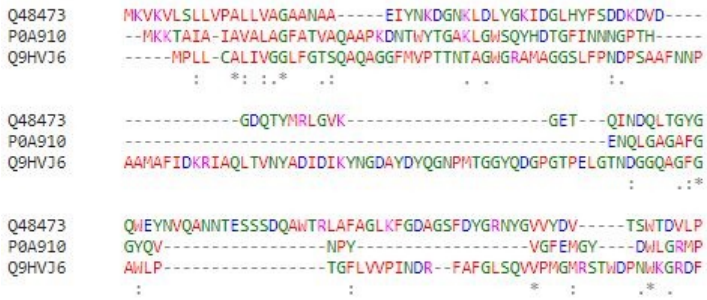

Figure 2.4: Sample Multiple Sequence Alignment. This figure shows the alignment of "Q48473", "P0A910" and "Q9HVJ6" with each other. This alignment is produced from Clustal Omega [51] program. "Q48473", "P0A910" and "Q9HVJ6" are Uniprot IDs [39].

In Equation 1, S is a set of m sequences where m is a whole number and $m \geq 3$. i.e. S_i where i=1,2,...,m.

$$S := \left\{ \begin{array}{c} (\mathrm{S}_1 = S_{11}, S_{12}, ..., S_{1n_1}), \\ (\mathrm{S}_2 = S_{21}, S_{22}, ..., S_{2n_2}), \\ ..., \\ (\mathrm{S}_m = S_{m1}, S_{m2}, ..., S_{mn_m}) \end{array} \right\} \tag{1}$$

$$S' := \left\{ \begin{array}{c} (\mathrm{S}'_1 = S'_{11}, S'_{12}, ..., S'_{1L}), \\ (\mathrm{S}'_2 = S'_{21}, S'_{22}, ..., S'_{2L}), \\ ..., \\ (\mathrm{S}'_m = S'_{r1}, S'_{r2}, ..., S'_{rL}) \end{array} \right\} \tag{2}$$

Then, an MSA, given in Equation 2 is produced by inserting gaps into sequences S_i in S so that the modified sequences S'_i in set S' conform to length L. $L \geq max\{x_i | i = 1, ..., r\}$ and no column in sequences of S can contain only gaps. Thus, removing all gaps from sequence S'_i gives sequence S_i.

MSA can be computed either by performing global or local alignments and can be an extension of pairwise alignment. Whichever technique is employed, the aim is to get the best alignment possible. Although there is no single best technique to finding the best alignment [5], it is non trivial to find alignment with biological correctness. MSAs should aim to maximise matches while permitting minimum gaps. Since more gaps make non homologous sequences to be aligned, it is undesired.

The most commonly studied method that computes score for characters in the alignment column is Sum-of-Pairs (SP) [9]. Consider the multiple sequence alignment S' of S, then

$$SP(a_1, ..., a_m) = \sum_{1 \leq i < j \leq m} \delta(a_i, a_j) \tag{3}$$

where, a_i can be any character or a space in a column of the alignment; $\delta(a_i, a_j)$ is a value

assigned to matches or mismatches or insertions or deletions, a numeric value. The SP score of S' is given in Equation 4.

$$SP(S') = \sum_{x\ column} SP(S'_1[x], ..., S'_m[x]) \tag{4}$$

The SP score is computed by assigning numeric values to matches, mismatches, insertions or deletions. Matches can be given positive values and mismatches, insertions, deletions can be penalized. Insertion/deletion can be further penalized when gaps are opened or extended (gap open and gap extension penalty) depending on the objective function used. Hence, points assigned to mismatches, matches, gaps and gap extensions is decided based on the biological significance and the scoring functions used. Ultimately, it has to provide the optimal alignment. The SP score is not appropriate when many sequence fragments lead to an MSA with columns that are not complete [40]. The following example shows how the SP score is calculated. We now consider 4 sequences that are aligned as given below,

S_1	=	A	L	G	-	-	G	A	G	A
S_2	=	-	L	G	V	V	G	A	L	-
S_3	=	A	L	-	V	-	G	A	-	A
S_4	=	L	L	G	V	V	L	A	L	-

We can look at the aligned sequences in the form of a matrix. We assume that a match = 1, a mismatch = -1, and a gap = -2. Then, Sum-of-Pairs score for the first column in the MSA is calculated as:

$$\begin{aligned} SP(A, -, A, L) &= \delta(A, -) + \delta(A, L) + \delta(A, A) + \delta(L, -) \\ &= -2 + 1 + -1 + -2 + -2 + -1 \\ &= -7 \\ SP(S_1 - S_4) &= -7 + 6 + -3 + -3 + -7 + 0 + 6 + -7 + -7 \\ &= -22 \end{aligned}$$

There are 9 columns in the alignment above. Note that δ(-,-) is given a value equal to 0.

This scoring scheme is applied to all such possible combinations while looking for optimal

multiple sequence alignment of a set of sequences. An alignment with higher score is considered to have a better alignment. However, it becomes complex when the number of sequences and length of sequences increase. In the last three decades, many MSA techniques were developed. In practise, MSA nowadays use heuristics to get optimal alignment. The following techniques are popular ones:

- Dynamic Programming
- Center star method
- Progressive alignment
- Iterative method

2.4.1 Dynamic Programming

Generally dynamic programming [41] is used for global pairwise alignment but can be used to perform alignment on multiple sequences. For aligning protein sequences, the parameters used are gap penalty and substitution matrix. Dynamic programming requires construction of n dimensional matrix for n sequences. Thus, the search space grows and the time to compute it grows exponentially with size of sequence and increase in number of sequences. Dynamic programming can described as the following

$$Align(S_i^1, S_j^2) = max \begin{cases} Align(S_{i-1}^1, S_{j-1}^2) + s[a_i, a_j] \\ Align(S_{i-1}^1, S_j^2) - g \\ Align(S_i^1, S_{j-1}^2) - g \end{cases} \tag{5}$$

where, S^1,S^2,...S^k is a set of sequences; S_i^1, S_j^2 are residue of S^1 and S^2 respectively; $s(a_i, a_j)$ is the weight residue a_i and a_j from a substitution matrix and g is the gap penalty (penalty applied when gaps occur in alignment).

2.4.2 Center star method

Optimal multiple sequence alignment using dynamic programming takes exponential time. The center star method [24] employs steps that can minimize SP distance score, and can compute alignments in polynomial time. The steps are as follows; (i) Find string S_c in the sequence and compute pairwise alignments. (ii) Convert pairwise alignment to MSA. The Figure 2.5 shows how pairwise alignment is converted to MSA in center star method.

S_1	=	M	L	G	-	-	G	A	G	A
S_2	=	-	L	G	V	V	G	A	C	-

S_1	=	M	L	G	-	-	G	A	G	A
S_3	=	M	L	-	V	-	G	A	-	A

S_1	=	M	L	G	-	-	G	A	G	A
S_2	=	-	L	G	V	V	G	A	C	-
S_3	=	M	L	-	V	-	G	A	-	A

S_1	=	M	L	G	-	-	G	A	G	A
S_4	=	L	L	G	V	V	C	A	C	-

S_1	=	M	L	G	-	-	G	A	G	A
S_2	=	-	L	G	V	V	G	A	C	-
S_3	=	M	L	-	V	-	G	A	-	A
S_4	=	L	L	G	V	V	C	A	C	-

Figure 2.5: Center Star Method. Pairwise alignments S_1,S_2 and S_1,S_3 is converted to MSA S_1,S_2,S_3. Then, pairwise alignment of S_1,S_4 and MSA S_1,S_2,S_3 is converted to MSA S_1,S_2,S_3,S_4.

2.4.3 Progressive alignment

Progressive methods are based on tree or hierarchical methods that are heuristic. The first progressive method of sequence alignment was developed by Da Fei Feng and Russell F. Doolittle [18]. The idea of this technique is to align the most related or groups of related sequences first and follow through less related ones. It depends on initial pairwise alignments. This technique uses guide trees to produce the alignments in the order that the tree grows. Clustering methods such as UPGMA [55] or Neighbor-Joining [48] are used to produce the guide trees. This method employs modified scoring matrices based on a weighting scheme so that the alignments are not picked randomly. They are not optimal

global alignments because when errors are made, they cannot be fixed in the later stages and this might influence the alignment. Generally, this technique to not appropriate for distantly related sequences but modern applications have fixed that problem (For example T-Coffee [43]). The Clustal family of alignment tools [11] also employ progressive methods. Progressive alignments are usually faster than other methods.

2.4.4 Iterative method

Iterative methods are like progressive alignment but rectifies its drawbacks. The idea of iterative method is to produce progressive alignment and then fix the errors in the initial alignment iteratively. So, the goal here is to improve the alignment score by maximising the objective function used. In this method, whole pairwise alignments produced in the beginning or sections of it can be revisited and modified to get a better score. This process makes it iterative. While progressive alignment methods compromise on accuracy and improve efficiency, iterative methods are the opposite. Two programs based on this method are PRRN/PRRP [22] and DIALIGN [37].

2.5 Substitution/Scoring Matrices

During evolution, biological sequences (protein or nucleotide) undergo mutation; i.e gradual changes in the sequence. A substitution matrix is a matrix of log-odds of the probability of change in biological sequences. It is represented in a 20x20 matrix of log-odds indexed by the amino acid letter. They are used to score the alignment of residues in a column. Scoring matrices should consider the physio-chemical properties of the amino acid and the probability that the amino acid is substituted. A positive score is given when the frequency of change in amino acid residue is less likely to be by random chance and a zero score is given when the frequency is equal to that expected by chance. A negative score is given when the frequency is less to that expected by chance. There are several variants of substitution matrices [25, 12]. The mathematical definition of a general substitution matrix is as follows:

$$S_{ij} = \log \frac{p_i.K_{i,j}}{p_i.p_j} = \log \frac{K_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}} \tag{6}$$

where $K_{i,j}$ is probability that amino acid i transforms into amino acid j, and p_i, p_j are the frequencies of amino acids i and j. There are several variants of substitution matrices most of which are a series of matrices; PAM [12], BLOSUM [25], GONNET (GON) [21].

2.5.1 BLOSUM (Block Substitution Matrix)

BLOSUM [25] is a series of matrices derived from gapless alignments of sequences that are distantly related. Different BLOSUM matrices are BLOSUM45, BLOSUM62, BLOSUM80. Numbers are added to the end of the matrix name and refers to the minimum sequence identity. Greater numbers mean closer distance between sequences. Possibility of errors in these matrices arise from the alignments that are erroneous.

2.5.2 PAM (Point Accepted Mutation)

PAM matrices [12] are substitution matrices based on global alignment of closely related sequences and their evolutionary model with divergence close to 15%. Unlike BLOSUM, PAM extrapolates the evolutionary distance to create more matrices. Higher the number, greater the evolutionary distance. Since PAM matrices are extrapolated, errors get scaled when dealing with PAM for higher distance. Different PAM matrices are PAM250, PAM160, PAM120 etc.

2.5.3 GON (GONNET)

The Gonnet matrix [21] is a substitution matrix based on pairwise alignments of sequences from protein databases. It uses classical distance matrices to produce an alignment of proteins. The alignment is used to produce a new distance matrix to refine the alignment, iteratively. The Gonnet matrix is normalized to PAM250, hence, they suggest its usage in conjunction with PAM matrices. Different Gonnet matrices are GON200, GON250, GON350 etc.

BLOSUM, PAM and GONNET (GON) matrices have different approaches to construction, yet they are comparable based on whether they are designed for proteins that are closely, distantly related or somewhere in between. Table 2.3 shows comparable BLOSUM, PAM and GON matrices [38].

BLOSUM	**PAM**	**GON**
BLOSUM45	PAM250	GON200
BLOSUM62	PAM160	GON250
BLOSUM80	PAM120	GON350

Table 2.3: Comparable BLOSUM, PAM and GON matrices which are similar [26].

2.6 Alignment Tools

2.6.1 Blast

Sequence homology is the presence of shared ancestry between sequences. This is important in identifying previously characterized sequences, finding phylogenetically related sequences, identifying possible functions based on similarities to known sequences etc. Blast [1] is a fast program used to search for such homology from databases. Hence, it uses substitution matrices and gap penalties to compare a query sequence with a database of sequences. The algorithm is: (i) Low complexity regions or sequences with repeats are removed so that it does not influence the scoring system to produce high score for insignificant sequences from the database. (ii) Create words from query sequence (typically word size is 3 for protein sequence) and scan the database for these words (called a 'hit'). For each word, high scoring matches are produced from the database. (iii) All high scoring words are categorized for later comparisons. (iv) A High-scoring Segment Pair (HSP) [71] is a local alignment with no gaps that achieves one of the highest alignment scores in a given search. The matches are extended to the left and right of the word in the sequence to form HSPs until the score starts to reduce. (v) All HSPs that have a high score are evaluated to determine

the significance in terms of e-value. (vi) Smith-Waterman local alignment [53] between the query and the match is produced for all matches that are beyond the e-value threshold.

Blast can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

2.6.2 Clustal Omega

Clustal Omega [51] is one of the sequence alignment tools in the Clustal family [11]. It is used for computing multiple sequence alignment of nucleotide or amino acid sequences The algorithm of Clustal Omega starts with computing pairwise alignment and then calculating distance matrices using UPGMA [55] /Neighbor-Joining methods [48]. Guide trees produced from this are used to compute the final multiple sequence alignment. Just like the other tools, this is a progressive alignment. Clustal Omega implements a modified version of mBed [8] and uses HHalign [54] for computing its pairwise alignment.

Clustal Omega be downloaded from http://www.clustal.org/omega/.

2.6.3 MAFFT

MAFFT [28] is a method for rapid multiple sequence alignment based on fast Fourier transform. It can be used to generate multiple sequence alignment of nucleotide or amino acid sequences. For the alignment, each amino acid is converted into a vector of two values, volume and polarity. After the algorithm calculates the relationship between the sequences, it is transformed into fourier signal information. MAFFT applies standard dynamic programming and optimally rearranges homologous regions in the sequences. It also implements an iterative approach. There are three modes of MAFFT: (i) progressive method, (ii) iterative refinement with Weighted Sum-of-Pairs score (WSP) [23], (iii) iterative refinement with WSP and consistency score [44].

MAFFT can be downloaded from https://mafft.cbrc.jp/alignment/software/.

2.6.4 MUSCLE

MUSCLE [16] is a multiple sequence alignment tool for proteins that computes progressive alignment of the sequences and then refines it. The stages of the algorithm are as follows: (i) Computing k-mer distance matrix of each pair of input sequence and then clustering the matrix with UPGMA [55] to produce a binary tree. (ii) A progressive MSA is computed from this tree. (iii) For refining the multiple sequence alignment, the new tree from stage (ii) is divided in to sub trees and a profile is generated for each tree. This stage involves improving the alignment produced in the previous stage. Here, Kimura distance [29] is used to re-estimate the tree from previous stage to create a new one and uses the multiple sequence alignment as input. An optimized alignment is computed using progressive alignment for sub branches that have changed in the new tree relative to the previous tree. Finally, the profiles are aligned to check if the SP score has improved, if not the alignment is discarded. The steps are repeated until a user defined threshold or convergence. Figure 2.6 shows the overview of MUSCLE.

MUSCLE can be downloaded from `https://www.drive5.com/muscle/`.

2.6.5 T-Coffee

T-Coffee[43], which stands for "Tree-based Consistency Objective Function for Alignment Evaluation" is an MSA method that uses progressive alignment in a broader sense. Progressive alignment tend to be greedy in nature and mistakes made during the beginning of the step by step alignment cannot be rectified during later steps. Although, T-Coffee uses the progressive alignment strategy, it minimizes the drawbacks of such algorithm. T-Coffee generates pairwise alignments and these are used to generate a library of alignments. This library is further used to produce the final MSA. The pairwise alignments help in determining how well the sequences align with each other. Thus, it considers the relationship between all the sequences and not just the sequence that is being currently aligned.

T-Coffee algorithm can be divided into three parts: (1) Pairwise Alignment; (2) T-Coffee

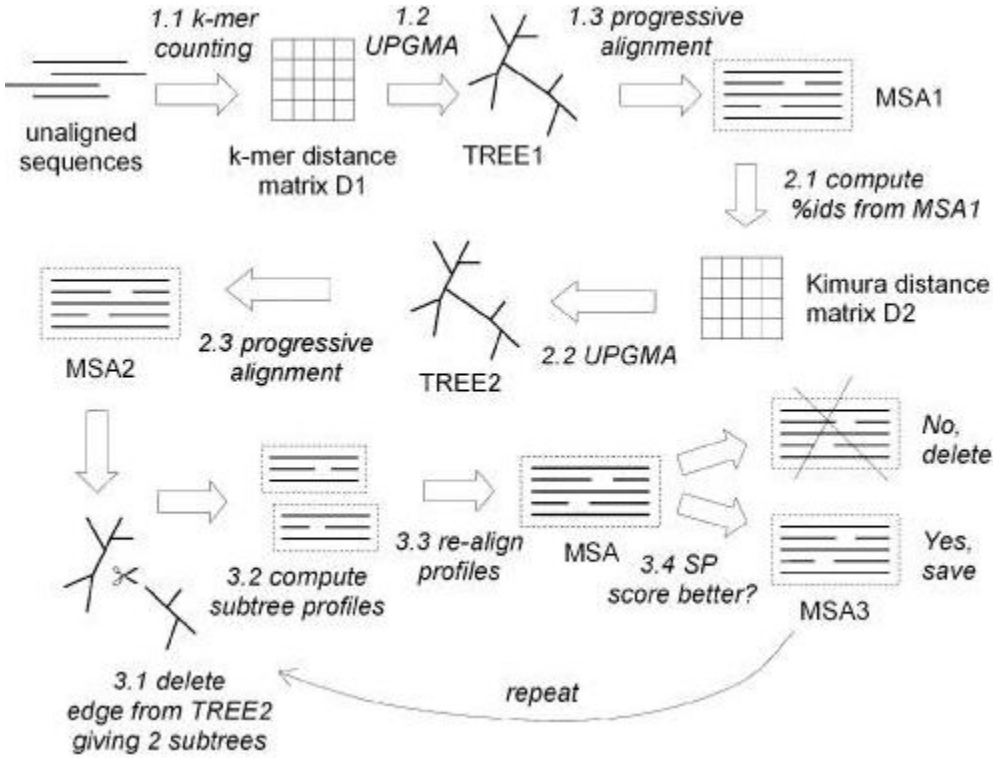

Figure 2.6: Overview of MUSCLE program. The figure summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple sequence alignment is available at the completion of each stage, at which point the algorithm may terminate. [16].

Library Construction is used as a position specific scoring scheme in the next part; (3) T-Coffee Progressive Alignment using the library. The library is a set of pairwise alignments that contains the weight of all possible matches of residues in those alignments. This library is used while multiply aligning sequences in the progressive alignment step. Figure 2.7 shows overview of T-Coffee method.

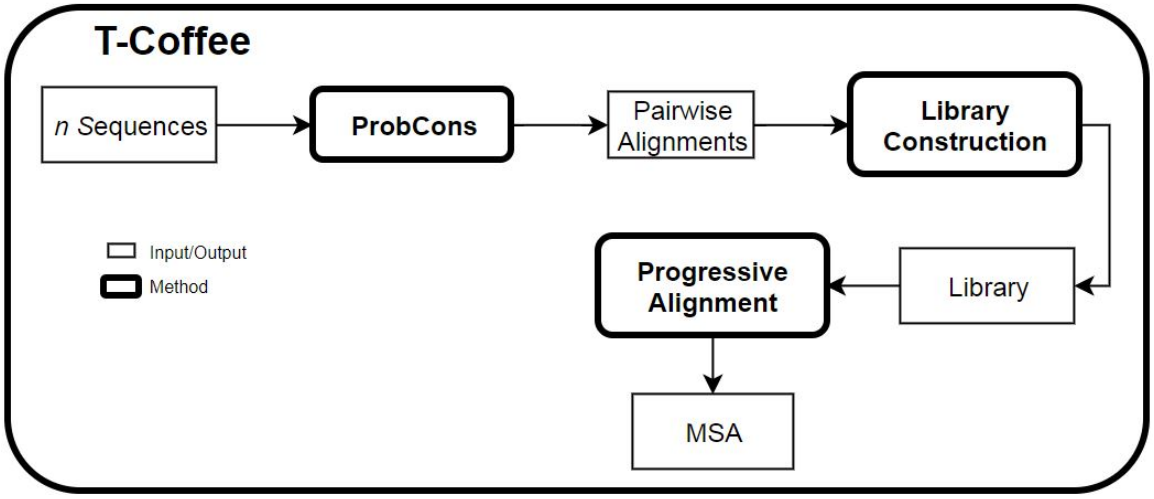

Figure 2.7: Overview of T-Coffee Alignment Program. The figure shows the main steps required to compute a multiple sequence alignment using the T-Coffee[43] method. Square blocks indicate input or output while rounded boxes indicate the method used.

(1) Pairwise Alignment

The algorithm starts with taking the sequences to be aligned as input and produces their pairwise alignments. It produces pairwise alignments of all pairs of sequences. The original T-Coffee uses a mix of ClustalW and Lalign pairwise alignment programs to compute global and local pairwise alignments respectively. However, a variety of pairwise alignment tools can be incorporated into T-Coffee package. T-Coffee is very flexible in terms of the methods that an be used along with it. The new default pairwise alignment program however is adapted from the Probcons package [13] and is called proba_pair. For the sake of simplicity we consider certain elements in the algorithm and has been described in Table 2.4. It takes a list of input sequences called "seqList". For each pair of sequences in this list, pairwise alignments are computed using proba_pair (we call it "probcons_pair") and stored

in “Lalignments”. Lalignments stores a list of all pairwise alignments (pairwise alignment from all pairs of input sequences).

(2) T-Coffee Library Construction

Step (2) can be further divided into two parts, construction of (a) Primary Library and (b) Extension of Primary Library. A weight “w” is computed for each pairwise alignment. The weight “w” is calculated as $weight = |(match * 100)/(ungapped\_alignment)|$. Thus, weight is the absolute value of percentage identity in the pairwise alignment. Percentage identity is used as weight because it is known to be an indicator for the accuracy of sequence alignment. Algorithm 2 shows the construction of T-Coffee library. To compute the weight, “Lmatches” contains all residue-residue matches for the pairwise alignment and we store the number of matches in “nbMatches”. The function count() finds the total number of residue-residue matches in “Lmatches” for an alignment. The function compute_ungapedAlignment() counts the number of columns in the alignment without gaps in them. The primary library consists of pairwise alignments, the weights of the pairwise alignments and the matched residues.

T-Coffee uses a triplet approach for extending the library. The triplet approach is performed to combine some of the information contained in the whole primary library such that the final weight for a pair of residues in a pairwise alignment contains weight from other pairs of residues. In the triplet approach, each alignment is considered with a triplet of sequences. Figure 2.8 shows how the T-Coffee library is computed using weights and shows its extension. For example: In Figure 2.8, a) shows alignment of residues SeqA(G) and SeqB(G) in the word GARFIELD and the weight computed for it (Prim. Weight=88). Let us call this weight pw.

In Figure 2.8, b) we see that the residues SeqA(G) and SeqC(G) are aligned, as well as, residues SeqC(G) and SeqB(G). Thus, there is an alignment of residues SeqA(G) and SeqB(G) through SeqC. So, the weight of the triplet is $min(w_1, w_2)$ where w_1 is w(SeqA(G),SeqC(G)) and w_2 is w(SeqC(G),SeqB(G)). Let us call this weight ew, where,

$ew = min(77, 100)$.Assuming this triplet, the weight of SeqA(G) and SeqB(G) in the extended library is updated to $Sum(pw, ew)$=88+77=165. In the complete extended library such triplets are considered for SeqA(G) and SeqB(G) through all other sequences. In summary, the weight associated with a pair of residues is the sum of all the weights gathered through the examination of all the triplets involving that pair. This is computed for each pair of residues of SeqA and SeqB. The steps are carried out for all such pairs of sequences [43].

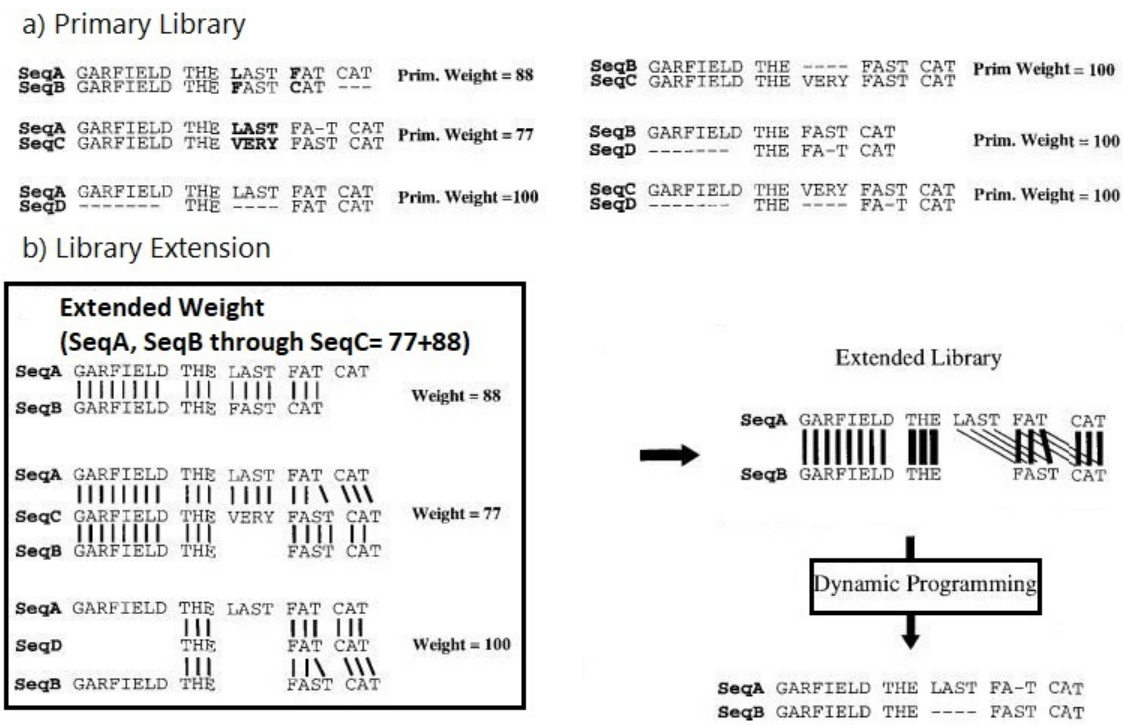

Figure 2.8: T-Coffee Library Construction. This picture is taken from the original T-Coffee paper [43]. a) Primary Library: If pairs of sequences are aligned as shown here, the weight can be calculated as absolute value of (100*(number of identical residues/total number of ungaped columns within the complete alignment). i.e For SeqA and SeqB, number of identical residues is 16 and ungapped columns is 18. Therefore, weight is 88. b) Extended Library: It takes each aligned residue pair from the library and check the alignment of the two residues with residues from the remaining sequences. For example; Consider alignment of residues in SeqA and SeqB through SeqC. The new weight is calculated as (min((weight of SeqA and SeqC),(weight of SeqB and SeqC))+(original weight of SeqA and SeqB)). i.e (min(77,100)+88)=165. In final library, the weight for residues in SeqA, SeqB through all other sequences are added up. The more intermediate sequences supporting an aligned pair of residues, the higher its weight. Note: Depending on the pairwise alignment program used, the weights may differ.

(3) T-Coffee Progressive Alignment

While performing progressive alignment with T-Coffee, the extended library is used as a position specific scoring scheme. Hence, T-Coffee uses this scheme instead of BLOSUM or PAM matrices for constructing the multiple sequence alignment. In progressive alignment, the pairwise alignments are made to produce distance trees. This distance tree is used to construct a new neighbor-joining distance matrix. This is used to construct a guide tree. The guide tree is used to produce multiple sequence alignments. Algorithm 1 shows T-Coffee program, Algorithm 2 shows how the T-Coffee library is constructed. Figure 2.9 shows the steps in function T-Coffee_progressiveAlignment( ) in Algorithm 3

Element	Description	Used In
Alignment	It is an object type to store pairwise alignment	Algorithm1; Line1
Lalignment	List to store pairwise alignment	Algorithm1; Line2
seqList	List of protein sequences	Algorithm1; Line3
probcons_pairwise( )	Sub routine adapted from Probcons package. It is used to produce the pairwise alignment of two sequences seq_1, seq_2	Algorithm1; Line5
p_align	Variable of type list to store pairwise alignment for probcons_pairwise	Algorithm1; Line5
store_pairwise _alignment( )	Sub routine used to store pairwise alignment in p_align to the Alignment object	Algorithm1; Line6
aln	Instance of the Alignment object	Algorithm1; Line7
count( )	Sub routine used to count the number of matches in pairwise alignment. It takes list of matches Lmatches in the alignment object as input	Algorithm2; Line3
nbMatches	Variable to store number of matches	Algorithm2; Line3
compute_ ungappedAlignment( )	Sub routine to remove gaps in the alignment. i.e It removed gapped columns in the alignment	Algorithm2; Line4
ungapped_align	Variable to store number of ungapped alignments	Algorithm2; Line4
get_pairwise( )	Sub routine to collect pairwise alignment of particular sequences from instance "aln"	Algorithm2; Line13,Line14
LB_{20x20}	Matrix to store the weights of pairwise aligments of all residue pairs. Matrix is of size 20x20 because there are 20 amino acid residues	Algorithm2; Line23
r1 and r2	Residues in the sequences for which matches have been identified and weighted	Algorithm2; Line26
D[i,j]	Distance matrix D contains distances of pairwise alignments in Lalignment.	Algorithm3; Line1
getnbMatches( ) countSeq( )	Sub routine to get number of matches Sub routine to count the number of sequences in seqList	Algorithm3; Line3
n	Integer n stores the number of sequences in seqList	Algorithm3; Line3
neighborJoining( )	Sub routine to produce guide tree T using Neighbor-Joining method. It takes distance matrix and number of sequences as input	Algorithm3; Line4

Table 2.4 continued from previous page

Element	Description	Used In
T	stores guide tree produced by Neighbor Joining method	Algorithm3; Line4
chooseclosestSeq( )	Sub routine to choose the closest two sequence in the guide tree T	Algorithm3; Line6
A_1, A_2	Variables to store closest two sequences in guide tree T	Algorithm3; Line6
Align(A_1, A_2)	Sub routine is a Dynamic Programming method to compute alignments. It is explained in Equation 4. Here however, the T-Coffee library LB[a_i,a_j] is used instead of substitution matrix and gap penalty is kept zero.	Algorithm3; Line7
A_n	stores the alignment produced by A(A_1,A_2) and the alignment grows as it proceeds (more and more sequences get aligned). It grows to become the MSA	Algorithm3; Line7
root	root of guide tree T	Algorithm3; Line9
Blastp( )	Sub routines that performs protein Blast with the parameters described in the description of Algorithm. It takes input list of sequences and database	Algorithm4; Line4 Algorithm5: Line4
gappedHomologseq	Variable to store list of homologous sequences (i.e Blast hits or result)	Algorithm4; Line4 Algorithm5: Line4
add( )	Adds homologous sequences to the query sequence. This step is described in figure 2.10	Algorithm4; Line5 Algorithm5; Line5
preProfile_seqs	Variable to store the list of seq and its homologs	Algorithm4; Line5 Algorithm5; Line5
MpreProfile_seqs	matrix to store seq and its homologs	Algorithm4; Line6 Algorithm5; Line6
PseqProfile	Variable to store list of profiles	Algorithm4,5; Line20, Line22
pairofProfile	Variable to store list of pairwise alignments	Algorithm4; Line24 Algorithm5; Line24
storepairwiseAlignment( )	Sub routine to store pairwise alignments to alignment object. Takes input a pairwise alignment	Algorithm4; Line25 Algorithm5; Line25
hmmtop( )	Sub routine hmmtop to predict transmembrane topology. Takes input sequence list, profile of that sequence	Algorithm4; Line31
topol	Variable to store topology of seq from seqList	Algorithm4; Line31
updateMSA( )	Sub routine to color code MSA according to the predicted topology. It takes input MSA and topology	Algorithm4; Line32

Table 2.4: Elements of the Algorithms

Using the T-Coffee package

After installing the T-Coffee package, the following command assumes the default parameters for the input *sequencefile*.

```
>t_coffee sequencefile
```

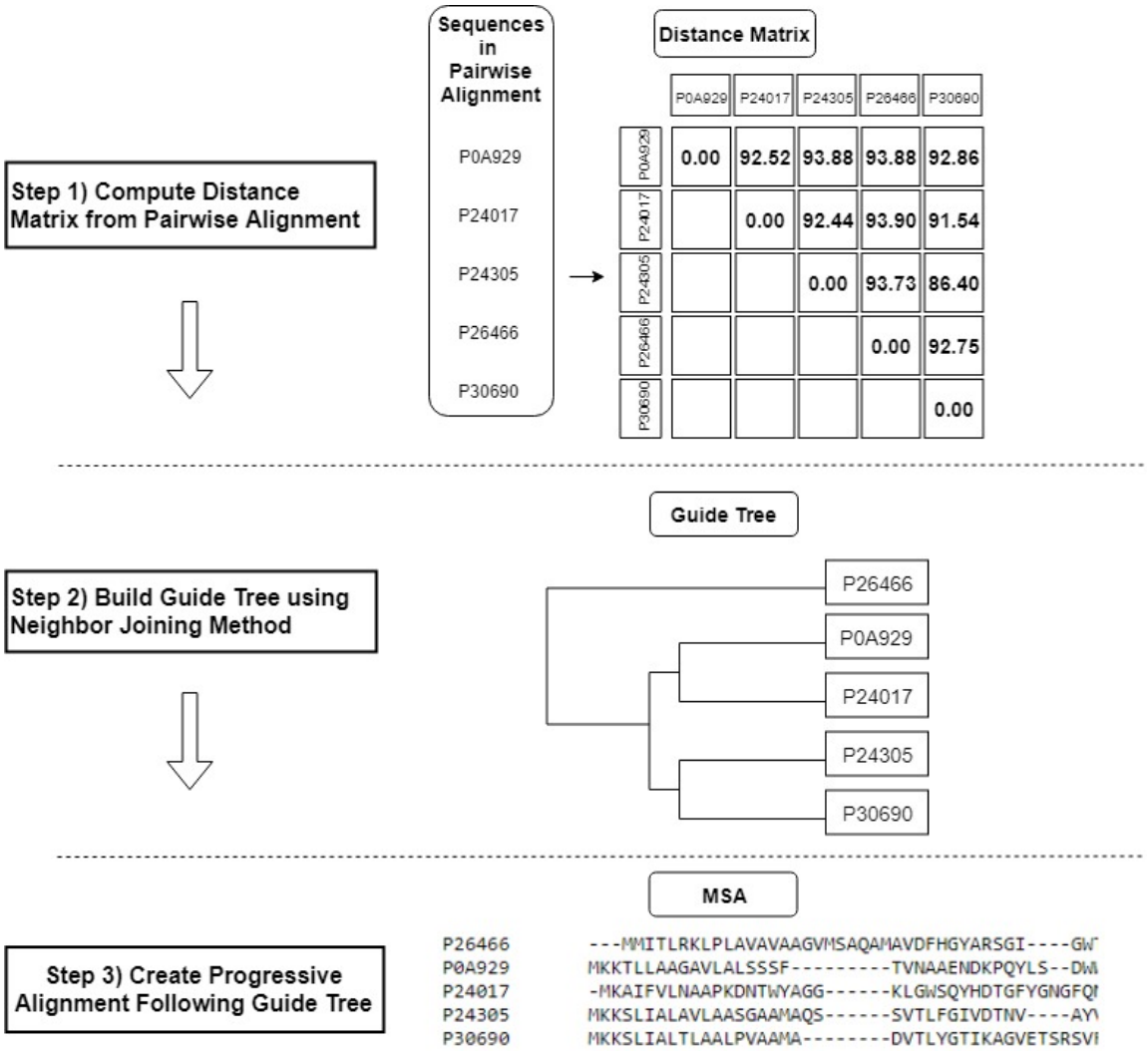

Figure 2.9: T-Coffee Progressive Alignment Steps. Step (1) shows distance matrix D for pairwise alignments from 5 different sequences. Step (2) shows guide tree produced for those 5 sequences based on the distance matrix D. Here we see that "P0A929" and "P24017" are closely related. So is "P24305" and "P30690". "P26466" is the distantly related sequence. Guide tree T is a binary tree. Step (3) shows the final multiple sequence alignment.

Algorithm 1 T-Coffee

Input: list of sequences *seqList*
Output: Multiple sequence alignment M

```
Creating an object type "Alignment" to store the pairwise alignment
Initializing the list of pairwise alignment, Lalignment := null
for seq1 in seqList do
    for seq2 in seqList do
        p_align := probcons_pairwise(seq1, seq2)
        # Storing pairwise alignment "p_align" in the Alignment object "aln" using sub routine "store_pairwiseAlignment()"
        aln := store_pairwiseAlignment(p_align)
        Lalignment := Lalignment.append(aln)
    end for
end for
LB, Lalignment := T-Coffee_Library(seqList,Lalignment)
M := T-Coffee_progressiveAlignment(seqList, LB, Lalignment)
return M
```

Parameters can specified by adding flags. The web server of T-Coffee uses the following command and it can be used for the local installation of T-Coffee as well.

```
>t_coffee -in=sequencefile -mode=regular -output=score_html clustalw_aln \
fasta_aln score_ascii phylip -maxnseq=150 -maxlen=10000 -case=upper \
-seqnos=off -outorder=input -run_name=result -multi_core=4 -quiet=stdout
```

T-Coffee can be downloaded from http://www.T-Coffee.org/Packages/Stable/Latest/.

2.6.6 TM-Coffee

TM-Coffee [19, 10] is a multiple sequence alignment method for alpha helical transmembrane protein sequences. This method combines homology extension and consistency-based progressive approach from T-Coffee to optimally align the multiple sequences. It uses PSI-Coffee mode of the T-Coffee package and performs Blast on a reduced database of transmembrane proteins for homology extension. The database used for TM-Coffee contains transmembrane helices and transmembrane helical segments. Thus the library that is produced contain these transmembrane elements and supports the progressive alignment

Algorithm 2 T-Coffee_Library

Input: list of sequences *seqList*, list of pairwise alignments *Lalignment*
Output: library *LB*, list of pairwise alignments *Lalignment*

Initializing the list of matches, *Lmatches* := null
for *aln* in *Lalignment* **do**
    nbMatches := count(*aln.Lmatches*)
    ungapped_align := compute_ungappedAlignment(*aln.p_align*)
    Computing weight w of pairwise alignment as:

$$w := \left| \frac{(nbMatches * 100)}{ungapped\_align} \right|$$

    # Update the list of matches
    for *match* in *aln.Lmatches* **do**
        match.w := w
    end for
end for
    # Construct library extension
for *aln* in *Lalignment* **do**
    #Collect all sequences not in pairwise alignment; store them in "LotherSeq"
    Lotherseq := *seqList* \ {$aln.seq_1$, $aln.seq_2$}
    for *seq* in *LotherSeq* **do**
    # Collect the pairwise alignments from Lalignment
        aln_1 := get_pairwise($aln.seq_1, seq, Lalignment$)
        aln_2 := get_pairwise($aln.seq_2, seq, Lalignment$)
    end for
    # Update the weight of current pairwise alignment
    $aln.w := min(aln_1.w + aln_2.w) + aln.w$
    for *match* in aln.*Lmatches* **do**
        if match in ($aln_1.Lmatches$) OR match in ($aln_2.Lmatches$) **then**
            $match.w := match.w + aln.w$
        end if
    end for
end for
    # Build library
Initilialize matrix LB_{20x20} to zero matrix.
for *aln* in *Lalignment* **do**
    for match in *aln.Lmatches* **do**
        $LB[match.r_1, match.r_2]$:= LB$[match.r_1, match.r_2]$ + $match.w$
    end for
end for
return *LB*, *Lalignment*

Algorithm 3 T-Coffee_progressiveAlignment

Input: list of sequences *seqList*, TCoffee library *LB*, list of pairwise alignment *Lalignment*

Output: Multiple sequence alignment M

1: **for** *aln* in *Lalignment* **do**

$$\text{Distance matrix D } [aln.seq_1, aln.seq_2] := 1 - \frac{get\_nbMatches(aln.Lmatches)}{min(|aln.Seq_1|, |aln.Seq_2|)}$$

2: **end for**

#Neighbor-Joining method produces a tree T

3: T := neighborJoining(D,countSeq(seqList))

4: **do**

5: A_1, A_2 := choose_closestSeq(T)

6: A_n := Align(A_1, A_2)

7: Update T with A_n

8: **while** not root(T)

9: $M := A_n$

return M

process better. The default database used for TM-Coffee is a reduced uniref50 with entries that contain the keyword "transmembrane". It also allows substitution of these databases with "uniref90" and "uniref100", both with the keyword "transmembrane", hence a database specific for transmembrane alpha helical proteins.

The algorithm can be divided into four parts: (1) Homology Extension; (2) T-Coffee Library Construction; (3) T-Coffee Progressive Alignment; (4) Topology Prediction. The TM-Coffee algorithm is shown in Algorithm 4 and an overview in Figure 2.10. Elements used in this algorithm are explained in Table 2.4.

(1) Homology Extension

In this step, Blast is run on each sequence from the input sequence list. Default parameters (evalue=10, gap open penalty=11, gap extension penalty=1, matrix=Blosum62) are used for running Blast. Blast hits with percent identity between 50%–90% and query coverage above 70% are filtered and the remaining hits are discarded. Then the corresponding query sequence is added to the filtered hits. These hits are the homologs of the sequence in consideration. Table 2.5 lists the parameters and filters used for Blast. This filtered result

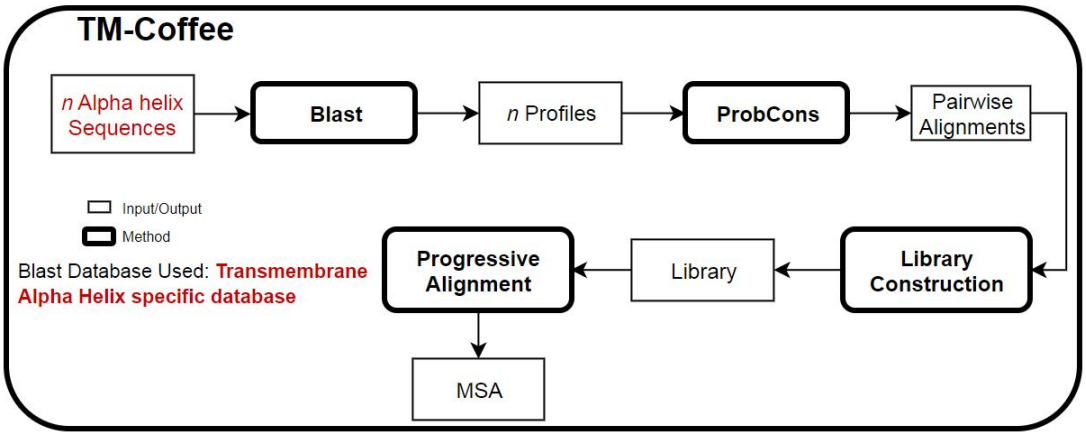

Figure 2.10: Overview of TM-Coffee. The figure shows the TM-Coffee strategy; the main steps required to compute a multiple sequence alignment using the TM-Coffee[19] method. Square boxes indicate input or output while rounded boxes indicate the method used.

is called "Preprofiles" in the algorithm. Figure 2.12, a) shows the Preprofile. It contains gapped columns (unaligned positions in the hits with respect to the query sequence) and unmatched positions in the hits with respect to the query sequence.

These gapped columns are removed and any unmatched position is replaced with a gap. Figure 2.12, b) illustrates the result of this step. It shows how a Profile looks like after gaps are removed and unmatched positions are replaced by gaps. The Profile looks like a one-to-all MSA (i.e seq-to-hits MSA) [19, 10].

(2) T-Coffee Library Construction

Pairwise alignment is performed on all pairs of Profiles produced in the previous step using the Probcons program adapted for TM-Coffee [13, 19]. These pairwise alignments are stored in the list "Lalignment" in the algorithm. The list of sequences "seqList" and "Lalignment" is passed as parameters to the function TCoffee_Library( ). This function is shown separately in Algorithm 2 and described in Section 2.6.5. Unlike the original T-Coffee program, TM-Coffee creates a T-Coffee library from the pairwise alignment of Profiles rather than pairwise alignment of sequences.

(3) T-Coffee Progressive Alignment

This step follows regular T-Coffee progressive alignment except that the T-Coffee library used has homology extension as a precursor as explained above. In the algorithm, the MSA is produced with the function T-Coffee_progressiveAlignment( ). It takes input the list of sequences "*seqList*", T-Coffee Library "*LB*" and pairwise alignment of Profiles "*Lalignment*" This function is explained separately in Algorithm 3 and described in Section 2.6.5.

(4) Topology Prediction

The HMMTOP [67] program produces topology predictions of each sequence. The MSA is color coded according to its topology as "IN", "HEL" and "OUT" where each stand for "inner loop", "transmembrane helix", "outer loop" respectively. function updateMSA( ) performs this step. An updated MSA is shown in Figure 2.11. Note that topology prediction does not influence the multiple sequence alignment method. It is only for visualization purposes.

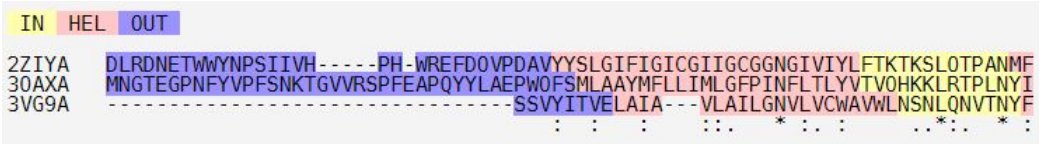

Figure 2.11: TM-Coffee Topology Prediction. This figure shows the TM-Coffee MSA color coded according to the transmembrane topology. Here, IN is inner loop, HEL is transmembrane helix, OUT is outer loop.

Parameter/Filter Type	Parameter/Filter Value
Substitution Matrix	Blosum62
Gap Opening Penalty (gop)	11
Gap Extension Penalty (gep)	1
Expectation Value (evalue)	10
Percentage Identity (pident)	$\geq 50\%$ & $\leq 90\%$
Query Coverage (qcovs)	70%
Blast Database (blastdb)	uniref50 with keyword "transmembrane"

Table 2.5: Default parameters and filters used in Blast for TM-Coffee.

a) Blast Output

Sequence id	Sequences																													
seq	**M**	**M**	**K**	**R**	**-**	**-**	**N**	**I**	**L**	**A**	**-**	**V**	**-**	**-**	**I**	**V**	**P**	**A**	**L**	**-**	**L**	**-**	**-**	**V**	**-**	**A**	**-**	**-**	**G**	**-**	**T**
15583	M	M	K	R	-	-	N	I	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
213754	I	V	K	R	-	-	N	L	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
43219	M	M	K	R	-	-	N	I	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
38230	I	M	K	R	-	-	K	V	L	A	-	M	-	-	L	V	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
43981	M	M	K	R	-	-	N	I	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
106204	M	M	K	R	-	-	N	I	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	V	-	A	-	-	G	-	A
150604		M	K	R	-	-	N	I	L	A	-	V	-	-	V	I	P	A	L	-	L	-	-	A	-	T	-	-	S	-	V
2313		M	K	R	-	-	N	I	L	A	-	I	-	-	L	I	P	T	L	-	L	-	-	V	-	A	-	-	T	-	T

b) Profile

Sequence id	Sequences																	
seq	**M**	**M**	**K**	**R**	**N**	**I**	**L**	**A**	**V**	**I**	**V**	**P**	**A**	**L**	**L**	**V**	**A**	**G**	**T**
15583	M	M	K	R	N	I	L	A	V	-	-	P	A	L	L	V	A	G	-
213754	-	-	K	R	N	-	L	A	V	-	-	P	A	L	L	V	A	G	-
43219	M	M	K	R	N	I	L	A	V	-	-	P	A	L	L	V	A	G	-
38230	-	M	K	R	K	-	L	A	-	-	V	P	A	L	L	V	A	G	-
43981	M	M	K	R	N	I	L	A	V	-	-	P	A	L	L	V	A	G	-
106204	M	M	K	R	N	I	L	A	V	-	-	P	A	L	L	V	A	G	-
150604	-	M	K	R	N	I	L	A	V	-	-	P	A	L	L	-	-	-	-
2313	-	M	K	R	N	I	L	A	-	-	-	P	-	L	L	V	A	-	T

Figure 2.12: Blast output to profile. a) Shows output of Blastp. Sequence used as input to Blastp has the Uniprot id "P02931" and is represented by "seq". This is only for illustrative purpose and the full length of sequences are not shown here. Here "seq" represents the input/query sequence to Blastp and the remaining rows are the Blast hits produced for that input sequence by Blastp for "seq". The Blast hits are represented by unique numeric Sequence ids. These Blast hits are similar sequences to the input sequence "seq". The output format for Blast that returns this type of output is called "FlatQueryAnchoredNoIdentities". b) Shows a profile looks like. Blastp output is turned into a profile by removing all columns corresponding to positions unaligned to the query sequence (i.e. gaps in the query sequence "seq") and by filling with gaps query positions unmatched by Blastp. Gaps are represented by "-" and have been shaded in the profile for visibility. [10, 19].

The TM-Coffee web-server takes input protein sequences in fasta format and produces an MSA in clustalW "aln" format by default, along with the guide tree used and TCscore. TM-Coffee uses EBI Blast/ NCBI Blast as the Local Blast program. The command used by the TM-Coffee web-server is shown below.

```
>tmcoffee.sh -in data_5dcfd219.in -mode psicoffee -blast_server LOCAL \
--search-db 'UniRef50 -- Very Fast/Rough' --search-type ' -prot_min_sim 50 \
-prot_max_sim 90 -prot_min_cov 70' --search-out 'clustalw_aln fasta_aln \
score_ascii phylip score_html' -maxnseq 1000 -maxlen 5000 -case upper \
-seqnos off -outorder input -run_name result -multi_core 4 -quiet=stdout
```

Figure 2.13: Evaluation of TM-Coffee. This figure shows steps in the evaluation of TM-Coffee. Square boxes represent input or output and rounded box represents the method used.

Evaluation of TM-Coffee Algorithm

TM-Coffee is evaluated by its authors using the well curated gold standard Balibase [59] of reference alignments. TM-Coffee is evaluated based on BAliBASE2-ref7 MSA. They use a dataset of 435 transmembrane alpha helices from reference 7 of BAliBASE2 [3]. The MSA generated by TM-Coffee is compared to the reference alignment using BaliScore program to get a score based on Total Column score and Sum-of-pairs score [10]. Figure 2.13 illustrates overview of steps involved in the evaluation of TM-Coffee.

A web server for TM-Coffee is available at
http://T-Coffee.crg.cat/apps/T-Coffee/do:tmcoffee.
TM-Coffee is a part of T-Coffee and can be downloaded from
http://www.T-Coffee.org/Packages/Stable/Latest/

Algorithm 4 TM-Coffee

Input: list of sequences *seqList*, blast database *blastdb*
Default parameters and Filters for Blastp: Given in Table 2.5
Output: multiple sequence alignment M with color coded topology

```
Create an object type "Alignment" to store the pairwise alignment
Initialize the list of pairwise alignments, Lalignment := null, empty matrix
  MpreProfile_seqs, list PseqProfile := null
for seq in seqList do
    gappedHomolog_seq := Blastp(seq, blastdb)
    preProfile_seqs := seq.add(gappedHomolog_seq)
    MpreProfile_seqs := matrix form of preProfile_seqs
    for j = 1 to N do
        if MpreProfile_seqs[1,j] is a gap "–" then
            Remove column j from MpreProfile_seqs
        else
            for i=2 to M do
                if MpreProfile_seqs[1,j] ≠ MpreProfile_seqs[i,j] then
                    Place gap "–" at MpreProfile_seqs[i,j]
                end if
            end for
        end if
    end for
    Store all MpreProfile_seqs in list PseqProfile
end for
for each pair (profile_1, profile_2) in PseqProfile do
    pairofProfiles := probcons_pairwise(profile_1, profile_2)
    aln := store_pairwiseAlignment(pairofProfiles)
    Lalignment := Lalignment.append(aln)
end for
LB, Lalignment := T-Coffee_Library(seqList,Lalignment)
M := T-Coffee_progressiveAlignment(seqList, LB, Lalignment)
topol := hmmtop(seqList, seqProfile)
M := updateMSA(M, topol)
return M
```

2.7 Other Tools and Resources

2.7.1 NorMD

NorMD [60] is a reliable objective scoring function for MSA. It stands for normalized mean distance and based on the objective function used by ClustalX[57]. It combines a column scoring method and residue similarity scores. To incorporate residue similarity, it uses matrices such as PAM[12], BLOSUM[25] or GONNET[21]. NorMD also considers ab-initio sequence information such as the length, number, and similarity of sequences being compared. It can also detect badly aligned sequences in the alignment. In NorMD, column score is calculated for each column in the alignment and summed over the full length of the alignment. This column score used on NorMD is based on [68].

Mean distance column scores are computed for the core blocks in the alignment. As MD column scores are normalized in the range of 0 to 100, a threshold can be set above which columns are considered to have statistically significant scores.

The NorMD score is calculated as

$$NorMD = \frac{MD - GAPCOST}{MaxMD * LQRID} \tag{7}$$

where, MD is mean pairwise distance between sequences in a continuous sequence space, $GAPCOST$ is the mean gap cost for each pairwise alignment in the MSA, $MaxMD$ is the maximum possible MD score and $LQRID$ is the lower quartile range of the pairwise hash score.

This objective scoring function is based on the sequences and can be used when there are no benchmark or reference alignments available. NorMD provides various advantages such as being unbiased towards the type of MSA program used during evaluation. NorMD allows the e-value for pairwise alignments in the MSA to be as high as 10. NorMD can be computed on single MSA and does not need alternate MSAs to form a consensus like in other object functions for MSA. For an alignment, NorMD score above 0.6 is considered to be a quality alignment. Higher the score is above the cutoff, better the MSA. NorMD

is tested on Balibase reference alignment and a cut-off 0.6 is the lowest score obtained by reliable alignments in general[60], [61].

NorMD can be downloaded from http://www.bork.embl.de/Docu/AQUA/latest/

2.7.2 Blast Database

A database of sequences and other related information in a Blast recognisable format is called a Blast database. A set of sequences can be converted to a Blast database. Popular protein databases that are available in Blast format are "NR", "Swissprot", "NT", "Refseq_protein", "PDB" etc. For this research, we construct and use local databases recognisable by Blast. These databases can be downloaded or created locally. Popular Blast databases can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/.

2.7.3 Homology Extension

Homology extension [52, 10] is a method by which evolutionary information required for the alignment is enriched by database searches for homologous sequence. Instead of considering the sequence itself, its homologs are taken to consider the evolutionary variability. A threshold is set for selecting homologs. Only sequences that fall beyond the threshold are considered. This is a pre-alignment step. The set of sequences obtained by homology extension are converted to a profile. Finally many such profiles are aligned to produce an MSA.

2.7.4 HMMTOP

HMMTOP [67, 66] is a program to predict transmembrane topology as well as localization of the helical segments in them. It uses an HMM showing that the maxima of the likelihood function on the space of all possible topologies of a given amino acid sequence with an experimentally established topology. The states for the HMM are; inside loop (I), inside tail (i), membrane helix (h), outside tail (o) and outside loop (O). Licence for HMMTOP can be requested at http://www.enzim.hu/hmmtop/html/download.html

Chapter 3

TMB-Coffee

Transmembrane beta-barrel proteins are an important class of proteins. It is important to study MSA of available sequence data on TMBB proteins to help identify, annotate function and also understand their structure. In this chapter, we discuss about how we construct a dataset of beta barrel proteins, and adapt TM-Coffee, an MSA method based on homology extension followed by progressive alignment using T-Coffee, to develop a tool TMB-Coffee for MSA of TMBB proteins. Finally, TMB-Coffee method is evaluated for the quality of alignments produced.

3.1 Construction of Dataset

There is no benchmark dataset of TMBB proteins or reference alignments for them. Literature review suggests that the work that is being done on transmembrane beta-barrel proteins is based only a select few well curated proteins that have known structure represented in PDB. Out of the 4034 redundant transmembrane structures, less than 12% percent of them are beta barrels (according to PDBTM [64] version 18th January 2019) and the remaining structures are alpha helices. After clustering to remove identical sequence up to 10% dissimilar, we are left with 93 beta barrels and 1169 alpha helices. Hence, there is a need to construct a larger dataset with available data that includes proteins that have known structures or have been reviewed.

PDB is the most popular 3D structure database for proteins yet it does not have any entries specifically annotated as transmembrane beta-barrel. Orientations of Proteins in Membranes database (OPM database) [35] is another database that claims to have spatial arrangement of transmembrane protein in the lipid bi-layer that has been obtained from theoretical and compared with experimental data. This database has 233 TMBB entries with co-referenced Uniprot IDs. TMBB sequence data that is available is either relatable to PDB IDs or Uniprot IDs. Whenever we found PDB IDs associated with TMBBs, corresponding Uniprot IDs are found to maintain consistency in the type of data that we have.

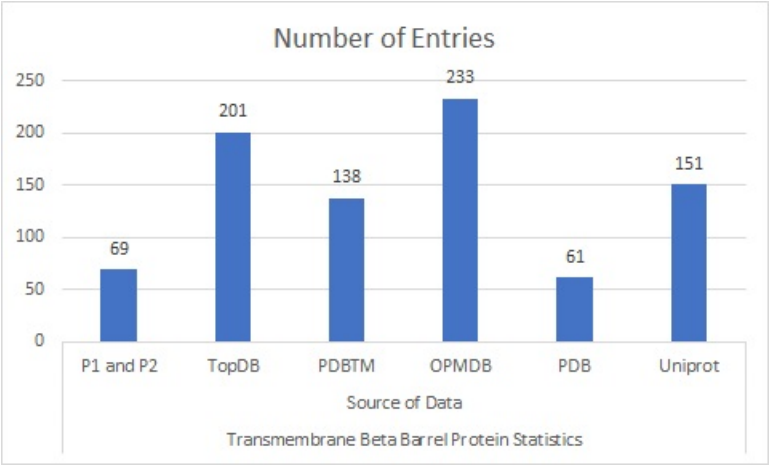

Figure 3.1: Statistics of Uniprot entries of beta barrel proteins. This represents the total number of Uniprot IDs corresponding to TMBB entries in each source of data (P1andP2- It comes from Paper1 [17] and Paper2 [30] citing TMBB structures, TopDB [65], PDBTM [64], OPMDB [63], PDB [6], Uniprot [4]).

Sequences that are considered for MSAs usually need to meet the following general requirements; (i) it would be ideal if sub-groups of the sequences that are too similar be pre-aligned separately, and (ii) one member of each subgroup be included in the sequence considered for final multiple sequence alignment, (iii) remove that are too dissimilar before computing the MSA.

Due to the lack of a benchmark dataset or alignment for TMBB, we carefully constructed a dataset. (1) We reviewed many papers that cited beta barrel proteins with known 3D

structure. (2) We chose Paper1[17] and Paper2[30]. (3) Both combined gave us TMBBs with available structures that has been empirically annotated. (4) PDB IDs [6] are extracted from these papers. (5) The corresponding Uniprot IDs are found for these PDB IDs. (6) We have a list of Uniprot IDs[4] of TMBB that have 3D structure. (7) We extract Uniprot IDs of all TMBB proteins from TopDb [65]. (8) Then, we get all TMBB from OPMDB[63]. It contains a list of PDB structures of the proteins. (9) The PDB IDs are extracted and corresponding Uniprot IDs are found.

(10) Now, we have a list of Uniprots IDs corresponding to entries from all the different sources of transmembrane beta barrel proteins (TMBB) (Paper1 (P1)[17], Paper2 (P2)[30], TopDB[65], OPMDB[63], PDB[6]), Uniprot. (11) Duplicate Uniprot IDs are removed from the list. (12) Keep only those entries in list that had "reviewed" status according to Uniprot and discard the rest. (13) Retrieve the fasta sequence of all the entries produced in the previous step. This is the list of entries/sequences that is used as the dataset for testing our MSA method. We call our dataset "datasetTMB" and it contains 159 entries/sequences. The complete list of Uniprot IDs of the sequences in datasetTMB is given in the Table A.3 in Appendix.

3.2 Construction of Local Blast Database for TMB-Coffee

MSA based on homology extension is usually more accurate than other general alignment techniques but suffers poor performance because of the time taken to iteratively perform Blast searches to get homologs. TM-Coffee is a multiple sequence alignment tool for alpha helical transmembrane proteins and performs homology extension using a database of transmembrane helices. This is highlighted in red in Table 2.10. We adapt their technique and use it to align TMBB. We use databases that contain beta barrel transmembrane proteins, beta barrel transmembrane segments, and outer membrane proteins. Since we have used reduced databases of TMBB proteins, the time taken to perform homology extension is reduced compared to the time taken for a larger database like Swissprot. For Blast we also consider Swissprot database, a database that is minimally redundant but contains a mix of

all proteins, alpha helices, beta barrels and other proteins. We use this as a control to check if the beta barrel databases that we consider makes a difference. We use standalone Blast+ version 2.9 to construct a local database that serves the purpose of similarity searches. First, we download non redundant beta barrel sequences from OMPDB [63]. These entries have a sequence identity of up to 70%. It is downloaded in fasta format. We use the following script to create a non redundant database in blast format. We call this database "OMPDB70" where OMPDB is the name of the database from where the data is downloaded and the number 70 represents sequence identity threshold.

```
> makeblastdb -in OMPDBsequence.fasta -dbtype prot -title ``OMPdb70" \
  - parse_seqids
```

In the above command, "makeblastdb" is the command used in Blast+ (a version of NCBI Blast package that contains many standalone programs for working with sequences) for the construction of local Blast database. The flag "-in" takes the input in fasta file format. The flag "-dbtype" is used to specify the type of database that "makeblastdb" command has to create. Here we create a protein sequence database, hence the keyword "prot". The flag, "-parse_seqids" is used to create a list of fasta headers of the sequences in the database that can be used for sequence look-ups. The final flag is "-title", to have a user-specified title for the database.

We construct another database, unirefOMBB100 in the same way; also named after the source and sequence identity threshold. The unirefOMBB100 database contains sequences from Uniprot Uniref database with keyword, "outer membrane beta barrel". The number 100 means it has sequences that are up to 100% identical and hence is a redundant database. Finally, we also download and convert sequences in Swissprot to a blast database . Swissprot contains well curated but minimally redundant amino acid sequences. Local versions of these databases are created in Blast format for further use in TMB-Coffee. Table 3.1 shows the number of sequences in each locally created blast database.

Local Blast Database	Number of Sequences
OMPdb70	22556
unirefOMBB100	57350
swissprot	471914

Table 3.1: Local databases and the number of sequences in them

3.3 TMB-Coffee

This section explains the TMB-Coffee algorithm. We call it, "TMB-Coffee" to distinguish it from TM-Coffee. TMB-Coffee differs from TM-Coffee with respect to the database used for homology extension and the input sequences. The differences in TMB-Coffee w.r.t TM-Coffee is highlighted in red in Table 3.2.

TMB-Coffee can be divided into three parts: 1) Homology extension 2) T-Coffee Library construction 3) T-Coffee progressive alignment. The TMB-Coffee algorithm is shown in Algorithm 5 and an overview in Figure 3.2. Elements used in this algorithm are explained in Table 2.4. Although the following steps are same as TM-Coffee, the input, blast database, the library and the final output is different.

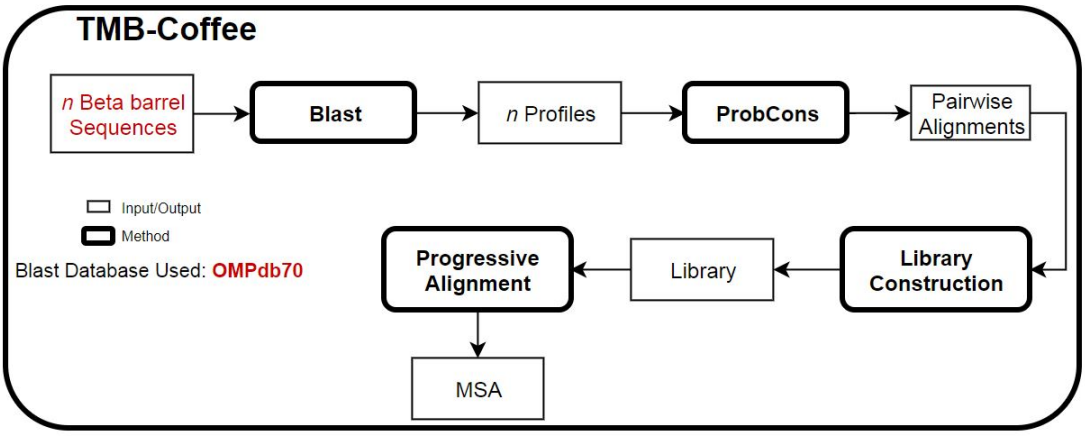

Figure 3.2: Overview of TMB-Coffee. The figure shows the TMB-Coffee strategy; the main steps required to compute a multiple sequence alignment using the TMB-Coffee method. Square blocks indicate input or output while rounded boxes indicate the method used. Input sequences and sequences in database used here are all beta-barrel sequences. Red highlights present the changes in TMB-Coffee.

(1) Homology Extension

The major difference between TMB-Coffee and TM-Coffee is the Blast database that is used for homology extension. TM-Coffee uses a transmembrane alpha helix specific database, whereas, TMB-Coffee uses TMBB specific database.

a) When we have a set of TMBB protein sequences to be aligned, the Homology Extension step takes each sequence in the set and performs Blast on it. A sequence that is used as an input to Blast program is called a query sequence. Blast returns a set of similar sequences as its output for each input/query sequence. These homologs are retrieved from a TMBB specific local blast database. b) The query sequence is stacked on its homologs. This is called a "Preprofile" and looks like an MSA. c) Unaligned columns in the Preprofile (gapped positions) corresponding to the query sequence is removed and unmatched positions are filled with gaps, to convert a Preprofile to a Profile. Steps a, b and c are repeated on all remaining sequences in the initial set of sequences to be aligned. In the end, we have one profile for each sequence from the input set. Next, pairwise alignment is performed using ProbCons Pair-HMM for all pairs of profiles.

Parameter/Filter Type	Parameter/Filter Value
Substitution Matrix	Blosum62
Gap Opening Penalty (gop)	11
Gap Extension Penalty (gep)	1
Expectation Value (evalue)	10
Blast Database (blastdb)	TMBB specific database

Table 3.2: Parameters and Filters used in Blast for TMB-Coffee

2) T-Coffee Library Construction

This step follows the same procedure as in the original T-Coffee program. The pairwise profile alignments are weighted to produce primary library. Then this library is extended to produce the T-Coffee library. The steps to compute the primary library and its extension to create a T-Coffee library is explained in detail in Section 2.6.5. While TM-Coffee uses transmembrane alpha helices to create a T-Coffee Library, TMB-Coffee uses TMBB for the same purpose. Therefore, the TMB-Coffee library constitutes weights of beta barrel

Features	MSA Method		
	T-Coffee	**TM-Coffee**	**TMB-Coffee**
Input	Protein sequences	TM Alpha Helix sequences	TMBB sequences
Homology Extension	No	Yes	Yes
Use of Database	No	Yes (TM Alpha Helix)	Yes (TMBB)
Pairwise Alignment	Sequence	Profile	Profile
T-Coffee Library Construction	Yes	Yes	Yes
T-Coffee Progressive Alignment	Yes	Yes	Yes

Table 3.3: Differences and Similarities between T-Coffee, TM-Coffee and TMB-Coffee.

residues computed from the pairwise alignments of beta barrel sequences.

3) T-Coffee Progressive Alignment

This step also follows the same procedure as the original T-Coffee program. Here, the pairwise alignments are made to produce distances. The distance between the pairs are used to compute a guide tree using the Neighbor-Joining method. The guide tree is followed for performing multiple sequence alignment using dynamic programming and uses T-Coffee library to determine which residues should align in a pair or sequences or sub-alignments. This step is explained in detail in Section 2.6.5.

Table 3.3 shows the differences between T-Coffee, TM-Coffee and TMB-Coffee methods.

3.4 Evaluation of TMB-Coffee

Generally, MSAs produced by various MSA programs are compared to reference/benchmark MSA to determine their performance or calculate accuracy. In the case of TM-Coffee, Balibase-Ref7 (Balibase MSA for transmembrane alpha helices) benchmark is used for the evaluation. Balibase is a manually curated MSA with the help of sequence and structural information. Since there is no benchmark or reference alignment like Balibase available for TMBB proteins, a comparison based evaluation is not possible for TMB-Coffee.

Instead of comparison based evaluation of MSA, there are programs that can estimate the quality of the MSA based of various criteria such as residue similarity, alignment of residues in a column of the MSA, etc. MUMSA[32], AL2CO[46], COFFEE Objective

Algorithm 5 TMB-Coffee

Input: list of sequences *seqList*, blast database *blastdb*
Default parameters and Filters for Blastp: Given in Table 3.2
Output: multiple sequence alignment M

```
Creating an object type "Alignment" to store the pairwise alignment
Initializing the list of pairwise alignment, Lalignment := null, empty matrix
  MpreProfile_seqs, list PseqProfile := null
for seq in seqList do
    gappedHomolog_seq := Blastp(seq, blastdb)
    preProfile_seqs := seq.add(gappedHomolog_seq)
    Store preProfile_seqs in matrix form to MpreProfile_seqs
     # Creating Profile from MpreProfile_seqs
    for j=1 to N do
        if MpreProfile_seqs[1,j] is a gap"–" then
            Remove column j from MpreProfile_seqs
        else
            for i=2 to M do
                if MpreProfile_seqs[1,j] ≠ MpreProfile_seqs[i,j] then
                    Place gap "–" at MpreProfile_seqs[i,j]
                end if
            end for
        end if
    end for
    Store all MpreProfile_seqs in list PseqProfile
end for
for each pair (profile_1, profile_2) in PseqProfile do
    pairofProfiles := probcons_pairwise(profile_1, profile_2)

    aln := store_pairwiseAlignment(pairofProfiles)
    Lalignment := Lalignment.append(aln)
end for
LB, Lalignment := T-Coffee_Library(seqList,Lalignment)
M := T-Coffee_progressiveAlignment(seqList, LB, Lalignment)
return M
```

function[44], NorMD [60] can be used to determine the quality of MSA when there is a lack of reference alignments. NorMD produces a score to evaluate the quality of the alignments based on the sum of columns scores and other ab-initio information.

We use NorMD to measure the quality of MSA in this research for the various advantages it has over other programs mentioned above. The strategy used in MUMSA requires several MSAs of the same sequences to be compared. It looks for the average overlap and the multiple overlap scores in the MSAs. This can be useful for comparing MSAs produced by different MSA programs but not to estimate the quality an individual MSA. The COFFEE objective function, on the other hand, looks for consistency between a library of pairwise alignments and the MSA produced from it. This method is believed to be biased towards TMB-Coffee. AL2CO looks for conserved regions and computes a conservation index. We use NorMD because the objective measure is based on the unaligned sequences and this does not have a bias towards any MSA program. It can also evaluate a single MSA to determine whether the alignment is reliable. In addition, NorMD permits the automatic identification of misaligned regions in the MSA.

From Equation 7, the range of NorMD score depends on $MaxMD$. It is the maximum possible MD score and is dependant on the longest sequence and number of sequences in MSA in consideration. For our research the lowest NorMD score is 0.098 and the highest is 4.490. However, majority of the scores ranges from 0.000 to 1.000. According to [60], [61], it is observed that alignment with lower NorMD score yet a reliable quality has a NorMD score of 0.6. It is also observed that higher NorMD scores mean better MSA.

Figure 3.3 shows the overview of steps of evalaution of TMB-Coffee. Following are the steps of evaluation of TMB-Coffee in detail: (1) We take one sequence in datasetTMB and perform Blast with OMPdb70 database. We use default Blast parameters for the input sequence. (2) Similar sequences obtained from Blast is used as input to TMB-Coffee. Blast in TMB-Coffee uses the same database for homology extension. (3) The MSA obtained from TMB-Coffee is scored using NorMD to get a numeric score. An MSA in this context contains one sequence from datasetTMB and its homologs. Steps 1, 2, 3 are repeated for all other sequences in datasetTMB. Note that when the initial Blast does not return any

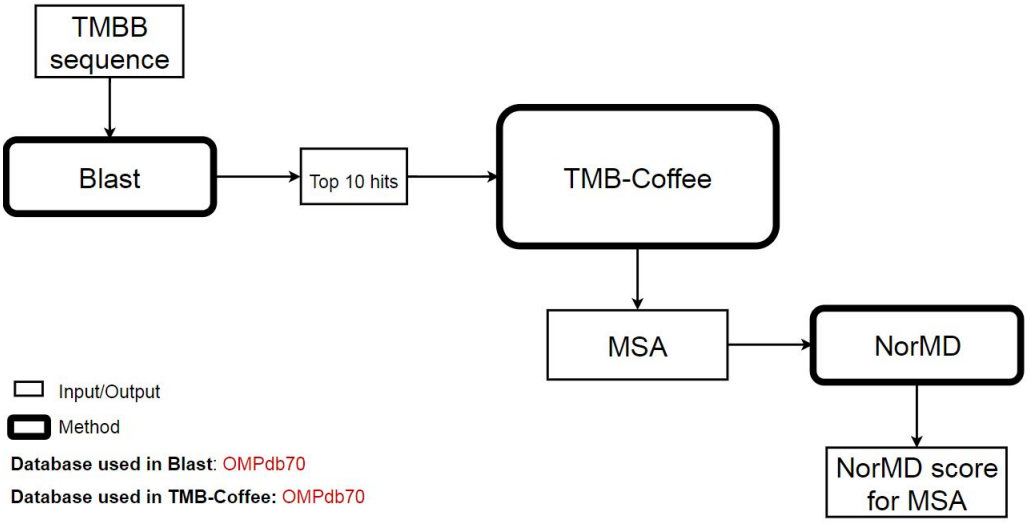

Figure 3.3: **Steps in the Evaluation of TMB-Coffee**. A TMBB sequence obtained from datasetTMB is used as input to the initial Blast. Top 10 blast hits/results are used as input to TMB-Coffee. Quality of the resulting MSA is determined using NorMD score. Note: The database used for initial Blast and Blast within TMB-Coffee is "OMPdb70".

hits, such input sequences are not processed further.

(A) We use three different local blast databases: OMPdb70, unirefOMBB100 and swissprot. While OMPdb70 and unirefOMBB100 are beta-barrel specific databases, swissprot is a general purpose protein database.

(B) We use three thresholds for the number of blast hits: 10, 25, and 50.

Table 3.4 shows the number of sequences in datasetTMB that obtained Blast results according to the above filters. Hence, for each sequence in datasetTMB that met the above criteria, TMB-Coffee is used to produce an MSA. NorMD is computed on each MSA to determine the quality of the MSA. Out of 159 sequences in datasetTMB, only some sequences produced Blast results with the thresholds (10, 25 and 50). Blast hits for the remaining sequences produced hits less than the threshold. Such sequences are not used in the experiments. For swissprot database, even though, Blast produced sufficient hits, TMBCoffee could not produce alignments for all sequences. This is due to the reason that swissprot database is redundant. When TMB-Coffee detects redundant sequences during the alignment step, it is considered as a single sequence. We see this difference in the

Database	Number of sequences with at least 10 hits	25 hits	50 hits
OMPdb70	137	110	106
unirefOMBB100	88	52	46
swissprot	142	106	57

Table 3.4: Number of sequences from datasetTMB that meets the Blast criteria for number of hits. This also tells how many MSA we have in each category except swissprot. Swissprot has 124/142, 85/106, 32/57 MSAs for 10 hits, 25 hits and 50 hits respectively

number of alignments produced while using swissprot because of the reason that too many redundant sequences are reduced to a single one and less than 3 sequences cannot produce an alignment. Only 124/142, 85/106, 32/57 MSAs are produced for 10 hits, 25 hits and 50 hits respectively. To make the performance evaluation and comparison of TMB-Coffee consistent, we consider only those sequences that meet the above criteria.

3.5 Environment Used

All programs are run on virtual machine using Ubuntu 18.10 64bit OS with Intel Core i7-4790 @ 3.6 Ghz processor, 1 core, 8 GB DDR3 RAM. Scripting is based on Python 3.6.8 in Jupyter Notebook 5.7.6 and GNU Bash 4.4.19(1) on Ubuntu 18.10. Virtual environment is built on the host Windows 7 Enterprise SP1 using OracleVM Virtual Box 5.1.14.

3.6 Results

This section contains results of the evaluation of TMB-Coffee. In Figure 3.3 we see the NorMD evaluation of TMB-Coffee using the database “OMPdb70” to get at least 10 hits from Blast. Hits in the initial Blast determines the number of sequences in the final MSA produced by TMB-Coffee. In this section, we show the results of not just “At least 10 hits” but also “At least 25” and “At least 50” hits from the initial Blast. We also present the results obtained for top hits from Blast while using the local Blast databases “OMPdb70”, “unirefOMBB100” and “swissprot”. Finally, we compare the performance of TMB-Coffee

	Average NorMD score for MSA from TMB-Coffee		
Blast hits	**Database: swissprot Average NorMD**	**Database: OMPdb70 Average NorMD**	**Database: unirefOMBB100 Average NorMD**
At least 10	0.996	0.644	0.811
At least 25	0.851	0.692	0.710
At least 50	0.727	0.683	**0.574**

Table 3.5: Average NorMD score for MSAs from TMB-Coffee for Top Blast hits while using different Blast databases (Swissprot, OMPdb70, unirefOMBB100).

with other general purpose MSA programs. According to [60], MSA with NorMD score above 0.6 is considered to be a reliable one. Therefore, 0.6 is the cut off used to determine if TMB-Coffee could generate reliable MSAs. In our study, the lowest NorMD score that we obtained is 0.098 for P31243 using swissprot (10 hits) and highest is 4.490 for **P09616** using unirefOMBB100 (25 hits).

Table 3.5 shows the average NorMD scores obtained for "At least 10", "At least 25", "At least 50" hits from Blast while using databases swissprot, OMPdb70 and unirefOMBB100. This is the average taken from the NorMD scores for individual MSA produced for each sequence in datasetTMB. Note that all NorMD scores are computed based on defaults parameters (substitution matrix= Blosum62 , gap open penalty= 0 and gap extension penalty= 0).

Table 3.6 shows the comparison of TMB-Coffee with other general purpose MSA programs. The saame set of sequences are used to obtain MSAs and its NorMD scores. This table shows the average of NorMD scores sequences with at least 10 Blast hits. Similarly, Table 3.7 and Table 3.8 shows the comparison of TMB-Coffee with other general purpose MSA programs for 25 and 50 hits respectively. The individual NorMD scores are for each category is given in Section A.4 in the Appendix.

3.7 Discussion

In the results we show the average NorMD scores obtained for MSAs from each MSA program on various databases with different input sizes. The individual NorMD scores are

MSA Method	Comparison of MSA Programs using NorMD score (10 hits)		
	Database: Swissprot Average NorMD	**Database: OMPdb70 Average NorMD**	**Database: unirefOMBB100 Average NorMD**
Clustal Omega	0.967	0.647	0.801
MAFFT	<u>1.096</u>	<u>0.715</u>	<u>1.131</u>
MUSCLE	0.996	0.655	0.988
T-Coffee	0.976	0.653	0.845
TMB-Coffee	**0.996**	**0.644**	**0.811**

Table 3.6: Comparison of MSA programs - Average NorMD score for MSA of sequences from datasetTMB with 10 hits. Highest scores are underlined. Scores for TMB-Coffee are bold.

MSA Method	Comparison of MSA Programs using NorMD score (25 hits)		
	Database: Swissprot Average NorMD	**Database: OMPdb70 Average NorMD**	**Database: unirefOMBB100 Average NorMD**
Clustal Omega	0.877	0.694	0.706
MAFFT	<u>0.924</u>	<u>0.715</u>	<u>0.869</u>
MUSCLE	0.899	0.704	0.764
T-Coffee	0.587	0.693	0.753
TMB-Coffee	**0.851**	**0.692**	**0.710**

Table 3.7: Comparison of MSA programs - Average NorMD score for MSA of sequences from datasetTMB with 25 hits. Highest scores are underlined. Scores for TMB-Coffee are bold.

MSA Method	Comparison of MSA Programs using NorMD score (50 hits)		
	Database: Swissprot Average NorMD	**Database: OMPdb70 Average NorMD**	**Database: unirefOMBB100 Average NorMD**
Clustal Omega	0.708	0.676	0.602
MAFFT	<u>0.804</u>	<u>0.691</u>	<u>0.713</u>
MUSCLE	0.748	0.683	0.599
T-Coffee	0.734	0.685	0.619
TMB-Coffee	**0.727**	**0.683**	**0.574**

Table 3.8: Comparison of MSA programs - Average NorMD score for MSA of sequences from datasetTMB with 50 hits. Highest scores are underlined. Scores for TMB-Coffee are bold.

for each category is given in Section A.4 in the Appendix. In order to evaluate the performance of TMB-Coffee we use NorMD score. According to [60], MSA with NorMD score above 0.6 is considered to be a reliable one. Therefore, 0.6 is the cut off used to determine if TMB-Coffee could generate reliable MSAs. In our study, the lowest NorMD score that we obtained is 0.098 for P31243 using swissprot (10 hits) and highest is 4.490 for **P09616** using unirefOMBB100 (25 hits).

TMB-Coffee generally produces reliable MSAs

As we see in Table 3.5, the average NorMD score for each category is above the cut-off except for 50 hits using unirefOMBB100. We also see that the scores for unirefOMBB100 for 10 hits and 25 hits have scores 0.811 and 0.710 respectively. From this we can conclude that, although, usage of this database has enabled MSAs to get reliable scores for 10 and 25 hits, it does not for 50 hits. NorMD scores for MSA from TMB-Coffee are above 0.6, conveying that TMB-Coffee generally produces reliable MSAs.

The Swissprot database results in better MSAs

The average scores for swissprot and OMPdb70 has met the cut-off score for NorMD in all categories. This suggests that both these databases can be used for producing reliable MSAs using TMB-Coffee, but, Swissprot database results in better MSAs. The assumption that using OMPdb70 and unirefOMBB100 (two beta-barrel specific databases) would produce better alignments is not supported by the results.

MAFFT performed better than TMB-Coffee

We also compared the performance of different MSA programs with that of TMB-Coffee. We performed experiments with three different input sizes 10 hits, 25 hits and 50 hits. The average NorMD scores obtained for various MSA programs for the different input sizes are reported in Table 3.6, Table 3.7 and Table 3.8 respectively.

From Table 3.6, we see that MAFFT has scored better than TMB-Coffee in all the categories. We can also see that T-Coffee has also performed better than TMB-Coffee

Paired t-test	ClustalO	MAFFT	Muscle	T-Coffee	TMB-Coffee
ClustalO	NA	**4.53**	1.52	0.45	1.69
MAFFT		NA	**-4.22**	**-5.24**	**-4.70**
Muscle			NA	-1.01	0.03
T-Coffee				NA	1.36
TMB-Coffee					NA

Table 3.9: Paired t-test (Swissprot - 10 hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant

based on higher NorMD scores. We see similar trend for the other input sizes in Table 3.7 and Table 3.8. In all the experiments, MAFFT gets better NorMD score than any other MSA program including TMB-Coffee.

The NorMD scores seems very close to each other. Therefore to check the statistical significance of the difference, we performed a paired t-test. We perform paired T-Test for MSA using Swissprot as it obtained top scores.

From Table 3.9 we see that the score obtained by MAFFT is significant compared to all other programs. Whereas, the score obtained by TMB-Coffee is not significant compared to the score from other programs. All remaining t-test tables given in Appendix Section A.5 yield similar results.

From literature review we know that T-Coffee is the state of the art and is considered to have one of the most accurate MSAs. However from the comparisons we see that MAFFT outperforms T-Coffee and TMB-Coffee. Therefore, we wanted to check if MAFFT aligns better in the TMBB region of the sequences than TMB-Coffee.

Figure 3.4 shows the alignment in the TMBB region by TMB-Coffee and Figure 3.5 shows the alignment in the TMBB region by MAFFT. The aligned regions are marked by a black box in both figures. To determine how well the TMBB regions have been aligned, we count the number of conserved, similar and weakly similar residues in aligned regions. After counting the aligned regions, we see that TMB-Coffee has aligned with a count of 21, whereas, MAFFT has a count of 17. Higher count for the aligned region suggests better alignment in the TMBB region.

Although, the alignment from MAFFT has a higher NorMD score of 1.000 than that of TMB-Coffee 0.701, TMB-Coffee has a better alignment of the TMBB region. This

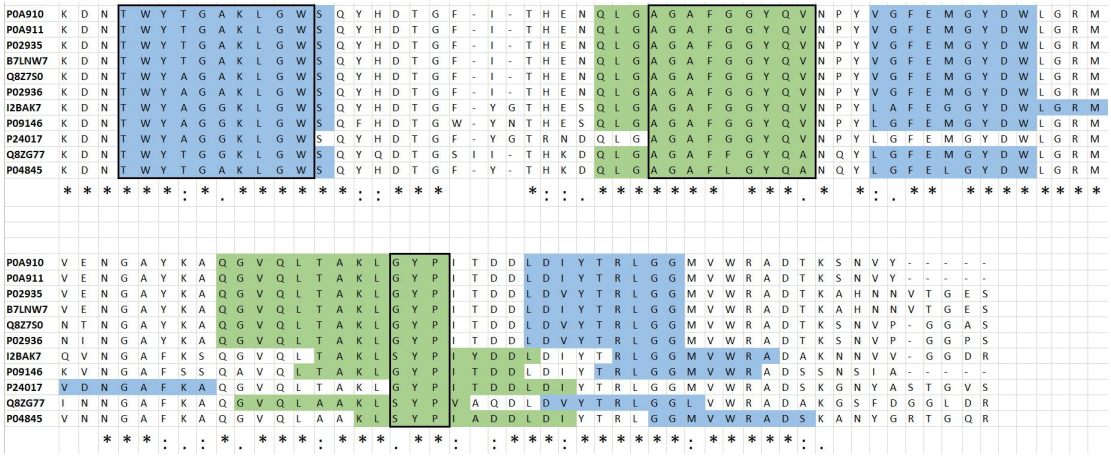

Figure 3.4: Illustration of alignment in TMBB region of sequences using TMB-Coffee. Blue and Green are consecutive TMBB regions. Black box shows area within the TMBB regions that have aligned well. "*" represents conserved residues, ":" represents slightly similar residues and "." represents weakly similar residues. NorMD score for this alignment (P0A910- Swissprot 10 hits) is 0.701.

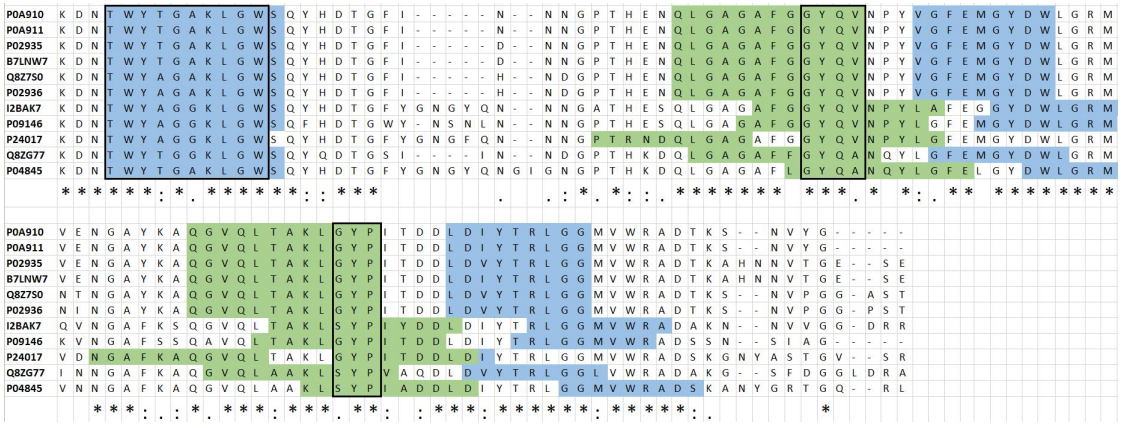

Figure 3.5: Illustration of alignment in TMBB region of sequences using MAFFT. Blue and Green are consecutive TMBB regions. Black box shows area within the TMBB regions that have aligned well. "*" represents conserved residues, ":" represents slightly similar residues and "." represents weakly similar residues. NorMD score for this alignment (P0A910- Swissprot 10 hits) is 1.000.

suggests that TMB-Coffee may align TMBB regions better than MAFFT. It also conveys that NorMD score may not reflect the alignment of beta-barrels. However, more work is to be done to check if this holds true for the majority of the other alignments in the experiment.

Chapter 4

Conclusion

In conclusion, MSA program MAFFT outperforms T-Coffee and TMB-Coffee in terms of NorMD score. However, TMB-Coffee can better align TMBB regions in the sequences. Well curated general purpose database, Swissprot results better alignment in TMB-Coffee. This chapter presents the work done to develop TMB-Coffee, an MSA method to align TMBB protein sequences. It also summarizes contributions and presents the limitations of this work. The aim of this research is to develop an MSA method to align TMBB protein sequences and to see how well existing methods perform in comparison to our method, TMB-Coffee.

4.1 Contributions

Our Contributions are:

(1) the creation of dataset of TMBB proteins as gold standard;

(2) adaption of TM-Coffee for the context of TMBB proteins, as TMB-Coffee tool;

(3) evaluation and comparison of TMB-Coffee.

We cleaned and consolidated the data that is available on TMBB into a single dataset. This dataset that we created has 159 TMBB proteins.

We adapted TM-Coffee to be TMB-Coffee for aligning TMBB protein sequences using beta-barrel specific sequence databases and library of TMBB.

We evaluated TMB-Coffee by scoring MSAs produced using NorMD. Finally, we compared the performance of TMB-Coffee with state of the art multiple sequence alignment tools.

We use two TMBB specific databases to test their effect on the MSA. However we get better alignments while using swissprot, a general purpose protein database and not a TMBB specific database. While we compare TMB-Coffee with state of the art MSA methods, we see that MAFFT produces better alignment based on NorMD score in all cases. Therefore we learn that, though the same method works for Transmembrane alpha helices, it does not hold true for TMBB. However, we also see that TMB-Coffee aligns TMBB regions better than MAFFT in select cases.

The dataset, databases and scripts used in this research are available on GitHub. The following URL can be used to reach the repository, https://github.com/akhiljobby/msa_transmembraneBetaBarrel.

4.2 Limitations

This work only considers sequence based approaches for constructing the MSA rather than a structure based one. It is shown in literature when sequence and structure based techniques are combined, it produces more accurate alignments. Hence, in future structure based alignments methods can be produced for TMBBs in combination with homology extension to check if that can improve the quality of multiple sequence alignments. Further study is required in determining if TMB-Coffee aligns TMBB regions better than state of the art methods. We also do not use topological information in this research. Such information may supplement the alignment process and in turn lead to better MSA.

Appendix A

Supplementary Info

A.1 Files on Github

The files related to our work can be found at https://github.com/akhiljobby/msa_transmembraneBetaBarrel. It has a README file with instructions and details of files. The list of files are presented in Table A.1.

FILES	DESCRIPTION
matrices.tar	Tar file contains various substitution matrices
datasetTMB.fasta	Fasta file contains the dataset of TMB used as a gold standard
AllIDs.xlsx	Excel file contains Uniprot ID of TMBB extracted from various sources
NorMDscores.xlsx	Excel file contains the NorMD scores for all MSA programs used separated by size of input and database used
alignments.tar	Tar file contains all MSAs produced by MSA programs based on various categories
extractFastaSeqfromBlastDB.ipynb	Jupyter Notebook file contains script to retreive Fasta sequences from local Blast databases
loopBlast.ipynb	Jupyter Notebook file contains script to perform iterations of Blastp
loopTMBCoffee.ipynb	Jupyter Notebook file contains script to loop TMBCoffee script
makeProfile.ipynb	Jupyter Notebook file contains script to create Profile from Blast output
removeCSVCol.ipynb	Jupyter Notebook file contains script to remove columns from csv output of Blast
removeIdsLessThresh.ipynb	Jupyter Notebook file contains script to filter Blast output based on thresholds
splitBlastOut.ipynb	Jupyter Notebook file contains script to separate output from command line blast into blast results of individual sequences

Table A.1 continued from previous page

FILES	DESCRIPTION
splitFasta.py.ipynb	Jupyter Notebook file contains script to split one fasta file with multiple sequences into multiple fasta files
topBlastResult.ipynb	Jupyter Notebook file contains script to filter top Blast results based on minimum number of hits
tmbcoffee.ipynb	Jupyter Notebook file contains script for TMBCoffee
AlignmentToProfile.ipynb	Jupyter Notebook file contains script to convert an alignment to a profile
combineFastaFiles.ipynb	Jupyter Notebook file contains script to combine several fasta files to one fasta file
README	File contains instructions and description of contents in the repository

Table A.1: List of files on GitHub

A.2 Tools Used

Tools	Version
T-Coffee	12.00.7fb08c2
Blast+	2.9
CDHIT	4.8
Clustal Omega	1.2.4
HMMTOP	2.0
MAFFT	7.428
MUSCLE	3.8.31
NORMD	1.3

Table A.2: Version of the tools used in this research

A.3 Dataset of TMBB - "datasetTMB"

Table A.3: Entries in datasetTMB

Information Related to Entries in datasetTMB					
Uniprot ID	**Entry name**	**Protein names**	**TCDB**	**Organism**	Length
A0QR29	MSPA_MYCS2	Porin MspA	1.B.24.1.1	Mycobacterium smegmatis	211
A1JUB7	YADA_YERE8	Adhesin YadA		Yersinia enterocolitica serotype O:8 / biotype 1B	422
A5F934	OMPU_VIBC3	Porin OmpU		Vibrio cholerae serotype O1	341
E3PJ86	GSPD2_ECOH1	Secretin GspD 2		Escherichia coli O78:H11	686
E3PJ88	ASPS2_ECOH1	Pilotin AspS 2		Escherichia coli O78:H11	136
E6MXW0	OMPA_NEIMH	Major outer membrane protein P.IA		Neisseria meningitidis serogroup B / serotype 15	392
O18423	TXL_EISFE	Lysenin (efL1)	1.C.43.1.1	Eisenia fetida	297
O33407	ESTA_PSEAE	Esterase EstA (Autotransporter esterase EstA)	1.B.12.5.9	Pseudomonas aeruginosa	646
O88093	HBP_ECOLX	Hemoglobin-binding protease hbp autotransporter		Escherichia coli	1377
P00646	CEA3_ECOLX	Colicin-E3		Escherichia coli	551
P00747	PLMN_HUMAN	Plasminogen		Homo sapiens	810
P01031	CO5_HUMAN	Complement C5		Homo sapiens	1676
P02748	CO9_HUMAN	Complement component C9		Homo sapiens	559
P02787	TRFE_HUMAN	Serotransferrin (Transferrin)		Homo sapiens	698
P02929	TONB_ECOLI	Protein TonB	2.C.1.1.1	Escherichia coli	239
P02930	TOLC_ECOLI	Outer membrane protein TolC	1.B.17.1.1	Escherichia coli	493
P02931	OMPF_ECOLI	Outer membrane porin F	1.B.1.1.1	Escherichia coli	362
P02932	PHOE_ECOLI	Outer membrane porin PhoE	1.B.1.1.2	Escherichia coli	351
P02943	LAMB_ECOLI	Maltoporin (Lambda receptor protein)	1.B.3.1.1	Escherichia coli	446
P04419	CEA2_ECOLX	Colicin-E2	1.C.1.4.1	Escherichia coli	581
P05430	OMPA_NEIGO	Major outer membrane protein P.IA	1.B.1.5.1	Neisseria gonorrhoeae	326
P05695	PORP_PSEAE	Porin P (Outer membrane protein D1)	1.B.5.1.1	Pseudomonas aeruginosa	440
P05825	FEPA_ECOLI	Ferrienterobactin receptor	1.B.14.1.22	Escherichia coli	746
P06129	BTUB_ECOLI	Vitamin B12 transporter BtuB	1.B.14.3.1	Escherichia coli	614
P06716	CEIA_ECOLX	Colicin-Ia	1.C.1.1.1	Escherichia coli	626
P06970	FAED_ECOLX	Outer membrane usher protein FaeD	1.B.11.1.1	Escherichia coli	812
P06971	FHUA_ECOLI	Ferric hydroxamate receptor	1.B.14.1.2	Escherichia coli	747
P06996	OMPC_ECOLI	Outer membrane porin C	1.B.1.1.3	Escherichia coli	367
P07110	PAPC_ECOLX	Outer membrane usher protein PapC	1.B.11.2.1	Escherichia coli	836
P07357	CO8A_HUMAN	Complement component C8 alpha chain	1.C.39.3.1	Homo sapiens	584
P07358	CO8B_HUMAN	Complement component C8 beta chain		Homo sapiens	591
P07360	CO8G_HUMAN	Complement component C8 gamma chain		Homo sapiens	202
P08189	FIMF_ECOLI	Protein FimF		Escherichia coli	176
P08190	FIMG_ECOLI	Protein FimG		Escherichia coli	167
P08191	FIMH_ECOLI	Type 1 fimbrin D-mannose specific adhesin (Protein FimH)		Escherichia coli	300
P09167	AERA_AERHY	Aerolysin	1.C.4.1.1	Aeromonas hydrophila	493
P09169	OMPT_ECOLI	Protease 7 (EC 3.4.23.49) (Omptin) (Outer membrane protein 3B)	9.B.50.1.1	Escherichia coli	317
P09545	HLYA_VIBCH	Hemolysin	1.C.14.1.1	Vibrio cholerae serotype O1	741

Table A.3 continued from previous page

Information Related to Entries in datasetTMB					
Uniprot ID	**Entry name**	**Protein names**	**TCDB**	**Organism**	Length
P09616	HLA_STAAU	Alpha-hemolysin (Alpha-HL) (Alpha-toxin)	1.C.3.1.1	Staphylococcus aureus	319
P09883	CEA9_ECOLX	Colicin-E9	1.C.1.4.1	Escherichia coli	582
P0A071	HLGA_STAAM	Gamma-hemolysin component A (H-gamma-2) (H-gamma-II)		Staphylococcus aureus	309
P0A074	HLGA_STAAU	Gamma-hemolysin component A (H-gamma-2) (H-gamma-II)	1.C.3.4.2	Staphylococcus aureus	309
P0A077	HLGB_STAAU	Gamma-hemolysin component B (H-gamma-1) (H-gamma-I)	1.C.3.4.2	Staphylococcus aureus	325
P0A232	PA1_SALTI	Phospholipase A1 (Detergent-resistant phospholipase A)		Salmonella typhi	289
P0A263	OMPC_SALTY	Outer membrane porin C (Porin OmpC)		Salmonella typhimurium	378
P0A264	OMPC_SALTI	Outer membrane porin C		Salmonella typhi	378
P0A903	BAMC_ECOLI	Outer membrane protein assembly factor BamC	1.B.33.1.3	Escherichia coli	344
P0A910	OMPA_ECOLI	Outer membrane protein A (OmpA)	1.B.6.1.1	Escherichia coli	346
P0A911	OMPA_ECO57	Outer membrane protein A		Escherichia coli O157:H7	346
P0A915	OMPW_ECOLI	Outer membrane protein W	1.B.39.1.6	Escherichia coli	212
P0A916	OMPW_SHIFL	Outer membrane protein W		Shigella flexneri	212
P0A917	OMPX_ECOLI	Outer membrane protein X	1.B.6.2.1	Escherichia coli	171
P0A918	OMPX_ECOL6	Outer membrane protein X		Escherichia coli O6:H1	171
P0A919	OMPX_ECO57	Outer membrane protein X		Escherichia coli O157:H7	171
P0A920	OMPX_SHIFL	Outer membrane protein X		Shigella flexneri	171
P0A921	PA1_ECOLI	Phospholipase A1(OMPLA)		Escherichia coli	289
P0A922	PA1_ECO57	Phospholipase A1 (Outer membrane phospholipase A)		Escherichia coli O157:H7	289
P0A923	PA1_SHIFL	Phospholipase A1 (OMPLA)		Shigella flexneri	289
P0A927	TSX_ECOLI	Nucleoside-specific channel-forming protein Tsx	1.B.10.1.1	Escherichia coli	294
P0A928	TSX_ECO57	Nucleoside-specific channel-forming protein Tsx		Escherichia coli O157:H7	294
P0A929	TSX_SHIFL	Nucleoside-specific channel-forming protein Tsx		Shigella flexneri	294
P0A937	BAME_ECOLI	Outer membrane protein assembly factor BamE	1.B.33.1.3	Escherichia coli	113
P0A940	BAMA_ECOLI	Outer membrane protein assembly factor BamA (Omp85)		Escherichia coli	810
P0A941	BAMA_ECOL6	Outer membrane protein assembly factor BamA		Escherichia coli O6:H1	810
P0A942	BAMA_ECO57	Outer membrane protein assembly factor BamA		Escherichia coli O157:H7	810
P0A943	BAMA_SHIFL	Outer membrane protein assembly factor BamA	1.B.33.1.3	Shigella flexneri	810
P0AC02	BAMD_ECOLI	Outer membrane protein assembly factor BamD	1.B.33.1.3	Escherichia coli	245
P0ADC1	LPTE_ECOLI	LPS-assembly lipoprotein LptE (Rare lipoprotein B)	1.B.42.1.2	Escherichia coli	193
P0ADE4	TAMA_ECOLI	Translocation and assembly module subunit TamA (Autotransporter assembly factor TamA)	1.B.33.2.4	Escherichia coli	577
P0ADE5	TAMA_ECO57	Translocation and assembly module subunit TamA (Autotransporter assembly factor TamA)		Escherichia coli O157:H7	577
P0AEA2	CSGG_ECOLI	Curli production assembly/transport component CsgG	1.B.48.1.1	Escherichia coli	277
P0C2W0	YADA2_YEREN	Adhesin YadA	1.B.40.1.1	Yersinia enterocolitica	422
P0C6Q6	OMPU_VIBCH	Outer membrane protein U (Porin OmpU)	1.B.1.1.15	Vibrio cholerae serotype O1	341
P0DH58	OMPA_NEIMB	Major outer membrane protein P.IA (PIA) (Protein IA) (Class 1 protein)		Neisseria meningitidis serogroup B	392
P10384	FADL_ECOLI	Outer membrane FadL protein(Outer membrane flp protein)	1.B.9.1.1	Escherichia coli	446

Table A.3 continued from previous page

Information Related to Entries in datasetTMB					
Uniprot ID	**Entry name**	**Protein names**	**TCDB**	**Organism**	Length
P10643	CO7_HUMAN	Complement component C7		Homo sapiens	843
P11922	INVA_YERPS	Invasin	1.B.54.1.2	Yersinia pseudotuberculosis serotype I	985
P12643	BMP2_HUMAN	Bone morphogenetic protein 2 (BMP-2)		Homo sapiens	396
P13036	FECA_ECOLI	Fe(3+) dicitrate transport protein FecA	1.B.14.1.20	Escherichia coli	774
P13423	PAG_BACAN	Protective antigen (PA) (Anthrax toxins translocating protein)	1.C.42.1.1	Bacillus anthracis	764
P13671	CO6_HUMAN	Complement component C6	1.C.39.3.3	Homo sapiens	934
P13794	PORF_PSEAE	Outer membrane porin F	1.B.6.1.2	Pseudomonas aeruginosa	350
P15319	PAPD_ECOLX	Chaperone protein PapD		Escherichia coli	239
P16869	FHUE_ECOLI	FhuE receptor (Fe(III)-ferrioxamine B and Fe(III)-rhodotrulic acid)	1.B.14.1.1	Escherichia coli	729
P17315	CIRA_ECOLI	Colicin I receptor	1.B.14.1.4	Escherichia coli	663
P17811	COLY_YERPE	Coagulase/fibrinolysin (Plasminogen activator)		Yersinia pestis	312
P18195	OMPB1_NEIGO	Major outer membrane protein P.IB (PIB) (Protein IB) (Porin)		Neisseria gonorrhoeae	348
P18895	ALGE_PSEAE	Alginate production protein AlgE	1.B.13.1.1	Pseudomonas aeruginosa	490
P19809	EAE_ECO27	Intimin (Attaching and effacing protein) (Eae protein)		Escherichia coli O127:H6	939
P21796	VDAC1_HUMAN	Voltage-dependent anion-selective channel protein 1 (VDAC-1) (hVDAC1)		Homo sapiens	283
P22340	SCRY_SALTM	Sucrose porin	1.B.3.1.2	Salmonella typhimurium	505
P24017	OMPA_KLEPN	Outer membrane protein A (Outer membrane porin A)	1.B.6.1.11	Klebsiella pneumoniae	344
P24305	OMP32_DELAC	Outer membrane porin protein 32 (OMP32)	1.B.1.6.1	Delftia acidovorans	351
P24391	TOM40_NEUCR	Mitochondrial import receptor subunit tom40 (Protein MOM38)		Neurospora crassa	349
P26466	LAMB_SALTY	Maltoporin (Maltose-inducible porin)		Salmonella typhimurium	452
P30130	FIMD_ECOLI	Outer membrane usher protein FimD	1.B.11.3.9	Escherichia coli	878
P30690	OMPB1_NEIMB	Major outer membrane protein P.IB (PIB) (Porin)		Neisseria meningitidis serogroup B	331
P31243	PORI_RHOCA	Porin		Rhodobacter capsulatus	301
P31554	LPTD_ECOLI	LPS-assembly protein LptD	1.B.42.1.2	Escherichia coli	784
P31697	FIMC_ECOLI	Chaperone protein FimC		Escherichia coli	241
P31780	GSPD_AERHY	Secretin ExeD (General secretion pathway protein D)		Aeromonas hydrophila	678
P32722	PORD_PSEAE	Porin D (Imipenem/basic amino acid-specific outer membrane pore) (Outer membrane protein D2)	1.B.25.1.1	Pseudomonas aeruginosa	443
P32977	PORO_PSEAE	Porin O	1.B.5.1.2	Pseudomonas aeruginosa	438
P35077	FHAC_BORPE	Filamentous hemagglutinin transporter protein FhaC (TpsB transporter)	1.B.20.1.6	Bordetella pertussis	584
P35672	INVG_SALTY	Protein InvG	1.B.22.3.2	Salmonella typhimurium	562
P35818	GSPD_PSEAE	Secretin XcpQ (General secretion pathway protein D) (T2SS protein D)	1.B.22.1.2	Pseudomonas aeruginosa	658
P35916	VGFR3_HUMAN	Vascular endothelial growth factor receptor 3 (Fms-like tyrosine kinase 4) (FLT-4)		Homo sapiens	1363

Table A.3 continued from previous page

Information Related to Entries in datasetTMB					
Uniprot ID	**Entry name**	**Protein names**	**TCDB**	**Organism**	Length
P37001	PAGP_ECOLI	Lipid A palmitoyltransferase PagP (EC 2.3.1.251) (Lipid A acylation protein)		Escherichia coli	186
P37432	OMPF_SALTY	Outer membrane porin F (Outer membrane protein 1A) (Outer membrane protein B)		Salmonella typhimurium	363
P39767	PORI_RHOBL	Porin	1.B.7.1.3	Rhodobacter blasticus	289
P42512	FPTA_PSEAE	Fe(3+)-pyochelin receptor (Fe(III)-pyochelin receptor)	1.B.14.1.8	Pseudomonas aeruginosa	720
P43261	EAE_ECO57	Intimin (Attaching and effacing protein) (Eae protein) (Gamma-intimin)	1.B.54.1.1	Escherichia coli O157:H7	934
P45758	GSPD_ECOLI	Putative secretin GspD	1.B.22.1.3	Escherichia coli	650
P45779	GSPD_VIBCH	Secretin GspD (Cholera toxin secretion protein EpsD)		Vibrio cholerae serotype O1	674
P46359	FYUA_YERPE	Pesticin receptor (IRPC)		Yersinia pestis	673
P48632	FPVA_PSEAE	Ferripyoverdine receptor	1.B.14.1.6	Pseudomonas aeruginosa	815
P49767	VEGFC_HUMAN	Vascular endothelial growth factor C (VEGF-C) (Flt4-L)	9.B.88.2.1	Homo sapiens	419
P69434	PGAA_ECOLI	Poly-beta-1,6-N-acetyl-D-glucosamine export protein (PGA export protein)	1.B.55.1.1	Escherichia coli	807
P69856	NANC_ECOLI	NanR-regulated channel (Porin NanC)	1.B.35.2.1	Escherichia coli	238
P69857	NANC_ECO57	NanC (Porin NanC)		Escherichia coli O157:H7	238
P69858	NANC_SHIFL	NanC (Porin NanC)		Shigella flexneri	238
P75780	FIU_ECOLI	Catecholate siderophore receptor Fiu (TonB-dependent receptor Fiu)	1.B.14.1.9	Escherichia coli	760
P76045	OMPG_ECOLI	Outer membrane porin G (Outer membrane protein G)	1.B.21.1.1	Escherichia coli	301
P77211	CUSC_ECOLI	Cation efflux system protein CusC	1.B.17.3.5	Escherichia coli	457
P77774	BAMB_ECOLI	Outer membrane protein assembly factor BamB	1.B.33.1.3	Escherichia coli	392
Q03155	AIDA_ECOLX	AIDA-I autotransporter (AIDA)	1.B.12.1.1	Escherichia coli	1286
Q04884	OPAH_NEIGO	Opacity protein opA60 (Fragment)		Neisseria gonorrhoeae	238
Q05098	PFEA_PSEAE	Ferric enterobactin receptor	1.B.14.1.5	Pseudomonas aeruginosa	746
Q06584	PYS2_PSEAE	Pyocin-S2 (EC 3.1.-.-) (Killer protein)	1.C.1.4.2	Pseudomonas aeruginosa	689
Q16853	AOC3_HUMAN	Membrane primary amine oxidase (Copper amine oxidase) (HPAO) (Semicarbazide-sensitive amine oxidase) (VAP-1)		Homo sapiens	763
Q2FFA2	LUKL2_STAA3	Uncharacterized leukocidin-like protein 2		Staphylococcus aureus	351
Q2FFA3	LUKL1_STAA3	Uncharacterized leukocidin-like protein 1		Staphylococcus aureus	338
Q45340	BRKA_BORPE	BrkA autotransporter	1.B.12.2.3	Bordetella pertussis	1010
Q48473	OMPC_KLEPN	Outer membrane porin C (Outer membrane protein C) (Porin OmpC) (Porin ompk36)		Klebsiella pneumoniae	363
Q51397	OPRJ_PSEAE	Outer membrane protein OprJ		Pseudomonas aeruginosa	479
Q51487	OPRM_PSEAE	Outer membrane protein OprM	2.A.6.2.21	Pseudomonas aeruginosa	485
Q54450	HASA_SERMA	Hemophore HasA (Heme acquisition system protein A)	1.B.14.5.1	Serratia marcescens	188
Q5Y4Y6	GSDA3_MOUSE	Gasdermin-A3 (Gasdermin-3)		Mus musculus	464
Q60932	VDAC1_MOUSE	Voltage-dependent anion-selective channel protein 1 (mVDAC5)		Mus musculus	296
Q7BCK4	ICSA_SHIFL	Outer membrane protein IcsA autotransporter	1.B.12.1.2	Shigella flexneri	1102
Q7BSW5	ESPP_ECO57	Serine protease EspP (EC 3.4.21.-)	1.B.12.4.3	Escherichia coli O157:H7	1300
Q7CJV2	LPTE_YERPE	LPS-assembly lipoprotein LptE		Yersinia pestis	207

Table A.3 continued from previous page

Information Related to Entries in datasetTMB					
Uniprot ID	**Entry name**	**Protein names**	**TCDB**	**Organism**	Length
Q83LX4	LPTE_SHIFL	LPS-assembly lipoprotein LptE		Shigella flexneri	193
Q83SQ0	LPTD_SHIFL	LPS-assembly protein LptD		Shigella flexneri	784
Q8CVI4	LAMB_ECOL6	Maltoporin (Maltose-inducible porin)		Escherichia coli O6:H1	446
Q8CVW1	OMPC_ECOL6	Outer membrane porin C (Outer membrane protein 1B)		Escherichia coli O6:H1	375
Q8ZIK3	LPTD_YERPE	LPS-assembly protein LptD		Yersinia pestis	780
Q8ZPC9	ABDH_SALTY	Gamma-aminobutyraldehyde dehydrogenase (ABALDH)		Salmonella typhimurium	474
Q8ZQZ7	LPTE_SALTY	LPS-assembly lipoprotein LptE		Salmonella typhimurium	196
Q8ZRP0	BAMA_SALTY	Outer membrane protein assembly factor BamA		Salmonella typhimurium	804
Q8ZRW0	LPTD_SALTY	LPS-assembly protein LptD		Salmonella typhimurium	786
Q934G3	KDGM_DICD3	Oligogalacturonate-specific porin KdgM	1.B.35.1.1	Dickeya dadantii	236
Q99RL1	HLGC_STAAM	Gamma-hemolysin component C		Staphylococcus aureus	315
Q9HVD1	PAGL_PSEAE	Lipid A deacylase PagL		Pseudomonas aeruginosa	173
Q9I5U2	LPTD_PSEAE	LPS-assembly protein LptD		Pseudomonas aeruginosa	924
Q9JZN9	Y964_NEIMB	Probable TonB-dependent receptor NMB0964	1.B.14.2.9	Neisseria meningitidis serogroup B	758
Q9K0U9	TBP1_NEIMB	Transferrin-binding protein 1		Neisseria meningitidis serogroup B	915
Q9TUM0	TRFL_CAMDR	Lactotransferrin (Lactoferrin)		Camelus dromedarius	708
Q9X2V7	MCJA_ECOLX	Microcin J25 (MccJ25)	9.A.52.1.1	Escherichia coli	58

Table A.4: Entries in datasetTMB with 3D structure in PDB.

Uniprot ID	PDB ID	Uniprot ID	PDB ID	Uniprot ID	PDB ID
A0QR29	1UUN	P0A919		P35916	4BSJ
A1JUB7	2LME	P0A920		P37001	1MM4
A5F934	6EHB	P0A921	1FW2	P37432	3NSG
E3PJ86	3OSS	P0A922		P39767	1BH3
E3PJ88	5ZDH	P0A923		P42512	1XKW
E6MXW0	2MPA	P0A927	1TLW	P43261	2ZQK
O18423	3ZX7	P0A928		P45758	5WQ7
O33407	3KVN	P0A929		P45779	5WQ8
O88093	1WXR	P0A937	2KM7	P46359	4EPA
P00646	1E44	P0A940	2QCZ	P48632	1XKH
P00747	1B2I	P0A941		P49767	2X1W
P01031	1CFA	P0A942		P69434	4Y25
P02748	5FMW	P0A943		P69856	2WJQ
P02787	1A8E	P0AC02	2YHC	P69857	
P02929	1IHR	P0ADC1	4NHR	P69858	
P02930	1EK9	P0ADE4	2LY3	P75780	6BPM
P02931	1BT9	P0ADE5		P76045	2F1C
P02932	1PHO	P0AEA2	4UV2	P77211	3PIK
P02943	1AF6	P0C2W0	3H7X	P77774	2YH3
P04419	2YSU	P0C6Q6	5ONU	Q03155	4MEE
P05430		P0DH58		Q04884	2MAF
P05695	2O4V	P10384	1T16	Q05098	5M9B
P05825	1FEP	P10643	2WCY	Q06584	4QKO
P06129	1NQE	P11922	1CWV	Q16853	1PU4
P06716	1CII	P12643	1ES7	Q2FFA2	
P06970		P13036	1KMO	Q2FFA3	
P06971	1BY3	P13423	1ACC	Q45340	3QQ2
P06996	2J1N	P13671	3T5O	Q48473	1OSM
P07110	2KT6	P13794	4RLC	Q51397	5AZS
P07357	2QOS	P15319	1N0L	Q51487	1WP1
P07358	3OJY	P16869	6E4V	Q54450	1B2V
P07360	1IW2	P17315	2HDF	Q5Y4Y6	5B5R
P08189	2JMR	P17811	2X4M	Q60932	3EMN
P08190	3BFQ	P18195		Q7BCK4	3ML3
P08191	1KIU	P18895	3RBH	Q7BSW5	2QOM
P09167	1PRE	P19809	1E5U	Q7CJV2	5IXM
P09169	1I78	P21796	2JK4	Q83LX4	4Q35
P09545	1XEZ	P22340	1A0S	Q83SQ0	4Q35
P09616	3M2L	P24017	2K0L	Q8CVI4	2VDA
P09883	1BXI	P24305	1E54	Q8CVW1	2XE1
P0A071	3B07	P24391	5O8O	Q8ZIK3	5IXM
P0A074	2QK7	P26466	1MPR	Q8ZPC9	6C43
P0A077	1LKF	P30130	1ZDV	Q8ZQZ7	4N4R
P0A232	5DQX	P30690	3WI4	Q8ZRP0	5OR1
P0A263		P31243	2POR	Q8ZRW0	4N4R

Table A.4 continued from previous page

Uniprot ID	PDB ID	Uniprot ID	PDB ID	Uniprot ID	PDB ID
P0A264	1IIV	P31554	4RHB	Q934G3	4FQE
P0A903	2LAE	P31697	1BF8	Q99RL1	4P1X
P0A910	1BXW	P31780	6I1X	Q9HVD1	2ERV
P0A911		P32722	2ODJ	Q9I5U2	5IVA
P0A915	2F1T	P32977	4RJW	Q9JZN9	4RDR
P0A916		P35077	3NJT	Q9K0U9	3V89
P0A917	1ORM	P35672	2Y9K	Q9TUM0	1DTZ
P0A918		P35818	4E9J	Q9X2V7	1PP5

A.4 Individual NorMD Scores

MSA	NorMD Score for MSAs from Different Programs (for Top10 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
A0QR29	0.324	0.492	0.408	0.491	0.443
A1JUB7	0.455	0.607	0.533	0.592	0.601
A5F934	1.000	1.000	1.000	1.000	1.000
E3PJ86	0.758	0.739	0.727	0.684	0.722
E6MXW0	0.913	0.926	0.912	0.914	0.913
O18423	0.181	0.439	0.209	0.216	0.185
O33407	0.667	0.674	0.676	0.676	0.668
O88093	0.706	0.703	0.708	0.708	0.704
P00646	0.233	0.594	0.369	0.281	0.254
P00747	0.733	0.768	0.653	0.696	0.778
P01031	0.237	0.468	0.272	0.228	0.199
P02929	0.669	1.016	0.574	0.579	0.564
P02930	0.661	0.655	0.653	0.655	0.652
P02931	0.959	0.830	0.915	0.966	0.961
P02932	0.727	0.710	0.740	0.708	0.714
P02943	0.387	0.514	0.422	0.403	0.416
P05430	0.828	0.836	0.844	0.845	0.834
P05695	0.839	0.838	0.860	0.844	0.838
P05825	0.709	0.701	0.729	0.737	0.733
P06129	0.667	0.673	0.673	0.672	0.670
P06716	0.184	0.386	0.228	0.177	0.173
P06970	0.684	0.685	0.685	0.686	0.689
P06971	0.705	0.704	0.726	0.668	0.667
P06996	0.728	0.766	0.823	0.770	0.771
P07110	0.696	0.698	0.697	0.700	0.698
P08189	0.315	0.564	0.332	0.427	0.328
P08190	0.396	0.543	0.345	0.461	0.427
P08191	0.275	0.492	0.281	0.295	0.219
P09167	0.282	0.545	0.319	0.382	0.355
P09169	0.555	0.566	0.570	0.571	0.563
P09545	0.250	0.490	0.274	0.292	0.294
P0A071	0.179	0.494	0.197	0.193	0.162
P0A074	0.179	0.494	0.197	0.193	0.162
P0A077	0.182	0.364	0.203	0.261	0.232

Table A.5 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
P0A232	0.695	0.706	0.696	0.702	0.702
P0A263	0.892	0.865	0.948	0.783	0.902
P0A264	0.892	0.865	0.948	0.783	0.902
P0A903	0.187	0.342	0.231	0.277	0.239
P0A910	0.795	0.814	0.773	0.802	0.795
P0A911	0.795	0.814	0.773	0.802	0.795
P0A915	0.714	0.717	0.719	0.726	0.733
P0A916	0.714	0.717	0.719	0.726	0.733
P0A917	0.793	0.805	0.787	0.783	0.768
P0A918	0.793	0.805	0.787	0.783	0.768
P0A919	0.793	0.805	0.787	0.783	0.768
P0A920	0.793	0.805	0.787	0.783	0.768
P0A921	0.696	0.707	0.699	0.703	0.703
P0A922	0.696	0.707	0.699	0.703	0.703
P0A923	0.696	0.707	0.699	0.703	0.703
P0A927	0.705	0.713	0.715	0.716	0.715
P0A928	0.705	0.713	0.715	0.716	0.715
P0A929	0.705	0.713	0.715	0.716	0.715
P0A937	1.407	1.545	1.276	1.439	1.174
P0A940	0.746	0.746	0.714	0.746	0.746
P0A941	0.746	0.746	0.714	0.746	0.746
P0A942	0.746	0.746	0.714	0.746	0.746
P0A943	0.746	0.746	0.714	0.746	0.746
P0AC02	0.263	0.546	0.266	0.299	0.229
P0ADE4	0.724	0.632	0.741	0.613	0.611
P0ADE5	0.724	0.632	0.741	0.613	0.611
P0AEA2	0.707	0.712	0.693	0.717	0.712
P0C2W0	0.621	0.687	0.592	0.656	0.641
P0C6Q6	1.000	1.000	1.000	1.000	1.000
P0DH58	0.847	0.871	0.876	0.855	0.849
P10384	1.000	1.000	0.946	0.957	0.957
P11922	0.968	1.039	1.030	1.003	0.963
P13036	0.757	0.756	0.765	0.760	0.761
P13423	0.183	0.376	0.286	0.246	0.197
P13794	0.674	0.965	0.952	0.781	0.778
P16869	0.659	0.673	0.656	0.670	0.665
P17315	0.819	0.833	0.829	0.830	0.821
P17811	0.661	0.668	0.651	0.667	0.666
P18195	0.821	0.878	0.828	0.861	0.828
P18895	0.720	0.732	0.733	0.733	0.726
P19809	0.615	0.753	0.739	0.734	0.744
P21796	0.283	0.432	0.422	0.279	0.236
P22340	0.875	0.881	0.848	0.880	0.796
P24017	0.746	0.783	0.776	0.769	0.746
P24305	0.831	0.848	0.838	0.845	0.831
P24391	0.177	0.450	0.229	0.256	0.220
P26466	1.095	0.976	0.418	0.522	0.426
P30130	0.687	0.690	0.648	0.692	0.695
P30690	0.782	0.685	0.804	0.827	0.787
P31243	0.775	0.706	0.665	0.664	0.776

Table A.5 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
P31554	0.736	0.745	0.747	0.735	0.735
P31697	0.251	0.453	0.268	0.222	0.162
P31780	0.733	0.736	0.735	0.735	0.735
P32722	0.669	0.669	0.670	0.671	0.652
P32977	0.987	0.989	0.998	0.993	0.988
P35077	0.635	0.651	0.645	0.649	0.645
P35672	0.778	0.781	0.777	0.781	0.778
P35818	0.821	0.743	0.834	0.788	0.817
P37001	0.791	0.663	0.659	0.520	0.519
P37432	0.826	0.717	0.837	0.864	0.857
P39767	0.722	0.740	0.729	0.728	0.728
P42512	0.644	0.645	0.588	0.649	0.649
P43261	0.926	0.995	0.984	0.935	0.950
P45758	0.783	0.784	0.786	0.786	0.782
P45779	0.747	0.749	0.748	0.750	0.747
P46359	0.715	0.731	0.719	0.720	0.719
P48632	0.748	0.746	0.747	0.750	0.747
P69434	0.238	0.449	0.306	0.320	0.297
P69856	0.605	0.636	0.603	0.632	0.622
P69857	0.605	0.636	0.603	0.632	0.622
P69858	0.605	0.636	0.603	0.632	0.622
P75780	0.675	0.671	0.675	0.662	0.656
P76045	0.793	0.801	0.793	0.803	0.796
P77211	0.643	0.645	0.629	0.639	0.642
P77774	0.157	0.386	0.260	0.203	0.150
Q03155	1.004	1.062	1.031	0.973	1.004
Q04884	1.111	1.152	1.156	1.065	1.146
Q05098	0.773	0.768	0.774	0.768	0.775
Q16853	0.191	0.441	0.243	0.229	0.227
Q2FFA3	0.158	0.462	0.241	0.199	0.176
Q45340	0.746	0.772	0.763	0.758	0.753
Q48473	0.954	0.958	0.954	0.958	0.959
Q51397	0.724	0.724	0.724	0.728	0.727
Q51487	0.722	0.724	0.717	0.710	0.725
Q54450	0.261	0.713	0.346	0.274	0.239
Q5Y4Y6	0.274	0.492	0.252	0.257	0.240
Q60932	0.241	0.590	0.295	0.286	0.262
Q7BCK4	0.802	0.860	0.860	0.819	0.805
Q7BSW5	0.727	0.747	0.740	0.736	0.733
Q7CJV2	0.217	0.393	0.243	0.235	0.177
Q83SQ0	0.732	0.741	0.747	0.732	0.731
Q8CVI4	0.387	0.954	0.379	0.403	0.416
Q8CVW1	0.715	0.756	0.697	0.752	0.757
Q8ZIK3	0.741	0.739	0.729	0.740	0.741
Q8ZPC9	0.266	0.677	0.355	0.381	0.398
Q8ZRP0	0.763	0.780	0.726	0.763	0.763
Q8ZRW0	0.733	0.744	0.747	0.731	0.733
Q934G3	0.739	0.767	0.731	0.741	0.740
Q99RL1	0.253	0.535	0.365	0.319	0.277
Q9HVD1	0.685	0.687	0.686	0.696	0.697

Table A.5 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
Q9I5U2	0.784	0.868	0.786	0.870	0.871
Q9JZN9	0.715	0.719	0.723	0.723	0.719
Q9K0U9	0.780	0.797	0.798	0.797	0.791
Average	0.648	0.716	0.656	0.654	0.645
Variance	0.061	0.030	0.052	0.052	0.056

Table A.5: NorMD Score for MSAs from Different Programs (for Top10 using OMPdb70)

	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
MSA	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A1JUB7	2.090	2.852	2.619	2.538	2.388
A5F934	1.224	1.253	1.313	1.173	1.187
E3PJ86	0.988	1.006	1.000	1.002	1.004
E6MXW0	1.062	1.113	1.059	1.077	1.077
O18423	0.280	1.829	1.448	1.060	0.941
O33407	0.963	1.000	1.026	1.026	1.002
O88093	0.932	0.952	0.946	0.939	0.934
P00646	2.219	1.963	1.627	1.747	1.842
P00747	1.390	0.177	1.830	0.178	0.164
P01031	1.138	1.174	1.311	1.151	1.143
P02748	0.963	0.989	0.972	0.980	0.980
P02787	0.830	0.828	0.850	0.830	0.832
P02929	0.504	0.757	0.488	0.451	0.455
P02930	0.539	0.524	0.545	0.592	0.591
P02931	0.787	0.779	0.777	0.778	0.784
P02932	0.886	0.874	0.884	0.884	0.889
P04419	1.804	2.770	1.466	2.334	2.158
P05430	1.143	1.167	1.026	1.149	1.149
P05825	1.250	1.441	1.248	1.322	1.329
P06716	0.811	0.874	0.840	0.806	0.795
P06970	0.740	0.761	0.755	0.760	0.753
P06971	0.698	0.777	0.759	0.755	0.743
P06996	1.257	1.265	1.124	1.269	1.269
P07110	0.730	0.749	0.750	0.755	0.754
P07357	0.948	1.015	0.979	1.011	0.985
P07358	1.008	1.091	1.018	1.099	1.087
P07360	0.928	0.995	1.031	0.951	0.962
P08189	0.721	0.726	0.714	0.759	0.749
P08190	0.638	0.692	0.680	0.766	0.685
P09167	0.535	1.484	0.891	0.776	0.634
P09169	0.301	1.082	0.687	0.609	0.513
P09883	1.579	1.760	1.166	1.854	1.759
P0A071	0.958	1.000	1.000	1.000	1.000
P0A074	0.958	1.000	1.000	1.000	1.000
P0A077	1.171	1.223	1.179	1.231	1.187
P0A263	0.956	0.953	0.965	0.961	0.962
P0A264	0.956	0.953	0.965	0.961	0.962

Table A.6 continued from previous page

	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
P0A903	0.794	0.838	0.835	0.817	0.812
P0A910	1.000	1.000	1.000	1.000	1.000
P0A911	1.000	1.000	1.000	1.000	1.000
P0A915	0.573	0.570	0.647	0.708	0.701
P0A916	0.573	0.570	0.647	0.708	0.701
P0A917	0.835	0.950	0.513	0.465	0.597
P0A918	0.835	0.950	0.513	0.465	0.597
P0A919	0.835	0.950	0.513	0.465	0.597
P0A920	0.835	0.950	0.513	0.465	0.597
P0A927	2.128	2.675	2.203	1.836	2.730
P0A928	2.128	2.675	2.203	1.836	2.730
P0A929	2.128	2.675	2.203	1.836	2.730
P0A937	0.779	0.772	0.782	0.824	0.836
P0AC02	0.567	0.572	0.572	0.593	0.591
P0AEA2	0.713	1.523	0.967	0.906	0.456
P0C2W0	2.088	3.023	2.790	2.780	2.534
P0C6Q6	1.572	1.585	1.428	1.574	1.522
P0DH58	1.062	1.113	1.059	1.077	1.077
P10384	1.013	1.074	1.062	1.034	1.036
P10643	1.038	1.054	1.088	1.076	1.068
P11922	0.927	1.132	1.000	1.059	1.042
P12643	1.091	1.122	1.055	1.113	1.092
P13036	0.905	0.996	0.953	0.967	0.938
P13671	1.001	1.024	1.091	1.048	1.046
P15319	0.753	0.765	0.765	0.783	0.782
P16869	0.762	0.785	0.774	0.782	0.783
P17315	1.274	1.487	1.267	1.391	1.369
P18195	1.048	1.086	1.026	1.054	1.054
P18895	0.438	0.670	0.495	0.518	0.601
P19809	1.303	1.517	1.496	1.516	1.513
P21796	1.000	1.000	1.000	1.000	1.000
P22340	1.313	2.577	1.886	1.317	1.328
P24017	1.000	1.000	1.000	1.000	1.000
P24305	1.405	1.657	1.549	1.531	1.483
P24391	1.003	0.972	1.039	1.030	1.028
P26466	1.000	1.000	1.000	1.000	1.000
P30130	0.768	0.778	0.609	0.396	0.776
P30690	1.053	1.091	1.025	1.059	1.059
P31243	0.143	0.276	0.179	0.133	0.098
P31554	1.000	1.000	1.000	1.000	1.000
P31697	0.744	0.745	0.747	0.757	0.757
P31780	0.959	0.973	0.884	0.973	0.978
P32722	0.372	0.513	0.319	0.336	0.328
P35077	0.412	0.450	0.435	0.422	0.417
P35672	0.705	0.736	0.701	0.713	0.725
P35818	0.959	0.980	0.968	0.976	0.978
P35916	0.869	0.923	0.874	0.876	0.873
P37001	1.000	1.000	1.000	1.000	1.000
P37432	0.808	0.817	0.797	0.805	0.809
P42512	0.836	0.867	0.864	0.861	0.858
P43261	0.953	1.164	1.005	1.062	1.043

Table A.6 continued from previous page

	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
P45758	0.868	0.877	0.857	0.877	0.878
P45779	0.988	1.006	0.891	1.005	1.005
P46359	0.842	0.922	0.827	0.917	0.917
P48632	0.787	0.825	0.797	0.821	0.820
P49767	1.140	1.386	1.504	1.267	1.241
P69856	0.609	0.932	0.568	0.675	0.660
P69857	0.609	0.932	0.568	0.675	0.660
P69858	0.609	0.932	0.568	0.675	0.660
P75780	0.785	0.816	0.807	0.820	0.819
P77211	0.841	0.859	0.931	0.847	0.846
P77774	0.668	0.674	0.649	0.677	0.671
Q03155	1.041	0.893	0.630	0.575	0.595
Q04884	1.000	1.000	1.000	1.000	1.000
Q05098	1.147	1.258	1.171	1.174	1.188
Q06584	0.918	1.310	0.604	0.667	0.789
Q16853	1.013	1.051	1.025	1.012	1.012
Q2FFA2	1.408	1.439	1.384	1.423	1.416
Q2FFA3	1.221	1.277	1.179	1.258	1.230
Q45340	0.617	0.686	0.626	0.616	0.615
Q48473	1.018	1.033	1.022	0.944	1.023
Q51397	0.756	0.761	0.770	0.756	0.757
Q51487	0.762	0.767	0.769	0.762	0.762
Q54450	0.760	2.253	1.520	1.033	1.103
Q5Y4Y6	0.729	0.776	0.757	0.774	0.755
Q60932	1.000	1.000	1.000	1.000	1.000
Q7BSW5	0.917	0.926	0.890	0.436	0.921
Q7CJV2	0.841	0.877	0.903	0.928	0.891
Q8CVW1	1.020	0.943	0.948	0.893	1.026
Q8ZIK3	0.881	0.878	0.886	1.027	0.883
Q8ZPC9	1.000	1.000	1.000	0.882	1.000
Q8ZRP0	1.000	1.000	1.000	1.000	1.000
Q8ZRW0	1.000	1.000	1.000	1.000	1.000
Q99RL1	0.958	1.000	1.000	1.000	1.000
Q9I5U2	0.949	0.946	0.950	1.000	0.951
Q9JZN9	1.104	1.279	1.121	0.950	1.144
Q9TUM0	0.970	0.972	1.058	1.178	0.972
Average	0.967	1.096	0.996	0.976	0.996
Variance	0.134	0.254	0.167	0.161	0.197

Table A.6: NorMD Score for MSAs from Different Programs (for Top10 using swissprot)

MSA	**NorMD Score for MSAs from Different Programs (for Top10 using unnirefOMBB100)**				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A0QR29	0.171	0.377	0.176	0.227	0.190
A1JUB7	1.403	1.188	1.601	1.397	1.363
A5F934	0.407	0.592	0.405	0.384	0.348
E6MXW0	0.430	2.114	1.705	0.979	0.954
O33407	0.966	0.966	0.969	0.966	0.966

Table A.7 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
O88093	1.000	1.000	1.000	1.000	1.000
P00747	1.054	1.252	1.162	1.010	0.987
P02929	0.595	1.114	0.799	0.686	0.620
P02931	0.514	0.746	0.522	0.347	0.330
P02932	0.210	0.384	0.428	0.332	0.338
P02943	0.531	1.150	0.947	0.651	0.571
P05430	1.138	1.416	0.968	1.068	1.111
P05695	0.295	0.674	0.406	0.364	0.314
P05825	0.417	0.603	0.338	0.326	0.428
P06129	0.859	1.013	0.400	0.598	0.599
P06971	0.525	0.769	0.517	0.516	0.465
P08190	0.586	1.117	0.769	0.672	0.580
P09169	1.175	1.205	1.205	1.184	1.184
P09616	2.767	4.398	4.029	3.375	3.197
P0A232	0.843	1.480	1.714	0.745	0.418
P0A263	0.300	0.697	0.468	0.295	0.327
P0A264	0.300	0.697	0.468	0.295	0.327
P0A903	0.668	0.808	0.749	0.722	0.669
P0A910	0.693	0.744	0.762	0.753	0.712
P0A911	0.693	0.744	0.762	0.753	0.712
P0A915	0.779	0.849	0.820	0.828	0.810
P0A916	0.779	0.849	0.820	0.828	0.810
P0A917	1.000	1.000	1.000	1.000	1.000
P0A918	1.000	1.000	1.000	1.000	1.000
P0A919	1.000	1.000	1.000	1.000	1.000
P0A920	1.000	1.000	1.000	1.000	1.000
P0A921	2.438	3.595	4.590	3.040	2.777
P0A922	2.438	3.595	4.590	3.040	2.777
P0A923	2.438	3.595	4.590	3.040	2.777
P0A937	0.822	0.820	0.787	0.824	0.839
P0A940	1.147	1.155	1.161	1.159	1.148
P0A941	1.147	1.155	1.161	1.159	1.148
P0A942	1.147	1.155	1.161	1.159	1.148
P0A943	1.147	1.155	1.161	1.159	1.148
P0AC02	0.621	0.629	0.616	0.631	0.631
P0ADE4	0.716	0.736	0.772	0.707	0.745
P0ADE5	0.716	0.736	0.772	0.707	0.745
P0C2W0	1.260	1.076	1.424	1.224	1.224
P0C6Q6	0.217	0.608	0.305	0.268	0.198
P0DH58	0.416	2.025	1.652	0.949	0.925
P10384	0.612	0.636	0.621	0.634	0.633
P11922	0.347	0.794	0.501	0.441	0.440
P13036	0.805	1.351	0.999	0.919	0.724
P13794	0.712	0.745	0.677	0.656	0.612
P16869	0.381	1.083	0.550	0.377	0.419
P17315	0.754	0.860	0.678	0.648	0.627
P17811	1.113	1.130	1.121	1.121	1.121
P18195	0.195	0.593	0.295	0.289	0.282
P18895	0.567	1.151	0.701	0.713	0.637
P19809	0.365	0.767	0.470	0.423	0.409
P24017	0.657	0.790	0.655	0.643	0.647

Table A.7 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
P24305	0.725	0.950	1.069	0.669	0.499
P26466	0.377	0.786	0.568	0.389	0.353
P30130	0.424	1.190	0.690	0.613	0.470
P30690	0.370	0.843	0.682	0.562	0.444
P35818	2.023	3.251	1.041	0.580	0.355
P43261	0.273	0.597	0.386	0.322	0.307
P46359	0.465	0.819	0.604	0.551	0.546
P48632	0.605	0.832	0.541	0.553	0.553
P69434	0.293	1.022	0.641	0.480	0.623
P69856	0.233	0.314	0.221	0.247	0.255
P69857	0.233	0.314	0.221	0.247	0.255
P69858	0.233	0.314	0.221	0.247	0.255
P75780	0.284	0.465	0.286	0.301	0.369
P76045	0.201	0.522	0.461	0.335	0.234
P77774	0.713	0.727	0.771	0.735	0.735
Q03155	1.374	0.848	0.881	0.761	1.266
Q04884	1.000	1.000	1.000	1.000	1.000
Q05098	0.602	0.712	0.469	0.465	0.470
Q2FFA2	2.464	4.229	3.355	3.105	3.041
Q45340	1.017	1.085	1.033	1.020	0.956
Q48473	0.127	0.419	0.197	0.209	0.240
Q54450	2.345	2.741	2.376	2.596	2.019
Q5Y4Y6	0.990	1.235	1.103	1.015	1.022
Q7BCK4	1.000	1.000	1.000	1.000	1.000
Q7BSW5	1.000	1.000	1.000	1.000	1.000
Q8CVI4	0.425	1.177	0.706	0.477	0.420
Q8CVW1	0.623	1.027	0.795	0.482	0.475
Q8ZRP0	1.133	1.148	1.253	1.143	1.144
Q934G3	0.389	0.638	0.315	0.335	0.310
Q9HVD1	0.173	0.469	0.342	0.366	0.373
Q9I5U2	0.853	2.118	1.269	0.938	0.944
Q9JZN9	0.332	0.880	0.605	0.400	0.338
Average	0.802	1.131	0.989	0.845	0.811
Variance	0.334	0.680	0.806	0.445	0.385

Table A.7: NorMD Score for MSAs from Different Programs (for Top10 using unirefOMBB100)

MSA	NorMD Score for MSAs from Different Programs (for Top25 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A0QR29	0.250	0.368	0.286	0.337	0.351
A1JUB7	0.382	0.460	0.467	0.456	0.466
A5F934	0.833	0.830	0.806	0.840	0.832
E3PJ86	0.706	0.721	0.699	0.679	0.678
E6MXW0	0.887	0.981	0.856	0.877	0.877
O33407	0.648	0.661	0.658	0.658	0.657
O88093	0.725	0.743	0.727	0.750	0.742
P02929	0.308	0.645	0.456	0.451	0.443
P02930	0.699	0.671	0.697	0.660	0.660

Table A.8 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
P02931	0.762	0.768	0.772	0.769	0.767
P02932	0.756	0.752	0.759	0.756	0.751
P02943	0.793	0.805	0.762	0.760	0.760
P05430	0.887	0.939	0.854	0.858	0.858
P05695	0.844	0.893	0.879	0.872	0.861
P05825	0.768	0.768	0.775	0.770	0.769
P06129	0.724	0.723	0.728	0.730	0.726
P06970	0.771	0.773	0.781	0.778	0.778
P06971	0.673	0.675	0.674	0.671	0.667
P06996	0.780	0.786	0.787	0.790	0.789
P07110	0.705	0.706	0.707	0.709	0.706
P09169	0.590	0.602	0.593	0.605	0.598
P0A232	0.715	0.718	0.729	0.733	0.731
P0A263	0.725	0.764	0.737	0.737	0.733
P0A264	0.725	0.764	0.737	0.737	0.733
P0A910	0.722	0.723	0.724	0.729	0.728
P0A911	0.722	0.723	0.724	0.729	0.728
P0A915	0.745	0.823	0.994	0.578	0.634
P0A916	0.745	0.823	0.994	0.578	0.634
P0A917	0.654	0.672	0.680	0.678	0.679
P0A918	0.654	0.672	0.680	0.678	0.679
P0A919	0.654	0.672	0.680	0.678	0.679
P0A920	0.654	0.672	0.680	0.678	0.679
P0A921	0.724	0.734	0.733	0.740	0.733
P0A922	0.724	0.734	0.733	0.740	0.733
P0A923	0.724	0.734	0.733	0.740	0.733
P0A927	0.679	0.711	0.708	0.694	0.690
P0A928	0.679	0.711	0.708	0.694	0.690
P0A929	0.679	0.711	0.708	0.694	0.690
P0A940	0.648	0.652	0.645	0.651	0.648
P0A941	0.648	0.652	0.645	0.651	0.648
P0A942	0.648	0.652	0.645	0.651	0.648
P0A943	0.648	0.652	0.645	0.651	0.648
P0ADE4	0.625	0.625	0.625	0.622	0.618
P0ADE5	0.625	0.625	0.625	0.622	0.618
P0AEA2	0.727	0.731	0.722	0.737	0.729
P0C2W0	0.357	0.497	0.453	0.481	0.489
P0C6Q6	0.830	0.849	0.843	0.863	0.828
P0DH58	0.872	0.981	0.854	0.873	0.873
P10384	0.800	0.800	0.799	0.806	0.806
P11922	0.923	0.741	0.729	0.740	0.739
P13036	0.828	0.821	0.825	0.829	0.828
P13794	0.707	0.745	0.763	0.740	0.751
P16869	0.669	0.685	0.683	0.683	0.671
P17315	0.778	0.788	0.783	0.787	0.771
P17811	0.637	0.651	0.640	0.649	0.645
P18195	0.851	0.939	0.847	0.862	0.855
P18895	0.786	0.748	0.813	0.789	0.800
P19809	0.613	0.608	0.631	0.588	0.610
P21796	0.167	0.372	0.244	0.215	0.218
P22340	0.731	0.747	0.739	0.739	0.737

Table A.8 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
P24017	0.714	0.709	0.717	0.707	0.713
P24305	0.816	0.837	0.866	0.836	0.869
P26466	0.796	0.821	0.759	0.775	0.769
P30130	0.722	0.725	0.699	0.725	0.723
P30690	0.799	0.955	0.784	0.808	0.809
P31243	0.804	0.849	0.827	0.816	0.814
P31554	0.632	0.635	0.635	0.637	0.633
P31780	0.805	0.806	0.807	0.810	0.806
P32722	0.693	0.707	0.705	0.707	0.700
P32977	0.809	0.841	0.832	0.850	0.837
P35077	0.625	0.637	0.629	0.637	0.633
P35672	0.788	0.796	0.791	0.793	0.792
P35818	0.823	0.834	0.825	0.832	0.822
P37001	0.743	0.683	1.126	0.513	0.512
P37432	0.763	0.764	0.758	0.762	0.760
P39767	0.795	0.796	0.758	0.805	0.807
P42512	0.602	0.607	0.572	0.608	0.609
P43261	0.661	0.631	0.648	0.633	0.650
P45758	0.761	0.766	0.764	0.765	0.762
P45779	1.075	0.782	0.517	0.501	0.499
P46359	0.701	0.715	0.708	0.712	0.710
P48632	0.755	0.758	0.758	0.760	0.757
P69434	0.157	0.255	0.202	0.162	0.164
P69856	0.524	0.538	0.536	0.560	0.528
P69857	0.524	0.538	0.536	0.560	0.528
P69858	0.524	0.538	0.536	0.560	0.528
P75780	0.721	0.691	0.671	0.675	0.672
P76045	0.252	0.446	0.363	0.339	0.362
P77211	0.693	0.695	0.692	0.699	0.696
Q03155	0.727	0.761	0.756	0.722	0.723
Q04884	0.809	0.865	0.840	0.857	0.862
Q05098	0.779	0.781	0.778	0.779	0.778
Q45340	0.553	0.601	0.580	0.571	0.600
Q48473	0.772	0.784	0.771	0.777	0.770
Q51397	0.706	0.711	0.708	0.711	0.710
Q51487	0.709	0.714	0.708	0.716	0.714
Q60932	0.163	0.308	0.260	0.226	0.229
Q7BCK4	0.646	0.676	0.679	0.662	0.654
Q7BSW5	0.831	0.837	0.817	0.845	0.836
Q83SQ0	0.631	0.634	0.635	0.635	0.632
Q8CVI4	0.787	0.807	0.752	0.759	0.758
Q8CVW1	0.798	0.809	0.779	0.791	0.804
Q8ZIK3	0.634	0.637	0.635	0.639	0.636
Q8ZRP0	0.649	0.654	0.644	0.651	0.648
Q8ZRW0	0.633	0.638	0.635	0.637	0.634
Q934G3	0.719	0.707	0.748	0.726	0.728
Q9HVD1	0.683	0.687	0.682	0.692	0.692
Q9I5U2	0.910	0.943	0.920	0.924	0.920
Q9JZN9	0.814	0.818	0.826	0.830	0.823
Q9K0U9	0.782	0.696	0.700	0.675	0.675
Average	0.694	0.715	0.704	0.693	0.692

Table A.8 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top10 using swissprot)				
Variance	0.023	0.016	0.020	0.018	0.018

Table A.8: NorMD Score for MSAs from Different Programs (for Top25 using OMPdb70)

MSA	NorMD Score for MSAs from Different Programs (for Top25 using swissprot)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A1JUB7	0.339	1.083	0.927	0.607	0.506
A5F934	1.356	1.421	1.201	1.297	1.351
E3PJ86	0.948	1.014	0.981	0.985	0.990
O33407	1.099	1.161	1.086	1.207	1.195
O88093	1.254	1.454	1.268	1.374	1.329
P01031	1.091	1.139	1.108	1.118	1.086
P02748	0.841	0.860	0.846	0.861	0.857
P02787	0.939	0.941	0.927	0.944	0.937
P02930	0.289	0.639	0.436	0.391	0.398
P02931	0.826	0.829	0.826	0.835	0.823
P02932	0.842	0.844	0.845	0.847	0.843
P05430	0.246	0.623	0.564	0.513	0.484
P06970	0.679	0.709	0.716	0.704	0.708
P06971	0.676	0.717	0.723	0.730	0.740
P06996	0.875	0.908	0.865	0.913	0.908
P07110	0.649	0.656	0.658	0.661	0.662
P07357	0.848	0.886	0.818	0.866	0.877
P07358	0.831	0.854	0.873	0.853	0.846
P07360	0.824	0.832	0.792	0.804	0.809
P08189	0.645	0.671	0.683	0.702	0.701
P08190	0.585	0.553	0.584	0.598	0.596
P0A071	1.097	1.161	1.203	1.113	1.116
P0A074	1.097	1.161	1.203	1.113	1.116
P0A077	0.941	0.970	1.134	0.961	0.949
P0A263	0.890	0.897	0.876	0.900	0.891
P0A264	0.890	0.897	0.876	0.900	0.891
P0A910	1.080	1.270	1.079	1.162	1.118
P0A911	1.080	1.270	1.079	1.162	1.118
P0A937	0.486	0.680	0.526	0.683	0.671
P0AEA2	0.184	0.504	0.320	0.333	0.297
P0C2W0	1.308	1.707	1.682	1.293	1.150
P0C6Q6	1.355	1.421	1.193	1.294	1.351
P0DH58	0.549	0.744	0.553	0.573	0.582
P10643	0.828	0.859	0.821	0.845	0.857
P11922	1.462	1.027	1.514	0.398	0.423
P12643	1.045	1.084	1.042	1.078	1.077
P13036	1.098	1.229	1.138	1.034	1.065
P13671	0.839	0.873	0.852	0.856	0.869
P15319	0.785	0.788	0.805	0.826	0.824
P16869	0.696	0.740	0.743	0.731	0.736
P19809	1.693	1.286	1.974	0.514	0.497
P21796	0.777	0.782	0.760	0.804	0.776

Table A.9 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top25 using swissprot)				
P22340	1.536	1.797	1.386	1.583	1.560
P24017	0.797	0.944	0.814	0.868	0.842
P24391	0.472	0.636	0.568	0.521	0.492
P26466	1.000	1.000	1.000	1.000	1.000
P30130	0.674	0.626	0.598	0.493	0.493
P30690	0.636	0.574	0.553	0.537	0.530
P31554	0.899	0.905	0.941	0.903	0.904
P31697	0.725	0.722	0.732	0.757	0.749
P31780	1.103	1.175	1.090	1.152	1.146
P35672	0.622	0.679	0.700	0.678	0.673
P35818	0.869	0.892	0.886	0.882	0.881
P37001	0.851	0.828	0.863	0.909	0.829
P37432	0.835	0.839	0.854	0.842	0.834
P42512	0.808	0.857	0.820	0.861	0.850
P43261	1.710	1.282	1.527	0.457	0.497
P45758	0.945	0.978	0.984	0.969	0.968
P45779	0.989	1.018	1.037	1.012	1.009
P46359	0.778	0.806	0.754	0.801	0.777
P48632	0.761	0.820	0.766	0.795	0.799
P49767	0.887	1.010	1.018	0.941	0.948
P75780	0.721	0.789	0.783	0.775	0.773
P77211	0.559	0.610	0.624	0.621	0.586
P77774	0.497	0.544	0.546	0.526	0.533
Q03155	0.574	0.678	0.563	0.509	0.516
Q04884	1.540	0.850	0.769	0.811	0.865
Q16853	0.878	0.908	0.924	0.903	0.900
Q2FFA2	1.092	1.172	1.138	1.133	1.130
Q2FFA3	0.870	0.932	0.946	0.906	0.911
Q45340	0.548	0.688	0.585	0.637	0.649
Q48473	0.868	0.869	0.857	0.874	0.869
Q51397	0.593	0.609	0.606	0.625	0.621
Q51487	0.585	0.593	0.601	0.623	0.610
Q60932	0.792	0.794	0.760	0.815	0.788
Q7BSW5	1.229	1.345	1.139	1.254	1.254
Q83SQ0	0.901	0.907	0.941	0.905	0.906
Q8CVW1	0.861	0.862	0.865	0.867	0.862
Q8ZIK3	0.900	0.907	0.946	0.904	0.905
Q8ZPC9	1.000	1.000	1.000	1.000	1.000
Q99RL1	1.080	1.147	1.129	1.105	1.104
Q9I5U2	0.726	0.749	0.727	0.743	0.748
Q9K0U9	1.210	1.278	1.170	1.264	1.239
Q9TUM0	0.880	0.889	0.929	0.883	0.882
Average	0.877	0.924	0.899	0.857	0.851
Variance	0.088	0.067	0.076	0.060	0.059

Table A.9: NorMD Score for MSAs from Different Programs (for Top25 using swissprot)

MSA	NorMD Score for MSAs from Different Programs (for Top25 using unirefOMBB100)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
O33407	0.842	0.842	0.822	0.892	0.843
O88093	1.000	1.000	1.000	1.000	1.000
P00747	0.422	0.840	0.811	0.487	0.468
P02929	0.563	0.957	0.670	0.640	0.648
P05430	1.485	1.655	1.327	1.459	1.362
P05825	0.457	0.515	0.286	0.362	0.407
P06129	0.467	0.386	0.331	0.478	0.498
P09169	0.466	0.699	0.543	0.636	0.645
P09616	2.344	4.490	3.994	3.790	3.180
P0A903	0.756	0.795	0.811	0.807	0.792
P0A910	0.695	0.773	0.688	0.727	0.728
P0A911	0.695	0.773	0.688	0.727	0.728
P0A915	0.428	0.457	0.458	0.399	0.399
P0A916	0.428	0.457	0.458	0.399	0.399
P0A917	0.809	0.813	0.797	0.813	0.816
P0A918	0.809	0.813	0.797	0.813	0.816
P0A919	0.809	0.813	0.797	0.813	0.816
P0A920	0.809	0.813	0.797	0.813	0.816
P0A937	1.097	1.145	1.072	1.170	1.186
P0A940	0.792	0.803	0.814	0.798	0.803
P0A941	0.792	0.803	0.814	0.798	0.803
P0A942	0.792	0.803	0.814	0.798	0.803
P0A943	0.792	0.803	0.814	0.798	0.803
P0AC02	0.657	0.659	0.652	0.665	0.666
P0ADE4	0.841	0.914	0.977	0.895	0.878
P0ADE5	0.841	0.914	0.977	0.895	0.878
P10384	0.701	0.738	0.704	0.764	0.719
P13036	0.402	0.795	0.675	0.464	0.467
P13794	0.560	0.643	0.640	0.606	0.599
P17315	0.436	0.509	0.290	0.431	0.431
P17811	0.296	0.525	0.558	0.393	0.436
P24017	0.657	0.869	0.698	0.746	0.753
P24305	0.185	0.383	0.275	0.258	0.292
P30690	0.518	0.596	0.547	0.188	0.264
P35818	1.823	2.793	1.543	2.364	0.451
P46359	0.244	0.549	0.344	0.268	0.257
P69434	0.277	0.430	0.415	0.360	0.332
P69856	0.169	0.272	0.211	0.209	0.208
P69857	0.169	0.272	0.211	0.209	0.208
P69858	0.169	0.272	0.211	0.209	0.208
P76045	0.245	0.438	0.323	0.218	0.230
P77774	0.576	0.583	0.582	0.590	0.588
Q03155	1.399	0.875	0.930	0.920	1.294
Q04884	1.000	1.000	1.000	1.000	1.000
Q05098	0.454	0.504	0.315	0.404	0.449
Q45340	0.949	1.008	0.935	0.979	0.951
Q7BCK4	1.000	1.000	1.000	1.000	1.000
Q7BSW5	1.000	1.000	1.000	1.000	1.000
Q8ZRP0	0.796	0.806	0.806	0.800	0.802
Q9HVD1	0.178	0.418	0.267	0.250	0.238
Q9I5U2	1.082	2.188	1.470	1.253	1.126

Table A.10 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top25 using unirefOMBB100)				
Q9JZN9	0.541	1.001	0.790	0.413	0.450
Average	0.706	0.869	0.764	0.753	0.710
Variance	0.173	0.447	0.304	0.328	0.208

Table A.10: NorMD Score for MSAs from Different Programs (for Top25 using unirefOMBB100)

MSA	NorMD Score for MSAs from Different Programs (for Top50 using OMPdb70)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A1JUB7	0.378	0.539	0.527	0.497	0.504
A5F934	0.807	0.830	0.794	0.808	0.807
E3PJ86	0.720	0.724	0.728	0.725	0.723
E6MXW0	0.813	0.831	0.804	0.822	0.826
O33407	0.667	0.692	0.691	0.693	0.686
O88093	0.811	0.747	0.724	0.703	0.698
P02929	0.181	0.465	0.316	0.255	0.270
P02930	0.662	0.661	0.669	0.665	0.665
P02931	0.726	0.731	0.727	0.733	0.726
P02932	0.706	0.716	0.698	0.708	0.710
P02943	0.718	0.735	0.707	0.718	0.710
P05430	0.807	0.816	0.820	0.822	0.821
P05695	0.689	0.713	0.737	0.721	0.717
P05825	0.760	0.765	0.764	0.763	0.762
P06129	0.724	0.731	0.741	0.733	0.740
P06970	0.689	0.693	0.702	0.697	0.696
P06971	0.611	0.615	0.615	0.616	0.613
P06996	0.710	0.740	0.715	0.723	0.720
P07110	0.710	0.713	0.712	0.716	0.711
P09169	0.554	0.565	0.560	0.564	0.563
P0A232	0.712	0.717	0.722	0.712	0.709
P0A263	0.713	0.731	0.718	0.722	0.717
P0A264	0.713	0.731	0.718	0.722	0.717
P0A910	0.743	0.735	0.752	0.744	0.748
P0A911	0.743	0.735	0.752	0.744	0.748
P0A915	0.630	0.637	0.633	0.614	0.611
P0A916	0.630	0.637	0.633	0.614	0.611
P0A917	0.670	0.671	0.662	0.681	0.682
P0A918	0.670	0.671	0.662	0.681	0.682
P0A919	0.670	0.671	0.662	0.681	0.682
P0A920	0.670	0.671	0.662	0.681	0.682
P0A921	0.713	0.721	0.707	0.717	0.714
P0A922	0.713	0.721	0.707	0.717	0.714
P0A923	0.713	0.721	0.707	0.717	0.714
P0A927	0.712	0.731	0.728	0.726	0.721
P0A928	0.712	0.731	0.728	0.726	0.721
P0A929	0.712	0.731	0.728	0.726	0.721
P0A940	0.640	0.644	0.641	0.641	0.641
P0A941	0.640	0.644	0.641	0.641	0.641
P0A942	0.640	0.644	0.641	0.641	0.641

Table A.11 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top50 using OMPdb70)				
P0A943	0.640	0.644	0.641	0.641	0.641
P0ADE4	0.650	0.645	0.647	0.649	0.642
P0ADE5	0.650	0.645	0.647	0.649	0.642
P0AEA2	0.927	1.158	0.978	1.102	0.972
P0C2W0	0.349	0.520	0.543	0.497	0.516
P0C6Q6	0.807	0.823	0.811	0.808	0.807
P0DH58	0.801	0.818	0.803	0.814	0.816
P10384	0.747	0.731	0.749	0.739	0.751
P11922	0.744	0.669	0.705	0.684	0.676
P13036	0.758	0.751	0.762	0.764	0.763
P13794	0.602	0.614	0.611	0.621	0.602
P16869	0.661	0.665	0.665	0.668	0.664
P17315	0.755	0.759	0.759	0.765	0.760
P17811	0.583	0.594	0.590	0.596	0.594
P18195	0.810	0.826	0.808	0.824	0.826
P18895	0.559	0.577	0.605	0.606	0.613
P19809	0.627	0.633	0.654	0.626	0.635
P22340	0.561	0.601	0.589	0.582	0.575
P24017	0.744	0.732	0.747	0.749	0.752
P24305	0.846	0.845	0.848	0.851	0.848
P26466	0.719	0.735	0.711	0.723	0.718
P30130	0.696	0.698	0.685	0.699	0.697
P30690	0.794	0.800	0.803	0.805	0.804
P31243	0.728	0.741	0.739	0.727	0.721
P31554	0.646	0.655	0.651	0.649	0.648
P31780	0.749	0.752	0.753	0.753	0.750
P32722	0.592	0.604	0.601	0.603	0.601
P32977	0.633	0.675	0.675	0.649	0.640
P35077	0.540	0.540	0.545	0.556	0.556
P35672	0.754	0.753	0.756	0.761	0.758
P35818	0.777	0.776	0.780	0.784	0.776
P37001	0.659	0.660	0.670	0.675	0.675
P37432	0.764	0.758	0.741	0.758	0.757
P39767	0.691	0.699	0.668	0.714	0.711
P42512	0.585	0.592	0.569	0.596	0.589
P43261	0.621	0.624	0.643	0.614	0.621
P45758	0.739	0.743	0.742	0.744	0.740
P45779	0.731	0.734	0.740	0.735	0.733
P46359	0.651	0.662	0.662	0.667	0.664
P48632	0.666	0.674	0.672	0.671	0.665
P69856	0.478	0.514	0.505	0.523	0.521
P69857	0.478	0.514	0.505	0.523	0.521
P69858	0.478	0.514	0.505	0.523	0.521
P75780	0.661	0.664	0.663	0.664	0.659
P76045	0.145	0.270	0.225	0.271	0.266
P77211	0.674	0.679	0.680	0.681	0.680
Q03155	0.641	0.675	0.663	0.636	0.632
Q04884	0.718	0.766	0.729	0.719	0.732
Q05098	0.753	0.756	0.753	0.757	0.756
Q45340	0.577	0.620	0.602	0.615	0.616
Q48473	0.711	0.721	0.710	0.716	0.710

Table A.11 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top50 using OMPdb70)				
Q51397	0.642	0.645	0.643	0.648	0.648
Q51487	0.671	0.679	0.677	0.681	0.680
Q7BCK4	0.644	0.663	0.661	0.644	0.641
Q7BSW5	0.815	0.719	0.613	0.674	0.691
Q83SQ0	0.632	0.645	0.641	0.636	0.635
Q8CVI4	0.718	0.733	0.711	0.718	0.711
Q8CVW1	0.735	0.746	0.729	0.738	0.741
Q8ZIK3	0.637	0.644	0.642	0.641	0.639
Q8ZRP0	0.640	0.644	0.640	0.641	0.640
Q8ZRW0	0.647	0.657	0.651	0.651	0.650
Q934G3	0.604	0.616	0.611	0.635	0.630
Q9HVD1	0.716	0.731	0.723	0.733	0.726
Q9I5U2	0.717	0.716	0.726	0.726	0.728
Q9JZN9	0.807	0.805	0.814	0.821	0.817
Q9K0U9	1.011	0.847	0.803	0.788	0.779
Average	0.676	0.691	0.683	0.685	0.683
Variance	0.014	0.010	0.009	0.010	0.009

Table A.11: NorMD Score for MSAs from Different Programs (for Top50 using OMPdb70)

MSA	NorMD Score for MSAs from Different Programs (for Top50 using swissprot)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
A5F934	0.864	0.955	0.597	0.797	0.709
E3PJ86	0.514	0.650	0.583	0.594	0.577
P01031	1.146	1.174	1.133	1.152	1.148
P02787	1.116	1.133	1.096	1.171	1.161
P02931	0.363	0.498	0.377	0.406	0.375
P02932	0.263	0.516	0.383	0.340	0.356
P06971	0.637	0.693	0.680	0.689	0.686
P07360	0.440	0.501	0.478	0.435	0.396
P08189	0.471	0.581	0.585	0.629	0.610
P0A910	0.736	0.768	0.748	0.819	0.812
P0A911	0.736	0.768	0.748	0.819	0.812
P10643	0.957	1.370	1.073	0.224	0.235
P12643	0.753	0.806	0.796	0.769	0.754
P16869	0.683	0.721	0.732	0.778	0.766
P24017	0.745	0.772	0.750	0.829	0.820
P26466	0.873	1.077	0.906	0.917	0.904
P31554	0.737	0.752	0.754	0.749	0.749
P37001	0.775	0.766	0.750	0.792	0.785
P45758	0.707	0.731	0.709	0.731	0.728
P45779	0.528	0.679	0.563	0.591	0.584
P48632	0.673	0.683	0.709	0.703	0.701
P49767	0.772	1.127	0.887	0.876	0.913
P75780	0.763	0.878	0.849	0.867	0.840
P77774	0.416	0.677	0.641	0.595	0.618
Q7BSW5	0.313	0.622	0.504	0.453	0.453
Q83SQ0	0.735	0.750	0.754	0.747	0.747

Table A.12 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top50 using swissprot)				
Q8ZIK3	0.756	0.766	0.758	0.762	0.762
Q8ZPC9	0.726	0.730	0.744	0.737	0.732
Q8ZRW0	0.747	0.756	0.771	0.752	0.752
Q9I5U2	0.626	0.636	0.631	0.656	0.656
Q9K0U9	1.002	1.064	1.004	1.048	1.043
Q9TUM0	1.094	1.114	1.229	1.076	1.067
Average	0.708	0.804	0.748	0.734	0.727
Variance	0.048	0.046	0.041	0.047	0.047

Table A.12: NorMD Score for MSAs from Different Programs (for Top50 using swissprot)

MSA	NorMD Score for MSAs from Different Programs (for Top50 using unirefOMBB100)				
ID	**ClustalO**	**MAFFT**	**Muscle**	**T-Coffee**	**TMB-Coffee**
O33407	0.844	0.841	0.825	0.842	0.845
O88093	1.000	1.000	1.000	1.000	1.000
P02929	0.394	0.642	0.537	0.488	0.469
P05430	0.963	1.189	1.091	0.848	0.815
P05825	0.376	0.311	0.303	0.354	0.376
P06129	0.344	0.312	0.379	0.354	0.358
P09169	0.451	0.773	0.635	0.404	0.540
P0A903	0.716	0.742	0.727	0.755	0.763
P0A910	0.548	0.686	0.596	0.595	0.600
P0A911	0.548	0.686	0.596	0.595	0.600
P0A915	0.150	0.307	0.265	0.308	0.279
P0A916	0.150	0.307	0.265	0.308	0.279
P0A917	0.732	0.742	0.733	0.743	0.742
P0A918	0.732	0.742	0.733	0.743	0.742
P0A919	0.732	0.742	0.733	0.743	0.742
P0A920	0.732	0.742	0.733	0.743	0.742
P0A937	0.847	0.890	0.906	0.891	0.974
P0A940	0.588	0.630	0.504	0.623	0.625
P0A941	0.588	0.630	0.504	0.623	0.625
P0A942	0.588	0.630	0.504	0.623	0.625
P0A943	0.588	0.630	0.504	0.623	0.625
P0AC02	0.627	0.634	0.626	0.655	0.649
P0ADE4	0.733	0.782	0.771	0.736	0.746
P0ADE5	0.733	0.782	0.771	0.736	0.746
P10384	0.251	0.439	0.332	0.322	0.299
P13036	0.226	0.498	0.387	0.319	0.310
P13794	0.470	0.607	0.588	0.527	0.568
P17315	0.404	0.428	0.335	0.418	0.422
P17811	0.257	0.336	0.322	0.313	0.283
P24017	0.605	0.730	0.627	0.663	0.649
P30690	0.519	0.638	0.308	0.156	0.168
P35818	2.610	3.040	1.368	2.595	0.457

Table A.13 continued from previous page

MSA	NorMD Score for MSAs from Different Programs (for Top50 using unirefOMBB100)				
P69856	0.137	0.260	0.197	0.193	0.204
P69857	0.137	0.260	0.197	0.193	0.204
P69858	0.137	0.260	0.197	0.193	0.204
P76045	0.205	0.336	0.260	0.212	0.230
P77774	0.469	0.563	0.586	0.457	0.496
Q03155	0.787	0.790	0.711	0.597	0.585
Q04884	1.000	1.000	1.000	1.000	1.000
Q05098	0.404	0.448	0.310	0.399	0.410
Q45340	0.682	0.762	0.730	0.704	0.674
Q7BCK4	1.000	1.000	1.000	1.000	1.000
Q7BSW5	0.862	0.871	0.863	0.868	0.864
Q8ZRP0	0.628	0.657	0.520	0.666	0.653
Q9HVD1	0.127	0.315	0.230	0.195	0.197
Q9I5U2	1.069	2.199	1.230	1.149	1.035
Average	0.602	0.713	0.599	0.619	0.574
Variance	0.163	0.228	0.082	0.153	0.062

Table A.13: NorMD Score for MSAs from Different Programs (for Top50 using unirefOMBB100)

A.5 Paired t-test

Paired t-test	OMPdb70 (10 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**6.36**	1.17	0.97	-0.47
MAFFT		NA	**-5.90**	**-6.30**	**-6.39**
MUSCLE			NA	-0.43	**-2.61**
T-Coffee				NA	**-2.73**
TMB-Coffee					NA

Table A.14: Paired t-test (OMPdb70 - 10 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	unirefOMBB100 (10 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**7.05**	**3.64**	1.64	0.36
MAFFT		NA	**-3.58**	**-6.85**	**-6.84**
MUSCLE			NA	**-4.16**	**-4.32**
T-Coffee				NA	**-2.83**
TMB-Coffee					NA

Table A.15 continued from previous page

Paired t-test	unirefOMBB100 (10 Hits)

Table A.15: Paired t-test (unirefOMBB100 - 10 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	swissprot (25 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**2.57**	1.5	-0.73	-1.02
MAFFT		NA	-1.74	**-3.87**	-4.15
MUSCLE			NA	-1.58	-1.82
T-Coffee				NA	**-2.27**
TMB-Coffee					NA

Table A.16: Paired t-test (Swissprot - 25 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	OMPdb70 (25 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**3.38**	1.33	-0.17	-0.31
MAFFT		NA	-1.61	**-4.08**	**-4.61**
MUSCLE			NA	-1.47	**-1.69**
T-Coffee				NA	-0.79
TMB-Coffee					NA

Table A.17: Paired t-test (OMPdb70 - 25 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	unirefOMBB100 (25 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**3.20**	1.59	1.46	0.13
MAFFT		NA	**-3.53**	**-4.20**	**-2.84**
MUSCLE			NA	-0.49	-1.77
T-Coffee				NA	-1.09
TMB-Coffee					NA

Table A.18: Paired t-test (unirefOMBB100 - 25 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	swissprot (50 Hits)				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**4.81**	**2.71**	1.00	0.69
MAFFT		NA	**-3.26**	-1.85	-2.09

Table A.19 continued from previous page

Paired t-test	**swissprot (50 Hits)**			
MUSCLE		NA	-0.45	-0.76
T-Coffee			NA	**-2.18**
TMB-Coffee				NA

Table A.19: Paired t-test (Swissprot - 50 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	**OMPdb70 (50 Hits)**				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**3.01**	1.44	**2.19**	1.61
MAFFT		NA	**-3.07**	**-2.2**	**-2.81**
MUSCLE			NA	1.47	0.16
T-Coffee				NA	-1.97
TMB-Coffee					NA

Table A.20: Paired t-test (OMPdb70 - 50Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Paired t-test	**unirefOMBB100 (50 Hits)**				
	ClustalO	**MAFFT**	**MUSCLE**	**T-Coffee**	**TMB-Coffee**
ClustalO	NA	**4.14**	-0.11	1.44	-0.57
MAFFT		NA	-0.57	**-3.41**	**-2.27**
MUSCLE			NA	0.69	-1.06
T-Coffee				NA	-0.95
TMB-Coffee					NA

Table A.21: Paired t-test (unirefOMBB100 - 50 Hits). NULL hypothesis: x and y have identical performance. Highlighted values are significant.

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[2] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[3] A. Bahr, J. D. Thompson, J.-C. Thierry, and O. Poch, "BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations," *Nucleic Acids Research*, vol. 29, no. 1, pp. 323–326, 2001.

[4] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane *et al.*, "The universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D154–D159, 2005.

[5] G. J. Barton, "Sequence alignment for molecular replacement," *Acta Crystallographica Section D: Biological Crystallography*, vol. 64, no. 1, pp. 25–32, 2008.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[7] T. J. Beveridge, "Structures of gram-negative cell walls and their derived membrane vesicles," *Journal of Bacteriology*, vol. 181, no. 16, pp. 4725–4733, 1999.

[8] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins, "Sequence embedding for fast construction of guide trees for multiple sequence alignment," *Algorithms for Molecular Biology*, vol. 5, no. 1, p. 21, 2010.

[9] H. Carrillo and D. Lipman, "The multiple sequence alignment problem in biology," *SIAM Journal on Applied Mathematics*, vol. 48, no. 5, pp. 1073–1082, 1988.

[10] J.-M. Chang, P. Di Tommaso, J.-F. Taly, and C. Notredame, "Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee," *BMC Bioinformatics*, vol. 13, no. 4, p. S1, 2012.

[11] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, "Multiple sequence alignment with the Clustal series of programs," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3497–3500, 2003.

[12] M. Dayhoff, R. Schwartz, and B. Orcott, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352, 1978.

[13] C. B. Do, M. S. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: Probabilistic consistency-based multiple sequence alignment," *Genome Research*, vol. 15, no. 2, pp. 330–340, 2005.

[14] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[15] R. V. Eck, "Non-randomness in amino-acid 'alleles'," *Nature*, vol. 191, no. 4795, p. 1284, 1961.

[16] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[17] J. W. Fairman, N. Noinaj, and S. K. Buchanan, "The structural biology of β-barrel membrane proteins: a summary of recent reports," *Current Opinion in Structural Biology*, vol. 21, no. 4, pp. 523–531, 2011.

[18] D.-F. Feng and R. F. Doolittle, "Progressive alignment and phylogenetic tree construction of protein sequences," *Methods in Enzymology*, vol. 183, pp. 375–387, 1990.

[19] E. W. Floden, P. D. Tommaso, M. Chatzou, C. Magis, C. Notredame, and J.-M. Chang, "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 Research*, vol. 44, no. W1, pp. W339–W343, 2016.

[20] S. Galdiero, M. Galdiero, and C. Pedone, "β-barrel membrane bacterial proteins: structure, function, assembly and interaction with lipids," *Current Protein and Peptide Science*, vol. 8, no. 1, pp. 63–82, 2007.

[21] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database," *Science*, vol. 256, no. 5062, pp. 1443–1445, 1992.

[22] O. Gotoh, "Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments," *Journal of Molecular Biology*, vol. 264, no. 4, pp. 823–838, 1996.

[23] O. Gotoh, "A weighting system and algorithm for aligning many phylogenetically related sequences," *Bioinformatics*, vol. 11, no. 5, pp. 543–551, 1995.

[24] D. Gusfield, "Efficient methods for multiple sequence alignment with guaranteed error bounds," *Bulletin of Mathematical Biology*, vol. 55, no. 1, pp. 141–154, 1993.

[25] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[26] S. Henikoff and J. G. Henikoff, "Performance evaluation of amino acid substitution matrices," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 49–61, 1993.

[27] B. K. Jap and P. J. Walian, "Structure and functional mechanism of porins," *Physiological Reviews*, vol. 76, no. 4, pp. 1073–1088, 1996.

[28] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[29] M. Kimura, *The neutral theory of molecular evolution.* Cambridge University Press, 1983.

[30] J. H. Kleinschmidt, "Folding of β-barrel membrane proteins in lipid bilayers-Unassisted and assisted folding and insertion," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1848, no. 9, pp. 1927–1943, 2015.

[31] R. Koebnik, K. P. Locher, and P. Van Gelder, "Structure and function of bacterial outer membrane proteins: barrels in a nutshell," *Molecular Microbiology*, vol. 37, no. 2, pp. 239–253, 2000.

[32] T. Lassmann and E. L. Sonnhammer, "Automatic assessment of alignment quality," *Nucleic Acids Research*, vol. 33, no. 22, pp. 7120–7128, 2005.

[33] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, no. 4693, pp. 1435–1441, 1985.

[34] K. L. Loening, "IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) Nomenclature and Symbolism for Amino Acids and Peptides," *European Journal of Biochemistry*, vol. 138, no. 1, pp. 9–37, 1984.

[35] M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, and A. L. Lomize, "OPM database and PPM web server: resources for positioning of proteins in membranes," *Nucleic Acids Research*, vol. 40, no. D1, pp. D370–D376, 2011.

[36] J. E. Merritt and K. L. Loening, "IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) Nomenclature of Tetrapyrroles Recommendations 1978," *European Journal of Biochemistry*, vol. 108, no. 1, pp. 1–30, 1980.

[37] B. Morgenstern, A. Dress, and T. Werner, "Multiple DNA and protein sequence alignment based on segment-to-segment comparison," *Proceedings of the National Academy of Sciences*, vol. 93, no. 22, pp. 12 098–12 103, 1996.

[38] D. W. Mount, "Comparison of the PAM and BLOSUM amino acid substitution matrices," *Cold Spring Harbor Protocols*, vol. 3, no. 6, 2008.

[39] N. R. Movva, K. Nakamura, and M. Inouye, "Gene structure of the OmpA protein, a major surface protein of Escherichia coli required for cell-cell interaction," *Journal of Molecular Biology*, vol. 143, no. 3, pp. 317–328, 1980.

[40] J. Muller, C. J. Creevey, J. D. Thompson, D. Arendt, and P. Bork, "AQUA: Automated quality improvement for multiple sequence alignments," *Bioinformatics*, vol. 26, no. 2, pp. 263–265, 2009.

[41] Ö. U. Nalbantoğlu, "Dynamic programming," in *Multiple Sequence Alignment Methods*. Springer, 2014, pp. 3–27.

[42] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[43] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

[44] C. Notredame, L. Holm, and D. G. Higgins, "Coffee: an objective function for multiple sequence alignments." *Bioinformatics (Oxford, England)*, vol. 14, no. 5, pp. 407–422, 1998.

[45] F. S.-M. Pais, P. de Cássia Ruy, G. Oliveira, and R. S. Coimbra, "Assessing the efficiency of multiple sequence alignment programs," *Algorithms for Molecular Biology*, vol. 9, no. 1, p. 4, 2014.

[46] J. Pei and N. V. Grishin, "AL2CO: calculation of positional conservation in a protein sequence alignment," *Bioinformatics*, vol. 17, no. 8, pp. 700–712, 2001.

[47] W. Pirovano, K. A. Feenstra, and J. Heringa, "Praline^{TM}: A strategy for improved multiple alignment of transmembrane proteins," *Bioinformatics*, vol. 24, no. 4, pp. 492–497, 2008.

[48] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[49] G. E. Schulz, "Transmembrane β-barrel proteins," *Advances In Protein Chemistry*, vol. 63, pp. 47–70, 2003.

[50] Y. Shafrir and H. R. Guy, "STAM: Simple transmembrane alignment method," *Bioinformatics*, vol. 20, no. 5, pp. 758–769, 2004.

[51] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Science*, vol. 27, no. 1, pp. 135–145, 2018.

[52] V. Simossis, J. Kleinjung, and J. Heringa, "Homology-extended sequence alignment," *Nucleic Acids Research*, vol. 33, no. 3, pp. 816–824, 2005.

[53] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[54] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2004.

[55] R. R. Sokal, "A statistical method for evaluating systematic relationship," *University of Kansas Science Bulletin*, vol. 28, pp. 1409–1438, 1958.

[56] L. K. Tamm, H. Hong, and B. Liang, "Folding and assembly of β-barrel membrane proteins," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1666, no. 1-2, pp. 250–263, 2004.

[57] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, no. 24, pp. 4876–4882, 1997.

[58] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[59] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 127–136, 2005.

[60] J. D. Thompson, F. Plewniak, R. Ripp, J.-C. Thierry, and O. Poch, "Towards a reliable objective function for multiple sequence alignments," *Journal of Molecular Biology*, vol. 314, no. 4, pp. 937–951, 2001.

[61] J. D. Thompson, J.-C. Thierry, and O. Poch, "Rascal: Rapid scanning and correction of multiple sequence alignments," *Bioinformatics*, vol. 19, no. 9, pp. 1155–1161, 2003.

[62] W. Tian, M. Lin, K. Tang, J. Liang, and H. Naveed, "High-resolution structure prediction of β-barrel membrane proteins," *Proceedings of the National Academy of Sciences*, vol. 115, no. 7, pp. 1511–1516, 2018.

[63] K. D. Tsirigos, P. G. Bagos, and S. J. Hamodrakas, "OMPdb: a database of β-barrel outer membrane proteins from gram-negative bacteria," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D324–D331, 2010.

[64] G. E. Tusnady, Z. Dosztanyi, and I. Simon, "PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank," *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D275–D278, 2005.

[65] G. E. Tusnady, L. Kalmar, and I. Simon, "TOPDB: topology data bank of transmembrane proteins," *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D234–D239, 2007.

[66] G. E. Tusnady and I. Simon, "Principles governing amino acid composition of integral membrane proteins: application to topology prediction," *Journal of Molecular Biology*, vol. 283, no. 2, pp. 489–506, 1998.

[67] G. E. Tusnady and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001.

[68] M. Vingron and P. R. Sibbald, "Weighting in sequence space: a comparison of methods in terms of generalized sequences," *Proceedings of the National Academy of Sciences*, vol. 90, no. 19, pp. 8777–8781, 1993.

[69] W. C. Wimley, "Toward genomic identification of β-barrel membrane proteins: Composition and architecture of known structures," *Protein Science*, vol. 11, no. 2, pp. 301–312, 2002.

[70] Ö. Yildiz, K. R. Vinothkumar, P. Goswami, and W. Kühlbrandt, "Structure of the monomeric outer-membrane porin OmpG in the open and closed conformation," *The EMBO Journal*, vol. 25, no. 15, pp. 3702–3713, 2006.

[71] H. Zhang, "Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm," *Bioinformatics*, vol. 19, no. 11, pp. 1391–1396, 2003.