

Comparison of Membrane Proteins Using Computational Programs

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Biochemie, Chemie und Pharmazie
der Johann Wolfgang Goethe-Universität
Frankfurt am Main

von
Marcus Stamm
aus Frankfurt am Main

Frankfurt am Main, 2015

This work was carried out in the computational structural biology group at the Max Planck Institute of Biophysics in Frankfurt am Main and accepted as a dissertation in the department of Biochemistry, Chemistry and Pharmacy at the Johann Wolfgang Goethe University of Frankfurt.

Dean: Prof. Dr. Michael Karas

First Advisor: Prof. Dr. Volker Dötsch

Second Advisor: Prof. Dr. Hartmut Michel

Date of disputation:

Acknowledgements

I am very grateful to everybody who made it possible to create this dissertation and supported me in researching and writing.

In particular, I would like to thank Lucy Forrest very much for giving me the opportunity to carry out a thesis with an interesting topic. In particular, I thank her very much for all her scientific advice and help as well as her motivational support. I really appreciated the work with the Computational Structural Biology group. I also want to thank Volker Dötsch and Hartmut Michel for being my supervisors and part of my thesis committee.

I would like to thank all members of the Computational Structural Biology Group for their help during my PhD and the very nice and friendly atmosphere in the group during and aside from work: Cristina Fenollar-Ferrer, Ahmadreza Mehdipour, Thomas Crisman, Caroline Koshy and Desirée Kaufmann. Special thanks go to Dr. Rene Staritzbichler and Dr. Kamil Khafizov for their great ideas and inspirations in many exhaustive and interesting discussions about AlignMe. A special thanks goes also to Sebastian Radestock for proofreading my summaries.

Additionally, I would also like to thank all members of the Theoretical Molecular Biophysics Group for all their support: José Faraldo-Gómez, Wenchang Zhou, Claudio Anselmi, Alexander Krah, Diana Garzon, Fabrizio Marinelli, Vanesa Leone and Davide Branduardi.

Finally, I am also deeply grateful to Alexander Schüler for all his abundance of patience, support and encouragement during my PhD time and also for proofreading my thesis.

List of publications

Khafizov, K., *et al.* (2010) A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe, *Biochemistry*, **49**, 10702-10713.

Stamm, M., *et al.* (2013) Alignment of Helical Membrane Protein Sequences Using AlignMe, *PloS one*, **8**, e57731.

Stamm, M., *et al.* (2014) AlignMe - a membrane protein sequence alignment web server, *Nucleic acids research*, **42**, W246-W251

Stamm, M. and Forrest, L. (2015) Structure alignment of membrane proteins: Accuracy of available tools and a consensus strategy, *Proteins*, **83**, 1720-1732.

Table of Contents

Abbreviations	I
List of Tables	II
List of Figures	III
Abstract	VI
Zusammenfassung	XI
1 Introduction	1
1.1 Globular & Membrane Proteins	1
1.2 Protein Properties & Structures of Membrane Proteins	4
1.2.1 General	4
1.2.2 α -helical Structures	5
1.2.3 β -strands, -sheets & -barrels	7
1.2.4 Less Frequent Secondary Structure Types	9
1.2.5 Tertiary Structures	11
1.2.6 Quaternary Structures	12
1.3 Sequence & Structural Homology Between Proteins	12
1.3.1 Definition of Homology	12
1.3.2 Sequence-based Comparison of Proteins in General	13
1.3.3 Sequence-based Comparison of Membrane Proteins	15
1.3.4 Sequence to Structure Modeling	15
1.3.5 Structure-based Comparisons of Proteins	17
1.3.6 Rating Structural Accuracy and Similarity of Protein Models	18
1.4 Databases for Proteins and Membrane Proteins	19
1.5 Outline of this Work (Research Question, Problem Statement)	21

2	Automated Generation of Homologous Membrane Protein Data Sets (HOME2 & HOME3)	23
2.1	Introduction	23
2.2	PDB_TM Database as a Starting Set for Clustering Membrane Proteins	26
2.3	Common Principles of Clustering Membrane Proteins to a Set of Homologous Proteins for HOME2 and HOME3	28
2.4	Initial Generation of HOME2	28
2.4.1	Modifications of HOME2	31
2.5	HOME3	31
3	AlignMe – an Optimized Program for Aligning Helical Membrane Protein Sequences	34
3.1	Introduction	34
3.2	Methods	38
3.2.1	General Description of Similarity Between Homologous Proteins	38
3.2.2	Inputs Tested that Define Similarity Between a Pair of Proteins	39
3.2.3	Alignment Difference Score (AD score) as an Alignment Accuracy Measure	43
3.2.4	Optimization of Gap Penalty Sets	45
3.2.5	Parameters of Other Alignment Methods Tested	47
3.2.6	Evaluations Based on Homology Modeling	49
3.2.7	BALiBASE Reference 7 Test Set	50
3.3	Results	52
3.3.1	Single Input Descriptors	52
3.3.2	Combined Input Descriptors	57
3.3.3	Comparison of AlignMe with Other Sequence Alignment Methods on HOME2	61
3.3.4	Evaluation of Alignment Accuracy Based on Homology Modeling	66
3.3.5	Comparison of AlignMe with Other Methods on the BALiBASE Reference 7 Test Set	69
3.4	Discussion	73

4	A Web Server for Aligning Membrane Protein Sequences	79
4.1	Introduction	79
4.2	The AlignMe PW Sequence-to-Sequence Alignment Mode	80
4.2.1	Standard Parameter Sets	80
4.2.2	Available Input Descriptors	83
4.2.3	Outputs of the AlignMe Web Server	84
4.3	Other Online Servers for the Alignment of Membrane Protein Sequences	85
4.4	Alignment Accuracy of the AlignMe PW Mode Compared to Other Web Servers	86
4.5	Alignment of Family-Averaged Hydropathy Profiles (HP) Using two Multiple Sequence Alignments	88
4.5.1	Inputs for the HP Mode	89
4.5.2	Standard Parameter Sets	89
4.5.3	Outputs	92
4.5.4	Example Applications of HP alignments	93
4.6	Conclusions	95
5	Evaluation of Structural Alignment Methods on HOMEP3	96
5.1	Introduction	96
5.2	Methods	97
5.2.1	Overview of Structural Alignment Methods Tested	97
5.2.2	Evaluation by Modeling Using Structural Similarity Scores	99
5.2.3	Strategies for Ranking the Accuracy of Structural Alignment Methods	103
5.2.4	Model Selection	105
5.2.5	Consistency of Alignments Between Homologous Protein Sequences	105
5.2.6	Statistical Analysis	106
5.3	Results	106
5.3.1	Selection of a Representative Model	107
5.3.2	Correlation of Structure Similarity Scores	107

5.3.3	Structural Alignment Methods with Length-Independent Scoring Schemes Generate Better Alignments	109
5.3.4	Rigid Superimpositions Compared to Fragment-Based Superimpositions	111
5.3.5	Conformational Flexibility of Membrane Proteins	112
5.3.6	Alignment Coverage	116
5.3.7	Self-Consistency of Alignments for Homologous Proteins	117
5.3.8	A Consensus Approach to Obtain Confidence Values for Aligned Positions	120
5.4	Discussion	122
6	Single Insertions and Deletions (InDels) within Membrane Segments and Their Influence on the Function of Homologous Membrane Proteins	125
6.1	Introduction	125
6.2	Methods	128
6.2.1	Computational Methods for Detecting Secondary Structure Elements	128
6.2.2	Computational Methods to Detect π -helical Structure Elements	129
6.2.3	Consensus Structural Alignments for the Reliable Identification of (Single) InDels in HOME3	131
6.3	Results	133
6.3.1	Occurrence of Secondary Structure Elements in HOME3	133
6.3.2	Single InDels in Confident Positions of Consensus Alignments	136
6.3.3	Single InDels in TM2 and TM5 of G-Protein Coupled Receptors	137
6.3.4	Single InDels in the Proton Pathways of Cytochrome C Oxidase Subunit I	141
6.3.5	A Single InDel in TM1 of LeuT Compared to AdiC	144
6.3.6	A Single InDel in TM5 of ECF Transporters	145
6.4	Discussion	147
7	Conclusion & Future Work	149
7.1	Improved Accuracy of Computational Methods for Membrane Proteins can be Achieved by Including Membrane Specific Information	149
7.2	Using Anchors on Known Conserved Residues in Pairwise Alignments of AlignMe could Improve the Alignment Accuracy	150

7.3	Novel Membrane Protein Descriptors are Available and could be Tested Using AlignMe	151
7.4	Gap Penalties could be Optimized for AlignMe on a Set of Consensus Alignments with Confidence Values	152
7.5	Several Structural Similarity Scores could be Tested for Their Ability to Align Membrane Protein Structures	152
7.6	π -helices should be Given More Importance in Computational Methods	153
7.7	Membrane Protein Reference Data Sets Require Constant Updates and Reliable Structural Alignment Methods	154
7.8	The AlignMe Web Server Needs to be Updated with the Latest Developments	155
A.	Appendix	156

Abbreviations

Short term	Explanation
Å	Angstrom
AD score	Alignment Difference Score
CASP	Critical Assessment of Techniques for Protein Structure Prediction
DSSP	Define Secondary Structure of Proteins program
ECF	Energy Coupling Factors
FIRL	Five-Helix Inverted-Repeat LeuT-like Fold
GDT_TS	Global Distance Test (Total Score)
GPCR	G-protein Coupled Receptor
HOMEPI	Data set of Homologous Membrane Proteins
MSA	Multiple Sequence Alignment
OPM	Orientations of Proteins in Membranes Database
PDB	Protein Data Bank
PDB_TM	Protein Data Bank of Transmembrane Proteins
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSD score	Protein Structural Distance score
PSSM	Position Specific Substitution Matrix
RMSD	Root Mean Square Deviation
SST	Protein Secondary Structure Assignment Program
TM	Transmembrane
TM-score	Template Modeling Score

Amino acid	One-letter code	Three-letter code
Alanine	A	Ala
Arginine	R	Arg
Aspartate	D	Asp
Cysteine	C	Cys
Glutamate	E	Glu
Glutamine	Q	Gln
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val

List of Tables

Table 2.1 Overview of membrane protein databases	25
Table 3.1 Total AD scores of 20 optimization runs for AlignMePST mode	47
Table 3.2 Percentage of positions in ‘core’ regions predicted as TM segments by MEMSAT-SVM	51
Table 3.3 Accuracy of alignments generated using different methods on the HOMEP2 data set	62
Table 3.4 Percentage of transmembrane segments in the HOMEP2 set that are correctly aligned by each method	65
Table 3.5 Accuracy of homology models constructed based on HOMEP2 data set alignments	67
Table 3.6 Percentage of residues that are correctly aligned in pairwise sequence alignments from the BALiBASE reference set 7, sorted by sequence identity of the protein families	70
Table 3.7 Percentage of residues that are correctly aligned in pairwise sequence alignments assigned to the same subgroup within the BALiBASE reference set 7, sorted by sequence identity of the alignments in each protein family	71
Table 3.8 Percentage of residues that are correctly aligned in pairwise sequence alignments assigned to different subgroups within the BALiBASE reference set 7, sorted by sequence identity of the alignments in each protein family	72
Table 3.9 Percentage of residues that are correctly aligned in the predicted transmembrane regions of pairwise sequence alignments from the BALiBASE reference set 7, sorted by protein family name	75
Table 3.10 Average shift error in pairwise alignments of the BALiBASE reference set 7	77
Table 3.11 Average shift error in pairwise alignments assigned to the same subgroup within the BALiBASE reference set 7	77
Table 3.12 Average shift error in pairwise alignments assigned to different subgroups within the BALiBASE reference set 7	78
Table 4.1 Overview of commonly used webservers for analyzing (membrane) protein sequences	80
Table 4.2 Overview of input parameters available as ‘user-defined parameters’ on the AlignMe web server	83
Table 4.3 Overview of recent webservers for aligning (membrane) protein sequences	85
Table 4.4 Percentage of residues aligned correctly in pairwise sequence alignments from the BALiBASE reference set 7, sorted by sequence identity of the protein families	86
Table 4.5. Average shift error in pairwise alignments of the BALiBASE reference set 7	87
Table 5.1 Overview of pairwise structural alignment methods	98
Table 5.2 Explaining model quality by the agreement of different ranking schemes	104
Table 5.3 Correlation between model quality assessment scores	108

Table 5.4 Ranking of structural alignment methods for the subset of α -helical membrane proteins	110
Table 5.5 Ranking of structural alignment methods for the subset of β -barrel membrane proteins	111
Table 5.6 Comparison of models in the MFS transporter family based on alignments generated using FATCAT in flexible and rigid-body fitting modes for structures in similar or different conformational states.	114
Table 5.7 Comparison of models in the MFS transporter family based on alignments generated using FATCAT in flexible and rigid-body fitting modes for structures in similar or different conformational states considering membrane-spanning segments only	115
Table 5.8 Alignment coverage in the alignments generated using different structure alignment programs	117
Table 5.9 Self-consistency of non-gapped positions in the alignments generated using different structure alignment programs.	119
Table A.1 Proteins in the HOME2 data set, listed by family	156
Table A.2 Sequence identities between pairs of proteins in the same HOME2 family, based on their SKA structural alignments	159
Table A.3 α -helical proteins in the HOME3 data set, listed by family	163
Table A.4 β -barrel-like proteins in the HOME3 data set, listed by family	168

List of Figures

Figure 1.1 Generic structure of amino acids	2
Figure 1.2 Generic structure of a peptide chain	2
Figure 1.3 Classification of the localization of proteins that are targeted by pharmaceutical drugs	3
Figure 1.4 Torsion angles in a polypeptide chain	4
Figure 1.5 Generic structure of an α -helix	5
Figure 1.6 Generic structure of β -strands	7
Figure 1.7 Chain A of the NanC porin (PDB code: 2WJR) as an example for a β -barrel-like structure	8
Figure 1.8 Structural differences of 3_{10} , α - and π -helices	10
Figure 2.1 Statistics about unique membrane protein structures that have been solved according to the database "Membrane Proteins of known 3D structure"	26
Figure 2.2 Total number of membrane proteins (including single-spans) that are in the PDB_TM database as a function of time	27

Figure 2.3 Clustering principle used for generation of the HOME _{P2} data set	30
Figure 2.4 Composition of the HOME _{P3} data set of homologous membrane protein structures	33
Figure 3.1 Determination of the fraction of correctly aligned positions	44
Figure 3.2 Determination of the Alignment Difference (AD) score	45
Figure 3.3 Correlations between alignment accuracy measures	50
Figure 3.4 Comparison of alignment accuracy when using single input descriptors in AlignMe	53
Figure 3.5 Comparison of alignment accuracy when using hydrophobicity scales as input descriptors in AlignMe	55
Figure 3.6 Comparison of alignment accuracy when using multiple input descriptors in AlignMe	58
Figure 3.7 Profiles of the predicted membrane propensity from the three different transmembrane helix prediction methods tested for AlignMe	59
Figure 3.8 Accuracy of HOME _{P2} alignments generated by different methods	64
Figure 3.9 Accuracy of homology models built from HOME _{P2} alignments generated by different methods	68
Figure 4.1 Screenshot of the AlignMe website (Jan, 2015) showing the option to select 4 different predefined modes for aligning α -helical membrane proteins.	82
Figure 4.2 Screenshot of the results page of the AlignMe website (Jan, 2015) showing the aligned hydrophobicity profiles based upon an alignment using the example sequences	84
Figure 4.3 Hydrophobicity profiles based on the alignment of the family-averaged hydropathy profiles (HP) of the two repeat units (repeat unit 1, left and repeat unit 2, right) in NaPi-II (red line) and VcINDY (black line) using AlignMe	88
Figure 4.4 Screenshot of the AlignMe website (Jan, 2015) showing the querylet for the HP mode of AlignMe	91
Figure 4.5 Screenshot of the results page of the AlignMe website (Jan, 2015) showing the aligned hydrophobicity profiles of the first sequences from each submitted multiple sequence alignment	92
Figure 4.6 Family hydropathy profile alignments	93
Figure 4.7 Family-averaged hydropathy profiles of the internal repeat units 1 (black) and 2 (red) for the following families: (A) APC, (B) BCCT, (C) NCS1, (D) NSS, (E) SSS	94
Figure 5.1 Example workflow for generating and evaluating a homology model	99
Figure 5.2 An alignment of Proteins A and B is derived from two other alignments (AC and BC) with Protein C as a reference. This derived alignment is then compared to its original alignment.	106
Figure 5.3 Alternate conformations in the family of major facilitator superfamily transporters	113

Figure 5.4 A consensus structure-based alignment fragment with confidence values.	121
Figure 5.5 Correlation of residue accuracy with confidence values based on a consensus of FR-TM-align, FATCAT, MATT and DaliLite alignments	122
Figure 6.1 Based on consensus alignments (A-C) between cytochrome c oxidase from <i>Rhodobacter sphaeroides</i> (PDB code: 2GSM), <i>cbb₃</i> cytochrome c oxidases (PDB code: 3MK7) and a nitric oxide reductase (PDB code: 3O0R) a multiple sequence alignment was manually created by assigning corresponding residues to each other (D)	132
Figure 6.2 Venn Diagram of assigned secondary structure states by SST, DSSP and STRIDE	134
Figure 6.3 Venn Diagram of residues that were predicted by SST and/or π -Detector to be in a π -helix and/or in a membrane-spanning segment according to definitions taken from the PDB_TM database.	135
Figure 6.4 Consensus alignment of the central segment of TM1 of LeuT (2A65) and AdiC (3OB6) belonging to the family of the FIRL fold	137
Figure 6.5 Manually created sequence alignment of GPCRs based on visual analysis of their structures only	138
Figure 6.6 Manually created multiple sequence alignment of transmembrane helix 2 (TM2) of G-protein coupled receptors of known structure based on pairwise consensus structural alignments	139
Figure 6.7 Manually created multiple sequence alignment of membrane helix 5 of G-protein coupled receptors based on pairwise consensus alignments	140
Figure 6.8 Protein structures and manually created multiple sequence alignment of cytochrome c oxidases subunit I based on pairwise consensus alignments	143
Figure 6.9 Residues in TM1 of LeuT (2A65) and AdiC (3OB6)	145
Figure 6.10 Residues in TM5 of Energy-Coupling Factor (ECF) Transporters	146

Abstract

Proteins are biological macromolecules that are encoded within an organism's genome. They are responsible for essential functions within an organism. Each protein consists of a set of amino acids that are connected together into a sequence in a peptide chain. Based upon this amino acid sequence and external factors (solvent, membrane etc.), a protein adopts a three-dimensional shape with specific functional properties. Proteins can be classified based on their surrounding into globular proteins located in the cytosol and membrane proteins located in a cell's membrane. Globular proteins contain only a few hydrophobic amino acids, whereas membrane proteins contain significantly more hydrophobic residues that interact with the hydrophobic membrane bilayer. Furthermore, the set of membrane proteins can be subdivided into two sub-classes that are defined by the regular structural elements with which the proteins cross the membrane: α -helical and β -barrel-like. Other structural elements like 3_{10} - or π -helices were also shown to occur in membrane proteins but only in local segments and not as longer regular structural elements. The spatial arrangement of secondary structure elements is called tertiary structure. Proteins can be clustered to families sharing a common fold that describes a generalized tertiary structure among a set of proteins. The use of spatial information from tertiary structures allows for a detection of transport pathways, binding pockets or bonding interactions. An arrangement of several tertiary structures of protein subunits into larger complexes is described by the quaternary structure of a protein. The interaction between different subunits has been shown to be crucial for a binding and transport of substrates through the membrane. In general, the location of membrane proteins in the membrane makes them responsible for a cell's function (e.g., transport), thus they are prominent drug targets. However, the extraction of membrane proteins from the hydrophobic membrane bilayer to determine high-resolution crystal structures is still a difficult task; thus only 2 % of all solved proteins structures are membrane proteins. Computational methods that allow for the detection of evolutionarily related protein sequences, structures or of important sequence patterns may help to gain deeper insights into membrane protein structures and their functions. This study will give an overview of such computational methods on a representative set of membrane proteins and will provide ideas for future computational and experimental research on membrane proteins.

A reliable and recent data set of membrane proteins was required to analyze and understand homology and evolutionary events between membrane proteins on a sequence and structural level. For this purpose I have updated an earlier, manually-curated data set of homologous membrane proteins (HOME1) to more recent versions in 2010 (HOME2) and 2013 (HOME3), using an automated clustering approach (chapter 2). For both data sets, all membrane proteins listed in the

PDB_TM database were downloaded as coordinate files from the Protein Data Bank and separated according to their overall fold: α -helical and β -barrel-like. Only crystal structures with a resolution better than 3.5 Å were considered to exclude non-reliable protein structures with a low resolution. Each protein structure was then split into its individual chains. For all pairs of proteins sharing the same number of membrane helices, a pairwise structural alignment was generated (using SKA for HOME2 and SKA as well as TM-align for HOME3). For the resulting structural alignments and their underlying sequence alignments specific similarity scores (PSD score of SKA, TM-score of TM-align) were calculated and threshold values were used for these scores to cluster protein chains to a common family. This clustering process resulted in a set of 81 α -helical proteins within 22 families and 177 alignments for HOME2. The newer data set HOME3 (generated 3 years later than HOME2) used an updated clustering approach and contains 152 α -helical proteins in 40 families with 354 alignments and 68 β -barrel-like proteins in 8 families with 319 alignments. Both data sets were used as a standard gold reference set for subsequent work.

In a first step (chapter 3), α -helical membrane protein sequences were used to determine descriptors that are suitable to describe an evolutionary relationship between homologous α -helical membrane proteins. So far, most sequence alignment methods were optimized on general data sets and only few sequence alignment methods were designed for membrane proteins, which have distinct evolutionary and structural properties different from globular proteins. These methods typically considered information about the membrane by using membrane-specific substitution matrices or by assigning different gap penalties in membranous and non-membranous segments. In this study, I have updated and applied the sequence alignment program AlignMe, which was initially created as a basic version in my diploma thesis. The program was extended to allow for position-specific substitution matrices and membrane propensity predictors as an input. Single input descriptors (substitution matrices, hydrophobicity, secondary structure and membrane propensities) were tested alone and in combination with each other in different modes of AlignMe by optimizing gap penalties on the HOME2 data set, which was used as a reference set. Most accurate alignments on the HOME2 data set were observed when using position-specific substitution information (P), secondary structure propensities (S) and transmembrane propensities (T) in the AlignMePST mode. These alignments were even more accurate than those of other commonly used sequence alignment methods but that was not surprising since AlignMe was optimized on HOME2. Thus, homology models were built for all protein pairs of HOME2 and evaluated using structural similarity scores. Homology models based upon alignments using the different AlignMe modes were shown to be more accurate than those based on alignments of other alignment methods. Moreover, AlignMePST, AlignMePS and AlignMeP modes were then tested together with other sequence alignment methods also on an independent reference set of membrane protein sequence alignments from the BALiBASE

collection. The combination of secondary-structure propensities (S) in combination with evolutionary information in form of position-specific substitution matrices (P) in the AlignMePS mode resulted in the most accurate alignments over a broad range of sequence similarities when compared to available methods. The application of transmembrane predictions (T) in addition to evolutionary information and secondary-structure predictions in the AlignMePST mode improved the alignment accuracy significantly for distantly-related proteins for which sequence information is less informative but resulted in less accurate alignment of closely-related proteins from the BALiBASE set relative to AlignMePS. The open source code of AlignMe is available at <http://www.forrestlab.org> and <https://sourceforge.net/projects/alignme/>, along with an online server and the HOME2 data set. This work was published in March 2013 in PLOS ONE.

The majority of frequently used computational methods for chemists and biologists are made available with a web server allowing for an easy access of those programs. In order to also allow an easier usage of the AlignMe program, I have implemented a web server of AlignMe (chapter 4) that provides the optimized settings and gap penalties for the AlignMeP, PS and PST modes. The server requires two sequences in fasta format as input and combines information about each sequence from multiple sources to produce a pair-wise alignment (PW mode). In addition, the alignment accuracy of the AlignMe web server is compared with those of other recent webservers on the set of membrane protein sequence alignments from the BALiBASE reference 7 set. Again, the alignments of AlignMe are shown to be more accurate than those of other programs, especially for very distantly related proteins for which the inclusion of membrane protein information has been shown to be suitable. Another mode that is provided on the AlignMe website allows for the alignment of two multiple sequence alignments to create family-averaged hydrophobicity profile alignments (HP mode). Each input multiple sequence alignment is converted into a hydrophobicity profile based upon a hydrophobicity scale and is then averaged over the provided set of sequence homologs. The two profiles are then aligned with each other. The HP mode enables a qualitative comparison of transmembrane topologies (and therefore potentially of 3D folds) of two membrane proteins, which can be useful if the proteins have low sequence similarity. In summary, the AlignMe web server provides user-friendly access to a set of tools for the analysis and comparison of membrane protein sequences. An access is available at <http://www.bioinfo.mpg.de/AlignMe>. This work was published in the NAR web server issue in July 2014.

Besides sequence similarity, also structural similarity can be applied to detect evolutionarily and functionally related positions or fragments of two protein structures (chapter 5). Although many structural alignment methods are available, there is a lack of programs that are optimized on and for membrane proteins. All available studies have assessed alignment accuracy and consistency only on

general protein data sets and did not explicitly consider the distinct class of membrane proteins. Thus, the choice of a suitable program is not apparent to a user who wants to generate structural alignments of membrane proteins. Consequently, I compared 13 widely-used pairwise structural alignment methods on an updated reference set of homologous membrane protein structures (HOME3) that includes α -helical and β -barrel-like membrane proteins. Each pair of protein structures was aligned and the underlying sequence alignment was then used to construct homology models as in chapter 3. The model accuracy compared to the known structures was assessed using scoring functions that were not used by the tested structural alignment methods (e.g., AL4 or CAD-score). The analysis shows that fragment-based approaches such as FR-TM-align are the most useful ones for aligning structures of membrane proteins but none of the fragment-based approaches was clearly superior to all other methods. Moreover, fragment-based approaches are more suitable for a comparison of protein structures that have undergone large conformational changes, whereas rigid approaches were more suitable for proteins that were solved in the same or a similar state but again no method showed a significantly higher accuracy than all other. Additionally, all methods lack a measure to rate the reliability of the accuracy for a specific position within a structure alignment and thus a user does not know if a position is confidently aligned or not. In order to solve these problems, I propose a consensus-type approach by combining alignments from four different methods, namely FR-TM-align, Dalilite, MATT and FATCAT. A confidence value that describes the agreement between the methods is assigned to each position of the alignment. This work has been published in August 2015 in the journal "PROTEINS: structure, function and bioinformatics".

Consensus alignments were then generated for each pair of proteins of the HOME3 data set and subsequently analyzed for single evolutionary events within membrane-spanning segments (chapter 6). In addition, I checked all membrane proteins of the HOME3 data set for irregular structures (e.g., 3_{10} - and π -helices) using structural assignment methods and a custom script to detect π -helices. Interestingly, single insertions and deletions (InDels) were observed in different families of membrane-spanning segments in which an α -helix in one protein was aligned with a gap to a π -helix in the other protein. In agreement with a recent study, a single gap was observed in G-Protein coupled receptors in TM2 and TM5 but in slightly different positions than proposed before. In both cases, binding specificity might be influenced by the presence or absence of an α - or a π -helix. A single InDel was also observed in the proton pathways of proteins belonging to the family of cytochrome c oxidase subunit I. The activity of the D- and K-pathways might be dependent on the presence and absence of π -helices in TM2 and TM9. Also among proteins belonging to the FIRL fold, a single gap was observed in TM1 of AdiC compared to LeuT. An additional arginine is present in LeuT that points into the extracellular pathway at a position that is known to be important for the protein's function. Last, consensus structural alignments of energy coupling factor (ECF) transporters

also reveal a single gap with high confidence in a π -helical segment of BioY that is aligned to α -helical segments in ThiT and RibU. Both proteins contain an amino acid with a large side chain that points into a cavity that is crucial for transport specificity whereas in BioY a glycine is present at this position.

This study shows that computational methods need to be adapted and optimized for membrane proteins in order to achieve results with a higher accuracy. Membrane-specific information has been shown to be suitable for aligning distantly related proteins on a sequence level. Such information is not incorporated into structural alignment programs so far. However, structural alignment methods that allow for fragment-based flexibility were shown to be a suitable choice for membrane proteins that undergo conformational changes. Interestingly, single insertions and deletions could be observed with the help of consensus alignments in the conserved membrane-spanning segments of membrane proteins. The detection of such single InDels might help to identify crucial residues for a protein's function.

Zusammenfassung

Proteine sind biologische Makromoleküle, die im Genom eines Organismus kodiert sind und im Organismus für essentielle Funktionen verantwortlich sind. Jedes Protein besteht aus einer Sequenz miteinander verbundener Aminosäuren, welche eine Peptidkette formen. Basierend auf dieser Aminosäuresequenz und externen Faktoren (Solvens, Membran etc.) nimmt ein Protein eine dreidimensionale Struktur mit spezifischen Eigenschaften an. Schon anhand ihrer Umgebung können Proteine in zwei Klassen eingeteilt werden. Es gibt globuläre Proteine, die sich im Zytosol befinden und Membranproteine, die sich in der Zellmembran befinden. Im Gegensatz zu globulären Proteine beinhalten Membranproteine signifikant mehr hydrophobe Reste, die mit der hydrophoben doppelschichtigen Zellmembran interagieren. Die Klasse der Membranproteine kann weiterhin basierend auf den vorwiegend vorherrschenden Sekundärstrukturelementen in α -helikale und β -Fass-ähnliche Proteine unterteilt werden. Die räumliche Anordnung aller Sekundärstrukturelemente wird Tertiärstruktur genannt. Mit Hilfe einer generalisierten Tertiärstruktur (einer so genannten Faltung, „fold“) lassen sich Membranproteine in unterschiedliche funktionelle Familien einteilen. Eine räumliche Anordnung verschiedener Tertiärstrukturen von Proteinuntereinheiten wird Quartärstruktur genannt. Durch die Interaktion zwischen verschiedenen Proteinuntereinheiten von Membranproteinen sind Bindung und Transport von Substraten durch die Membran möglich, weswegen Membranproteine Hauptziele für die Wirkstoffentwicklung sind. Allerdings ist die Bestimmung hochauflösender Kristallstrukturen für Membranproteine immer noch eine schwierige Aufgabe, bedingt durch ihre Lage in der hydrophoben doppelagigen Membran, weswegen nur 2% aller kristallisierten Proteine zur Klasse der Membranproteine gehören. Computergestützte Anwendungen zur Erkennung von evolutionär verwandten Proteinsequenzen, -strukturen oder wichtiger Sequenzmuster können dabei helfen weitere Einblicke in Strukturen von Membranproteinen und deren Funktionen zu gewinnen. Diese Arbeit wird sowohl eine Übersicht über die Genauigkeit solcher computergestützten Methoden geben als auch eine Anwendung zeigen, um funktionell wichtige Proteinsegmente in Membranproteinen zu finden und zu analysieren.

Zur Bestimmung evolutionärer Ereignisse zwischen Membranproteinen auf Sequenz- und Strukturebene wird ein zuverlässiger und aktueller Datensatz von homologen Membranproteinen benötigt. In Kapitel 2 werden zwei Aktualisierungen eines manuell erstellten Datensatzes homologer Membranproteine (HOME) beschrieben: HOME2 und HOME3. Die Aktualisierungen wurden mittels eines automatisierten Klassifizierungsverfahrens durchgeführt (Kapitel 2). Für beide Datensätze wurden alle Membranproteine (gelistet in der PDB_TM Datenbank) als Strukturdateien von der Protein Data Bank heruntergeladen und in α -helikale und β -Fass-ähnliche Proteine unterteilt. Anschließend wurde jede Proteinstruktur in ihre individuellen Ketten aufgeteilt. Für alle Proteinpaare

mit der gleichen Anzahl von Transmembransegmenten wurde ein paarweises strukturelles Alignment generiert. Die Proteinketten wurden basierend auf Ähnlichkeitswerten, die für die strukturellen Alignments (PSD-score von SKA und TM-score von TM-align) und deren zugrunde liegenden Sequenzalignments berechnet wurden, in verschiedene Familien eingeteilt. Dies resultierte in einem Satz von 81 α -helikalen Proteinen in 22 Familien und 177 Alignments für HOME2. Für den aktuelleren Datensatz HOME3 wurde ein verbesserter Klassifizierungsansatz verwendet, der zu 152 α -helikalen Proteinen in 40 Familien mit 354 Alignments und 68 β -Fass-ähnlichen Proteinen in 8 Familien mit 319 Alignments führte. Beide Datensätze wurden als Gold-Standard-Referenz für anschließende Analysen benutzt.

In einem ersten Schritt (Kapitel 3) wurde nach Deskriptoren gesucht mittels welcher das evolutionäre Verhältnis zwischen homologen α -helikalen Membranproteinen passend beschrieben werden kann. Bisher wurde der Großteil der Programme für Sequenzalignments auf generellen Datensätzen optimiert, obwohl Membranproteine deutlich andere evolutionäre und strukturelle Eigenschaften besitzen als globuläre Proteine. Einige Programme wurden speziell für Membranproteine entwickelt, doch diese berücksichtigten bisher nicht Hydrophobizitätsprofile oder Transmembranwahrscheinlichkeitswerte als Eingabe für Alignments. Im Rahmen dieser Arbeit wurde das speziell für Membranproteine erstellte Alignmentprogramm AlignMe, von dem eine erste Version bereits im Rahmen meiner Diplomarbeit erstellt wurde, zu einer neueren Version erweitert und anschließend angewendet. Die Erweiterung von AlignMe erlaubt nun Alignments basierend auf positions-spezifischen Substitutionsmatrizen (PSSMs) und Transmembranwahrscheinlichkeitsvorhersagen. Einzelne Proteindeskriptoren (z.B. Substitutionsmatrizen oder Sekundärstrukturvorhersagen) wurden sowohl einzeln als auch in Kombination miteinander in verschiedenen Modi von AlignMe getestet. Hierfür wurden Strafwerte für Lücken im Alignment („gap penalties“) auf dem HOME2-Datensatz optimiert. Die akkuratesten Alignments wurden erzielt, wenn eine positionsspezifische Substitutionsmatrix (P) in Kombination mit einer Sekundärstrukturvorhersage (S) und Transmembranwahrscheinlichkeiten (T) im AlignMePST-Modus verwendet wurden. Es war nicht überraschend, dass diese Alignments akkurater waren als die anderer, häufig verwendeter Programme, da die Referenzalignments des HOME2-Datensatzes sowohl für die Optimierung als auch für die Auswertung genutzt wurden. Deshalb wurden anschließend Homologie-Modelle für alle Proteinpaare des HOME2-Datensatzes basierend auf den zugrunde liegenden Alignments der verschiedenen getesteten Programme generiert. Homologie-Modelle der verschiedenen AlignMe Modi waren akkurater als die Homologie-Modelle anderer Programme. In einem letzten Schritt wurden die AlignMe Modi zusammen mit anderen Programmen auf einem unabhängigen Referenzdatensatz bestehend aus Sequenzalignments von Membranproteinen (BALiBASE Referenz 7) getestet. Die Kombination einer

Sekundärstrukturvorhersage (S) mit einer positionsspezifischen Substitutionsmatrix (P) im AlignMePS Modus resultierte in den akkuratesten Alignments über eine große Spanne von Sequenzähnlichkeitswerten im Vergleich zu anderen verfügbaren Methoden. Die zusätzliche Anwendung von Transmembranwahrscheinlichkeiten (T) im AlignMePST-Modus verbesserte die Akkuratheit der Alignments signifikant für entfernt verwandte Proteine, resultierte aber relativ zu AlignMePS in weniger akkuraten Alignments von nah verwandten Proteinen. Der Open Source Code von AlignMe ist verfügbar unter <http://www.forrestlab.org>, zusammen mit dem HOMEPE2-Datensatz. Dieses Kapitel wurde im März 2013 in PLOS ONE veröffentlicht.

Die Mehrheit der häufig verwendeten Computerprogramme für Chemiker und Biologen ist über Webserver verfügbar, die einen einfachen Zugriff auf diese Programme ermöglichen. Um ebenso die Verwendung von AlignMe zu vereinfachen, wurde ein Webserver für AlignMe programmiert (siehe Kapitel 4), welcher optimierte Einstellungen für die AlignMeP-, PS- und PST-Modi sowie einen schnellen Modus zur Verfügung stellt. Für ein paarweises Sequenzalignment werden zwei Aminosäuresequenzen im Fasta-Format benötigt. Diese Sequenzen werden dann basierend auf den ausgewählten Modi und den entsprechenden Proteindeskriptoren miteinander verglichen und aligniert. Ein Vergleich mit anderen Webservern basierend auf den BALiBASE-Referenz-7-Alignments zeigte, dass die Alignments der verschiedenen Modi des AlignMe Webserver akkurat sind, insbesondere bei sehr entfernt verwandten Proteinen, für welche die Einbeziehung von Transmembranwahrscheinlichkeiten vorteilhaft ist. Neben dem Modus für paarweise Alignments (PW-Modus) steht auf der AlignMe Webseite auch ein Modus zur Verfügung, mit welchem man ein Alignment von zwei multiplen Sequenzalignments basierend auf einem gemittelten Hydrophobizitätsprofil generieren kann (HP Modus). Hierzu wird jedes der beiden eingegebenen multiplen Sequenzalignments basierend auf einer Hydrophobizitätsskala in ein Hydrophobizitätsprofil übersetzt und anschließend über den Satz von Sequenzhomologen gemittelt. Diese beiden Profile werden dann miteinander aligniert. Dieser HP-Modus ermöglicht einen qualitativen Vergleich der Transmembrantopologie von zwei Membranproteinen. Der AlignMe Webserver ist verfügbar unter: <http://www.bioinfo.mpg.de/AlignMe/>. Dieses Kapitel wurde im Juli 2014 im jährlichen NAR Webserver Issue veröffentlicht.

Eine Detektion evolutionär und funktionell verwandter Sequenzpositionen kann nicht nur basierend auf der Aminosäuresequenz geschehen, sondern auch auf struktureller Ebene (Kapitel 5). Ebenso wie für Sequenzalignment-Programme besteht auch bei Programmen für Strukturalignments ein Mangel an Programmen, die für Membranproteine optimiert sind. Daher ist die Wahl eines passenden Programms für einen Benutzer, der Strukturalignments von Membranproteinen generieren möchte, nicht einfach. Aus diesem Grund wurden 13 häufig verwendete Programme für paarweise

Strukturalignments auf einem aktualisierten Referenzdatensatz von homologen Membranproteinstrukturen (HOME3) getestet und verglichen. Für jedes Paar von Proteinstrukturen wurde ein Strukturalignment generiert, und das zugrundeliegende Sequenzalignment wurde dann benutzt, um Homologiemodelle zu generieren, deren Genauigkeit anschließend ausgewertet wurde. Dazu wurden die Homologiemodelle mit den bereits bekannten Kristallstrukturen verglichen und deren Ähnlichkeit wurde mit Hilfe von Bewertungsfunktionen ausgewertet (z.B. AL4 oder CAD-score). Diese Analyse zeigte, dass fragmentbasierte Methoden (z.B. FR-TM-align) akkurate Strukturalignments generieren. Allerdings ist keines der getesteten Programme signifikant besser als alle anderen Programme. Des Weiteren zeigte sich, dass sich fragmentbasierte Programme zum Vergleich von Proteinstrukturen eignen, die in unterschiedlichen Konformationen kristallisiert sind. Im Gegensatz dazu sind rigide Programme besser geeignet für Proteine, die in einer ähnlichen Konformation kristallisiert wurden. Allerdings besitzt keines der getesteten Programme einen Messwert für eine positionsspezifische Zuverlässigkeit der generierten Alignments, so dass ein Benutzer nicht weiß, welche Fragmente akkurat aligniert sind oder welche eher problematisch sind. Um dieses Problem zu lösen, wurde ein Konsensus-ähnlicher Ansatz vorgeschlagen, der die Alignments von vier verschiedenen Methoden miteinander kombiniert: FR-TM-align, DaliLite, MATT und FATCAT. Für jede Position des Konsensus-Alignments wird ein Konfidenzwert berechnet, welcher die Übereinstimmung der Alignments der verschiedenen Programme beschreibt. Dieses Kapitel wurde 2015 im Journal "PROTEINS: structure, function and bioinformatics" veröffentlicht.

Mit Hilfe von Konsensus-Alignments wurden dann alle Proteinpaare des HOME3-Datensatzes analysiert, um einzelne evolutionäre Vorkommnisse in Transmembransegmenten zu entdecken (Kapitel 6). Zusätzlich wurden die Proteine von HOME3 auf irreguläre Strukturelemente untersucht mit Hilfe von Methoden, die zur Bestimmung von Sekundärstrukturelementen (z.B. SST) dienen, und einem eigenen Skript, um π -helikale Elemente zu finden. Interessanterweise traten einzelne evolutionäre Insertionen und Deletionen (InDels) in den Transmembransegmenten von vier verschiedenen Proteinfamilien auf. In allen Fällen war eine α -Helix in einem Protein zu einer π -Helix im jeweils anderen Protein aligniert. In Übereinstimmung mit einer aktuellen Studie wurden einzelne Gaps in den Membranhelices 2 und 5 der Familie der G-Protein-gekoppelten Rezeptoren gefunden. An diesen Stellen der Proteine könnte die Spezifität der Substratbindung abhängig sein von der An- oder Abwesenheit einer α - oder π -Helix. Allerdings scheinen die Gaps an einer leicht verschobenen Stelle zu sein als bisher angenommen. Ein einzelner InDel wurde ebenfalls in den Protonenpfaden der Untereinheit I der Cytochrom-C-Oxidasen entdeckt. Die Aktivität der D- und K-Pfade könnte abhängig von der An- oder Abwesenheit von π -Helices in den Membranhelices 2 und 9 sein. Auch in Proteinen mit einer FIRL-Faltung wurde in Membranhelix 1 von AdiC im Vergleich zu LeuT ein einzelner InDel gefunden. In LeuT ist ein zusätzliches Arginin an einer funktionell wichtigen

Position vorhanden. Ein letztes Beispiel für einen einzelnen InDel lässt sich in der Familie der „energy coupling factor (ECF)“-Transporter finden, in welcher ein π -helikales Segment von BioY zu einem α -helikalen Segment in ThiT und RibU aligniert ist. ThiT und RibU besitzen eine Aminosäure mit langer Seitenkette, die in eine für die Transportspezifität wichtige Vertiefung deutet, wohingegen in BioY an dieser Stelle ein Glycin vorhanden ist.

Diese Doktorarbeit verdeutlicht, dass computergestützte Methoden an den speziellen Satz von Membranproteinen angepasst und optimiert werden müssen, um akkurate Ergebnisse zu erhalten. Insbesondere für Sequenzalignments von entfernt verwandten Proteinen zeigte sich eine membranspezifische Information als nützlich. Für Strukturalignments wurde eine solche Information bisher allerdings nicht verwendet, doch fragment-basierte flexible Programme für Strukturalignments zeigten sich als gute Wahl für Alignments von Membranproteinstrukturen, die in unterschiedlichen Konformationen kristallisiert wurden. Interessanterweise konnten mit Hilfe von Konsensus-Alignments einzelne InDels von Aminosäuren in Transmembransegmenten von Membranproteinen gefunden werden. Die Detektion solcher InDels kann helfen, funktionell wichtige Aminosäurereste eines Proteins zu finden.

1 Introduction

1.1 Globular & Membrane Proteins

Proteins are biological macromolecules that are located within all cells being responsible for essential functions within an organism (i.e. enzymatic catalysis, control of cell growth and differentiation by cell signaling, transport of molecules through cells etc.). Each protein is encoded within an organism's genome and can be synthesized by a translation of the genomic code to a set of amino acids connected with each other. In general, all proteins are composed of twenty standard amino acids (see Figure 1.1) that are connected together in a peptide chain by peptide bonds between the amino and carboxyl groups of adjacent amino acids (see Figure 1.2). Next to the twenty standard amino acids, there are also certain organisms that contain selenocysteine (Johansson, et al., 2005) and certain archaea that contain pyrrolysine (Srinivasan, et al., 2002) within their protein sequences. Based upon the amino acid sequence and external factors (chaperone, solvent, membrane), proteins are able to adopt a three-dimensional shape and specific functional properties. This evolutionarily differentiation allows for a classification of proteins to gain deeper insights into their evolutionarily and functional relationships. A major distinction between proteins can be made by classifying them based upon their environment, which has a major influence on the amino acids composition and three-dimensional structure of a protein. Proteins can be classified into a large set of globular proteins located in the cytosol and to a smaller set of membrane proteins that are located within a cell's membrane. Globular proteins are surrounded by a hydrophilic environment that causes a low content of hydrophobic residues within these proteins. The majority of hydrophobic residues in globular proteins is typically located within the inside of the proteins but there are also some hydrophobic residues at the surface of globular proteins for providing interaction patterns with other proteins or molecules (Moelbert, et al., 2004). In contrast to globular proteins, membrane proteins are located within a hydrophobic environment of the membrane lipid bilayer. Therefore, they contain significantly more hydrophobic amino acids than globular proteins (Gromiha and Suwa, 2003; Ulmschneider and Sansom, 2001). These hydrophobic amino acids are located predominantly on surfaces of the membrane protein facing the membrane bilayer. Hydrophilic amino acids are typically observed within loop segments outside of the membrane or at interaction sites within the membrane protein (e.g., ligand binding or helix-helix packing).

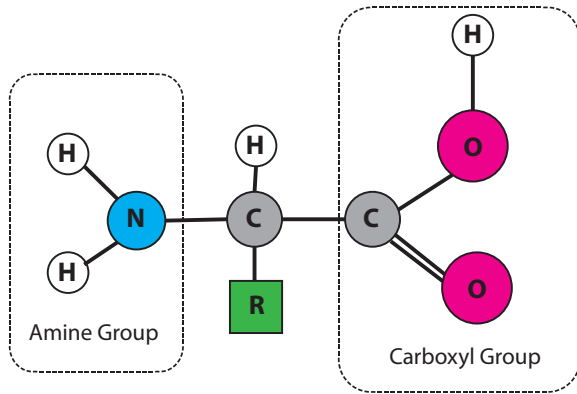


Figure 1.1 Generic structure of amino acids. All amino acids contain an amine group (NH_2) and a carboxyl group (CO_2H). The rest (R) represents a side chain that is specific for each amino acid type (e.g., $\text{R} = \text{H}$ for Alanine or $\text{R} = \text{CH}_2$ for Valine).

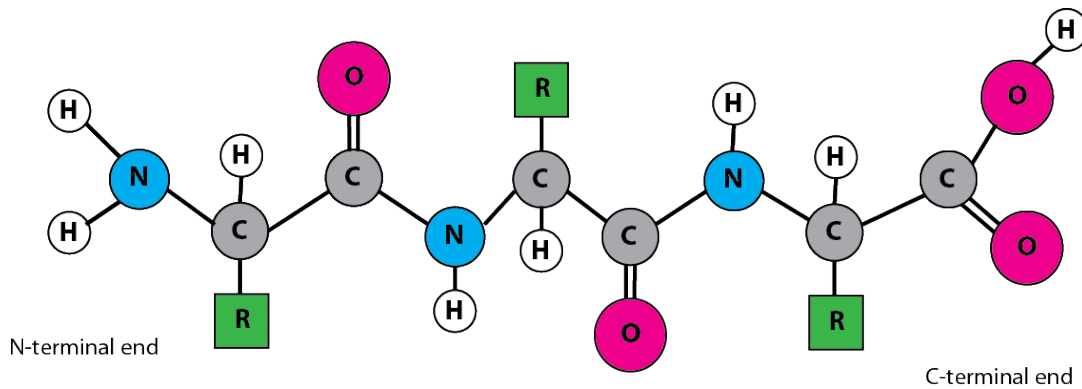


Figure 1.2 Generic structure of a peptide chain. Amino acids are connected together into a peptide chain by peptide bonds between the amino and carboxyl groups of adjacent amino acids.

The impermeable membrane bilayer does not only influence the composition and structure of proteins, it also makes membrane proteins a crucial class of proteins that are responsible for important cell functions like transport, signaling and cell adhesion, as well as recognition. This functional importance is reflected by the fact that $\sim 30\%$ of all proteins of genomes, that have been sequenced so far, belong to the class of membrane proteins (Jones, 1998; Krogh, et al., 2001; Nugent and Jones, 2009). Accordingly, membrane proteins constitute $>50\%$ of targets for active pharmacological drugs on the market (Drews, 2000; Hopkins and Groom, 2002; Uhlen, et al., 2015) (see Figure 1.3).

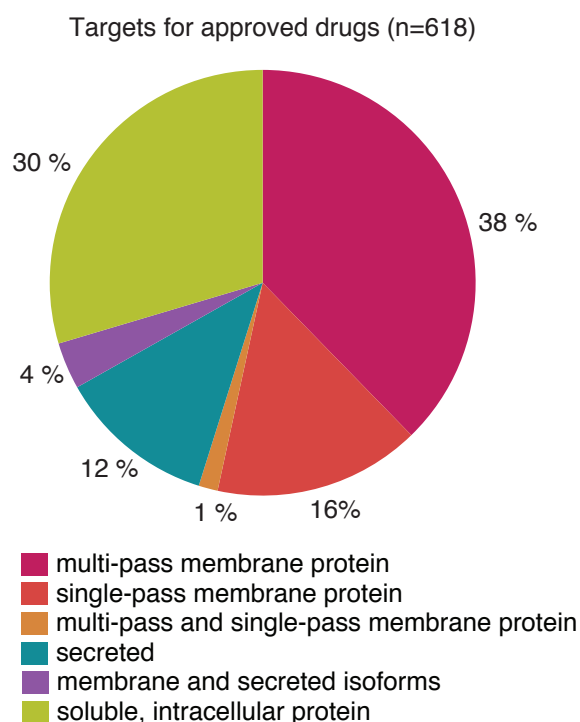


Figure 1.3. Classification of the localization of proteins that are targeted by pharmaceutical drugs, which are approved by the FDA. Figure taken from (Uhlen, et al., 2015).

However, there is a lack of structural data for membrane proteins because only 2 % of all solved protein structures belong to the class of membrane proteins. They are represented by ~1700 high-resolution structures in the Protein Data Bank (Berman, et al., 2003) of which ~540 are unique (<http://blanco.biomol.uci.edu/mpstruc/>, May 2015). The over-expression and extraction of membrane proteins from the hydrophobic membrane-bilayer to determine experimentally high-resolution X-ray structures is still a difficult task (Ostermeier and Michel, 1997) but there is a steady increase of the number of crystallized high-resolution membrane protein structures and complexes (Bill, et al., 2011; White, 2004). In this context, computational methods may help to understand and gain deeper insights into important features of membrane proteins. They allow for the detection of evolutionarily related proteins, structure predictions, the identification of important sequence patterns and many more aspects for enlightening a protein's properties (Arinaminpathy, et al., 2009). Consequently, computational methods are commonly used during structure elucidation and for explaining the occurrence and function of a solved structure. However, computational methods are typically developed for proteins in general and do not differentiate between globular or membrane proteins, despite their differences. This lack of computational methods for membrane proteins is addressed by this study that will give an overview of recent developments of computational methods on membrane proteins and will provide ideas as well as improvements for future research on membrane proteins.

1.2 Protein Properties & Structures of Membrane Proteins

1.2.1 General

First of all, the main properties of membrane proteins have to be detected for being able to apply computational methods upon this class of proteins. Membrane proteins themselves can be divided into two sub-classes that are defined by local regular protein structures: α -helical and β -barrel-like membrane proteins. These regular structures describing local three-dimensional protein shapes are called secondary structure elements and were discovered within three-dimensional structures of crystallized peptide structures (Pauling, et al., 1951). The hydrogen-bonding patterns as well as the dihedral angles of a segment of consecutive amino acids define the type of secondary structure element. A Ramachandran plot (Ramachandran, et al., 1963) allows for a fast overview of a protein's phi- (torsion angle around the N-C $_{\alpha}$ bond) and psi- (torsion angle around the C $_{\alpha}$ -C bond) angles. Specific combinations of those angles correspond to specific secondary structure types (see Figure 1.4). The most common well-ordered secondary structure elements are α -helices followed by β -sheets, 3_{10} - and π -helices. Each of these secondary structure types has unique features that will be described in the next chapters. Besides these regular structures, there are also unordered secondary structure elements that all belong to the very common group of coils.

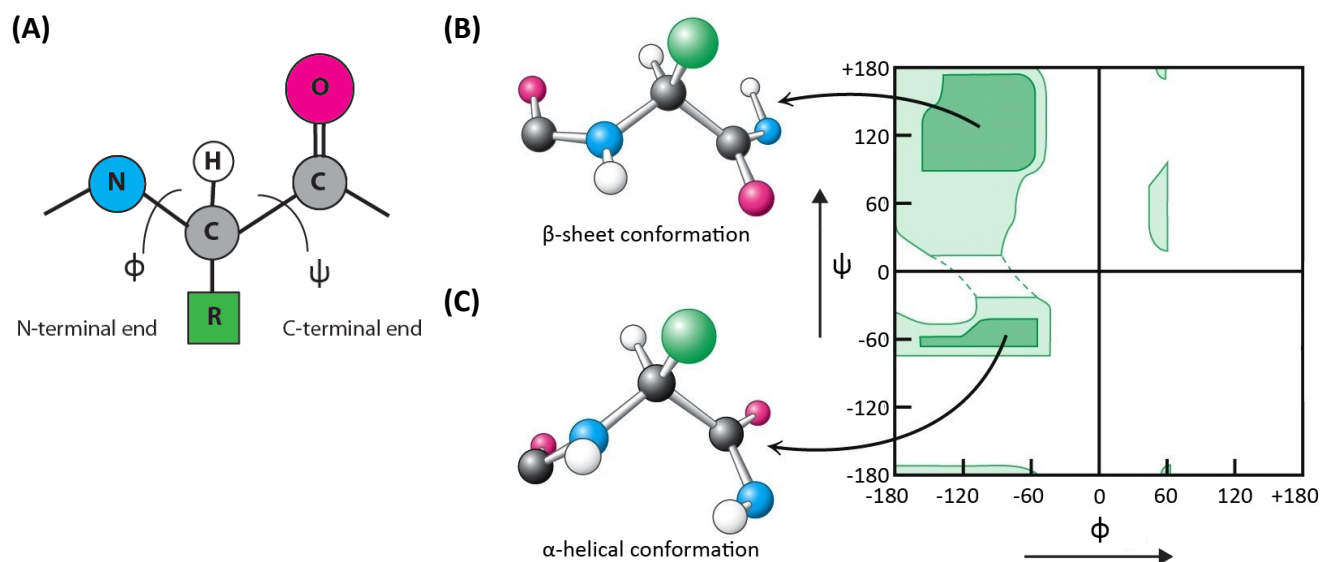


Figure 1.4 Torsion angles in a polypeptide chain. (A) Two angles that define the conformation of the polypeptide chain are the phi-torsion angle (ϕ) around the N-C $_{\alpha}$ bond and the psi-torsion angle (ψ) of the C $_{\alpha}$ -C bond. (B) Torsion angles of a β -sheet. (C) Torsion angles of an α -helix. Figures B and C are adapted from (Berg, 2010).

1.2.2 α -helical Structures

The most prevalent secondary structure type is the α -helical structure. α -helices are amino acid peptides that are arranged in a right-handed helical structure with 3.6 amino acids per turn, an average rise of 1.56 Å per residue and an hydrogen bond between the carbonyl oxygen of every i and the amid proton of every $i+4$ residue (see Figure 1.5). Such a regular intra-chain hydrogen-bonding pattern stabilizes the α -helical structures in an energetically favorable conformation with a high number of electrostatic dipole-dipole interactions.

The relative frequencies and propensities of specific amino acids in helical segments were analyzed for predicting α -helical structures in proteins (Chou and Fasman, 1978; Pace and Scholtz, 1998). Including evolutionarily information (Jones, 1999) as well as the usage of more complex computational methods (i.e. machine learning) (Karplus, 2009) improved the accuracy of predicting helical segments and other secondary structure types.

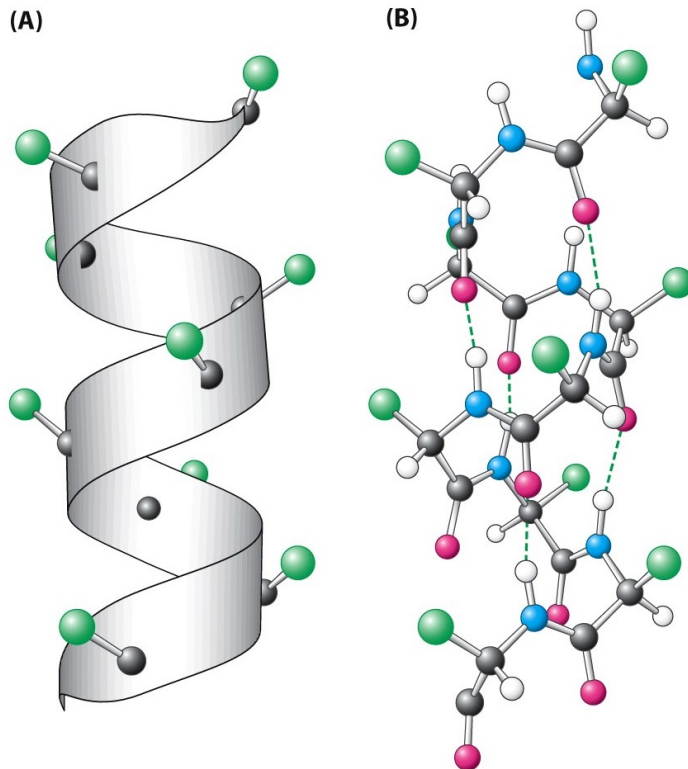


Figure 1.5 Generic structure of an α -helix. (A) Cartoon representation of an α -helix. (B) Detailed representation showing a hydrogen bonding pattern between the carbonyl oxygen of every i and the amid proton of every $i+4$ residue. Figure taken from (Berg, 2010).

Membrane-spanning segments adopting an α -helical structure were shown to contain more non-polar (alanine, valine, glycine) or large hydrophobic (phenylalanine, leucine, isoleucine) amino acids compared to non-membranous helical segments (Senes, et al., 2000). In interfacial segments at the border of the membrane, aromatic residues (tyrosine, phenylalanine) or residues that have the ability to form H-bonds (histidine) were observed more frequently than usual. Charged or polar residues (aspartate or glutamate) occur less frequently in membrane-spanning segments in general, and if so then they require interactions that satisfy their binding energy (e.g., Q32 of chain J from the chicken ubiquinol cytochrome c oxidoreductase is part of a membrane helix and might interact with a lipid head group) (Ulmschneider and Sansom, 2001). The basic amino acids arginine and lysine are also rarely observed in membrane-spanning segments. Interestingly, arginine and lysine have been observed to occur (three to four times) more often in the cytoplasmic domain than in the periplasmic domain. This amino acid distribution within non-membranous loop segments is also known as the positive-inside rule (von Heijne, 1989). Alterations (mutations or deletions) in terminal non-membranous loop segments were shown to change the orientation in which a protein is inserted into a membrane (e.g., adding four positively charged lysine residues to the N-terminus of a leader peptidase caused a topology change from a $N_{out}-C_{out}$ topology to a $N_{in}-C_{in}$ topology (von Heijne, 1989).

The distribution of specific amino acids within α -helices was furthermore examined based upon experimental and computational analysis and several hydrophobicity scales were developed (Koehler, et al., 2009) and used for predicting α -helical segments crossing the membrane (Dobrowolski, et al., 2007). Additionally, more complex computational methods were developed that include evolutionarily information for predicting membrane-spanning helices (Nugent and Jones, 2009; Viklund and Elofsson, 2008). However, the correct assignment of the first and last residue of a membranous α -helix is still a challenging task caused by several properties of the protein and the membrane itself. First of all, membrane helices do not cross the membrane in a straight way. Aside from being tilted and not straight in the membrane, they can contain twists, kinks or even broken, unwound fragments (Werner and Church, 2013). Second, there are re-entrant helices that enter and leave the membrane on the same side (Viklund, et al., 2006). At last, the membrane can vary in its size from 27 – 42 Å (Ulmschneider and Sansom, 2001). All these aspects cause membrane helices to vary between 22 and 32 residues. Aside from these residues within the membrane, a membrane-spanning helix also extends into non-membranous segments, which is another factor that contributes to the length of the helix.

1.2.3 β -strands, -sheets & -barrels

The second most common secondary structure elements are β -strands. A β -stranded structure is a single polypeptide chain with a backbone in an almost fully stretched conformation. Several β -strands located next to each other and connected by hydrogen bonds are called β -sheets. Three types of β -sheets exist that are all named according to their hydrogen-bonding pattern: parallel, antiparallel or mixed β -sheets (see Figure 1.6).

Transmembrane proteins with β -barrels mostly appear in the asymmetric outer membrane of bacteria, chloroplasts and mitochondria. In gram-negative bacteria, only β -barrel proteins are located within the outer membrane whereas in mitochondria and chloroplasts there are α -helical as well as β -barrel-like proteins (Wimley, 2003). However, there are also two atypical β -TM proteins (MspA and α -hemolysin) in gram-positive bacteria that typically do not contain β -barrel-like proteins (Remmert, et al., 2010). The β -strands are arranged predominantly in antiparallel β -sheets like a barrel with a hydrogen-bonding between the first and the last strand (see Figure 1.7).

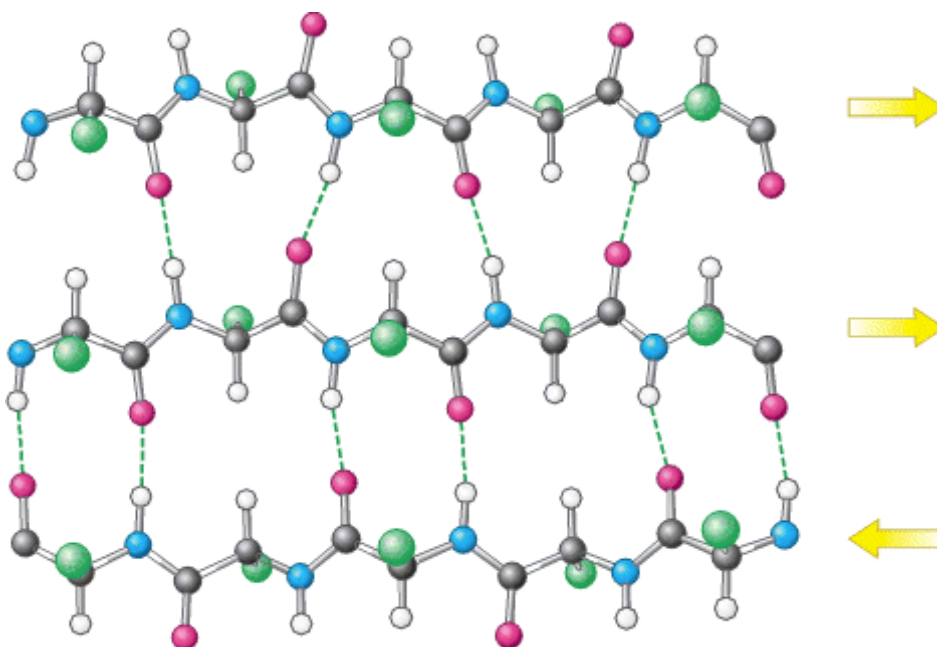


Figure 1.6 Generic structure of β -strands. β -strands that are connected with hydrogen bonds to β -sheets are called either parallel or anti-parallel β -sheets dependent on the orientation of the β -strands to each other. A protein that contains both types (parallel and anti-parallel) has overall mixed β -sheets. Figures taken from (Berg, 2010).

The amino acid composition of membrane-spanning segments of β -barrel-like proteins is mainly influenced by the hydrophobic membrane-bilayer and the hydrophilic pore or plug domain inside of the barrel. Thus, membrane-spanning segments in β -barrels have a high propensity of charged residues (asn, asp, gln, glu, arg, lys) located at positions that are facing the pore and hydrophobic residues at positions that are facing the membrane bilayer (Gromiha and Suwa, 2005). This observation of alternating amino acids was applied for predicting membrane-spanning β -strands using Hidden Markov Models (e.g., PRED-TMBB (Bagos, et al., 2004), PROFtmb (Bigelow, et al., 2004)).

Another difference between α -helical and β -barrel-like proteins is the membrane they are located in. α -helical proteins are located in symmetric phospholipid bilayers (e.g., eukaryotic membranes and prokaryotic inner membranes) whereas β -barrel-like proteins are inserted in an asymmetric membrane (e.g., prokaryotic outer membranes) with phospholipids on one and lipopolysaccharides on the other side. Accordingly, there is also a rule for the location of positively charged amino acids in non-membranous loop segments of β -barrels. Positively charged amino acids (e.g., R, K) occur prevalently in the extracellular cap (Jackups and Liang, 2005) and were shown to interact with non-membranous fragments of lipopolysaccharides (LPS) (Ferguson, et al., 2000; Kukkonen, et al., 2004) resulting in a positive-outside rule for β -barrel-like proteins opposite to the positive-inside rule for α -helical proteins (Jackups and Liang, 2005).

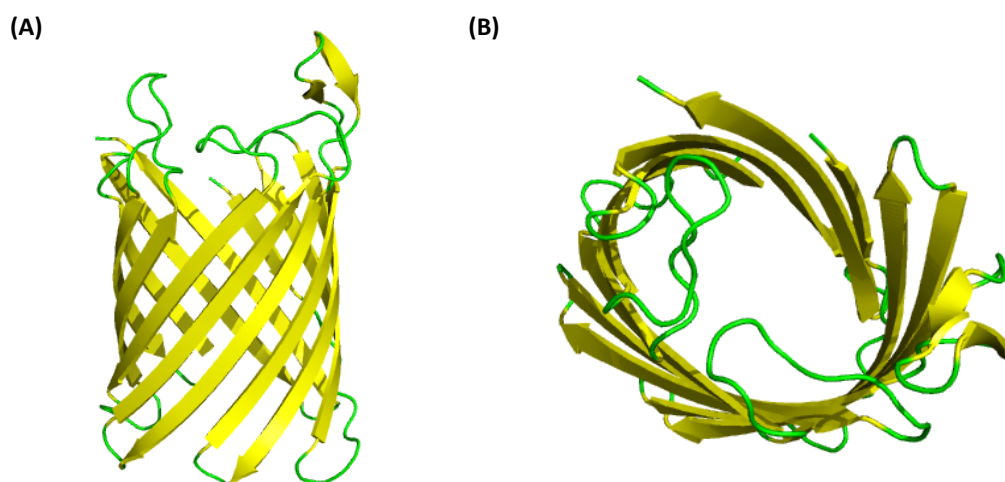


Figure 1.7 Chain A of the NanC porin (PDB code: 2WJR) as an example for a β -barrel-like structure. (A) View from the side. (B) View from top.

As for α -helical proteins, the observed β -sheet forming propensities (Fooks, et al., 2006; Minor and Kim, 1994) of amino acids were used in computational methods to predict membrane-spanning segments of β -barrel-like proteins (Hayat and Elofsson, 2012; Randall, et al., 2008). However, an accurate prediction of membrane-spanning β -strand segments is more difficult than the prediction of α -helical transmembrane segments. The simple usage of hydrophobicity is not sufficient because the hydrophobicity pattern is not perfectly regular due to twists within the sheets (Remmert, et al., 2010). Additionally, the β -strands of β -barrel-like proteins are much shorter than α -helices with a length between 8 and 15 residues and a rise of $2.7 \pm 0.5 \text{ \AA}$ per residue in a membrane of size 19 - 35 \AA depending on the thickness of the membrane bilayer (Ulmschneider and Sansom, 2001; Wimley, 2002).

1.2.4 Less Frequent Secondary Structure Types

Around 10 % of helical structures are 3_{10} -helices. They contain, as the name suggests, 3 residues and 10 atoms per turn. Their radius of 1.9 \AA is smaller than the one of an α -helix whereas the average rise per residue is with 2.0 \AA higher than the rise of an α -helix (Enkhbayar, et al., 2006; Pal, et al., 2002) (see Figure 1.7). Accordingly, packing and van-der-Waals contacts between residues within 3_{10} -helices are not optimal. This explains a less frequent occurrence of 3_{10} -helical segments compared to α -helical segments (Vieira-Pires and Morais-Cabral, 2010). Hence, 3_{10} -helices are assumed to occur in important protein segments like binding sites (i.e. copper or heme) or to be involved in signal transduction (Pal and Basu, 1999). Similar to α -helices and β -strands, 3_{10} -helices are commonly annotated in protein structure files of the Protein Data Bank and they can be detected by all standard secondary structure assignment programs (DSSP (Kabsch and Sander, 1983), STRIDE (Heinig and Frishman, 2004) etc.).

A secondary structure element that has been observed only rarely in proteins (<1%), is the so-called π -helical element (Riek, et al., 2001; Weaver, 2000). π -helices contain 4.4 residues and 16 atoms per turn, resulting in a helix with a radius of 2.8 \AA and an average rise per residue of 1.1 \AA (Fodje and Al-Karadaghi, 2002; Riek, et al., 2008) (see Figure 1.8). The high diameter creates a loss of van-der-Waals contacts between amino acids within a π -helix causing the π -helical elements to be less stable and energetically unfavorable compared to regular α -helical segments (Riek and Graham, 2011). Similar to 3_{10} -helices, π -helical elements are only observed in small fragments of a helix. Therefore, π -helical structures are typically missed by the assignment of secondary structure programs and are falsely declared as α -helices (Cooley, et al., 2010). However, π -helices have been discovered to occur at functional sites within proteins (Cartailler and Luecke, 2004; Gonzalez, et al., 2012; Riek and

Graham, 2011; Riek, et al., 2001; Weaver, 2000) and a recently developed secondary structure assignment method (SST (Konagurthu, et al., 2012)) considers π -helices explicitly.

Finally, there are kinks that are defined as a single amino acid or as a stretch of amino acids that causes a change of the orientation of a helix. Typically, kinks are induced by the occurrence of a proline within an α -helical segment (Huang and Chen, 2012). Proline residues have been shown to be strong helix breakers (i.e. cytochrome c oxidoreductases) (Ulmschneider and Sansom, 2001). Although there are computational methods (i.e. TMkink (Meruelo, et al., 2011), Helanal (Langelaan, et al., 2010) etc.) to discover kinks within helices, there is still no standard definition of assigning the residue located at the center of the kink or defining a cut-off of a minimal kink that distinguishes a kink from a regular α -helical structure (Werner and Church, 2013).

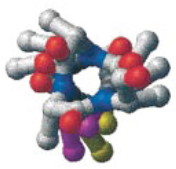
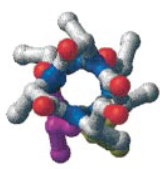
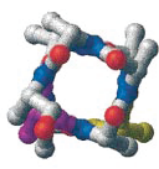
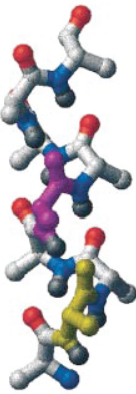
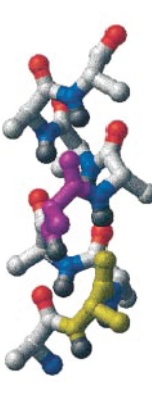
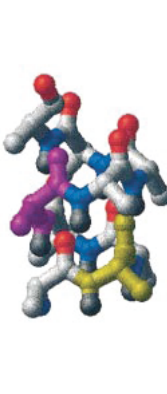
Type	3_{10} -helix	α -helix	π -helix
Residues per turn	3	3.6	4.4
Atoms per turn	10	13	16
View from top			
Helical radius	1.9 Å	2.3 Å	2.8 Å
View from side			
Rise per residue	2.0 Å	1.56 Å	1.1 Å
Hydrogen bonds	$i \rightarrow i+3$	$i \rightarrow i+4$	$i \rightarrow i+5$

Figure 1.8 Structural differences of 3_{10} , α - and π -helices. Structure figures are taken from (Riek, et al., 2001)

1.2.5 Tertiary Structures

The spatial arrangement of secondary structure elements is called tertiary structure. A general example of a tertiary structure is the β -barrel described in chapter 1.2.3. More specific is the term “fold” that describes in a generalized form a tertiary structure being common among a set of evolutionarily related proteins (e.g., FIRM for proteins having a five transmembrane-helix inverted topology repeat, LeuT-like fold). A definition of a new fold or an assignment of proteins to different folds typically requires human knowledge since a fold is a rather visual definition. Thus, accurate classifications of proteins into different fold families (e.g., SCOP (Murzin, et al., 1995), CATH (Orengo, et al., 1997) or HOMER (Forrest, et al., 2006)) are not fully automated and still rely on manual assignments but there are also less error-sensitive approaches for mapping proteins automatically to different fold families (Holm and Sander, 1996).

The use of spatial information from tertiary structures allows for a better detection of specific protein domains, bonding interactions and pathways within a protein than it is possible by applying only sequence or secondary structure information. This increase of dimensionality in information consequently also increases the complexity and difficulty of predicting the tertiary structure of a protein. Two computational principles are applied to address this issue: template-based approaches (e.g., homology modeling) and non-template-based methods (e.g., ab initio modeling).

Template-based methods require the presence of at least one evolutionarily related (homologous) protein to the protein of interest that shares a similar tertiary structure as the protein of interest. This template protein has to be detected (e.g., by a database search method) followed by a more detailed detection of homologous amino acids in the protein sequences of the template and target proteins (e.g., by an alignment method). Further details about homology and homology modeling are described the next chapter.

If a homologous template structure is not present, homology modeling is not suitable and there are several other methods that can be applied. First, there are ab initio methods referring to fragment based approaches. The three-dimensional shape of a protein is predicted by building up the structures from fragment pieces. Second, there are template-free methods using evolutionarily couplings (Hopf, et al., 2014). Correlated evolutionarily events within a protein were observed to be suitable for identifying residues that are close in space. Based on the couplings between such residue pairs, a three-dimensional model can be build up. Last, threading algorithms also do not require detailed information of the homology to be known beforehand but they need a template structure.

These threading algorithms apply statistical knowledge deduced from known information about evolutionarily and structural events (Gront, et al., 2012; Wu and Zhang, 2008).

1.2.6 Quaternary Structures

The quaternary structure of a protein describes the arrangement of three-dimensional structural elements in larger complexes. A protein can be composed of single units (monomers with a single polypeptide chain) as well as of multiple units (2: dimer, 3: trimer, etc. with multiple polypeptide chains). Multiple protein units can be stabilized with each other (e.g., via hydrogen bonds) and/or can be interacting with each other (e.g., via van der Waals forces). The similarity between multiple subunits is described by the suffixes *homo* for units that are identical and *hetero* for units that differ from each other. The interaction of these units has been shown to be crucial for binding and transport of substrates through the membrane. An example from the HOME3 data set for a quaternary structure is the homotrimeric structure of the multidrug transporter AcrB (PDB code: 4DX5 (Eicher, et al., 2012), Table A.3) consisting of three different subunits with a similar sequence that are connected with each other by loops which also stabilize the protein. Each of the subunits can be in one of three states: loose, tight and open. A substrate (e.g., acridine) can be bound to a monomer of AcrB being in the loose state, is transported through the protein in the tight state and finally released in the open state. During the conformational change of one monomer, a synchronous conformational change of the two other monomers is observed reflecting an alternating access mechanism that allows for the transport of protons and substrates through the membrane (Eicher, et al., 2014; Pos, 2009).

1.3 Sequence & Structural Homology Between Proteins

1.3.1 Definition of Homology

A structural or sequence-based similarity between proteins can arise from distinct evolutionary events. Analogous structures share a similar function or structure but evolved independently from each other in a process called “convergent evolution”. These proteins are similar but do not share a common ancestor. In contrast, two proteins are called homologs if they share a common ancestor during evolution. There are two types of homologous proteins: paralogs and orthologs. Paralogs are created by a duplication event within the genome of an organism whereas orthologs are genes in different species that descended from the same ancestral sequence present in the last common ancestor of both proteins. Single or multiple evolutionary events (insertions, deletions, mutations) within the sequence of the common ancestor cause differences in sequence and structure of

homologous proteins. Despite the large changes on the sequence level, the overall fold of homologous proteins typically stays similar during evolution as does the function (Shakhnovich, et al., 2005). A homology between two or more proteins can be detected by a comparison of the sequences and/or structures of the proteins of interest. For this purpose, each amino acid of a sequence has to be assigned to the evolutionarily corresponding amino acid of the other sequence(s) in a so-called alignment. This assignment of evolutionarily-related amino acids can be carried out on a sequence (Altschul, et al., 1990; Needleman and Wunsch, 1970), structural (Holm and Sander, 1993) or both levels (Tang, et al., 2003). Several (computational) methods have been developed for comparing protein sequences by generating an alignment and assessing the quality of the alignment being generated. The choice of the method for comparing/aligning proteins is dependent on the information that is available for the proteins of interest (Fiser, 2010).

1.3.2 Sequence-based Comparison of Proteins in General

The lowest informational content is stored in the amino acid sequence of a protein, followed by its secondary structure, tertiary and quaternary structure. Nonetheless, protein comparisons on a sequence level have been shown to be useful for detecting homology between proteins.

In the case of sequence alignments, the amino acid sequences of both proteins are represented as strings and evolutionarily related amino acids are assigned to each other based upon their similarity. Evolutionarily related amino acids that are assigned to each other are called matches if they are similar or mismatches if they are not identical (i.e. point mutations). Insertions or deletions within a sequence are represented by so-called gaps.

The main target of sequence alignment methods is an accurate assignment of evolutionarily related residues. For this, several computational methods were developed that each detects similarity between two (or more) protein sequences in a distinct way. All of these alignment approaches only require the amino acid sequence as an input for their alignment.

The first sequence alignment method for matching two protein sequences was the algorithm of Needleman and Wunsch (Needleman and Wunsch, 1970). The idea of the Needleman-Wunsch algorithm is to maximize the number of identical or similar amino acids by using a dynamic programming approach. For this, similarity between two positions is defined in a generalized similarity matrix (i.e. BLOSUM, PAM – see chapter 3.2.2.1) that has positive values for similar and negative values for non-similar amino acids. Evolutionarily insertions and deletions (gaps) are penalized by negative values. The final alignment is the one that leads to the highest similarity score. More details about alignment concepts can be found in chapter 3. Similar to this approach to align

the full sequences of proteins, there is also the Smith-Waterman algorithm (Smith and Waterman, 1981) that was developed for local alignments of protein sequences. Local alignments are also applied by the BLAST (Basic Local Alignment Search Tool) algorithm (Altschul, et al., 1990) that locates short matches between two proteins using a heuristic seed-based method. All these algorithms require a high sequence similarity between two proteins for obtaining an accurate assignment of evolutionarily related sequence positions. However, proteins have been shown to be homologs despite a low sequence identity (e.g., a low number of conserved residues during evolution) by sharing a similar tertiary structure and/or function. Consequently, more advanced computational methods were developed that include more (evolutionarily) information and/or more flexible algorithms.

First, position-specific evolutionary information was included in the process of comparing proteins. A commonly used database search method that uses such information is PSI-BLAST (Altschul, et al., 1997) that applies several iterations to obtain evolutionarily related protein sequences. The first iteration of PSI-BLAST is identical to a standard BLAST search. Based upon the highest-scoring results of this search and the underlying multiple sequence alignment, a profile in the form of a position-specific substitution matrix (PSSM) is generated. A PSSM contains for each sequence position substitution values to all other amino acids and therewith allows assigning different substitution rates to strongly conserved and highly flexible sequence positions. The obtained PSSM is then used in the next iteration instead of the substitution matrix that was used in the step before. This process is iteratively continued for a defined number of steps.

The usage of three or more evolutionarily related sequences in a multiple sequence alignment is based on a similar idea. Including evolutionarily related sequences into an alignment increases the likelihood of identifying evolutionarily conserved sequence positions, which then helps to generate an accurate alignment (Edgar, 2004; Liu, et al., 2010; Notredame, et al., 2000).

Another idea is the increase of flexibility within the sequence alignment algorithm. Position-specific gap penalties were introduced to account for the fact that insertions or deletions have a different likeliness to occur at specific protein segments. Such protein segments can be defined by properties inherent of a protein like its hydrophobicity (Thompson, et al., 2002) or by the evolutionary composition of each sequence position as implemented in alignment methods based upon Hidden Markov Models (HMMs) (Eddy, 2011). However, evolutionarily information might not be adequate to describe the relationship between two distantly related proteins that share a low sequence identity but have an overall similar fold. For the incorporation of structural elements during evolution, methods were developed that also include secondary structure information in the form of secondary

structure predictions or known three-dimensional information along with evolutionarily information for the alignment process (Remmert, et al., 2011; Yang and Honig, 2000).

Interestingly, all these methods are commonly used for alignments of membrane proteins although they were developed upon general data sets including all kinds of proteins. However, membrane proteins have a different amino acid composition and distinct evolutionary patterns within their membrane-spanning segments due to the interactions of these segments with the membrane-bilayer. This lack of membrane specific information can result in inaccurate alignments especially in the case of distantly related homologs (Forrest, et al., 2006). Consequently, I developed a membrane-specific alignment method for membrane proteins that was optimized on a reference set of membrane protein alignments, which will be described in the next chapters.

1.3.3 Sequence-based Comparison of Membrane Proteins

One of the main characteristics of membrane-spanning proteins, the annotation of their membrane-spanning segments, has been applied by two multiple sequence alignment methods (STAM (Shafrir and Guy, 2004) and PRALINETM (Pirovano, et al., 2008)). Both methods apply different evolutionary substitution rates for membranous (i.e. PHAT (Ng, et al., 2000)) and non-membranous protein segments (i.e. BLOSUM62 (Henikoff and Henikoff, 1992)) in order to improve alignment accuracy. During the alignment process, STAM first separates out the transmembrane segments and aligns them independently whereas PRALINETM keeps the sequences undivided. STAM and PRALINETM both only consider evolutionarily information in the alignment process. Additional information like membrane propensities, hydrophobicity or secondary structure probabilities are not considered. However, the application of hydrophobicity for comparing membrane protein sequences has been shown to be successful for detecting homology between membrane proteins (Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998; Lolkema and Slotboom, 2005). Unfortunately, this principle to align proteins by their hydrophobicity profiles has neither been automated nor been tested systematically on a set of membrane protein families.

1.3.4 Sequence to Structure Modeling

Generating a three-dimensional structure of a sequence for which a structure is not known is possible by applying discovered homology between a sequence without known structure (target) and a sequence with known structure (template) in a process called homology modeling. The idea of homology modeling is that evolutionarily-related amino acids of two proteins are located in a similar spatial arrangement. Therefore, accurate alignments are required to ensure a correct detection of a template structure as well as a correct assignment of evolutionarily-related residues. Based upon an

underlying sequence alignment, amino acids of the target protein are assigned to amino acids of the template protein and modeled according to this assignment using spatial constraints, energy functions or other biophysical parameters (i.e. surface-accessibility, van-der Waals interactions etc.) (Eswar, et al., 2006; Kelm, et al., 2010; Šali and Blundell, 1993). The final result of homology modeling is a three-dimensional model of the target protein for which a structure was not known so far.

Homology models can give insights into important structural features of a protein that cannot be deduced directly from a protein's sequence. The structural information of a homology model enables the prediction of binding sites, transport pathways or protein-protein interaction sites. These predictions can help to guide the design of constructs for elucidating a protein's structure (i.e. X-ray crystallography, NMR) and can offer new insights into the understanding of a protein's function (Faraldo-Gómez and Forrest, 2011; Radestock and Forrest, 2011; Schushan, et al., 2012).

A shortcoming of homology modeling is the lack of structural information for sequence fragments without an underlying template due to insertion or deletion events between target and template sequence during evolution. Such insertions or deletions typically occur less frequently in conserved segments than in variable protein segments. The modeling of loop segments is addressed by special modeling programs, which consider the characteristic amino acid composition and evolutionarily divergence of loop segments. There are two strategies for loop modeling: modeling loops *ab initio* (e.g., ModLoop that relies on the loop modeling routine in Modeller (Fiser and Sali, 2003)) or using a database search method to detect fragments that can be inserted as loops into the protein (e.g., FREAD (Choi and Deane, 2010)). Nonetheless, the accuracy of segments modeled without underlying template structure is lower than the one of segments that are modeled based on a template structure. Another factor that weakens the quality of a homology model is the similarity of the target and template sequences. A lower sequence similarity decreases the alignment accuracy so that less accurate homology models may be generated (Forrest, et al., 2006).

Several homology modeling approaches have been developed and are tested bi-annually in CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments (Kryshtafovych, et al., 2013). Recent results of CASP show that a reliable detection of homologous proteins, accurate alignments and a good score for constructing the model improve the model quality but that there are advantages and disadvantages for all methods on all steps of the homology modeling process. Consequently, an overall perfect homology modeling strategy that performs best for all types of proteins does not exist (Kryshtafovych, et al., 2013).

1.3.5 Structure-based Comparisons of Proteins

The availability of two homologous protein structures enables a comparison on a structural level by structural alignment programs. Recent studies have evaluated methodologies and accuracies of a diverse set of structural alignment methods (Berkalk, et al., 2009; Kolodny, et al., 2005; Sadowski and Taylor, 2012; Slater, et al., 2012).

All structural alignment methods tested are based on the assignment of evolutionarily-related amino acids of homologous proteins that are in a similar secondary structure type and orientation in three-dimensional space. Consequently, spatial information of protein structures or fragments is used in combination with a distance measure describing the similarity of those proteins or protein fragments in order to superimpose their three-dimensional protein structures. Similar to sequence alignment methods, the output consists of a set of aligned amino acids as well as a similarity measure of the two proteins, but this similarity measure is structure- and not sequence-based. An additional output of structural alignment programs is a structure superimposition that can be used for a visualization of the alignment.

Challenges for all structural alignment programs are the protein's structural flexibility and dynamics. A protein solved in a given state is represented by a single conformation (rigid) for an alignment although the protein is dynamic and has a tendency to adopt a set of different conformations (flexible). Homologous protein structures that are solved in different states are harder to align than those that are crystallized in exactly the same state (Menke, et al., 2008; Ye and Godzik, 2003). In general, homologous proteins are not solved exactly in the same conformation or even not in the same state. Consequently, users of structural alignment programs have to be aware of the conformation and structural state of the proteins to be aligned. And again, alignment accuracy is dependent on the similarity of the protein structures used, with an increased accuracy for proteins with a higher similarity.

Consequently, all structural alignment programs have their advantages and disadvantages according to all evaluation studies of structural alignment programs mentioned above. There is no outstanding program that generates accurate structural alignments for all types of proteins.

1.3.6 Rating Structural Accuracy and Similarity of Protein Models

The accuracy of a structural protein model can be measured by either internal scores rating the energy of a protein structure or by comparing the model to its original X-ray structure.

First, there are scores that rate the energy of a protein structure (e.g., DOPE score (Discrete Optimized Protein Energy score) from Modeller (Eswar, et al., 2006)). Models that are assigned to have a low energy score are more likely to be correct than those that have received a high energy score. Unfortunately, such scoring functions were optimized on general data sets that contain predominantly soluble proteins. This results in mean force potentials that are not suitable for membrane proteins (Heim and Li, 2012; Ray, et al., 2012). There are also some membrane-protein potentials but they are either not precise (e.g., ProQM (Ray, et al., 2010)) or not user-friendly (Rosetta-Membrane (Yarov-Yarovoy, et al., 2006)).

Another approach for comparing structures of proteins (models) with each other is implemented by distance-based geometric scores. In the CASP experiments, homology models are compared to the original X-ray structures of their corresponding proteins using a variety of structural similarity scores (Kryshtafovych, et al., 2013). Global superposition scores (i.e. RMSD, GDT_TS, AL4, TM-score, see chapter 5.2.2) compare the arrangement of the backbone atoms of a homology model and X-ray structure with each other. Model structures are assigned to be more accurate, the more similar the arrangement of their backbone atoms is to those of the original X-ray structures. In contrast, superposition-free model scores do not directly calculate spatial distances between the two proteins of interest. An example of such a score is the CAD-score (Olechnovic, et al., 2012) that compares differences in residue-residue contact areas of two protein structures.

Structural as well as sequential similarity annotations can be used to define similarity between proteins for generating protein data sets. Besides providing an overview of the relationship between protein sequences, such a protein data set of homologous protein sequences can also be applied for assessing the quality of a specific method or score. Moreover, protein data sets are commonly used as a reference for optimizing a method or a score on these data sets. Different types of protein data sets are described in the next chapter.

1.4 Databases for Proteins and Membrane Proteins

Much information about proteins (amino acid sequence, secondary and tertiary structure) as well as the relationship between different classes of proteins is stored in a variety of databases; each of them addressing a specific need.

First of all, there are databases that are collections of proteins without explicit information about the relationship between those proteins. The most comprehensive databases are those containing protein sequences with and without known structures (e.g., UniProtKB (UniProt, 2013) with a size of 547599 protein sequences (status: 4-Feb-2015), see web.expasy.org/docs/reNotes/relstat.html). The UniRef100 (UniProt Reference Clusters) database merges identical sequences from UniProtKB into single UniRef entries (Suzek, et al., 2007) and is therefore non-redundant. UniRef100 sequences are also clustered at a 90% or 50% level using CD-HIT (Li and Godzik, 2006) to generate the corresponding UniRef90 and UniRef50 databases. Aside from these databases, which contain sequence information only, structural information about proteins that have been solved by X-ray crystallography, NMR or EM is stored in the Protein Data Bank (PDB) (Berman, et al., 2003). Derivations of the Protein Data Bank are the OPM (Orientations of Proteins in Membranes) (Lomize, et al., 2006) and PDB_TM (Protein Data Bank of transmembrane proteins)(Kozma, et al., 2013; Tusnady, et al., 2005) data banks that both contain only membrane proteins that were extracted from the PDB. Both databases contain similar proteins but those proteins sometimes have slightly different membrane annotations caused by the distinct computational approaches that were used for generating these database. For the OPM database, a protein is handled as a rigid body that floats in a hydrophobic membrane bilayer and an energy function that describes the spatial arrangement of the protein structure in the lipid bilayer is minimized for this protein. This energy function considers the accessible surface area of each atom, solvation parameters for each atom and an interfacial water concentration profile (Lomize, et al., 2006). For the PDB_TM database, the detection of membrane-spanning segments is divided in two steps. In the first step, the BIOMOLECULE record of the PDB file is analyzed in order to detect internal symmetry between protein subunits. Symmetry in a protein can help to guide the detection of the membrane axis because the rotational axis might be parallel to the membrane bilayer normal. In a second step, an objective function is applied to detect membrane-spanning protein fragments by considering three protein properties: water-accessibility, hydrophilicity and a structure factor, which considers the straightness of a secondary structure, turns and chain ends (Tusnady, et al., 2004). More details about the PDB_TM database can be found in chapter 2.2.

Next, there are databases that have classified proteins into families sharing a similar fold or function. PROSITE (Sigrist, et al., 2010; Sigrist, et al., 2013) is a database of protein domains, families and functional sites. Each protein family or domain is represented by associated patterns and/or profiles. Patterns are manually defined observations that describe a protein family whereas profiles are statistical descriptors of a multiple sequence alignment of the family similar to a Hidden Markov Model. These profiles contain position-specific scores for amino acids and position-specific gap penalties (for opening and extending a gap). Thus, each family is represented by a multiple sequence alignment as well as by a Hidden Markov Model. The SCOP (Structural Classifications of Proteins) database (Murzin, et al., 1995) contains structural classifications of protein structures in general based upon structural superimpositions of those proteins, similar to the HOMSTRAD (Mizuguchi, et al., 1998; Stebbings and Mizuguchi, 2004), FSSP (Holm and Sander, 1996) (Holm and Sander, 1996) and CATH (Class, Architecture, Topology, Homology) (Orengo, et al., 1997) databases. More specialized is the data set of “membrane proteins of known 3D structure” that only contains membrane protein structures that are classified into specific functional families (<http://blanco.biomol.uci.edu/mpstruc/>).

Moreover, some databases also contain direct information about the relationship between two or more proteins described in the form of sequence alignments of each family. In Pfam (Punta, et al., 2012), for example, protein families are represented by motifs in form of multiple sequence alignments and Hidden Markov Models. BALIBASE (Bahr, et al., 2001; Thompson, et al., 2005; Thompson, et al., 1999) is a manually refined database of multiple sequence alignments for a diverse set of protein families (including membrane proteins, repeats, circular permutations etc.). Pairwise information in the form of pairwise alignments and homology models based upon those alignments is stored in ModBase and HOMEPEP. ModBase is a database of annotated protein structure models derived from an automated modeling pipeline that relies on PSI-BLAST and Modeller (Pieper, et al., 2014). More specific is the manually created HOMEPEP data set with a set of homologous membrane proteins, pairwise alignments and corresponding homology models (Forrest, et al., 2006).

Finally, there are specialized databases for more specific requests like searching for protein motifs (i.e. PRINTS (Attwood, et al., 2003; Scordis, et al., 1999), protein-protein interactions (i.e. STRING (von Mering, et al., 2005)) and many more protein-specific properties or relationships among proteins.

1.5 Outline of this Work (Research Question, Problem Statement)

Computational approaches for proteins (i.e. alignments, databases) were typically developed, optimized and evaluated using general protein data sets, although there is the distinct class of membrane proteins. This dissertation addresses this issue by studying and evaluating computational approaches and ideas to understand homology and evolutionarily events among membrane proteins on a sequence and structural level. The outline of this work is as follows:

First, a reliable and recent data set of membrane proteins is required to be able to understand membrane proteins. Thus in chapter 2, I updated a an earlier data set of homologous membrane proteins that I used in my diploma thesis (Stamm, 2010) to more recent versions in 2010 (HOMEP2) and 2013 (HOMEP3) using automated clustering approaches.

In a next step (chapter 3), the challenge of aligning α -helical membrane proteins accurately is addressed. A novel sequence alignment package for membrane proteins (AlignMe), which I developed together with colleagues and reported on in my diploma thesis (Stamm, 2010), has been developed further to allow for more input options (e.g., position-specific substitution matrices or membrane prediction propensities) and was optimized using these new inputs on the HOMEP2 data set. The applicability and accuracy of different modes of AlignMe were then tested and compared to other recent sequence alignment methods using a sequence-based comparison to reference alignments of the HOMEP2 data set. This study includes more alignment methods (e.g., MSAProbs or ProbCons) than my diploma thesis. In addition, alignment accuracy was also assessed using homology modeling on HOMEP2 and a sequence-based evaluation on an independent benchmark set of membrane protein alignments (BALIBASE reference 7), which has not been done previously. This work was published in March 2013 in PLOS ONE (Stamm, et al., 2013).

The implementation of a webserver for AlignMe is described in chapter 4. The alignment accuracy of AlignMe is compared to those of other recent webserver for aligning protein sequences. Additionally, the alignment of family-averaged hydropathy profiles using AlignMe (Khafizov, et al., 2010) and the corresponding implementation into the AlignMe web server is described. This work was published in the NAR web server issue in July 2014 (Stamm, et al., 2014).

Chapter 5 contains an assessment of the accuracy of structural alignment methods for membrane proteins using the HOMEP3 data set that includes α -helical and β -barrel-like proteins. As in chapter 3, homology modeling is used as an assessment criterion for alignment accuracy and additionally, a deeper insight into structural similarity scores (i.e. superposition-dependent vs. superposition free

scores) is provided. This work was published in the journal "PROTEINS: structure, function and bioinformatics " in August 2015 (Stamm and Forrest, 2015).

Evolutionarily events within transmembrane segments of membrane proteins are examined in chapter 6. Four different structural alignment methods that were shown to generate accurate alignments (see chapter 5.3.8) were used to generate a consensus alignment for each protein pair within HOME3. These consensus alignments included confidence values for each alignment position and enabled the identification of reliable single insertions and deletions (InDels) within several membrane protein families. Detected InDels were examined for changes within secondary structure elements and their effect on protein-specific functions such as ligand specificity.

Finally, all results are discussed and the key findings, as well as the impact of this dissertation on membrane protein research, are summarized in chapter 7. Possible improvements of the current work are also discussed. Additionally, an outlook for future perspectives on the topic of membrane protein alignment is provided.

2 Automated Generation of Homologous Membrane Protein Data Sets (HOME2 & HOME3)

2.1 Introduction

Protein data sets are collections of membrane protein sequences, structures or other biological information about proteins. The inherent information that is stored in a protein data set is dependent on the topic that a data set addresses like information retrieval or the optimization or evaluation of computational methods (i.e. database searches, clustering methods etc.). Accordingly, there are general protein data sets as well as data sets that are composed only of membrane protein structures that will be discussed in the following chapter. Similar to the general data sets, which are mentioned in chapter 1.4, different types of membrane protein data sets exist (see Table 2.1 for an overview of membrane protein databases), which are described here in more detail:

First, there are membrane protein data sets that are collections of membrane proteins without an evolutionarily annotation between the proteins. Annotations of membrane-spanning segments (i.e. number of TM segments, location of TM segments etc.) are stored in several databases: CGDB, OPM and PDB_TM. For CGDB (coarse-grained model database), coarse-grained simulations of proteins, lipids and water were applied to assemble the lipids around the protein (Sansom, et al., 2008) whereas for OPM (Orientations of Proteins in Membranes) (Lomize, et al., 2006) an implicit solvent model of the lipid bilayer is used (Lomize, et al., 2006; Lomize, et al., 2007) and for PDB_TM (Protein Data Bank of transmembrane proteins) (Kozma, et al., 2013; Tusnády, et al., 2004; Tusnády, et al., 2005) an objective function that considers solvent accessibility, hydrophobicity and structural features is applied for inserting a membrane protein into a membrane (Tusnady, et al., 2005). In contrast to CGDB, which is a fixed set of proteins, OPM and PDB_TM are updated on a regular basis. Besides structural information, also functional and experimental information can be retrieved from the Membrane Protein Data Bank (MPDB) (Raman, et al., 2006). Record entries of the MPDB contain structural, functional and experimental information of integral, anchored and peripheral membrane proteins and peptides that have been revealed by X-ray, NMR or EM experiments. Direct information of the relation between the proteins is not provided by these data banks.

Another type of membrane protein databases includes evolutionarily information between the proteins they consist of. The Transporter Classification Database (TCDB) contains sequence, structural, functional and evolutionarily information about (putative) transporter proteins (Saier, et al., 2006). Similarly, TransportDB is composed of cytoplasmic membrane transporters and outer

membrane channels (Ren, et al., 2007). A more specific database is the GPCRDB (Horn, et al., 2003) that stores large amounts of heterogeneous data (e.g., sequence information, binding constants, homology models etc.) about G-Protein coupled receptors. Also for β -barrel-like proteins, there are specific databases like OMPdb (Tsirigos, et al., 2011) that contains classifications of integral β -barrel outer membrane proteins from gram-negative bacteria into distinct families based upon functional and structural criteria. However, all these databases contain only a subset of all membrane protein classes, hence they are too specific for a general analysis on membrane proteins.

Other data sets are specialized on certain organisms like the ARAMEMNON database that provides membrane protein data of nine plant species (Schwacke, et al., 2003). A more general data set for all types of membrane proteins is the set of “membrane proteins of known 3D structure” in which all types of membrane protein structures that have been solved are classified to specific functional families (<http://blanco.biomol.uci.edu/mpstruc/>). More specialized by considering only NMR structures is the set of “membrane proteins of known structure determined by NMR” (<http://www.drорlist.com/nmr/MPNMR.html>).

More explicit information about the evolutionarily relationship between proteins is contained in data sets that provide alignments of their proteins. The HOMEPEP (HOMologous MEMbrane Protein) data set from 2006 (Forrest, et al., 2006) contains 36 proteins in 11 families including homology models and underlying alignments of different quality (good models vs. decoy models) for each pair of proteins within a family. This data set has been used to test scoring functions for membrane protein structures (Heim and Li, 2012) or to test the quality of database searches (Bernsel, et al., 2008). Another data set of membrane proteins that addresses the alignment accuracy of membrane proteins is the Reference 7 set of BALiBASE (Bahr, et al., 2001) generated in 2001, which was shown to be adequate for evaluating the alignment quality of a membrane protein alignment method (Chang, et al., 2012). This set contains 435 membrane proteins in 8 superfamilies, namely 7tm, acr, photo, dtd, ion, msl, Nat and ptga, each superfamily aligned in a multiple sequence alignment.

A shortcoming of both data sets (HOMEPEP and BALiBASE reference 7) is their age, meaning a lack of recent sequence and structure information of newly solved membrane proteins. For this reason, new versions of HOMEPEP were created that will be described in this chapter. The HOMEPEP2 data set of 2010 was used for the optimization of AlignMe (see chapter 3). In 2013, an updated version of HOMEPEP2 called HOMEPEP3 was applied for assessing the accuracy of structural alignment methods (see chapter 5) and for locating single evolutionarily deletions and insertions (see chapter 6).

Table 2.1 - Overview of Membrane Protein Databases

Type of database	Name	URL
Databases with annotations on membrane-spanning segments	PDB_TM	http://pdbtm.enzim.hu/
	OPM	http://opm.phar.umich.edu/
	CGDB	http://sbc.bioch.ox.ac.uk/cgdb/
Structural, functional and experimental data of membrane proteins	MPDB	http://www.mpdb.tcd.ie/
Membrane protein family-specific databases with evolutionarily information	TCDB	http://www.tcdb.org/
	TransportDB	http://www.membranetransport.org/
	GPCRDB	http://www.gpcr.org/7tm/
	OMPdb	http://www.ompdb.org/
Membrane protein data of nine plant species	ARAMEMNON	http://aramemnon.uni-koeln.de/
Pairwise alignments and homology models of membrane proteins	HOMEp, HOMEp2, HOMEp3	http://www.forrestlab.org/software_databases/
Multiple alignments of membrane proteins	BALiBASE reference 7	http://www.lbgi.fr/balibase/

2.2 PDB_TM Database as a Starting Set for Clustering Membrane Proteins

A manual assignment of proteins to specific membrane protein families has become too time-consuming with the increase in unique membrane proteins structures that have been solved (see Figure 2.1). Consequently, an automated clustering method is needed to regenerate an updated set of membrane protein families. Additionally, a computational approach might discover protein relationships that are not detectable by a manual annotation.

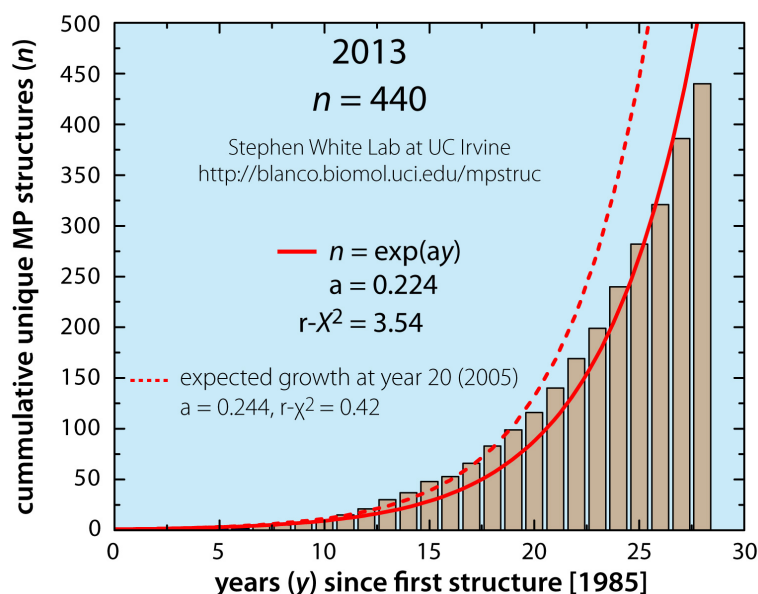


Figure 2.1 Statistics about unique membrane protein structures that have been solved according to the database “Membrane Proteins of known 3D structure”, <http://blanco.biomol.uci.edu/mpstruc/>

The Protein Data Bank of Transmembrane Proteins (PDB_TM) (Kozma, et al., 2013; Tusnady, et al., 2005) is a good starting set for clustering membrane proteins to families of homologous sequences. This database was created with the TMDDET algorithm (Tusnady, et al., 2004) scanning all PDB entries in regular time intervals resulting in a steady increase of the data bank (see Figure 2.2). In a first step, the TMDDET algorithm filters the PDB data bank by excluding virus and pilus proteins as well as nucleotide sequences. For all other sequences, the biological oligomer is then built by using the BIOMOLECULE record stored within the PDB File. Non-biological contacts, which result from the crystallization process, are removed. This generation of an oligomer helps to identify symmetry within the protein that facilitates the search of a membrane location based upon an objective function. Three protein properties are examined and combined into this objective function. First, the water-accessible surface area in the structure is calculated for each amino acid within the protein sequence. Second, all amino acids are classified into hydrophilic (A, C, D, E, H, K, N, P, Q, R, S) and hydrophobic (F, G, I, L, M, V, W, Y) residues, which was shown to be as adequate as using hydrophobicity scales. Last, a structure factor is incorporated by analyzing the three-dimensional

fragments of the protein for their likeliness to contain turn or ends of chains as well as for their straightness. Membrane-spanning segments are assumed to contain hydrophobic residues that are exposed to the solvent (in case of membrane proteins, this is the membrane) and are assumed to form a regular structure.

For proteins that are detected as membrane proteins, the number of membrane segments is annotated as well as a residue-specific annotation of whether amino acids are inside or outside of the membrane. All information observed is stored in an XML-file that can be easily accessed and evaluated by custom computational scripts. For this study, I developed several scripts to process data from the PDB_TM data bank using C++, Perl and Bash.

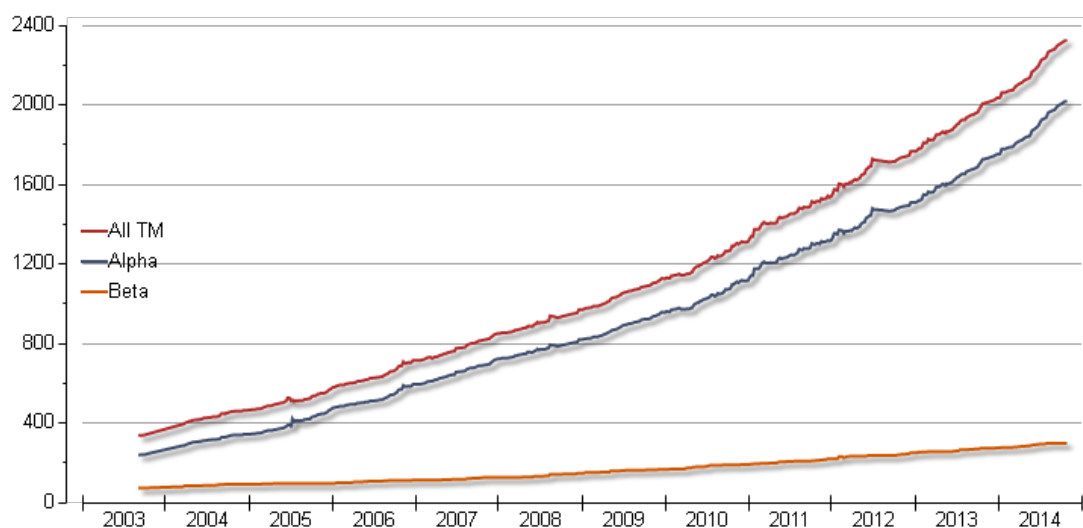


Figure 2.2. Total number of membrane proteins (including single-spans) that are in the PDB_TM database as a function of time. Statistics taken from http://pdbtm.enzim.hu/?_=/statistics/growth

2.3 Common Principles of Clustering Membrane Proteins to a Set of Homologous Proteins for HOME2 and HOME3

The HOME2 and HOME3 data sets are both based upon the membrane proteins that are listed in the PDB_TM database. Several principles that are applied for clustering membrane proteins are similar for the HOME2 and the HOME3 data sets. For both HOME data sets, all protein structure coordinate files listed in the PDB_TM database were downloaded from the Protein Data Bank and filtered in order to include only high-resolution structures with a resolution <3.5 Å. NMR structures, theoretical models or structures with a resolution >3.5 Å were therefore discarded. The extracted proteins were then separated according to their overall fold: α -helical or β -barrel-like; and each protein structure was split into its individual chains. For all pairs of proteins sharing the same number of membrane segments pairwise structural alignments were generated using a structural alignment method. The resulting structural alignments as well as the underlying sequence alignment of each protein pair were subsequently analyzed for their similarity using specific similarity scores. Threshold values were then used for clustering proteins with adequate similarity scores to a common family. In case of identical proteins or protein chains, the protein (chain) with the better resolution or if the resolution was equal, the one with the smaller R-factor was chosen to be the representative for that type of protein (chain) within a family.

2.4 Initial Generation of HOME2

In the case of HOME2, structural alignments for each protein pair were generated using SKA (Petrey and Honig, 2003; Yang and Honig, 2000) and their similarity was rated using a structural similarity score (PSD) as well as two sequence similarity scores (one based upon structure alignment, the other on a pure sequence-based alignment).

Typically, the RMSD score (Kabsch, 1976) is used to rate the similarity between two protein structures. For each amino acid (i), the structural differences between corresponding C_{α} -atoms (v_i and w_i) of the two superimposed structures of length L_{aln} is calculated:

$$RMSD(v, w) = \sqrt{\frac{1}{L_{aln}} \sum_{i=1}^{L_{aln}} (v_i - w_i)^2} \quad (2.1)$$

The squaring term makes the RMSD score very sensitive to large local deviations and differences in the lengths of the template (Moult, et al., 1997; Moult, et al., 1995). Therefore, the PSD (Protein Structural Distance) score calculated by SKA was used as a measure of the structural similarity

between two superimposed structures instead. In contrast to the RMSD score, the PSD score is less sensitive to large outliers because it averages them down. The PSD score is based upon the number of secondary structure elements a for Protein A and b Protein B , the alignment score of the secondary structure elements $s(A,A)$ for the self-alignment of Protein A , and $s(A,B)$ for the alignment of Protein A with Protein B , an RMSD term and two adjustable parameters (x,y) for the sensitivity and accuracy of the PSD score. The parameters x and y were set to the 3 Å and 5 Å respectively in the SKA program because these values were shown to result in PSD scores that can be used to classify proteins from the SCOP database accurately with a low number of positives, especially for low PSD thresholds (Yang and Honig, 2000).

$$PSD(A, B) = \left(\frac{\log \left[\left(\frac{a}{\max(a, b)} \right) \left(\frac{s(A, B)}{s(A, A)} \right) \right]}{\log x} \right)^2 + \left(\frac{RMSD}{y} \right)^2 \quad (2.2)$$

The PSD score approaches zero for two identical proteins and increases for more distant protein structures. Two protein structures are assumed to be close homologs if their PSD score is below 0.4 and distant homologs if their PSD score is below 1.2.

For the sequence similarity, two similarity scores were calculated based either on a structural alignment or on a sequence alignment. In general, structural alignments are more accurate than sequence alignments but structural alignments might be less accurate in cases of proteins solved in different states. In such cases, structural alignment programs might align residues that are close in space but are evolutionarily not related. Sequence only methods do not face this issue. Accordingly, the sequence similarity of two proteins is measured here based on a structural and on a sequence alignment. First, the sequence similarity of a protein is calculated based on the sequence alignment that is also generated by SKA alongside the generation of the structural alignment. This sequence alignment is based on the spatial assignment of amino acids from the structural alignment. In a second approach, the sequence similarity is calculated based upon a pairwise protein sequence alignment without spatial information. Therefore, a simple Needleman-Wunsch algorithm and a BLOSUM62 substitution matrix are applied for generating the sequence alignment. Two protein structures are assumed to be identical if they had a sequence identity above 85 % according to the Needleman-Wunsch alignment and above 95 % according to an alignment obtained with SKA.

Using these three scores, a protein and two other proteins were clustered to a common group in a first step if they were close or distant structural homologs (i.e. PSD score <1.2 for close (i.e. within the same SCOP superfamily) or <0.4 for distant homologs) without being identical (i.e. sequence

similarity <95% based on a structural alignment). A group was then assigned as a family if all proteins within that group fulfilled the clustering criteria to all other proteins of that group (see blue dots in Figure 2.3). Otherwise, a second clustering step was applied to filter out incorrectly clustered proteins. Therefore, all close homologs (i.e. PSD value below 0.4) were clustered to a common family (e.g., the same SCOP family, red and green dots in Figure 2.3). Subsequently, distant homologs were added if they were homologous (PSD score below 1.2) to all other proteins of that group (see yellow dots in Figure 2.3). In the final step, pairwise alignments of all sequences in each family generated using AlignMe (using BLOSUM62, and standard gap opening and extension penalties of 10 and 1, respectively) were used to identify redundant sequences (>85% sequence identity): for two chains of the same protein, only the one with the longest sequence, or the lower total B-factor was retained; for two chains of different proteins, only the one with the higher resolution, or with the smaller R-factor, was retained.

This clustering strategy resulted in a data set of 81 proteins within 22 families and 177 alignments. This data set is called HOME2 and captures a variety of protein folds and diversity among proteins (i.e. transporters, channels, signal transduction proteins etc.). Additionally, HOME2 captures a large range of protein chain sizes from 2 to 12 membrane-spanning helices with the smallest protein chain sequence containing 101 and the longest sequence containing 535 amino acids. A detailed overview of HOME2 is shown in Table A.1 and Table A.2. Due to this amount and diversity of proteins, HOME2 was used as a representative membrane protein data set for optimizing the gap penalties of a novel alignment program called AlignMe and for assessing the different alignment input descriptors that were used to generate the alignments, see chapter 3.

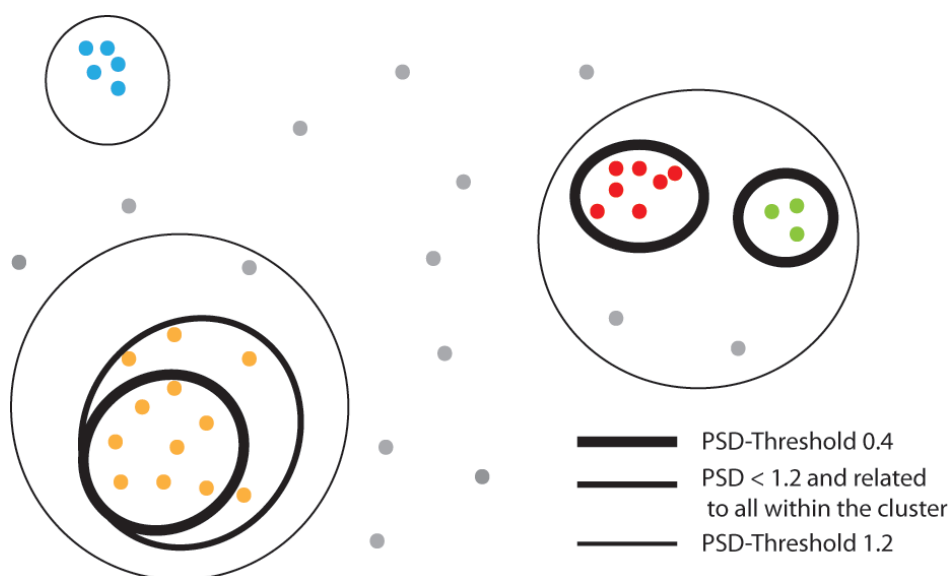


Figure 2.3 Clustering principle used for generation of the HOME2 data set. Proteins are represented by dots. Dots with the same color are assigned to the same family. Grey dots do not belong to any family. The detailed clustering procedure is described in chapter 2.4.

2.4.1 Modifications of HOME2

After an initial optimization of the gap-penalties for AlignMe using substitution matrices (see chapter 3 for more details), I found some sequence-based alignments of HOME2 to be highly different from the reference alignments of HOME2. Even alignments of proteins that were found to be close homologs were concerned. Thus, I had a closer look at these alignments to find the cause of this inaccuracy.

Since the structural alignment method of SKA applies only spatial information, some positions can be aligned due to their close spatial distance although they are obviously evolutionarily not related. Some alignments contained a large internal gap flanked by one or two terminal residues instead of a large N- or C-terminal gap due to the misleading alignment of SKA. Those terminal residues were manually moved to the other end of the gap, resulting instead in a long terminal gap.

Additionally, a large cytoplasmic domain inserted in the human β -2-adrenergic receptor (PDB code: 2RH1) for the purposes of enhancing crystallization was removed from the sequence of the corresponding protein chain because this insertion domain does not belong to the original sequence of human β -2-adrenergic receptor and caused an unusual large internal gap in alignments of the human β -2-adrenergic receptor to other G-protein coupled receptors. Such large internal gaps were not representative for the alignments in HOME2.

2.5 HOME3

More recently, a steady increase of available structural information on membrane proteins required an update of the HOME2 data set, which was generated in 2010. The clustering strategy was changed to address two issues that had subsequently been detected in HOME2:

First of all, inaccuracies in alignments generated with SKA were observed for the family of G-protein coupled receptors (see chapter 3). SKA had difficulties aligning homologous GPCRs that were crystallized in different states but the alignments of SKA did not contain a measure of their accuracy for specific pairs of amino acids. Consequently, all positions in the alignment of SKA were treated as being correct and accurate although they were incorrect. This issue might lead to errors in clustering as well as to errors in optimizing or evaluating programs on HOME. A solution to this problem is the application of another structural alignment method. TM-align was shown to produce accurate structural alignments (Sadowski and Taylor, 2012) including the reliable TM-score (Template Modeling score (Zhang and Skolnick, 2004)) that can be applied to clustering of proteins (Dai and Zhou, 2011).

Similar to the PSD score, the TM-score accounts for large outliers by using a distance-dependent weighting scheme in which the contribution from largely deviating residue pairs is reduced.

$$\text{TM-score} = \max \left[\frac{1}{L_{\text{target}}} \sum_i^{L_{\text{aln}}} \frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} \right] \quad (2.3)$$

With L_{target} being the number of residues in the target and d_i being the distance between the i^{th} pair of residues. Additionally, the TM-score includes the term d_0 for normalizing the match difference so that the distance-downweighting varies with the protein size, resulting in a score that is less size-dependent than scores that square differences (e.g., RMSD):

$$d_0 = 1.24 \cdot \sqrt[3]{L_{\text{target}} - 15} - 1.8 \quad (2.4)$$

The TM-score ranges between 0 (worst case) and 1 (perfect match). Protein structures were shown to share a similar fold if they share a TM-score (Template Modeling score) above 0.5 (Xu and Zhang, 2010). Hence, TM-align was applied alongside SKA with its PSD score (see chapter 2.4) for generating structural alignments of all protein pairs that were extracted from the PDB_TM database and that share the same number of TM segments (see chapter 2.3).

Another drawback of HOMEP2 is that the clustering principle generates only families with at least three homologous proteins and misses protein families that consist of only two proteins. A change in the clustering strategy for HOMEP3 allows for the detection of such small families and thereby also the inclusion of more diverse biological information to the data set. Two different hierarchical clustering steps were applied. First, similar to HOMEP2, all proteins that were assumed to be homologs were clustered in an agglomerative step into a group and a group was then assigned as a family if all proteins within that group fulfilled the clustering criteria to all other proteins of that group. Instead of using another agglomerative clustering step with a higher identity cut-off at this stage as for HOMEP2, a divisive hierarchical clustering was applied that allows for families containing only two proteins. Specifically, in cases of families with proteins that were not related to all other proteins of that family, the protein that fulfilled the clustering criteria to the least number of other proteins was removed from that group. In cases, two or more proteins fulfilled the clustering criteria to the least number of other proteins, the protein that had on average the most inaccurate structural similarity score compared to all other proteins of the family was removed from that group. This procedure was repeated until all proteins within a group fulfilled the homology criteria with each

other, and these were then assigned to a new family. All proteins that were removed from that group were then assigned to a new group and checked for their homology criteria as mentioned before. This hierarchical clustering principle was repeated until either no more proteins were excluded from a group or all excluded proteins shared no homology criteria with each other.

Two data sets were clustered, one set based upon structural alignments of SKA including the PSD score and one set using structural alignments by TM-align and its TM-score. These two sets were compared with each other for their similarity. In cases of families containing different proteins, a manual inspection was necessary to inspect the cause of the difference. This process ensures that the clustering method is not biased by a certain structural alignment method or structural similarity score and is moreover robust against errors inherent to structural alignment programs. For testing the accuracy of structural alignment methods (see chapter 5) HOME3 was used as a set of clustered proteins without any reference alignments. Using the four most accurate structural alignment methods out of this analysis, reference alignments with confidence values were then generated for the HOME3 data set (see chapter 6).

The new clustering principle resulted in a data set called HOME3 containing 152 α -helical proteins in 40 families with 354 alignments and 68 β -barrel-like proteins in 8 families with 319 alignments (see Figure 2.4). Aside from the new set of β -barrel-like proteins, HOME3 contains 15 new families of α -helical proteins and 78 proteins that are either added newly to the data set or that were solved at a higher resolution. More details about the HOME3 data set are available in Table A.3 and Table A.4.

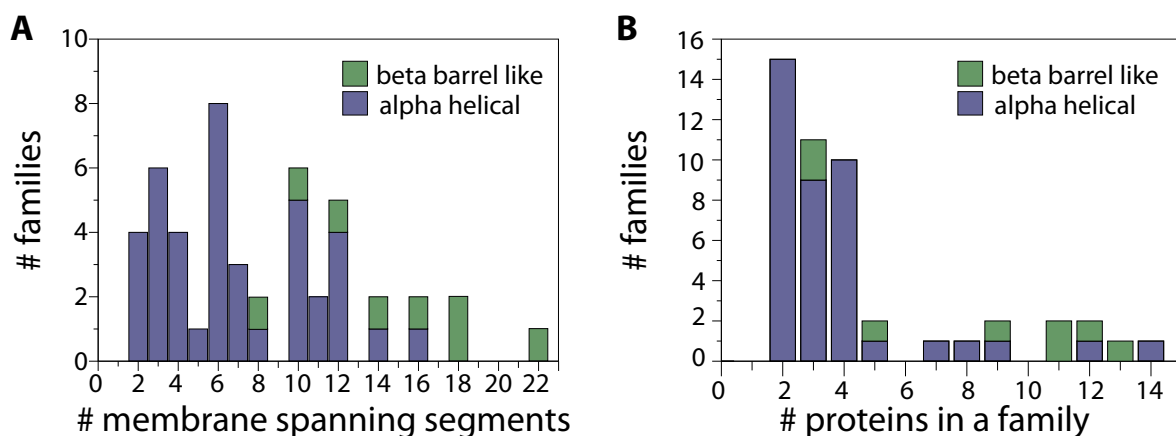


Figure 2.4 Composition of the HOME3 data set of homologous membrane protein structures. A. Distribution of family size for a given number of membrane-spanning segments for α -helical (blue) and β -barrel-like proteins (green). B. Distribution of families with different numbers of proteins. Most families of α -helical proteins (blue) contain 2-4 known protein structures, whereas the β -barrel-like families (green) contain more proteins per family.

3 AlignMe - an Optimized Program for Aligning Helical Membrane Protein Sequences

3.1 Introduction

Sequence alignment methods allow for the identification of evolutionarily related amino acids in homologous protein sequences (see chapter 1.3.2). The application of a simple generalized substitution matrix using the Needleman-Wunsch algorithm for generating a pairwise sequence alignment results in accurate results for homologous protein sequences that share a high sequence identity (>60% identical residues). However, the alignment accuracy decreases with decreasing pairwise sequence similarity since evolutionarily information is not sufficient to describe relationships of distantly related proteins (e.g., with a sequence identity <25 %) (Forrest, et al., 2006; Kryshatovych, et al., 2005; Tress, et al., 2005). Consequently, more advanced alignment methods that include additional information about the protein and/or are based on a more sophisticated alignment algorithm were developed for the alignment of low-similarity protein sequences.

The inclusion of the evolutionarily information contained in a set of homologous sequences is implemented in various multiple protein sequence alignment (MSA) methods like T-Coffee (Notredame, et al., 2000), MUSCLE (Edgar, 2004), ProbCons (Do, et al., 2005) and MSAProbs (Liu, et al., 2010) and has been shown to be successful in generating accurate alignments even for protein pairs within the twilight zone (e.g., sequence identity of 25-30 %). Two of these multiple sequence methods, T-Coffee and MUSCLE, were tested for their accuracy on the original HOMEPEP data set and were found to produce relatively accurate alignments for membrane protein sequence pairs (Forrest, et al., 2006).

A more accurate alignment method for protein sequence pairs of HOMEPEP is the profile-to-profile alignment program HMAP (Tang, et al., 2003) that creates profiles for each sequence before aligning them. These profiles include evolutionarily information in the form of substitution matrices, homologous sequences, structural propensities and structural relationships. Similarly, PSI-Coffee (Chang, et al., 2012) uses profiles of multiple sequences for generating a multiple sequence alignment. The alignment of profiles is also provided by the BCL::Align program (Dong, et al., 2008). A weighted scoring function in BCL::Align combines several protein properties like general and position-specific scoring matrices, secondary structure predictions and a variety of chemical properties. This scoring function of BCL::Align was optimized on a general data set and did not consider membrane proteins explicitly although a term for hydrophobicity is included. Complex

information about a protein's properties can also be represented in the form of a Hidden Markov Model (HMM). HHalign (Söding, 2005) is one of the most readily-available HMM-alignment methods used to align HMMs that were generated by a database search method. In addition to evolutionarily information, this approach also provides the option to include a secondary structure prediction from PSIPRED. With this combination of protein information, HHalign has been shown to generate accurate alignments based upon which accurate models of water-soluble proteins can be generated. This approach has been implemented in the HHpred structure prediction protocol (Hildebrand, et al., 2009).

None of the above mentioned methods include membrane-specific information during the alignment process, although this environment was shown to be a crucial determinant of the amino acid composition of membrane proteins (see chapter 1.2.2). The hydrophobic character of α -helical membrane segments inspired the early use of hydropathy profiles for locating TM segments in a protein sequence (Kyte and Doolittle, 1982). In these hydropathy profiles, a value describing the hydrophobicity according to a particular scale is assigned to each sequence position. A variety of such scales have been developed using biochemical, biophysical or theoretical considerations (e.g., the HWvH (Hessa, et al., 2005) or UHS (Koehler, et al., 2009) scales, see chapter 3.2.2.3 for more details). Depending on the scale that is used, slight differences occur in the profile for a given protein, but all scales allow for the visual identification of hydrophobic membrane-spanning regions in a hydropathy plot. Typically, window averaging is used to smooth out details of these profiles and to simplify a visual inspection of the hydrophobicity profiles (Rose, 1978). Based upon this observation of hydrophobicity profiles, an alignment strategy for membrane proteins was developed for which membrane proteins were aligned based upon their hydropathy profiles (Lolkema and Slotboom, 1998). This strategy allowed for the detection of structurally homologous proteins despite their low pairwise sequence identity (Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998; Lolkema and Slotboom, 2005).

Other methodologies developed specifically for membrane proteins consider the distinct amino acid composition of membranous- and non-membranous protein segments. The MSA methods STAM (Shafir and Guy, 2004) and PRALINETM (Pirovano, et al., 2008), which were developed specifically for membrane proteins, apply both a membrane-specific substitution matrix (PHAT (Ng, et al., 2000)) for membrane-spanning segments and a generalized non-specific substitution matrix (BLOSUM62 (Henikoff and Henikoff, 1992)) for segments outside of the membrane. The difference between the two methods is their approach to handle membranous and non-membranous segments during the alignment. In STAM, transmembrane segments are first separated out and aligned independently whereas in PRALINETM such a division is not present and the sequences are kept undivided.

Using specific substitution rates depending on a protein's environment has also been shown to improve accuracy of pairwise-sequence-to-structure alignments. Substitution rates that depend on secondary structure, membrane propensities and solvent accessibility were applied in the program "membrane FUGUE". Membrane protein alignments generated using "membrane FUGUE" were more accurate than those generated with FUGUE, which is the equivalent approach for water-soluble proteins (Hill, et al., 2011). A recent development of membrane FUGUE, called MP-T, also incorporates homologous membrane protein sequences into a multiple sequence alignment in order to guide and improve pairwise sequence-to-structure alignments (Hill and Deane, 2012). MP-T was shown to perform well compared to other standard alignment methods tested using a membrane protein data set as a reference. At the time of the development and evaluation of AlignMe, MP-T was not available and thus has not been compared to AlignMe for its accuracy.

Two more recently developed methods that were optimized for membrane protein sequences (TM-Coffee and KalignP) unfortunately hold significant drawbacks. TM-Coffee is a version of PSI-Coffee that includes evolutionarily information in the form of evolutionarily profiles, which were obtained by a database search (Chang, et al., 2012). In contrast to PSI-Coffee, TM-Coffee includes only membrane protein sequences in the profiles. Tests on the BALiBASE reference set 7 for membrane proteins revealed a slightly better performance of TM-Coffee by some measures compared to other programs (i.e. MSAProbs) but TM-Coffee was also significantly slower. In contrast, the fast and low-memory usage method Kalign2, which is an update of Kalign, can handle position-specific gap penalties (i.e. in TM segments) but has been shown to be less accurate on the reference set 7 of BALiBASE than ProbCons or T-Coffee (Shu and Elofsson, 2011).

These drawbacks of alignment methods optimized for membrane proteins (e.g., high computational costs, no significant improvement of alignment accuracy, or non-applicability for local calculations) result in a less frequent usage of these programs. Additionally, multiple sequence alignment methods require a set of homologous protein sequences that might not be present for some membrane protein families. Currently, a common practice for membrane protein researchers is to apply standard (multiple) sequence alignment methods to gain insights into proteins of their interest.

This chapter addresses the issue of pairwise alignments for membrane protein sequences. The initial version (1.0) of a novel membrane protein alignment program, AlignMe, was designed and written in 2010 in cooperation with Dr. Rene Staritzbichler, with the help of important discussions with Dr. Kamil Khafizov, to test properties of a protein that might be useful for generating an accurate pairwise sequence alignment even for evolutionarily distantly-related proteins. A preliminary analysis was done in my diploma thesis (Stamm, 2010), showing that a combination of a substitution matrix

together with a secondary structure prediction and a membrane prediction generates accurate alignments. However, this evaluation was too much biased towards the reference data set. The gap penalties for a certain input combination for AlignMe were optimized on the same reference data set on which they were then subsequently evaluated. Thus, the high accuracy of AlignMe was not surprising and an evaluation on external data sets was outstanding. In addition, several other descriptors for membrane proteins (e.g., membrane predictors, position-specific substitution matrices etc.) were not tested so far. Thus, I updated the AlignMe program to allow for more sophisticated protein descriptors (e.g., membrane predictions by other computational methods or position-specific substitution matrices) resulting in version 1.1 of AlignMe. I used this new version of AlignMe to test the applicability of generalized or position-specific substitution matrices, hydrophobicity scales, secondary structure and membrane predictions for aligning membrane proteins. Gap penalties for the dynamic programming algorithm of AlignMe were optimized for each set of alignment inputs, using alignments of the HOME2 data set (see chapter 2.4) as a reference. The best of the different AlignMe strategies were then compared to a set of other commonly used methods using homology models of the HOME2 data set as well as to independent reference alignments of the BALiBASE reference 7 set of membrane proteins.

The work that is presented in this chapter was published in PLOS One in March 2013 (Stamm, et al., 2013). The open source code of AlignMe is available at <https://sourceforge.net/projects/alignme/>, and at <http://www.forrestlab.org/>, along with an online server and the HOME2 data set.

3.2 Methods

3.2.1 General Description of Similarity Between Homologous Proteins

AlignMe (for Alignment of Membrane proteins) is a pairwise protein sequence alignment tool implemented in C++ that has been developed for and tested upon membrane protein sequences. The underlying algorithm in AlignMe is a Needleman-Wunsch dynamic programming algorithm with a set of affine gap penalties. Initially, the optimal descriptors for an alignment of membrane proteins were not known and thus, AlignMe was designed to support single and multiple protein descriptors of different types for explaining similarity between two (membrane) protein sequences.

The similarity Sim between two residues (i, j) at a given alignment position is defined as a linear combination of substitution rates from M input substitution matrices (S), and differences between property values (V) from N input profiles at that specific sequence position:

$$Sim_{i,j} = \sum_m^M (w_m * S_{i,j}) - \sum_n^N (w_n * |V_i - V_j|) \quad (3.1)$$

This formalism allows for a flexible combination of any number of substitution rates (e.g., generalized or position-specific) with any number of profile values. In this study, transmembrane location propensities (e.g., predictions or hydrophobicity values) or secondary structure predictions are used in form of profiles. A profile assigns a value that describes a specific protein property (e.g., likelihood of being within the membrane) to each sequence position. The similarity term Sim also provides weights (w) for each input with which inputs can be scaled relative to each other in order to minimize any bias towards a specific input during the alignment process. For example, a hydrophobicity scale containing values from -3.0 to 1.0 (i.e., a range of 4.0) would be assigned $w = 5$ when used in combination with a substitution matrix whose values range from -5 to 15 (i.e., a range of 20). In this example, the range of the hydrophobicity scale and the substitution matrix are assigned a 1:1 ratio so that both input properties contribute equally to the alignment process. Such a weighting ratio of 1:1 was applied for all tested input combinations in this work.

Another parameter that is required for global alignments using a Needleman-Wunsch algorithm to define similarity between two protein sequences, aside from mutation rates, is the possibility of insertions and deletions. These evolutionarily aspects are reflected by gap penalties. A gap-opening penalty describes the likelihood that an insertion starts and a gap-extension penalty is the likelihood that an existing gap is extended. The basic Needleman-Wunsch algorithm has only these two types of gap penalties, but substitution rates at the C- or N-terminal domains of protein sequence have been

observed to occur with a different likelihood than point mutations in the middle of a protein's sequence. These differences are caused by the insertion or deletion of additional terminal domains (Dong, et al., 2008; Huang, 1994; Thompson, et al., 1994). AlignMe addresses this observation with the possibility to assign different gap-opening and gap-extension penalties according to whether a gap is at a terminus ($p_o^{terminal}$ and $p_e^{terminal}$, respectively), or not (p_o and p_e , respectively) (Edgar, 2004; Thompson, et al., 1994).

These alignment options might be sufficient for aligning soluble protein sequences, but in case of membrane protein sequences, different evolutionary rates were observed for TM segments and for non-membranous segments, with a higher conservation rate for TM segments (Stevens and Arkin, 2001). Insertions and deletions are less likely to occur in conserved segments and consequently the gap penalties should be higher in these segments (Thompson, et al., 2002). Accordingly, AlignMe has the option to apply different gap penalties for conserved and non-conserved protein segments based on a certain input criteria that can be defined by the user (i.e. hydrophobicity values or membrane propensities). Given a threshold value for one of the input parameters, gap penalties are then defined by the threshold (i.e., either above or below the threshold). In case of hydrophobicity scales, positions with values above the threshold (i.e., hydrophobic) receive different gap penalties (p_o^{above} and p_e^{above}) than those below the threshold (p_o^{below} and p_e^{below}) representing hydrophilic positions. This scheme assigns six gap penalty types in total, namely p_o^{above} , p_e^{above} , p_o^{below} , p_e^{below} , $p_o^{terminal}$ and $p_e^{terminal}$.

3.2.2 Inputs Tested that Define Similarity Between a Pair of Proteins

An accurate alignment of two membrane protein sequences requires an adequate description of the similarity between these two proteins (see previous chapter). Therefore, multiple properties of a protein were tested for their suitability to describe similarity between a pair of membrane proteins: (position-specific) substitution matrices, hydrophobicity scales, secondary structure and membrane propensity predictions. In a first step, all inputs were tested on their own for their suitability to generate accurate pairwise alignments. Subsequently, the protein descriptors that resulted on their own in the most accurate alignments were then tested in combination with each other to evaluate whether the inclusion of a second additional similarity criteria increases the alignment accuracy or not. Finally, three different alignment input descriptors were used in combination with each other.

Significance between alignments of different approaches (or AlignMe inputs) were measured using the Wilcoxon signed ranked test (Wilcoxon, 1946) and were deemed to be significant when $p < 0.05$.

3.2.2.1 *Substitution Matrices*

Substitution matrices describe the probability that an amino acid is replaced by another amino acid or stays conserved during evolution. Seven different substitution matrices were compared, each of them optimized on or for a special type of proteins.

First of all, there are substitution matrices that reflect a specific evolutionary divergence within a general protein data set they are constructed upon. The BLOSUM matrix series was derived from observed amino acid exchanges in block alignments of sequences with a certain degree of evolutionary divergence (Henikoff and Henikoff, 1992). Similarly, the PAM matrices are also described by a series of matrices. PAM matrices were trained on a set of closely homologous sequences using a Markovian model of amino-acid replacement (Dayhoff, et al., 1978). This training principle was also applied for generating the VTML matrices, but in contrast to the PAM matrices, a large set of distantly related homologs was used for training the VTML matrices (Müller, et al., 2002; Müller and Vingron, 2000).

Unlike those matrices that were optimized on general (water-soluble) protein data sets, the JTT (Jones, et al., 1994) and PHAT (Ng, et al., 2000) matrices were optimized for α -helical membrane proteins by taking substitution rates between membrane protein sequences into account. For the JTT matrix, an approach similar to the one for the BLOSUM matrix series was applied by using blocks of aligned transmembrane protein sequences sharing a certain degree of divergence. Even more specialized is the PHAT matrix that considers only substitution rates calculated from alignments of either predicted transmembrane segments or hydrophobic regions from proteins stored in the BLOCKS+ database.

Moreover, I tested the bbTM matrix (Jimenez-Morales, et al., 2008), which was constructed from a set of β -barrel-like protein sequences. The bbTM matrix was included to test whether substitution rates for β -barrel-like membrane proteins could also be suitable for aligning α -helical membrane protein sequences.

3.2.2.2 *Position-Specific Substitution Matrices (PSSMs)*

Generalized substitution matrices contain position-independent substitution rates and assume that evolutionarily substitution rates are the same for all amino acids within a protein sequence. However, evolutionarily rates may differ in conserved protein segments compared to non-conserved protein fragments. This idea of evolutionary variability in different positions along the sequence is provided by position-specific substitution matrices (PSSMs).

The position-specific substitution rates (S) of two proteins (i and j) were compared with each other in AlignMe using a simple approach. The similarity between an amino acid (A) in the first sequence (i) and an amino acid (B) in the second sequence (j) is the average value of the corresponding mutation from A to B (stored in the PSSM of the first sequence) and the reverse substitution B to A (stored in the PSSM of the second sequence):

$$Sim_{i,j} = 0.5 * (S_{A \rightarrow B}^i + S_{B \rightarrow A}^j) \quad (3.2)$$

For each sequence of the HOME2 data set, a PSSM was generated by a PSI-BLAST search on the Uniref90 database dated 28th April 2009 that was performed during the PSIPRED predictions for each sequence.

3.2.2.3 Hydrophobicity Scales

As for evolutionary rates, similarity between amino acids can be described by similarities or differences of hydrophobicity stored within the amino acid sequence. For this analysis, six different hydrophobicity scales were tested, each addressing a different aspect of membrane proteins. Several of those scales were derived from experimental free energies of transfer of amino-acids between ethanol and water (Nozaki and Tanford, 1971), including the scales reported by Hopp and Woods (HW) (Hopp and Woods, 1981) and by Wimley and White (WW) (Wimley and White, 1996). A combination of such transfer-free energies with known structural properties or theoretical considerations was applied for constructing the scales of Kyte and Doolittle (KD) (Kyte and Doolittle, 1982) and the Goldman, Engelman and Steitz scale (GES) (Engelman, et al., 1986), while Eisenberg and Weiss (EW) created a consensus of five other scales (Eisenberg, et al., 1982). White, von Heijne and colleagues (HWvH), derived a hydrophobicity scale from probabilities of α -helical segments inserting into a biological membrane (Hessa, et al., 2005), whereas the knowledge-based unified hydrophobicity scale (UHS) (Koehler, et al., 2009) was constructed from the distribution of amino acid types in known protein structures. When using hydrophobicity scales, any position with $V_i \geq 0$ was assigned to the membrane and this threshold was also used for assigning gap penalties to hydrophobic and hydrophilic segments.

The membrane-spanning segments of a protein also contain non-hydrophobic residues that are involved in binding sites, protein translocation or protein-protein interaction. Thus, hydrophobicity profiles considering each sequence position separately are very fuzzy. A manual visual detection or annotation of membrane-spanning segments is not possible. However, single outliers can be averaged out by creating a smoothed profile using a sliding window approach. In such smoothed

hydrophobicity profiles, a value at a given residue is replaced with an average over a window of residues centered at that position, and then that window is processed along the protein sequence (Kyte and Doolittle, 1982). Here, rectangular, triangular or sinusoidal windows of length $L = 13$ were tested (Koehler, et al., 2009). The sinusoidal shape mimics the amphipathic periodicity of a transmembrane helix, so that values that are 3.6 positions away from the center are given equal weight, while other positions contribute less. More details about sliding windows can be found in my diploma thesis, pages 19ff (Stamm, 2010).

3.2.2.4 Membrane Propensity Predictions

Although using hydrophobicity information within alignments can be useful for detecting homology between a pair of membrane protein sequences, hydrophobicity alone cannot predict reliably the location of transmembrane helices because hydrophobic residues can also be present within hydrophobic patches outside of the membrane and hydrophilic patches can be present within membrane-spanning segments. Hence, the challenge of predicting accurately membrane-spanning segments was addressed by more sophisticated predictors which were also tested here for their applicability to sequence alignments.

Three different predictors for α -helical transmembrane segments were evaluated: TMHMM (Krogh, et al., 2001), OCTOPUS (Viklund and Elofsson, 2008) and MEMSAT-SVM (Nugent and Jones, 2009). Each of them applies a different machine-learning algorithm with Hidden Markov Models for TMHMM, neural networks for OCTOPUS and a support vector machine (used for binary classification) that is applied by MEMSAT-SVM. Whereas predictions of TMHMM rely on the raw sequence only, those of OCTOPUS and MEMSAT-SVM consider additional evolutionarily information included in PSSMs. These PSSMs were obtained from a PSI-BLAST search against the corresponding recommended database, namely the Uniprot_Sprot database (on 1st August 2010) for MEMSAT-SVM and a version of Uniref90 filtered for transmembrane proteins (from 4th August 2010) for OCTOPUS (Viklund and Elofsson, 2008).

The per-residue membrane propensity was extracted from the results of each program and used as a profile input for AlignMe. Predictions of OCTOPUS and TMHMM contain a 0 to 1 propensity of a residue to be located within the membrane (1) or not (0). For both these programs, positions with per-residue membrane propensities >0.5 were defined as being in the membrane. In case of MEMSAT-SVM, there are no absolute borders that define the membrane propensity, but there is a threshold of 0; positions are assigned as being in the membrane if their propensities are >0 , or not if their propensities are negative.

3.2.2.5 Secondary Structure Predictions

Two different secondary structure predictors were tested: Jufo (Meiler and Baker, 2003) and PSIPRED (Jones, 1999) with version 2.6 and 3.2 of PSIPRED. Both predictors are based upon the results of a PSI-BLAST search for the protein of interest. Thus, PSI-BLAST searches were run for each method on the corresponding recommended database, i.e., Uniprot_Sprot (from 1st August 2010) and Uniref90 (from 28th April 2009), respectively.

Each method produces a three-state prediction of the probability of a position being in a coil, α -helix or β -sheet. This probability can vary between 0 (secondary structure type not present at that position) and 1 (secondary structure type present at that position). A position was assigned to be in a certain state (e.g., α -helical) if the predicted probability thereof was >0.5 .

All three probabilities were used as input profiles for AlignMe alone as well as in combination with each other, with each state contributing one third of the whole.

3.2.3 Alignment Difference Score (AD score) as an Alignment Accuracy Measure

The accuracy of a sequence alignment is often evaluated using a score that counts the fraction of correctly aligned positions with respect to the reference alignment (Edgar, 2009; Edgar, 2010) (see Figure 3.1). However, this simple score becomes less useful for assessing the alignment accuracy of more distantly related protein sequences due to its lack of discriminating between different degrees of mismatches. A structural element (e.g., an α -helix) that is partially misaligned to an equivalent helix (e.g., to gaps; see Figure 3.1c) would receive the same score like in the case of a helix being shifted by single residue only (see Figure 3.1b). A more useful starting point for analysis and homology modeling is clearly the latter example that contains more residues with an underlying template structure of a similar secondary structure type like the original structure. Consequently, other scores consider also the shift size, defined as the number of positions that a residue in the test alignment is displaced from its aligned column in the reference alignment. For example, the fraction of positions aligned within a certain shift size has been used (Tang, et al., 2003), with the disadvantage that it introduces an arbitrary cut-off in the accuracy measure. In a more advanced strategy, the Cline score penalizes shifts asymptotically, so that it emphasizes residues that are close to their correct position and undervalues errors of greater than four positions (Cline, et al., 2002). A drawback of all these approaches is that none of them takes into account residues that should be aligned to a gap (e.g., evolutionarily insertions or deletions of segments). Thus, I developed a new scoring scheme (AD score), which takes into account shifted residues and considers insertions and deletions more explicitly.

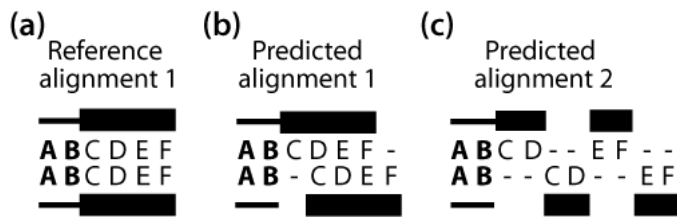


Figure 3.1 Determination of the fraction of correctly aligned positions. A reference alignment (a) between two sequences is used to score two test alignments (b, c). The alignment of the residues in bold is correct in both test alignments, and thus the fraction of correctly-aligned positions will be 2/7 or 28,6% in both cases. However, the alignment of the other residues is more useful in predicted alignment 1 (b) than in predicted alignment 2 (c), although this difference is not captured when considering only the correctly aligned positions. Hypothetical helical segments in each sequence are marked with thick black bars.

The Alignment Difference (AD) score that I developed and use here is similar to the mean shift error (MSE) score (Moult, et al., 1998) or the position shift error (PSE) score (Raghava, et al., 2003), and takes into account the full extent of any shifts. Residues aligned to other amino acids (and not to gaps) in the reference alignment (bold residues in Figure 3.2a) are assessed within the test alignment and their shift value is calculated. The shift value describes how many positions a residue in a test alignment is shifted compared to its position in the reference alignment. A score of zero is assigned to a sequence position if its amino acids are correctly aligned in the test alignment, whereas shifted positions are penalized by the shift value, as in the MSE score (see Figure 3.2b).

In the AD score, gap-containing columns are considered explicitly in the test alignment, but these columns are treated differently than those that contain aligned amino acids: the shift value of such columns is defined as the mean of the shift values for the two residues either side of the aligned gap (see Figure 3.2c and d). The final total AD score is the sum of the (negative) shift values of all columns of the reference alignment divided by its length. Thus, a perfect alignment has a total AD score of zero, while more negative values represent less accurate alignments. The AD score correlates with the fraction of correctly aligned positions, but the two measures deviate at low values, and thus the AD score provides distinct information in that realm (see Figure 3.3).

However, I have to note that although the AD score contains information about shift size, information regarding the direction of the shift is missing. Consequently, a mis-alignment of two α -helices in which one helix is disrupted by the insertion of single gaps (Figure 3.2f) receives the same AD score as an alignment in which the entire helix is shifted by a single position (Figure 3.2g). However, single or short gaps that disrupt a secondary structure element were so far not observed in alignments of membrane-spanning segments and thus were typically also penalized highly even within the alignment process itself (e.g., CLUSTAL W (Thompson, et al., 2002)). The gap penalty optimization itself favors to match frequently occurring patterns (e.g., helix aligned against a helix,

Figure 3.2f) over less frequent patterns (e.g., gaps in helices, Figure 3.2g); so the drawback caused by the lack of directionality of the AD score is negligible.

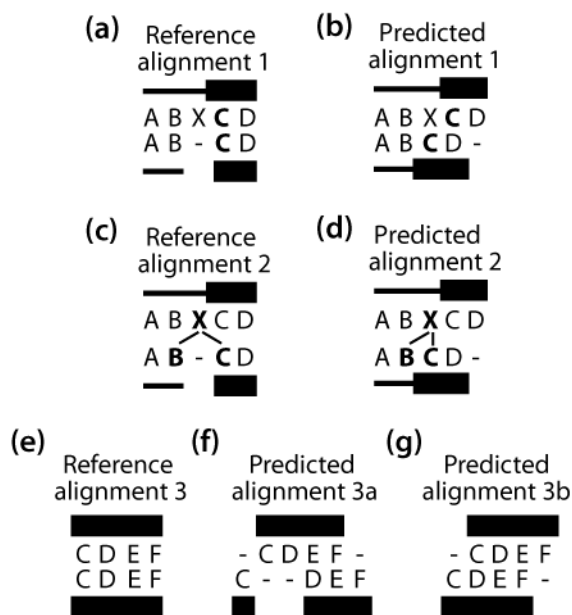


Figure 3.2 Determination of the Alignment Difference (AD) score. (a, b) Two residues (bold) that are aligned in the reference alignment (a) have a shift value of 1 in the test alignment (b). (c, d) The accuracy of a column containing a gap (c) is determined by identifying the two residues adjacent to the gap (residues B and C), and calculating the mean of their shift values in the test alignment (d). For residue B the shift value is 0, and for residue C the shift value is 1, so the shift value of this gapped column is 0.5. (e, f, g) Note that the AD score does not penalize the direction of the shift. Thus, two structural elements (residues CDEF) aligned in a reference alignment (e) have the same AD score for the test alignment, whether the element is divided (f) or shifted as a whole (g).

3.2.4 Optimization of Gap Penalty Sets

A Needleman-Wunsch algorithm requires suitable gap penalties reflecting the likelihood of evolutionarily insertions or deletions in addition to suitable descriptors that describe evolutionarily mutations (see chapter 3.2.1). However, suitable gap penalties are not known for aligning membrane proteins or for using hydrophobicity, secondary structure or membrane predictions for an alignment process. For AlignMe, the optimization of 4 or 6 gap penalties (see chapter 3.2.1) is required, but a systematic optimization (Edgar, 2009) is not computationally feasible for optimizing such a large number of gap penalties. In the case of six different gap penalties: if each gap penalty were allowed to range from 0 to 10 in increments of 0.1, a systematic search would require $100^6 = 10^{12}$ alignments. Consequently, a Monte Carlo scheme was used to optimize the gap penalty values (e.g., 4 or 6 gap penalties) for each input and each combination of inputs tested.

In each step of the optimization process, pairwise alignments for all pairs of proteins of HOME2 were created using the current set of gap penalties and subsequently analyzed for their accuracy using the AD score. A drawback during this optimization process could be the fact that families within the HOME2 data set have different sizes, causing the optimization to be biased towards large families. Therefore, all AD scores of protein-pairs within a family of HOME2 were averaged by the size of each family (m). The overall alignment accuracy for a given set of gap penalty parameters is the sum over all scores for each family (n) in HOME2.

$$Alignment\ accuracy(total) = \sum_1^n \frac{\sum_1^m Alignment\ accuracy(pairwise)}{m} \quad (3.3)$$

Starting with a randomly selected set of values (between 0 and 30) for each gap penalty parameter, the search procedure then involved random modifications of one or more gap penalty values from those values, or from the optimal values identified so far. The range of allowed modifications was initially set to be very small (with a maximal step size of 0.06) to encourage a detailed examination of the score landscape around the current optimal gap penalty combination. A given combination of gap penalties was accepted if the overall alignment accuracy score was better than the best score found so far, in which case the maximal step size was reset to its initial (minimal) value. Otherwise, that combination of gap penalties was rejected and the search space was expanded by increasing the maximum step size by 0.06. However, the gap penalty values were limited to the range from 0 to 30, with a maximum step size of 30 (e.g., 500 rejections). If no improvements were found after reaching the maximum step size, the search was repeated, starting with the initial maximal step size of 0.06. The search was finished if new gap penalty sets were rejected more than 1000 times in a row because then the search space was examined in detail twice without finding a better gap penalty combination than the current optimal gap penalty combination.

For each set of input descriptors, this optimization procedure was repeated 20 times in parallel with different initial gap penalty values, which was found to be sufficient for reasonable convergence (see Table 3.1). The gap penalty parameters for which the alignments had the best alignment accuracy score were then used for that set of input descriptors. As a validation of this optimization idea, optimized gap penalty sets were compared to those that were reported previously for a specific input descriptor. The optimal gap penalties obtained with my optimizing procedure using the JTT membrane substitution matrix were $p_o = 16.3$, $p_e = 1.3$, $p_o^{terminal} = 1.7$ and $p_e^{terminal} = 0.6$, consistent with typical values (e.g. (Saigo, et al., 2006)), providing confidence in the optimization procedure.

Table 3.1. Total AD scores of 20 optimization runs for AlignMePST mode

# of test run	total AD score	p_o^{below}	p_o^{above}	p_e^{below}	p_e^{above}	$p_o^{terminal}$	$p_e^{terminal}$
5	-17.39	2.14	2.96	3.10	3.06	0.07	1.18
17	-17.42	2.13	2.77	2.95	2.97	0.04	1.24
16	-17.43	2.23	3.09	3.18	3.19	0.13	1.27
9	-17.45	2.16	2.88	3.03	3.04	0.20	1.25
4	-17.49	1.84	2.70	3.05	2.93	1.26	1.18
11	-17.60	2.09	2.70	2.91	2.93	4.47	1.16
12	-17.63	2.03	2.53	2.60	2.63	3.44	1.12
13	-17.69	1.73	2.62	3.39	2.93	0.64	1.13
19	-17.73	1.88	2.02	2.68	2.52	1.60	1.10
8	-17.73	1.62	2.04	2.77	2.42	1.01	1.08
1	-17.95	2.42	5.83	3.12	3.23	0.05	1.39
2	-17.96	1.92	5.56	2.70	2.81	0.01	1.32
3	-17.98	2.28	5.24	2.98	3.20	0.02	1.37
20	-18.23	1.89	8.57	3.22	3.11	1.95	1.51
7	-18.25	2.03	14.60	2.83	2.11	0.25	1.06
6	-18.31	2.31	14.49	3.30	2.58	0.10	1.46
15	-18.42	2.14	23.85	3.52	2.87	0.04	1.60
10	-18.59	2.21	24.27	3.75	2.20	0.05	1.25
14	-20.75	15.13	3.92	2.83	3.22	6.77	1.88

Total alignment accuracy (AD) scores were calculated for optimized gap penalties of 20 runs using a PSSM, a secondary structure prediction of PSIPRED and a membrane prediction of OCTOPUS as an input. Higher total AD scores correspond to more accurate alignments (with 0 being a perfect alignment) for all protein pairs tested. Similar gap penalties were observed for the top-ranking optimization runs.

Other parameters that could be optimized are the weights that are assigned to each input descriptor and that define the contribution of that input descriptor to the final alignment. However, an optimization of these weights was found to be computationally impractical because the search increases by the power of N , with N being the number of weights to be optimized. Moreover, initial tests of optimizing weight parameters did not converge reliably.

3.2.5 Parameters of Other Alignment Methods Tested

Alignments were also calculated with HMAP (Tang, et al., 2003), T-Coffee v8.9.1 (Notredame, et al., 2000), MUSCLE v3.7 (Edgar, 2004), ProbCons v1.12 (Do, et al., 2005), MSAProbs v0.9.4 (Liu, et al., 2010) and HHalign v1.5.0 (Söding, 2005). For MSAs, sequence homologues for each of the sequences were identified using a PSI-BLAST search on the non-redundant (nr) database dated 4th August 2010, with five iterations, an E-value cut-off of 10^{-4} and a maximum of 2500 sequences. Sequences in the PSI-BLAST results that were more than twice the length of the query were filtered out. The remaining sequences were clustered using UCLUST (Edgar, 2010) with the original sequence taken as the

representative of the first cluster and a sequence identity cut-off of 65 %. This is a reasonable sequence identity cut-off that falls into the limits of 50% and 70% pairwise sequence identity, which were shown to be the optimal cut-offs for clustering a database (Park, et al., 2000). Additionally, this cut-off was already applied successfully in an earlier study (Tang, et al., 2003). This clustering principle to reduce the number of input sequences was necessary for T-Coffee, ProbCons and MSAProbs, which are extremely memory- and cpu-intensive, in order to make the test over the whole HOME2 dataset computationally tractable, and so, for all the tested MSA methods (including MUSCLE) I used the suggested T-Coffee protocol, namely selecting the 25 “most-informative” homologues of each sequence (including the query) from the UCLUST clustered results (Notredame, et al., 2000).

There are two different possible approaches for generating a MSA from two query sequences and their respective homologues. In the standard approach, all results of both PSI-BLAST searches (including the two query sequences) are combined and aligned as a single large MSA, before extracting the two query sequences for scoring. The second approach, which is called the “profile-profile” strategy, is to create MSAs for each query and its homologues. The resulting two MSAs or “profiles” are then aligned to one another to create a single MSA, from which the query sequences are then extracted for scoring.

A similar strategy to the MSA “profile-to profile” approach was applied by HAlign and HMAP, which both construct profiles for each sequence and then subsequently generate profile-to-profile alignments based upon these profiles. In HAlign, each query is described by a Hidden Markov Model (HMM) based on the results from a PSI-BLAST search (as for AlignMePSSMs, see chapter 3.2.2.2), as well as by secondary structure predictions from PSIPRED, generated as described above (see chapter 3.2.2.5). Those HMMs were then globally aligned to each other by using the “-mact 0.0” maximum accuracy flag and by assigning all other parameters their default values. In HMAP, one of the sequences was assigned to be the query, and its profile included evolutionarily information from a PSI-BLAST search (obtained as for HAlign and AlignMePSSMs, see chapter 3.2.2.2) combined with a predicted secondary structure from PSIPRED v3.2; the other sequence was assigned to be the template, and its profile was similar except that the secondary structure was assigned from the known spatial information using DSSP, where available. The two profiles were then globally aligned using HMAP using the flag for global alignments to allow for a direct comparison with the global alignment method AlignMe.

I also tested TM-Coffee (Chang, et al., 2012), but found the computational cost prohibitive for the large number of pairwise alignments in the BALiBASE set (see chapter 3.2.7). STAM, PRALINETM and

MP-T were not available for local installation, and therefore could also not be tested on the large reference data sets.

3.2.6 Evaluations Based on Homology Modeling

An assessment regarding the alignment accuracy of all methods tested here relies on building homology models based on pairwise alignments, and comparing them to their native crystal structures. This evaluation principle is independent from the reference alignments that were used for optimizing the gap penalties of AlignMe and so is able to detect whether AlignMe is over-optimized on the HOME2 data set or if the optimization process resulted in parameters that can be generalized. However, it has to be noted that the gap penalties of AlignMe are based on an optimization against the HOME2 data set, which might have a positive influence on the ability of AlignMe to generate alignments that are useful for building accurate homology models of those same proteins.

For every pair of protein sequences from HOME2, each protein was modeled using the structure of the other protein as a template based on an alignment from each alignment method tested here. In each case, five models were created using Modeller v9.9 with default settings (Šali and Blundell, 1993). The model with the best (lowest) DOPE energy score was chosen as the top model for subsequent analysis. With this selection procedure, the influence of Modeller to generate a single accurate model is averaged down. The top model was then evaluated using two different measures of structural similarity: the GDT_TS score (global distance test total score) and the AL4 score (% of residues aligned within 10 Å) (Forrest, et al., 2006; Zemla, 2003).

The GDT_TS score, which stands for global distance test (total score), is defined as the percentage of C_α-atom pairs from the model and the native structure averaged over four different cut-off distances (i.e. 1, 2, 4 and 8 Å) and correlates well with the percentage of correctly aligned residues (Figure 3.3b). The advantage of the GDT_TS score is that only correctly modeled positions are rewarded, without a penalty for inaccurately modeled regions. Nonetheless, the score is still dependent on the size of the protein.

$$\text{GDT_TS}(\%) = \frac{1}{4} \sum_{c=1,2,4,8} \left[\frac{G(c)}{L_{\text{target}}} \times 100 \right] \quad (3.4)$$

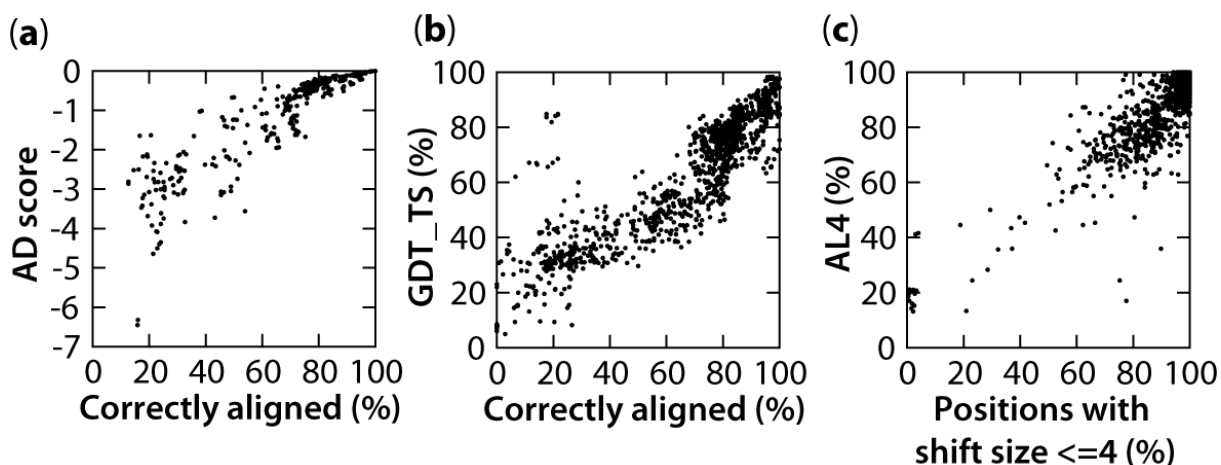


Figure 3.3. Correlations between alignment accuracy measures. All measures were calculated based on HOMEP2 alignments generated with AlignMePST. (a, b) Fraction of correctly aligned positions plotted against either the AD score (a) or the corresponding model quality measured using GDT_TS (b). (c) The percentage of residues correctly aligned within a shift of 4 positions is plotted against the corresponding model quality measured using the AL4 score. Correlation coefficients are: (a) 0.90, (b) 0.92 and (c) 0.88.

By contrast the AL4 score allows a clearer discrimination between low-accuracy models than the GDT_TS score, since it identifies the largest subset of C_{α} -atoms of the model that can be superimposed with the target structure (of size L_{target}) below a certain cut-off distance $G(c)$ of 10 Å, corresponding to an approximate shift of four alignment positions (Figure 3.3c).

$$AL(\%) = \frac{G(c)}{L_{target}} \times 100 \quad (3.5)$$

For helical membrane proteins, such small shift errors of four positions can still be readily overcome by manual adjustments to the alignment, and therefore it is deemed to be a useful cut-off for the analysis. In other words, the AL4 is a suitable measure of model quality because it focuses on all models that can be refined manually.

3.2.7 BALiBASE Reference 7 Test Set

Another independent assessment of alignment accuracy was obtained using the Reference 7 set of BALiBASE (Bahr, et al., 2001; Bahr, et al., 2001) as a gold standard for membrane protein alignments. This set contains 435 membrane proteins in 8 superfamilies, namely 7tm, acr, photo, dtd, ion, msl, Nat and ptga, each multiply aligned. The first three of these families are represented to some extent in the GPCR, multidrug efflux and (bacterio)rhodopsin families, respectively, of HOMEP2 (Table A.1). During the evaluation, alignments were generated for all pairs of sequences in each of the 8 superfamilies. Since I evaluate pairwise sequence alignments, I calculated the fraction of correctly

aligned residues as well as the average shift error for each pairwise alignment, rather than SP (Sum of Pairs) or TC (Total Column) scores, which are sometimes used to describe the accuracy of MSAs and require more than two sequences in an alignment.

The so-called ‘core’ regions provided by BALiBASE were not analysed explicitly, as they have been shown to correspond only weakly to conserved secondary structures (Edgar, 2010) or membrane-spanning elements in this set (see Table 3.2). Instead, I analysed segments in each pairwise alignment that were predicted to be membrane-embedded in both sequences by MEMSAT-SVM (e.g., membrane propensity values higher than 0). MEMSAT-SVM was chosen as a confident predictor because it was shown to be the most accurate among all membrane predictors tested in this study (see chapter 3.3.1.4). Additionally, by using MEMSAT-SVM a bias in the analysis towards one of the alignment methods that uses OCTOPUS is avoided (AlignMePST; see chapter 3.3.2).

Table 3.2. Percentage of positions in ‘core’ regions predicted as TM segments by MEMSAT-SVM

	7tm	acr	dtd	ion	msl	photo	ptga
MEMSAT-SVM	91,0%	84,4 %	80,0%	33,4%	98,2%	80,0%	22,0%

Overlap between BALiBASE definition of ‘core’ region, and predicted membrane-spanning segments according to a prediction from MEMSAT-SVM.

3.3 Results

I first describe the alignment accuracy of all input descriptors tested using AlignMe individually, and then describe the selection of input descriptors tested in combination with each other for the identification of the optimal combination of alignment descriptors for the alignment of membrane protein sequences. To make the comparison between descriptors as fair as possible, gap penalties were optimized and evaluated for their alignment accuracy for each single descriptor and each set of combined input descriptors tested. So far, I first constructed an updated set of homologous membrane protein structures (HOMEP2; see chapter 2.4), and structure-based pairwise sequence alignments of these proteins were used as a gold standard for assessing alignment accuracy. Input descriptors of AlignMe were considered to be effective and accurate if the AlignMe alignments had both a high number of correctly aligned positions and a small shift error in relation to the reference alignments of HOMEP2, measured as less negative AD scores (see chapters 3.2.3 and 3.2.4). By also considering the shift error, I expect to help identify methods that are effective for very distantly related proteins. The findings are described for single inputs (chapter 3.3.1) and then for combinations of inputs (chapter 3.3.2). Finally, I compare three of the optimized AlignMe strategies with available alignment programs using reference alignments of the HOMEP2 set (chapter 3.3.3) as well as homology models (chapter 3.3.4) and the BALiBASE reference set 7 (chapter 3.3.5).

3.3.1 Single Input Descriptors

A set of substitution matrices, hydrophobicity scales (with and without window-averaging), secondary structure and transmembrane predictions were tested as individual inputs (see chapters 3.2.2.1 - 3.2.2.5). In all cases, the highest alignment accuracy score (AD score), which was found during the optimization, is shown.

3.3.1.1 Alignment Accuracy Using (Position-Specific) Substitution Matrices

Comparing alignments constructed using different general substitution matrices (Figure 3.4a), the closest agreement with the reference alignments was obtained with the membrane-specific JTT matrix, followed by the general-purpose VTML matrix, although the differences between JTT and the others were not very significant ($p = 0.01$ to 0.33 ; Figure 3.4a). The substitution rates of both matrices are suitable for aligning proteins of the HOMEP2 data set because they were obtained from a set of protein sequences similar to those in HOMEP2. In the case of JTT, the values for the matrix were obtained using blocks of aligned transmembrane sequences of diverse sequence identity and for the VTML matrix, a large set of distantly related homologs were used for obtaining substitution rates.

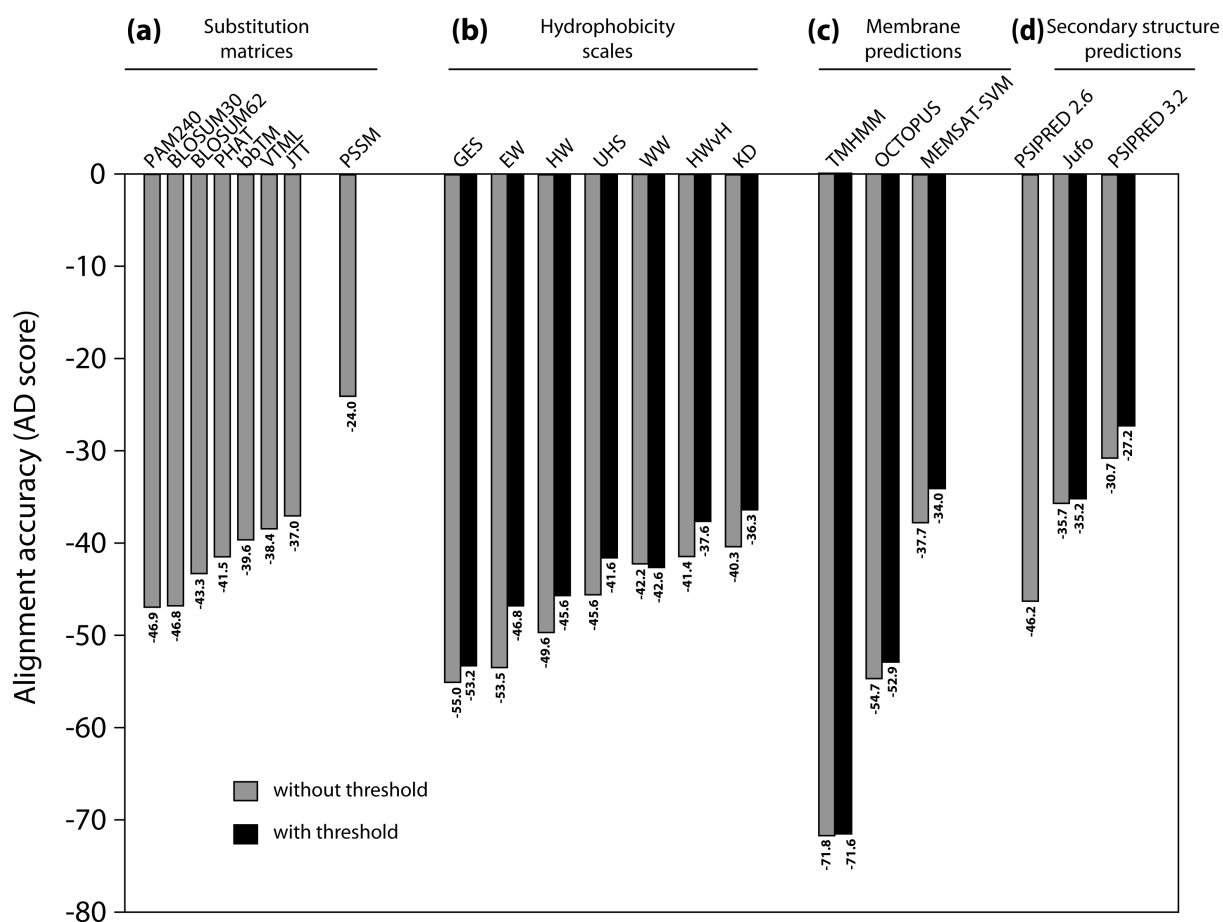


Figure 3.4 Comparison of alignment accuracy when using single input descriptors in AlignMe. The total alignment accuracy score (AD score) for all α -helical proteins in the HOMEP2 dataset is plotted for each of the input descriptors using their optimized gap penalties, and arranged according to increasing score for different (a) substitution matrices, (b) hydrophobicity scales (with no smoothing), (c) other transmembrane predictions or (d) secondary structure predictions. Sequence segments with hydrophobic, helical or transmembrane scores above a given threshold could be assigned the same (gray bars; without threshold) or different (black bars; with threshold) gap penalty values from segments below that threshold (see chapter 3.2 for definition of threshold values and abbreviations).

However, alignments using position-specific substitution rates from PSSMs were significantly more accurate (AD score = -24.0 ; $p < 10^{-9}$) than those generated using general substitution matrices, which do not account for position-flexible substitution rates. This result underlines the fact that evolutionarily rates vary along a protein sequence.

3.3.1.2 Alignment Accuracy Using Hydrophobicity Scales

Representing sequences by their hydrophobicity values (without averaging their values over a sliding window) is, in essence, equivalent to using a substitution matrix, except with a focus on one specific physicochemical property. Applying a set of four gap penalties, which treats all non-terminal gaps as equivalent, (gray bars) for constructing alignments using hydrophobicity scales (Figure 3.4b) results in slightly (not significantly) less accurate alignments than those generated with the best of the

generalized substitution matrices. The accuracy increased if non-terminal gap penalties were allowed to differ within the transmembrane segments (e.g., different gap penalties in hydrophobic and in hydrophilic fragments), but again the differences to the most accurate generalized substitution matrices were not statistically significant (Figure 3.4b black bars; see chapter 3.2.1). The alignments generated using the KD, HWvH and WW hydrophobicity scales were significantly more accurate ($p > 0.05$) than those generated with other hydrophobicity scales (Figure 3.4.b, black bars), but not significantly different from one another ($p < 0.05$).

An interesting observation is that the KD and HWvH scales resulted in the most accurate alignments, since the KD scale, one of the first hydrophobicity scales developed, is based on a consensus of biophysical and structural data (Kyte and Doolittle, 1982), whereas the HWvH scale was obtained more recently from biochemical studies of helix insertions into membrane bilayers (Hessa, et al., 2005). The similarity of the results indicates that these two distinct strategies are both able to describe membrane proteins well.

Next, all hydrophobicity scales were used for testing the effect of window averaging on the alignment accuracy. Independent of the shape of the sliding window used (e.g., rectangular, triangular or sinusoidal – see chapter 3.2.2.3), alignments using window averaging were less accurate compared to those without window averaging and treating all non-terminal gaps as equivalent (see Figure 3.5). Allowing the gap penalties to differ in hydrophobic and hydrophilic fragments (e.g., a set of 6 gap penalties) improved the alignment accuracy, but it still remained lower than without averaging. This decrease in alignment accuracy is presumably caused due to a loss of position-specific information that was shown to increase the alignment accuracy for substitution matrices significantly (see Figure 3.4a).

Overall, the usage of hydrophobicity values as in a hydrophobicity plot (see Figure 3.5) did not significantly improve the alignment accuracy compared to alignments using generalized substitution matrices and moreover resulted in significantly less accurate alignments compared to those generated using PSSMs (see Figure 3.4a).

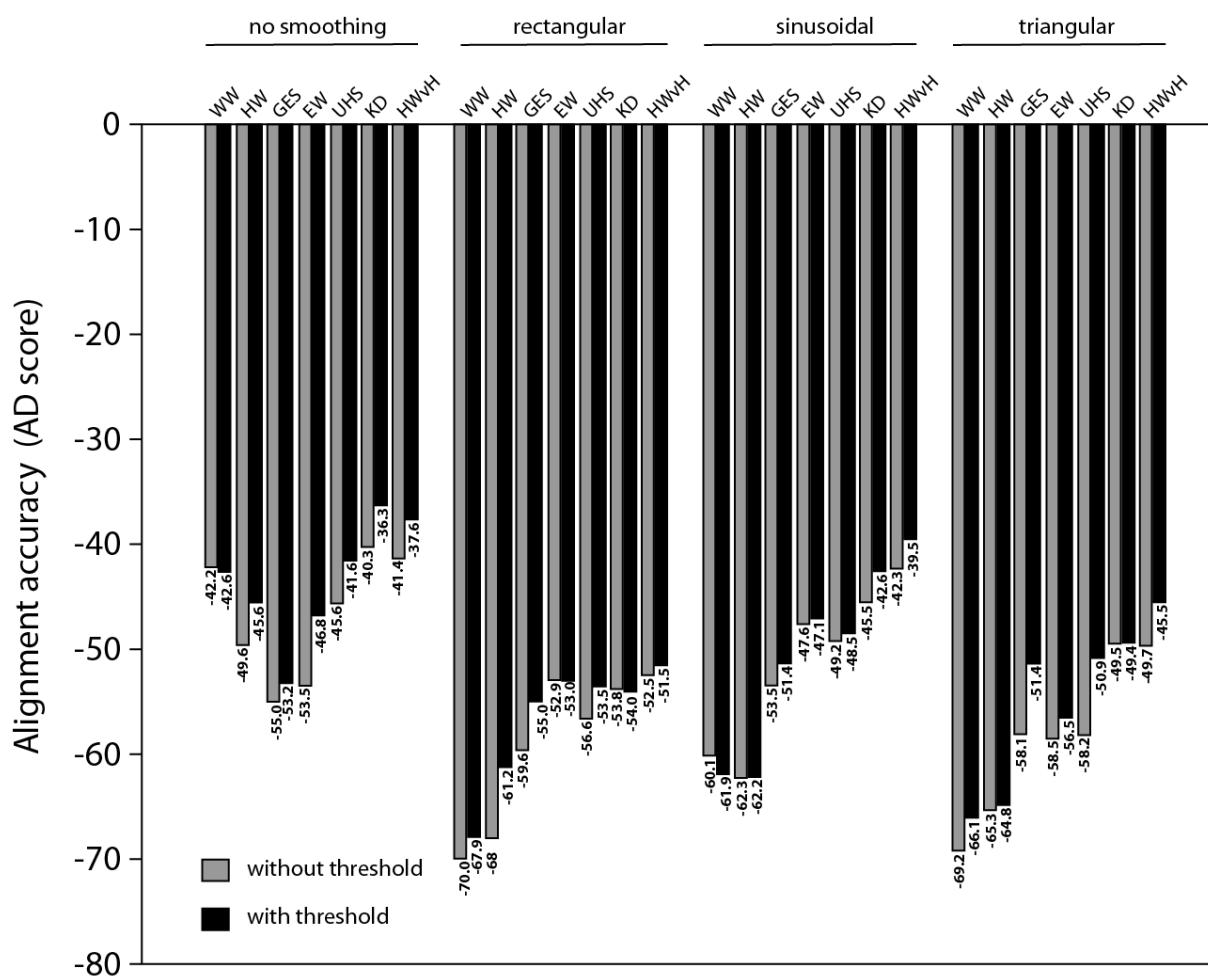


Figure 3.5 Comparison of alignment accuracy when using hydrophobicity scales as input descriptors in AlignMe. The different hydrophobicity scales were either not smoothed (“no smoothing”) as shown also in Figure 3.4, or window-averaged using a rectangular-, sinusoidal- or triangular-shaped window for averaging. See legend to Figure 3.4 for more details.

3.3.1.3 Alignment Accuracy Using Secondary Structure Predictions

When representing the sequences as profiles of predicted secondary structure types, the alignments in closest agreement with the reference alignments were obtained using PSIPRED3.2 predictions (Figure 3.4d) compared to those of PSIPRED 2.6 and Jufo. However, only the difference between Jufo and PSIPRED3.2 is statistically significant ($p = 0.01$, gray bars) whereas the differences between the different PSIPRED versions are not significant ($p > 0.05$). This non-significant difference is caused by a disproportionate contribution of good PSIPRED2.6 alignments in the (large) aquaporin family, which have a major influence on the Wilcoxon-signed rank test for calculating significance; this contribution is not reflected in the total AD scores in Figure 3.4d because AD scores are averaged over families (see chapter 3.2.3). Interestingly, the alignment accuracy relative to the reference alignments correlates with the accuracy of the corresponding transmembrane prediction method, with PSIPRED3.2 being more accurate (75.3% of residues are correctly predicted) than the other methods

tested (74.0% for PSIPRED2.6, and 70.4% for Jufo; $p < 0.05$) for the HOME2 protein set, using DSSP assignments as a reference (Kabsch and Sander, 1983). Notably, allowing the penalties for gaps in predicted α -helical structure elements to differ from those in other regions improved the alignment accuracy significantly (black bars, Figure 3.4d).

3.3.1.4 Alignment Accuracy Using Membrane Propensity Predictions

More sophisticated methods for predicting the location of membrane-spanning helices than the usage of hydrophobicity scales, are those of TMHMM, OCTOPUS and MEMSAT-SVM. Here, again, the alignment accuracy correlates with that of the underlying prediction method. Using PDB_TM assignments as a reference for a position-specific membrane propensity, MEMSAT-SVM is a significantly more accurate predictor (88.2% of the residues in HOME2 are correctly predicted), followed by OCTOPUS (86.4%) and TMHMM (83.0%) (compare with Figure 3.4c).

Treating all non-terminal gaps as equivalent, alignments generated using MEMSAT-SVM alone were not significantly more accurate than those obtained using a hydrophobicity scale or a substitution matrix. As for the hydrophobicity scale based alignments, the MEMSAT-SVM and OCTOPUS-based alignments became significantly more accurate when penalties were assigned differently to gaps in membrane and non-membrane segments (black bars, Figure 3.4c). The alignments generated based upon predictions of MEMSAT-SVM were also significantly ($p < 10^{-4}$) more accurate than those using other membrane propensity predictors and those generated using the best of the hydrophobicity scales (KD, Figure 3.4b).

Comparing all the alignments generated with a single input descriptor, I find that the alignments that were significantly most similar to the reference alignments, and therefore also most accurate, were obtained using position-specific matrices (PSSMs), followed by secondary structure predictions (PSIPRED3.2, $p = 2 \times 10^{-5}$), and transmembrane predictions (MEMSAT-SVM, $p = 5 \times 10^{-6}$) (Figure 3.4). This finding reflects the more detailed information included in the evolutionarily profiles compared to the secondary structure and transmembrane predictions.

3.3.2 Combined Input Descriptors

Using the results for single inputs, I next tested alignments for which the best two or three input descriptors were used in combination, since inclusion of complementary information is expected to progressively improve alignment accuracy (see, e.g., (Forrest, et al., 2006; Kelley, et al., 2000)). Therefore, I tested evolutionarily information combined with transmembrane and secondary structure predictions, a combination of membrane and secondary structure predictions and a combination of all three of them.

3.3.2.1 PSSMs Combined with a Transmembrane Prediction

A potentially useful combination for membrane proteins is evolutionarily information with the addition of transmembrane information containing membrane likeliness propensities for each sequence position that is stored in a smoothed profile. The latter can be in the form of either a hydrophobicity value (e.g., using a sliding window approach) or a transmembrane prediction propensity generated by a sophisticated prediction method. Interestingly, in AlignMe, nearly all such combinations of evolutionarily information with membrane propensities resulted in significantly more accurate alignments than those based on the corresponding individual input parameters. However, this improvement required an extended gap-penalty scheme that allowed gap penalties to differ between membrane and non-membrane regions (black bars, Figure 3.6a). Otherwise, a significant improvement of accuracy was not observed if all non-terminal gaps are treated as equivalent. Surprisingly, alignments based on PSSMs were significantly more accurate when combined with OCTOPUS (total AD score of -20.4 , Figure 3.6a) than with MEMSAT-SVM (total AD score of -22.8 , Figure 3.6a), even though MEMSAT-SVM predictions are more accurate *per se* (chapter 3.3.1.4). This apparent contradiction can be explained by several factors that influence the alignment accuracy, which are listed here in the order of their likeliness. One possible explanation may be that OCTOPUS predictions of two related proteins match one another better in an alignment than those of MEMSAT-SVM. Another explanation may be that the OCTOPUS predictions have a simpler form, perhaps providing more orthogonal (complementary) information to the PSSMs than the more detailed and smoothed profiles obtained from MEMSAT-SVM (Figure 3.7). Alternatively, the fact that the MEMSAT-SVM values are more evenly distributed over a wider range of values than the OCTOPUS scores (Figure 3.7) and are thus given a smaller weighting (see chapter 3.2.1) could mean that the MEMSAT-SVM scores can have less influence on the alignments.

3.3.2.2 PSSMs Combined with a Secondary Structure Prediction

Combining secondary structure with evolutionarily information has already been shown to improve alignment quality of profile-to-profile alignments for water-soluble proteins (e.g., (Tang, et al., 2003)). Using AlignMe, a similar improvement is observed in the HOMEPE2 alignments when combining evolutionarily rates of PSSMs with secondary structure information of PSIPRED predictions. Compared to the most accurate alignments using secondary structure predictions (total AD score of -27.2) or evolutionarily information (total AD score of -24.0) as a single input, the combination of both produces significantly more accurate alignments (total AD score of -21.6 , $p = 0.04$; Figure 3.6b, black bars).

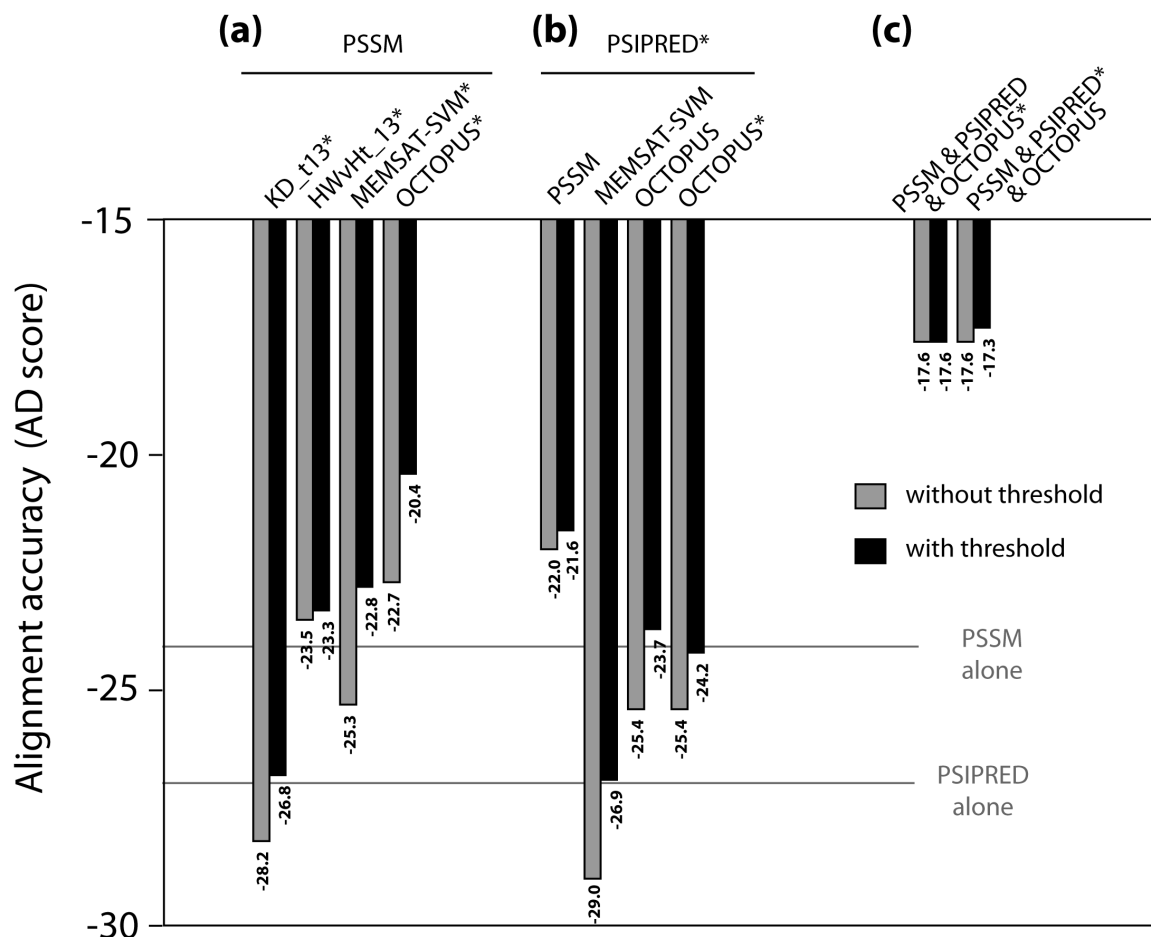


Figure 3.6 Comparison of alignment accuracy when using multiple input descriptors in AlignMe. Combinations included: (a) PSSMs with hydrophobicity descriptors or transmembrane predictions; (b) secondary structure prediction with PSSMs or transmembrane predictions; or (c) PSSMs, PSIPRED and OCTOPUS together. The scores obtained using PSSMs or PSIPRED alone are indicated with gray lines for reference. Gap penalties were assigned differently to sequence segments above or below a threshold (black bars), and the threshold was defined using the inputs marked by *. For example, in the PSIPRED* & OCTOPUS combination, the threshold was assigned using PSIPRED. See legend for Figure 3.4 for further details.

3.3.2.3 Secondary Structure Prediction Combined with a Transmembrane Prediction

Also a combination of secondary structure likeliness with membrane propensities might add additional information to the alignment since only a combination of both methods can distinguish between membrane-spanning α -helices, non-membranous α -helices and other types of secondary structure elements. Alignments using a combination of a secondary structure prediction with a transmembrane prediction were also significantly more accurate (total AD score of -23.7 for PSIPRED combined with OCTOPUS; Figure 3.6b) than alignments using each descriptor on its own (total AD score of -27.2 for PSIPRED and -52.9 for OCTOPUS; Figure 3.4), with OCTOPUS again being the best choice of transmembrane predictor (Figure 3.6b). In these combinations, when assigning gap penalties differently to structured regions (chapter 3.2.1), the latter may be defined using either secondary structure or membrane propensity. I found that using α -helix positions for this distinction (OCTOPUS and PSIPRED*, total AD score of -23.7 , Figure 3.6b) led to significantly ($p = 0.03$) more accurate alignments than when using the transmembrane positions for assigning the thresholds (OCTOPUS* and PSIPRED, total AD score of -24.2 , Figure 3.6b).

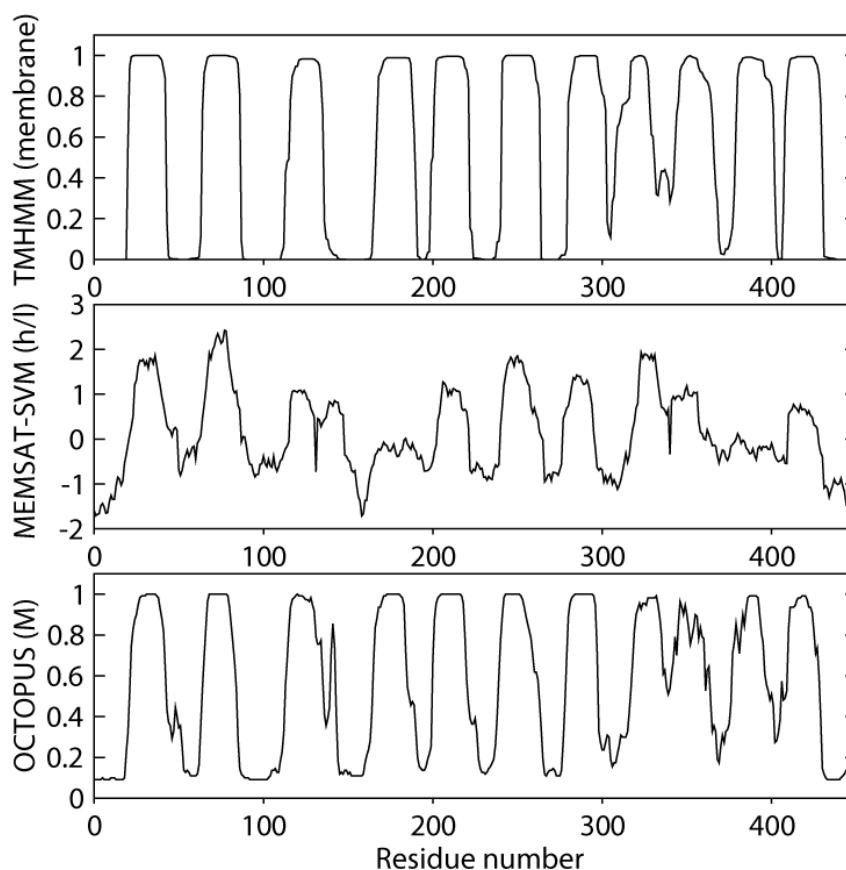


Figure 3.7 Profiles of the predicted membrane propensity from the three different transmembrane helix prediction methods tested for AlignMe. Predictions were generated for the chloride channel protein with PDB identifier 1KPL. The y-axis label contains the header of the corresponding column in the output from the prediction method.

Two major observations can be obtained from these results. First, the increase in alignment accuracy proves the assumption that that secondary structure and transmembrane predictions contain complementary information, consistent with the fact that not all secondary structure elements in a membrane protein are located within the membrane. Indeed, among the 60% of the residues that are outside the membrane in HOME2 structures (as defined by PDB_TM), 46.2% of residues are α -helical, and 7.4% are in a β -strand. Moreover, not all transmembrane segments are fully helical (Kauko, et al., 2008), and include segments of coil (7.5% of residues) and even β -sheet (0.1% of residues). A combination of both methods includes all information whereas each method on its own misses either membrane information (e.g., secondary structure predictors) or additional secondary structure information of coils and sheets within membranous and non-membranous segments (e.g., membrane predictors). Another observation is that gap insertion should be disfavored in all structured regions, whether in the membrane or not. Consequently, a gap-penalty scheme that is based on secondary structure propensities allows for more accurate alignments. Nevertheless, no matter how the different regions are assigned to define an extended gap-penalty scheme, the alignments are significantly more accurate using an extended gap penalty scheme than when the same gap penalty values are used in both structured and unstructured regions (Figure 3.6b).

3.3.2.4 Combinations of PSSMs with Secondary Structure and Transmembrane Predictions

All combinations of two input descriptors were shown to result in more accurate alignments. Consequently, I tested the three protein descriptors with the most useful and complementary information (PSSM, OCTOPUS and PSIPRED3.2) in combination with each other. This combination increased the alignment accuracy significantly (total AD score = -17.6 , $p < 0.05$, Figure 3.6c) even though a simple gap penalty was used, which only differentiates between terminal and non-terminal open and extension gap penalties. A further significant increase was obtained by assigning gap penalties according to secondary structure propensity (total AD score = -17.3 , $p < 0.05$, Figure 3.6c), but not by transmembrane position (total AD score = -17.6 , $p = 0.32$, Figure 3.6c).

Interestingly, input descriptors that led to the most accurate alignments when used alone were not always the most effective when they were used in combination with complementary input descriptors of another type (e.g., OCTOPUS contributed more in combination than alone, and the converse was found for MEMSAT-SVM; Figure 3.4 and Figure 3.6). Presumably, single input descriptors may contain detailed information to produce an accurate alignment, whereas in combination that information may become redundant or even conflicting with the other input descriptor that is used. Clearly this suggests that it would be desirable to optimize the parameters on

all combinations of all descriptors as well as optimizing the contribution (e.g., weights) of each input descriptor to the alignment, but unfortunately this is not computationally tractable at this time.

For subsequent evaluations I decided to compare three different versions of AlignMe for their accuracy on HOME2 as well as on the BALiBASE reference 7 set. All three modes of AlignMe contain different input descriptors to test whether the inclusion of additional information increases the alignment accuracy also significantly on an independent data set or not. For reference, I tested one version, called AlignMeP, that uses only evolutionarily information (PSSM), with gap penalties of $p_o = 15.36$, $p_e = 0.88$, $p_o^{terminal} = 1.69$ and $p_e^{terminal} = 0.25$. In the AlignMePS version, secondary structure information (PSIPRED3.2) was used besides evolutionarily information of PSSMs, with optimized gap penalties of $p_o^{above} = 6.80$, $p_e^{above} = 2.28$, $p_o^{below} = 6.22$, $p_e^{below} = 1.37$, $p_o^{terminal} = 0.29$ and $p_e^{terminal} = 0.86$. Finally, I tested the effect of including transmembrane information from OCTOPUS in addition to PSSMs and PSIPRED3.2 within AlignMePST. For this combination, gap penalties of $p_o^{above} = 2.96$, $p_e^{above} = 3.06$, $p_o^{below} = 2.14$, $p_e^{below} = 3.06$, $p_o^{terminal} = 0.07$ and $p_e^{terminal} = 1.18$ were shown to generate accurate alignments for that combination. In both AlignMePS and AlignMePST versions, α -helicity was used to define the gap penalty assignment threshold of 0.5 based on the α -helical propensities obtained with PSIPRED 3.2.

3.3.3 Comparison of AlignMe with Other Sequence Alignment Methods on the HOME2 Data Set

I compared the three AlignMe versions, i.e., evolutionarily information without (AlignMeP) or with secondary structure propensities (AlignMePS) or with additional matching of membrane probabilities (AlignMePST), to several available multiple-sequence alignment programs (e.g., ProbCons, MSAProbs, T-Coffee and Muscle), as well as the pairwise profile-to-profile alignment program HMAP, and the HMM-HMM alignment program HAlign (see chapter 3.2.5). Here, I first assess alignment accuracy using sequence alignments of the structure-based HOME2 reference dataset that was used for training. Then I evaluate the accuracy of homology models built from those alignments as a reference-independent measure of alignment quality. In both cases, the HOME2 reference data set is split up into three groups of close homologues (>30% identical residues), homologues with moderate similarity (15-30%) and those with low similarity (<15%).

At first sight, alignments generated with AlignMe contain the highest number of correctly aligned residues and the lowest shift errors for all three different similarity levels within the HOME2 data set compared to other sequence alignment methods. However, the different modes of AlignMe (P,

PS and PST) are ranked differently depending on the sequence similarity of the proteins that are aligned. Additionally, the significance between the different modes and all MSA methods tested has to be considered.

For pairs of membrane protein sequences in the HOME2 set within the group of low (0-15 %) and moderate (15-30 %) similarity, AlignMe alignments have significantly more (~2 %) correctly-aligned positions than all other methods (Table 3.3 and Figure 3.8a, c, e) with all three modes being not significantly different from each other for the group of low similarity proteins and with AlignMeP and PS being most accurate for the group of moderate similarity. In the case of close homologues (>30% identical residues) in the HOME2 set, alignments generated by AlignMeP and PS have the highest number of correctly aligned residues but AlignMePST and HAlign alignments also have a high fraction of correctly aligned residues, for example (Table 3.3 and Figure 3.8c).

Table 3.3 Accuracy of alignments generated using different methods on the HOME2 data set

	0-15 % (44)		15-30 % (71)		30-85 % (62)	
	% correct	shift	% correct	shift	% correct	shift
AlignMeP	30.1*	4.31	72.0	1.15	88.2	0.25*
AlignMePS	30.6*	3.35	71.5	1.16	87.9*	0.28
AlignMePST	30.7	2.73	70.4	0.85	87.5	0.21
AlignMePST x-fold	30.3	2.89	70.4	0.89	87.3	0.30
MSAProbs	28.3	7.22	68.6	1.08	85.7	0.24*
HAlign	17.3	10.50	61.8	1.75	86.5*	0.29*
HMAP	24.9	7.00	68.6	1.27	85.3	0.32
MUSCLE	26.4	9.41	68.5	1.13	85.5	0.31
MUSCLE profile-profile	25.6	9.77	63.6	1.65	75.6	0.86
ProbCons	26.7	8.30	67.0	1.34	84.2	0.31
T-Coffee	25.3	7.55	66.5	1.27	83.4	0.32
T-Coffee profile-profile	14.5	35.22	55.9	2.25	70.7	1.09

Results are sorted according to the level of sequence similarity of the sequence pair, in percentage identity. The number of pairwise alignments is shown in parentheses. The percentage of correctly aligned residues (% correct) and average shift error size (shift) with respect to the structure-based reference alignments (see chapter 3.2.3) are reported. *Values marked with an asterisk in this and all other subsequent tables are not significantly different from those of AlignMePST (p -value > 0.05) based on a pairwise Wilcoxon signed rank test. All other values are significantly different from those of AlignMePST. Entries in bold in this table, and all subsequent tables, indicate the highest or best scores in that column, including all values that are not significantly different from the best scores.

For all similarity groups, misaligned residues are shifted for fewer positions in AlignMe alignments, particularly when transmembrane information is included (see Table 3.3). A significant difference in the average shift error from AlignMePST to all other methods is observed for proteins sharing a low and moderate sequence identity but not for protein pairs sharing a high sequence identity for which the average shift error is similarly low for HAlign, MSAProbs, AlignMeP and AlignMePST alignments (see Table 3.3 and Figure 3.8b, d, and f). The significant reduction in shift errors reflects optimization of the gap penalties to the shift-size sensitive AD score.

It has to be noted that the reduction in shift errors obtained by matching transmembrane predictions (AlignMePST *cf.* AlignMePS) does come at the cost of some correctly-aligned positions, especially for sequences with moderate similarity (15-30 %). As mentioned above, for homology modeling of distantly-related pairs of proteins with low sequence identity it can be useful to reduce the magnitude of large shift errors since manual adjustment of an alignment can be aided relatively easily by conservation mapping once the helices are approximately aligned. For similar reasons, it is also interesting to know whether homologous transmembrane helices have been matched to some extent, as many (although not all) functional residues (e.g., residues involved in ligand binding or transport) lie in these regions. The matching of transmembrane helices in the HOME2 set by AlignMe appears to be particularly effective: using AlignMePS and AlignMePST, $\geq 97\%$ of the known transmembrane helices overlap by at least half of their residues, and $\geq 62\%$ of the helices (at least 10% more than the next best method) overlap by at least 90% of their residues (Table 3.4). These enhancements are achieved largely by the inclusion of secondary structure information (compare AlignMePS to AlignMeP), and to some extent by the matching of transmembrane predictions (compare AlignMePS to AlignMePST). However, even without transmembrane predictions, AlignMePS also matches these membrane-spanning segments of distant homologs (0-15%) significantly better (8-12% more overlap by at least half of their residues) than another method that considers secondary structure (HMAP), at least on the HOME2 training set.

All these results based on HOME2 are perhaps unsurprising given that the gap penalties are optimized for this data set. Consequently, an independent evaluation is required that does not rely on the data that AlignMe is optimized upon.

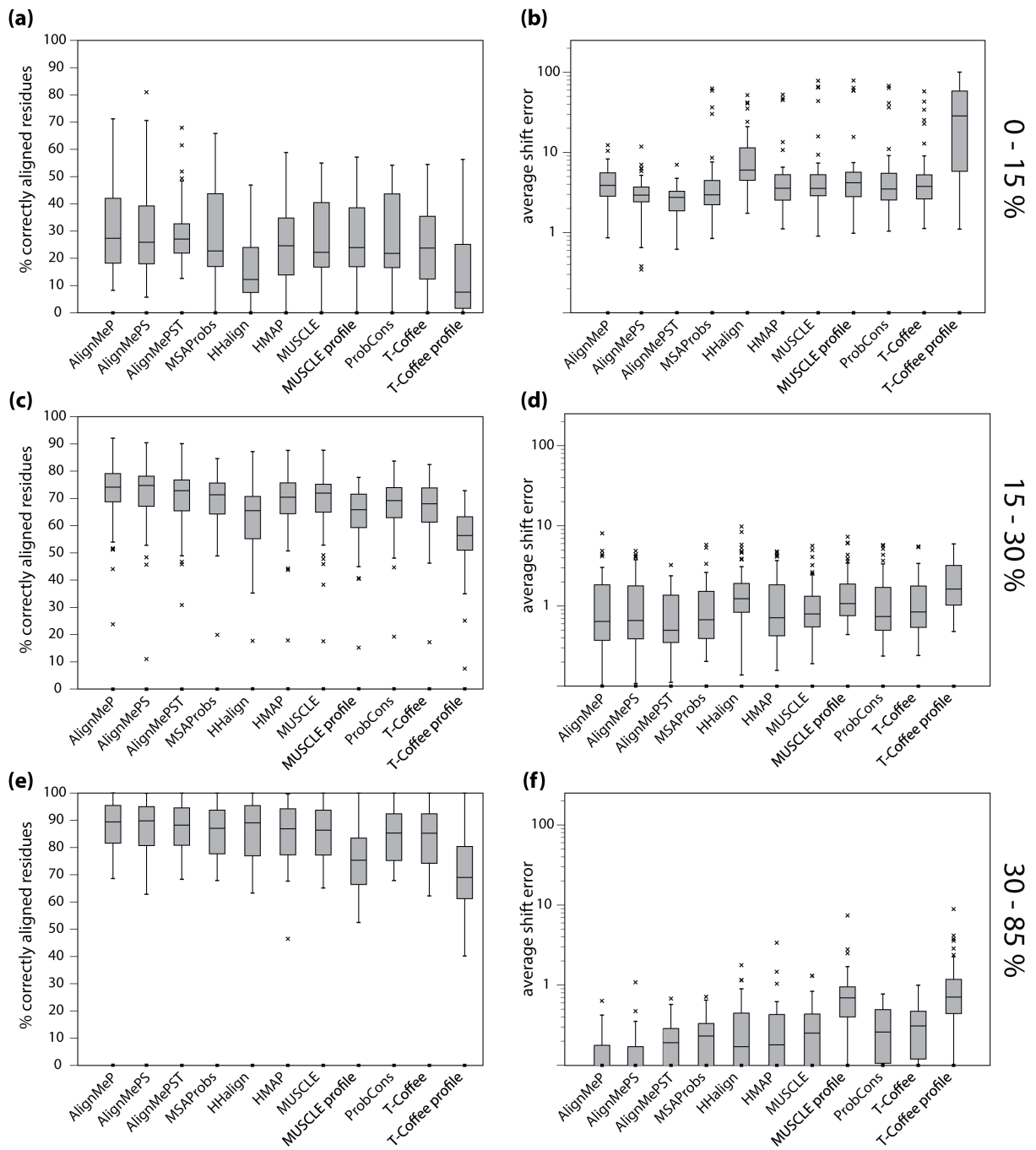


Figure 3.8 Accuracy of HOME2 alignments generated by different methods. Accuracy is measured using (a, c, e) the percentage of correctly-aligned residues, or (b, d, f) the average shift error, shown for all alignments as a box and whisker plot. Alignments are sorted according to the percentage identity between the sequences, namely (a, b) <15%, (c, d) 15-30% or (e, f) >30%.

Table 3.4 Percentage of transmembrane segments in the HOMEP2 set that are correctly aligned by each method

	0-15 % (44)		15-30 % (71)		30-85 % (62)	
	f^{50}	f^{90}	f^{50}	f^{90}	f^{50}	f^{90}
AlignMeP	93.65	52.80	98.64	95.54	100.00*	99.31*
AlignMePS	97.00	62.43*	99.49*	96.85*	100.00*	99.08*
AlignMePST	98.32	63.73	100.00	97.17	100.00	99.77
MSAProbs	90.42	53.01	99.49*	95.90*	100.00*	99.31*
HHalign	70.50	28.61	97.05	76.97	100.00*	95.72*
HMAP	85.83	54.31	99.49*	96.87*	100.00*	99.08*
MUSCLE	82.92	49.59	99.60*	93.89	100.00*	99.04
MUSCLE profile-profile	82.20	48.90	98.08	86.30	99.46*	88.16
ProbCons	89.73	52.17	99.49*	95.78*	100.00*	99.04
T-Coffee	88.02	51.18	99.49*	95.42	100.00*	98.85
T-Coffee profile-profile	38.32	18.75	95.56	66.68	97.33	73.12

Transmembrane segment definitions are taken from the structures according to the PDB_TM database (see Methods); matching is defined as correct if 50% (f^{50}) or 90% (f^{90}) of the residues are aligned. Results are sorted according to the level of sequence similarity of the sequence pair. The number of pairwise alignments is shown in parentheses. See legend for Table 3.3 for further details.

An obvious concern regarding the robustness of the AlignMe method(s) is an overtraining of the gap penalties for specific input descriptors due to the limited number of membrane protein structures available in the HOMEP2 data set. I first tested the robustness of the gap penalty optimization process by using cross-validation. The optimization of a set of 6 gap penalties using the input descriptors of the AlignMePST method (PSSM, secondary structure prediction by PSIPRED 3.2 and a transmembrane prediction of OCTOPUS) was repeated 11 times by leaving out 2 of the 22 families of HOMEP2 and thus using only 20 of the 22 families for the optimization process. In each case, the resultant gap penalties were used to evaluate the alignment accuracy of the remaining two families. As shown in Table 3.3 (see AlignMePST x-fold), the accuracy of the alignments using these gap penalties was similar to that obtained by training and testing on the whole HOMEP2 set. Moreover, the mean and standard deviation of the gap penalties of the cross-validation for the AlignMePST combination indicates relatively small variations between the different cross-folds (i.e., after optimization on different subsets): $p_o^{above} = 3.00 \pm 0.27$, $p_e^{above} = 3.16 \pm 0.46$, $p_o^{below} = 2.06 \pm 0.16$, $p_e^{below} = 2.86 \pm 0.25$, $p_o^{terminal} = 0.97 \pm 1.48$ and $p_e^{terminal} = 1.23 \pm 0.06$. These results suggest that the gap penalties are not significantly over-trained on a particular family of the HOMEP2 dataset, and

thus should be applicable to other membrane protein sequences, which are not included in the HOME2 data set yet.

3.3.4 Evaluation of Alignment Accuracy Based on Homology Modeling

Another concern addressing the alignment accuracy of the different AlignMe modes is the optimization of their gap penalties and the subsequent evaluation based on the same reference sequence alignments. An independent assessment criterion is necessary to ensure that AlignMe is not over-optimized on the HOME2 data set and that the alignment parameters of AlignMe can also be applied to sequences that are not reflected directly by the HOME2 data set. Accordingly, structural (homology) models were generated with Modeller v.9.9 using all pairwise sequence alignments that were generated so far by all methods (AlignMe, MSA methods, profile-profile methods). The models that were generated by Modeller were then compared to the native structures by calculating structural similarity scores (e.g., GDT_TS and AL4 values – see chapter 3.2.6), and also compared to “gold standard” models built using the reference sequence alignments extracted from the SKA structural alignments.

First, the percentage of correctly modeled protein fragments was examined using the GDT_TS score, which is also closely correlated to the number of correctly aligned residues. According to the GDT_TS score, several methods (e.g., AlignMe, MSAProbs etc.) have a similar accuracy on average (Table 3.5 and Figure 3.3). However, alignments generated with any of the AlignMe modes result in fewer very poor models (with GDT_TS < 20%), while there are models based on alignments of other methods whose GDT_TS scores are as low as 5% for distantly related proteins of the HOME2 set (Figure 3.9a). Next, the AL4 score that correlates well with the average shift error (see Figure 3.3) was applied to analyze homology model accuracy. This score discriminates better between low-accuracy models (Figure 3.9b). Models based on AlignMePS and AlignMePST alignments have significantly higher AL4 scores (up to 5% higher) than the best of the other alignment methods (see Table 3.5).

Both results show that alignments of all AlignMe modes contain a low shift error (cf. Table 3.3) due to the optimization towards less negative AD scores. Additionally, I have to note that models built from the structure-based reference alignments are the most accurate (SKA; Table 3.5 and Figure 3.9). This indicates that an optimization against these reference alignments was a useful procedure but also that there remains room for improvement in alignment methods.

Table 3.5 Accuracy of homology models constructed based on HOME2 data set alignments

	0-15 % (88)		15-30 % (142)		30-85% (124)	
	GDT_TS	AL4	GDT_TS	AL4	GDT_TS	AL4
AlignMeP	34.74	73.97	67.53*	90.75	83.94*	97.65
AlignMePS	38.06	79.97*	67.40*	90.52	83.79*	97.33
AlignMePST	36.30	80.48	67.36	92.19	83.96	98.03
MSAProbs	36.71*	75.00	67.33*	90.81	84.17*	97.76
HHalign	25.08	59.06	61.38	87.71	83.12	97.63
HMAP	36.33*	74.97	67.31*	90.44	83.25	97.04
MUSCLE	32.95	69.02	66.00	90.66	82.89	97.31
Muscle profile-profile	32.56	69.35	62.19	88.82	75.75	94.24
ProbCons	35.28*	72.78	67.16*	90.22	83.29	97.46
T-Coffee	35.30*	72.20	66.78	90.42	83.38	97.57
T-Coffee profile-profile	18.27	37.85	59.30	86.58	73.03	92.95
SKA structure-based ^a	46.38	85.42	71.12	93.99	85.51	98.18*

^aReference alignments generated by the structure alignment program, SKA. The number of models is shown in parentheses. See legend for Table 3.3 for further details.

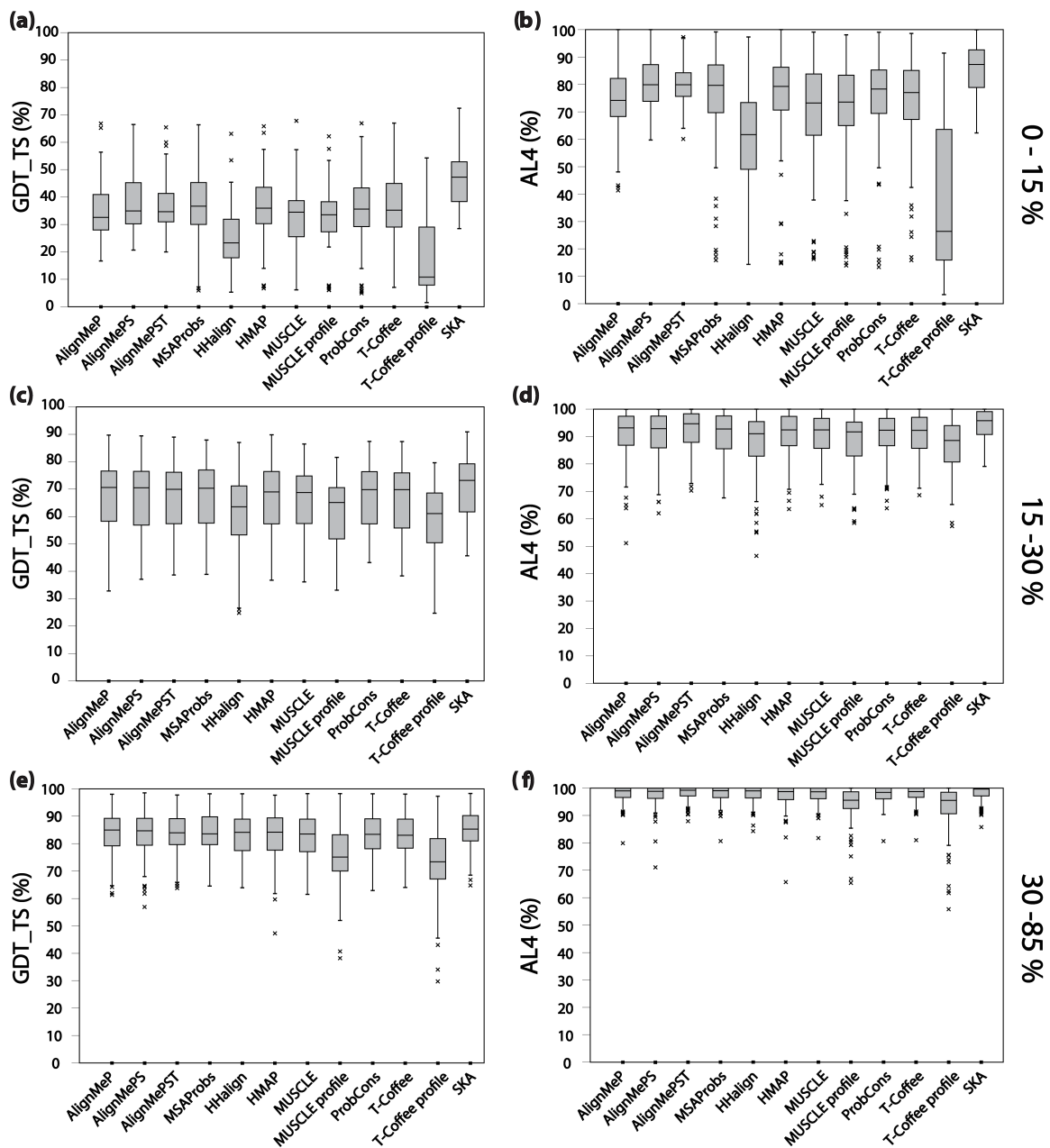


Figure 3.9 Accuracy of homology models built from HOME2 alignments generated by different methods. Accuracy is measured using the (a, c, e) GDT_TS and (b, d, f) AL4 scores of homology models compared to the known structures. Alignments are sorted according to the percentage identity between the sequences, namely (a, b) <15%, (c, d) 15-30% or (e, f) >30%.

3.3.5 Comparison of AlignMe with Other Sequence Alignment Methods on the BALiBASE Reference 7 Test Set

The alignment accuracy of the various alignment methods was also evaluated on an independent data set of membrane protein sequences (reference set 7 of BALiBASE; see chapter 3.2.7). This second assessment is also required to prove that the alignment parameters of AlignMe are not over-optimized on the HOME2 data set and that AlignMe also generates accurate alignments for protein sequences not included in HOME2. The BALiBASE reference 7 dataset contains manually-curated multiple-sequence alignments, based on PFAM alignments and was optimized to improve amino acid and secondary structure matching. At the time of the data set construction, no structural information was available to help guide the dataset generation (Bahr, et al., 2001) and such structural data is still missing for the majority of the proteins within that set. Consequently, I analyze the accuracy of all pairwise alignments in BALiBASE, but I am not able to generate homology models for this data set. Finally, I separate out the results for closely- and distantly-related protein sequences to gain a deeper insight into the influence of sequence similarity on the alignment quality.

Interestingly, the results vary depending on the average sequence identity of the proteins within each family. For the “ion” family, which has a very low average sequence identity (11.7 % - see Table 3.6), significantly more accurate alignments were obtained with AlignMePST showing that matching of transmembrane segments is favorable for protein pairs sharing a low sequence identity.

For the next 5 families, which follow in the order of increasing average sequence identity, AlignMePS alignments were found to statistically have the significantly highest number of correctly aligned residues. The inclusion of secondary structure-specific information seems to be useful at those sequence identity levels (14.3 % to 26.9 % - see Table 3.6), whereas membrane-specific information becomes less useful. Finally, alignments of AlignMeP are shown to be most accurate for the families with the highest sequence identity (e.g., 27.3 % of photo and 35.3 % of msl – see Table 3.6) but those alignments are not significantly better than those of AlignMePS for the “msl” family. The application of PSSMs alone seems to be useful only if the sequence identity is high enough, because the alignment accuracy of AlignMeP drops significantly compared to other methods the lower the sequence identity gets. Overall, AlignMePS alignments have the most correctly aligned residues in BALiBASE set 7 on average, including those not represented in the training set, followed by AlignMePST and the profile-to-profile method HMAP, which has also high-ranking scores for four out of the eight families.

Table 3.6 Percentage of residues that are correctly aligned in pairwise sequence alignments from the BALiBASE reference set 7, sorted by sequence identity of the protein families

	ion	Nat	ptga	7tm	dtd	acr	photo	msl	mean
AlignMeP	38.9	43.5	42.1	42.5	67.1	87.0	87.9	82.5	61.4
AlignMePS	45.2	66.2	64.8	65.9	76.0	89.7	87.6	82.3	72.2
AlignMePST	48.1	58.6	58.8	59.4	71.2	86.3	82.9	76.5	67.7
MSAProbs	24.5	53.3	45.9	54.7	64.4	89.0	73.4	70.6	59.5
HHalign	39.1	48.9	42.3	38.4	42.7	49.5	67.3	59.9	48.5
HMAP	32.8	61.9	54.9	61.4	65.3	87.6	83.4	78.5	65.7
MUSCLE	27.9	56.8*	48.4	56.6	70.3	89.5	80.5	76.1*	46.7
MUSCLE profile-profile	18.5	47.1	39.7	48.2	67.4	88.5	70.4	64.1	55.5
ProbCons	23.8	52.0	44.1	54.4	63.7	88.7	69.3	66.8	57.9
T-Coffee	25.5	50.6	44.2	55.1	63.7	88.8	67.5	67.5	57.9
T-Coffee profile-profile	10.8	14.5	27.0	40.2	52.9	86.2	52.1	53.0	42.1
Number ^a	1326	1711	1275	8128	1485	903	528	91	
Sequence identity (%) ^b	11.7± 13.8	14.3± 10.8	15.9± 12.1	18.2± 9.7	18.7± 11.5	26.9± 11.3	27.3±1 6.9	35.3± 13.5	

Mean = mean percentage of correctly-aligned residues over averages for eight families. ^aNumber of pair-wise alignments. ^bMean (\pm standard deviation) of the percentage sequence identity between pairs of alignments in each family. See legend for Table 3.3 for further details.

All families of BALiBASE reference 7 consist of different subgroups of proteins that are more closely related to each other than to proteins of other subgroups (Bahr, et al., 2001). Accordingly, I split up the BALiBASE set into sequences assigned to the same subgroup (Table 3.7) or to different subgroups (Table 3.8). The high ranking of the various AlignMe methods (with AlignMePS being the most accurate method) and of HMAP remains for alignments of both closely and distantly-related sequence pairs (Table 3.7 and Table 3.8). The average sequence similarity is higher for sequences assigned to the same subgroup (Table 3.7) than for sequences assigned to different subgroups (Table 3.8). Accordingly, alignments of AlignMeP, that are based on evolutionarily information only, are more accurate for closely related sequences from the same subgroup (ranked 2nd, Table 3.7) than for those from different subgroups (ranked 7th, Table 3.8). Indeed, AlignMeP alignments are significantly more accurate for the most similar sequences (in the dtd and photo families, Table 3.7) showing that membrane profile or secondary structure matching is not beneficial for closely related sequences. Nevertheless, secondary structure and transmembrane information becomes progressively more useful as the similarity decreases, especially for those assigned to different subgroups (Table 3.8) whereas the accuracy of alignments based upon evolutionarily information only (AlignMeP) decreases significantly as the similarity decreases.

Table 3.7 Percentage of residues that are correctly aligned in pairwise sequence alignments assigned to the same subgroup within the BALiBASE reference set 7, sorted by sequence identity of the alignments in each protein family

	ion	ptga	7tm	Nat	acr	msl	dtd	photo	mean
AlignMeP	62.8	83.4	67.6	80.6	93.4	82.0	90.3	94.7	81.8
AlignMePS	64.9	83.9	74.2	81.8	93.9	81.7	89.6	94.0	83.0
AlignMePST	62.9	81.7	68.4	79.3	92.4	78.3	86.9	91.4	80.2
MSAProbs	44.3	67.5	62.5	71.1	92.5*	74.4	84.5	88.8	73.2
HHalign	51.6	52.0	43.9	64.8	56.0	58.6	66.4	84.3	59.7
HMAP	50.6	75.2	69.2*	77.5*	91.7	80.9	82.8	90.6*	77.3
MUSCLE	47.0	72.3	62.6	72.4	93.0	78.0*	85.0	88.6	74.9
MUSCLE profile-profile	25.1	60.8	53.5	54.3	91.6	62.6	74.7	74.1	62.1
ProbCons	43.8	66.5	62.1	69.7	92.2	69.9	83.7	83.6	71.4
T-Coffee	45.9	69.8	64.7	72.5	92.2	76.8	85.2	87.0	74.3
T-Coffee profile-profile	45.3	66.3	63.5	70.4	92.1	71.4	84.1	83.6	72.1
Number	551	559	1082	282	420	51	84	122	
Sequence identity (%)	22.1± 16.6	26.7± 11.0	28.0± 20.0	31.3± 16.7	34.4± 12.9	43.6± 12.7	49.5± 19.1	52.2± 18.1	

See legend to Table 3.6 for more details.

Table 3.8 Percentage of residues that are correctly aligned in pairwise sequence alignments assigned to different subgroups within the BALiBASE reference set 7, sorted by sequence identity of the alignments in each protein family

	ion	ptga	Nat	7tm	dtd	photo	acr	msl	mean
AlignMeP	21.9	9.9	36.2	38.6	65.7	85.9	81.4	83.3	52.9
AlignMePS	31.2	49.9	63.1	64.7	75.2	85.7	86.0	83.0	67.3
AlignMePST	37.5	41.0	54.5	58.0	70.3	80.3	81.0	74.2	62.1
MSAProbs	10.5	29.0	49.8	53.5	63.2	68.8	85.9	65.9	53.3
HHalign	30.2	34.8	45.8	37.6	41.3	62.2	43.8	61.6	44.6
HMAP	20.1	39.2	58.9	60.2	64.3	81.3	83.9	75.5*	60.4
MUSCLE	14.3	29.8	53.7	55.7	69.4*	78.1	86.5	73.7*	57.7
MUSCLE profile-profile	13.7	23.2	45.7	47.4	67.0	69.3	85.7	66.1	52.3
ProbCons	9.5	26.6	48.6	53.2	62.5	65.0	85.8	62.9	51.7
T-Coffee	13.5	34.3	46.5	55.3	63.6	72.8	86.1	69.7	55.2
T-Coffee profile-profile	11.5	26.9	46.7	53.8	62.5	62.6	85.9	62.7	51.6
Number	775	716	1429	7046	1401	406	483	40	
Sequence identity (%)	4.3± 1.0	7.5± 1.6	10.9± 3.9	16.7± 5.4	16.8± 7.6	19.8± 5.5	20.4± 1.6	24.7± 3.3	

See legend to Table 3.6 for more details.

3.4 Discussion

In this work, I have applied and improved a sequence alignment method called AlignMe, for which I trained gap penalty sets in combination with specific input descriptors on a dataset of membrane protein structural homologues (HOME2). Three different strategies (AlignMeP, AlignMePS and AlignMePST) were assessed in comparison with other available alignment methods using HOME2 for an initial evaluation for alignment accuracy, homology modeling accuracy for a second evaluation and the independent BALiBASE membrane protein dataset (set 7) for a final evaluation. The results of the comparisons to the HOME2 data set were not surprising since the AlignMe parameters were optimized on that data set and consequently, all AlignMe modes were suitable for aligning (and modeling) proteins from the HOME2 data set. However, the independent analysis on the reference 7 subset from BALiBASE suggests that versions of AlignMe that match secondary structure prediction profiles may be generally useful for aligning membrane proteins (AlignMePS and AlignMePST; Table 3.6 - Table 3.12) and that membrane-specific information is suitable as additional information to an alignment if the pairwise sequence identity between two proteins decreases below 10 % (see Table 3.6 - Table 3.12). Overall, AlignMePS alignments are more accurate than those of the profile-to-profile methods HMAP and HHalign, both of which also use secondary structure information directly, indicating that the training of AlignMePS specifically on a membrane protein dataset was advantageous and that the discovered parameters can be generalized to proteins that are not included in the HOME2 data set.

For closely-related sequence pairs within BALiBASE, the usage of secondary structure and membrane-specific information (AlignMePS and AlignMePST) decreased the pairwise alignment accuracy compared to alignments based upon evolutionarily information only. Thus, I checked the matching of the transmembrane profiles that were used for AlignMePST alignments. Differences between the two membrane propensity profiles, which were generated by OCTOPUS, were calculated at every position in each alignment, summed up, normalized by calculating the total difference by the alignment length, and finally all differences were averaged over all HOME2 alignments. Indeed, the difference between the membrane-propensity profiles of two proteins is smaller (0.056) when using the transmembrane predictions in AlignMePST than without (in AlignMePS; 0.085), confirming that the predicted transmembrane profiles match more closely in AlignMePST alignments. The fact that transmembrane matching does not improve alignment accuracy for the closely-related BALiBASE sequence pairs may be caused by the error rate of >10 % inherent in the membrane propensity prediction by OCTOPUS (see chapter 3.2.2.4). Consequently, incorrectly predicted protein positions are matched in the AlignMePST mode for obtaining a low difference between the two membrane propensity profiles. Indeed, the matching of OCTOPUS

predictions in the reference structure-based alignments is almost as poor (profile difference score of 0.079) as the matching in the AlignMePS alignments showing that OCTOPUS profiles contain errors and that a perfect matching of them might be unfavorable. Such prediction errors can potentially be cancelled out in the context of a sequence alignment if the predictions for both sequences are incorrect in the same way, but this is not always the case, and the likelihood of errors canceling diminishes as the sequences diverge in similarity. An update of AlignMe with future membrane prediction methods that have an improved accuracy might solve or at least diminish that problem.

As mentioned above, another source of errors for the AlignMePST strategy (especially in the transmembrane regions) may be discrepancies between the secondary structure and transmembrane predictions. Quantifying the matching of secondary structure prediction profiles as described above indicates that the secondary structure profiles match less well in alignments generated with transmembrane predictions (profile difference score for AlignMePST is 0.060) than those without (profile difference score for AlignMePS is 0.055). In other words, transmembrane matching occurs at the expense of secondary structure matching.

A third possible cause of the reduced accuracy for closely-related sequences using AlignMePST is that including a third parameter (the transmembrane prediction besides evolutionarily information and secondary structure prediction) in the score for each position diminishes the contribution of the PSSM to the total alignment in a deleterious way.

The above discussion notwithstanding, the BALiBASE results indicate that incorporating transmembrane matching is useful for very distantly-related proteins (e.g., sequence identity < 15 %), particularly for reducing the overall shift error (Table 3.10 - Table 3.12). The observations for the accuracy in the transmembrane segments, however, are somewhat contradictory: the overlap of the known transmembrane regions in the HOME2 alignments is increased significantly by including transmembrane profiles (Table 3.4), whereas in the predicted transmembrane regions of the BALiBASE alignments there were fewer correctly aligned positions than with, e.g. T-Coffee (Table 3.9). Again, this may reflect conflicts between the secondary structure and transmembrane predictions, which might be addressed in future by adjusting the procedure so that secondary structure information is used only in regions not predicted to be in the membrane. However, such an approach might fail to align kinks in membrane segments properly since information about non-helical elements within membrane-spanning segments is missing. Unfortunately, I do not yet have sufficient data at low sequence identities to test this hypothesis more thoroughly and must await the availability of larger reference sets.

Table 3.9 Percentage of residues that are correctly aligned in the predicted transmembrane regions of pairwise sequence alignments from the BALiBASE reference set 7, sorted by protein family name

	7tm	acr	dtd	ion	msl	Nat	photo	ptga	mean
AlignMeP	54.6	96.0	76.5	36.1	96.7	44.6	91.8	40.3	67.1
AlignMePS	92.6	98.0	90.1	58.3	97.1	73.6	96.0	67.2	84.1
AlignMePST	87.0	95.6	86.2	57.8	95.7	64.2	93.9	58.1	79.8
MSAProbs	95.8	98.0	89.5	62.7	96.5	69.5	91.7	72.3	84.5
HHalign	51.9	37.6	51.8	37.1	76.3	50.0	71.6	31.5	51.0
HMAP	95.1	97.6	82.8	61.5	96.0*	72.4	96.7	69.3	83.9
MUSCLE	89.5	97.6	89.1	49.7	95.0*	64.9	91.7	57.2	79.3
MUSCLE profile-profile	79.9	97.4	89.0	30.2	92.9	53.9	85.8	47.6	72.1
ProbCons	95.7	97.9	89.6	61.6	96.5	67.9	90.6	69.8	83.7
T-Coffee	95.8	98.1	89.9	65.7	96.4	66.5	88.2	69.8	83.8
T-Coffee profile-profile	75.5	98.0	83.9	12.3	91.2	18.2	71.8	49.0	62.5

Mean = mean over averages for eight families. See legend for Table 3.3 for further details.

A concern about the current study is the fact that no structural information was available to aid with the alignments when the BALiBASE reference set 7 was constructed, and therefore it is possible that these alignments contain errors whose effect I cannot yet know (Edgar, 2010). Likewise, there might also be errors in the HOMEP2 alignments that were generated by a single structural alignment method (SKA), which generates a rigid superimposition of the protein structures. The accuracy of structural alignments for flexible protein segments was shown to be higher if a fragment-based method like FR-TM-align was used (Pandit and Skolnick, 2008). Therefore, I would recommend using a flexible structural alignment method for generating an updated reference data set.

Nevertheless, the relatively consistent ranking of the different methods on both the BALiBASE and HOMEP2 sets, i.e., with AlignMePS, MSAProbs and HMAP frequently high-ranking, and the profile-profile MSA methods ranked towards the bottom, suggests that my findings are reasonably robust.

Of the other available sequence alignment methods tested, alignments of the profile-to-profile alignment method HMAP were most frequently ranked towards the top (Table 3.3 - Table 3.12), and T-Coffee and MSAProbs alignments were also frequently very accurate, particularly in the transmembrane regions of the BALiBASE set (Table 3.9). Recently, MSAProbs and ProbCons were tested on this same BALiBASE reference 7 set (Chang, et al., 2012); however, in that study, they were assessed for their ability to construct MSAs rather than pair-wise alignments, which are in the focus here. It should also be reiterated that when testing the MSA methods on BALiBASE, I did not

construct a single MSA containing only the BALiBASE sequences, but rather, for each pair of sequences, I additionally included 24 homologues of those sequences identified by PSI-BLAST and clustered by UCLUST, in order to make the results comparable to those of AlignMe, HAlign and HMAP (see chapter 3.2.5). A consequence of this approach was that TM-Coffee, a slower method also shown to perform well for MSA of BALiBASE set 7 (Chang, et al., 2012), was too computationally expensive to be tested in the current study.

The profile-to-profile alignments strategy used with MUSCLE and T-Coffee typically resulted in fewer correctly-aligned positions and larger shifts compared to alignments of the other methods tested (Table 3.3 - Table 3.12). Also profile-to-profile alignments of HAlign for the BALiBASE set had surprisingly low fractions of correctly-aligned positions (Table 3.6 - Table 3.9), although the shift errors in the alignments for this method were among the smallest (Table 3.10 - Table 3.12) and the scores on the low sequence-identity ion family were consistently high-ranking (Table 3.6 - Table 3.8 and Table 3.10 - Table 3.12). This low performance of the profile-profile methods may reflect that greater deviations are present in the two profiles than in the sequences themselves, making them more difficult to align. Since the selection of sequence homologues appears to be an important parameter (Hill and Deane, 2012), in future work I plan to analyze the influence of the database search parameters on the accuracy of the different alignment methods, and to test not only evolutionarily information generated by PSI-BLAST but also those generated by programs such as SHRIMP (Bernsel, et al., 2008), HMMER3 (Eddy, 2011), and HHblits (Remmert, et al., 2011).

Finally, I have to point out that this study focuses on α -helical membrane protein sequences, so that gap penalties were obtained that are optimal for long helices and are not biased by the inclusion of short β -stranded regions (Hill and Deane, 2012). Optimization against β -barrel proteins is likely to lead to different gap penalty sets, and may result in methods that are particularly useful for that membrane protein architecture. And as the size of the database of (α -helical and β -barrel-like) membrane protein structures grows, further assessment of pairwise and multiple sequence alignment methods will be useful. Nevertheless, the results presented here suggest that there is potential for using the specific properties of membrane proteins for training and design in a way that aids the alignment of their sequences.

Table 3.10 Average shift error in pairwise alignments of the BALiBASE reference set 7

	ion	Nat	ptga	7tm	dtd	acr	photo	msl	mean
AlignMeP	29.92	48.71	33.98	47.58	9.83	1.09	0.31*	0.59*	15.38
AlignMePS	28.83	2.46	3.12	3.67	1.71	0.33	0.36	0.42	5.11
AlignMePST	13.83	3.24	5.39	11.82	3.46	0.42	0.31	0.47	4.87
MSAProbs	37.00	2.42*	5.99	5.17	4.29	0.34	1.36	0.84	6.87
HHalign	15.89	4.81	7.96	9.91	6.37	1.61	0.84	1.78	6.15
HMAP	35.66	1.95	6.18	4.61	6.84	0.31	0.52	0.58	7.08
MUSCLE	49.39	6.01	12.97	10.42	3.31	0.34	0.73	0.64	10.48
MUSCLE profile-profile	57.33	11.53	18.23	22.06	3.86	0.40	1.28	1.20	14.49
ProbCons	41.46	3.20	7.91	5.60	4.78	0.35*	1.70	1.09	8.22
T-Coffee	39.93	4.62	6.69	4.50	4.73	0.35*	1.60	1.09	7.90
T-Coffee profile-profile	64.15	42.50	12.03	17.50	8.48	0.45	2.15	2.22	18.69

Families are sorted by the average sequence identity (see Table 3.6). Mean = mean over averages for eight families. See legend for Table 3.3 for further details.

Table 3.11 Average shift error in pairwise alignments assigned to the same subgroup within the BALiBASE reference set 7

	ion	ptga	7tm	Nat	acr	msl	dtd	photo	mean
AlignMeP	12.35	0.79	16.19	1.45*	0.16*	0.72	0.62	0.18	4.06
AlignMePS	6.69	0.73	2.38	1.35	0.16	0.46*	0.58*	0.20*	1.57
AlignMePST	5.57	0.65	8.44	1.45	0.16	0.44	1.09	0.16	2.24
MSAProbs	21.91	2.97	3.90	1.85	0.19	0.70	1.25	0.48	4.16
HHalign	6.03	3.14	8.56	2.37	1.32	1.94	2.66	0.29	3.29
HMAP	17.91	2.03	2.93	1.40	0.20	0.55*	3.96	0.26	3.66
MUSCLE	17.67	5.73	9.13	3.56	0.19	0.63	0.99	0.37	4.78
MUSCLE profile-profile	42.37	8.06	15.17	10.81	0.23	1.31	2.36	1.01	10.16
ProbCons	23.82	3.98	4.40	2.44	0.22	1.01	1.55	0.65	4.76
T-Coffee	19.90	1.98	3.42	2.23	0.21	0.58	1.08	0.56	3.74
T-Coffee profile-profile	23.86	3.11	3.62	2.66	0.22	0.79	1.13	0.62	4.50

Families are sorted by the average sequence identity (see Table 3.7). Mean = mean over averages for eight families. See legend for Table 3.3 for further details.

Table 3.12 Average shift error in pairwise alignments assigned to different subgroups within the BALiBASE reference set 7

	ion	ptga	Nat	7tm	dtd	photo	acr	msl	mean
AlignMeP	42.41	59.90	58.04	52.40	10.38	0.35	1.90	0.43	28.23
AlignMePS	44.56	4.99	2.67	3.87	1.78	0.40	0.48	0.35	7.39
AlignMePST	19.71	9.10	3.60	12.34	3.61	0.35	0.65	0.50	6.23
MSAProbs	47.73	8.35	2.53	5.37	4.47	1.62	0.47	1.01	8.94
HHalign	22.90	11.73	5.29	10.12	6.60	1.01	1.86	1.56	7.63
HMAP	48.28	9.43	2.06	4.87	7.01	0.60	0.41	0.60*	9.16
MUSCLE	71.94	18.63	6.49	10.61	3.45	0.83	0.48	0.65	14.13
MUSCLE profile-profile	67.96	26.17	11.67	23.11	3.95	1.36	0.55	1.05	16.98
ProbCons	54.01	10.98	3.35	5.81	4.97	2.01	0.46	1.20	10.35
T-Coffee	34.91	5.23	4.90	4.32	4.05	1.67	0.44	0.75	7.03
T-Coffee profile-profile	51.36	9.48	5.01	4.62	4.95	1.90	0.47	1.46	9.91

Families are sorted by the average sequence identity (see Table 3.8). Mean = mean over averages for eight families. See legend for Table 3.3 for further details.

4 A Web Server for Aligning Membrane Protein Sequences

4.1 Introduction

The majority of frequently used computational programs for biologists and chemists are available via a web server that enables easy access to these programs. Web servers do not require the user to have specific hardware or to install specific software locally. All users will receive the same results of their input queries and will always use the latest version of the software including recent bug fixes and improvements of the software. Examples of prominent web servers that are commonly used for the analysis of membrane protein sequences are those of database search methods like PSI-BLAST or HHblits, secondary structure prediction methods like PSIPRED or membrane prediction methods like MEMSAT-SVM or OCTOPUS (see Table 4.1). Other servers that allow for the alignment of multiple protein sequences like MUSCLE, T-Coffee, ClustalW or the more recent version Clustal Ω (Sievers, et al., 2011) are also often used for membrane protein sequences although they were not developed with membrane proteins in mind (see Table 4.3).

In fact, only a few sequence alignment programs have been developed and/or optimized for the alignment of pairs of, or multiple, membrane protein sequences yet (see chapter 3). Consequently, I created a website for two different modes of AlignMe to provide an accessible interface to the AlignMe software with the intention that more users will use this program than if it was only available as a standalone installation package. The interface allows for two different types of alignments (see Figures 4.1 and 4.4). In the first mode, accurate pair-wise (PW) sequence alignments can be generated, such as those required for comparative modeling. The user has to provide two sequences in the standard fasta format as input and AlignMe will combine information about each sequence from multiple sources, producing a pairwise (PW) alignment. The second mode allows for an alignment of two sets of sequence homologues, by a comparison of their family-averaged hydrophobicity profiles (HP); this mode is based on the methodology of Lolkema and Slotboom (Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998), which has been shown to be useful for analysis of transmembrane topologies (Fenollar-Ferrer, et al., 2014; Khafizov, et al., 2010; Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998). These two different modes of AlignMe and their implementation into a website will be described in this chapter. This work was also published in the 2014 edition of the Nucleic Acid Research annual web server issue (Stamm, et al., 2014).

Table 4.1. Overview of commonly used webservers for analyzing (membrane) protein sequences

Type of method	Name	URL
Database search method	PSI-BLAST	http://www.ebi.ac.uk/Tools/sss/psiblast/
	HHblits	http://toolkit.tuebingen.mpg.de/hhblits/
Membrane propensity predictor	OCTOPUS	http://octopus.cbr.su.se/
	MEMSAT-SVM	http://bioinf.cs.ucl.ac.uk/psipred/?memsatsvm=1
Secondary structure propensity predictor	PSIPRED 3.2	http://bioinf.cs.ucl.ac.uk/psipred/

4.2 The AlignMe PW Sequence-to-Sequence Alignment Mode

AlignMe uses the standard Needleman-Wunsch algorithm in serial C/C++ code for pairwise alignments, and the only required input is two protein sequences in standard fasta format. For the PW sequence alignment mode, four different optimized parameter sets are provided that have each been shown to be suited for aligning sequences at a specific similarity level (see chapter 3). These default settings utilize different types of inputs alone or in combination with each other: (position-specific) substitution matrices; secondary structure predictions; and transmembrane propensities in the form of transmembrane predictions or hydrophobicity scales. AlignMe has also been designed to be flexible in handling other input descriptors reflecting protein properties for describing similarity between two proteins. Such similarity measures will then be used to guide the pairwise alignment. The web server provides a high level of flexibility so that the user may customize the inputs and alignment parameters.

4.2.1 Standard Parameter Sets

For standard usage, the web server provides four optimized sets of gap penalties and input parameters that result in accurate alignments dependent on the similarity between the two protein sequences of interest:

- 1) AlignMePST: This mode is useful for aligning distantly related proteins, with a sequence identity <15 % (see chapter 3.3). Based upon the two input sequences, a set of inputs is generated by the server that is then used for the alignment process: a Position-Specific Substitution Matrix (P) based upon a PSI-BLAST search on the UniRef90 database (see chapter 3.2), a secondary structure prediction (S) of PSIPRED 3.2 and a transmembrane prediction (T) from OCTOPUS. An alignment with these predictions typically takes minutes, with PSI-BLAST being the most time-consuming step. The duration of the alignment is dependent on the length of the protein sequences that are aligned.

- 2) AlignMePS: This mode is recommended for aligning low-homology proteins (~15-45% sequence identity) (see chapter 3.3). This version is similar to AlignMePST but omits the membrane prediction.
- 3) AlignMeP: With this mode, closely related proteins (>45%) can be aligned accurately (see chapter 3.3). This approach only considers sequence information since it uses only the PSSM with none of the structure predictions.
- 4) Fast: The fast mode is useful for situations in which a quick response is required, i.e. for detecting if two proteins could be related to each other. For this mode, the time-consuming PSI-BLAST search is omitted and thus, alignments are generated in less than 5 seconds. However, these alignments are less accurate because they are based only upon a general substitution matrix (VTML (Müller, et al., 2002; Müller and Vingron, 2000)) combined with a hydrophobicity scale (HWvH (Hessa, et al., 2005)). This combination was the most accurate of the fast strategies tested (i.e., of the combinations that do not require results from PSI-BLAST search) (see chapter 3.3).

The screenshot displays the AlignMe website interface. At the top, there is a navigation bar with the mpibp logo, the title 'AlignMe', and a decorative graphic of a signal waveform with amino acid labels (A, L, I, G, N, M, E). Below the navigation bar, there are four tabs: 'AlignMe Home', 'Sequence to Sequence Alignment' (which is selected), 'Alignment of two Multiple Sequence Alignments', and 'AlignMe FAQ'.

The main content area is divided into three sections:

- 1) Sequences**: This section contains two input fields for 'Enter a sequence in fasta format' and 'Enter another sequence in fasta format'. Below each field is a 'Choose File' button and the text 'no file selected'.
- 2) Usage of own or optimized predefined parameters**: This section is titled 'α-helical membrane proteins' and lists four predefined modes:
 - AlignMe PST**: Most accurate alignments for very distantly related proteins (<15% identity) (~9 mins for a 179 and 215-residue sequence pair)
 - AlignMe PS**: Most accurate alignments for low-homology proteins (~15-45% identity) (~5 mins for a 179 and 215-residue sequence pair)
 - AlignMe P**: Most accurate alignments for very closely related proteins (>45% identity) (~5 mins for a 179 and 215-residue sequence pair)
 - Fast, but less accurate alignments (~3 sec.)**: based on a substitution matrix and a hydrophobicity scale
 Below these options is a link for 'Usage of own alignment parameters' and a 'user defined parameters' checkbox. At the bottom of this section is a green arrow icon and the text 'Show detailed alignment parameters!'.
- 3) Submission**: This section prompts the user to 'Enter an e-mail address if you wish to receive an e-mail with your result' and 'Recommended for "Most accurate alignments" and alignments with automated PSSM calculations, PsiPred or OCTOPUS predictions'. It includes an email input field and two sets of 'Submit' and 'Reset' buttons.

Figure 4.1 Screenshot of the AlignMe website (Jan, 2015) showing the option to select 4 different predefined modes for aligning α -helical membrane proteins.

4.2.2 Available Input Descriptors

Aside from the described default parameter sets, the web server allows also the usage of the same input parameters that can be used with the local version of AlignMe. The web users can upload their own alignment parameters (e.g., custom substitution matrices, hydrophobicity scales or predictions) or can choose between three types of input descriptors that are provided by the server for the alignment: (position-specific) substitution matrices; hydrophobicity scales; or per-residue profiles, such as transmembrane predictions (Table 4.2). In addition, hydrophobicity scales can be either used similar to a substitution matrix with a single-residue specific substitution rate or they can be window-averaged in different ways (rectangular, triangular, zigzag, sinusoidal) over any length of a window. There are no limits on the number of matrices, scales or profiles that can be combined. However, the different input parameters should be weighted according to the range of values within that scale to prevent bias (see chapter 3.1); details are provided in the user manual.

Table 4.2 Overview of input parameters available as ‘user-defined parameters’ on the AlignMe web server

Substitution matrices	Hydrophobicity scales	(Predicted) profiles
Custom matrix	Custom scale	Custom profile
BLOSUM62	Eisenberg & Weiss	PSIPRED 3.2 prediction
PHAT	Hessa, White & von Heinje	OCTOPUS prediction
SLIM	Kyte & Doolittle	
VTML	Wimley & White	
PSSM from PSIPRED		

For details see chapter 3.2.2.

4.2.3 Outputs of the AlignMe Web Server

The output from the PW sequence-to-sequence mode includes the pair-wise alignment of the two amino-acid sequences provided in ClustalW format, the corresponding sequence identity, and the percentage of matched positions. For each prediction (e.g., membrane or secondary structure prediction) or hydrophobicity scale used for the alignment process, also a plot (generated by gnuplot) is shown. This plot illustrates per-residue profiles and thereby provides a simple representation of the similarity between the two proteins that are aligned (see Figure 4.2). Aside from this visual representation of the alignment, the table of hydrophobicity and/or prediction values for each alignment position is also displayed at the bottom of the results page, allowing the user to use these values for representing the data in different custom formats locally. Finally, a summary of the input parameters (e.g., weights and types of inputs) that were used for the alignment is also provided. All the output files can be downloaded separately or together as a single (archive) file. Results are stored for 14 days on the server and can be retrieved using a Job Identifier (ID), which is provided on the results page.

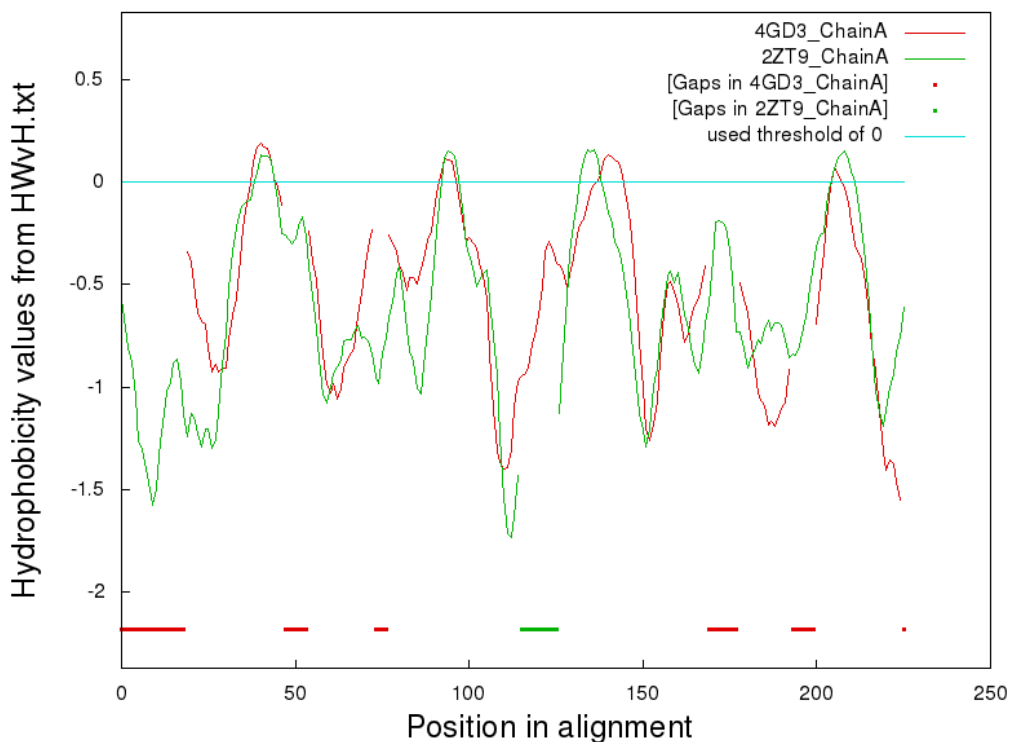


Figure 4.2 Screenshot of the results page of the AlignMe website (Jan, 2015) showing the aligned hydrophobicity profiles based upon an alignment using the example sequences (two proteins of the cytochrome b_6f family from HOME2, PDB codes: 4GD3 chain A and 2ZT9 chain A) with the alignment parameters from the fast mode of AlignMe. Gaps are shown as bars in corresponding colors underneath the profile.

4.3 Other Online Servers for the Alignment of Membrane Protein Sequences

Since the development of AlignMe in 2010/11 and the evaluation of the alignment accuracy of AlignMe in 2011/12, several other alignment programs were developed whose accuracy has not yet been analyzed as described in chapter 3. Among these programs is MP-T, which uses environment-specific substitution matrices for generating sequence-to-structure alignments (Hill and Deane, 2012). MP-T is limited to alignments for which at least one structure of the two proteins of interest is known and therefore is not applicable for a pairwise alignment that is only based on sequence information. As yet, MP-T cannot directly be used and evaluated via a webserver, but it is implemented within an online membrane protein homology modeling workflow called Memoir (Ebejer, et al., 2013)

Two other programs specifically designed with membrane proteins in mind are PRALINETM (Pirovano, et al., 2008) and TM-Coffee (Chang, et al., 2012) (see Table 4.3). Neither program could be compared with AlignMe previously, as they are available exclusively as webservers and so could not be tested on the large HOME2 and BALiBASE reference 7 data sets (see chapter 3.3.3 and 3.3.5). Also a number of sequence alignment web servers not designed specifically for membrane proteins were developed recently, including Clustal Ω (Sievers, et al., 2011), PicXAA (Sahraeian and Yoon, 2010; Sahraeian and Yoon, 2011) and PSI-Coffee (Chang, et al., 2012) (see Table 4.3). For all these programs, it is not yet clear whether any of them is able to generate more accurate pairwise alignments than AlignMe for membrane protein sequences. Therefore, I will also demonstrate in this chapter that the accuracy of alignments obtained with the AlignMe webserver is comparable or better than those of alignments generated with the following currently-available webservers: Clustal Ω , PicXAA, PRALINETM, ProbCons, PSI-Coffee, TM-Coffee.

Table 4.3 Overview of recent webservers for aligning (membrane) protein sequences

Type of method	Name	URL
Pairwise sequence alignment of membrane proteins	AlignMe	http://www.bioinfo.mpg.de/AlignMe/
Multiple sequence alignment of membrane proteins	PRALINE TM	http://www.ibi.vu.nl/programs/pralinewww/
	TM-Coffee	http://tcoffee.crg.cat/apps/tcoffee/do:tmcoffee
Multiple sequence alignment methods optimized on general protein data sets	MUSCLE	http://www.drive5.com/muscle/
	T-Coffee	http://www.tcoffee.org
	PSI-Coffee	http://tcoffee.crg.cat/apps/tcoffee/do:psicoffee
	ClustalW	http://www.clustal.org/clustal2/
	Clustal Ω	http://www.clustal.org/omega/
	PicXAA	http://gsp.tamu.edu/picxaa/

4.4 Alignment Accuracy of the AlignMe PW Mode Compared to Other Web Servers

As described before in chapter 3.3.5, AlignMe performed well (Stamm, et al., 2013) compared to other sequence alignment programs such as MSAProbs (Liu, et al., 2010) or HMAP (Tang, et al., 2003) on the BALIBASE reference 7 set of membrane proteins (Bahr, et al., 2001; Thompson, et al., 1999). However, neither MSAProbs nor HMAP are available as web servers whereas other alignment programs of interest were exclusively available as web servers (see chapter 4.3). In this chapter, I compare the accuracy of alignments based on the AlignMe PW sequence-to-sequence mode to those from other available alignment web servers. These methods include ProbCons, the third most accurate alignment method on the BALIBASE set 7 (Stamm, et al., 2013), Clustal Ω , PSI-Coffee, TM-Coffee, PicXAA and PRALINETM. For each of these multiple sequence alignment methods, all sequences of a family were submitted as one single input set of sequences and aligned with the default parameters provided on the website of each server to obtain one multiple sequence alignment (MSA) per family, for a total of 8 multiple-sequence alignments. From these MSAs, pairwise sequence alignments were extracted and accuracy scores were then calculated for each pair of sequence within a family (Tables 4.4 and 4.5)

Table 4.4 Percentage of residues aligned correctly in pairwise sequence alignments from the BALIBASE reference set 7, sorted by sequence identity of the protein families

	ion	Nat	ptga	7tm	dtd	acr	photo	msl	mean
AlignMeP	38.90	43.50	42.10	42.50	67.10	87.00	87.90	82.50	61.4
AlignMePS	45.20	66.20	64.80	65.90	76.00	89.70	87.60	82.30*	72.2
AlignMePST	48.10	58.60	58.78	59.40	71.20	86.30	82.90	76.50	67.7
AlignMe fast	40.94	53.33	54.86	60.93	66.67	82.13	79.94	77.45	64.5
Clustal Ω	45.75	64.26	63.54	65.17	74.89	88.89	88.78	81.49	71.6
PicXAA	36.96	61.27	59.48	60.25	68.36	88.74	81.06	80.31	67.1
PRALINE TM	31.98	56.94	63.19	61.60	73.07	87.74	81.39	78.33	66.8
ProbCons	26.32	52.84	45.43	54.37	63.41	85.88	77.90	74.25	60.0
PSI-Coffee	27.11	51.01	47.97	57.09	64.82	89.19	77.43	73.80	61.1
TM-Coffee	25.84	49.65	47.38	55.25	65.42	88.44	76.39	68.40	59.6
Number ^a	1326	1711	1275	8128	1485	903	528	91	
Sequence identity (%) ^b	11.7± 13.8	14.3± 10.8	15.9± 12.1	18.2± 9.7	18.7± 11.5	26.9± 11.3	27.3±1 6.9	35.3± 13.5	

Entries in bold in all tables, indicate the highest or best scores in that column. *Values marked with an asterisk in all tables are not significantly different from the highest/best score in a column according to a pairwise Wilcoxon signed rank test. Mean = mean percentage of correctly-aligned residues over averages for eight families. ^aNumber of pair-wise alignments. ^bMean (\pm standard deviation) of the percentage sequence identity between pairs of alignments in each family.

In general, the different modes of AlignMe resulted in the most accurate alignments, with significantly higher fractions of correctly aligned residues (Table 4.4) and low average shift errors (Table 4.5), although the alignments of Clustal Ω also had low average shift errors for several families.

The alignment accuracy obtained using a given AlignMe mode depends on the sequence similarity of the proteins being aligned; the results shown here are consistent with the ranges described in chapter 3.3.5. For all the sequence identity ranges mentioned, the percentage of correctly aligned residues is significantly higher using the corresponding AlignMe version than any other program tested. The “fast” version of AlignMe is generally less accurate than the three slower versions but it appears to provide a compromise between the PST, PS and P parameter sets. On average, the fast mode is still ranked 6th among 10 approaches, suggesting that it can provide a useful first pass approach, for example, to approximate the sequence identity. Among the other web servers tested, alignments generated by Clustal Ω contained the highest proportion of correctly aligned residues and the smallest shift errors, followed by PicXAA and PRALINETM.

Table 4.5 Average shift error in pairwise alignments of the BALiBASE reference set 7

	ion	Nat	ptga	7tm	dtd	acr	photo	msl	mean
AlignMeP	29.92	48.71	33.98	47.58	9.83	1.09*	0.31	0.59	21.50
AlignMePS	28.83	2.46	3.12	3.67	1.71	0.33	0.36*	0.42*	5.11
AlignMePST	13.83	3.24	5.39*	11.82	3.46	0.42	0.31	0.47	4.87
AlignMe fast	28.18	4.21	10.10	4.27	4.14	0.84	0.58	0.71	6.63
Clustal Ω	20.77	2.71	2.87	3.63	3.87	0.38	0.61*	0.40	4.40
PicXAA	28.73	3.18	4.36	4.90	5.12	0.37	0.86	0.43	5.99
PRALINE TM	64.03	7.89	3.01	4.27	7.28	0.95*	0.89	0.43	11.09
ProbCons	36.99	2.86	5.97	5.05	10.87	0.37	1.30	0.56	8.00
PSI-Coffee	27.01	4.30	5.15	4.05	5.81	0.32	1.18	0.70	6.06
TM-Coffee	27.66	5.18	6.21	4.61	8.95	0.38	1.39	0.70	6.89
Number ^a	1326	1711	1275	8128	1485	903	528	91	
Sequence identity (%) ^b	11.7±	14.3±	15.9±	18.2±	18.7±	26.9±	27.3±1	35.3±	
	13.8	10.8	12.1	9.7	11.5	11.3	6.9	13.5	

The shift error is calculated as the number of positions by which a given residue is misaligned summed over the length of the alignment and averaged over all alignments. Families are sorted by the average sequence identity. Mean = mean over averages for eight families. See legend for Table 4.4 for more details.

4.5 Alignment of Family-Averaged Hydropathy Profiles (HP) Using two Multiple Sequence Alignments

Another mode that is provided on the AlignMe website for aligning membrane protein sequences is the HP mode that allows for the alignment of family-averaged hydropathy profiles (HP) which has been shown to be useful for detecting evolutionary relationships between distantly related membrane protein sequences (Fenollar-Ferrer, et al., 2014; Khafizov, et al., 2010). The shape of a hydrophobicity profile is based on the transmembrane topology of a membrane protein with strong peaks in the most hydrophobic transmembrane segments. These hydrophobic transmembrane helices are generally conserved during evolution and thus, corresponding averaged hydropathy profiles among a protein family can contain similar global features even in very distantly related proteins (see Figure 4.3). Despite lacking detailed position-specific information and a significance score for similarity, a comparison of hydropathy profiles (HPs) can provide an intuitive overview of the similarities between the transmembrane topologies of two proteins (Khafizov, et al., 2010; Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998). Averaging each of the input profiles over a set of sequence homologues, in a so-called family-averaged HP, can smooth out noise and sequence-specific detail, making comparisons much clearer (Khafizov, et al., 2010; Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998). To date, the ability to generate these aligned HPs has not been readily available to the community. The AlignMe web server provides a simple interface to such alignments in the profile-to-profile alignment mode (see Figure 4.4).

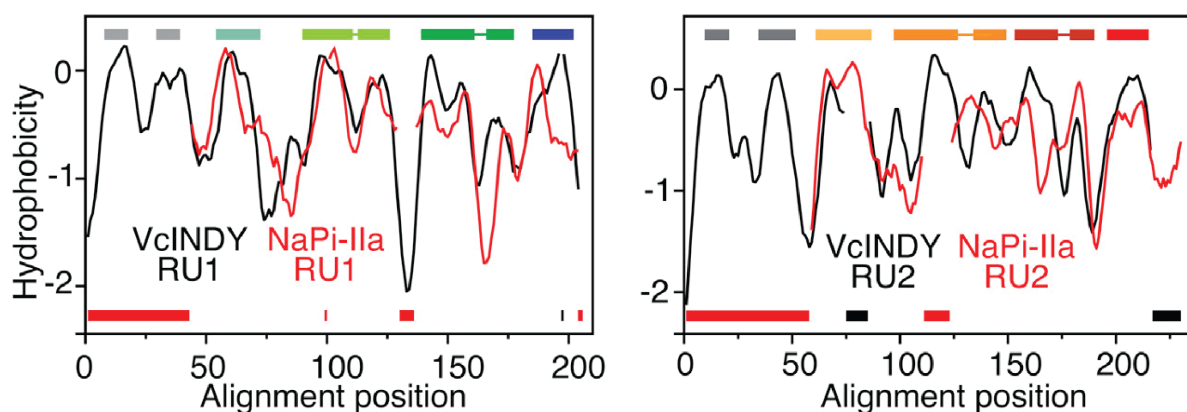


Figure 4.3 Hydrophobicity profiles based on the alignment of the family-averaged hydropathy profiles (HP) of the two repeat units (repeat unit 1, left and repeat unit 2, right) in NaPi-II (red line) and VcINDY (black line) using AlignMe showed that they share a common fold despite their low sequence identity (7% similarity of NaPi-IIa and VcINDY based on a AlignMePS alignment) (Fenollar-Ferrer, et al., 2014). Figure taken from (Fenollar-Ferrer, et al., 2014).

4.5.1 Inputs for the HP Mode

The alignment of family-averaged hydropathy profiles (HP) requires two reference sequences of interest as well as corresponding sequence homologs that are already aligned to those reference sequences in form of a multiple sequence alignment. In a first step, the user would typically carry out a database search for sequence homologues of each query protein (e.g., with PSI-BLAST) if no sequence homologs are previously known, and in the next step align each set of homologues with a multiple-sequence alignment program, such as Clustal Ω . If homologous sequences are already known then the user can also use those for generating a MSA. Both steps are independent of the AlignMe server and the user has the free choice of using any method for generating MSAs. The only requirement of the HP mode is that all protein sequences (including gaps) in a MSA have the same length. After submitting the two MSAs to the HP mode of AlignMe server an alignment will be generated based upon a Needleman-Wunsch algorithm and a hydrophobicity scale. In general, a single sequence may also be provided but the lack of homologous sequences is likely to result in a less accurate alignment. For pairwise sequence alignments, the PW mode of the AlignMe server is the better choice.

4.5.2 Standard Parameter Sets

The pre-defined default parameter set that is provided on the web server for HP alignments is based upon a previous study for the identification of five transmembrane helix domains using HP alignments (Khafizov, et al., 2010). In this study, the protein sequences were aligned using profiles based on values from the HWvH hydrophobicity scale (Hessa, et al., 2005) that are smoothed using a 13-residue long, triangular sliding window. Using these hydropathy profiles, a systematic optimization procedure of the gap penalties was executed in a first step in order to detect the best combination of gap penalties for a search of homologous five transmembrane helix domains. The ideal gap penalties were assumed to be the ones that ranked hydropathy profile alignments of known five transmembrane sequences with each other higher than alignments of these protein sequences with random sequences. In a second step, the obtained gap penalties were slightly modified in order to align accurately family-averaged hydropathy profiles of sequences from the FIRL fold. The value for the terminal gap penalties was decreased because terminal gaps occur frequently in alignments of sequences that differ in length or number of membrane helices. The gap penalties that were used are as follows: $p_o^{above} = 2.5$, $p_e^{above} = 1.0$, $p_o^{below} = 1.0$, $p_e^{below} = 0.85$, $p_o^{terminal} = 0.25$ and $p_e^{terminal} = 0.25$, where the hydrophobicity threshold used to assign “above” and “below” was -0.5 (see Experimental Procedures section in (Khafizov, et al., 2010)). These gap

penalties differ from those of the AlignMe PW modes (see chapter 3.3.2) because a different input was used to describe similarity between the protein sequences.

An additional parameter for HP alignments is the fraction of allowed gaps that checks for the percentage of gaps within a specific column of the MSA (e.g., a fraction of allowed gaps of 0.8 means at least 80% of positions within that column have an amino acid and at the most 20 % have a gap). Only columns that have more amino acids than defined by the threshold of the fraction of allowed gaps are considered for the alignment process. All other columns are discarded and not considered anymore. With this parameter, low confidence columns in the input MSA that contain a high number of gaps can be omitted from the alignment process. This value can vary between 1 (a column contains no gaps) and a value close to 0 (a column contains only one amino acid aligned to gaps in all other homologous sequences).

With the gap penalties and the value for the fraction of allowed gaps, AlignMe calculates the average hydrophobicity value for every confident position (i.e. column) in the MSA, resulting in a family-averaged hydropathy profile. However, if the family contains many insertions and deletions in a given column of the alignment (specified by the fraction of allowed gaps), then that column is not considered for the alignment and also not displayed in the resulting output plots.

Similar to the PW mode, the HP mode also allows users to apply different hydrophobicity scales (see Table 4.2) and/or custom gap penalties as well as their own values for the fraction of allowed gaps.

The screenshot shows the AlignMe website interface. At the top, there is a navigation bar with the logo 'mpibp' on the left, the title 'AlignMe' in the center, and a decorative graphic on the right consisting of a blue and green waveform with labels 'A A', 'L L', 'I I', 'G G', 'N N', 'M M', and 'E E' stacked vertically. Below the navigation bar are four tabs: 'AlignMe Home', 'Sequence to Sequence Alignment', 'Alignment of two Multiple Sequence Alignments' (which is the active tab), and 'AlignMe FAQ'.

Below the tabs is a 'Help/Examples:' section with a 'Generate example input' button and a link to the 'FAQ!'.

The main content area is divided into three sections:

- 1) Multiple Sequence Alignments**: This section contains two input fields for 'Enter a multiple sequence alignment with sequences in fasta format' and 'Enter another multiple sequence alignment with sequences in fasta format'. Each field has a 'Choose File' button and the text 'no file selected' below it.
- 2) Usage of own or optimized predefined parameters (optional)**: This section includes a link for ' α -helical membrane proteins', a checkbox for 'fast ~ 3 sec. (substitution matrix & hydrophobicity scale)', a link for 'Usage of own alignment parameters', a checkbox for 'user defined parameters', and a green arrow icon with the text 'Show detailed alignment parameters!'.
- 3) Submission**: This section contains a text input field for 'Enter a E-Mail address if you want to have an E-Mail with your results (not required)'. Below the input field are two pairs of 'Submit' and 'Reset' buttons.

Figure 4.4 Screenshot of the AlignMe website (Jan, 2015) showing the querylet for the HP mode of AlignMe

4.5.3 Outputs

The output of the HP mode of AlignMe consists of a sequence alignment in ClustalW format and a hydropathy plot. An alignment of the two reference sequences of both MSAs is shown in ClustalW format. Gaps in the alignment that were present in the original MSAs are represented by a “.” symbol, whereas gaps introduced during the alignment of the averaged hydropathy profiles are indicated with a “-” symbol. Based upon this alignment, a hydropathy plot is generated that displays the position-specific alignment of the underlying hydrophobicity of the reference sequences as in the fast alignment option of the PW mode (Figure 4.5). Similar to the PW mode, also in the HP mode, parameters and results can be downloaded separately or together in a single file.

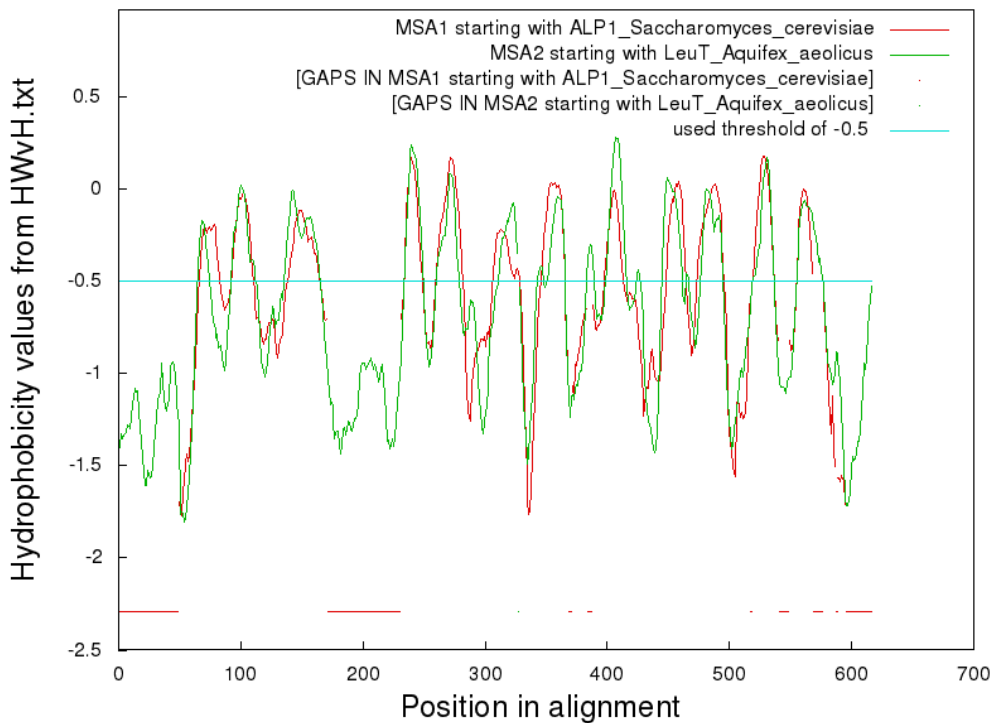


Figure 4.5 Screenshot of the results page of the AlignMe website (Jan, 2015) showing the aligned hydrophobicity profiles of the first sequences from each submitted multiple sequence alignment. The alignment was generated with the default parameters using the example multiple sequence alignments with the protein sequence of ALP1 from *Saccharomyces cerevisiae* as a representative of the first multiple sequence alignment and the protein sequence of LeuT from *aquifex aeolicus* as a representative of the second multiple sequence alignment.

4.5.4 Example Applications of HP alignments

Hydrophobicity profiles have been used in a number of studies to assess the topological similarity of two proteins with very low or undetectable sequence similarities. For example, evolutionary relationships have been illustrated between neurotransmitter:sodium symporters (NSS), sodium:solute symporters (SSS) and members of the amino-acid/polyamine/organocation (APC) superfamily (Lolkema and Slotboom, 2008). The crystal structures of the Na^+ /galactose symporter vSGLT (SSS family), the leucine transporter LeuT (NSS family) and the amino acid transporter AroP (APC family) were shown to share a common five-helix inverted repeat fold despite their low sequence identities among each other (<20 % sequence similarity). Members of those families were collected by BLAST searches and aligned using ClustalW. Based upon those alignments, an averaged 'family hydrophathy profile' was calculated using a hydrophobicity scale and a sliding window approach. Those profiles were then aligned using the MemGen alignment approach (Lolkema and Slotboom, 1998) with manually defined penalties for insertions of gaps (see Figure 4.6) (Lolkema and Slotboom, 2008). This idea was already applied previously for the classification of numerous secondary transporters (Dobrowolski, et al., 2007; Lolkema and Slotboom, 2003; Lolkema and Slotboom, 2005).

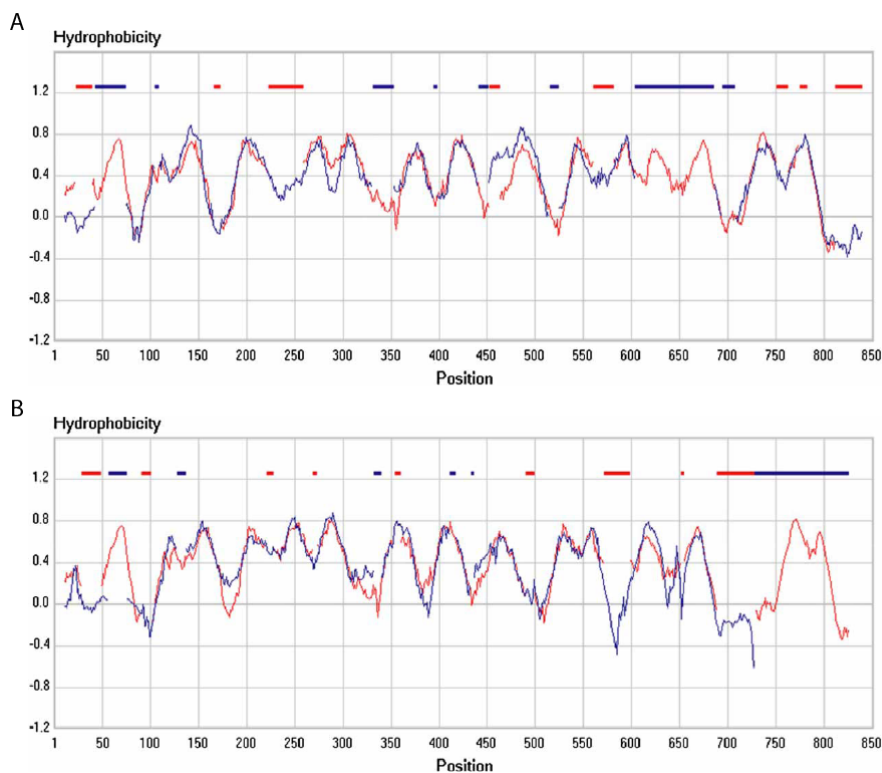


Figure 4.6 Family hydrophathy profile alignments (not generated by AlignMe). (A) Hydrophathy profile of the SSS family in red aligned to those of the NSS family in blue (B) Hydrophathy profile of the SSS family in red aligned to those of the APC family in blue. Gaps are shown as boxes in corresponding colors above the profile. This figure is taken from (Lolkema and Slotboom, 2008).

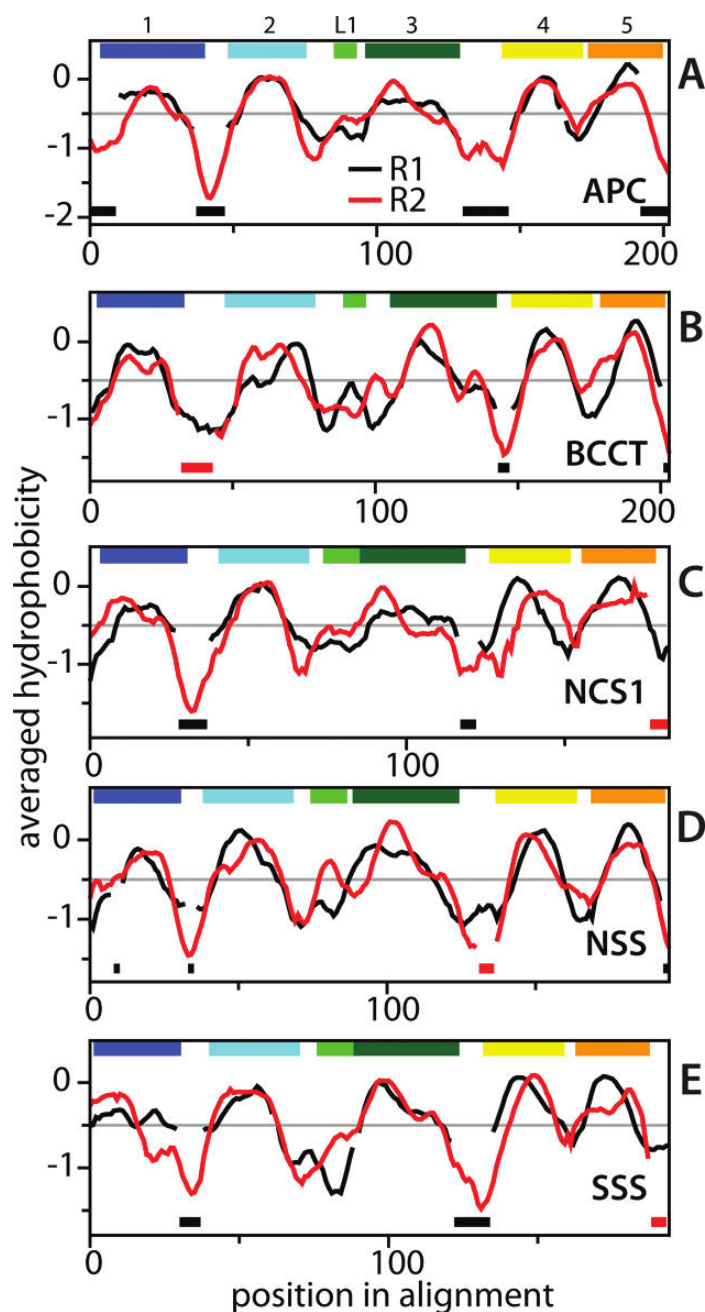


Figure 4.7 Family-averaged hydropathy profiles of the internal repeat units 1 (black) and 2 (red) for the following families: (A) APC, (B) BCCT, (C) NCS1, (D) NSS, (E) SSS. Gray horizontal lines represent the threshold above which a position is defined as membrane-spanning for the assignment of gaps. Gaps in the alignments are shown as bars underneath the profiles, in corresponding colors. This Figure is taken from (Khafizov, et al., 2010).

Hydropathy profiles were also used to compare and analyse internal structural repeats of similar transporters (see Figure 4.7) (Fenollar-Ferrer, et al., 2014; Khafizov, et al., 2010). Protein sequences for each family were again identified with a database search (i.e. PSI-BLAST) and aligned with a multiple sequence alignment method (i.e. MUSCLE). However, in this study, the profiles were aligned using the HP mode of AlignMe with the default parameters mentioned in chapter 4.5.2 (Khafizov, et al., 2010).

Furthermore, evolutionarily relationships have been shown between the 2-hydroxycarboxylate transporters (2HCT) and so-called ESS transporter families (Dobrowolski, et al., 2007; ter Horst and Lolkema, 2012); between a multidrug and toxin extrusion (MATE) transporter and the inner membrane flippase Wzx (Islam, et al., 2012); as well as between SLC34 and SLC13 families of transporters (Fenollar-Ferrer, et al., 2014) by using hydropathy profiles. The same basic approach was also used in a comparison of members of the sodium-phosphate transporter family NaPi-II (Forster, et al., 2002) and to identify a putative ancestral half-transporter (Khafizov, et al., 2010).

4.6 Conclusions

With these two modes, the AlignMe web server provides user-friendly access for the analysis and comparison of membrane protein sequences.

In the PW option, the user can readily compute pair-wise membrane protein sequence alignments suitable, e.g., for homology modeling. A comparison to recent web servers for aligning protein sequences showed that a good accuracy is achieved by using different PW alignment parameter sets depending on the sequence identity range of the proteins of interest (see Table 4.4, Table 4.5 and chapter 4.4), which agrees with the observations that were made for AlignMe beforehand (see chapter 3). For an estimation of the sequence identity or structural homology of the protein pairs, the user might first use the *fast mode* of AlignMe, or compare the hydropathy profiles by aligning two MSAs, respectively.

The second functionality provided by the AlignMe web server, namely HP alignments, allows for qualitative comparison of transmembrane topologies (and therefore potentially the 3D folds) based upon the hydrophobicity profiles of two multiple sequence alignments of membrane proteins, even for protein pairs with a low sequence similarity. This HP mode provides for the first time a user-friendly interface for the method originally developed by Lolkema and Slotboom (Lolkema and Slotboom, 1998; Lolkema and Slotboom, 1998).

The web server can be accessed at <http://www.bioinfo.mpg.de/AlignMe> and supports all major web browsers (Mozilla Firefox, Google Chrome, Internet Explorer, Edge, Safari). A login to the website is not required but an email address can be submitted if users wish to receive their alignment results via email. In addition, the AlignMe manual and Unix source code are available for download at <http://www.bioinfo.mpg.de/AlignMe/download/>.

5 Evaluation of Structural Alignment Methods on HOME3

5.1 Introduction

Structural alignment programs allow for a comparison of one or more proteins on a structural level by including spatial and secondary structure information. This enables structural alignment programs to be used for the detection of evolutionarily-related protein segments (i.e. repeats), for clustering and classifying large data sets of protein structures and for the generation of reference data sets containing structure as well as sequence alignments. An accurate detection and assignment of evolutionarily and homologous residues or fragments is based on the accuracy of the method being used and thus various structural alignment approaches that are available have been assessed by recent studies for their accuracy (Berbalk, et al., 2009; Kolodny, et al., 2005; Sadowski and Taylor, 2012; Slater, et al., 2012). However, these studies agree with the observation that there is no outstanding structural alignment program a user can rely on. Consequently, the authors advise to use several structural alignment programs, especially for low-similarity protein structures, in parallel and then evaluate the quality of the alignment by analyzing the structural alignment manually (Slater, et al., 2012), by assessing the underlying sequence alignment (Berbalk, et al., 2009) or by using geometrical match measures on the aligned structures (Kolodny, et al., 2005) to determine the most accurate alignment among a set of alignments. Membrane proteins were not explicitly considered in those studies or by any of the tested structure alignment methods although they consist of major fold classes (e.g., distinct SCOP superfamilies) that are all influenced by the hydrophobic surrounding of the membrane bilayer. Integral membrane proteins might therefore favor structural conformations and local interactions that are energetically discouraged in soluble proteins and vice versa. Within the group of membrane proteins, two major fold types are present: α -helical and β -barrel-like membrane proteins. Consequently, one cannot assume that structural alignment programs that perform well on general data sets of proteins are also accurate on the particular fold types that are unique to membrane proteins. All these observations address the problem that an analysis of the accuracy of structural alignment programs for their performance on both major folds of membrane proteins (e.g., α -helical and β -barrel-like) is required to identify reliable structural alignment methods for membrane proteins and to understand if there is still room for improvement in advanced methods (e.g., by using membrane specific information to guide the alignment process). For such an analysis, a recent membrane protein data set is required.

My evaluation of structural alignment methods is based upon the HOME3 data set including 40 α -helical and 8 β -barrel-like membrane protein families (see Chapter 2 for more details on HOME3). In

this chapter, the accuracy of 13 structural alignment methods (see Table 5.1) is assessed using this dataset by generating homology models based upon alignments from all methods. These models were then ranked using geometrical and physicochemical measures to ensure a comprehensive evaluation. Finally, the four best structural alignment methods are used in a consensus approach that allows an estimation of the confidence level of each alignment position.

5.2 Methods

5.2.1 Overview of Structural Alignment Methods Tested

In this study, 13 structural alignment methods differing in their superimposition approaches, internal scoring schemes, and handling of flexible regions were tested and compared with each other: CE (Shindyalov and Bourne, 1998), DaliLite (Holm and Park, 2000; Holm and Rosenstrom, 2010), FATCAT (Veeramalai, et al., 2008; Ye and Godzik, 2004), FR-TM-align (Pandit and Skolnick, 2008), LovoAlign (Martinez, et al., 2007), MAMMOTH (Ortiz, et al., 2002), MATT (Menke, et al., 2008), PPM (Csaba, et al., 2008), SABERTOOTH (Teichert, et al., 2007), SAP (Taylor, 1999; Taylor, 2000), SHEBA (Jung and Lee, 2000), SKA (Yang and Honig, 2000), and TM-align (Zhang and Skolnick, 2005). All these methods differ substantially in their superimposition approach (i.e. rigid vs. fragment-based), internal scoring scheme (RMSD vs. TMScore) or in their handling of flexible regions (see Table 5.1). For FATCAT, two different alignments were generated by either setting the flag “-flexible” to false for the rigid mode (FATCAT rigid) or to true for the flexible mode (FATCAT).

These methods were all available for installation, commonly used, and/or shown to out-perform other available methods. For each method, pairwise structure alignments were generated for all pairs of proteins within each family of HOME3. Because some methods produce different alignments depending on which protein is listed first, alignments were generated using both combinations of each pair of proteins.

Some methods provided structural alignments and the corresponding sequence alignments whereas other methods provided only a structural alignment. If a sequence alignment was provided then that alignment was used for subsequent analysis. Otherwise, the underlying sequence alignment was extracted from the structural alignment with a customized script that constructs a sequence alignment based on the assignment of matched residues in the structural alignment. Non-matched residues were aligned to gaps. For CE, DaliLite, FATCAT, MAMMOTH, PPM and SAP some C- and N-terminal residues were missing in the structural alignments. For those alignments, C- and N-terminals had to be added to the alignment by aligning them against gaps.

Based upon these alignments, homology models were then generated and finally evaluated using structural similarity scores that describe the quality of the model (see Figure 5.1).

Table 5.1 Overview of pairwise structural alignment methods

Method	^a Fragments	DP?	Score	Alignment
CE	8-residue	N	RMSD	Combinatorial extension of locally aligned fragment pairs based on intra-structural distances
DaliLite	6-residue	N	Dali	Joins optimally-matched fragments based on a Monte Carlo search
FATCAT	8-residue		S	Flexible chaining of aligned fragment pairs allowing for twists
FR-TM-align	Y	Y	TM-score	Matching aligned fragment pairs
LovoAlign		Y	STRUCTAL & RMSD	“Low Order Value Optimization” using dynamic programming
MAMMOTH			RMSD	Matching molecular models obtained from theory
MATT	5-8-residue		RMSD	Aligning fragment pairs allowing temporarily for twists and translations
PPM			PPM	Phenotypic plasticity applied to measure the cost of morphing structures
SABERTOOTH				Matching profiles of vectorial representations of two protein structures
SAP		Y	Intra RMSD	Iterated double dynamic programming of matrix of intra-structure residue-residue distance differences
SHEBA				Comparing a list of primary, secondary and tertiary structural profiles
SKA		Y	PSD	Double dynamic programming to align secondary structure alignments
TM-align		Y	TM-score	Optimize intra-structure residue-residue distance matrix using dynamic programming

Methods are listed in alphabetical order. TM-score: template modeling score. DP = methods using dynamic programming. ^aFor methods that use fragments in the optimization phase, the fragment length is provided. The DALI score measures the difference between intra-structure residue-residue distances. Methods also differ in the construction of initial alignments, which are then refined to identify better-scoring alignments. Flexible aligners typically use a sum of the similarities of all aligned fragment pairs, to which a penalty is added for each breakage between fragments.

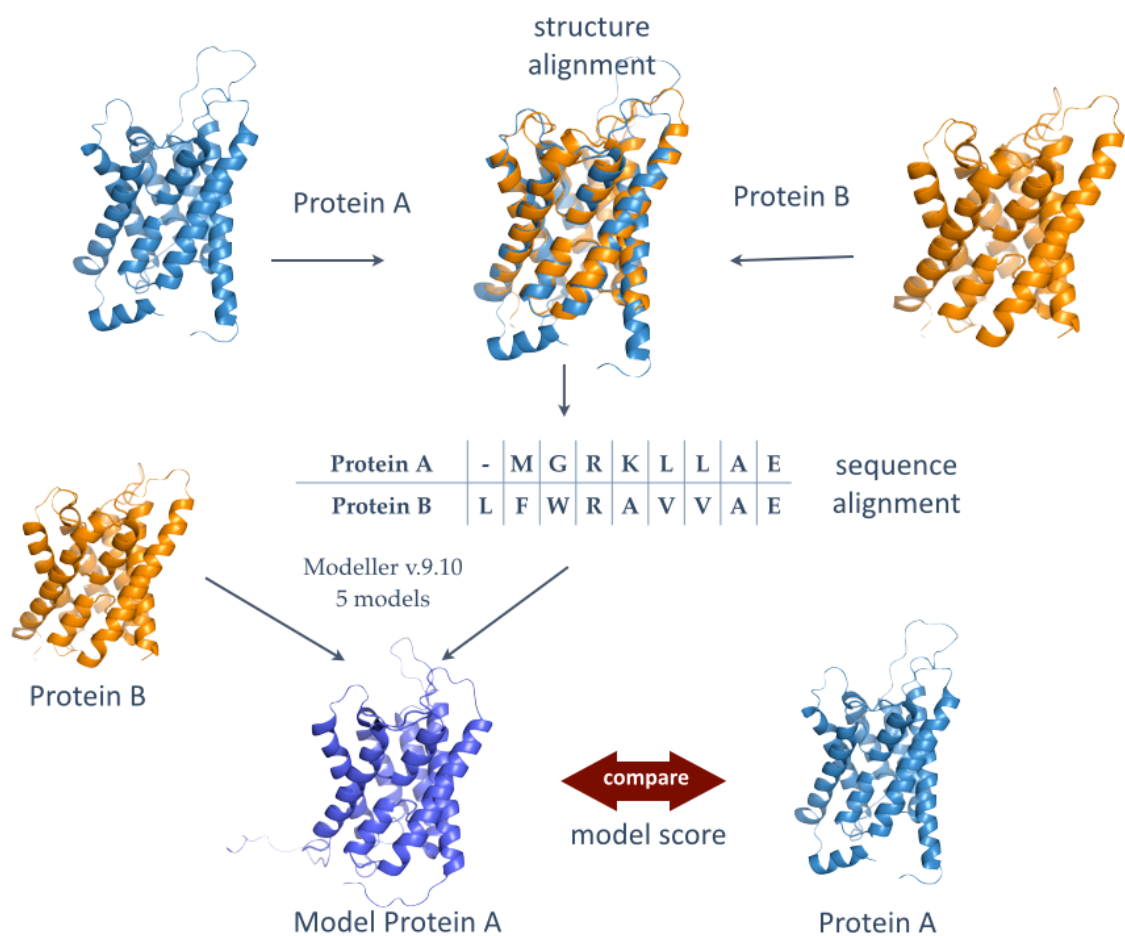


Figure 5.1 Example workflow for generating and evaluating a homology model. Two proteins (A and B) are aligned with each other using a structural alignment method. The underlying sequence alignment is then submitted to Modeller with a template structure of Protein B to generate a model of Protein A. The quality of the model is evaluated by a comparison of the model structure with its original X-ray structure using a model score (e.g., GDT_TS).

5.2.2 Evaluation by Modeling Using Structural Similarity Scores

The major challenge in assessing structural alignment accuracy is the lack of a standard score that ranks the quality of structural alignments. Although efforts have been made to address this issue recently (Collier, et al., 2014), those strategies are not yet publicly available. Here, I rely on the fact that a homology model based upon an accurate alignment results in a more accurate model compared to the original crystal structure than a model based upon a less accurate alignment. Thus, for each pair of sequences, each sequence was modeled using the other as a template, and vice versa (see Figure 5.1). Models were generated using five cycles of optimization in Modeller v9.10 (Šali and Blundell, 1993), and the best of these five models was selected for comparison based on the GDT_TS score of the model relative to the known structure (see below). This approach is similar to the

evaluation of sequence alignment methods (see chapter 3.2.6) but for this purpose, more structural similarity measures (AL0, AL4, GDT_TS, RMSD, TM-score, CAD score) were used to rate and rank the model quality.

5.2.2.1 AL0 & AL4 scores

A simple structural similarity score that has been used before for assessing homology models is the AL4 score (see chapter 3.2.6 for more details). The AL4 score identifies the largest subset of C_α-atoms of the model that can superimposed with target structure below a cut-off distance of 10 Å. This cut-off corresponds to the shift of 4 residues in an alignment or one helical turn in the protein structure. Besides the AL4 score, also the AL0 score (Kopp, et al., 2007) was applied in this study. The threshold of the AL0 score is 3.8 Å and corresponds roughly to the distance between adjacent C_α-atoms in a regular α-helical protein structure.

A characteristic of both scores is their insensitivity to small and very extreme structural distances by omitting explicit differences of distances between two amino acids in the score. For example, two amino acids that are only 1 Å away are treated similar to those that are 9 Å away by the AL4 score.

5.2.2.2 Global Distance Test – Total Score (GDT_TS)

Another score that has been used before is the GDT_TS score, which stands for global distance test (total score). The GDT_TS score identifies the number of structurally-equivalent pairs of atoms below four different specific distance thresholds ($G(c) = 1, 2, 4$ and 8 Å) The advantage of the GDT_TS score is that only correctly modeled positions are rewarded, without a penalty for inaccurately modeled regions. Nonetheless, the score is still dependent on the size of the protein (see chapter 3.2.6 for more details).

A variant of the GDT_TS score with a better resolution for high-accuracy models by focusing more on correctly modeled segments is the GDT_HA score (Kopp, et al., 2007), that applies cut-off distances being the half of those of the GDT_TS score, with $G(c)$ being 0.5, 1, 2 and 4:

$$\text{GDT_HA}(\%) = \frac{1}{4} \sum_{c=0.5,1,2,4} \left[\frac{G(c)}{L_{\text{target}}} \times 100 \right] \quad (5.1)$$

The GDT_TS, GDT_HA as well as the AL0 and AL4 were so far only used for assessing homology model quality and have not been applied in any of the tested structural alignment methods.

5.2.2.3 *RMSD (Root Mean Square Deviation) Score*

A score that has been applied not only for ranking the similarity between two protein structures with identical amino acid sequences but also for superimposing two distinct protein structures in a structural alignment method (e.g., in CE or SAP) is the RMSD (Root Mean Square Deviation) score (Kabsch, 1976). However, the RMSD score is sensitive to large outliers because it squares the distance differences between the pairs of equivalent C_α-atoms (see chapter 2.4 for more details). This domination of small outliers makes the RMSD score powerful for closely related proteins in which small structural deviations should be detected. Structures in which the differences are distributed evenly among the entire protein sequences receive smaller and thus better RMSD scores than structures that are overall similar but have small fragments that differ significantly. Consequently, the RMSD score is less useful for more distantly related proteins, especially if there are segments modeled without an underlying template structure (i.e. insertions or deletions).

However, the HOMEP data set contains many pairs of proteins of different lengths, as well as proteins that share a low sequence identity, resulting in difficult alignment cases. In both cases, parts of the target protein are poorly modeled because evolutionarily related residues are misaligned or an underlying template structure fragment is missing. The sensitivity to similarity of template and target and to the length of the sequences might dominate the score and mislead the assessment of homology models. These issues were discovered during the first CASP experiments (Moult, et al., 1997; Moult, et al., 1995). The RMSD score is therefore not useful for evaluating the alignment accuracy of structural alignments based upon the HOMEP data set.

5.2.2.4 *Template Modeling (TM)-Score*

Another score that has been applied for the superimposition of two protein structures is the TM-score that is applied in the structural alignment programs TM-align and FR-TM-align. In contrast to the RMSD score, the TM-score is sensitive to small distance differences and also accounts for large outliers with a distance-dependent weighting scheme that averages down the contribution of largely deviating pairs of residues to the final score. The TM-score has also been applied for the generation of the HOMEP3 data set (see chapter 2.5 for more details on the TM-score). The TM-score ranges between 0 (worst case) and 1 (perfect match) and a TM-score higher than 0.5 suggests that two structures might be homologous (Xu and Zhang, 2010).

5.2.2.5 Packing-based Contact Area Difference (CAD) Score

A drawback of geometrical scores (e.g., GDT_TS, AL0) is that they only consider the correctness of the C_α-backbone and lack information about the correctness of residue side chain modeling although they could in principle also include all atoms of a protein into their calculation. However, the correct orientation of the side chains is crucial for hydrophobic interactions within the membrane-spanning segments of membrane proteins in which hydrophobic amino acids are facing the lipids of the membrane bilayer. Hydrophobic interactions occur also at protein-protein contacts (in water-soluble and in membrane proteins) being crucial for the packing of a protein as well. Additionally, the orientation of charged (or other types of) amino acids is important in binding sites or transport pathways of membrane proteins. Steric clashes of side chains or uncommon distortions should therefore be penalized by a geometrical score that accounts for their biological inappropriateness. The CAD (Contact Area Difference) score (Olechnovic, et al., 2012) considers the orientation of side chains by computing residue-residue contact surface areas using Voronoi tessellation for both the model (*M*) and template reference structure (*T*). The residue-residue contact areas of each residue pair (*i,j*) in the model (*M*) and template (*T*) are compared to each other. Rearrangements of domains, fragments or side chains are penalized without any threshold. Consequently, the CAD score captures essential geometrical properties of the template and the modeled structure:

$$CAD_{(i,j)} = \min\left(\left|T_{(i,j)} - M_{(i,j)}\right|, T_{(i,j)}\right) \quad (5.5)$$

and the CAD-score of the model is:

$$CAD\text{-score} = 1 - \frac{\sum_{(i,j) \in G} CAD_{(i,j)}}{\sum_{(i,j) \in G} T_{(i,j)}} \quad (5.6)$$

A CAD score of 1 indicates that all residues in the model have the same contact surface areas and thus exactly the same orientation as in the reference structure, whereas a score of 0 corresponds to a total disagreement without any contact surface areas being in common between model and template structure. In this study, the CAD-score that is based upon the surface area of all atoms (AA-CAD) is applied because it has been shown to be most accurate among all variants of the CAD score (Olechnovic, et al., 2012).

In general, the CAD score could also be used for optimizing a structural alignment procedure but so far, the CAD score is only used as an assessment score and has not been implemented in a structural alignment method.

The RMSD, GDT_TS, AL0 and AL4 scores were calculated using the LGA package (Zemla, 2003), and the CAD, GDT_HA and TM-score scores were obtained using the CADscore package (Olechnovic, et al., 2012).

5.2.3 Strategies for Ranking the Accuracy of Structural Alignment Methods

The structural alignment programs can be ranked by the reliability and accuracy of the alignments they produce. Whereas accuracy describes the average quality of a set of alignments, reliability reflects the deviation of the accuracy among a set of alignments. Both measures are valuable parameters for users who want to have high quality alignments (high accuracy) for all alignments (high reliability). Consequently, two different ranking schemes are applied for testing the performance of the structural alignment methods used in this study.

The first ranking scheme (R_{mean}) reflects an accuracy measure for all structural alignment methods. For each of the aforementioned structural similarity scores (S), each structural alignment method was assigned a mean score (S_{mean}) over the scores S_m of all models m in the set of M models (i.e., all aligned pairs of protein structures from the HOMEP3 data set):

$$S_{mean} = \frac{1}{M} \sum_{m=1}^M S_m \quad (5.7)$$

Subsequently, a rank (R_{mean}) is assigned to each method by comparing the S_{mean} values of all methods with each other. The best rank with the value 1 is assigned to the most accurate method with the highest S_{mean} value. All other methods are sorted in an ascending order so that an increasing rank R_{mean} reflects a decreasing alignment accuracy showing the relative overall accuracy of a method among others. The R_{mean} ranking is equivalent to the so-called AR score (Reddy Ch, et al., 2006) and has been renamed for clarity.

The second ranking scheme ($R_{reliability}$) reflects the reliability of a method and assesses whether there are, among a set of alignments, large outliers (positive as well as negative) of the alignment accuracy for a specific method. For every model m (each alignment pair), a structural similarity score was calculated, and based upon this score a rank R_m was assigned to each structural alignment method based on their scores for that model. These rankings were computed and averaged over all M models to obtain $R_{reliability}$ for each method:

$$R_{reliability} = \frac{1}{M} \sum_{m=1}^M R_m \quad (5.8)$$

Again, the best rank with the value 1 ($R_{reliability} = 1$) is assigned to the most reliable structural alignment method, which is on averaged ranked higher than all other structural alignment methods. An increasing rank corresponds to less reliable methods. This measure ($R_{reliability}$) is equivalent to the RA score from Reddy (Reddy Ch, et al., 2006).

This combination of the two measures R_{mean} and $R_{reliability}$ allows for a confident quality assessment of structural alignment methods. If a method is ranked relatively well (e.g., a good rank of 1 out of 14 methods) or poorly (e.g., a poor rank of 14) according to both ranking schemes, then it can be assumed that the rank is confident without any large outliers. If a method is accurate (e.g., $R_{mean} = 2$) and reliable (e.g., $R_{reliability} = 2$) then all corresponding models are assumed to be accurate, whereas a high rank of both scores (e.g., $R_{reliability}=9$, $R_{mean}=9$) shows that a method performs poorly for all protein pairs (see Table 5.2). However, the two measures R_{mean} and $R_{reliability}$ do not have necessarily to agree with each other and two cases can be observed.

First, a structural alignment method can have a good rank for R_{mean} (e.g., 2), but a poor rank for $R_{reliability}$ (e.g., 9). Such a method generates poor models in most cases (poor $R_{reliability}$ value) but has some overtly accurate models as outliers that raise the R_{mean} score. Second, a structural alignment method can have a poor rank for R_{mean} (e.g., 9), but a good rank for $R_{reliability}$ (e.g., 2). Such a method generates accurate models in most cases (good $R_{reliability}$ value) but has some overtly poor outliers lowering the R_{mean} score (see Table 5.2).

Table 5.2 Explaining model quality by the agreement of different ranking schemes

	Good $R_{reliability}$ rank	Poor $R_{reliability}$ rank
Good R_{mean} rank	Overall accurate models	Poor models in most cases but a few overtly accurate models in a few cases
Poor R_{mean} rank	Accurate models in most cases with a few overtly poor outliers	Overall poor models

Similar R_{mean} and $R_{reliability}$ values identify methods that perform overall accurate (good ranks) or inaccurate (poor ranks) without outliers, whereas inconsistent R_{mean} and $R_{reliability}$ values are present for methods that have some (positive or negative) outliers in the alignment accuracy.

5.2.4 Model Selection

As mentioned before (chapter 5.2.1), for each alignment that was obtained, five models were generated using Modeller v.9.10 to explore different modeling solutions using different random seeds in order to not get stuck with a bad model. These five models were then compared to the original crystal structure using a geometrical measure and the model with the best (highest) score was then used as a representative for the underlying alignment (see Figure 5.1).

However, the selection of the representative model might also be dependent on the geometrical score that was applied to obtain the “best” model. To identify the most reasonable, representative model out of 5 models for representing an alignment, three structural similarity scores (GDT_TS, AL4 and CAD) as well as the DOPE score, which is a statistical energy function, were considered. All structural alignment methods were ranked four times using in each case one of these scores for selecting the best models based upon their alignments. The structural alignment methods were then ranked using the measures R_{mean} and $R_{reliability}$, which have been described in the previous chapter. Subsequently, the rankings R_{mean} and $R_{reliability}$ based upon the four different geometrical scores and their representatives were compared to check whether the process of selecting the best model influences significantly the ranking of the structural alignment methods or not. Results are described in chapter 5.3.1 in more detail.

5.2.5 Consistency of Alignments Between Homologous Protein Sequences

Accurate structural alignment methods are expected to align evolutionarily-related protein sequence positions to each other. Consequently, an attribute of an accurate method is that the alignment denoted AB of a pair of homologs (protein A with protein B) can be deduced from alignments of the two proteins with a third homolog (AC and BC). Based upon the alignments AC and BC , an alignment of AB is derived using protein C as a reference sequence. The derived alignment AB is then compared to the original alignment of AB . From the original alignment of AB , each amino acid of Protein A and B are checked if they are aligned to the same position in the derived alignment AB (see Figure 5.2). The number of consistently aligned residues is counted and divided by the lengths of Proteins A and B . This procedure is repeated for all possible combinations of those three proteins.

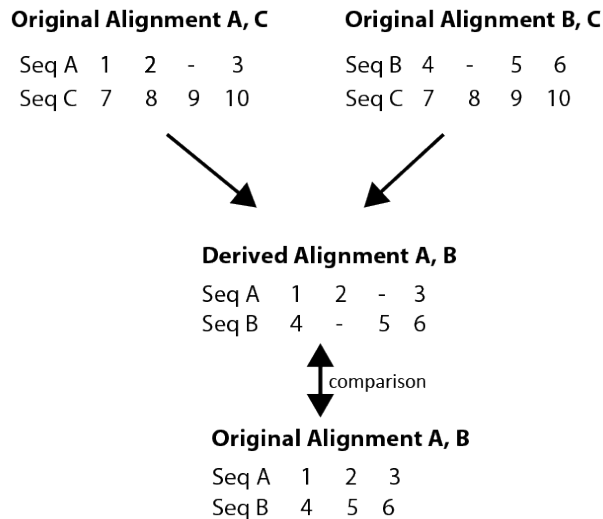


Figure 5.2 An alignment of proteins A and B is derived from two other alignments (AC and BC) with Protein C as a reference. This derived alignment is then compared to its original alignment.

Another useful metric is the “average shift error” of the inconsistent positions. For consistent positions, the average shift error is 0. In case of inconsistent positions (A_i aligned to B_j with j being different in the derived alignment compared to the original alignment: $j^{derived} \neq j^{original}$), the difference between the aligned positions is calculated: $|j^{derived} - j^{original}|$. Here I considered only ungapped positions. To obtain the average shift error $E(AB/C)$ for the alignment AB , relative to AC and BC , the shift error is summed over all inconsistent positions, and divided by the sum of L_A and L_B .

5.2.6 Statistical Analysis

Differences between inputs were assessed using the Wilcoxon signed ranked test (Wilcoxon) and were deemed to be significant when $p < 0.05$.

5.3 Results

A first step before evaluating the accuracy of homology models is the choice of a representative model from a set of models that were generated based upon an alignment. Chapter 5.3.1 describes the selection of a specific score that ensures that the best representative model is selected from a set of models. For each structural alignment method, a representative model was chosen for each protein pair and based upon these models subsequent analysis was carried out. In chapter 5.3.2, structural similarity scores were compared and correlated with each other. Subsequently, three structural similarity scores were used for the analysis of the models that were generated by each structural alignment method. Three features of structural alignment programs were shown to

contribute to their alignment accuracy: their inherent scoring scheme for which results are described in chapter 5.3.3; their ability to allow for a fragment-based superimposition described in chapter 5.3.4; and the usage of conformational rigidity or flexibility in their alignments which is described in chapter 5.3.5. Additionally, the alignments were checked for their alignment coverage in chapter 5.3.6 and their self-consistency for alignments of homologous proteins in chapter 5.3.7. Last, a consensus approach that allows for obtaining confidence values by fusing multiple structural alignments to a single alignment is presented in chapter 5.3.8.

5.3.1 Selection of a Representative Model

The geometrical measure that is used for selecting a representative model out of 5 models obtained with Modeller was analyzed for its influence on the subsequent ranking of the different structural alignment methods. A structural alignment method was never ranked more than 3 positions higher or lower according to its R_{mean} that was based upon models that were selected by either the structural similarity scores AL4, GDT_TS, CAD or the Modeller internal DOPE score. R_{mean} values were identical in 83.7% and 88.9% of cases for α -helical and β -barrel proteins, respectively, and the R_{mean} rankings for a given method never deviated by more than three ranks. Similar results were obtained if the structural alignment methods were ranked according to their $R_{reliability}$ value. $R_{reliability}$ values were identical in 70.7% and 77.4% of all cases for α -helical and β -barrel proteins, respectively, and also never deviated more than three positions.

Accordingly, the AL4, GDT_TS, CAD and DOPE scores are all similarly suitable for selecting the most adequate model among a set of initial models. Since the GDT_TS score has been shown to be a useful measure for assessing protein similarity in previous CASP studies, all models that were evaluated in the subsequent chapters were selected among a set of 5 models generated by Modeller v.9.10 by using the model with the highest GDT_TS score.

5.3.2 Correlation of Structure Similarity Scores

The usage of different similarity scoring schemes does not guarantee that they evaluate different quality aspects of a homology model. Consequently, an efficient and effective comparison of the different structure alignment methods is best achieved by using complementary structure similarity scores. For each structural alignment method, a representative model was chosen for each pair of proteins in the HOME3 data set and all similarity scores were calculated by comparing the model to its original X-ray structure. The generated structural similarity scores for each protein pair were then correlated by calculating a Pearson correlation coefficient between them to check whether they contain complementary information or not.

The geometrical distance-based scores GDT_TS, GDT_HA and AL0 are strongly correlated with one another for both α -helical and β -barrel-like proteins (see Table 5.3) reflecting the fact that these three scores all consider short-range similarity, in a length-independent way (e.g., considering the fraction of residues that are aligned within a threshold of 4 Å). The TM-score is also highly correlated with the GDT_TS and AL0 score due to its inclusion of global as well as of local information. However, the usage of global information results in a weaker correlation to the GDT_HA score. The AL4 score is less well correlated than the TM-score with the other threshold-dependent scores because of the 10 Å threshold that makes the AL4 score being dominated by long-range distance differences over detailed local spatial information. Not surprisingly, the RMSD score is not correlated at all with any of these distance-dependent scores (see Table 5.3) due to its tendency to penalize small fractions of outliers too strongly. Aside from these geometrical dependent distance measures, there is the CAD score that includes environment-based information by its usage of contact areas between atoms. Interestingly, the CAD score is better correlated with the threshold-based modeling scores GDT_TS and GDT_HA that focus on highly-accurately aligned positions (< 2Å) than with the AL0 (cut-off of 4 Å) or AL4 score (cut-off of 10 Å), both of which consider longer-range differences (see Table 5.3).

Based on these results, three different scores were selected for subsequent analysis: the GDT_TS score which is sensitive for local deviations and averages out large outliers, the AL4 score which counts all correctly aligned residues below a threshold of 10 Å and the CAD score which compares the contact areas of all atoms between two homology models, although there may be some overlap between the conclusions based on CAD and GDT_TS.

Table 5.3 Correlation between model quality assessment scores

		β -barrels						
		GDT _TS	GDT _HA	AL0	TM-score	AL4	CAD	RMSD
α -helical proteins	GDT_TS		0.98	0.98	0.94	0.86	0.97	-0.43
	GDT_HA	0.98		0.93	0.88	0.76	0.97	-0.37
	AL0	0.96	0.90		0.96	0.90	0.93	-0.44
	TM-score	0.91	0.85	0.92		0.86	0.90	-0.48
	AL4	0.79	0.68	0.83	0.91		0.79	-0.56
	CAD	0.90	0.90	0.85	0.79	0.71		-0.42
	RMSD	-0.57	-0.60	-0.60	-0.60	-0.67	-0.56	

Scores are correlated for either β -barrels (upper right), or α -helical proteins (lower left). Structural similarity scores were correlated using a Pearson correlation coefficient. Entries in bold indicate scores that are highly correlated with each other (Pearson correlation coefficient >0.9).

5.3.3 Structural Alignment Methods with Length-Independent Scoring Schemes Generate Better Alignments

A correlation of the structural similarity scores allows for an assessment of the different structural alignment methods using three of these scores: CAD, AL4 and GDT_TS. The structural alignment methods differ in their inherent scoring scheme for obtaining the optimal superimposition of two protein structures as well as in their algorithms for optimizing this score by a superimposition (see Table 5.1).

Structural alignment methods that apply a score that down-weights the contribution of incorrectly modeled fragments to the total score (e.g., TM-score in TM-align, Dali score in DaliLite) were generally more accurate and were ranked better than methods that apply a score that squares the spatial differences between two structures for a superimposition (e.g., URMS in Mammoth, RMSD score in CE and SAP; Table 5.4 and Table 5.5).

The negative influence on the alignment accuracy of squaring spatial distances between two protein structures is exemplified by CE, which applies simultaneously a fragment-based approach as well as an RMSD score. Although fragment-based approaches are shown to be advantageous for generating accurate alignments, the inclusion of an RMSD score results in the least accurate and worst ranked alignments (and models) of the fragment-based approaches for both α -helical (Table 5.4) and β -barrel-like proteins (Table 5.5).

Similar to the results for a general protein data set (Sadowski and Taylor, 2012), the Template Modeling score (TM-score) seems to be most useful for aligning membrane proteins, since both methods that apply the TM-score (FR-TM-align and TM-align) were the highest ranking and most accurate methods for α -helical proteins. Another score that seems adequate to obtain accurate structural alignments is the Dali score used in DaliLite, which down-weights large outliers similar to the TM-score. However, the Dali score is more suitable for aligning β -barrel-like proteins (rank of 3 for DaliLite comparing all methods with each other, Table 5.5) than for aligning α -helical proteins (rank of 8 of DaliLite among all methods, Table 5.4).

Table 5.4 Ranking of structural alignment methods for the subset of α -helical membrane proteins

Score	Type	FR-TM-align	TM-align	FATCAT rigid	FATCAT	MATT	LovoAlign	SKA	Dallite	SABERTOOTH	SHEBA	CE	SAP	PPM	MAMMOTH
CAD ^a	R_{mean}	1	4	3	2	5	8	10	9	7	6	11	12	13	14
	$R_{\text{reliability}}$	1	2	4	3	8	5	7	6	11	10	12	9	13	14
	mean (%)	63.3	63.3	63.3	63.3	63.3	62.8	62.4	62.7	62.8	62.9	62.3	62.0	60.7	59.7
	stdev (%)	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.10	0.09	0.09
AL4	R_{mean}	2	1	3	6	4	9	5	11	7	10	8	14	12	13
	$R_{\text{reliability}}$	1	2	4	6	7	5	3	8	10	11	9	12	14	13
	mean (%)	94.8	94.8	94.5	93.9	94.2	93.1	94.2	92.8	93.6	92.9	93.4	90.0	90.9	90.4
	stdev (%)	7.0	6.9	7.2	9.3	8.1	12.9	8.3	11.2	8.3	9.2	8.3	17.0	10.9	14.7
GDT	R_{mean}	1	2	3	5	4	9	10	8	6	7	11	12	13	14
	$R_{\text{reliability}}$	1	2	7	8	6	3	4	5	10	11	9	12	13	14
	mean (%)	71.1	71.0	70.6	70.1	70.4	69.5	69.7	69.7	70.0	69.6	69.4	67.2	66.2	63.3
	stdev (%)	14.8	14.8	14.8	16.0	15.8	17.9	16.1	16.5	15.6	15.8	15.9	19.6	16.7	18.4
ave ^b	R_{mean}	1.3	2.3	3.0	4.3	4.3	8.7	8.3	9.3	6.7	7.7	10.0	12.7	12.7	13.7
	$R_{\text{reliability}}$	1.0	2.0	5.0	5.7	7.0	4.3	4.7	6.3	10.3	10.7	10.0	11.0	13.3	13.7

The structural alignment methods are sorted according to the sum of their average R_{mean} and $R_{\text{reliability}}$ rankings, with the most accurate alignments on the left side of the table. The mean and standard deviation (stdev) of each score over all pairs of alignments are given. ^aCAD score multiplied by 100. ^bMean ranking over all three scores. Entries in bold indicate the highest or best scores in that column and those that are not significantly different from the highest/best score, according to the Wilcoxon signed rank test, with $p < 0.05$.

Table 5.5 Ranking of structural alignment methods for the subset of β -barrel membrane proteins

Score	Type	FR-TM-align	TM-align	DaliLite	FATCAT rigid	FATCAT	MATT	SKA	LovoAlign	SHEBA	SABERTOOTH	CE	PPM	SAP	MAMMOTH
CAD	R_{mean}	2	4	1	5	3	6	8	9	7	10	11	12	14	13
	$R_{\text{reliability}}$	1	2	5	3	4	7	6	8	9	10	12	13	11	14
	mean	54.4	54.1	54.4	54.1	54.2	54.1	53.1	52.6	53.6	52.6	51.5	47.5	46.9	47.1
	stdev	0.09	0.10	0.09	0.09	0.09	0.09	0.11	0.11	0.10	0.11	0.11	0.10	0.15	0.11
AL4 (%)	R_{mean}	1	2	4	3	6	5	8	9	7	10	11	12	14	13
	$R_{\text{reliability}}$	1	2	7	3	5	8	4	6	11	10	9	13	12	14
	mean	90.5	90.1	89.3	89.8	88.9	89.1	87.0	86.1	87.7	85.7	85.7	82.7	66.8	79.9
	stdev	10.5	11.2	11.2	11.0	12.6	11.5	17.6	17.6	12.4	16.9	15.5	14.7	35.9	17.7
GDT_TS	R_{mean}	1	3	2	4	6	5	8	10	7	9	11	12	14	13
	$R_{\text{reliability}}$	1	2	3	4	5	6	8	7	10	11	9	13	12	14
	mean	63.9	63.1	63.4	62.9	62.0	62.6	59.5	58.7	61.0	59.0	58.3	50.4	46.1	47.4
	stdev	14.9	15.7	15.3	15.5	16.6	15.9	19.4	20.9	16.4	19.3	19.2	19.0	30.3	20.3
ave.	R_{mean}	1.3	3.0	2.3	4.0	5.0	5.3	8.0	9.3	7.0	9.7	11.0	12.0	14.0	13.0
	$R_{\text{reliability}}$	1.0	2.0	5.0	3.3	4.7	7.0	6.0	7.0	10.0	10.3	10.0	13.0	11.7	14.0

See legend to Table 5.4 for details.

5.3.4 Rigid Superimpositions Compared to Fragment-Based Superimpositions

Besides the inherent scoring scheme, another major difference between the tested structural alignment approaches is their algorithms for optimizing the inherent similarity score by a superimposition procedure.

Some methods measure structural similarity based upon a rigid superimposition of the total structures (e.g., CE or TM-align) whereas other methods allow for a flexible protein structure alignment by focusing more on an optimal match of local protein substructures than on the total structure overall (e.g., FATCAT flexible mode, FR-TM-align, MATT). Interestingly, 3 out of 5 top-ranking methods for α -helical proteins are fragment based methods (FR-TM-align, FATCAT and MATT, Table 5.4). Comparison of the results obtained for α -helical proteins using the TM-align and FR-TM-align methods (Table 5.4) clearly demonstrates the increased accuracy obtained using fragments, since all other aspects of these methods, including the scoring function (both use the

Template Modeling score), are the same. Models based upon alignments from FR-TM-align are more accurate and ranked higher than their counterparts based on TM-align alignments for α -helical-proteins (Table 5.4). The only case in which TM-align is ranked higher than FR-TM-align is according to the R_{pairwise} ranking based upon an assessment using the AL4 score (Table 5.4). However, that difference is only marginal and the results are not statistically significantly different from each other.

Similar results as for α -helical proteins were obtained for β -barrel-like proteins (Table 5.5). Again, fragment-based approaches (FR-TM-align, FATCAT, DaliLite) were ranked higher than methods that apply a rigid superimposition, with FR-TM-align again being the best choice among all methods. Intriguingly, the DaliLite method appears to be better suited to β -barrel proteins than to α -helical proteins (Table 5.4 and Table 5.5). I speculate that this is because DaliLite compares the intra-structural distance matrices of two proteins: the distance matrices of β -barrel proteins are likely to be a distinctive mixture of small and large distances, unlike the matrix of many short distances that would be characteristic of α -helical proteins.

5.3.5 Conformational Flexibility of Membrane Proteins

All structural alignment methods consider only one state of a membrane protein during the alignment process. This assumption fits to very rigid proteins like cytochrome bc_1 or Light Harvesting Complexes, which change the electronic state of their bound co-factor during function. In contrast to those rigid proteins, there are also some flexible membrane proteins that can adopt distinct conformations to be able to transport substrates or to transmit signals. Two examples for such conformations are the inward- and the outward-facing states of a transporter. In the inward-facing state of a protein, its binding side is exposed to the inside of a cell and allows the uptake of ligands and substrates from the cell or their release into the cell. The outward-facing state is the counterpart to the inward-facing state and allows the uptake and release of ligands and substrates from the exterior environment of a cell. These two conformational states are exemplified in HOME3 by the Major Facilitator Superfamily (MFS), which contains two structures (GlpT and LacY; PDB codes: 1PW4 and 2CFQ) that have been determined in inward-facing states, and two structures (FucP and Xyle; PDB codes: 3O7Q and 4GC0) solved in outward-facing states. The differences in these states mainly arise from the repositioning of two six-transmembrane-helix domains relative to one another, so as to open a pathway into the membrane from one or other side of the membrane (Figure 5.3).

The results obtained when aligning these MFS structures indicate that the performance depends on the question being asked. I demonstrate this by comparing the results obtained using FATCAT with the rigid-body fitting option, or in its 'flexible' (fragment-based) mode, using a spatial structural

similarity score (GDT_TS) or an environmental-based structural similarity score (CAD score). For comparison of structures in the same conformation, using FATCAT with the rigid-body fitting option resulted in more accurate models than using FATCAT in 'flexible' mode (Table 5.6a). For those cases, I also found that the distance-threshold based GDT_TS scores and the packing-based CAD score both described the results well, and agreed well with one another (Table 5.6a).

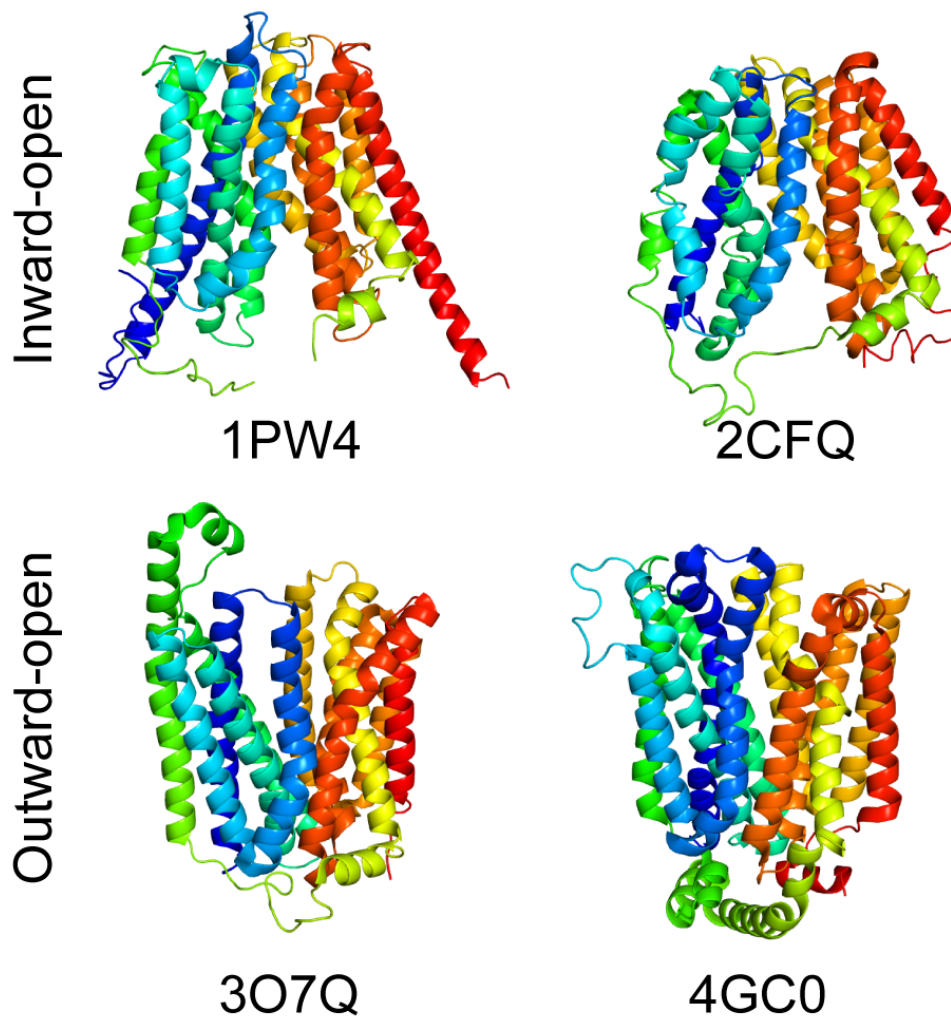


Figure 5.3 Alternate conformations in the family of Major Facilitator Superfamily transporters. Two structures reflect inward-facing conformations (GlpT and LacY; PDB codes: 1PW4 and 2CFQ) and two reflect outward-facing conformations (FucP and XyleE; PDB codes: 3O7Q and 4GC0). The proteins are shown as cartoon helices, viewed from along the plane of the membrane with the outside of the cell toward the top, and colored according to the rainbow, from blue (N-terminal) to red (C-terminal).

Table 5.6 Comparison of models in the MFS transporter family based on alignments generated using FATCAT in flexible and rigid-body fitting modes for structures in similar or different conformational states.

a) same states		GDT_TS		CAD	
template	model	rigid	flexible	rigid	flexible
1PW4	2CFQ	48.5	44.3	49.8	48.8
2CFQ	1PW4	46.4	42.3	49.3	48.2
3O7Q	4GC0	50.1	45.5	50.3	49.6
4GC0	3O7Q	56.0	51.0	54.2	53.4
b) different states		GDT_TS		CAD	
1PW4	3O7Q	37.3	40.2	47.5	50.9
1PW4	4GC0	38.8	32.4	45.8	46.7
2CFQ	3O7Q	34.9	37.3	46.5	49.8
2CFQ	4GC0	33.2	30.9	45.0	45.2
3O7Q	1PW4	36.1	38.3	49.0	50.2
3O7Q	2CFQ	36.3	35.3	46.9	50.3
4GC0	1PW4	41.5	33.7	50.1	50.1
4GC0	2CFQ	38.9	34.6	48.4	49.2

Structural similarity scores for MFS transporter proteins that have been reported in an inward-facing conformation (GlpT and LacY; PDB codes: 1PW4 and 2CFQ) or an outward-facing conformation (FucP and XylE; PDB codes: 3O7Q and 4GC0). Entries in bold indicate the best score of the flexible or rigid-body mode. Chain A is used in all cases.

The applicability of structural similarity scores like the GDT_TS or CAD score is also affected by conformational changes because they compare two rigid structures with each other. For all structural similarity scores, the highest score is achieved if two protein structures match each other perfectly with the assumption that these structures are in the same conformational state. The more the conformation of two structures changes, the more their structural similarity can change (e.g., increase of C_{α} - C_{α} distances or changes of contact areas between molecule spheres). Consequently, two structures of the same protein that are in very different states might not be detected as being structurally similar by a structural similarity scoring scheme.

During the homology modeling process the model adopts the conformation of the template structure even if the expected target structure would be in a different conformation. In such cases, a spatial-based structural similarity score (e.g., GDT_TS) might be inappropriate because helical segments may be correctly modeled as helices but may be arranged in a different orientation towards each other.

An environmental based structural similarity score like the CAD score might be expected to better capture the changes inherent in repositioning large domains relative to one another. Consistent with this expectation, the models are ranked differently when using the GDT_TS and CAD scores (Table 5.6b). According to the CAD score, the flexible mode of FATCAT results in more accurate alignments for structures solved in different conformations than the mode of FATCAT that relies on rigid-body fitting (Table 5.6a). Similar results are also obtained if only the membrane-spanning segments of MFS transporters are considered (Table 5.7).

This example illustrates that introducing fragmentation and therewith flexibility is particularly useful for aligning proteins that are in different conformations, whereas proteins solved in a similar state are preferably aligned using a rigid-based superimposition.

Table 5.7 Comparison of models in the MFS transporter family based on alignments generated using FATCAT in flexible and rigid-body fitting modes for structures in similar or different conformational states considering membrane-spanning segments only

a) same states		GDT_TS		CAD	
template	model	rigid	flexible	rigid	flexible
1PW4	2CFQ	52.0	48.5	0.525	0.532
2CFQ	1PW4	54.6	51.4	0.532	0.526
3O7Q	4GC0	64.5	60.4	0.581	0.563
4GC0	3O7Q	67.4	64.0	0.598	0.596
b) different states		GDT_TS		CAD	
template	model	rigid	flexible	rigid	flexible
1PW4	3O7Q	49.4	51.9	0.532	0.562
1PW4	4GC0	51.6	44.9	0.532	0.527
2CFQ	3O7Q	44.6	48.6	0.489	0.539
2CFQ	4GC0	45.7	43.4	0.510	0.525
3O7Q	1PW4	48.6	51.4	0.538	0.550
3O7Q	2CFQ	45.7	45.4	0.517	0.550
4GC0	1PW4	53.2	45.3	0.568	0.528
4GC0	2CFQ	47.3	43.3	0.523	0.531

See legend to Table 5.6 for more details.

Unfortunately, none of the programs includes an option that recognizes the states of the proteins (e.g., by defining the states of the input structures) although that might be useful for considering the degree of flexibility that has to be introduced for an accurate superimposition of the protein structures. FATCAT and MATT have the option to set a flag that allows for more levels of flexibility but they are not able to active or deactivate this flag on their own. Currently, the decision of allowing for flexibility has to be made ad hoc by the user.

5.3.6 Alignment Coverage

The accuracy of structural alignment methods is directly reflected by their ability to correctly insert gaps at evolutionarily variant positions (e.g., insertions or deletions). Accordingly, structural alignment methods that are not able to capture homology correctly may align too many residues (over-align) that are evolutionarily not related, resulting in a rather short alignment, or they might miss correct relationships between evolutionarily related positions and insert too many gaps (under-align).

To test whether the different methods tend to over- or under-align, two values were calculated: the alignment coverage and the average alignment length (Table 5.8). The alignment coverage is the average percentage of the two structure lengths that are aligned. For the average alignment length, the lengths of all alignments for a specific alignment method were summed up and divided by the number of alignments. The values of the percentage of aligned residues and the average alignment length were compared to that of the most accurate method, which in this case is FR-TM-align (evaluation shown in Table 5.4 and Table 5.5). Structural alignment programs that produce shorter alignments than those generated using FR-TM-align are therefore assumed to likely over-align whereas methods that produce longer alignments tend to under-align amino acids. For both α -helical and β -barrel proteins, SAP and PPM tended to significantly under-align membrane protein structures, whereas MAMMOTH and SHEBA tended to over-align them (Table 5.8), which explains their overall poor to average performance in terms of model accuracy (see Table 5.4 and Table 5.5).

Table 5.8 Alignment coverage in the alignments generated using different structure alignment programs

α -helical proteins			β -barrels		
	%aln	length		%aln	length
MAMMOTH	85.4%	333.8	MAMMOTH	79.0%	455.2
SHEBA	84.8%	336.0	SHEBA	76.0%	461.7
FATCAT	84.3%	336.8	LovoAlign	75.5%	462.9
LovoAlign	84.1%	337.3	FATCAT	75.5%	463.2
FATCAT rigid	83.9%	337.9	FATCAT rigid	75.2%	463.8
TM-align	82.5%	340.5	FR-TM-align	73.4%	468.2
FR-TM-align	82.5%	340.5	TM-align	73.2%	468.6
SABERTOOTH	79.4%	346.9	SABERTOOTH	73.8%	468.6
SKA	78.4%	349.1	DaliLite	69.4%	481.0
CE	78.1%	350.0	MATT	66.2%	489.8
MATT	77.7%	351.3	SKA	66.4%	490.2
DaliLite	78.5%	356.8	CE	65.6%	496.8
SAP	82.7%	359.2	PPM	59.1%	512.0
PPM	70.5%	366.6	SAP	57.5%	602.4

%aln: percentage of structure that is aligned. Values of FR-TM-align used as a reference are shown in bold.

5.3.7 Self-Consistency of Alignments for Homologous Proteins

Moreover, I checked whether a method that produces accurate alignments also produces self-consistent alignments, or not, by comparing the correct consistency and the average shift error of consistency in triplets of homologous sequences (as described in chapter 5.2.5) with structural similarity scores for the models of the corresponding alignments. In general, the correct consistency, measured as the percentage of positions that are consistent, was found to be moderately correlated with model quality, assessed using either the CAD, AL4 or GDT_TS scores. This observation was made both for α -helical proteins (Pearson correlations of 0.80, 0.66, and 0.78, respectively for CAD, AL4 and GDT_TS) and for β -barrel-like proteins (Pearson correlation of 0.79, 0.67, and 0.78, respectively). The average shift error of consistency is even less correlated to the quality of the homology models

for α -helical proteins (Pearson correlations of -0.36, -0.56, and -0.38, respectively for CAD, AL4 and GDT_TS) as well as for β -barrel-like proteins (-0.49, -0.75 and -0.55, respectively). Consequently, the accuracy of a set of alignments cannot be deduced from their shift error of consistency but the value of their correct consistency gives a good hint for the accuracy of the alignments since the correlation is not too bad.

Next, the ranking of the structural alignment methods based on their accuracy was compared to their ranking according to their consistency. In agreement with Sadowski (Sadowski and Taylor, 2012), alignments generated using SAP contained a higher proportion of self-consistent positions for α -helical proteins (second highest after DaliLite, see Table 5.9) even though they scored poorly in terms of accurate alignments and models according to structural similarity scores (see Table 5.4 and Table 5.5). This discrepancy can be explained by the observation that when amino acids are inconsistently aligned, the average shift error in their position tends to be large (Table 5.9). SAP therefore seems to incorrectly align some parts of the structures, but for all pairs of homologs, the errors in the structure alignments are the same. Interestingly, methods that produced the most accurate alignments according to the model scores (e.g., FR-TM-align and FATCAT rigid) also exhibited low shift errors (Table 5.9), suggesting that even though there are inconsistencies between the alignments, these errors are quite small, with the correct residue only one to four positions away. In α -helical proteins, such errors may reflect a subtle shift in the pitch of individual helices, since the repetitive nature of a helix may lead to multiple similar solutions with the helix shifted up and down by a turn.

For the set of β -barrel-proteins, the alignments were generally less consistently aligned than those of the α -helical proteins, and the average shift error was significantly higher (Table 5.9). These differences between the two folds may be explained by the higher number of residues that are loosely-packed in β -barrels, because the reduced number of constraints on their positions means that they tend to be less consistently aligned than those that are well-packed or buried, as noted previously (Sadowski and Taylor, 2012). Moreover, the internal pseudo-symmetry of a β -barrel could lead to many possible solutions that are shifted by one or two β -strands, and therefore lead to very large shift errors. Compared to the set of α -helical proteins, the ranking of the structural alignment methods according to their consistency and average shift error of consistency agrees better with the ranking observed using modeling and structural similarity scores (compare Table 5.5 with Table 5.9). DaliLite generates the most accurate alignments that are also the most consistent with the lowest shift in consistency among all structural alignment methods tested.

Table 5.9 Self-consistency of non-gapped positions in the alignments generated using different structure alignment programs.

α -helical proteins			β -barrels		
	%correct	<i>E</i>		%correct	<i>E</i>
SABERTOOTH	85.9%	0.59	DaliLite	74.2%	2.84
FR-TM-align	87.2%	0.60	FR-TM-align	71.9%	2.95
TM-align	87.2%	0.60	FATCAT	72.0%	3.32
SHEBA	87.7%	0.60	MATT	71.0%	4.46
FATCAT rigid	88.4%	0.61	SHEBA	69.5%	4.68
SAP	88.5%	0.76	PPM	48.2%	4.79
MATT	86.7%	0.84	TM-align	68.8%	5.19
DaliLite	88.7%	1.05	FATCAT rigid	68.1%	5.23
FATCAT	87.4%	1.09	MAMMOTH	51.6%	5.59
SKA	80.6%	1.36	LovoAlign	59.6%	9.17
PPM	79.4%	1.46	SAP	45.5%	9.79
MAMMOTH	69.4%	1.54	SKA	58.6%	10.06
CE	76.7%	1.79	SABERTOOTH	57.1%	10.60
LovoAlign	79.2%	2.25	CE	43.6%	26.57

Results are sorted according to the average shift error, *E*. Entries in bold indicate the highest scores in that column.

In general, consistency and accuracy were shown to be two distinct measurements that evaluate different aspects of an alignment and are not necessarily correlated with each other. A method can generate highly consistent alignments of a set of homologous protein sequence although those alignments are inaccurate but the inaccuracies can be distributed similarly among all alignments (e.g., SAP). Nonetheless, accurate alignment methods (e.g., FR-TM-align or FACAT rigid) were also shown to generate fairly consistent alignments. Interestingly, alignments of α -helical proteins were more self-consistent than those of β -barrel-proteins.

5.3.8 A Consensus Approach to Obtain Confidence Values for Aligned Positions

Structural alignments of membrane proteins have been used as reference data sets for assessing different computational approaches, like sequence alignment methods (Hill and Deane, 2012). In all these studies, a single structural alignment from a single structural alignment program was used as a reference for each pair of sequences. However, my comparisons and analysis indicate that none of the structural alignment methods performs significantly better than all other methods in producing accurate alignments neither for α -helical proteins (Table 5.4) nor for β -barrel-like proteins (Table 5.5). This observation is consistent with those that were made for the accuracy of structural alignment programs for water-soluble proteins (Berbalk, et al., 2009; Sadowski and Taylor, 2012; Slater, et al., 2012). Applying a single method to generate a reference data set may therefore result in some errors in the reference data and mislead the results for optimizations and assessments on that set. Specifically, incorrectly assigned positions would be treated as correct (false positives) and correctly-aligned positions would be treated as incorrect (false negatives). Unfortunately, such errors in the reference alignment cannot be identified easily by a manual inspection of the alignment.

In order to avoid inaccuracies in structural (reference) alignments, I propose a consensus-based structural alignment approach that considers the alignments of four structural alignment methods to generate a consensus alignment, similar to consensus approach used for transmembrane helix prediction in TOPCONS (Bernsel, et al., 2008), which was shown to produce some of the most accurate predictions. In a consensus approach, an observation (e.g., alignment of two residues) is assumed to be more likely a true observation, the more methods agree with each other on that observation. For a consensus approach to structural alignments, I selected the structural alignment methods that were shown to generate the most accurate alignments according to the GDT_TS and CAD scores: FR-TM-align, FATCAT rigid, MATT and DaliLite (see Tables 5.4 and 5.5). TM-align was excluded because the underlying algorithm and score is too similar to that of FR-TM-align and the models of FR-TM-align and TM-align were not significantly different ($p > 0.5$) from each other. Thus, TM-align does not introduce any additional information to the alignment that is already contained within FR-TM-align. Similarly, for FATCAT, the rigid-body mode was chosen instead of the flexible mode, because the former produced more accurate alignments overall (see Table 5.4 and Table 5.5).

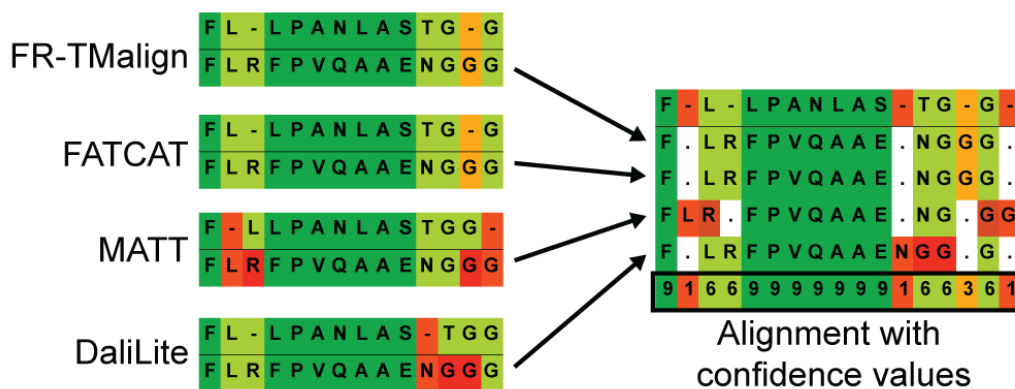


Figure 5.4 A consensus structure-based alignment fragment with confidence values. Two protein structures were aligned with four different structural alignment methods: FR-TM-align, FATCAT, MATT and DaliLite. The resulting alignments were then fused using the sequence of one of the protein structures as a reference. Depending on the agreement between the four methods, confidence values were assigned as very strong (i.e., all methods concur, confidence value of 9, *dark green*), strong (three methods agree, confidence value of 6, *pale green*), moderate (two methods agree, confidence of 3, *orange*), and weak (only one method found this solution, confidence value of 1, *red*).

For all structure pairs of HOME3, I calculated the agreement between the results of four structural alignment methods using a custom computational program. This program fuses the underlying sequence alignments of the structural alignments to a consensus alignment and calculates a confidence score for each alignment position depending on the agreement of the four alignments used. This confidence score is a measure of the reliability at each alignment position and ranges from very strong (i.e., all methods concur, confidence value of 9), strong (three methods agree, confidence value of 6), moderate (two methods agree, confidence of 3) to weak (only one method found this solution, confidence value of 1), see Figure 5.4. This script is available for download at www.bioinfo.mpg.de/AlignMe/download/ConsensusAlignment.zip.

It has to be noted that an agreement between methods is not necessarily a good reflection of the accuracy of a position, since they could all be incorrect. However, correlating the different confidence levels with the position-specific model accuracy of the corresponding positions shows that alignment positions with the highest agreement between the four methods typically correspond to accurately modeled positions, with an error in position $<4 \text{ \AA}$ (Figure 5.5). Moreover, as the confidence level decreases, so does the model accuracy (Figure 5.5). Consequently, positions with a high confidence level are the most reliable; these values could be useful for constructing “gold standard” reference alignments for evaluations of other methods on membrane proteins. Alignment positions with low confidence values should be treated with caution and potentially checked manually for their correctness.

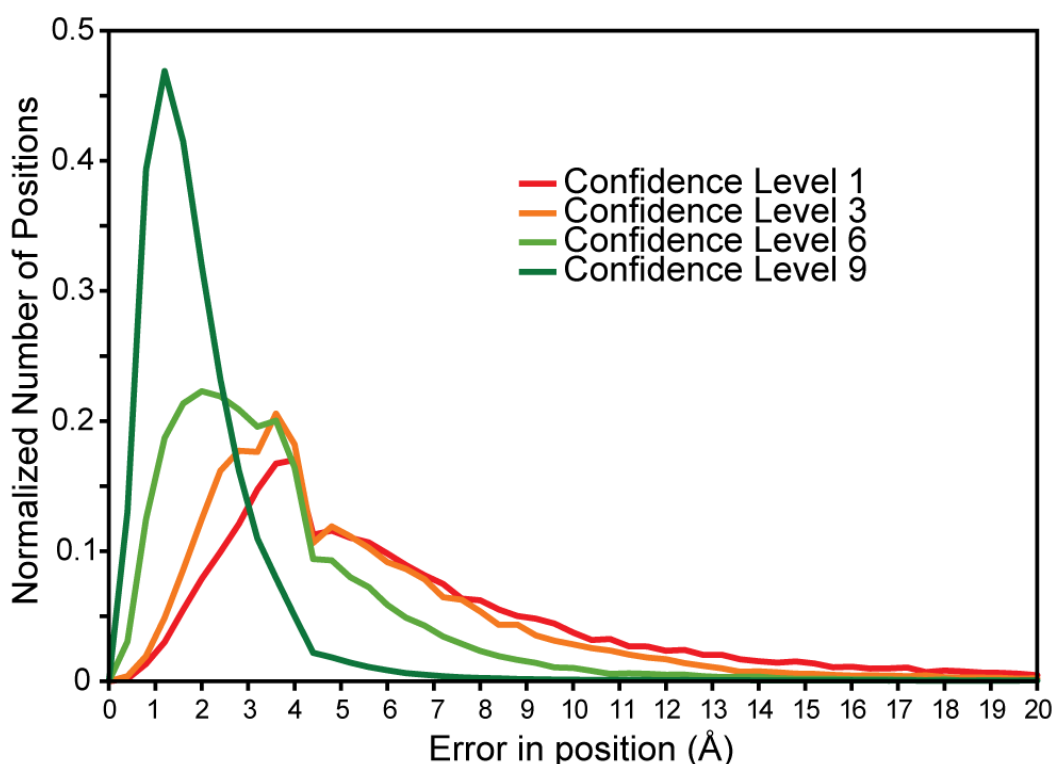


Figure 5.5 Correlation of residue accuracy with confidence values based on a consensus of FR-TM-align, FATCAT, MATT and DaliLite alignments. From all consensus alignments of the α -helical subset of HOME3, the position-specific confidence level was extracted for positions in which amino acids were aligned, i.e., excluding gapped positions. For each considered position, the distance (in Å) of the corresponding C_{α} -atom in the homology model to that in the native X-ray structure was calculated. This value then was averaged over the models built based on each of the four structural alignment methods. The plot contains the normalized distribution of averaged C_{α} -distances for each of the confidence levels (see Figure 5.4).

5.4 Discussion

In this chapter, I compared structural alignment methods in order to identify a method that is suitable for reliably aligning membrane protein structures, which should be useful for many studies of membrane protein structure prediction or analysis and shows future challenges for the structure-based alignment of membrane-protein structures. Overall, the evaluation showed that there is no single outstanding method that generates more accurate alignments than all other methods tested. This result agrees with studies on more general data sets including mainly water-soluble proteins.

In using structural similarity scores for comparing the different methods, several trends became clear. First, the selection of representative homology model out of a set of 5 models was not dependent on the similarity score (AL4, GDT_TS, CAD or DOPE score) that was used to find the best model. This result is not surprising since none of the scores was optimized on and for membrane proteins. An assessment score for homology models that considers explicitly membrane protein-

specific attributes might be more sensitive to differences between good and poor models of membrane protein structures and thus be a more efficient score for finding a good model out of a set of models.

Next, a correlation of all structural similarity scores showed that three different scores are useful for rating the structural similarity between membrane protein structures: the threshold-based AL4 score, the distant-dependent GDT_TS score that averages out large outliers and the contact-area based CAD score that considers the agreement of spheres between two protein structures. However, all of these scores showed to be correlated to a certain degree. The RMSD score, which was only weakly correlated to all other scores, was left out for a subsequent analysis because it was shown to be an inaccurate similarity measure for protein structures. Again, none of these structural similarity scores was explicitly designed for membrane proteins in mind although such membrane-specific information could give insights into the correctness of the modeled membrane-spanning segments.

Moreover, two properties of structural alignment programs were shown to have a major influence on their alignment accuracy: length-dependent scoring schemes and fragment-based alignments. First, the methods that use length-independent scores that minimize the contribution of large outliers for optimizing their superimpositions such as TM-score produced the most accurate alignments of membrane proteins. Scores that squared spatial distances were shown to generate less accurate alignments (e.g., RMSD score in CE). This observation could be biased by the use of GDT_TS, AL4 and CAD scores for assigning the accuracy of the different methods, because these scores are quite strongly correlated with the TM-score.

A second striking observation is that the fragment-based approaches typically resulted in overall more accurate alignments than the rigid-body fitting methods. This was particularly clear when comparing the alignment accuracy of the fragment-based structural alignment method FR-TM-align with its rigid-based counterpart TM-align; both being similar in all other aspects. In most cases, FR-TM-align is similar or more accurate (ranked the same or better) than TM-align, showing that using fragments for aligning structures is an effective strategy. FATCAT, in contrast typically gave more accurate alignments when it was used in its “rigid” mode, rather than in its “flexible” mode. Importantly, the fragment-based flexible mode of FATCAT was most useful for comparing structures with very large conformational differences, such as the inward- versus outward-facing conformations of the major facilitator superfamily transporters, whereas the rigid-body mode resulted in more accurate alignments when comparing structures of similar conformational states. Thus, introducing a high flexibility seems to be more useful for cases in which template and target structure are in very different conformations rather than applying it for alignments of structures that are in the same

state. Unfortunately, none of the programs includes an option that recognizes the states of the proteins (e.g., by defining the states of the input structures) although that might be useful for considering the degree of flexibility that has to be introduced for an accurate superimposition of the protein structures. FATCAT and MATT have the option to set a flag that allows for greater levels of flexibility, but they are not able to activate or deactivate this flag on their own.

I conclude that, as discussed previously for globular proteins (Collier, et al., 2014), there is room for improvement in structure alignment programs, both in terms of alignment accuracy and alignment consistency (Sadowski and Taylor, 2012). Introducing membrane information has been shown to result in accurate sequence alignments of membrane proteins (AlignMePST mode, chapter 3.3.2 and Figure 3.6) and thus might be a suitable protein descriptor that could be used within the superimposition procedure of membrane proteins for improving the alignment quality. Unfortunately, such membrane-specific information has not been applied so far in any structural alignment program. For example, for each structure, the transmembrane-spanning segments could be identified, and used as an additional criterion in the fitting. Also the application of the packing based CAD score, that has been shown to be suitable to assess protein structures in different conformations, could help to improve the alignment accuracy of membrane proteins solved in distinct conformations.

In the meantime, a consensus approach using FR-TM-align, FATCAT, MATT and DaliLite provides a useful strategy for evaluating structural alignment quality since there is no superior method that outperforms all other methods in its accuracy. Confidence scores are assigned to each alignment position and were shown to give insights into the alignment accuracy. Using such a consensus alignment with confidence values allows for producing the most accurate possible alignments of membrane protein structures. A data set of membrane proteins similar to HOMEPEP could also be built using a set of consensus alignments with confidence values. Optimizations of programs (e.g., alignment programs or membrane propensity predictors) on such a data set could include the confidence values into their calculations and the optimization could be driven by protein segments that are aligned with high confidence values. This would minimize the contribution of erroneous alignment segments into the optimization process. Moreover, the usage of confidence values allows for confirming the identification of mutations and insertions/deletions that rarely occur in proteins (e.g., gaps in membrane-spanning segments).

6 Single Insertions and Deletions (InDels) Within Membrane Segments and Their Influence on the Function of Homologous Membrane Proteins

6.1 Introduction

Membrane-spanning segments are typically treated as being evolutionarily conserved (i.e. free of insertions or deletions) by sequence alignment methods (Pirovano, et al., 2008; Shafrir and Guy, 2004) or during homology modeling (Fenollar-Ferrer, et al., 2014; Radestock and Forrest, 2011) because of their hydrophobicity and inaccessibility to aqueous solution, which are both caused by the location and interaction of the membrane protein with the hydrophobic membrane bilayer (see chapter 1.2.1). Indeed, a fully regular hydrogen-bonding pattern that stabilizes the protein backbone is typical in membrane proteins and occurs either in the form of α -helices (see chapter 1.2.2) or β -sheets (see chapter 1.2.3).

However, in a recent study, α -helical G-protein coupled receptors (GPCRs) were analyzed for structural anomalies such as bulges and kinks within their membrane-spanning segments (Gonzalez, et al., 2012). Single as well as double evolutionarily insertions or deletions (InDels) were observed when comparing homologous GPCRs within two membrane segments (TM2 and TM5). Accordingly, the twist angles of these fragments showed irregularities from the angles typical for α -helices and instead were similar to those of 3_{10} - or π -helices (see Figure 1.8 in chapter 1.2.4). In TM2, sequence fragments with twist angles ($\sim 120^\circ$) characteristic of a 3_{10} -helix and an $i \rightarrow i+3$ hydrogen bonding pattern were found in chemokine receptor (CXCR4) and μ -opioid receptor (mOR), whereas other proteins (e.g., β_1 adrenoreceptor (B1AR), β_2 adrenoreceptor (B2AR)) had twist angles ($\sim 75^\circ$) characteristic of a π -helix and an $i \rightarrow i+5$ hydrogen bonding pattern. In squid rhodopsin (s-Rhod), an even longer π -helical segment with a twist angle of 40° was found. The authors of this evaluation (Gonzalez, 2012 #53) did not report homologous proteins with corresponding residues in an α -helical conformation. As a result, alignments of TM2 containing these five proteins contained gaps in the middle of the membrane. Similarly, TM5 of one known GPCR structure (i.e. α -2 adrenergic receptor (A2AR)) contains π -helical twist angles and an $i \rightarrow i+5$ hydrogen bonding pattern at a sequence position for which in all other sequences an α -helix with an $i \rightarrow i+4$ hydrogen bonding pattern is present; therefore the A2AR sequence of TM5 is aligned to a gap in the other sequences. Both helices TM2 and TM5 have been shown to contribute to GPCR function. In TM2, a conserved S/TxP motif is

responsible for receptor activation and modulation (Govaerts, et al., 2001); the proline is crucial for receptor activation and the amino acids two positions before proline (threonine or serine) modulate the activation signal. TM5 contains residues that are involved in ligand binding and that contribute to the stabilization of the active state (Cherezov, et al., 2007). Consequently, single InDels can influence characteristic properties of a protein if they are located at such functionally-important protein segments. Additionally, single InDels cause a change in the hydrogen bonding pattern and alter the secondary structure state from the typical α -helical structure to a tighter 3_{10} -helix or a wider π -helical structure. This changes the spatial orientation of the residues involved, allowing their side chains to point into different regions than can be possible with a canonical helix.

π -helices have also been discovered at functional sites in other protein families (Cartailler and Luecke, 2004; Gonzalez, et al., 2012; Riek and Graham, 2011; Riek, et al., 2001; Weaver, 2000) like those of globular proteins (e.g., human squalene synthase, PDB code: 1EZF) or membrane proteins (e.g., cytochrome c oxidase, PDB code: 2OCC) despite their rare occurrence (< 5 %, (Cartailler and Luecke, 2004)) in protein structures. Their rarity is a consequence of their energetically less favorable conformation than the one of α -helices (Riek, et al., 2001; Weaver, 2000). The decrease of stability in π -helices is caused by a loss of side-chain to side-chain interactions in comparison to α -helices due to the enlarged radius of the helix (2.8 Å compared to 2.3 Å of an α -helix) (see Figure 1.8). Additionally, the dihedral angles of the backbone associated with π -helices (ϕ from -47 to -71 and ψ from -70 to -41) (Cartailler and Luecke, 2004; Fodje and Al-Karadaghi, 2002) are less favorable than those of α -helices. An analysis of 10 proteins that contained a π -helical structural element revealed that in 8 out of those 10 proteins the π -helical fragment was directly involved in specific binding events (Weaver, 2000). For the two other proteins, the π -helices were not directly involved in binding but were also expected to contain important amino acids (e.g., for communication between active sites or the stabilization of the protein state). Consequently, a π -helix might have a significant impact on a proteins function being worth its energetic cost compared to an energetically more favorable α -helix.

Similarly, 3_{10} -helices are also reported to be present in functional segments (e.g., copper or heme binding sites) or are assumed to be involved in signal transduction (Pal and Basu, 1999) although their packing properties are also not optimal because of non-optimal van-der-Waals contacts between residues within a 3_{10} -helix (Vieira-Pires and Morais-Cabral, 2010). However, all these studies analyzed only structural propensities of proteins and not the evolutionarily relationship between homologous proteins or whether there are InDels that are associated with irregularities of α -helices. Thus, these studies did not observe a connection between 3_{10} -helices and InDels or π -helical elements and InDels.

To date, a barrier to an automated and detailed analysis of structural irregularities in membrane-spanning segments has been the lack of high-resolution crystal structures of membrane proteins. An exact and confident spatial positioning of the amino acids is required to define structural elements as either 3_{10} -, α - or π -helical (or any other state). Additionally, to make connections between structural propensities and evolutionarily events requires the knowledge of structures of one or more homologous proteins, ideally in the same functional state. The latter is important because changes of conformation can also alternate α -helical elements into π -helical elements or 3_{10} -helices and vice versa (Armen, et al., 2003).

The issue of missing high-resolution structures assigned to homologous families is addressed by the set of homologous high-resolution membrane protein structures in HOME3 (see chapter 2.5). Moreover, my recent work shows that a combination of 4 different structural alignment methods can be used to generate a consensus alignment with confidence values, which helps to ensure that only reliable InDels will be considered for the analysis (see chapter 5.3.8). The application of a consensus alignment addresses the issue that a single structural alignment method does not align all structures correctly (see chapter 5.3). Aligned amino acids and gaps were deemed to be reliable if at least three structural alignment methods agreed with their assignment of an amino acid to another amino acid of the homologous protein sequence or to a gap. Although there still is a risk that only a single method might find the correct alignment and all other structural alignment methods fail to find a correct superimposition, positions with a high confidence score were shown to correspond to more accurately modeled residues than those with a low confidence score (see Figure 5.5).

In this chapter, all homologous proteins of the HOME3 data set are analyzed for structural irregularities that are caused by single InDels between evolutionarily related proteins (e.g., conversion of α -helical elements to 3_{10} - or π -helical structures). Subsequently, all detected InDels were analyzed for possible contribution to a protein's function (e.g., enhancement or inhibition of characteristic protein functions such as protein transport or ligand binding).

6.2 Methods

6.2.1 Computational Methods for Detecting Secondary Structure Elements

There are two ways to analyze the secondary structure propensity of residues in a protein computationally: analyzing the known three-dimensional amino acid positions of a crystal structure using a secondary structure assignment method or predicting secondary structure states based on a sequence using a prediction method that is based upon generalized properties obtained from a set of protein structures. The most commonly used secondary structure assignment methods are DSSP (Kabsch and Sander, 1983) and STRIDE (Frishman and Argos, 1995). The assignments by DSSP are based upon a “Dictionary of Secondary Structures of Proteins” that relies on hydrogen bonding patterns between consecutive nitrogen and carbonyl groups of the backbone. For each hydrogen bond, a value describing the energy of that bond is calculated using a Coulomb approximation and hydrogen bond is defined when the bond energy is below -0.5 kcal/mol. STRIDE also applies a hydrogen-bond energy function, but additionally considers the dihedral torsion angles of the backbone in order to discard hydrogen bond patterns if their ϕ and ψ angles are unfavorable. Moreover, STRIDE elongates structural elements if their adjacent amino acids have favorable angles. Both programs (DSSP and STRIDE) were optimized for the detection of regular secondary structure elements such as α -helices and β -sheets. Such consecutive and consistent regular structures (e.g., long α -helices or β -strands) are energetically preferred compared to irregular structures (e.g., coils, 3_{10} - or π -helices) or to secondary structure elements that contain only a single or a few amino acids. Long secondary structure elements were assumed to occur more frequently in a protein than very short consecutive fragments (e.g., of less than three amino acids) of alternating states. Consequently, single structural irregularities might not be detected by secondary structure assignment methods like DSSP or STRIDE. However, a recent secondary structure assignment method called SST (Konagurthu and Lesk, 2013) considers explicitly secondary structure elements that occur less frequently in proteins: 3_{10} -helices and π -helices. For a secondary structure assignment of a protein, SST applies a Bayesian method of minimum message length (MML) inference (Konagurthu, et al., 2012). Bayes’ theorem describes the probability of an event based on specific conditions. In SST, the coordinates of the protein correspond to the event and the conditions are proposed secondary structure elements based on ideal models following the Pauling and Corey geometry (Pauling, et al., 1951), which are used as a hypothesis to explain the observed data. In addition, the amount of information is minimized that is required to explain the three-dimensional coordinates of a specific amino acid in relation to the three-dimensional location of other amino acids in the protein of interest. After the detection of regular secondary structure elements like α -helices or β -strands, SST applies a

subsequent step to detect 3_{10} - and π -helices. This step is not present in DSSP and STRIDE and therefore, SST might provide useful insights into the occurrence of 3_{10} - and π -helical segments in membrane proteins.

Secondary structure prediction methods (e.g., PSIPRED or Jufo) can also be applied for obtaining information about the secondary structure properties of a protein, but those predictors are less accurate since they do not explicitly include the known three-dimensional information of the protein of interest and provide typically only a three-state prediction of a residue being in a helix, sheet or coil. There are also two predictors (SSpro8 and another eight-class structure predictor) that explicitly consider π -helices but their accuracy was still lower than 70 % and they had difficulties to predict 3_{10} - or π -helical elements (Wang, et al., 2011). Consequently, secondary structure prediction methods were not applied in this work, which accordingly was focused on atomic-resolution structures instead of sequence data.

6.2.2 Computational Methods to Detect π -helical Structure Elements

A reliable detection of π -helices is required for analyzing their influence on a protein's structure. The recent secondary structure assignment program SST supports the detection of π -helical fragments but has not been tested for its accuracy on a membrane protein data set. Another computational method that was designed explicitly for detecting π -helices in protein crystal structures is a perl script called π -HUNT (Cooley, et al., 2010). π -HUNT assigns a residue to be in a π -helical conformation if a secondary structure assignment of DSSP for that protein has found at least two sequential residues with the most likely hydrogen bonding pattern of $i \rightarrow i+5$ with strengths of ≤ 0.5 kcal/mol. Additionally, the torsion angles of those residues were required to be within the wide range of valid α -helical angles ($-180^\circ < \phi < 0^\circ$, $-120^\circ < \psi < 45^\circ$) (Pauling and Corey, 1951) to avoid incorrect annotations of π -helices in coiled segments. Unfortunately, this script was not available for download or upon request.

Consequently, I developed an in-house automated script, written in perl called π -Detector that is similar to the π -HUNT script, but that applies some more rules for detecting π -helices, in order to make a more conservative and confident assignment. π -Detector requires the three-dimensional coordinates of the protein (taken from the PDB) and an analysis of hydrogen bonding patterns and angles by DSSP of the same PDB as an input. An amino acid of a protein is assigned by π -Detector to be in a π -helical conformation if the following four observations are true:

- 1) A hydrogen bonding pattern of $i \rightarrow i+5$ (values of 5 and -5) is present in at least one of the four H-bond interaction columns in the DSSP raw output file. Two of those columns list the relative positions of the hydrogen donor interaction partners (N-H \rightarrow O) and two other columns list the relative positions of the hydrogen acceptor interaction partners (O \rightarrow H-N). The first two columns describe the most likely interactions, and the second two columns describe the second most likely hydrogen bonding patterns.
- 2) All backbone angles of the segment from $i-4$ to $i+4$ are within the range of ϕ (-180° to 0°) and ψ (-120° to 45°) according to DSSP calculations.
- 3) The C_α - C_α distances of $i \rightarrow i+2$ and $i \rightarrow i+3$ are larger than those of an ideal α -helix ($>5.4 \text{ \AA}$ for $i \rightarrow i+2$ and $>5.3 \text{ \AA}$ for $i \rightarrow i+3$ with the values in \AA representing those of a canonical α -helix).
- 4) The C_α - C_α distances of $i \rightarrow i+4$ and $i \rightarrow i+5$ are smaller than those of an ideal canonical α -helix ($<6.8 \text{ \AA}$ for $i \rightarrow i+4$ and $<9.2 \text{ \AA}$ for $i \rightarrow i+5$ with the values in \AA representing those of a canonical α -helix).

The first two rules are similar to those of the π -HUNT script. In (1) π -Detector considers not only the likeliest, but the two most likely hydrogen bonding patterns, because DSSP has a preference over an $i \rightarrow i+4$ hydrogen bonding pattern compared to an $i \rightarrow i+5$ hydrogen bonding pattern. Although DSSP is able to predict an $i \rightarrow i+5$ hydrogen bonding pattern for some residues, it assigns the first and the last residues of a π -helix as being part of the surrounding α -helices. Consequently, the minimal length of a π -helix is not met because there are too few residues left in a potential π -helical state and the π -helix stays undetected. For π -Detector, no further restrictions are made regarding the DSSP assignments that are used as input, whereas the π -HUNT script is more restrictive and requires at least two sequential residues with an $i \rightarrow i+5$ hydrogen bonding pattern with strengths of ≤ 0.5 kcal/mol according to the DSSP output (Cooley, et al., 2010). Next, in (2), the surrounding backbone dihedral angles are required to correspond to α -helical structures in order to exclude coiled segments that adopt by chance a structure similar to a π -helix. The threshold values are the same as those of the π -HUNT script (Cooley, et al., 2010). Then two more rules are applied (3 and 4), which were not used in the π -HUNT script, in order to make the assignment more reliable: C_α - C_α distances of consecutive residues have been shown to be correlated with the occurrence of π -helices in a previous study (Riek, et al., 2008).

Secondary structure assignments were generated for all proteins in the HOME3 data set using DSSP 2.0.4, STRIDE and SST. Additionally, π -Detector was used for discovering π -helices in HOME3 to test

whether DSSP and STRIDE really fail to correctly detect π -helices and to confirm the assigned π -helical positions in proteins by SST. An explicit detection of 3_{10} -helices was not implemented because it can be assumed that DSSP and STRIDE assign 3_{10} -helices properly; reports contradicting that assumption could not be found.

6.2.3 Consensus Structural Alignments for the Reliable Identification of (Single) InDels in HOME3

A meaningful comparison of secondary structure elements between homologous proteins requires a reliable alignment of those proteins. However, the evaluation of the accuracy of structural alignment programs on general protein data sets (Sadowski and Taylor, 2012; Slater, et al., 2012) as well as on HOME3 (see chapter 5) showed that all structural alignment methods had their pros and cons and that there is no superior method that can be universally relied on.

As concluded in Chapter 5.3.8, a consensus approach combining results from several structural alignment methods might therefore help to identify amino acids that are aligned consistently by several structural alignment methods. For the present analysis, consensus alignments were generated based on pairwise structural alignments that were obtained for all protein pairs in the 40 families from the α -helical subset of the HOME3 data set using DaliLite, FR-TM-align, FATCAT (rigid mode) and MATT. For the first consensus alignment, the first sequence (*a*) was used as a reference sequence to which the second sequence (*b*) from each of the four alignments was added. Differences between the four alignments of a given pair sometimes required the insertion of additional gaps into one of the second sequences (*b*). Those additional gaps are represented by a “.” within the final consensus alignment (rather than “-” for a gap from each pairwise alignment). This procedure was then repeated using the second sequence (*b*) as a reference to which the first sequences (*a*) were added.

In each consensus alignment, an alignment of two amino acids (or of an amino acid with a gap) that are matched in the same way by all structural alignment methods receives a high confidence value (i.e. 9 if all methods agree) reflecting a high probability of being correctly aligned, since the different scoring and superimposition procedures of the different structural alignment methods all agree on the alignment of that pair of amino acids. The less the structural alignment methods agree with each other, the lower the confidence value of the concerned aligned amino acids (i.e. 1 if only one method had that pair of amino acids aligned – see red colored columns in Figure 6.1A, B and C).

In a subsequent step, position specific membrane-propensity assignments from the PDB_TM database were used as a reference to define membranous and globular protein segments for each

protein sequence. All consensus alignments were then analyzed to identify single insertions and deletions (InDels) that were within membrane-spanning segments and received a high confidence score in the consensus alignment (i.e. at least 3 out of 4 methods agree with each other).

Families that contained single InDels in membrane-spanning segments were further analyzed. Based on the pairwise consensus alignments, a multiple sequence alignment was manually generated. In cases for which the consensus alignments differed from each other, the corresponding crystal structures were checked manually and a multiple sequence alignment was created upon these results (Figure 6.1D).

A		B	
2GSM,A	F T V G G V T G	2GSM,A	F T V G - G - V T G
3MK7,A - FR-TMalign	Y G M S T F E G	3OOR,B - FR-TMalign	A F L G A G . V W G
3MK7,A - FATCAT rigid	Y G M S T F E G	3OOR,B - FATCAT rigid	A F L G A G . V W G
3MK7,A - MATT	Y G M S T F E G	3OOR,B - MATT	A F L G A G . V W G
3MK7,A - DaliLite	Y G M S T F E G	3OOR,B - DaliLite	A F L G . A G V W G
Confidence	9 9 9 9 9 6 9 9	Confidence	9 9 9 9 6 6 1 9 9 9

C		D	
3MK7,A	Y G M - S - T F E G	2GSM,A	F T V G - G V T G
3OOR,B - FR-TMalign	A F L G A . G V W G	3MK7,A	Y G M S - T F E G
3OOR,B - FATCAT rigid	A F L G A . G V W G	3OOR,B	A F L G A G V W G
3OOR,B - MATT	A F L . G A G V W G		
3OOR,B - DaliLite	A F L G A . G V W G		
Confidence	9 9 9 6 6 1 9 9 9 9		

Figure 6.1 Based on consensus alignments (A-C) between cytochrome c oxidase from *Rhodobacter sphaeroides* (PDB code: 2GSM), cbb_3 cytochrome c oxidases (PDB code: 3MK7) and a nitric oxide reductase (PDB code: 3OOR) a multiple sequence alignment was manually created by assigning corresponding residues to each other (D). In case of differences between the consensus alignments, manual visual investigations on a structural level were done in order to understand the relationships between the proteins for generating the final multiple sequence alignment.

6.3 Results

6.3.1 Occurrence of Secondary Structure Elements in HOME3

The overall percentages of secondary structure assignments of DSSP, STRIDE and SST are quite similar for residues identified as α -helical (DSSP: 64.9%, STRIDE: 67.6%, SST: 67.7%) or β -strand conformations (DSSP: 3.7%, STRIDE: 3.9%, SST: 4.3%) in the structures of HOME3. On the residue level, all three methods also agree with each other for most of the protein residues that are assigned as being α -helical (29241 sequence positions are assigned by all three programs in an α -helix, Figure 6.2A, out of a total of 36477 assigned by any method as α -helical) or being in a β -strand (1842 residues out of 3741, Figure 6.2B). Interestingly, the overlap of residues being assigned to two methods is higher for DSSP and STRIDE than their agreements with SST.

All methods are also capable of detecting 3_{10} -helices (DSSP: 3.0%, STRIDE: 2.8%, SST: 1.8%) but the overlap between the three programs is lower than for residues being assigned as being α -helical or in a β -strand (241 out of 2379 assignments; compare Figure 6.2C with Figure 6.2A and B). Again, there is a strong agreement between DSSP and STRIDE and a weaker agreement of these methods with SST, which assigns fewer residues to 3_{10} -helices (865 total) than the other methods (1481 and 1577 residues, respectively).

As a fourth secondary structure type, π -helical residues were only rarely annotated by DSSP and STRIDE. DSSP assigned 11 π -helices (0.12%, 57 residues) and STRIDE only 3 π -helices (0.02%, 15 residues) that were also detected by the SST program (Figure 6.2D) that in contrast assigned 2377 residues (5.1%) from 136 proteins as being in a π -helical state (Figure 6.2D).

The π -Detector script assigned even more protein residues as being π -helical (8.2%, 3958 residues, Figure 6.3). This difference presumably reflects the fact that the π -Detector script assigns a minimal length of eight amino acids for a π -helix whereas SST also allows for smaller π -helical fragments, with as few as three. π -Detector might therefore incorrectly assign α -helical positions adjacent to a π -helix as being π -helical. SST and π -Detector agree nonetheless for 466 residues in membrane-spanning segments and 255 residues outside of the membrane. Interestingly, SST detects more non-membranous π -helical segments than π -Detector (Figure 6.3). This reflects the fact that SST does not require π -helical segments to be surrounded by α -helical segments and thus, a few coiled regions are (incorrectly) assigned as being π -helical by SST.

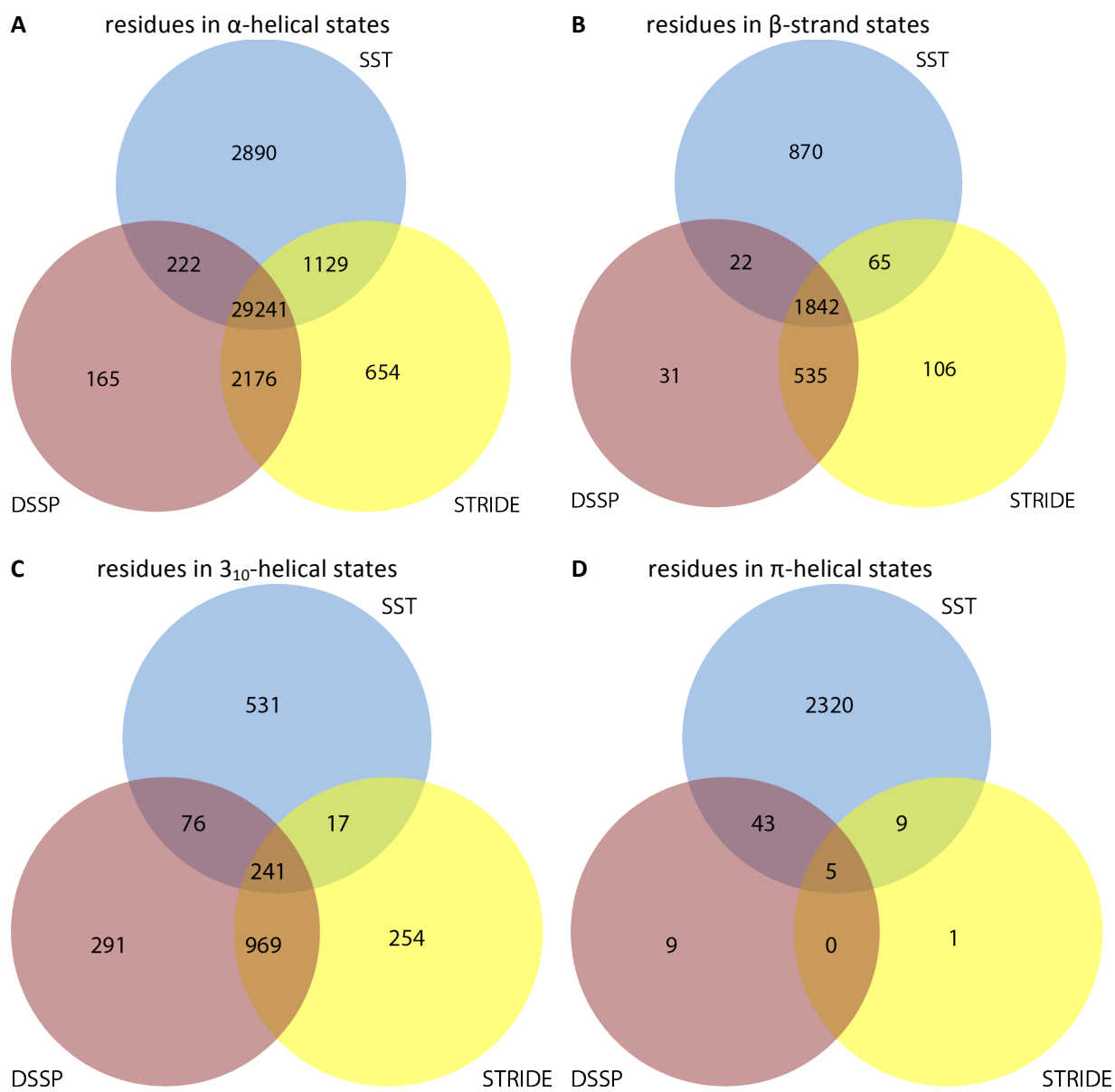


Figure 6.2 Venn Diagram of assigned secondary structure states by SST, DSSP and STRIDE. The numbers of residues were counted that were assigned by one or more secondary structure assignment programs to be in (A) an α -helix, (B) a β -strand, (C) a 3_{10} -helix or (D) a π -helix.

The usage of either π -Detector or SST has both advantages and disadvantages. The π -Detector script has strict criteria in order to detect π -helices only flanked by α -helical elements, but might incorrectly assign too many residues as a π -helical (false positives), including residues that are in a α -helical state flanking a π -helix rather than in the π -helix itself. SST is able to assign π -helical propensities to small fragments, but might also incorrectly assign residues of a coil to be in a π -helical conformation, because SST does not check the secondary structure elements adjacent to the π -helix. For these reasons, a combination of SST and the π -Detector script was used for subsequent analysis to ensure that each one of the identified π -helical elements is reliable. Specifically, secondary structure elements were treated as π -helical if they were identified as being π -helical by both approaches.

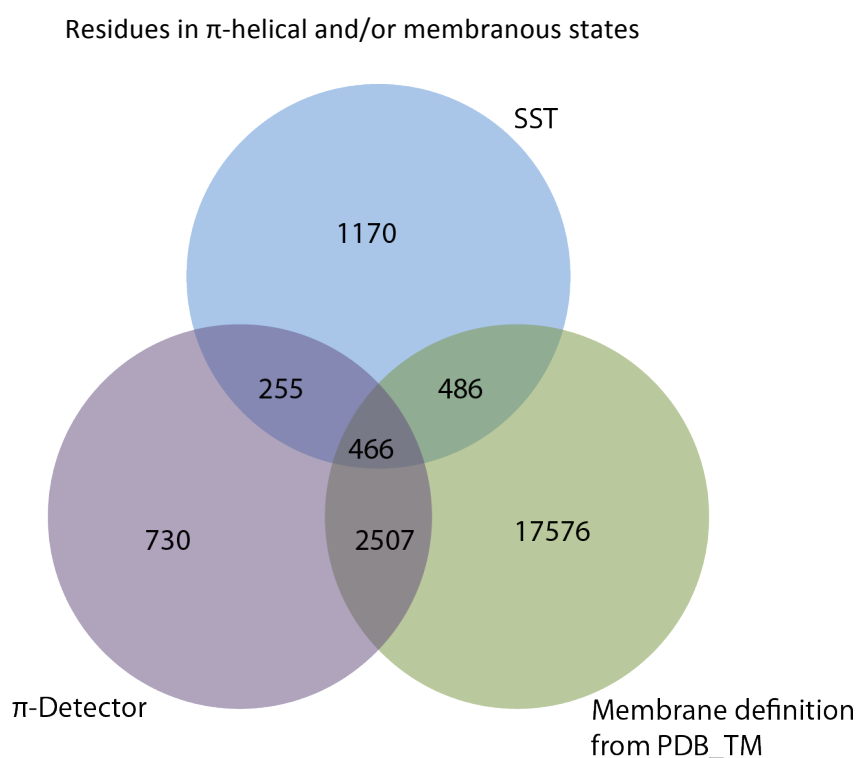


Figure 6.3 Venn Diagram of residues that were predicted by SST and/or π -Detector to be in a π -helix and/or in a membrane-spanning segment according to definitions taken from the PDB_TM database.

6.3.2 Single InDels in Confident Positions of Consensus Alignments

A reliable analysis of evolutionarily events between secondary structure elements of homologous proteins requires knowledge of the quality of the underlying alignment of those proteins. Confidently-aligned positions (i.e. confidence value ≥ 6 where at least 3 out of 4 methods agree on the alignment of two residues or of a residue against a gap) were detected for 70.4% of positions of the consensus HOME3 alignments. Considering only positions in the alignments that contain two aligned amino acids (i.e. excluding positions that are aligned against a gap) in the reference sequence, this confidence value increases to 86.3%. Accordingly, amino acids that are aligned against a gap are on average less confidently-aligned because the structural alignment methods disagree especially in choosing the proper assignment of the start and the end of the gap. This issue also occurs for some single InDels in π -helices that receive lower confidence values (confidence value of 6, Figure 6.4) than the aligned amino acids of the surrounding α -helix (confidence value of 9, Figure 6.4).

Interestingly, 50.3% of all positions that were confidently aligned were in transmembrane segments, although only 37.9% of all protein residues are located within the membrane. Thus, 93.4% of all the membrane-spanning residues are confidently-aligned showing that membrane-spanning segments are conserved and thus easier to align (structurally) by all methods than non-membranous segments which also contain coiled fragments.

Within these confidently-aligned transmembrane regions, I found a total of 166 single InDels that were manually analyzed for their correctness by visually comparing the affected protein structures with each other on a structural level. Consistent with earlier observations (Gonzalez, et al., 2012), single inDels were found in TM2 and TM5 of the family of GPCRs. Moreover, I identified conserved single InDels in TM2 and TM9 of cytochrome c oxidases (subunit I), in TM1 of LeuT relative to AdiC, and in TM5 of ECF transporters. These single InDels were subsequently analyzed for their secondary structure states and for possible implications for the protein's functionality.

2A65, chain A	L	G	N	F	L	R	F	P	V	Q
3OB6, chain A - FR-TM-align	S	G	V	F	L	-	L	P	A	N
3OB6, chain A - FATCAT rigid	S	G	V	F	L	-	L	P	A	N
3OB6, chain A - MATT	S	G	V	F	-	L	L	P	A	N
3OB6, chain A - DaliLite	S	G	V	F	L	-	L	P	A	N
Confidence	9	9	9	9	6	6	9	9	9	9

Figure 6.4 Consensus alignment of the central segment of TM1 of LeuT (2A65) and AdiC (3OB6) belonging to the family of the FIRL fold. Four different structural alignment programs were applied to align the sequence of 3OB6 against the reference sequence of 2A65. The first and the last four columns are aligned consistently by all four methods and receive a confidence value of 9. The two columns in the middle that contain a single gap are only consistently aligned by three methods (FR-TM-align, FATCAT rigid, DaliLite) and thus receive a confidence value of 6.

6.3.3 Single InDels in TM2 and TM5 of G-Protein Coupled Receptors

A membrane protein family for which single InDels have previously been observed (see chapter 6.1) is the group of G-protein coupled receptors (GPCR). GPCRs constitute one of the largest and most well-studied families among integral membrane proteins. Their main function is the transfer of endogenous and exogenous signals across the membrane by their interaction with intracellular heterotrimeric G-proteins. Those signals can be induced by a wide variety of ligands like hormones, lipids, ions or by sensory stimuli. A common structural feature of all GPCRs is their conserved 7TM fold that is present across different sub-classes of GPCRs (e.g., opioid receptors or S1P1 lipid receptors) although they share a sequence identity of ~25% or lower. In a previous study, structural differences between different GPCRs were analyzed based on a manual superimposition of their structures. This analysis showed that single InDels occur in two membrane helices of those proteins: TM2 and TM5 (Gonzalez, et al., 2012). The generated consensus alignments of the GPCR family of HOME3 confirm both observations made by Gonzalez because the consensus alignments contain single InDels with high confidence values in TM2 and TM5. However, the results obtained with the consensus alignments differ slightly from those reported previously, and will be discussed below.

According to Gonzalez (Gonzalez, et al., 2012) TM2 contains a consecutive two-residue gap in a mouse opioid receptor (m- μ OR, PDB code: 4DKL) and a human S1P1 lipid receptor (h-S1PR, PDB code: 3V2Y), compared to the crystal structure of squid rhodopsin (s-Rhod, PDB code: 2Z73), with gaps being aligned to V86 and N87 of squid rhodopsin (Figure 6.5a). In contrast, my consensus alignment shows two single gaps in TM2 that are separated by two amino acids; the gaps are aligned to V86 and F89 of s-Rhod (see Figure 6.6a). An analysis of the crystal structures shows that TM2 adopts an α -helical structure in opioid receptors like m- μ OR and a long π -helix with two more residues in rhodopsins like s-Rhod. Interestingly, the residues V86 and F89 of s-Rhod are located in the bulges of the long π -helix and occupy space that is not occupied by the residues in the α -helix of

m- μ OR. Thus, the observation of two single InDels corresponds better to the crystal structures than two consecutive gaps.

A single InDel in TM2 is observed for β -adrenergic GPCRs (e.g., h- β 2AR, PDB code: 2RH1), a dopamine D₃ receptor (h-D3R, PDB code: 3PBL), muscarinic acetylcholine receptors (e.g., h-M3R, PDB code: 4DAJ) and an human histamine H1 receptor (h-H1R, PDB code: 3RZE) in comparison to the crystal structure of s-Rhod (PDB code: 2Z73) (Gonzalez, et al., 2012) (Figure 6.5a). Those proteins contain a short π -helical structure at this segment, which is aligned to a long π -helical segment with an additional residue in s-Rhod. However, my consensus alignments show that this single gap is aligned to F89 of rhodopsin (Figure 6.6a) rather than to V86 as stated previously (Gonzalez, et al., 2012) (Figure 6.5a).

Interestingly, residues in TM2 were shown to be spatially close to the ligand binding pocket for proteins that adopt an α -helical conformation or a long π -helical conformation in TM2. For the human kappa opioid receptor (h- κ OR, PDB code: 4DJH), residues V108, T111, Q115, V118 were shown to be <4.5 Å from the ligand binding pocket (Figure 6.6a), and receptor homology models suggested that V108 and V118 are crucial for selectivity of the antagonist JD_{Tic} (Wu, et al., 2012). Similarly, a residue (Q107) of the nociceptin/orphanin FQ peptide receptor (h-ORL, PDB code: 4EA3) that is located close to the π -helix was also shown to be located spatially close to the binding pocket, and mutations of Q107 to alanine caused a ten-fold-loss in binding (Thompson, et al., 2012). In h-S1PR, Y98 and Q101 are located in a long π -helical segment and are spatially closer than 4.0 Å to the h-S1PR agonist binding pocket (Hanson, et al., 2012). In contrast to those examples, residues in TM2 of proteins with a short π -helix in TM2 were previously not supposed to be involved in ligand binding (e.g., h-D3R or h- β 2AR, Figure 6.6a).

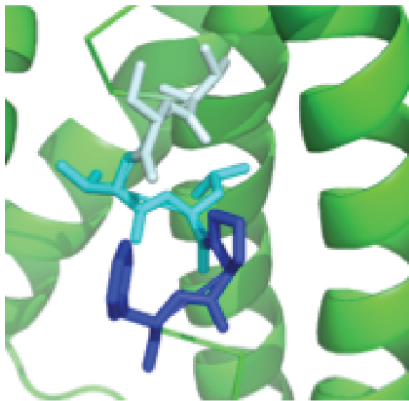
a)	alignment of membrane helix 2								b)	alignment of membrane helix 5							
4DKL	¹¹⁶ L	A	T	S	-	-	T	L	P ¹²²	3V2Y	²⁰⁷ T	-	T	V	F	T	L ²¹²
3V2Y	⁹³ L	A	G	V	-	-	A	Y	P ⁹⁹	2RH1	²⁰⁴ S	I	V	S	F	Y	V ²¹⁰
3PBL	⁷⁷ L	V	A	T	-	L	V	M	P ⁸⁴	3PBL	¹⁹³ S	V	S	S	F	Y	L ¹⁹⁹
4DAJ	¹¹⁵ I	I	G	V	-	I	S	M	N ¹²²	2Z73	²⁰⁵ F	I	L	G	F	F	G ²¹¹
3RZE	⁷⁵ I	V	G	A	-	V	V	M	P ⁸²	4DKL	²³⁷ F	I	F	A	F	I	M ²⁴³
2RH1	⁸¹ V	M	G	L	-	A	V	V	P ⁸⁸	3RZE	¹⁹⁵ A	I	I	N	F	Y	L ²⁰¹
2Z73	⁸² T	F	S	L	V	N	G	F	P ⁹⁰	4DAJ	²³⁵ A	I	I	A	F	Y	M ²⁴¹

Figure 6.5 Manually created sequence alignment of GPCRs based on visual analysis of their structures only; adapted from Figure 3 of Gonzalez et al (Gonzalez, et al., 2012). (a) Alignment of TM2 and (b) Alignment of TM5.

a) Alignment of membrane helix 2

m- μ OR (4DKL)	¹¹⁶ L A T S - T L - P F Q ¹²⁴
d- δ OR (4EJ4)	⁹⁷ L A T S - T L - P F Q ¹⁰⁵
h-S1PR (3V2Y)	⁹³ L A G V - A Y - T A N ¹⁰¹
h-ORL-1 (4EA3)	⁹⁹ L V L L - T L - P F Q ¹⁰⁷
h- κ OR (4DJH)	¹⁰⁷ L V T T - T M - P F Q ¹¹⁵
h-D3R (3PBL)	⁷⁷ L V A T L V M - P W V ⁸⁶
h-M3R (4DAJ)	¹¹⁵ I I G V I S M - N L F ¹²⁴
h-M2R (3UON)	⁷¹ I I G V F S M - N L Y ⁸⁰
h-H1R (3RZE)	⁷⁵ I V G A V V M - P M N ⁸⁴
h-A2aR (4EIY)	⁵⁴ A V G V L A I - P F A ⁶³
t- β 1AR (4AMJ)	⁸⁹ V V G L L V V - P F G ⁹⁸
h- β 2AR (2RH1)	⁸¹ V M G L A V V - P F G ⁹⁰
s-Rhod (2Z73)	⁸² T F S L V N G F P L M ⁹²

b) m- μ OR (4DKL)



c) s-Rhod (2Z73)

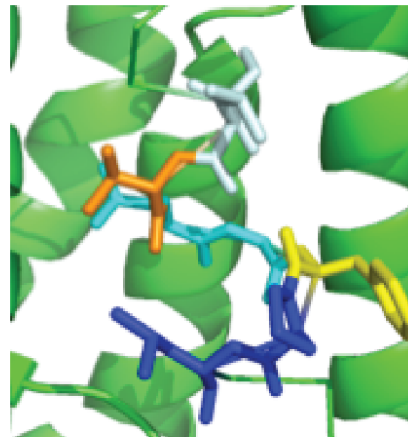


Figure 6.6 Manually created multiple sequence alignment of transmembrane helix 2 (TM2) of G-protein coupled receptors of known structure based on pairwise consensus structural alignments. (a) A sequence alignment based upon underlying structural alignments shows that single gaps occur within TM2 between different types of G-protein coupled receptors. Amino acids that were spatially close to the binding pocket are shown with a brownish background. (b) TM2 adopts an α -helical structure in opioid receptors like m- μ OR, (c) or contains a long π -helix in rhodopsins such as s-Rhod.

a) alignment of membrane helix 5

h-S1PR (3V2Y)	207	T	T	V	-	F	T	L	L	L	214
h-β2AR (2RH1)	204	S	I	V	S	F	Y	V	P	L	212
t-β1AR (4AMJ)	212	S	I	I	S	F	Y	I	P	L	220
h-D3R (3PBL)	193	S	V	V	S	F	Y	L	P	F	201
h-ORL-1 (4EA3)	220	F	L	F	S	F	I	V	P	V	228
s-Rhod (2Z73)	205	F	I	L	G	F	F	G	P	I	213
h-M2R (3UON)	191	A	I	A	A	F	Y	L	P	V	199
h-M3R (4DAJ)	235	A	I	A	A	F	Y	M	P	V	243
h-H1R (3RZE)	195	A	I	I	N	F	Y	L	P	T	203
h-A2aR (4E1Y)	182	F	F	A	C	V	L	V	P	L	190
h-κOR (4DJH)	231	F	I	F	A	F	V	I	P	V	239
m-μOR (4DKL)	237	F	I	F	A	F	I	M	P	V	245
d-δOR (4EJ4)	218	F	L	F	A	F	V	V	P	I	226

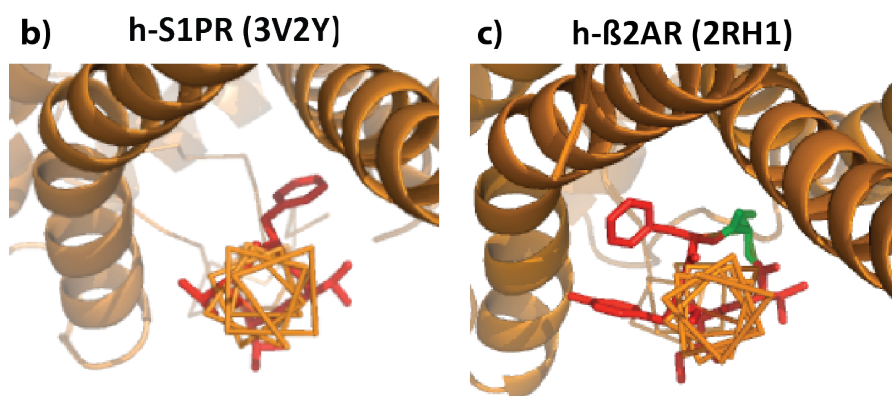


Figure 6.7 Manually created multiple sequence alignment of membrane helix 5 of G-protein coupled receptors based on pairwise consensus alignments. (a) A multiple sequence alignment based upon underlying structural alignments reveals that a h-S1PR (PDB code: 3V2Y) contains one amino acid less in TM5 compared to all other GPCRs. (b) h-S1PR contains an α -helix in TM2 whereas (c) all other GPCRs (i.e. 2RH1) contain a π -helix in TM5.

In TM5, h-S1PR (PDB code: 3V2Y) contains an α -helix which is aligned with a single gap to all other known GPCRs structures that instead contain a short π -helical segment in this helix. Interestingly, the consensus alignments show a confident gap in the sequence of the hS1PR between residues V209 and F210 (Figure 6.7a), whereas a single InDel was reported between T207 and T208 in the study of Gonzalez et al (Gonzalez, et al., 2012) (Figure 6.5b). A closer look at the crystal structures of the GPCRs shows that a single gap between V209 and F210 seems to be more reasonable. h-S1PR contains an α -helix with a phenylalanine side chain that is oriented toward the binding site (Figure 6.7b). A series of point mutations along the hydrophobic ligand binding pocket showed that a mutation F210L decreased CYM-5422 (ligand)-induced ERK phosphorylation and binding (Hanson, et al., 2012). In contrast, all other GPCRs contain a single additional residue prior to that phenylalanine, and therefore adopt a π -helical structure. The π -helix alters the orientation of the phenylalanine side

chain so that the residue prior to the phenylalanine can point into the binding pocket (e.g., S207 in β 2AR (2RH1), Figure 6.7c). In summary, this analysis confirms that single InDels can be identified reliably with consensus alignments and that an InDel can alter the orientation of side-chains in key regions of a membrane protein structure.

6.3.4 Single InDels in the Proton Pathways of Cytochrome C Oxidase Subunit I

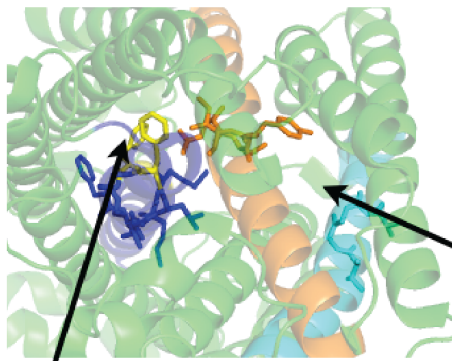
Another protein family in HOMEP3 in which two single InDels were observed based on the consensus alignments, is the family of cytochrome c oxidases (subunit I). The number of subunits in cytochrome c oxidases varies among species, but subunit I (with 12 TM helices) and subunit II (with 2 TM helices) are known to be highly conserved among various species. The cytochrome c oxidase represents complex IV of the respiratory electron transport chain and reduces oxygen to water using electrons carried by cytochrome c. The oxidase then transports protons through the membrane to establish an electrochemical potential that then can be used by the ATP synthases to synthesize ATP. At least two distinct proton pathways have been discovered so far in subunit I of the cytochrome c oxidases: the D- and K-pathways, which are both named for residues within subunit I whose mutations cause a blockage of the corresponding pathway. Specifically residue D132 (located in TM2) is the namesake of the D-pathway and residue K362 (located in TM9) is the namesake of the K-pathway (Fetter, et al., 1995; Garcia-Horsman, et al., 1995; Thomas, et al., 1993).

In the cytochrome c oxidase of *Rhodobacter sphaeroides* (PDB code: 2GSM), TM2 was shown to be involved in the active D-pathway for conducting protons to the active site and to the external bulk phase (Qin, et al., 2006). Interestingly, TM2 contains two consecutive π -helical segments (M106 to V110 and A114 to G118, Figures 6.8a and e). Similar observations can be made for the *bovine* cytochrome c oxidase (PDB code: 3AG3, (Muramoto, et al., 2010)) and the cytochrome c oxidase of *Paracoccus denitrificans* (PDB code: 3HB3, (Koepke, et al., 2009)) both also having an active D-pathway and two π -helical segments in TM2 at the evolutionarily related sequence positions (Figure 6.8e). Interestingly, the D-pathway is active neither in the *cbb₃* cytochrome c oxidases (PDB code: 3MK7, Figure 6.8b) (Buschmann, et al., 2010) that contains a single π -helix within TM2 (V64 to F68), nor in the nitric oxide reductases (PDB code: 3O0R, Figure 6.8c) (Hino, et al., 2010), which contains only regular α -helical segments in TM2. Thus, a structural variation in TM2 caused by InDels (Figure 6.8e) may be correlated with inactivation or activation of the D-pathway. However, this proposal has not been suggested before and so far experimental data is missing that shows an involvement of the π -helices in the activation of the D-pathway.

Similar results are observed for the K-pathway. The K-pathway is active in the cytochrome c oxidase of *Rhodobacter sphaeroides* (Qin, et al., 2006) and in the cbb_3 cytochrome c oxidase (Buschmann, et al., 2010), which both contain only α -helical residues in TM9 (Figure 6.8a, b and d). In contrast, the nitric oxide reductase contains a π -helical element (L320 to V324) and thus an additional amino acid (A322) in TM9 (Figure 6.8c and Figure 6.8d). This InDel may be related to inactivation of the K-pathway that is not active in nitric oxide reductase (Hino, et al., 2010), in contrast to cytochrome c oxidase and cbb_3 cytochrome c oxidase.

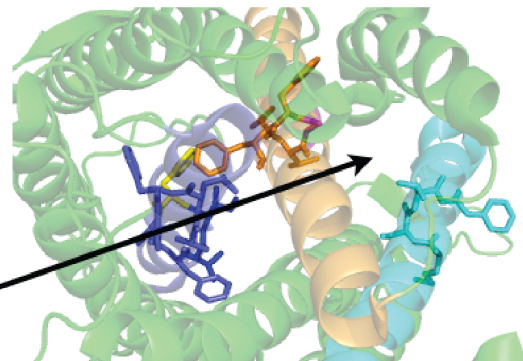
Interestingly, if these observations hold true, the cytochrome c oxidases would provide examples in which the insertion of a residue into a short π -helical segment of TM2 (e.g., in cbb_3 cytochrome c oxidase) or the insertion of two residues into a long α -helix (e.g., in nitric oxide reductase) results in a long π -helical segment with two consecutive π -helices and an activation of a pathway, whereas an insertion of a residue into an α -helical segment of TM9 results in a π -helical segment and an deactivated pathway. In other words, the insertion of a residue may lead to the opposite functional effect in terms of activation or deactivation of a pathway, depending on the position of the insertion.

a) cytochrome c oxidase (2GSM)



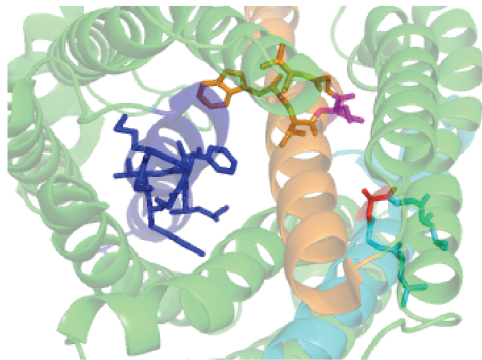
D
K-pathway: active
D-pathway: active

b) *cbb*₃ cytochrome c oxidase (3MK7)



K
K-pathway: active
D-pathway: inactive

c) nitric oxide reductase (3O0R)



K-pathway: inactive
D-pathway: inactive

d) alignment of membrane helix 9

2GSM	³⁹¹ F	T	V	G	-	G	V	T	G ³⁹⁸
3HB3	³⁸³ F	T	V	G	-	G	V	T	G ³⁹⁰
3AG3	³⁴⁸ F	T	V	G	-	G	L	T	G ³⁵⁵
3MK7	³¹⁷ Y	G	M	S	-	T	F	E	G ³²⁴
3O0R	³¹⁸ A	F	L	G	A	G	V	W	G ³²⁶

e) alignment of membrane helix 2

2GSM	¹⁰² H	G	I	L	M	M	F	F	V	V	I	P	A	L	F	G	G	F	G	N	Y ¹²²
3HB3	⁹⁴ H	G	V	L	M	M	F	F	V	V	I	P	A	L	F	G	G	F	G	N	Y ¹¹⁴
3AG3	⁶¹ H	A	F	V	M	I	F	F	M	V	M	P	I	M	I	G	G	F	G	N	W ⁸¹
3MK7	⁶⁰ H	T	N	A	V	I	F	A	F	G	G	C	A	L	-	F	A	T	S	Y	Y ⁷⁹
3O0R	⁶⁰ H	T	N	L	L	I	-	V	W	L	L	F	G	F	-	M	G	A	A	Y	Y ⁷⁸

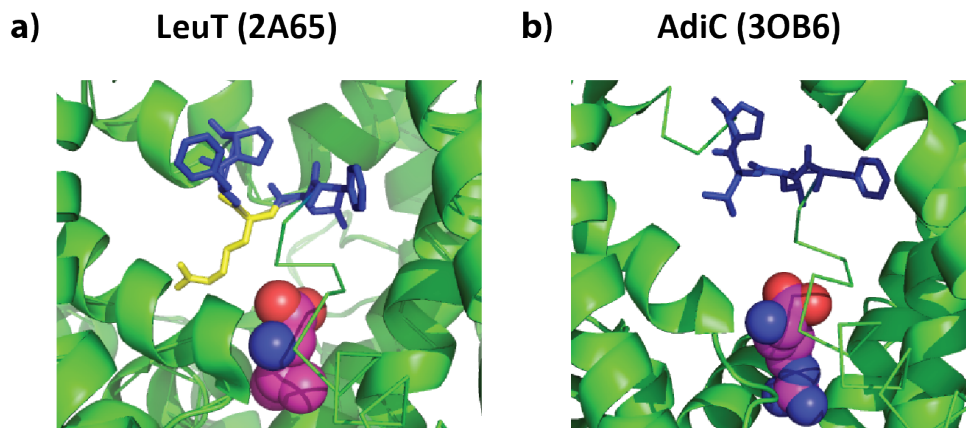
Figure 6.8 Protein structures and manually created multiple sequence alignment of cytochrome c oxidases subunit I based on pairwise consensus alignments. Structures are shown for (a) cytochrome c oxidase from *Rhodobacter sphaeroides* (PDB code: 2GSM), (b) *cbb*₃ cytochrome c oxidase (PDB code: 3MK7) and (c) a nitric oxide reductase (PDB code: 3O0R) with TM2 colored in dark blue, TM6 in orange and TM9 in light blue. Corresponding multiple sequence alignments of (d) TM9 and (e) TM2 suggest an influence of single InDels on the activation or deactivation of the D- and K-pathways.

6.3.5 A Single InDel in TM1 of LeuT Compared to AdiC

A single InDel in a membrane-spanning segment is also observed in the superfamily of the five transmembrane-helix inverted-repeat, LeuT-like (FIRL) fold proteins. As the name of the family suggests, these proteins consist of two sets of five transmembrane helices that are topologically related despite sharing a low sequence identity (Forrest and Rudnick, 2009). The inverted structural repeats allow for an alternating access mechanism and a pathway in which substrate transport across the membrane is coupled to the transport of sodium ions (e.g., LeuT, Mhp1 and vSGLT), betaine (e.g., CaiT), arginine (e.g., AdiC) or other ligands.

Two proteins sharing the FIRL fold are the bacterial L-arginine/agmatine antiporter AdiC (PDB code: 3OB6, (Kowalczyk, et al., 2011)), which belongs to the acid/polyamine/organocation (APC) transporter subfamily, and the sodium-coupled amino-acid transporter LeuT (PDB code: 2A65, (Yamashita, et al., 2005)), which belongs to the neurotransmitter/sodium symporter (NSS) subfamily. These proteins have been shown to be likely evolved from a common ancestor in a previous study (Khafizov, et al., 2010). An alignment of the first membrane helix (TM1) of the two proteins contains a single InDel with a high confidence value (Figure 6.9c). Both proteins have been crystallized in a similar but not identical conformation (LeuT: outward-facing occluded, AdiC: outward-facing open) and share a higher sequence identity of their first membrane helix (33.3% similarity) than overall (14.3% similarity based on an alignment using AlignMePST). This high sequence similarity indicates that TM1 is conserved in these two proteins and that InDels were not expected.

AdiC contains an α -helical structure in TM1 at positions S26 to N34 (Figure 6.9b) (Kowalczyk, et al., 2011), whereas there are π -helical structural elements assigned by SST for LeuT from L25 to F28 followed by α -helical elements and by π -Detector from R30 to E37 surrounded by α -helical elements. Interestingly, there is no overlap between the assignments from SST and π -Detector but they both detected a non-regular α -helical structure in TM1 of LeuT. This irregularity might be caused by the insertion of the additional arginine (R30) that points into the extracellular pathway and forms the so-called “extracellular gate” with D404 in TM10 (Figure 6.9a) (Yamashita, et al., 2005), whose mutation in other NSS transporters causes a complete loss of transport function (e.g., mutations of R69 in GAT-1, (Ben-Yona and Kanner, 2012) or R104C in SERT (Henry, et al., 2003)). Although there are many evolutionarily steps between AdiC and LeuT, the arginine in LeuT is known to be important for its function; its presence in a π -bulge was not previously commented upon, and this structural framework would seem to be important for its orientation into the pathway. The comparison with AdiC makes clear that this feature is specific to NSS transporters: a similar charged residue in AdiC might provide an electrostatic barrier to binding for its positively-charged residues.



c) alignment of membrane helix 1

LeuT (2A65)	²⁵ L	G	N	F	L	R	F	P	V	Q ³⁴
AdiC (3OB6)	²⁶ S	G	V	F	L	-	L	P	A	N ³⁴

Figure 6.9 Residues in TM1 of LeuT (2A65) and AdiC (3OB6). (a) LeuT contains an arginine that is located close to the ligand leucine (shown as spheres). (b) AdiC contains a gap in the same sequence position that is close to the ligand arginine (shown as spheres). (c) This InDel is also visible in the underlying sequence alignment that was taken from the corresponding pairwise consensus alignment.

6.3.6 A Single InDel in TM5 of ECF transporters

Energy coupling factor (ECF) transporters enable the uptake of vitamins and micronutrients (Erkens, et al., 2012) and consist of three subunits: a S-component, a T-component and an energy-coupling module. The S-component contains six transmembrane helices, provides substrate-binding specificity and is connected with a second membrane protein (the T-component) to two non-membranous nucleotide-binding domains (energy coupling modules) that are evolutionarily related to those of the ATP-binding cassette (ABC) transporters (Erkens, et al., 2012).

The HOMEP3 data set contains two structures of homologous proteins that are S-components of ECF transporters: ThiT (PDB code: 3RLB), which is specific for thiamin (Erkens, et al., 2011) and BioY (PDB code: 4DVE), which binds biotin (Berntsson, et al., 2012). Another ECF transporter structure that is not included in HOMEP3 due to its low (3.6 Å) resolution is RibU (PDB code: 3P5N), which has a binding site for riboflavin (Zhang, et al., 2010). Despite its low resolution, the structure of RibU still might reveal and/or confirm evolutionarily information in comparison to ThiT and BioY, which were both solved at high resolution (2.00 Å and 2.09 Å, respectively).

Consensus structural alignments of ThiT, RibU and BioY reveal a single gap with high confidence in a π -helical segment of TM5 of BioY that is aligned against α -helical segments of ThiT and RibU (Figure 6.10). Interestingly, K121 in TM5 of ThiT, which is located next to the single gap, has been proposed to be crucial for substrate translocation by forming a salt bridge with Q38 in loop 1 (Erkens, et al., 2011). Amino acids from TM4-6 and loop1 were also shown to interact in RibU in order to form hydrogen bonds for recognizing and binding riboflavin (Zhang, et al., 2010). The structures of both ThiT and RibU contain an amino acid with a large side chain that points into a cavity that could be crucial for transport specificity: K121 in ThiT (Figure 6.10b), and M123 in RibU (Figure 6.10c). In contrast, BioY contains a π -helical segment in TM5 with an amino acid having a small side chain (G129, Figure 6.10a and d) at a potentially important position for substrate translocation. However, the influence of TM5 on protein binding or transport has not been examined in BioY yet. I propose that the differences in TM5 contribute to the exquisite substrate transport specificity exhibited by the different ECF transporters. A biochemical analysis of the role of its additional phenylalanine (F128) via a single-point deletion, would be of interest to provide further support for this proposal.

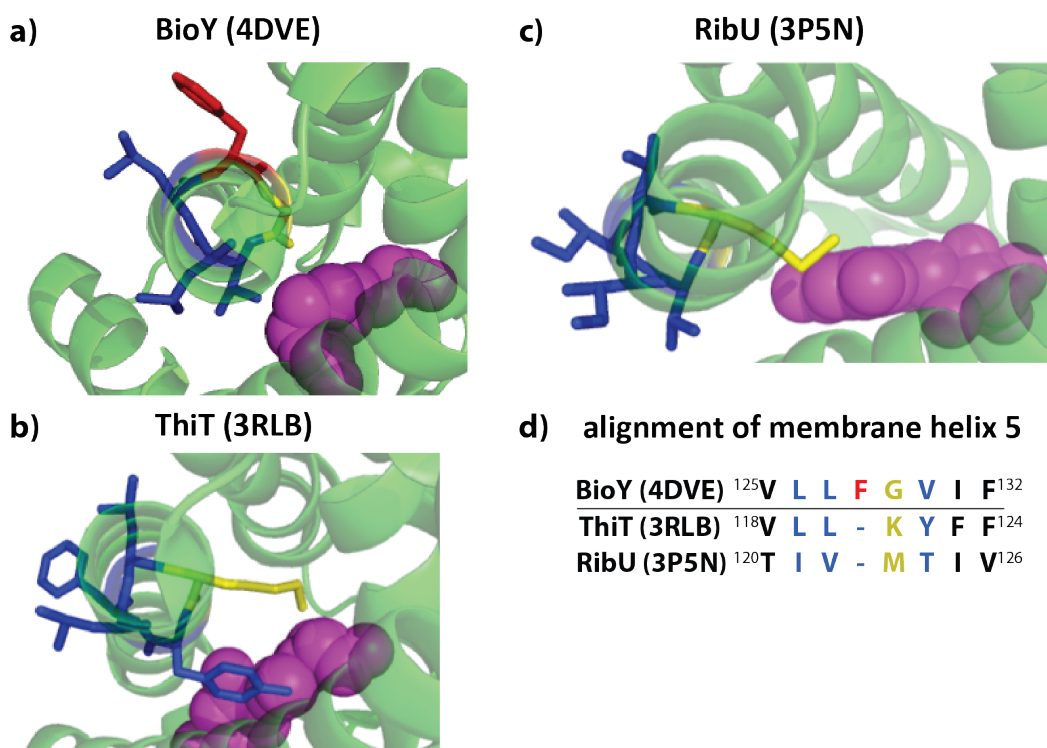


Figure 6.10 Residues in TM5 of Energy-Coupling Factor (ECF) Transporters. The bound ligand is shown as purple spheres. (a) TM5 adopts a π -helical shape in BioY (4DVE), whereas TM5 is α -helical in (b) ThiT and (c) RibU. (d) A manually created multiple sequence alignment based on pairwise consensus alignments shows the single InDel within TM5.

6.4 Discussion

To date, single evolutionarily events (InDels) have been assumed not to occur in conserved membrane-spanning segments of homologous membrane proteins. When constructing alignments for homology modeling, for example, single gaps are typically removed manually from TM segments to account for the conservation of those segments. My analysis of single InDels within TM segments reveals that this concept of fully-conserved membrane-spanning segments does not hold for all types of membrane proteins and that single InDels in transmembrane segments can be crucial for a protein's function. Consequently, membrane segments should be treated carefully during manual modifications of membrane protein alignments because single InDels may be responsible for altering a protein's function (i.e. binding, communication, transport).

The reliable detection of these InDels was made possible by the application of a consensus approach for structural alignments that allows for the identification of confidently aligned pairs of evolutionarily related amino acids of two homologous proteins. Moreover, a systematic analysis using consensus alignments on a large data set of homologous membrane proteins (HOMEP3) allowed for the identification of single InDels in TM segments. These InDels were shown to occur between homologous proteins in four different membrane protein families: GPCRs, cytochrome c oxidases, FIRL fold transporters and ECF transporters. In all cases, a single evolutionarily event (e.g., an insertion or deletion) is assumed to cause a change from an α -helical to a π -helical structure or vice versa.

These computational results are a preliminary step for further biological analysis of single InDels and π -helical structural elements in membrane proteins. Up to now, the detection of residues that are crucial for function, whether for binding affinity or pathway formation, has been tested via substitutions (e.g., by mutating large side-chains to small side-chains). The analysis of single InDels further suggests that functional properties may often be dependent on the presence of π -helical bulges that contain additional residues. The role of such residues on pathway closure could be tested using the family of cytochrome c oxidases. A residue could be inserted in TM9 of *cbb₃* cytochrome c oxidase (PDB code: 3MK7) to test whether this insertion results in a π -helical structure that deactivates the K-pathway as in the nitric oxide reductase (PDB code: 3OOR) that contains a π -helix in TM9 and but lacks a K-pathway. Also the deletion of a residue could be applied to the π -helical element in TM9 of nitric oxide reductase to test whether the structure changes to an α -helix and thereby activates a K-pathway. A role for InDels in binding specificity could be tested with BioY from the class of ECF transporters, where the phenylalanine in the π -helical segment in TM5 could be

deleted to test whether this influences the specificity of BioY to bind biotin and, if the structure of this mutant could be determined, whether the local structure changes from a π -helix to an α -helix.

More investigations could also be carried out on the relationship of the proteins belonging to the superfamily of the five transmembrane-helix inverted-repeat, LeuT-like (FIRL) fold proteins. In this study, a representative protein structure was chosen for a specific protein type without considering the conformational state of the protein. A more reliable comparison might be possible by collecting all proteins of the FIRL fold that are solved in exactly the same state (e.g., outward open). A subsequent structural comparison would then not be influenced by conformational changes. Additionally, the conservation (of functional residues) of the amino acid sequences in the FIRL fold could be checked using weblogos (evolutionarily conservation patterns) of the relevant regions if it is possible to collect a large diverse set of protein sequences for proteins of the FIRL fold.

Additionally, I recommend to test for other examples of functional properties conferred by InDels by searching for homologous protein sequences to a target protein that are highly similar (>99%) but that contain a single InDel in membrane-spanning segments. Such proteins could be tested for their specificity using transport or binding assays. Since there are only a few changes in those proteins, any differences in the results of the assay analysis are likely to be directly related to the InDels.

The increasing number of membrane proteins solved at high resolution also allows for a further detection of π -helices on an updated set of HOMEPEP in the future. This might also help to detect single InDels in other membrane protein families and to better understand the evolutionarily relationships between homologous membrane proteins.

7 Conclusion & Future Work

7.1 Improved Accuracy of Computational Methods for Membrane Proteins can be Achieved by Including Membrane Specific Information

So far, alignments of protein sequences and protein structures were optimized and tested only on general proteins data sets in which the distinct class of membrane-proteins was under-represented. This study examined the alignment accuracy of sequence (chapter 3) and structural alignment programs (chapter 5) on the important group of membrane proteins. For this, I created two different versions of the HOMEPEP data set (chapter 2), of which both were used to assess the alignment quality of membrane protein sequences or structures.

In chapter 3, I used structural reference alignments of HOMEPEP2 to identify protein descriptors that can be applied for aligning membrane protein sequences accurately. Results obtained from a database search (e.g., PSSMs from PSI-BLAST), secondary structure predictions (e.g., PSIPRED), membrane prediction methods (e.g., OCTOPUS) and other protein descriptors (e.g., substitution matrices, hydrophobicity scales) were tested for this purpose. The results showed that the inclusion of specific protein information for the alignment process should be dependent on the similarity of the proteins to be aligned. For closely related proteins, the usage of evolutionarily information in form of a PSSM (AlignMeP mode) is sufficient to generate accurate alignments, whereas proteins with lower similarity require more protein descriptors to be aligned properly. For low-homology proteins (15-45% sequence identity), the additional usage of secondary structure information besides evolutionarily information (AlignMePS mode) was shown to improve the alignment accuracy. Even more striking is the result that the additional usage of membrane-specific information in the AlignMePST mode increased the alignment accuracy for very distant homologs with a sequence identity below 15% (chapter 3.3). The alignments of AlignMe as well as models based upon them were shown to be more accurate (chapter 3.3) than those of more sophisticated alignment methods with more complex algorithms (e.g., alignments based on Hidden Markov Models from HHalign, chapter 3.2.5) although a simple Needleman-Wunsch algorithm was applied for AlignMe (chapter 3.2.1). Thus, it would be interesting to know if the inclusion of membrane-specific information in more complex computational methods (e.g., HHalign or HMAP) would increase their alignment accuracy for distant homologs.

A similar conclusion as for sequence alignment methods can be made for the structural alignment methods that were analyzed in chapter 5 for their ability to align membrane protein structures

accurately. Several programs were shown to generate accurate alignments but there was no outstanding method that performed significantly better than all other methods.

This result could be caused by the fact that none of these structural alignment methods included explicit information about membrane-spanning segments or was optimized on a set of membrane proteins. For AlignMe, the inclusion of membrane-specific information increased the alignment accuracy significantly and accordingly, I suggest to test if the inclusion of membrane-specific attributes (e.g., membrane propensities, hydrophobicity etc.) into the structural alignment process increases the alignment accuracy of membrane proteins. The fold of homologous membrane proteins in a family is mainly defined by the membrane-spanning helices. A preference of aligning residues of membrane-spanning helices with each other compared to residues located within coiled segments might help to narrow down the search space for a correct superimposition. Another reduction of the search space could be achieved by the definition that large coiled segments (more than 10 residues) could be excluded from being located in a membranous environment of the aligned structures. Short coiled segments should still be allowed in membranous segments because they have been observed in several membrane protein families (e.g., unwound segments in the middle of TM1 and TM6 in LeuT from the FIRL fold or two re-entrant loops in proteins belonging to the family of aquaporins).

7.2 Using Anchors on Known Conserved Residues in Pairwise Alignments of AlignMe could Improve the Alignment Accuracy

For some membrane proteins, important residues or motifs (e.g., those that are involved in ligand binding) are known to be conserved and thus are also known to be aligned with each other in homologous membrane proteins even without knowing the rest of the alignment. An example can be found in the family of the G-Protein coupled receptors that contain an E/DRY-motif that regulates ligand binding and is involved in a subsequent conformational change of the receptor (Rovati, et al., 2007). Conserved motifs in distantly related proteins can also be identified using the alignment of family-averaged hydropathy profiles, which I implemented in the AlignMe web server (chapter 4.5). In those family-averaged hydropathy alignments, local conserved segments are aligned reliably but in global alignments non-conserved segments might be incorrectly aligned.

Consequently, it would be of advantage if known conserved residues could be fixed for an alignment process by aligning these residues first and building the rest of the alignment using these restrictions. Currently, a modified version of AlignMe, in which anchors with varying strength are used to guide

the alignment to align pre-defined residues with each other, is tested by Rene Staritzbichler and Kamil Khafizov.

7.3 Novel Membrane Protein Descriptors are Available and could be Tested Using AlignMe

There also are continuous improvements of other computational methods that describe membrane protein properties. Thus, new programs are available that have not been tested for their applicability to describe membrane protein properties for alignments. Among these methods are database search programs like HHblits (Remmert, et al., 2011), the secondary structure prediction method SPINE-X (Faraggi, et al., 2012), membrane prediction methods like HMMpTM (Tsaousis, et al., 2014) or WRF-TMH (Hayat and Khan, 2013) and a combined secondary structure and membrane prediction method called BCL::Jufo9D (Leman, et al., 2013). I suggest to test whether the data they produce improve the alignment accuracy of AlignMe. The outputs of these programs should be tested alone as well as in combination with the current best alignment descriptors for AlignMe.

Additionally, combinations of two or more input descriptors of the same type could be tested. So far, only distinct input descriptors have been tested in combination with each other (e.g., a secondary structure prediction with a membrane prediction) but also two similar protein descriptors (e.g., the secondary structure predictors PSIPRED and Jufo) could be used at the same time for an alignment. A consensus prediction could be created based upon two predictions in which a position with a similar assignment by both predictors get a higher confidence value than a position for which the predictors disagree. Using confidence values for the inclusion of a prediction into the alignment process could emphasize consistently and therewith correctly predicted propensities.

Moreover, AlignMe was optimized for aligning α -helical membrane proteins and is therefore not suitable for aligning the distinct group of β -barrels, since their sequences and structures exhibit distinct properties (chapter 1.2.3). The membrane propensity prediction method OCTOPUS that is used in AlignMe is only able to predict α -helical membrane-spanning segments (chapter 3.2.2.4). For an alignment of β -barrel-like membrane proteins, different membrane propensity predictors like BOCTOPUS (Hayat and Elofsson, 2012) and TMBpro (Randall, et al., 2008) that are able to predict membrane-spanning segments for β -barrels need to be tested. Additionally, the evaluation of structural alignment methods also showed slightly different results between the groups of α -helical- and β -barrel-like membrane proteins. DaliLite was shown to be more suitable for aligning β -barrels rather than for aligning α -helical membrane proteins (chapter 5.3). Accordingly, a set of parameters

that is suitable for aligning β -barrels could be identified and optimized similarly as it has been done previously for α -helical proteins.

7.4 Gap Penalties could be Optimized for AlignMe on a Set of Consensus Alignments with Confidence Values

Besides protein descriptors, gap penalties are required for generating a pairwise sequence alignment with a Needleman-Wunsch algorithm as implemented by AlignMe. In chapter 3, gap penalties were optimized against a reference set of pairwise sequence alignments of the HOME2 data. Those sequence alignments were based on general structure alignments from the structural alignment method SKA (chapter 2.4).

An optimization process for gap penalties requires correct structural reference alignments. However, there were errors in the alignments of proteins belonging to the group of GPCRs and consequently similar errors were present in alignments of AlignMe for this protein family. This result suggests that the optimization of gap penalties should be carried out against a consensus alignment with confidence values so that erroneous data from the structural reference alignments does not influence the optimization process. For example, a change of gap penalties could be considered to increase the alignment accuracy only if amino acids from positions sharing a high confidence value are aligned more accurately (e.g., have a smaller shift) than before.

Aside from optimizing gap penalties, I also propose to optimize the weights that are used to adjust the input descriptors for the alignment (e.g., predictions) towards each other. The increase of a specific protein attribute (e.g., membrane propensity) might also increase the alignment accuracy (or vice versa). Unfortunately, a computational solution to address an optimization of weights and gap penalties at the same time is currently not available because the search space increases by the power of N , with N being the number of parameters to be optimized (as mentioned in chapter 3.2.4).

7.5 Several Structural Similarity Scores could be Tested for their Ability to Align Membrane Protein Structures

A property of structural alignment programs that has been shown to contribute significantly to their alignment accuracy is their internal scoring scheme that is applied during the superimposition process of two structures (chapter 5.3.3). My analysis on the HOME3 data set showed that scores that averaged out large outliers (e.g., TM-score, DaliLite score) were shown to be more adequate for

aligning proteins than scores that squared the differences between the two structures (e.g., RMSD, URMS).

Interestingly, the TM-score that was applied in FR-TM-align and TM-align is also available for assessing the quality of homology models and showed a high correlation to other structural similarity scores like the GDT_TS, AL0 or AL4 score (chapter 5.3.2). These other scores might also be suitable being used during the alignment process of proteins structures. Thus, I suggest to test these structural similarity scores for their ability to describe the similarity of two different protein structures during a superimposition process and to compare their accuracy with each other (e.g., by replacing the TM-score in FR-TM-align with the GDT_TS score).

Another score that showed a correlation with the TM-score is the CAD score (chapter 5.3.2) that compares contact areas of amino acids in two protein structures with each other. The CAD score was shown to capture conformational changes of proteins (i.e. repositioning of helices or domains relative to each other) better than a spatial-based structural similarity score (e.g., GDT_TS). A new approach for aligning structures of membrane proteins could be the inclusion of the CAD score into the alignment process of a protein structural alignment method. The application of such an environmental based structural similarity score could facilitate the superimposition process of two similar structures being solved in different states and could result in more accurate alignments. For membrane proteins, it would also be useful to have a CAD score that considers the membranous environment of membrane-spanning segments in its contact areas.

7.6 π -helices should be Given More Importance in Computational Methods

The updated HOME3 data set (chapter 2.5) and the combination of different structural alignments in a consensus alignment with confidence values (chapter 5.3.8) allowed the identification of single insertions and deletions (InDels) in membrane-spanning protein segments as described in chapter 6. Single InDels were observed at alignment positions in which π -helical segments were aligned to α -helical segments. Additionally, those single InDels were observed to occur at positions that contribute to a protein's function. These observations suggest that secondary structure assignment methods should account for these structural differences by including π -helical definitions in their assignments instead of averaging them out as in DSSP or STRIDE. Similarly, I recommend the inclusion of irregular secondary structure elements like 3_{10} -helices or π -helices in secondary structure prediction methods so that those predictors also account for these important structural elements.

Additionally, current homology model approaches or alignment processes that eliminate or forbid single gaps in membrane-spanning segments should be revised and checked for their biological

validity. Sequence alignment methods like AlignMe could also be tested for including more than the current three secondary structure states (helix, sheet and coil) into their alignment process (i.e. by also including π -helical or 3_{10} -helical propensities). A further extension of AlignMe could be the application of different gap penalties for all secondary structure types instead of using only one secondary structure type as it is currently implemented in the AlignMePST mode that uses six gap penalties dependent on the assignment of positions to be in a α -helical, non- α -helical or terminal segments (chapter 3.3.2). Additionally, it could make sense to favor an alignment of a π -helical segment against a single gap in an α -helix to better account for single insertions or deletions in the alignment. Similarly, structural alignment methods could consider structural irregularities by allowing single insertions or deletions in their alignments if they are surrounded by regular α -helical elements.

7.7 Membrane Protein Reference Data Sets Require Constant Updates and Reliable Structural Alignment Methods

A future evaluation of sequence and structural alignment methods might require an additional update of the HOMEP data sets. The HOMEP2 data set of 2010 already had to be updated in 2013 to the HOMEP3 data set due to the increased number of newly solved high-resolution structures of membrane proteins. However, since the generation of HOMEP3 in March 2013, 564 new non-unique membrane protein structures (519 α -helical and 45 β -barrel-like proteins) were added to the PDB_TM database (version: 2015-02-20), which have not been assigned so far to any existing or new family of the HOMEP data set. Regarding this steady increase of crystallized membrane protein structures, I suggest to update HOMEP on a regular basis for considering all available protein structure information. The semi-automated scripts that I developed to generate HOMEP allow for such regular updates, although the step of generating pairwise structural alignments between all proteins is computationally expensive and I recommend limiting updates to a yearly basis. In chapter 2.5, I also demonstrated that using two structural alignment methods (i.e. SKA and TM-align) consolidate the confidence of proteins being assigned to a specific family. One idea to improve this clustering approach further was the use of more structural alignment methods. In chapter 5.3.8, I described an approach to generate consensus alignments with confidence values for each pair of proteins in HOMEP3 and used them in chapter 6 to detect confident single insertions and deletions in in membrane-spanning segments. Accordingly, a consensus alignment could also be used for clustering proteins to a specific family if only confident alignment positions are used for calculating the similarity scores that are used for the clustering process (i.e. PSD and TM-score). The manual step for checking the correct assignment of proteins to a specific family might thereby become unnecessary. The analysis of structural alignments from proteins being solved in different conformations (chapter 5.3.5) suggests that another potential improvement during the clustering

process might be achieved by aligning proteins that were crystallized within the same conformation with rigid alignment methods (e.g. FATCAT nonflexible) and proteins that were crystallized within different states (e.g., inward facing vs. outward facing conformation) with flexible, fragment-based structural alignment methods (e.g., FR-TM-align). I suggest to use different structural alignment methods depending on the conformations of the aligned proteins in order to increase the number of aligned positions that receive a high confidence score resulting in an increased accuracy of the reference alignments. However, the knowledge about the state in which a protein is crystallized has to be known beforehand, but this information is not available by the data stored in the Protein Data Bank. I support the idea of a computational approach for detecting exact protein conformations (e.g., inward-facing vs. outward-facing) because such a method might help to gain deeper insights into the relationships between two protein structures.

7.8 The AlignMe Web Server Needs to be Updated with the Latest Developments

The AlignMe web server that I made available for an easy access to AlignMe and its different alignment modes (chapter 4) had 1330 visitors in 2014, including both new and returning users. Recently, the AlignMe web server was used to align terminal domains of serotonin transporters (Fenollar-Ferrer, et al., 2014), homologs of the vesicular monoamine transporter (Yaffe, et al., 2014) and for other proteins using the mode to align family-averaged hydropathy profiles (see chapter 4.5.4 for more examples). This shows it was useful to implement a web server for AlignMe to get potential users aware of AlignMe and also to use it. Consequently, the AlignMe web server also needs to be updated with all changes that will be applied to the AlignMe software or to the HOMEP data set in future so that visitors of the website are always able to use the most recent versions of AlignMe and HOMEP. Potential future updates of the AlignMe web server could include the application of anchors to pairwise alignments (chapter 7.2), updated alignment descriptors (chapter 7.3) or updated gap penalties (chapter 7.4). Additionally, it would be useful for the scientific community to get easy access via a web server to two approaches that were explained in this study. First, a web server could be set up that allows users to detect π -helical segments in protein structures. Users would have to submit a three-dimensional coordinate file of the protein (e.g., via a PDB file) to the server for which a detection of π -helical conformations is then performed using the π -Detector script (see chapter 6.2.2). The output could be a list of residues within π -helical conformations and a three-dimensional coordinate file with a highlighting on those residues. Next, a web server that easily allows for generating consensus alignments as described in chapter 5.3.8 would be useful. Users would have to submit two protein structures that are then aligned using four different structural alignment methods. The output could be a consensus alignment with confidence values as well as the four different structure alignments that were generated.

A. Appendix

Table A.1. Proteins in the HOME2 data set, listed by family

#	PDB ID	Name	Source	Res (Å)
0	2R6G	THE CRYSTAL STRUCTURE OF THE E. COLI MALTOSE TRANSPORTER	<i>ESCHERICHIA COLI K12</i>	2.80
0	3D31	MODBC FROM METHANOSARCINA ACETIVORANS	<i>METHANOSARCINA ACETIVORANS</i>	3.00
0	2ONK	ABC TRANSPORTER MODBC IN COMPLEX WITH ITS BINDING PROTEIN MODA	<i>ARCHAEOGLOBUS FULGIDUS</i>	3.10
1	2NQ2	AN INWARD-FACING CONFORMATION OF A PUTATIVE METAL-CHELATE TYPE ABC TRANSPORTER	<i>HAEMOPHILUS INFLUENZAE</i>	2.40
1	2QI9	ABC-TRANSPORTER BTUCD IN COMPLEX WITH ITS PERIPLASMIC BINDING PROTEIN BTUF	<i>ESCHERICHIA COLI</i>	2.60
2	3B9W	THE 1.3 Å RESOLUTION STRUCTURE OF NITROSOMONAS EUROPAEA RH50 AND MECHANISTIC IMPLICATIONS FOR NH ₃ TRANSPORT BY RHESUS FAMILY PROTEINS	<i>NITROSOMONAS EUROPAEA</i>	1.30
2	1U7G	CRYSTAL STRUCTURE OF AMMONIA CHANNEL AMTB FROM E. COLI	<i>ESCHERICHIA COLI</i>	1.40
2	2B2H	AMMONIUM TRANSPORTER AMT-1 FROM A. FULGIDUS (AS)	<i>ARCHAEOGLOBUS FULGIDUS</i>	1.54
3	2W2E	1.15 ÅNGSTROM CRYSTAL STRUCTURE OF P.PASTORIS AQUAPORIN, AQP1, IN A CLOSED CONFORMATION AT PH 3.5	<i>PICHIA PASTORIS</i>	1.15
3	2F2B	CRYSTAL STRUCTURE OF INTEGRAL MEMBRANE PROTEIN AQUAPORIN AQP1 AT 1.68 Å RESOLUTION	<i>METHANOTHERMOBACTER MARBURGENSIS STR.</i>	1.68
3	3GD8	CRYSTAL STRUCTURE OF HUMAN AQUAPORIN 4 AT 1.8 Å AND ITS MECHANISM OF CONDUCTANCE	<i>HOMO SAPIENS</i>	1.80
3	2B6O	ELECTRON CRYSTALLOGRAPHIC STRUCTURE OF LENS AQUAPORIN-0 (AQP0) (LENS MIP) AT 1.9 Å RESOLUTION, IN A CLOSED PORE STATE	<i>OVIS ARIES</i>	1.90
3	2O9G	CRYSTAL STRUCTURE OF AQPZ MUTANT L170C COMPLEXED WITH MERCURY.	<i>ESCHERICHIA COLI</i>	1.90
3	3D9S	HUMAN AQUAPORIN 5 (AQP5) - HIGH RESOLUTION X-RAY STRUCTURE	<i>HOMO SAPIENS</i>	2.00
3	3LLQ	AQUAPORIN STRUCTURE FROM PLANT PATHOGEN AGROBACTERIUM TUMEFACIENS	<i>AGROBACTERIUM TUMEFACIENS STR. C58</i>	2.01
3	3CN5	CRYSTAL STRUCTURE OF THE SPINACH AQUAPORIN SOPIP21 S115E, S274E MUTANT	<i>SPINACIA OLERACEA</i>	2.05
3	3C02	X-RAY STRUCTURE OF THE AQUAGLYCEROPORIN FROM PLASMIDIUM FALCIPARUM	<i>PLASMIDIUM FALCIPARUM</i>	2.05
3	1LDF	CRYSTAL STRUCTURE OF THE E. COLI GLYCEROL FACILITATOR (GLPF) MUTATION W48F, F200T	<i>ESCHERICHIA COLI</i>	2.10
3	3KLY	PENTAMERIC FORMATE CHANNEL	<i>VIBRIO CHOLERAE</i>	2.10
3	1J4N	CRYSTAL STRUCTURE OF THE AQP1 WATER CHANNEL	<i>BOS TAURUS</i>	2.20
4	1M0L	BACTERIORHODOPSIN/LIPID COMPLEX AT 1.47 Å RESOLUTION	<i>HALOBACTERIUM SALINARUM</i>	1.47
4	2JAF	GROUND STATE OF HALORHODOPSIN T203V	<i>HALOBACTERIUM SALINARIUM</i>	1.70
4	1H2S	MOLECULAR BASIS OF TRANSMEMBRANE SIGNALLING BY SENSORY RHODOPSIN II-TRANSDUCER COMPLEX	<i>NATRONOMONAS PHARAONIS</i>	1.93
4	1XIO	ANABAENA SENSORY RHODOPSIN	<i>NOSTOC SP. PCC 7120</i>	2.00
4	3A7K	CRYSTAL STRUCTURE OF HALORHODOPSIN FROM NATRONOMONAS PHARAONIS	<i>NATRONOMONAS PHARAONIS DSM 2160</i>	2.00
4	2E14	TRIMERIC COMPLEX OF ARCHAERHODOPSIN-2	<i>HALOBACTERIUM SP. AUS-2</i>	2.10
5	1OTS	STRUCTURE OF THE ESCHERICHIA COLI CLC CHLORIDE CHANNEL AND FAB COMPLEX	<i>ESCHERICHIA COLI</i>	2.51
5	1KPL	CRYSTAL STRUCTURE OF THE CLC CHLORIDE CHANNEL FROM S. TYPHIMURIUM	<i>SALMONELLA TYPHIMURIUM</i>	3.00
6	2ZT9	CRYSTAL STRUCTURE OF THE CYTOCHROME B6/F COMPLEX FROM NOSTOC SP. PCC 7120	<i>NOSTOC SP. PCC 7120</i>	3.00
6	1Q90	STRUCTURE OF THE CYTOCHROME B6/F (PLASTOHYDROQUINONE : PLASTOCYANIN OXIDOREDUCTASE) FROM CHLAMYDOMONAS REINHARDTII	<i>CHLAMYDOMONAS REINHARDTII</i>	3.10
7	3CX5	STRUCTURE OF COMPLEX III WITH BOUND CYTOCHROME C IN REDUCED STATE AND DEFINITION OF A MINIMAL CORE INTERFACE FOR ELECTRON TRANSFER.	<i>SACCHAROMYCES CEREVISIAE</i>	1.90
7	2A06	BOVINE CYTOCHROME BC1 COMPLEX WITH STIGMATELLIN BOUND	<i>BOS TAURUS</i>	2.10

7	2QJY	CRYSTAL STRUCTURE OF RHODOBACTER SPHAEROIDES DOUBLE MUTANT WITH STIGMATELLIN AND UQ2	<i>RHODOBACTER SPHAEROIDES</i>	2.40
7	3L70	CYTOCHROME BC1 COMPLEX FROM CHICKEN WITH TRIFLOXYSTROBIN BOUND	<i>GALLUS GALLUS</i>	2.75
8	1V54	BOVINE HEART CYTOCHROME C OXIDASE AT THE FULLY OXIDIZED STATE	<i>BOS TAURUS</i>	1.80
8	2GSM	CATALYTIC CORE (SUBUNITS I AND II) OF CYTOCHROME C OXIDASE FROM RHODOBACTER SPHAEROIDES	<i>RHODOBACTER SPHAEROIDES</i>	2.00
8	3HB3	HIGH RESOLUTION CRYSTAL STRUCTURE OF PARACOCCLUS DENITRIFICANS CYTOCHROME C OXIDASE	<i>PARACOCCLUS DENITRIFICANS</i>	2.25
9	1V54	BOVINE HEART CYTOCHROME C OXIDASE AT THE FULLY OXIDIZED STATE	<i>BOS TAURUS</i>	1.80
9	1M56	STRUCTURE OF CYTOCHROME C OXIDASE FROM RHODOBACTOR SPHAEROIDES (WILD TYPE)	<i>RHODOBACTER SPHAEROIDES</i>	2.30
9	1QLE	CRYO-STRUCTURE OF THE PARACOCCLUS DENITRIFICANS FOUR-SUBUNIT CYTOCHROME C OXIDASE IN THE COMPLETELY OXIDIZED STATE COMPLEXED WITH AN ANTIBODY FV FRAGMENT	<i>PARACOCCLUS DENITRIFICANS</i>	3.0
10	2DYZ	BOVINE HEART CYTOCHROME C OXIDASE AT THE FULLY OXIDIZED STATE	<i>BOS TAURUS</i>	1.80
10	2GSM	CATALYTIC CORE (SUBUNITS I AND II) OF CYTOCHROME C OXIDASE FROM RHODOBACTER SPHAEROIDES	<i>RHODOBACTER SPHAEROIDES</i>	2.00
10	3HB3	HIGH RESOLUTION CRYSTAL STRUCTURE OF PARACOCCLUS DENITRIFICANS CYTOCHROME C OXIDASE	<i>PARACOCCLUS DENITRIFICANS</i>	2.25
11	3B44	CRYSTAL STRUCTURE OF GLPG W136A MUTANT	<i>ESCHERICHIA COLI</i>	1.70
11	2NR9	CRYSTAL STRUCTURE OF GLPG, RHOMBOID PEPTIDASE FROM HAEMOPHILUS INFLUENZAE	<i>HAEMOPHILUS INFLUENZAE 86-028NP</i>	2.20
12	2RH1	HIGH RESOLUTION CRYSTAL STRUCTURE OF HUMAN B2-ADRENERGIC G PROTEIN-COUPLED RECEPTOR.	<i>HOMO SAPIENS, ENTEROBACTERIA PHAGE T4</i>	2.40
12	2Z73	CRYSTAL STRUCTURE OF SQUID RHODOPSIN	<i>TODARODES PACIFICUS</i>	2.50
12	3KJ6	CRYSTAL STRUCTURE OF A METHYLATED BETA2 ADRENERGIC RECEPTOR- FAB COMPLEX	<i>HOMO SAPIENS</i>	3.40
12	2VT4	TURKEY BETA1 ADRENERGIC RECEPTOR WITH STABILISING MUTATIONS AND BOUND CYANOPINDOLOL SUGAR FREE LACTOSE PERMEASE AT NEUTRAL PH	<i>MELEAGRIS GALLOPAVO</i>	2.7
13	2CFQ	SUGAR FREE LACTOSE PERMEASE AT NEUTRAL PH	<i>ESCHERICHIA COLI</i>	2.95
13	1PW4	CRYSTAL STRUCTURE OF THE GLYCEROL-3-PHOSPHATE TRANSPORTER FROM E.COLI	<i>ESCHERICHIA COLI</i>	3.30
14	2J8S	DRUG EXPORT PATHWAY OF MULTIDRUG EXPORTER ACRB REVEALED BY DARPIN INHIBITORS	<i>ESCHERICHIA COLI</i>	2.54
14	2V50	THE MISSING PART OF THE BACTERIAL MEXAB-OPRM SYSTEM: STRUCTURAL DETERMINATION OF THE MULTIDRUG EXPORTER MEXB	<i>PSEUDOMONAS AERUGINOSA PA01</i>	3.00
15	3EAM	AN OPEN-PORE STRUCTURE OF A BACTERIAL PENTAMERIC LIGAND- GATED ION CHANNEL	<i>GLOEOBACTER VIOLACEUS</i>	2.90
15	2VL0	X-RAY STRUCTURE OF A PENTAMERIC LIGAND GATED ION CHANNEL FROM ERWINIA CHRYSANTHEMI (ELIC)	<i>ERWINIA CHRYSANTHEMI</i>	3.3
16	1JB0	CRYSTAL STRUCTURE OF PHOTOSYSTEM I: A PHOTOSYNTHETIC REACTION CENTER AND CORE ANTENNA SYSTEM FROM CYANOBACTERIA	<i>SYNECHOCOCCUS ELONGATUS</i>	2.50
16	2WSC	IMPROVED MODEL OF PLANT PHOTOSYSTEM I	<i>PISUM SATIVUM</i>	3.30
17	1RZH	PHOTOSYNTHETIC REACTION CENTER DOUBLE MUTANT FROM RHODOBACTER SPHAEROIDES WITH ASP L213 REPLACED WITH ASN AND ARG M233 REPLACED WITH CYS IN THE CHARGE-NEUTRAL DQAQB STATE (TRIGONAL FORM)	<i>RHODOBACTER SPHAEROIDES</i>	1.80
17	1RZH	PHOTOSYNTHETIC REACTION CENTER DOUBLE MUTANT FROM RHODOBACTER SPHAEROIDES WITH ASP L213 REPLACED WITH ASN AND ARG M233 REPLACED WITH CYS IN THE CHARGE-NEUTRAL DQAQB STATE (TRIGONAL FORM)	<i>RHODOBACTER SPHAEROIDES</i>	1.80
17	2WJN	LIPIDIC SPONGE PHASE CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTRE FROM BLASTOCHLORIS VIRIDIS (HIGH DOSE)	<i>RHODOPSEUDOMONAS VIRIDIS</i>	1.86
17	2WJN	LIPIDIC SPONGE PHASE CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTRE FROM BLASTOCHLORIS VIRIDIS (HIGH DOSE)	<i>RHODOPSEUDOMONAS VIRIDIS</i>	1.86
17	1EYS	CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTER FROM A THERMOPHILIC BACTERIUM, THERMOCHROMATIUM TEPIDUM	<i>THERMOCHROMATIUM TEPIDUM</i>	2.20
17	1EYS	CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTER FROM A THERMOPHILIC BACTERIUM, THERMOCHROMATIUM TEPIDUM	<i>THERMOCHROMATIUM TEPIDUM</i>	2.20
17	3BZ1	CRYSTAL STRUCTURE OF CYANOBACTERIAL PHOTOSYSTEM II (PART 1 OF 2). THIS FILE CONTAINS FIRST MONOMER OF PSII DIMER	<i>THERMOSYNECHOCOCCUS ELONGATUS</i>	2.90
17	3BZ1	CRYSTAL STRUCTURE OF CYANOBACTERIAL	<i>THERMOSYNECHOCOCCUS</i>	2.90

		PHOTOSYSTEM II (PART 1 OF 2). THIS FILE CONTAINS FIRST MONOMER OF PSII DIMER	<i>ELONGATUS</i>	
18	1RH5	THE STRUCTURE OF A PROTEIN CONDUCTING CHANNEL	<i>METHANOCALDOCOCCLUS JANNASCHII</i>	3.20
18	2ZJS	CRYSTAL STRUCTURE OF SECYE TRANSLOCON FROM THERMUS THERMOPHILUS WITH A FAB FRAGMENT	<i>THERMUS THERMOPHILUS</i>	3.20
19	2H88	AVIAN MITOCHONDRIAL RESPIRATORY COMPLEX II AT 1.8 ANGSTROM RESOLUTION	<i>GALLUS GALLUS</i>	1.74
19	2H88	AVIAN MITOCHONDRIAL RESPIRATORY COMPLEX II AT 1.8 ANGSTROM RESOLUTION	<i>GALLUS GALLUS</i>	1.74
19	1ZOY	CRYSTAL STRUCTURE OF MITOCHONDRIAL RESPIRATORY COMPLEX II FROM PORCINE HEART AT 2.4 ANGSTROMS	<i>SUS SCROFA</i>	2.40
19	2WDQ	E. COLI SUCCINATE:QUINONE OXIDOREDUCTASE (SQR) WITH CARBOXIN BOUND	<i>ESCHERICHIA COLI</i>	2.40
19	1ZOY	CRYSTAL STRUCTURE OF MITOCHONDRIAL RESPIRATORY COMPLEX II FROM PORCINE HEART AT 2.4 ANGSTROMS	<i>SUS SCROFA</i>	2.40
19	2WDQ	E. COLI SUCCINATE:QUINONE OXIDOREDUCTASE (SQR) WITH CARBOXIN BOUND	<i>ESCHERICHIA COLI</i>	2.40
19	1KF6	E. COLI QUINOL-FUMARATE REDUCTASE WITH BOUND INHIBITOR HQNO	<i>ESCHERICHIA COLI</i>	2.70
20	2A65	CRYSTAL STRUCTURE OF LEUTAA, A BACTERIAL HOMOLOG OF NA ⁺ /CL ⁻ DEPENDENT NEUROTRANSMITTER TRANSPORTERS	<i>AQUIFEX AEOLICUS VF5</i>	1.65
20	3GIA	CRYSTAL STRUCTURE OF APCT TRANSPORTER	<i>METHANOCALDOCOCCLUS JANNASCHII</i>	2.32
20	2JLN	STRUCTURE OF MHP1, A NUCLEOBASE-CATION- SYMPORT-1 FAMILY TRANSPORTER	<i>MICROBACTERIUM LIQUEFACIENS</i>	2.85
20	3L1L	STRUCTURE OF ARG-BOUND ESCHERICHIA COLI ADIC	<i>ESCHERICHIA COLI</i>	3.00
20	2WIT	CRYSTAL STRUCTURE OF THE SODIUM-COUPLED GLYCINE BETAINE SYMPORTER BETP FROM CORYNEBACTERIUM GLUTAMICUM WITH BOUND SUBSTRATE	<i>CORYNEBACTERIUM GLUTAMICUM</i>	3.35
21	1WPG	CRYSTAL STRUCTURE OF THE SR CA ₂ ⁺ -ATPASE WITH MGF4	<i>ORYCTOLAGUS CUNICULUS</i>	2.30
21	2ZXE	CRYSTAL STRUCTURE OF THE SODIUM - POTASSIUM PUMP IN THE E2.2K+.PI STATE	<i>SQUALUS ACANTHIAS</i>	2.40

The numbers in the first column refer to the following families: (0) ABC transporters - efflux, (1) ABC transporters - influx, (2) Ammonium transporters, (3) Aquaporins, (4) Bacterial rhodopsins, (5) CIC transporters, (6) Cytochrome b₆f, (7) Cytochrome bc₁, (8) Cytochrome c oxidases - 2TM-helix subunit, (9) Cytochrome c oxidases - 7TM-helix subunit, (10) Cytochrome c oxidases - 12TM-helix subunit, (11) GlpG, (12) GPCR, (13) MFS transporters, (14) Multi-drug exporters, (15) Pentameric ion channels, (16) Photosystem I, (17) Photosystem II, (18) Protein conducting channels, (19) Reductases, (20) LeuT-fold transporters, (21) Sodium/potassium pumps.

Table A.2. Sequence identities between pairs of proteins in the same HOMEPEP family, based on their SKA structural alignments

Family	PDB1	Chain ID from PDB1	PDB2	Chain ID from PDB2	PSD value	% sequence identity
0	3D31	C	2R6G	G	0.47	19.5
0	2ONK	C	2R6G	G	0.50	20.5
0	3D31	C	2ONK	C	0.07	50.8
1	2NQ2	A	2QI9	B	0.34	31.9
2	1U7G	A	3B9W	A	0.26	18.3
2	3B9W	A	2B2H	A	0.28	18.6
2	1U7G	A	2B2H	A	0.13	34.1
3	2O9G	A	3KLY	C	0.57	7.3
3	1J4N	A	3KLY	C	0.58	8.3
3	3CN5	A	3KLY	C	0.50	9.3
3	1LDF	A	3KLY	C	0.52	9.7
3	2B6O	A	3KLY	C	0.57	9.7
3	2F2B	A	3KLY	C	0.38	9.9
3	3LLQ	A	3KLY	C	0.47	9.9
3	3C02	A	3KLY	C	0.47	10.4
3	3GD8	A	3KLY	C	0.52	10.4
3	2W2E	A	3KLY	C	0.44	11.1
3	3D9S	D	3KLY	C	0.47	13.1
3	2W2E	A	3C02	A	0.14	17.8
3	3CN5	A	3C02	A	0.22	20.5
3	3GD8	A	3C02	A	0.13	21.9
3	2W2E	A	1LDF	A	0.15	22.4
3	1J4N	A	3C02	A	0.17	23.2
3	2W2E	A	2F2B	A	0.11	23.2
3	3C02	A	2O9G	A	0.12	23.6
3	2W2E	A	2O9G	A	0.13	23.8
3	2B6O	A	2W2E	A	0.11	24.1
3	3C02	A	3D9S	D	0.15	24.2
3	2B6O	A	3C02	A	0.16	24.3
3	3LLQ	A	3C02	A	0.10	24.3
3	3GD8	A	1LDF	A	0.13	24.7
3	2W2E	A	3LLQ	A	0.12	24.8
3	2W2E	A	3D9S	D	0.19	25.2
3	3CN5	A	1LDF	A	0.16	25.3
3	1J4N	A	1LDF	A	0.14	25.4
3	2B6O	A	1LDF	A	0.17	25.4
3	3CN5	A	2O9G	A	0.13	26.4
3	1J4N	A	2F2B	A	0.17	26.7
3	3D9S	D	1LDF	A	0.13	26.9
3	2B6O	A	2F2B	A	0.09	27.1
3	2W2E	A	3GD8	A	0.13	27.2
3	3LLQ	A	1LDF	A	0.15	27.8
3	2W2E	A	1J4N	A	0.18	28.1
3	3CN5	A	2F2B	A	0.17	28.2
3	2O9G	A	1LDF	A	0.15	28.3
3	2W2E	A	3CN5	A	0.18	28.4
3	3GD8	A	2O9G	A	0.15	28.5
3	3C02	A	2F2B	A	0.15	28.7
3	3LLQ	A	3CN5	A	0.08	28.8

3	3GD8	A	3LLQ	A	0.12	29.7
3	2B6O	A	3CN5	A	0.09	30.5
3	2B6O	A	3LLQ	A	0.12	31.1
3	1J4N	A	2O9G	A	0.16	31.5
3	3LLQ	A	1J4N	A	0.12	31.9
3	1LDF	A	2F2B	A	0.14	32.1
3	2O9G	A	2F2B	A	0.14	32.6
3	3LLQ	A	2F2B	A	0.14	32.6
3	3C02	A	1LDF	A	0.11	32.7
3	3GD8	A	2F2B	A	0.10	32.9
3	3D9S	D	2F2B	A	0.12	33.1
3	3CN5	A	3D9S	D	0.11	33.6
3	2B6O	A	2O9G	A	0.15	33.8
3	3LLQ	A	3D9S	D	0.13	33.8
3	2O9G	A	3D9S	D	0.14	33.9
3	3GD8	A	3CN5	A	0.08	37.0
3	1J4N	A	3CN5	A	0.09	38.7
3	1J4N	A	3D9S	D	0.13	45.3
3	2B6O	A	3GD8	A	0.04	45.4
3	3GD8	A	1J4N	A	0.06	45.8
3	2B6O	A	1J4N	A	0.07	47.1
3	3GD8	A	3D9S	D	0.05	51.7
3	2B6O	A	3D9S	D	0.09	56.2
3	3LLQ	A	2O9G	A	0.02	72.5
4	1XIO	A	3A7K	A	0.14	20.2
4	1H2S	A	3A7K	A	0.10	21.5
4	2JAF	A	1XIO	A	0.13	24.4
4	1H2S	A	2JAF	A	0.13	25.3
4	1XIO	A	1M0L	A	0.15	27.3
4	1H2S	A	1M0L	A	0.09	27.7
4	1H2S	A	1XIO	A	0.11	28.5
4	1M0L	A	3A7K	A	0.07	29.1
4	2E14	A	1H2S	A	0.09	29.8
4	2E14	A	3A7K	A	0.11	30.2
4	2E14	A	1XIO	A	0.17	31.2
4	2E14	A	2JAF	A	0.10	31.2
4	2JAF	A	1M0L	A	0.09	31.5
4	2E14	A	1M0L	A	0.02	55.2
4	2JAF	A	3A7K	A	0.04	57.1
5	1KPL	B	1OTS	A	0.04	76.6
6	1Q90	D	2ZT9	B	0.03	76.6
7	2QJY	A	3CX5	N	0.09	45.0
7	2QJY	A	3L70	C	0.09	45.3
7	2A06	P	2QJY	A	0.06	46.4
7	2A06	P	3CX5	N	0.03	50.3
7	3CX5	N	3L70	C	0.03	51.2
7	2A06	P	3L70	C	0.02	73.6
8	1V54	B	3HB3	B	0.07	32.6
8	1V54	B	2GSM	B	0.07	34.0
8	3HB3	B	2GSM	B	0.03	57.1
9	1V54	C	1M56	C	0.04	48.1
9	1V54	C	1QLE	C	0.05	49.6
9	1QLE	C	1M56	C	0.06	70.6
10	2DYR	A	2GSM	A	0.04	51.5
10	2DYR	A	3HB3	A	0.04	52.4
10	2GSM	A	3HB3	A	0.02	82.3
11	3B44	A	2NR9	A	0.09	36.0
12	2Z73	A	3KJ6	A	0.59	6.3
12	2Z73	A	2RH1	A	0.48	18.0

12	2Z73	A	2VT4	A	0.35	20.5
12	3KJ6	A	2RH1	A	0.52	24.6
12	2VT4	A	3KJ6	A	0.24	54.4
12	2VT4	A	2RH1	A	0.11	61.5
13	1PW4	A	2CFQ	A	0.72	8.7
14	2J8S	A	2V50	B	0.36	68.5
15	2VLO	D	3EAM	B	0.29	21.7
16	2WSC	B	1JB0	B	0.39	74.3
17	3BZ1	D	1EYS	M	0.66	12.5
17	2WJN	M	3BZ1	A	0.37	13.1
17	1EYS	L	3BZ1	A	0.42	15.0
17	1EYS	M	3BZ1	A	0.52	15.0
17	3BZ1	D	1RZH	M	0.59	15.0
17	3BZ1	D	2WJN	M	0.43	15.1
17	1RZH	M	3BZ1	A	0.42	15.7
17	3BZ1	D	1EYS	L	0.39	15.8
17	2WJN	L	3BZ1	A	0.36	16.1
17	3BZ1	D	1RZH	L	0.41	16.1
17	3BZ1	D	2WJN	L	0.33	16.3
17	1RZH	L	3BZ1	A	0.44	16.6
17	2WJN	L	2WJN	M	0.19	23.2
17	1EYS	L	2WJN	M	0.21	23.5
17	1RZH	L	2WJN	M	0.19	25.5
17	1EYS	L	1EYS	M	0.24	25.8
17	1RZH	L	1EYS	M	0.23	26.7
17	2WJN	L	1EYS	M	0.18	26.7
17	1RZH	M	1RZH	L	0.22	27.5
17	1EYS	L	1RZH	M	0.24	28.6
17	2WJN	L	1RZH	M	0.22	29.3
17	3BZ1	D	3BZ1	A	0.11	29.6
17	1RZH	M	2WJN	M	0.06	46.8
17	2WJN	L	1RZH	L	0.02	58.1
17	2WJN	M	1EYS	M	0.04	59.0
17	1RZH	M	1EYS	M	0.05	59.5
17	2WJN	L	1EYS	L	0.03	65.1
17	1EYS	L	1RZH	L	0.02	65.6
18	1RH5	A	2ZJS	Y	0.58	15.8
19	2H88	C	2H88	D	0.65	6.7
19	1KF6	D	2H88	C	0.77	7.0
19	1KF6	D	1ZOY	D	0.52	7.2
19	2WDQ	K	2H88	D	0.11	7.2
19	2WDQ	L	2H88	C	0.87	7.4
19	1KF6	D	1ZOY	C	0.94	7.8
19	1ZOY	C	1ZOY	D	0.58	8.3
19	1ZOY	C	2H88	D	0.54	8.4
19	1KF6	D	2H88	D	0.56	9.1
19	2WDQ	K	1ZOY	D	0.09	9.9
19	1KF6	D	2WDQ	K	0.46	10.0
19	1ZOY	D	2H88	C	0.47	10.0
19	2WDQ	L	1ZOY	C	0.89	10.7
19	2WDQ	L	1ZOY	D	0.19	11.6
19	2WDQ	L	2H88	D	0.21	11.7
19	1KF6	D	2WDQ	L	0.38	14.3
19	2WDQ	K	2H88	C	0.43	14.6
19	2WDQ	K	2WDQ	L	0.60	15.9
19	2WDQ	K	1ZOY	C	0.33	18.5
19	1ZOY	C	2H88	C	0.03	70.0
19	1ZOY	D	2H88	D	0.01	76.8
20	2WIT	A	2JLN	A	1.06	5.5

20	2A65	A	2JLN	A	1.03	6.6
20	3GIA	A	2JLN	A	0.91	7.1
20	2WIT	A	3GIA	A	0.88	7.8
20	2WIT	A	2A65	A	0.96	8.0
20	2A65	A	3GIA	A	1.07	8.1
20	2JLN	A	3L1L	A	0.84	8.1
20	2WIT	A	3L1L	A	0.90	8.7
20	2A65	A	3L1L	A	1.19	9.4
20	3GIA	A	3L1L	A	0.49	16.8
21	2ZXE	A	1WPG	A	0.65	27.9

SKA structural alignments were generated between pairs of protein chains, and the PSD value and the percentage sequence identity for the corresponding sequence alignment were computed. See legend to Table A1 for more details.

Table A.3. α -helical proteins in the HOME3 data set, listed by family

#	PDB ID	Name	Source	Res (Å)
0	2YEV	STRUCTURE OF CAA3-TYPE CYTOCHROME OXIDASE	THERMUS THERMOPHILUS	2.36
0	2GSM	CATALYTIC CORE (SUBUNITS I AND II) OF CYTOCHROME C OXIDASE FROM RHODOBACTER SPHAEROIDES	RHODOBACTER SPHAEROIDES	2.00
0	3AG3	BOVINE HEART CYTOCHROME C OXIDASE IN THE NITRIC OXIDE-BOUND FULLY REDUCED STATE AT 100 K	BOS TAURUS	1.80
0	3HB3	HIGH RESOLUTION CRYSTAL STRUCTURE OF PARACOCCLUS DENITRIFICANS CYTOCHROME C OXIDASE	PARACOCCLUS DENITRIFICANS	2.25
1	2XQU	MICROSCOPIC ROTARY MECHANISM OF ION TRANSLOCATION IN THE FO COMPLEX OF ATP SYNTHASES	ARTHROSPIRA PLATENSIS	1.84
1	4F4S	STRUCTURE OF THE YEAST F1FO ATPASE C10 RING WITH BOUND OLIGOMYCIN	SACCHAROMYCES CEREVISIAE	1.90
1	2X2V	STRUCTURAL BASIS OF A NOVEL PROTON-COORDINATION TYPE IN AN F1FO-ATP SYNTHASE ROTOR RING	BACILLUS PSEUDOFIRMUS OF4	2.50
1	2WGM	COMPLETE ION-COORDINATION STRUCTURE IN THE ROTOR RING OF NA-DEPENDENT F-ATP SYNTHASE	ILYOBACTER TARTARICUS	2.35
2	2QKS	CRYSTAL STRUCTURE OF A KIR3.1-PROKARYOTIC KIR CHANNEL CHIMERA	BURKHOLDERIA XENOVORANS	2.20
2	3SPC	INWARD RECTIFIER POTASSIUM CHANNEL KIR2.2 IN COMPLEX WITH DIOCTANOYLGLYCEROL PYROPHOSPHATE (DGPP)	GALLUS GALLUS	2.45
2	3SYA	CRYSTAL STRUCTURE OF THE G PROTEIN-GATED INWARD RECTIFIER K+ CHANNEL GIRK2 (KIR3.2) IN COMPLEX WITH SODIUM AND PIP2	MUS MUSCULUS	2.98
3	4H33	CRYSTAL STRUCTURE OF A VOLTAGE-GATED K+ CHANNEL PORE MODULE IN A CLOSED STATE IN LIPID MEMBRANES, TETRAGONAL CRYSTAL FORM	LISTERIA MONOCYTOGENES	3.10
3	3LDC	HIGH RESOLUTION OPEN MTHK PORE STRUCTURE CRYSTALLIZED IN 100 MM K+	METHANOTHERMOBACTER THERMAUTOTROPHICUS	1.45
3	2IH3	ION SELECTIVITY IN A SEMI-SYNTHETIC K+ CHANNEL LOCKED IN THE CONDUCTIVE CONFORMATION	MUS MUSCULUS	1.72
3	3OUF	STRUCTURE OF A K+ SELECTIVE NAK MUTANT	BACILLUS CEREUS	1.55
4	2H88	AVIAN MITOCHONDRIAL RESPIRATORY COMPLEX II AT 1.8 ANGSTROM RESOLUTION	GALLUS GALLUS	1.74
4	1ZOY	CRYSTAL STRUCTURE OF MITOCHONDRIAL RESPIRATORY COMPLEX II FROM PORCINE HEART AT 2.4 ANGSTROMS	SUS SCROFA	2.40
5	3VR8	MITOCHONDRIAL RHODOQUINOL-FUMARATE REDUCTASE FROM THE PARASITIC NEMATODE ASCARIS SUUM	ASCARIS SUUM	2.81
5	1ZOY	CRYSTAL STRUCTURE OF MITOCHONDRIAL RESPIRATORY COMPLEX II FROM PORCINE HEART AT 2.4 ANGSTROMS	SUS SCROFA	2.40
5	2WDQ	E. COLI SUCCINATE:QUINONE OXIDOREDUCTASE (SQR) WITH CARBOXIN BOUND	ESCHERICHIA COLI	2.40
5	2H88	AVIAN MITOCHONDRIAL RESPIRATORY COMPLEX II AT 1.8 ANGSTROM RESOLUTION	GALLUS GALLUS	1.74
6	2ZT9	CRYSTAL STRUCTURE OF THE CYTOCHROME B6F COMPLEX FROM NOSTOC SP. PCC 7120	NOSTOC SP. PCC 7120	3.00
6	1Q90	STRUCTURE OF THE CYTOCHROME B6F (PLASTOQUINONE : PLASTOCYANIN OXIDOREDUCTASE) FROM CHLAMYDOMONAS REINHARDTII	CHLAMYDOMONAS REINHARDTII	3.10
7	2BHW	PEA LIGHT-HARVESTING COMPLEX II AT 2.5 ANGSTROM RESOLUTION	PISUM SATIVUM	2.50
7	3PL9	CRYSTAL STRUCTURE OF SPINACH MINOR LIGHT-HARVESTING COMPLEX CP29 AT 2.80 ANGSTROM RESOLUTION	SPINACIA OLERACEA	2.80
8	2VV5	THE OPEN STRUCTURE OF MSCS	ESCHERICHIA COLI	3.45
8	3UDC	CRYSTAL STRUCTURE OF A MEMBRANE PROTEIN	THERMOANAEROBACTER TENGCONGENSIS, ESCHERICHIA	3.35
9	1JB0	CRYSTAL STRUCTURE OF PHOTOSYSTEM I: A PHOTOSYNTHETIC REACTION CENTER AND CORE ANTENNA SYSTEM FROM CYANOBACTERIA	SYNECHOCOCCUS ELONGATUS	2.50
9	2WSC	IMPROVED MODEL OF PLANT PHOTOSYSTEM I	PISUM SATIVUM	3.30
10	3TLW	THE GLIC PENTAMERIC LIGAND-GATED ION CHANNEL LOOP2-21' OXIDIZED MUTANT IN A LOCALLY-CLOSED CONFORMATION (LC2 SUBTYPE)	GLOEOBACTER VIOLACEUS	2.60
10	3RQW	CRYSTAL STRUCTURE OF ACETYLCHOLINE BOUND TO A	DICKEYA DADANTII	2.91

		PROKARYOTIC PENTAMERIC LIGAND-GATED ION CHANNEL, ELIC		
10	3RHW	C. ELEGANS GLUTAMATE-GATED CHLORIDE CHANNEL (GLUCL) IN COMPLEX WITH FAB AND IVERMECTIN	CAENORHABDITIS ELEGANS	3.26
11	3UM7	CRYSTAL STRUCTURE OF THE HUMAN TWO PORE DOMAIN K ⁺ ION CHANNEL TRAAK (K2P4.1)	HOMO SAPIENS	3.31
11	3UKM	CRYSTAL STRUCTURE OF THE HUMAN TWO PORE DOMAIN POTASSIUM ION CHANNEL K2P1 (TWIK-1)	HOMO SAPIENS	3.40
12	3PCV	CRYSTAL STRUCTURE ANALYSIS OF HUMAN LEUKOTRIENE C4 SYNTHASE AT 1.9 ANGSTROM RESOLUTION	HOMO SAPIENS	1.90
12	2H8A	STRUCTURE OF MICROSOMAL GLUTATHIONE TRANSFERASE 1 IN COMPLEX WITH GLUTATHIONE	RATTUS NORVEGICUS	3.20
13	4GD3	STRUCTURE OF E. COLI HYDROGENASE-1 IN COMPLEX WITH CYTOCHROME B	ESCHERICHIA COLI	3.30
13	2ZT9	CRYSTAL STRUCTURE OF THE CYTOCHROME B6F COMPLEX FROM NOSTOC SP. PCC 7120	NOSTOC SP. PCC 7120	3.00
14	3ARC	CRYSTAL STRUCTURE OF OXYGEN-EVOLVING PHOTOSYSTEM II AT 1.9 ANGSTROM RESOLUTION	THERMOSYNECHOCOCCUS VULCANUS	1.90
14	2WJN	LIPIDIC SPONGE PHASE CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTRE FROM BLASTOCHLORIS VIRIDIS (HIGH DOSE)	RHODOPSEUDOMONAS VIRIDIS	1.86
14	1RZH	PHOTOSYNTHETIC REACTION CENTER DOUBLE MUTANT FROM RHODOBACTER SPHAEROIDES WITH ASP L213 REPLACED WITH ASN AND ARG M233 REPLACED WITH CYS IN THE CHARGE-NEUTRAL DQAQB STATE (TRIGONAL FORM)	RHODOBACTER SPHAEROIDES	1.80
14	1EYS	CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTER FROM A THERMOPHILIC BACTERIUM, THERMOCHROMATIUM TEPIDUM	THERMOCHROMATIUM TEPIDUM	2.20
14	1RZH	PHOTOSYNTHETIC REACTION CENTER DOUBLE MUTANT FROM RHODOBACTER SPHAEROIDES WITH ASP L213 REPLACED WITH ASN AND ARG M233 REPLACED WITH CYS IN THE CHARGE-NEUTRAL DQAQB STATE (TRIGONAL FORM)	RHODOBACTER SPHAEROIDES	1.80
14	1EYS	CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTER FROM A THERMOPHILIC BACTERIUM, THERMOCHROMATIUM TEPIDUM	THERMOCHROMATIUM TEPIDUM	2.20
14	2WJN	LIPIDIC SPONGE PHASE CRYSTAL STRUCTURE OF PHOTOSYNTHETIC REACTION CENTRE FROM BLASTOCHLORIS VIRIDIS (HIGH DOSE)	RHODOPSEUDOMONAS VIRIDIS	1.86
14	3ARC	CRYSTAL STRUCTURE OF OXYGEN-EVOLVING PHOTOSYSTEM II AT 1.9 ANGSTROM RESOLUTION	THERMOSYNECHOCOCCUS VULCANUS	1.90
15	3LLQ	AQUAPORIN STRUCTURE FROM PLANT PATHOGEN AGROBACTERIUM TUMERFACIENS	AGROBACTERIUM TUMERFACIENS STR.	2.01
15	3CN5	CRYSTAL STRUCTURE OF THE SPINACH AQUAPORIN SOPIP2 1 S115E, S274E MUTANT	SPINACIA OLERACEA	2.05
15	2F2B	CRYSTAL STRUCTURE OF INTEGRAL MEMBRANE PROTEIN AQUAPORIN AQPM AT 1.68A RESOLUTION	METHANOTHERMOBACTER MARBURGENSIS STR. MARBURG	1.68
15	2O9G	CRYSTAL STRUCTURE OF AQPZ MUTANT L170C COMPLEXED WITH MERCURY.	ESCHERICHIA COLI	1.90
15	2B6O	ELECTRON CRYSTALLOGRAPHIC STRUCTURE OF LENS AQUAPORIN-0 (AQP0) (LENS MIP) AT 1.9A RESOLUTION, IN A CLOSED PORE STATE	OVIS ARIES	1.90
15	2W2E	1.15 ANGSTROM CRYSTAL STRUCTURE OF P.PASTORIS AQUAPORIN, AQY1, IN A CLOSED CONFORMATION AT PH 3.5	KOMAGATAELLA PASTORIS	1.15
15	1LDF	CRYSTAL STRUCTURE OF THE E. COLI GLYCEROL FACILITATOR (GLPF) MUTATION W48F, F200T	ESCHERICHIA COLI	2.10
15	3NE2	ARCHAEOGLOBUS FULGIDUS AQUAPORIN	ARCHAEOGLOBUS FULGIDUS	3.00
15	3GD8	CRYSTAL STRUCTURE OF HUMAN AQUAPORIN 4 AT 1.8 AND ITS MECHANISM OF CONDUCTANCE	HOMO SAPIENS	1.80
15	1J4N	CRYSTAL STRUCTURE OF THE AQP1 WATER CHANNEL	BOS TAURUS	2.20
15	3D9S	HUMAN AQUAPORIN 5 (AQP5) - HIGH RESOLUTION X-RAY STRUCTURE	HOMO SAPIENS	2.00
15	3C02	X-RAY STRUCTURE OF THE AQUAGLYCEROPORIN FROM PLASMODIUM FALCIPARUM	PLASMODIUM FALCIPARUM	2.05
16	3ARC	CRYSTAL STRUCTURE OF OXYGEN-EVOLVING PHOTOSYSTEM II AT 1.9 ANGSTROM RESOLUTION	THERMOSYNECHOCOCCUS VULCANUS	1.90
16	3ARC	CRYSTAL STRUCTURE OF OXYGEN-EVOLVING PHOTOSYSTEM II AT 1.9 ANGSTROM RESOLUTION	THERMOSYNECHOCOCCUS VULCANUS	1.90
17	3QF4	CRYSTAL STRUCTURE OF A HETERODIMERIC ABC TRANSPORTER IN ITS INWARD- FACING CONFORMATION	THERMOTOGA MARITIMA	2.90
17	4AYT	STRUCTURE OF THE HUMAN MITOCHONDRIAL ABC TRANSPORTER, ABCB10	HOMO SAPIENS	2.85
17	3QF4	CRYSTAL STRUCTURE OF A HETERODIMERIC ABC TRANSPORTER IN ITS INWARD- FACING CONFORMATION	THERMOTOGA MARITIMA	2.90

17	4A82	FITTED MODEL OF STAPHYLOCOCCUS AUREUS SAV1866 MODEL ABC TRANSPORTER IN THE HUMAN CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR VOLUME MAP EMD-1966.	HOMO SAPIENS	2.00
18	3D31	MODBC FROM METHANOSARCINA ACETIVORANS	METHANOSARCINA ACETIVORANS	3.00
18	2ONK	ABC TRANSPORTER MODBC IN COMPLEX WITH ITS BINDING PROTEIN MODA	ARCHAEOGLOBUS FULGIDUS	3.10
18	3RLF	CRYSTAL STRUCTURE OF THE MALTOSE-BINDING PROTEIN/MALTOSE TRANSPORTER COMPLEX IN AN OUTWARD-FACING CONFORMATION BOUND TO MGAMPPNP	ESCHERICHIA COLI	2.20
19	3RLB	CRYSTAL STRUCTURE AT 2.0 Å OF THE S-COMPONENT FOR THIAMIN FROM AN ECF- TYPE ABC TRANSPORTER	LACTOCOCCUS LACTIS SUBSP. CREMORIS	2.00
19	4DVE	CRYSTAL STRUCTURE AT 2.1 Å OF THE S-COMPONENT FOR BIOTIN FROM AN ECF- TYPE ABC TRANSPORTER	LACTOCOCCUS LACTIS SUBSP. CREMORIS	2.09
19	3RGB	CRYSTAL STRUCTURE OF PARTICULATE METHANE MONOOXYGENASE FROM METHYLOCOCCUS CAPSULATUS (BATH)	METHYLOCOCCUS CAPSULATUS	2.80
20	4DXW	CRYSTAL STRUCTURE OF NAVRH, A VOLTAGE-GATED SODIUM CHANNEL	ALPHA PROTEOBACTERIUM HIMB114	3.05
20	3RVY	CRYSTAL STRUCTURE OF THE NAVAB VOLTAGE-GATED SODIUM CHANNEL (ILE217CYS, 2.7 Å)	ARCOBACTER BUTZLERI	2.70
20	3BEH	STRUCTURE OF A BACTERIAL CYCLIC NUCLEOTIDE REGULATED ION CHANNEL	MESORHIZOBIUM LOTI	3.10
20	2R9R	SHAKER FAMILY VOLTAGE DEPENDENT POTASSIUM CHANNEL (KV1.2-KV2.1 PADDLE CHIMERA CHANNEL) IN ASSOCIATION WITH BETA SUBUNIT	RATTUS NORVEGICUS	2.40
21	3KCU	STRUCTURE OF FORMATE CHANNEL	ESCHERICHIA COLI O157:H7	2.24
21	4FC4	FNT FAMILY ION CHANNEL	SALMONELLA ENTERICA SUBSP. ENTERICA SEROVAR	2.40
21	3TDS	CRYSTAL STRUCTURE OF HSC F1941	CLOSTRIDIUM DIFFICILE	1.98
21	3KLY	PENTAMERIC FORMATE CHANNEL	VIBRIO CHOLERAEE	2.10
22	2NR9	CRYSTAL STRUCTURE OF GLPG, RHOMBOID PEPTIDASE FROM HAEMOPHILUS INFLUENZAE	HAEMOPHILUS INFLUENZAE	2.20
22	2XOV	CRYSTAL STRUCTURE OF E.COLI RHOMBOID PROTEASE GLPG, NATIVE ENZYME	ESCHERICHIA COLI	1.65
23	3PBL	STRUCTURE OF THE HUMAN DOPAMINE D3 RECEPTOR IN COMPLEX WITH ETICLOPRIDE	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	2.89
23	4EJ4	STRUCTURE OF THE DELTA OPIOID RECEPTOR BOUND TO NALTRINDOLE	MUS MUSCULUS, ENTEROBACTERIA PHAGE T4	3.40
23	3RZE	STRUCTURE OF THE HUMAN HISTAMINE H1 RECEPTOR IN COMPLEX WITH DOXEPIN	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	3.10
23	4EA3	STRUCTURE OF THE N/O/FQ OPIOID RECEPTOR IN COMPLEX WITH A PEPTIDE MIMETIC	HOMO SAPIENS, ESCHERICHIA COLI	3.01
23	4DAJ	STRUCTURE OF THE M3 MUSCARINIC ACETYLCHOLINE RECEPTOR	RATTUS NORVEGICUS, ENTEROBACTERIA PHAGE T4	3.40
23	3UON	STRUCTURE OF THE HUMAN M2 MUSCARINIC ACETYLCHOLINE RECEPTOR BOUND TO AN ANTAGONIST	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	3.00
23	4DKL	CRYSTAL STRUCTURE OF THE MU-OPIOID RECEPTOR BOUND TO A MORPHINAN ANTAGONIST	MUS MUSCULUS, ENTEROBACTERIA PHAGE T4	2.80
23	1U19	CRYSTAL STRUCTURE OF BOVINE RHODOPSIN AT 2.2 Å RESOLUTION	BOS TAURUS	2.20
23	2RH1	HIGH RESOLUTION CRYSTAL STRUCTURE OF HUMAN B2-ADRENERGIC G PROTEIN- COUPLED RECEPTOR.	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	2.40
23	4DJH	STRUCTURE OF THE HUMAN KAPPA OPIOID RECEPTOR IN COMPLEX WITH JDIC	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	2.90
23	2Z73	CRYSTAL STRUCTURE OF SQUID RHODOPSIN	TODARODES PACIFICUS	2.50
23	4AMJ	TURKEY BETA1 ADRENERGIC RECEPTOR WITH STABILISING MUTATIONS AND BOUND BIASED AGONIST CARVEDILOL	MELEAGRIS GALLOPAVO	2.30
23	4EYI	CRYSTAL STRUCTURE OF THE CHIMERIC PROTEIN OF A2AAR-BRIL IN COMPLEX WITH ZM241385 AT 1.8Å RESOLUTION	HOMO SAPIENS, ESCHERICHIA COLI	1.80
23	3V2Y	CRYSTAL STRUCTURE OF A LIPID G PROTEIN-COUPLED RECEPTOR AT 2.80Å	HOMO SAPIENS, ENTEROBACTERIA PHAGE T4	2.80
24	1M56	STRUCTURE OF CYTOCHROME C OXIDASE FROM RHODOBACTER SPHAEROIDES (WILD TYPE)	RHODOBACTER SPHAEROIDES	2.30
24	3AG3	BOVINE HEART CYTOCHROME C OXIDASE IN THE NITRIC OXIDE-BOUND FULLY REDUCED STATE AT 100 K	BOS TAURUS	1.80
24	1QLE	CRYO-STRUCTURE OF THE PARACOCCUS DENITRIFICANS FOUR-SUBUNIT CYTOCHROME C OXIDASE IN THE COMPLETELY OXIDIZED STATE COMPLEXED WITH AN ANTIBODY FV FRAGMENT	PARACOCCUS DENITRIFICANS	3.00
25	3AM6	CRYSTAL STRUCTURE OF THE PROTON PUMPING RHODOPSIN AR2 FROM MARINE ALGA ACETABULARIA	ACETABULARIA ACETABULUM	3.20

ACETABULUM				
25	1XIO	ANABAENA SENSORY RHODOPSIN	NOSTOC SP. PCC 7120	2.00
25	1M0L	BACTERIORHODOPSIN/LIPID COMPLEX AT 1.47 Å RESOLUTION	HALOBACTERIUM SALINARUM	1.47
25	3QAP	CRYSTAL STRUCTURE OF NATRONOMONAS PHARAONIS SENSORY RHODOPSIN II IN THE GROUND STATE	NATRONOMONAS PHARAONIS	1.90
25	2JAF	GROUND STATE OF HALORHODOPSIN T203V	HALOBACTERIUM SALINARIUM	1.70
25	2EI4	TRIMERIC COMPLEX OF ARCHAERHODOPSIN-2	HALOBACTERIUM SP. AUS-2	2.10
25	3QBG	ANION-FREE BLUE FORM OF PHARAONIS HALORHODOPSIN	NATRONOMONAS PHARAONIS	1.80
25	3UG9	CRYSTAL STRUCTURE OF THE CLOSED STATE OF CHANNELRHODOPSIN	CHLAMYDOMONAS REINHARDTII	2.30
25	3DDL	CRYSTALLOGRAPHIC STRUCTURE OF XANTHORHODOPSIN, A LIGHT-DRIVEN ION PUMP WITH DUAL CHROMOPHORE	SALINIBACTER RUBER	1.90
26	3CX5	STRUCTURE OF COMPLEX III WITH BOUND CYTOCHROME C IN REDUCED STATE AND DEFINITION OF A MINIMAL CORE INTERFACE FOR ELECTRON TRANSFER.	SACCHAROMYCES CEREVISIAE	1.90
26	2A06	BOVINE CYTOCHROME BC1 COMPLEX WITH STIGMATELLIN BOUND	BOS TAURUS	2.10
26	2QJY	CRYSTAL STRUCTURE OF RHODOBACTER SPHAEROIDES DOUBLE MUTANT WITH STIGMATELLIN AND UQ2	RHODOBACTER SPHAEROIDES	2.40
26	3L70	CYTOCHROME BC1 COMPLEX FROM CHICKEN WITH TRIFLOXYSTROBIN BOUND	GALLUS GALLUS	2.75
27	3ND0	X-RAY CRYSTAL STRUCTURE OF A SLOW CYANOBACTERIAL CL-/H+ ANTIporter	SYNECHOCYSTIS	3.20
27	1OTS	STRUCTURE OF THE ESCHERICHIA COLI CLC CHLORIDE CHANNEL AND FAB COMPLEX	ESCHERICHIA COLI	2.51
27	1KPL	CRYSTAL STRUCTURE OF THE CLC CHLORIDE CHANNEL FROM S. TYPHIMURIUM	SALMONELLA TYPHIMURIUM	3.00
28	3K3F	CRYSTAL STRUCTURE OF THE UREA TRANSPORTER FROM DESULFOVIBRIO VULGARIS	DESULFOVIBRIO VULGARIS	2.30
28	4EZC	CRYSTAL STRUCTURE OF THE UT-B UREA TRANSPORTER FROM BOS TAURUS	BOS TAURUS	2.36
29	2NQ2	AN INWARD-FACING CONFORMATION OF A PUTATIVE METAL-CHELATE TYPE ABC TRANSPORTER.	HAEMOPHILUS INFLUENZAE	2.40
29	1L7V	BACTERIAL ABC TRANSPORTER INVOLVED IN B12 UPTAKE	ESCHERICHIA COLI	3.20
30	2ZJS	CRYSTAL STRUCTURE OF SECYE TRANSLOCON FROM THERMUS THERMOPHILUS WITH A FAB FRAGMENT	THERMUS THERMOPHILUS	3.20
30	1RH5	THE STRUCTURE OF A PROTEIN CONDUCTING CHANNEL	METHANOCALDOCOCCLUS JANNASCHII	3.20
30	3MP7	LATERAL OPENING OF A TRANSLOCON UPON ENTRY OF PROTEIN SUGGESTS THE MECHANISM OF INSERTION INTO MEMBRANES	PYROCOCCUS FURIOSUS	2.90
31	3AR7	CALCIUM PUMP CRYSTAL STRUCTURE WITH BOUND TNP-ATP AND TG IN THE ABSENCE OF CA2+	ORYCTOLAGUS CUNICULUS	2.15
31	2ZXE	CRYSTAL STRUCTURE OF THE SODIUM - POTASSIUM PUMP IN THE E2.2K+.PI STATE	SQUALUS ACANTHIAS	2.40
32	1U7G	CRYSTAL STRUCTURE OF AMMONIA CHANNEL AMTB FROM E. COLI	ESCHERICHIA COLI	1.40
32	2B2H	AMMONIUM TRANSPORTER AMT-1 FROM A. FULGIDUS (AS)	ARCHAEOGLOBUS FULGIDUS	1.54
32	3B9W	THE 1.3 Å RESOLUTION STRUCTURE OF NITROSOMONAS EUROPAEA RH50 AND MECHANISTIC IMPLICATIONS FOR NH3 TRANSPORT BY RHESUS FAMILY PROTEINS	NITROSOMONAS EUROPAEA	1.30
33	2WSC	IMPROVED MODEL OF PLANT PHOTOSYSTEM I	PISUM SATIVUM	3.30
33	1JB0	CRYSTAL STRUCTURE OF PHOTOSYSTEM I: A PHOTOSYNTHETIC REACTION CENTER AND CORE ANTENNA SYSTEM FROM CYANOBACTERIA	SYNECHOCOCCUS ELONGATUS	2.50
33	1JB0	CRYSTAL STRUCTURE OF PHOTOSYSTEM I: A PHOTOSYNTHETIC REACTION CENTER AND CORE ANTENNA SYSTEM FROM CYANOBACTERIA	SYNECHOCOCCUS ELONGATUS	2.50
33	2WSC	IMPROVED MODEL OF PLANT PHOTOSYSTEM I	PISUM SATIVUM	3.30
34	1PW4	CRYSTAL STRUCTURE OF THE GLYCEROL-3-PHOSPHATE TRANSPORTER FROM E. COLI	ESCHERICHIA COLI	3.30
34	3O7Q	CRYSTAL STRUCTURE OF A MAJOR FACILITATOR SUPERFAMILY (MFS) TRANSPORTER, FUCP, IN THE OUTWARD CONFORMATION	ESCHERICHIA COLI	3.14
34	4GC0	THE STRUCTURE OF THE MFS (MAJOR FACILITATOR SUPERFAMILY) PROTON:XYLOSE SYMPORTER XYLE BOUND TO 6-BROMO-6-DEOXY-D-GLUCOSE	ESCHERICHIA COLI	2.60
34	2CFQ	SUGAR FREE LACTOSE PERMEASE AT NEUTRAL PH	ESCHERICHIA COLI	2.95
35	2V50	THE MISSING PART OF THE BACTERIAL MEXAB-OPRM SYSTEM: STRUCTURAL DETERMINATION OF THE	PSEUDOMONAS AERUGINOSA PA01	3.00

MULTIDRUG EXPORTER MEXB				
35	4DX5	TRANSPORT OF DRUGS BY THE MULTIDRUG TRANSPORTER ACRB INVOLVES AN ACCESS AND A DEEP BINDING POCKET THAT ARE SEPARATED BY A SWITCH-LOOP	ESCHERICHIA COLI	1.90
35	3NE5	CRYSTAL STRUCTURE OF THE CUSBA HEAVY-METAL EFFLUX COMPLEX FROM ESCHERICHIA COLI	ESCHERICHIA COLI	2.90
36	3HB3	HIGH RESOLUTION CRYSTAL STRUCTURE OF PARACOCCLUS DENITRIFICANS CYTOCHROME C OXIDASE	PARACOCCLUS DENITRIFICANS	2.25
36	3AG3	BOVINE HEART CYTOCHROME C OXIDASE IN THE NITRIC OXIDE-BOUND FULLY REDUCED STATE AT 100 K	BOS TAURUS	1.80
36	3O0R	CRYSTAL STRUCTURE OF NITRIC OXIDE REDUCTASE FROM PSEUDOMONAS AERUGINOSA IN COMPLEX WITH ANTIBODY FRAGMENT	MUS MUSCULUS	2.70
36	3MK7	THE STRUCTURE OF CBB3 CYTOCHROME OXIDASE	PSEUDOMONAS STUTZERI	3.20
36	2GSM	CATALYTIC CORE (SUBUNITS I AND II) OF CYTOCHROME C OXIDASE FROM RHODOBACTER SPHAEROIDES	RHODOBACTER SPHAEROIDES	2.00
37	4DJK	STRUCTURE OF GLUTAMATE-GABA ANTIporter GADC	ESCHERICHIA COLI	3.10
37	2A65	CRYSTAL STRUCTURE OF LEUTAA, A BACTERIAL HOMOLOG OF NA ⁺ /CL ⁻ -DEPENDENT NEUROTRANSMITTER TRANSPORTERS	AQUIFEX AEOLICUS	1.65
37	4AIN	CRYSTAL STRUCTURE OF BETP WITH ASYMMETRIC PROTOMERS.	CORYNEBACTERIUM GLUTAMICUM	3.10
37	3GIA	CRYSTAL STRUCTURE OF APCT TRANSPORTER	METHANOCALDOCOCCLUS JANNASCHII	2.32
37	3OB6	STRUCTURE OF ADIC(N101A) IN THE OPEN-TO-OUT ARG ⁺ BOUND CONFORMATION	ESCHERICHIA COLI	3.00
37	2WSW	CRYSTAL STRUCTURE OF CARNITINE TRANSPORTER FROM PROTEUS MIRABILIS	PROTEUS MIRABILIS	2.29
37	3DH4	CRYSTAL STRUCTURE OF SODIUM/SUGAR SYMPORTER WITH BOUND GALACTOSE FROM VIBRIO PARAHAEMOLYTICUS	VIBRIO PARAHAEMOLYTICUS	2.70
38	3RKO	CRYSTAL STRUCTURE OF THE MEMBRANE DOMAIN OF RESPIRATORY COMPLEX I FROM E. COLI AT 3.0 ANGSTROM RESOLUTION	ESCHERICHIA COLI	3.00
38	3RKO	CRYSTAL STRUCTURE OF THE MEMBRANE DOMAIN OF RESPIRATORY COMPLEX I FROM E. COLI AT 3.0 ANGSTROM RESOLUTION	ESCHERICHIA COLI	3.00
39	4A01	CRYSTAL STRUCTURE OF THE H-TRANSLOCATING PYROPHOSPHATASE	VIGNA RADIATA	2.35
39	4AV3	CRYSTAL STRUCTURE OF THERMOTOGA MARITIMA SODIUM PUMPING MEMBRANE INTEGRAL PYROPHOSPHATASE WITH METAL IONS IN ACTIVE SITE	THERMOTOGA MARITIMA	2.60

The numbers in the first column refer to the following families: (0) Cytochrome c oxidases – 2TM-helix subunit, (1) ATP synthase rings – 2TM subunit, (2) KIR channels, (3) K⁺-channels, (4) Complex II chain C, (5) Complex II chain D, (6) Cytochrome b₆f, (7) Light-harvesting complexes II, (8) Small mechanosensitive channels, (9) Photosystem I - subunit XI, (10) Pentameric ligand-gated ion channels, (11) Two-pore K⁺ ion channels, (12) Glutathione transferases, (13) Cytochrome b-type subunits, (14) Photosynthetic reaction centres, (15) Aquaporins, (16) Photosystems II, (17) Heterodimeric ABC transporters, (18) ABC transporters – efflux, (19) ECF transporters, (20) Potassium channels, (21) Formate channels, (22) GlpG, (23) GPCR, (24) Cytochrome c oxidases - 7TM-helix subunit, (25) Bacterial rhodopsins, (26) Cytochrome bc₁, (27) CIC transporters, (28) Urea transporters, (29) ABC transporters – influx, (30) Protein conducting channels, (31) P-type ATPases, (32) Ammonium transporters, (33) Photosystem I, (34) MFS transporters, (35) RND transporters, (36) Cytochrome c oxidases - 12TM-helix subunit, (37) FIRL-fold transporters, (38) Complex I antiporter-like subunits, (39) Pyrophosphatases.

Table A.4. β -barrel-like proteins in the HOME3 data set, listed by family

#	PDB ID	Name	Source	Res (Å)
0	1QJ8	CRYSTAL STRUCTURE OF THE OUTER MEMBRANE PROTEIN OMPX FROM ESCHERICHIA COLI	ESCHERICHIA COLI	1.90
0	2ERV	CRYSTAL STRUCTURE OF THE OUTER MEMBRANE ENZYME PAGL	PSEUDOMONAS AERUGINOSA	2.00
0	2F1V	OUTER MEMBRANE PROTEIN OMPW	ESCHERICHIA COLI K12	2.70
0	1QJP	HIGH RESOLUTION STRUCTURE OF THE OUTER MEMBRANE PROTEIN A (OMPA) TRANSMEMBRANE DOMAIN	ESCHERICHIA COLI	1.65
0	1P4T	CRYSTAL STRUCTURE OF NEISSERIA SURFACE PROTEIN A (NSPA)	NEISSERIA MENINGITIDIS	2.55
0	3DZM	CRYSTAL STRUCTURE OF A MAJOR OUTER MEMBRANE PROTEIN FROM THERMUS THERMOPHILUS HB27	THERMUS THERMOPHILUS	2.80
0	3QRA	THE CRYSTAL STRUCTURE OF AIL, THE ATTACHMENT INVASION LOCUS PROTEIN OF YERSINIA PESTIS	YERSINIA PESTIS	1.80
0	3GP6	CRYSTAL STRUCTURE OF PAGP IN SDS/MPD	ESCHERICHIA COLI	1.40
1	1I78	CRYSTAL STRUCTURE OF OUTER MEMBRANE PROTEASE OMP7 FROM ESCHERICHIA COLI	ESCHERICHIA COLI	2.60
1	2X55	YERSINIA PESTIS PLASMINOGEN ACTIVATOR PLA (NATIVE)	YERSINIA PESTIS	1.85
1	2VDF	STRUCTURE OF THE OPCA ADHESION FROM NEISSERIA MENINGITIDIS DETERMINED BY CRYSTALLIZATION FROM THE CUBIC MESOPHASE	NEISSERIA MENINGITIDIS	1.95
2	3AEH	INTEGRAL MEMBRANE DOMAIN OF AUTOTRANSPORTER HBP	ESCHERICHIA COLI	2.00
2	3FID	LPXR FROM SALMONELLA TYPHIMURIUM	SALMONELLA TYPHIMURIUM	1.90
2	4E1S	X-RAY CRYSTAL STRUCTURE OF THE TRANSMEMBRANE BETA-DOMAIN FROM INTIMIN FROM EHEC STRAIN O157:H7	ESCHERICHIA COLI	1.85
2	3QQ2	CRYSTAL STRUCTURE OF THE BETA DOMAIN OF THE BORDETELLA AUTOTRANSPORTER BRKA	BORDETELLA PERTUSSIS	3.00
2	1UYN	TRANSLOCATOR DOMAIN OF AUTOTRANSPORTER NALP FROM NEISSERIA MENINGITIDIS	NEISSERIA MENINGITIDIS	2.60
2	2WJR	NANC PORIN STRUCTURE IN RHOMBOHEDRAL CRYSTAL FORM.	ESCHERICHIA COLI	1.80
2	4E1T	X-RAY CRYSTAL STRUCTURE OF THE TRANSMEMBRANE BETA-DOMAIN FROM INVASIN FROM YERSINIA PSEUDOTUBERCULOSIS	YERSINIA PSEUDOTUBERCULOSIS	2.26
2	3SLT	PRE-CLEAVAGE STRUCTURE OF THE AUTOTRANSPORTER ESPP - N1023S MUTANT	ESCHERICHIA COLI	2.46
2	1QD6	OUTER MEMBRANE PHOSPHOLIPASE A	ESCHERICHIA COLI	2.10
2	3KVN	CRYSTAL STRUCTURE OF THE FULL-LENGTH AUTOTRANSPORTER ESTA FROM PSEUDOMONAS AERUGINOSA	PSEUDOMONAS AERUGINOSA	2.50
2	1TLY	TSX STRUCTURE	ESCHERICHIA COLI	3.01
3	3DWO	CRYSTAL STRUCTURE OF A PSEUDOMONAS AERUGINOSA FADL HOMOLOGUE	PSEUDOMONAS AERUGINOSA	2.20
3	3BRY	CRYSTAL STRUCTURE OF THE RALSTONIA PICKETTII TOLUENE TRANSPORTER TBUX	RALSTONIA PICKETTII	3.20
3	3PGU	PHE3GLU MUTANT OF ECFADL	ESCHERICHIA COLI K-12	1.70
3	3BS0	CRYSTAL STRUCTURE OF THE P. PUTIDA TOLUENE TRANSPORTER TODX	PSEUDOMONAS PUTIDA	2.60
3	2X9K	STRUCTURE OF A E.COLI PORIN	ESCHERICHIA COLI	2.18
4	3NSG	CRYSTAL STRUCTURE OF OMPF, AN OUTER MEMBRANE PROTEIN FROM SALMONELLA TYPHI	SALMONELLA ENTERICA SUBSP. ENTERICA SEROVAR	2.79
4	3VY8	CRYSTAL STRUCTURE OF PORB FROM NEISSERIA MENINGITIDIS IN COMPLEX WITH CESIUM ION, SPACE GROUP P63	NEISSERIA MENINGITIDIS	2.12
4	4GEY	HIGH PH STRUCTURE OF PSEUDOMONAS PUTIDA OPRB	PSEUDOMONAS PUTIDA	2.70
4	1OSM	OSMOPORIN (OMPK36) FROM KLEBSIELLA PNEUMONIAE	KLEBSIELLA PNEUMONIAE	3.20
4	3UPG	LOOP DELETION MUTANT OF SALMONELLA TYPHI OSMOPORIN (OMPC):AN OUTER MEMBRANE PROTEIN.	SALMONELLA ENTERICA SUBSP. ENTERICA SEROVAR	3.20
4	2J1N	OSMOPORIN OMPC	ESCHERICHIA COLI	2.00
4	2FGQ	HIGH RESOLUTION X-RAY STRUCTURE OF OMP32 IN COMPLEX WITH MALATE	DELFTIA ACIDOVORANS	1.45
4	2POR	STRUCTURE OF PORIN REFINED AT 1.8 ANGSTROMS RESOLUTION	RHODOBACTER CAPSULATUS	1.80
4	2ZFG	STRUCTURE OF OMPF PORIN	ESCHERICHIA COLI	1.59

4	3PRN	E1M, A104W MUTANT OF RH. BLASTICA PORIN	RHODOBACTER BLASTICUS	1.90
4	1PHO	CRYSTAL STRUCTURES EXPLAIN FUNCTIONAL PROPERTIES OF TWO E. COLI PORINS	ESCHERICHIA COLI	3.00
4	2O4V	AN ARGININE LADDER IN OPRP MEDIATES PHOSPHATE SPECIFIC TRANSFER ACROSS THE OUTER MEMBRANE	PSEUDOMONAS AERUGINOSA PAO1	1.94
5	2MPR	MALTOPORIN FROM SALMONELLA TYPHIMURIUM	SALMONELLA TYPHIMURIUM	2.40
5	1AF6	MALTOPORIN SUCROSE COMPLEX	ESCHERICHIA COLI	2.40
5	1A0S	SUCROSE-SPECIFIC PORIN	SALMONELLA TYPHIMURIUM	2.40
6	3SZD	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OCCK2 (OPDF)	PSEUDOMONAS AERUGINOSA	2.31
6	3T0S	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OCCK4 (OPDL)	PSEUDOMONAS AERUGINOSA	2.20
6	3SYB	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OCCD3 (OPDP)	PSEUDOMONAS AERUGINOSA	2.70
6	3T24	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OPDQ	PSEUDOMONAS AERUGINOSA	2.40
6	3RBH	STRUCTURE OF ALGINATE EXPORT PROTEIN ALGE FROM PSEUDOMONAS AERUGINOSA	PSEUDOMONAS AERUGINOSA	2.30
6	3JTY	CRYSTAL STRUCTURE OF A BENF-LIKE PORIN FROM PSEUDOMONAS FLUORESCENS PF-5	PSEUDOMONAS FLUORESCENS PF-5	2.58
6	3SY7	IMPROVED CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OPRD	PSEUDOMONAS AERUGINOSA	2.15
6	2Y2X	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OPDK WITH VANILLATE	PSEUDOMONAS AERUGINOSA PA01	1.65
6	2Y0L	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OPDO	PSEUDOMONAS AERUGINOSA	2.59
6	3SZV	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OCCK3 (OPDO)	PSEUDOMONAS AERUGINOSA	1.45
6	3SY9	CRYSTAL STRUCTURE OF PSEUDOMONAS AERUGINOSA OCCD2 (OPDC)	PSEUDOMONAS AERUGINOSA	2.80
7	3EFM	STRUCTURE OF THE ALCALIGIN OUTER MEMBRANE RECEPTUR FAUA FROM BORDETELLA PERTUSSIS	BORDETELLA PERTUSSIS	2.33
7	1QFG	E. COLI FERRIC HYDROXAMATE RECEPTOR (FHUA)	ESCHERICHIA COLI	2.50
7	3QLB	ENANTIOPYCHELIN OUTER MEMBRANE TONB-DEPENDENT TRANSPORTER FROM PSEUDOMONAS FLUORESCENS BOUND TO THE FERRI-ENANTIOPYCHELIN	PSEUDOMONAS FLUORESCENS	3.26
7	3FHH	CRYSTAL STRUCTURE OF THE HEME/HEMOGLOBIN OUTER MEMBRANE TRANSPORTER SHUA FROM SHIGELLA DYSENTERIAE	SHIGELLA DYSENTERIAE, UNDEFINED	2.60
7	2HDI	CRYSTAL STRUCTURE OF THE COLICIN I RECEPTOR CIR FROM E.COLI IN COMPLEX WITH RECEPTOR BINDING DOMAIN OF COLICIN IA.	ESCHERICHIA COLI	2.50
7	1KMO	CRYSTAL STRUCTURE OF THE OUTER MEMBRANE TRANSPORTER FECA	ESCHERICHIA COLI K12	2.00
7	2W16	STRUCTURES OF FPVA BOUND TO HETEROLOGOUS PYOVERDINES	PSEUDOMONAS AERUGINOSA	2.71
7	1XKW	PYOCHELIN OUTER MEMBRANE RECEPTOR FPTA FROM PSEUDOMONAS AERUGINOSA	PSEUDOMONAS AERUGINOSA	2.00
7	2GUF	IN MESO CRYSTAL STRUCTURE OF THE COBALAMIN TRANSPORTER, BTUB	ESCHERICHIA COLI	1.95
7	4EPA	THE CRYSTAL STRUCTURE OF THE FERRIC YERSINIABACTIN UPTAKE RECEPTOR FYUA FROM YERSINIA PESTIS	YERSINIA PESTIS	3.20
7	1FEP	FERRIC ENTEROBACTIN RECEPTOR	ESCHERICHIA COLI K12	2.40
7	3CSL	STRUCTURE OF THE SERRATIA MARCESCENS HEMOPHORE RECEPTOR HASR IN COMPLEX WITH ITS HEMOPHORE HASA AND HEME	SERRATIA MARCESCENS	2.70
7	4B7O	FRPB IRON TRANSPORTER FROM NEISSERIA MENINGITIDIS (F5-1 VARIANT)	NEISSERIA MENINGITIDIS	
7	3V8X	THE CRYSTAL STRUCTURE OF TRANSFERRIN BINDING PROTEIN A (TBPA) FROM NEISSERIAL MENINGITIDIS SEROGROUP B IN COMPLEX WITH FULL LENGTH HUMAN TRANSFERRIN	NEISSERIA MENINGITIDIS SEROGROUP B	2.60

The numbers in the first column refer to the following families: (0) 8 TM Outer Membrane Proteins, (1) 10TM Outer Membrane Proteins, (2) 12TM Outer Membrane Proteins, (3) 14TM Outer Membrane Proteins, (4) 16TM Outer Membrane Proteins, (5) Sugar porins, (6) 18TM Outer Membrane Proteins (7) 22TM Outer Membrane Proteins.

References

- Altschul, S.F., *et al.* (1990) Basic Local Alignment Search Tool, *Journal of molecular biology*, **215**, 403-410.
- Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research*, **25**, 3389-3402.
- Arinaminpathy, Y., *et al.* (2009) Computational analysis of membrane proteins: the largest class of drug targets, *Drug discovery today*, **14**, 1130-1135.
- Armen, R., Alonso, D.O. and Daggett, V. (2003) The role of alpha-, 3(10)-, and pi-helix in helix->coil transitions, *Protein science : a publication of the Protein Society*, **12**, 1145-1157.
- Attwood, T.K., *et al.* (2003) PRINTS and its automatic supplement, prePRINTS, *Nucleic acids research*, **31**, 400-402.
- Bagos, P.G., *et al.* (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins, *BMC bioinformatics*, **5**, 29.
- Bahr, A., *et al.* (2001) BALiBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations, *Nucleic acids research*, **29**, 323-326.
- Bahr, A., *et al.* (2001) BALiBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations., *Nucl. Acids Res.*, **29**, 323-326.
- Ben-Yona, A. and Kanner, B.I. (2012) An acidic amino acid transmembrane helix 10 residue conserved in the neurotransmitter:sodium:symporters is essential for the formation of the extracellular gate of the gamma-aminobutyric acid (GABA) transporter GAT-1, *The Journal of biological chemistry*, **287**, 7159-7168.
- Berbalk, C., Schwaiger, C.S. and Lackner, P. (2009) Accuracy analysis of multiple structure alignments, *Protein science : a publication of the Protein Society*, **18**, 2027-2035.
- Berg, J.M., Tymoczko, John L., Stryer, Lubert (2010) *Biochemistry*. W. H. Freeman.
- Berman, H., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank, *Nature structural biology*, **10**, 980.
- Bernsel, A., Viklund, H. and Elofsson, A. (2008) Remote homology detection of integral membrane proteins using conserved sequence features, *Proteins*, **71**, 1387-1399.
- Berntsson, R.P., *et al.* (2012) Structural divergence of paralogous S components from ECF-type ABC transporters, *Proc Natl Acad Sci U S A*, **109**, 13990-13995.
- Bigelow, H.R., *et al.* (2004) Predicting transmembrane beta-barrels in proteomes, *Nucleic acids research*, **32**, 2566-2577.
- Bill, R.M., *et al.* (2011) Overcoming barriers to membrane protein structure determination, *Nature biotechnology*, **29**, 335-340.

- Buschmann, S., *et al.* (2010) The structure of cbb3 cytochrome oxidase provides insights into proton pumping, *Science*, **329**, 327-330.
- Cartailler, J.P. and Luecke, H. (2004) Structural and functional characterization of pi bulges and other short intrahelical deformations, *Structure*, **12**, 133-144.
- Chang, J.M., *et al.* (2012) Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee, *BMC bioinformatics*, **13 Suppl 4**, S1.
- Cherezov, V., *et al.* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor, *Science*, **318**, 1258-1265.
- Choi, Y. and Deane, C.M. (2010) FREAD revisited: Accurate loop structure prediction using a database search algorithm, *Proteins*, **78**, 1431-1440.
- Chou, P.Y. and Fasman, G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence, *Advances in enzymology and related areas of molecular biology*, **47**, 45-148.
- Cline, M., Hughey, R. and Karplus, K. (2002) Predicting reliable regions in protein sequence alignments, *Bioinformatics*, **18**, 306-314.
- Collier, J.H., *et al.* (2014) A new statistical framework to assess structural alignment quality using information compression, *Bioinformatics*, **30**, i512-518.
- Cooley, R.B., Arp, D.J. and Karplus, P.A. (2010) Evolutionary origin of a secondary structure: pi-helices as cryptic but widespread insertional variations of alpha-helices that enhance protein functionality, *Journal of molecular biology*, **404**, 232-246.
- Csaba, G., Birzele, F. and Zimmer, R. (2008) Protein structure alignment considering phenotypic plasticity, *Bioinformatics*, **24**, i98-104.
- Dai, L. and Zhou, Y. (2011) Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations, *Journal of molecular biology*, **408**, 585-595.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) *A Model of Evolutionary Change in Proteins*. Atlas of Protein Sequence and Structure
- Do, C.B., *et al.* (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res*, **15**, 330-340.
- Dobrowolski, A., Sobczak-Elbourne, I. and Lolkema, J.S. (2007) Membrane topology prediction by hydropathy profile alignment: membrane topology of the Na(+)-glutamate transporter GltS, *Biochemistry*, **46**, 2326-2332.
- Dong, E., *et al.* (2008) BCL::Align - Sequence alignment and fold recognition with a custom scoring function online, *Gene*, **422**, 41-46.
- Drews, J. (2000) Drug discovery: a historical perspective, *Science*, **287**, 1960-1964.
- Ebejer, J.P., *et al.* (2013) Memoir: template-based structure prediction for membrane proteins, *Nucleic acids research*, **41**, W379-383.

- Eddy, S.R. (2011) Accelerated Profile HMM Searches, *PLoS computational biology*, **7**, e1002195.
- Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC bioinformatics*, **5**, 1-19.
- Edgar, R.C. (2009) Optimizing substitution matrix choice and gap parameters for sequence alignment, *BMC bioinformatics*, **10**, 396.
- Edgar, R.C. (2010) Quality measures for protein alignment benchmarks, *Nucleic acids research*, **38**, 2145-2153.
- Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.
- Eicher, T., *et al.* (2012) Transport of drugs by the multidrug transporter AcrB involves an access and a deep binding pocket that are separated by a switch-loop, *Proc Natl Acad Sci U S A*, **109**, 5687-5692.
- Eicher, T., *et al.* (2014) Coupling of remote alternating-access transport mechanisms for protons and substrates in the multidrug efflux pump AcrB, *eLife*, **3**.
- Eisenberg, D., *et al.* (1982) Hydrophobic moments and protein structure, *Faraday Symposia of the Chemical Society*, **17**, 109-120.
- Engelman, D.M., Steitz, T.A. and Goldman, A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Biophys Chem*, **15**, 321-353.
- Enkhbayar, P., *et al.* (2006) 3(10)-helices in proteins are parahelices, *Proteins*, **64**, 691-699.
- Erkens, G.B., *et al.* (2011) The structural basis of modularity in ECF-type ABC transporters, *Nature structural & molecular biology*, **18**, 755-760.
- Erkens, G.B., *et al.* (2012) Energy coupling factor-type ABC transporters for vitamin uptake in prokaryotes, *Biochemistry*, **51**, 4390-4396.
- Eswar, N., *et al.* (2006) Comparative protein structure modeling using Modeller, *Curr Protoc Bioinformatics*, **Chapter 5**, Unit 5 6.
- Faraggi, E., *et al.* (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles, *Journal of computational chemistry*, **33**, 259-267.
- Faraldo-Gómez, J.D. and Forrest, L.R. (2011) Modeling and simulation of ion- and ATP-driven membrane proteins, *Current opinion in structural biology*, **21**, 1-7.
- Fenollar-Ferrer, C., *et al.* (2014) Structural fold and binding sites of the human Na(+)-phosphate cotransporter NaPi-II, *Biophysical journal*, **106**, 1268-1279.
- Fenollar-Ferrer, C., *et al.* (2014) Structure and regulatory interactions of the cytoplasmic terminal domains of serotonin transporter, *Biochemistry*, **53**, 5444-5460.
- Ferguson, A.D., *et al.* (2000) A conserved structural motif for lipopolysaccharide recognition by prokaryotic and eucaryotic proteins, *Structure*, **8**, 585-592.

- Fetter, J.R., *et al.* (1995) Possible proton relay pathways in cytochrome c oxidase, *Proc Natl Acad Sci U S A*, **92**, 1604-1608.
- Fiser, A. (2010) Template-based protein structure modeling, *Methods Mol Biol*, **673**, 73-94.
- Fiser, A. and Sali, A. (2003) ModLoop: automated modeling of loops in protein structures, *Bioinformatics*, **19**, 2500-2501.
- Fodje, M.N. and Al-Karadaghi, S. (2002) Occurrence, conformational features and amino acid propensities for the pi-helix, *Protein engineering*, **15**, 353-358.
- Fooks, H.M., *et al.* (2006) Amino acid pairing preferences in parallel beta-sheets in proteins, *Journal of molecular biology*, **356**, 32-44.
- Forrest, L.R. and Rudnick, G. (2009) The rocking bundle: a mechanism for ion-coupled solute flux by symmetrical transporters, *Physiology*, **24**, 377-386.
- Forrest, L.R., Tang, C.L. and Honig, B. (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins, *Biophysical journal*, **91**, 508-517.
- Forster, I.C., *et al.* (2002) Forging the link between structure and function of electrogenic cotransporters: the renal type IIa Na⁺/Pi cotransporter as a case study, *Progress in biophysics and molecular biology*, **80**, 69-108.
- Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment, *Proteins: Structure, Function, and Bioinformatics*, **23**, 566-579.
- Garcia-Horsman, J.A., *et al.* (1995) Proton transfer in cytochrome bo₃ ubiquinol oxidase of *Escherichia coli*: second-site mutations in subunit I that restore proton pumping in the mutant Asp135-->Asn, *Biochemistry*, **34**, 4428-4433.
- Gonzalez, A., *et al.* (2012) Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors, *Chembiochem : a European journal of chemical biology*, **13**, 1393-1399.
- Govaerts, C., *et al.* (2001) The TXP motif in the second transmembrane helix of CCR5. A structural determinant of chemokine-induced activation, *The Journal of biological chemistry*, **276**, 13217-13225.
- Gromiha, M.M. and Suwa, M. (2003) Variation of amino acid properties in all-beta globular and outer membrane protein structures, *International journal of biological macromolecules*, **32**, 93-98.
- Gromiha, M.M. and Suwa, M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy, *Bioinformatics*, **21**, 961-968.
- Gront, D., *et al.* (2012) BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles, *Nucleic acids research*, **40**, W257-262.
- Hanson, M.A., *et al.* (2012) Crystal structure of a lipid G protein-coupled receptor, *Science*, **335**, 851-855.
- Hayat, M. and Khan, A. (2013) WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids, *Amino acids*, **44**, 1317-1328.

- Hayat, S. and Elofsson, A. (2012) BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins, *Bioinformatics*, **28**, 516-522.
- Heim, A.J. and Li, Z. (2012) Developing a high-quality scoring function for membrane protein structures based on specific inter-residue interactions, *Journal of computer-aided molecular design*, **26**, 301-309.
- Heinig, M. and Frishman, D. (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins, *Nucleic acids research*, **32**, W500-502.
- Henikoff, S. and Henikoff, J.G. (1992) Amino-Acid Substitution Matrices from Protein Blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.
- Henry, L.K., *et al.* (2003) Serotonin and cocaine-sensitive inactivation of human serotonin transporters by methanethiosulfonates targeted to transmembrane domain I, *The Journal of biological chemistry*, **278**, 37052-37063.
- Hessa, T., *et al.* (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon, *Nature*, **433**, 377-381.
- Hildebrand, A., *et al.* (2009) Fast and accurate automatic structure prediction with HHpred, *Proteins*, **77 Suppl 9**, 128-132.
- Hill, J.R. and Deane, C.M. (2012) MP-T: improving membrane protein alignment for structure prediction, *Bioinformatics*.
- Hill, J.R., *et al.* (2011) Environment specific substitution tables improve membrane protein alignment, *Bioinformatics*, **27**, i15-23.
- Hino, T., *et al.* (2010) Structural basis of biological N₂O generation by bacterial nitric oxide reductase, *Science*, **330**, 1666-1670.
- Holm, L. and Park, J. (2000) DaliLite workbench for protein structure comparison, *Bioinformatics*, **16**, 566-567.
- Holm, L. and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D, *Nucleic acids research*, **38**, W545-549.
- Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices, *Journal of molecular biology*, **233**, 123-138.
- Holm, L. and Sander, C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins, *Nucleic acids research*, **24**, 206-209.
- Holm, L. and Sander, C. (1996) Mapping the protein universe, *Science*, **273**, 595-603.
- Hopf, T.A., *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes, *eLife*, **3**.
- Hopkins, A.L. and Groom, C.R. (2002) The druggable genome, *Nature reviews. Drug discovery*, **1**, 727-730.

- Hopp, T.P. and Woods, K.R. (1981) Prediction of Protein Antigenic Determinants from Amino-Acid-Sequences, *Proc Natl Acad Sci U S A*, **78**, 3824-3828.
- Horn, F., *et al.* (2003) GPCRDB information system for G protein-coupled receptors, *Nucl. Acids Res.*, **31**, 294-297.
- Huang, X. (1994) On global sequence alignment, *Comput Appl Biosci*, **10**, 227-235.
- Huang, Y.H. and Chen, C.M. (2012) Statistical analyses and computational prediction of helical kinks in membrane proteins, *Journal of computer-aided molecular design*, **26**, 1171-1185.
- Islam, S.T., *et al.* (2012) A cationic lumen in the Wzx flippase mediates anionic O-antigen subunit translocation in *Pseudomonas aeruginosa* PAO1, *Molecular microbiology*, **84**, 1165-1176.
- Jackups, R., Jr. and Liang, J. (2005) Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction, *Journal of molecular biology*, **354**, 979-993.
- Jimenez-Morales, D., Adamian, L. and Liang, J. (2008) Detecting remote homologues using scoring matrices calculated from the estimation of amino acid substitution rates of beta-barrel membrane proteins, *Conf Proc IEEE Eng Med Biol Soc*, **2008**, 1347-1350.
- Johansson, L., Gafvelin, G. and Arner, E.S. (2005) Selenocysteine in proteins-properties and biotechnological use, *Biochimica et biophysica acta*, **1726**, 1-13.
- Jones, D.T. (1998) Do transmembrane protein superfolds exist?, *Febs Lett*, **423**, 281-285.
- Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *Journal of molecular biology*, **292**, 195-202.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) A mutation data matrix for transmembrane proteins, *Febs Lett*, **339**, 269-275.
- Jung, J. and Lee, B. (2000) Protein structure alignment using environmental profiles, *Protein engineering*, **13**, 535-543.
- Kabsch, W. (1976) Solution for Best Rotation to Relate 2 Sets of Vectors, *Acta Crystallogr A*, **32**, 922-923.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.
- Karplus, K. (2009) SAM-T08, HMM-based protein structure prediction, *Nucleic acids research*, **37**, W492-497.
- Kauko, A., Illergard, K. and Elofsson, A. (2008) Coils in the membrane core are conserved and functionally important, *Journal of molecular biology*, **380**, 170-180.
- Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J. Mol. Biol.*, **299**, 499-520.

- Kelm, S., Shi, J. and Deane, C.M. (2010) MEDELLER: homology-based coordinate generation for membrane proteins, *Bioinformatics*, **26**, 2833-2840.
- Khafizov, K., *et al.* (2010) A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe, *Biochemistry*, **49**, 10702-10713.
- Koehler, J., *et al.* (2009) A unified hydrophobicity scale for multispan membrane proteins, *Proteins*, **76**, 13-29.
- Koepke, J., *et al.* (2009) High resolution crystal structure of *Paracoccus denitrificans* cytochrome c oxidase: new insights into the active site and the proton transfer pathways, *Biochimica et biophysica acta*, **1787**, 635-645.
- Kolodny, R., Koehl, P. and Levitt, M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures, *Journal of molecular biology*, **346**, 1173-1188.
- Konagurthu, A.S. and Lesk, A.M. (2013) Structure description and identification using the tableau representation of protein folding patterns, *Methods Mol Biol*, **932**, 51-59.
- Konagurthu, A.S., Lesk, A.M. and Allison, L. (2012) Minimum message length inference of secondary structure from protein coordinate data, *Bioinformatics*, **28**, i97-105.
- Kopp, J., *et al.* (2007) Assessment of CASP7 predictions for template-based modeling targets, *Proteins*, **69 Suppl 8**, 38-56.
- Kowalczyk, L., *et al.* (2011) Molecular basis of substrate-induced permeation by an amino acid antiporter, *Proc Natl Acad Sci U S A*, **108**, 3935-3940.
- Kozma, D., Simon, I. and Tusnady, G.E. (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years, *Nucleic acids research*, **41**, D524-529.
- Krogh, A., *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *Journal of molecular biology*, **305**, 567-580.
- Kryshtafovych, A., Fidelis, K. and Moulton, J. (2013) CASP10 results compared to those of previous CASP experiments, *Proteins*.
- Kryshtafovych, A., Monastyrskyy, B. and Fidelis, K. (2013) CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL, *Proteins*.
- Kryshtafovych, A., *et al.* (2005) Progress over the first decade of CASP experiments, *Proteins*, **61 Suppl 7**, 225-236.
- Kukkonen, M., *et al.* (2004) Lack of O-antigen is essential for plasminogen activation by *Yersinia pestis* and *Salmonella enterica*, *Molecular microbiology*, **51**, 215-225.
- Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *Journal of molecular biology*, **157**, 105-132.
- Langelaan, D.N., *et al.* (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors, *Journal of chemical information and modeling*, **50**, 2213-2220.

- Leman, J.K., *et al.* (2013) Simultaneous prediction of protein secondary structure and transmembrane spans, *Proteins*, **81**, 1127-1140.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.
- Liu, Y., Schmidt, B. and Maskell, D.L. (2010) MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities, *Bioinformatics*, **26**, 1958-1964.
- Lolkema, J.S. and Slotboom, D.-J. (2008) The major amino acid transporter superfamily has a similar core structure as Na⁺-galactose and Na⁺-leucine transporters, *Mol Membr Biol*, **25**, 567 - 570.
- Lolkema, J.S. and Slotboom, D.J. (1998) Estimation of structural similarity of membrane proteins by hydropathy profile alignment, *Mol Membr Biol*, **15**, 33-42.
- Lolkema, J.S. and Slotboom, D.J. (1998) Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins, *FEMS Microbiol Rev*, **22**, 305-322.
- Lolkema, J.S. and Slotboom, D.J. (2003) Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis, *Journal of molecular biology*, **327**, 901-909.
- Lolkema, J.S. and Slotboom, D.J. (2005) Sequence and hydropathy profile analysis of two classes of secondary transporters, *Mol Membr Biol*, **22**, 177-189.
- Lomize, A.L., *et al.* (2006) Positioning of proteins in membranes: a computational approach, *Protein science : a publication of the Protein Society*, **15**, 1318-1333.
- Lomize, A.L., *et al.* (2007) The role of hydrophobic interactions in positioning of peripheral proteins in membranes, *BMC structural biology*, **7**, 44.
- Lomize, M.A., *et al.* (2006) OPM: orientations of proteins in membranes database, *Bioinformatics*, **22**, 623-625.
- Martinez, L., Andreani, R. and Martinez, J.M. (2007) Convergent algorithms for protein structural alignment, *BMC bioinformatics*, **8**, 306.
- Meiler, J. and Baker, D. (2003) Coupled prediction of protein secondary and tertiary structure, *Proc Natl Acad Sci U S A*, **100**, 12105-12110.
- Menke, M., Berger, B. and Cowen, L. (2008) Matt: local flexibility aids protein multiple structure alignment, *PLoS computational biology*, **4**, e10.
- Meruelo, A.D., Samish, I. and Bowie, J.U. (2011) TMKink: a method to predict transmembrane helix kinks, *Protein science : a publication of the Protein Society*, **20**, 1256-1264.
- Minor, D.L., Jr. and Kim, P.S. (1994) Measurement of the beta-sheet-forming propensities of amino acids, *Nature*, **367**, 660-663.
- Mizuguchi, K., *et al.* (1998) HOMSTRAD: a database of protein structure alignments for homologous families, *Protein science : a publication of the Protein Society*, **7**, 2469-2471.

- Moelbert, S., Emberly, E. and Tang, C. (2004) Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins, *Protein science : a publication of the Protein Society*, **13**, 752-762.
- Moult, J., *et al.* (1998) Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, **29**, 2-6.
- Moult, J., *et al.* (1997) Critical assessment of methods of protein structure prediction (CASP): round II, *Proteins*, **Suppl 1**, 2-6.
- Moult, J., *et al.* (1995) A large-scale experiment to assess protein structure prediction methods, *Proteins*, **23**, ii-v.
- Müller, T., Spang, R. and Vingron, M. (2002) Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method, *Molecular biology and evolution*, **19**, 8-13.
- Müller, T. and Vingron, M. (2000) Modeling amino acid replacement, *J Comput Biol*, **7**, 761-776.
- Muramoto, K., *et al.* (2010) Bovine cytochrome c oxidase structures enable O₂ reduction with minimization of reactive oxygens and provide a proton-pumping gate, *Proc Natl Acad Sci U S A*, **107**, 7740-7745.
- Murzin, A.G., *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *Journal of molecular biology*, **247**, 536-540.
- Needleman, S.B. and Wunsch, C.D. (1970) A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins, *Journal of molecular biology*, **48**, 443-53.
- Ng, P.C., Henikoff, J.G. and Henikoff, S. (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane, *Bioinformatics*, **16**, 760-766.
- Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *Journal of molecular biology*, **302**, 205-217.
- Nozaki, Y. and Tanford, C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale., *J. Biol. Chem.*, **246**, 2211-2217.
- Nugent, T. and Jones, D.T. (2009) Transmembrane protein topology prediction using support vector machines, *BMC bioinformatics*, **10**, 159.
- Olechnovic, K., Kulberkyte, E. and Venclovas, C. (2012) CAD-score: A new contact area difference-based function for evaluation of protein structural models, *Proteins*, **81**, 149-162
- Orengo, C.A., *et al.* (1997) CATH--a hierarchic classification of protein domain structures, *Structure*, **5**, 1093-1108.
- Ortiz, A.R., Strauss, C.E. and Olmea, O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, *Protein science : a publication of the Protein Society*, **11**, 2606-2621.

- Ostermeier, C. and Michel, H. (1997) Crystallization of membrane proteins, *Current opinion in structural biology*, **7**, 697-701.
- Pace, C.N. and Scholtz, J.M. (1998) A helix propensity scale based on experimental studies of peptides and proteins, *Biophysical journal*, **75**, 422-427.
- Pal, L. and Basu, G. (1999) Novel protein structural motifs containing two-turn and longer 3(10)-helices, *Protein engineering*, **12**, 811-814.
- Pal, L., Basu, G. and Chakrabarti, P. (2002) Variants of 3(10)-helices in proteins, *Proteins*, **48**, 571-579.
- Pandit, S.B. and Skolnick, J. (2008) Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score, *BMC bioinformatics*, **9**, 531.
- Park, J., *et al.* (2000) RSDb: representative protein sequence databases have high information content, *Bioinformatics*, **16**, 458-464.
- Pauling, L. and Corey, R.B. (1951) Atomic coordinates and structure factors for two helical configurations of polypeptide chains, *Proc Natl Acad Sci U S A*, **37**, 235-240.
- Pauling, L., Corey, R.B. and Branson, H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain, *Proc Natl Acad Sci U S A*, **37**, 205-211.
- Petrey, D. and Honig, B. (2003) GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences, *Methods Enzymol.*, **374**, 492-509.
- Pieper, U., *et al.* (2014) ModBase, a database of annotated comparative protein structure models and associated resources, *Nucleic acids research*, **42**, D336-346.
- Pirovano, W., Feenstra, K.A. and Heringa, J. (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins, *Bioinformatics*, **24**, 492-497.
- Pos, K.M. (2009) Drug transport mechanism of the AcrB efflux pump, *Biochimica et biophysica acta*, **1794**, 782-793.
- Punta, M., *et al.* (2012) The Pfam protein families database, *Nucleic acids research*, **40**, D290-301.
- Qin, L., *et al.* (2006) Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase, *Proc Natl Acad Sci U S A*, **103**, 16117-16122.
- Radestock, S. and Forrest, L.R. (2011) The alternating-access mechanism of MFS transporters arises from inverted-topology repeats, *Journal of molecular biology*, **407**, 698-715.
- Radestock, S. and Forrest, L.R. (2011) Outward-facing conformation of MFS transporters revealed by inverted-topology repeats, *Journal of molecular biology*, **407**, 698-715.
- Raghava, G., *et al.* (2003) OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy, *BMC bioinformatics*, **4**, 47.
- Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations, *Journal of molecular biology*, **7**, 95-99.

- Raman, P., Cherezov, V. and Caffrey, M. (2006) The Membrane Protein Data Bank, *Cellular and molecular life sciences : CMLS*, **63**, 36-51.
- Randall, A., *et al.* (2008) TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins, *Bioinformatics*, **24**, 513-520.
- Ray, A., Lindahl, E. and Wallner, B. (2010) Model quality assessment for membrane proteins, *Bioinformatics*, **26**, 3067-3074.
- Ray, A., Lindahl, E. and Wallner, B. (2012) Improved model quality assessment using ProQ2, *BMC bioinformatics*, **13**, 224.
- Reddy Ch, S., *et al.* (2006) Homology modeling of membrane proteins: a critical assessment, *Computational biology and chemistry*, **30**, 120-126.
- Remmert, M., *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nat Methods*, **9**, 173-175
- Remmert, M., *et al.* (2010) Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin, *Molecular biology and evolution*, **27**, 1348-1358.
- Ren, Q., Chen, K. and Paulsen, I.T. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels, *Nucleic acids research*, **35**, D274-279.
- Riek, R.P., *et al.* (2008) Wide turn diversity in protein transmembrane helices implications for G-protein-coupled receptor and other polytopic membrane protein structure and function, *Molecular pharmacology*, **73**, 1092-1104.
- Riek, R.P. and Graham, R.M. (2011) The elusive pi-helix, *Journal of structural biology*, **173**, 153-160.
- Riek, R.P., *et al.* (2001) Non-alpha-helical elements modulate polytopic membrane protein architecture, *Journal of molecular biology*, **306**, 349-362.
- Rose, G.D. (1978) Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature*, **272**, 586-590.
- Rovati, G.E., Capra, V. and Neubig, R.R. (2007) The highly conserved DRY motif of class A G protein-coupled receptors: beyond the ground state, *Molecular pharmacology*, **71**, 959-964.
- Sadowski, M.I. and Taylor, W.R. (2012) Evolutionary inaccuracy of pairwise structural alignments, *Bioinformatics*, **28**, 1209-1215.
- Sahraeian, S.M. and Yoon, B.J. (2010) PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences, *Nucleic acids research*, **38**, 4917-4928.
- Sahraeian, S.M. and Yoon, B.J. (2011) PicXAA-Web: a web-based platform for non-progressive maximum expected accuracy alignment of multiple biological sequences, *Nucleic acids research*, **39**, W8-12.
- Saier, M.H., Jr., Tran, C.V. and Barabote, R.D. (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information, *Nucleic acids research*, **34**, D181-186.

Saigo, H., Vert, J.-P. and Akutsu, T. (2006) Optimizing amino acid substitution matrices with a local alignment kernel, *BMC bioinformatics*, **7**, 246.

Šali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints., *J. Mol. Biol.*, **234**, 779-815.

Sansom, M.S., Scott, K.A. and Bond, P.J. (2008) Coarse-grained simulation: a high-throughput computational approach to membrane proteins, *Biochemical Society transactions*, **36**, 27-32.

Schushan, M., *et al.* (2012) A model-structure of a periplasm-facing state of the NhaA antiporter suggests the molecular underpinnings of pH-induced conformational changes, *The Journal of biological chemistry*, **287**, 18249-18261.

Schwacke, R., *et al.* (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins, *Plant physiology*, **131**, 16-26.

Scordis, P., Flower, D.R. and Attwood, T.K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database, *Bioinformatics*, **15**, 799-806.

Senes, A., Gerstein, M. and Engelman, D.M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions, *Journal of molecular biology*, **296**, 921-936.

Shafir, Y. and Guy, H.R. (2004) STAM: simple transmembrane alignment method, *Bioinformatics*, **20**, 758-769.

Shakhnovich, B.E., *et al.* (2005) Protein structure and evolutionary history determine sequence space topology, *Genome Res*, **15**, 385-392.

Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein engineering*, **11**, 739-747.

Shu, N. and Elofsson, A. (2011) KalignP: improved multiple sequence alignments using position specific gap penalties in Kalign2, *Bioinformatics*, **27**, 1702-1703.

Sievers, F., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular systems biology*, **7**, 539.

Sigrist, C.J., *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotation, *Nucleic acids research*, **38**, D161-166.

Sigrist, C.J., *et al.* (2013) New and continuing developments at PROSITE, *Nucleic acids research*, **41**, D344-347.

Slater, A.W., *et al.* (2012) Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments, *Bioinformatics*.

Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences, *Journal of molecular biology*, **147**, 195-197.

Söding, J. (2005) Protein homology detection by HMM-HMM comparison, *Bioinformatics*, **21**, 951-960.

- Srinivasan, G., James, C.M. and Krzycki, J.A. (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA, *Science*, **296**, 1459-1462.
- Stamm, M. (2010) Design and Testing of Membrane Protein Sequence Alignment Tools. *Fachbereich Biowissenschaften*. Johann Wolfgang Goethe-Universitaet Frankfurt am Main.
- Stamm, M., *et al.* (2013) Alignment of Helical Membrane Protein Sequences Using AlignMe, *PloS one*, **8**, e57731.
- Stamm, M., *et al.* (2014) AlignMe - a membrane protein sequence alignment web server, *Nucleic acids research*, **42**, W246-W251
- Stamm, M. and Forrest, L. (2015) Structure alignment of membrane proteins: Accuracy of available tools and a consensus strategy, *Proteins*, **83**, 1720-1732.
- Stebbing, L.A. and Mizuguchi, K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database, *Nucleic acids research*, **32**, D203-207.
- Stevens, T.J. and Arkin, I.T. (2001) Substitution rates in alpha-helical transmembrane proteins, *Protein science : a publication of the Protein Society*, **10**, 2507-2517.
- Suzek, B.E., *et al.* (2007) UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics*, **23**, 1282-1288.
- Tang, C.L., *et al.* (2003) On the role of structural information in remote homology detection and sequence alignment: New methods using hybrid sequence profiles, *Journal of molecular biology*, **334**, 1043-1062.
- Taylor, W.R. (1999) Protein structure comparison using iterated double dynamic programming, *Protein science : a publication of the Protein Society*, **8**, 654-665.
- Taylor, W.R. (2000) Protein structure comparison using SAP, *Methods Mol Biol*, **143**, 19-32.
- Teichert, F., Bastolla, U. and Porto, M. (2007) SABERTOOTH: protein structural alignment based on a vectorial structure representation, *BMC bioinformatics*, **8**, 425.
- ter Horst, R. and Lolkema, J.S. (2012) Membrane topology screen of secondary transport proteins in structural class ST[3] of the MemGen classification. Confirmation and structural diversity, *Biochimica et biophysica acta*, **1818**, 72-81.
- Thomas, J.W., *et al.* (1993) Site-directed mutagenesis of highly conserved residues in helix VIII of subunit I of the cytochrome bo ubiquinol oxidase from Escherichia coli: an amphipathic transmembrane helix that may be important in conveying protons to the binuclear center, *Biochemistry*, **32**, 11173-11180.
- Thompson, A.A., *et al.* (2012) Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic, *Nature*, **485**, 395-399.
- Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. In, *Curr Protoc Bioinformatics*. pp. Unit 2 3.

- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucl. Acids Res.*, **22**, 4673-4680.
- Thompson, J.D., *et al.* (2005) BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark, *Proteins*, **61**, 127-136.
- Thompson, J.D., Plewniak, F. and Poch, O. (1999) BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs, *Bioinformatics*, **15**, 87-88.
- Tress, M., *et al.* (2005) Assessment of predictions submitted for the CASP6 comparative modeling category, *Proteins*, **61 Suppl 7**, 27-45.
- Tsaousis, G.N., Bagos, P.G. and Hamodrakas, S.J. (2014) HMMpTM: improving transmembrane protein topology prediction using phosphorylation and glycosylation site prediction, *Biochimica et biophysica acta*, **1844**, 316-322.
- Tsirigos, K.D., Bagos, P.G. and Hamodrakas, S.J. (2011) OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria, *Nucleic acids research*, **39**, D324-331.
- Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucleic acids research*, **33**, D275-278.
- Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2005) TMDet: web server for detecting transmembrane regions of proteins by using their 3D coordinates, *Bioinformatics*, **21**, 1276-1277.
- Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics*, **20**, 2964-2972.
- Tusnady, G.E., Dosztanyi, Z. and Simon, I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucl. Acids Res.*, **33**, D275-D278.
- Uhlen, M., *et al.* (2015) Proteomics. Tissue-based map of the human proteome, *Science*, **347**, 1260419.
- Ulmschneider, M.B. and Sansom, M.S. (2001) Amino acid distributions in integral membrane protein structures, *Biochimica et biophysica acta*, **1512**, 1-14.
- UniProt, C. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013, *Nucleic acids research*, **41**, D43-47.
- Veeramalai, M., Ye, Y. and Godzik, A. (2008) TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model, *BMC bioinformatics*, **9**, 358.
- Vieira-Pires, R.S. and Morais-Cabral, J.H. (2010) 3(10) helices in channels and other membrane proteins, *The Journal of general physiology*, **136**, 585-592.
- Viklund, H. and Elofsson, A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, *Bioinformatics*, **24**, 1662-1668.

Viklund, H., Granseth, E. and Elofsson, A. (2006) Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes, *Journal of molecular biology*, **361**, 591-603.

von Heijne, G. (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues, *Nature*, **341**, 456-458.

von Mering, C., *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic acids research*, **33**, D433-437.

Wang, Z., *et al.* (2011) Protein 8-class secondary structure prediction using conditional neural fields, *Proteomics*, **11**, 3786-3792.

Weaver, T.M. (2000) The pi-helix translates structure into function, *Protein science : a publication of the Protein Society*, **9**, 201-206.

Werner, T. and Church, W.B. (2013) Kink characterization and modeling in transmembrane protein structures, *Journal of chemical information and modeling*, **53**, 2926-2936.

White, S.H. (2004) The progress of membrane protein structure determination, *Protein Sci.*, **13**, 1948-1949.

Wilcoxon, F. (1946) Individual comparisons of grouped data by ranking methods, *Journal of economic entomology*, **39**, 269.

Wimley, C.W. and White, S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces, *Nature Struct. Biol.*, **3**, 842-848.

Wimley, W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures, *Protein science : a publication of the Protein Society*, **11**, 301-312.

Wimley, W.C. (2003) The versatile beta-barrel membrane protein, *Current opinion in structural biology*, **13**, 404-411.

Wu, H., *et al.* (2012) Structure of the human kappa-opioid receptor in complex with JDTC, *Nature*, **485**, 327-332.

Wu, S. and Zhang, Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information, *Proteins*, **72**, 547-556.

Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5?, *Bioinformatics*, **26**, 889-895.

Yaffe, D., *et al.* (2014) Functionally important carboxyls in a bacterial homologue of the vesicular monoamine transporter (VMAT), *The Journal of biological chemistry*, **289**, 34229-34240.

Yamashita, A., *et al.* (2005) Crystal structure of a bacterial homologue of Na⁺/Cl⁻-dependent neurotransmitter transporters, *Nature*, **437**, 215-223.

- Yang, A.S. and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance, *Journal of molecular biology*, **301**, 665-678.
- Yarov-Yarovoy, V., Schonbrun, J. and Baker, D. (2006) Multipass membrane protein structure prediction using Rosetta, *Proteins*, **62**, 1010-1025.
- Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists, *Bioinformatics*, **19 Suppl 2**, ii246-255.
- Ye, Y. and Godzik, A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching, *Nucleic acids research*, **32**, W582-585.
- Zemla, A. (2003) LGA: A method for finding 3D similarities in protein structures, *Nucleic acids research*, **31**, 3370-3374.
- Zhang, P., Wang, J. and Shi, Y. (2010) Structure and mechanism of the S component of a bacterial ECF transporter, *Nature*, **468**, 717-720.
- Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality, *Proteins*, **57**, 702-710.
- Zhang, Y. and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic acids research*, **33**, 2302-2309.