Computational Investigation of Structural and Thermodynamic Properties of Beta-Barrel Membrane Proteins

BY

WEI TIAN
B.S., Shanghai Jiao Tong University, 2009

THESIS
Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:
Jie Liang, Chair and Advisor
Thomas Royston
Yang Dai
Ao Ma
Hammad Naveed, National University of Computer and Emerging Sciences, Pakistan

ACKNOWLEDGMENTS

First and foremost, I would like to thank Professor Jie Liang, my adviser and mentor, for his continuous encouragement, support, and inspiration. This thesis would not be possible without Jie's patient guidance and unwavering trust in me during the past years. His immense knowledge and great sense of research directions always help me out when I got stuck in my research. His enthusiasm for science motivates me to work harder and helps me to develop interest in diverse research directions.

I also want to thank my former colleague and the defense committee member Dr. Hammad Naveed. Hammad provided tremendous amount of assistance and opportunities during my PhD. His knowledge and experience fueled my studies, and his encouragement led me through many frustrating moments. This thesis would have been impossible if Hammad were not there. I would like to thank the other defense committee members Professor Thomas Royston, Professor Yang Dai, and Professor Ao Ma for taking the time out of their busy schedules to serve on my preliminary exam committee and dissertation defense committee. Your teachings and insightful comments are greatly appreciated.

Very special thanks go to my wonderful former colleague and office–mate Dr. Gamze Gursoy. Gamze was a like a big sister to me, and offered enormous help and inspiring insights in both my research career and my life. I could not have asked for a better friend and colleague.

I am also very grateful to my (former) colleagues: Dr. Meishan Lin, Dr. Jieling Zhao, and Boshen Wang. Meishan was a great senior for me, and provided patient guidance when

ACKNOWLEDGMENTS (Continued)

I began my research project on β–barrel proteins, which really gave me a jump–start. Jieling was always willing to offer his time and efforts when I needed help or discussion in diverse projects. Boshen have been providing wonderful tech support for my computational work and stimulating discussions in research directions. I could not imagine how much more difficult my PhD research would have been without your companion.

I want to thank my fellow lab alumni: Dr. Youfang Cao, Dr. David Jimenez Morales, Dr. Yun Xu, and Dr. Ke Tang. I would also like to thank my fellow colleagues: Anna Terebus, Alan Perez–Rathke, Farid Manuchehrfar, Pourya Delafrouz, Lin Du, and Samira Mali. It has been a great pleasure and honor working alongside you all.

Finally, I want to offer my sincere thank to my wife Dr. Xue Lei, who supported me unwaveringly through sunny and gloomy days during the completion of my degree.

CONTRIBUTION OF AUTHORS

Chapter 1 is an introduction to the background of my dissertation research, which places my dissertation question in the context of the bigger field and highlights its significance. Chapter 2 represents a published manuscript of which I was the first author. Jie Liang, Hammad Naveed, and I designed the research. Meishan Lin and I contributed to the input dataset. Hammad Naveed and I performed the research. I prepared all the figures and tables. All the authors analyzed the data. Jie Liang, Hammad Naveed and I wrote the paper. Chapter 3 represents a published manuscript of which I was the first author. All the authors designed the research. I performed the research, and prepared all the figures and tables. Jie Liang and I analyzed the data, and wrote the paper. Chapter 3 represents a submitted manuscript of which I was the first author. Jie Liang and I designed the research. Meishan Lin and I contributed to the input dataset. I performed the research, and prepared all the figures and tables. Jie Liang, Hammad Naveed, and I analyzed the data. Jie Liang and I wrote the paper.

TABLE OF CONTENTS

CHAPTER PAGE

1 INTRODUCTION 1
1.1 Overview 1
1.2 Structures of β–barrel membrane proteins 2
1.3 Thermodynamical properties of β–barrel membrane proteins . 4

2 HIGH RESOLUTION STRUCTURE PREDICTION OF β-BARREL MEMBRANE PROTEINS 7
2.1 Introduction 7
2.2 Methods and Materials 9
2.2.1 Dataset 9
2.2.2 Workflow of 3D–BMPP 10
2.2.3 Secondary structure determination 10
2.2.4 Strand Register Prediction 12
2.2.4.1 Model for interstrand interactions 13
2.2.4.2 Model for sequence covariation 14
2.2.4.3 Parameter determination and cross–validation 15
2.2.4.4 Sidechain direction prediction 17
2.2.4.5 Global register optimization 18
2.2.5 Parametric model for 3D structures of barrel domains. 21
2.2.5.1 C_α trace construction 21
2.2.5.2 Parameter estimation 26
2.2.5.3 Construction of backbones and sidechains 27
2.2.6 Predict structures of βMP loops 28
2.3 Results 28
2.3.1 Register prediction 28
2.3.2 Predicted 3D structures of TM regions of βMPs 30
2.3.2.1 Intrinsic flexibility 31
2.3.2.2 Predicted structures of βMPs with novel folds. 32
2.3.2.3 Predicted structures of non-TM regions of βMPs 33
2.3.2.4 Detailed prediction results of the barrel domains of all 51 βMPs 33
2.3.2.5 Comparison with a previous study. 34
2.3.3 Structures of loops. 35
2.4 Discussion 39

3 EFFICIENT COMPUTATION OF TRANSFER FREE ENERGIES OF AMINO ACIDS IN BETA–BARREL MEMBRANE PROTEINS 42

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

3.1 Introduction 42
3.2 Methods and Materials 45
3.2.1 Dataset 45
3.2.2 Reduced state space 45
3.2.3 Computing the TFE of an amino acid residue 48
3.2.4 Exact algorithm of partition function calculation 49
3.2.4.1 Exact algorithm 49
3.2.4.2 Time complexity 50
3.2.5 Approximation algorithm of partition function calculation . . 52
3.2.5.1 Approximation algorithm 52
3.2.5.2 Strand reindexing 53
3.2.5.3 Time complexity 53
3.2.6 Approximation algorithm with histogram scheme 55
3.2.6.1 Histogram scheme 55
3.2.6.2 Time complexity 56
3.3 Results 59
3.3.1 Accuracy of the approximation algorithm 59
3.3.2 Efficiency of the approximation algorithms 63
3.3.3 Transfer free energy profile of a large βMP 63
3.4 Discussion 63

4 GETFEP: A GENERAL TRANSFER FREE ENERGY PROFILE OF TRANSMEMBRANE PROTEINS 68
4.1 Introduction 68
4.2 Materials and methods 70
4.2.1 Dataset 70
4.3 Results 70
4.3.1 GeTFEP: General Transfer Free Energy Profile 70
4.3.1.1 Computation of TFE profiles of βMPs 70
4.3.1.2 Derivation of the general TFE profiles of βMPs by clustering analysis. 71
4.3.1.3 Comparison with other hydrophobicity scales 73
4.3.2 Insertion of βMPs into membrane 76
4.3.2.1 βMP insertion as a thermodynamically driven spontaneous process 76
4.3.2.2 Importance of patterns of TM lipid–facing residues in membrane insertion 79
4.3.2.3 Roles of residues in different TM regions during membrane insertion 80
4.3.2.4 Prediction of βMP positioning and orientation in the membrane 81
4.3.3 Prediction of structurally and functionally important sites of βMPs 83

4.3.4 GeTFEP can predict TM region of α–helical membrane proteins 85
4.3.4.1 The validity of transfer free energy value of Pro in the GeTFEP 87
4.4 Conclusions and discussion 87

5 CONCLUSION 91
5.1 3D structure prediction of βMPs 91
5.2 Transfer free energy computation 92
5.3 GeTFEP: General Transfer Free Energy Profile for membrane proteins 92

APPENDICES 93

CITED LITERATURE 109

VITA 130

LIST OF TABLES

TABLE **PAGE**

I The groups of βMPs in this study. All the six groups are used in the construction of the empirical energy function. Structure predictions are made for only the first five groups. 16

II Values for α, β, γ, δ, ε, w_0, w_1, and w_2. 17

III Shear number of βMPs. 20

IV Comparison of different methods for strand register and 3D structure prediction for TM regions of βMPs. 3D-BMPP can predict strand registers with an accuracy of ∼85%, and 3D structures of TM regions with an average mainchain RMSD of 3.48Å and average all atom RMSD of 4.26Å for a much bigger dataset (51 vs 14 – 23 βMPs). . . 29

V Flexibility of TM regions of βMPs and the accuracy of the prediction of 3D-BMPP. $D^{\mathrm{TM}}_{s_1,s_2}$ is the average of the mutual C_α-RMSD between structures s_1 and s_2. * As no X-ray structures for these proteins are available, we used the first model of the NMR data. 32

VI Protein size and average mainchain RMSD using different prediction methods. Proteins with the number of strands ≤ 14 are grouped into the small dataset, those with > 14 and ≤ 20 strands are grouped into the medium dataset, and proteins with > 20 strands are grouped into the large dataset. In contrast to the other prediction methods, the quality of prediction of our methods, 3D-BMPP, does not deteriorate for large–sized proteins. 36

VII Comparison of the accuracy of loop prediction for βMPs. We are able to sample most of the loop conformations seen in the NMR structures with <3Å deterioration in C_α-RMSD. 38

VIII The PDB codes of the βMPs used in this study. 45

LIST OF TABLES (Continued)

TABLE **PAGE**

IX Average running time using the exact algorithm (Algorithm 1) and the approximation algorithm with the histogram scheme (Algorithm 3) to calculate all 20 substitutions for a given host position. These values are recorded from programs running on a 2600MHZ CPU written in C++ programming language except those marked by *, which are extrapolated based on the complexity analysis (Equation 3.4). . . . 51

X The average RMSEs between the results of the approximation algorithms and of the exact algorithm. The column BD shows the RMSEs of the TFEs of the residues on or neighboring a boundary strand, while the column non–BD shows the RMSEs of the residues in the interior of a half–barrel segment. Small RMSE values correspond to high accuracy. 60

XI The RMSEs between results of exact algorithms with and without the histogram scheme. Different bin sizes (BS) are tested. The unit of the bin sizes is *kcal*/*mol*. 61

XII Comparison between the method in this study and the other computational methods. The R^2 of the results of each computational method is calculated against experimentally measured values (1). The *original* column is calculated with all 20 amino acids, while the *filtered* column is calculated with Cys, Met, Thr, Trp, and Tyr excluded.
^{1}: Data of Cys not reported.
^{2}: Data of Trp and Tyr not reported.
^{3}: Data of Cys, Met, and Thr not reported. 61

XIII Comparison between the predicted positioning and orientation and experimental results (2) of βMPs. 83

XIV Predicted important sites in OmpLA, PagP and PagL by deviation analysis 85

XV Prediction of TM segments and residues αMPs. The mid-GeTFEP scale performs better than the other hydrophobicity scales. The first three scales are measured or derived in α–helical systems, the others in βMPs. 88

TABLE **PAGE**

XVI Dataset and prediction results. Strand register prediction and RMSD between the TM region and the TM+extended barrel regions of real and modeled structures of 51 non–homologous βMPs. (Continue on the next page) 94

XVI (Cont'd) Dataset and prediction results. Strand register prediction and RMSD between the TM region and the TM+extended barrel regions of real and modeled structures of 51 non–homologous βMPs. 95

XVII Examples of the 20D TFE vectors calculated using our methods. The RMSEs between the results calculated from the approximation algorithm (Algorithm 3) and the exact algorithm (Algorithm 1) are listed in the last row. The average RMSEs in Table X and Table XI in the main text are computed from this kind of vectors of corresponding residues. 99

XVIII The insertion TFEs of WT and lipid–facing residue–shuffled βMPs calculated with GeTFEP. The $\Delta\Delta G$ shows the differences between TFEs of the WT βMPs at step 0 and the average of the minimum TFEs of the lipid–facing residue–shuffled βMPs. (Continue on the next page) 101

XVIII (Cont'd) The insertion TFEs of WT and lipid–facing residue–shuffled βMPs calculated with GeTFEP. The $\Delta\Delta G$ shows the differences between TFEs of the WT βMPs at step 0 and the average of the minimum TFEs of the lipid–facing residue–shuffled βMPs. 102

XIX **T**he computed insertion TFEs of the 6 βMPs that do not have the minimum energy at position 0. However, their most stable position is close to 0, and the minimum TFEs are close to their TFEs at position 0. Nonetheless, we exclude these 6 βMPs in our other analysis of membrane insertion stability. 103

LIST OF FIGURES

FIGURE **PAGE**

1 **A.** OmpF (PDB code: 2omf) has a barrel–like structure formed by β–sheet. **B.** The top view of OmpF shows that its TM domain devides the space into lumen–side and lipid–side. **C.** The barrel domain of α–hemolysin (PDB code: 7ahl) is formed by repeated β–strand hairpin. **D.** OmpF forms the functional unit in the membrane in a homotrimeric state. 3

2 **A.** Interstrand interactions between β–strands of βMPs. Red dashed lines indicate the strong H–bonds, green the weak H–bonds, and blue the sidechain interactions. Figure adapted from Ref (3). **B.** An example hydrophobicity scale measured in Ref(4). Figure adapted from Ref (4). 5

3 The flowchart of βMP structure prediction method 3D–BMPP. The strand registers are predicted using a combination of empirical energy function and sequence covariation information. Global shear optimization is then performed upon the predicted register candidates. The 3D coordinates of C_α atoms of TM and non-TM residues are then predicted using a parametric structural model. We also predict ensembles of loop conformations. 11

4 Model for interstrand interactions between adjacent strands. 13

5 Model for calculating sequence covariation between adjacent strands. . 15

6 Shear number is the displacement of the relative positions in the TM strands if one starts to follow the backbone H–bond between strands, starting from strand 1 and returning after a full circle to the same strand. For example, the shear number for the 4 strand β-barrel shown above is $n - 1$. 19

7 Geometric properties of βMPs. (a) The C_α trace of a β-strand shows a zigzag pattern (red). The structure used here is TodX (PDB code:3bs0). (b) Distribution of distance between consecutive C_α atoms on the same strand. (c) Distribution of distance between C_α atoms of residues sharing a non–H–bond (green) and of those sharing a strong H–bond (blue) on adjacent strands. 23

LIST OF FIGURES (Continued)

FIGURE **PAGE**

8 The parametric structural model of intertwined zigzag coils. (a) One zigzag coil (blue) and the corresponding coil basis (black) wrap around the hypothetical cylinder (grey). (b) The relative position and the corresponding parameters of coil bases are shown after unwrapping the coil bases onto a plane. 24

9 Structure prediction of TM regions. (A) Predicted structures of the TM regions (green) superimposed on experimentally determined structures (cyan): OmpA (1bxw), TodX (3bs0), Porin (1prn), BamA (4n75), OpdO (3szv), and HasR (3csl). (B) Predicted structures of the TM regions of proteins with novel folds (green) superimposed on experimentally determined structures (cyan): VDAC (3emn), FimD (3rfz), PapC (2vqi), and LptD (4q35). PapC and LptD are shown in top view. 31

10 RMSDs of our prediction against the size of the proteins. Each blue dot represents one of the 51 predicted structures, while each red dot shows the average RMSD of predicted structures with the same corresponding strand number. 36

11 Our method on register prediction does not suffer from the limitation of requiring a large number ($\sim$ 1000) of available sequences for sequence covariation analysis. This figure shows how the number of available sequences affect register prediction accuracy. The numbers of sequences are found by HHblits (5). Each blue dot represents the register prediction for one protein using our model, while each red dot represents the prediction made by the sequence covariation analysis results alone (using Equation 2.4 and Figure 5). The blue and red curves are fitted from the corresponding dots, respectively. The inset shows the details of proteins when the available number of sequences is limited. In these cases, our model can still make accurate prediction (accuracy at $\sim 0.7 = 70\%$) while the prediction made using covariation analysis along is not reliable ($\sim 0.3 = 30\%$) 37

12 Structure prediction of loop regions. (A) Ensemble of predicted loop structures of OmpX (1qj8). (B & C) Examples of predicted loops on extracellular side (b, green) and on periplasmic side (c, green) superimposed on the corresponding NMR structure (cyan) (6). The black arrows indicate the big fluctuations in the barrel region. 39

13 OmpLA and the residue 210. Figure adapted from Ref (1) 43

FIGURE **PAGE**

14 The reduced state space and the enumeration of conformational states . **A**. A strand interacts with neighbor strands through three different types of interactions (3). Each strand can slide up or down l positions. **B**. In approximation algorithms, the TM region of a βMP is divided into two half–barrel segments (grey) and two boundary strands (red). . . . 47

15 TFE $\Delta\Delta G^{210}$s at the host position 210 of OmpLA. Computed TFEs are shown in blue bars. Experimentally measured values are shown in red bars. The results of the approximation method correlate well with the experimental data with an R^2 of 0.81 (inset). 62

16 The depth–dependent TFE profile of lipopolysaccharide transport proteins D (PDB code: 4q35). The profile was calculated by systematically substituting each residue at all lipid–facing host positions in the TM region to the other 19 amino acids. The averages and the standard deviations of the calculated TFEs of the same amino acid substitution at host positions of the same depth are plotted. The position index of the depth is 0 at the midplane, and -4 – -1 on the periplasmic leaflet and 1 – 4 on the extracellular leaflet. 64

17 **A.** Visualization of the βMP TFE profiles after reduced to a 3D space via PCA. **B.** The silhouette scores for different cluster numbers of the βMP TFE profiles. 72

18 Derivation of the GeTFEP. **A.** Results of hierarchical clustering shows that all βMPs in the dataset can be group into one cluster, except α– and γ–hemolysins (7ahl and 3b07). **B.** The distribution of distance between the sampled TFE profiles of the artificially constructed hemolysin–like βMPs and their average TFE profile. The distances of both α– and γ–hemolysins are at the 80th percentile of the distribution. **C.** The General Transfer Free Energy Profile (GeTFEP) of each residues (blue), and the corresponding curves fitted by 3rd degree polynomials (red). 74

19 Comparison between the GeTFEP and the experimentally measured MF–scale. **A.** The mid–GeTFEP scale agrees well with the MF–scale, except Pro and His. **B.** The depth–dependent TFEs of Arg and Leu of GeTFEP also agree well with the experimental measurements (1). . . 75

LIST OF FIGURES (Continued)

FIGURE **PAGE**

20 Illustration of the membrane insertion process of βMPs. **A.** The simplified βMP insertion model following Ref (7). **B.** The computed insertion energies of the OmpA protein at different depth positions. Contribution of different regions are color coded. **C.** Illustration on how free energies change with the position of βMP in the membrane. The dashed red segments show that lipid–facing residues in extracellular head group region sometimes become energetically unfavorable. 78

21 An example of the insertion TFEs (Omp32, PDB code:1e54). **A.** The insertion TFEs of the intact Omp32 shows a funnel pattern. **B.** The insertion TFEs of the residue–shuffled Omp32 regardless of sidechain directions. **C.** The insertion TFEs of the lipid–facing–residue–shuffled Omp32 80

22 Prediction of positioning and orientation of βMPs. **A.** The positioning and orientation of the βMP inside the membrane are determine by the rotation angles θ_x and θ_y, the translation displacement d_z, and the membrane thickness h. **B.** The funnel–like landscape of the stability of the BtuB protein. It shows how rotation angles affect the stability of BtuB when d_z and h are fixed. 82

23 Predicted important sites of PagP. Residues 27I and 125L are at the sites where the H–bonds between the β–strands are disrupt. 69L has interaction with the out–clamp α–helix of PagP. 86

24 Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan). (Continue on the next page) 96

24 (Cont'd) Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan). 97

25 Structures and TFE profiles of α– and γ–hemolysin. **A.** A typical TFE profile of βMPs (FptA, PDB code: 1xkw). **B.** The structures of α–hemolysin (PDB code: 7ahl) and γ–hemolysin (PDB code: 3b07) Both TM segments are constructed with repeated β–hairpins. **C.** The TFE profile of α–hemolysin. **D.** The TFE profile of γ–hemolysin. Since the structures are both repeated hairpin, there is only one data point for each amino acid residue in every depth of their profiles. 100

LIST OF FIGURES (Continued)

FIGURE PAGE

26 Comparison between the mid–GeTFEP scale and other hydrophobicity scales. The mid-GeTFEP scale agrees well with previously measured or derived hydrophobicity scales. 104

27 Hydropathy analysis with mid–GeTFEP. The blue segments are the known TM segments, while the red ones are predicted by the hydropathy analysis. The analysis was carried out using Membrane Protein Explorer (MPEx) (8) **A**. An example (AChR pore α subunit) shows both the TM region and the number of the TM segments are correctly predicted. **B**. An example (AChR pore γ subunit) shows the predicted number of the TM segments are wrong, though the TM regions are correctly predicted. 105

28 The sym–GeTFEP for symmetric membranes. This profile is derived by mirroring the left part (depth -4 to -1) of the original GeTFEP. 106

LIST OF ABBREVIATIONS

βMP	β–barrel membrane protein
TM	Transmembrane
OMP	Outer membrane protein
TFE	Transfer free energy
H–bond	hydrogen bond
Cryo–EM	cryogenic electron microscopy
VDAC	Voltage–dependent anion channel
OmpF	Outer membrane protein F
OmpLA	Outer membrane phospholipase A
NMR	Nuclear magnetic resonance
PDB	Protein Data Bank
RMSD	Root–mean–square deviation
3D–BBMP	3D Beta–barrel Membrane Protein Predictor
β–barrel assembly machinery component A	BamA
OmpA	Outer membrane protein A
OmpX	Outer membrane protein X
OmpG	Outer membrane protein G

LIST OF ABBREVIATIONS (Continued)

OmpW	Outer membrane protein W
OprH	Outer membrane protein H
DSSP	Define Secondary Structure of Proteins
LOOCV	Leave–one–out cross–validation
MD	Molecular Dynamics
ER	Endoplasmic reticulum
RMSE	Root–mean–square error
GeTFEP	General Transfer Free Energy Profile
PCA	Principal Component Analysis
OPM	Orientations of Proteins in Membranes
αMP	α–helical membrane protein
MPEx	Membrane Protein Explorer

SUMMARY

Transmembrane (TM) proteins are important proteins as they serve as gateways to permit substance transport and/or signaling transduction between interior and exterior of cells. β–barrel membrane proteins (βMPs) are a major type of TM proteins. They are solely found in the outer membranes of Gram–negative bacteria, mitochondria, and chloroplast. βMPs serve a multitude of essential cellular functions, including reaction catalysis, protein anchoring, metabolite transportation, and outer–membrane biogenesis. In bacteria, βMPs are also found to be responsible for the release of virulence factors and are implicated in multidrug resistance. Dysfunctional βMPs in mitochondria are related to neurodegenerative diseases as well. The effective pore formation ability and the high stability in the membrane of βMPs grant them large potential in applications of bionanotechnology. βMPs have drawn increasing attention in their promising application, including protein profiling, DNA sequencing, and small molecule detection. In order to investigate the roles of βMPs in biological and pathological processes and to develope engineering or designing methods of βMPs for biotechnical applications, it is critical for us to understand structural and thermodynamical properties of βMPs.

Despite the important roles of βMPs in cellualr process, limited availability of βMP structures hinders understanding of their structural properties and structure–function relationship. It was estimated that there exist hundreds of βMPs in each Gram–negative bacterium genome, while there are only around 60 non–homologous structures deposited in the Protein Data Bank when this study was conducted. This limitation is due to the great difficulty of experimental determination of TM protein structures because of their amphipathic nature. It is therefore important to develop accurate and efficient computational structure prediction methods for these proteins. We have developed a method to predict the 3D structures of βMPs. We predict strand registers and construct 3D structures of TM domains of βMPs accurately, including proteins for which no prediction has been attempted before. Our method also accurately predicts structures from protein families with a limited number of sequences and proteins with novel folds. An average mainchain RMSD of 3.48Å is achieved between predicted and experimentally resolved structures of TM domains, which is a significant improvement (>3Å) over a recent study. For βMPs with NMR structures, the deviation between predictions and experimentally solved structures is similar to the difference among the NMR structures, indicating excellent prediction accuracy. Moreover, we can now accurately model the extended β-barrels and loops in non-TM domains, increasing the overall coverage of structure prediction by $> 30\%$.

In additional to structural properties, it is also important to characterize thermodynamical properties of βMPs, which is important to understand their folding and stability, and may help in understanding the structure–function relationship. One major contributing factor to thermodynamic stability of membrane proteins is free energy of transferring amino acid sidechains from aqueous environment into lipid bilayers, known as transfer free energy (TFE). However, experimental measurement of TFEs of βMPs is challenging. A recent computational method has been developed to calculate TFEs, the results of which are in excellent agreement with experimentally measured values. However, the application of the method is limited to only small βMPs due to its computational complexity. We have improved this method and developed an

SUMMARY (Continued)

approximation method, which is comparably accurate but much faster than the original method. The new method enables the systematical calculation of TFEs of all βMP regardless of the size of the proteins. Based on the TFEs calculated from a representative set of βMPs, we further derived a TFE profile named General Transfer Free Energy Profile (GeTFEP). The GeTFEP agrees well with experimentally measured and computationally derived TFEs. Analysis based on the GeTFEP shows that residues in different regions of the TM segments of βMPs have different roles during the membrane insertion process. Results further reveal the importance of the sequence pattern of TM strands in stabilizing βMPs in the membrane environment. In addition, we show that GeTFEP can be used to predict the positioning and the orientation of βMPs in the membrane. We also show that GeTFEP can be used to identify structurally or functionally important amino acid residue sites of βMPs. Furthermore, the TM segments of α–helical membrane proteins can be accurately predicted with GeTFEP, suggesting that the GeTFEP captures fundamental thermodynamic properties of amino acid residues inside membrane, and is of general applicability in studying membrane protein.

The methods reported in this thesis require only sequence information, implying their general applications to genome -wide studies. The structure prediction and the TFE characterization methods provide ways to investigate properties of novel βMPs without conducting expensive wet lab experiments. They will also be useful in bionanotechnologies such as engineering existing βMPs and in design novel ones.

CHAPTER 1

INTRODUCTION

1.1 Overview

β–barrel membrane proteins (βMPs) are one major type of transmembrane (TM) proteins. They are solely found in the out membranes of Gram–negative bacteria, mitochondria, and chloroplast, so they are also known as outer membrane proteins (OMPs). βMPs serve a multitude of essential cellular functions, including reaction catalysis, protein anchoring, metabolite transportation, and outer–membrane biogenesis (9; 10; 11; 12; 13). In bacteria, βMPs are also found to be responsible for the release of virulence factors (14) and are implicated in multidrug resistance (15). Dysfunctional βMPs in mitochondria are also related to neurodegenerative diseases (16; 17). The ability for effective pore formation and the high stability in the membrane of βMPs have drawn increasing attention to βMPs due to their promise in a number of applications in bionanotechnology, including protein profiling (18), DNA sequencing (19), and small molecule detection (20).

In order to investigate the roles of βMPs in biological and pathological processes and to engineer or to design novel βMPs for biotechnical applications, it is critical for us to understand structural and thermodynamical properties of βMPs. Experimental approaches are usually costly and time consuming. It is therefore important to develop accurate and efficient computational methods to study these properties, which is the main focus of this thesis. There are four chapters in this thesis. Chapter 1 is a brief overview of β barrel membrane proteins, their structures and thermodynamical properties. Chapter 2 studies the structural properties of βMPs, where we developed a accurate 3D structure prediction method for βMPs. Chapter 3 focuses on the transfer free energies (TFEs), an important thermodynamical properties, of βMPs, where we improved a computational method for TFE calculation. In Chapter 4, we derived a general TFE profile for membrane proteins, and used the thermodynamical stability of βMPs inside the membrane.

1.2 Structures of β–barrel membrane proteins

The most prominent component of a βMP is its TM domain, formed by the β–strands that is assembled in the antiparallel arrangement forming a barrel–like shape (Figure 1A). Between adjacent β–strands, extensive hydrogen bonds (H–bonds) are formed among residues, providing the strong thermodynamical stability of the proteins (21). The barrel–like structures divide the TM space into two sides: the interior or the lumen– side, and the exterior or the lipid– side (Figure 1B). This distinction gives rise to the amphipathicity of the TM domains of βMPs. The residues on β–strands with lumen–facing sidechains are usually hydrophilic, while those with lipid–facing sidechains are usually hydrophobic since they are in the hydrophobic lipid environment (22).

βMPs vary in their sizes. The number of β–strands forming the barrel domains was thought to range from 8 to 26 when I was conducting the research of this thesis. With the development of the cryogenic electron microscopy (Cryo–EM) technology, now we have seen much more complex βMPs such as the Salmonella SPI-1 type III secretion injectisome secretin (InvG,

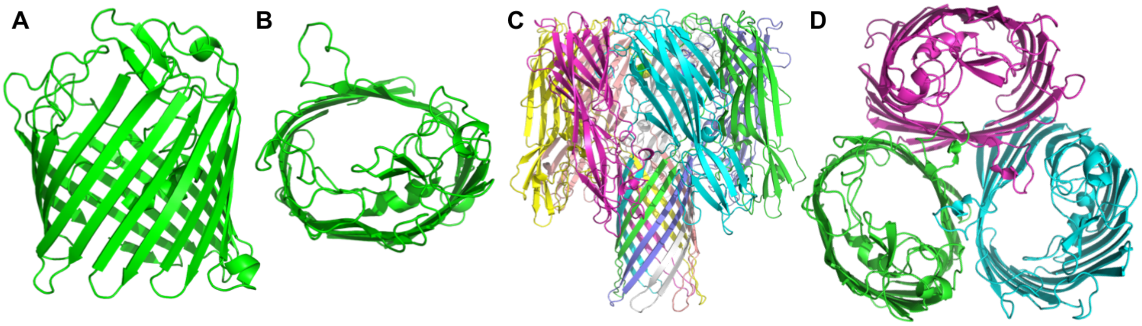


Figure 1: **A.** OmpF (PDB code: 2omf) has a barrel–like structure formed by β–sheet. **B.** The top view of OmpF shows that its TM domain devides the space into lumen–side and lipid–side. **C.** The barrel domain of α–hemolysin (PDB code: 7ahl) is formed by repeated β–strand hairpin. **D.** OmpF forms the functional unit in the membrane in a homotrimeric state.

PDB code: 6dv3) formed by 45 TM β–strands (23) and the Gasdermin A3 membrane pore (PDB code: 6cb8) formed by 108 TM β–strands (24). Concordant with the antiparallel nature of the β–sheets that formed the barrels, all βMPs have even numbers of β–strands, except the voltage–dependent anion channel (VDAC, PDB code: 3emn) (25), which has 19 TM β–strands. The first and the last strands of VDAC contact with each other in a parallel arrangement, while the other adjacent strands of VDAC remain the antiparallel arrangement.

The barrel domain of a βMP is usually formed by a single amino acid chain. For example, the barrel domain of the Outer membrane protein F (OmpF, PDB code: 2omf) (26) is formed by a single amino acid chain consisting of 16 β–strands (Figure 1A). There are also several multichain–barrels, formed by repeated subunits of β–sheets as in the case of efflux pump component TolC (PDB code: 1ek9) (27), or by repeated β–strand hairpins as in the case of bacterial toxin α–hemolysin (PDB code: 7ahl, Figure 1C) (12). These multichain βMPs are excreted as monomers, and assembled into multichain single β–barrels inside the membrane. Some single–chain βMPs also dwell in oligomeric status by forming homodimers or homotrimers (28; 26). The outer membrane phospholipase A (OmpLA, PDB code: 1qd6), an enzyme that catalyzes the phospholipids hydrolysis, has a homodimeric structure (28). Its enzymatic activity is regulated by the reversible dimerization (28). OmpF porins form homotrimeric biological units inside bacterial outer membranes (PDB code: 2omf, Figure 1D) (26).

The β–strands of the barrel domains are connected by long loops on the extracellular side, and by short turns on the periplasmic side of the outer membrane (29). Loops are the most flexible regions of βMPs, and are important for their functions (30). Nuclear magnetic resonance (NMR) structures of βMPs show that these loops adopt multiple conformations (31; 32), which likely contribute to the challenges in predicting binding affinity of βMP–ligand interactions (33). In addition to the β–barrel domains and the loops, these proteins often have in–plug and out–clamp domains for structural and functional purposes (34).

1.3 Thermodynamical properties of β–barrel membrane proteins

Thermodynamical studies utilize heating (35) and denaturants (1) to induce folding/unfolding processes to investigate factors stabilizing membrane proteins (36). One important factor revealed by such studies on α–helical peptides is backbone H–bonds among mainchain residues, which greatly reduce the energy cost of their insertion into membranes (36). There are two different types of H–bonds among adjacent β–strands in βMPs: one is C=O$\cdots$H–N called *strong hydrogen bonds*, and the other is C=)$\cdots$H–C called *weak hydrogen bonds* (Figure 2A).

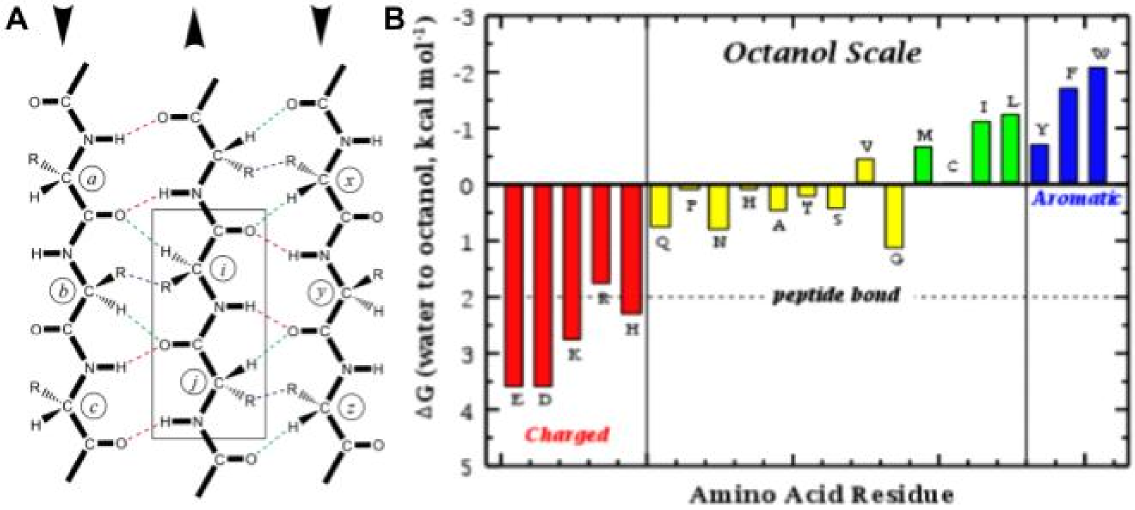


Figure 2: **A.** Interstrand interactions between β–strands of βMPs. Red dashed lines indicate the strong H–bonds, green the weak H–bonds, and blue the sidechain interactions. Figure adapted from Ref (3). **B.** An example hydrophobicity scale measured in Ref(4). Figure adapted from Ref (4).

Extensive H–bonds of these two types are formed among residues between adjacent β–strands, providing the strong thermodynamical stability of the proteins (21). Indeed in a thermal denaturation test, human mitochondrial protein–conducting channel protein hTom40A had an apparent melting temperature of 73°C (37).

A combinatorial analysis of mainchain H–bonds showed that there are characteristic preferences of different types of residue pairing even for the same type (strong or weak) H–bonds (3), implying that contribution of mainchain H–bonded pairing between different types of residues to the thermodynamical stability of the βMP could be different. Thus, different regions in a βMP may have different stability depending on the local amino acid composition. Following such rationale, a computational approach was developed to identify weakly stable regions of βMP barrel domains (34). It shows that these regions are usually stabilized by interactions from oligomerization or by interactions with other in–plug or out–clamp domains (34).

In addition to H–bonds, energy cost of transferring residue sidechains from water environment to the membrane environment is another important factor stabilizing βMPs (38; 36). Experiments have measured these energies using different host systems, such as peptides (4; 39) and proteins (1). Since these TFEs quantify relative hydrophobicity of amino acids, they are also called hydrophobicity scales (Figure 2B),which has generated considerably insights to our understanding of membrane proteins, such as identification of membrane proteins and their topology, interpretation of folding process, and prediction of structurally or functionally important sites (40; 41; 8; 42; 43).

CHAPTER 2

HIGH RESOLUTION STRUCTURE PREDICTION OF β-BARREL MEMBRANE PROTEINS

Adapted from Tian, W., Lin, M., Tang, K., Liang, J., and Naveed, H.: High–resolution structure prediction of β–barrel membrane proteins. Proceedings of the National Academy of Sciences, 115(7):1511–1516, 2018.

2.1 Introduction

A major obstacle in studies of βMPs is the limited availability of structural data. Only $\sim$ 320 βMP structures, of which $\sim$ 59 are nonhomologous, have been deposited in the Protein Data Bank (PDB) that contains $> 135,000$ protein structures (44). Computational studies have contributed to expand our knowledge of βMPs by successfully predicting βMP sequences at genome–wide scale (45; 46), identifying TM segments (47; 48), and uncovering sequence and spatial motifs (22; 49). The stability, oligomerization state, protein–protein interaction interfaces and the transfer free energy of residues in the TM regions of βMPs can also be accurately computed (34; 50; 51; 52; 42; 53; 43).

Template–based methods for structure prediction have been successfully applied in studies of globular proteins (54). They have also been employed to predict 3D structures of βMPs but have achieved limited success with novel folds due to the limited availability of templates for βMPs (55). Recently solved structures of the voltage–dependent anion channel (VDAC) found in mitochondria (25), the usher protein PapC from *P pilus* (56), and LPS-assembly protein LptD from *S flexneri* (57) contain 19, 24 and 26 TM β-strands, respectively. These structures are incompatible with the generally observed βMP topologies that consist of even TM strands, with the number ranging between 8 and 22 (29). Template based methods perform poorly on these novel structures, as homologous template structures do not exist for these proteins. Second, when homologous templates can be found for βMPs, the template protein sequences need to have exactly the same number of TM β-strands, so the radius of the barrel, which depends on the number of TM β-strands can be modeled accurately (58; 59). As an example, BamA with 16 TM segments (60) is a homologue of Toc75 protein, which is predicted to contain 18 segments (61). Therefore using the 3D structure of BamA as a template to model Toc75 protein will likely results in low quality predictions. General purpose template–free structure prediction methods do not generate accurate structures of βMPs, as these proteins can be large, with the number of residues reaching 800.

A recently published βMP specific method that combines sequence covariation for contact prediction with a machine learning based method achieved limited progress, with a mainchain RMSD of 6.66Å for predicted structures of TM regions, before it was adjusted to a better published value of 4.45Å when only a subset of residues were aligned instead of all TM residues (62). Another template–free βMP specific method, 3D-SPoT, can predict the TM regions of βMPs with an average mainchain RMSD of 4.14Å (61). Despite such progress, further improvement in prediction methods to generate accurate structural models is required to bridge the gap between identified βMP sequences and resolved βMP structures, so that modeled structures can be used directly for applications such as nanopore engineering and drug design/delivery.

In this study, we describe a template–free method for predicting 3D structures of βMPs, which provides significant improvement over previous methods. Our approach, named 3D–BMPP (3D Beta–barrel Membrane Protein Predictor), is based on a statistical mechanical model (63) that incorporates sequence covariation information, and is built upon a parametric structural model of intertwined zigzag coils. In a blind test of 51 nonhomologous βMPs, our prediction generates accurate 3D structures of TM regions with an average mainchain RMSD of 3.48Å. This represents a significant improvement of $\sim$ 3.1Å compared to a recent study (62) over a much bigger dataset (51 vs 17 proteins). In addition, predictions are expanded to include non-TM regions, including both extended β-sheets and loops, resulting in significant increase in the coverage of residues compared to previous methods. Furthermore, our method can be applied to model structures of βMPs with novel folds, including those from mitochondria of eukaryotes, as evidenced by the accurately modeled structures of VDAC, and FimD. Our method is general and can be applied to genome–wide structural prediction of βMPs.

2.2 Methods and Materials

2.2.1 Dataset

We use 59 non–homologous βMPs (resolution 1.45Å– 3.2Å) with less than 30% pairwise sequence identity for this study. The PDB codes of these proteins are 1a0s, 1bxw, 1e54, 1ek9, 1fep, 1i78, 1k24, 1kmo, 1nqe, 1p4t, 1prn, 1qd6, 1qj8, 1t16, 1thq, 1tly, 1uyn, 1xkw, 1yc9, 2erv, 2f1c, 2f1t, 2fcp, 2gr8, 2lhf, 2lme, 2mlh, 2mpr, 2o4v, 2omf, 2por, 2qdz, 2vqi, 2wjr, 2ynk, 3aeh, 3b07, 3bs0, 3csl, 3dwo, 3dzm, 3emn, 3fid, 3kvn, 3o44, 3pik, 3rbh, 3rfz, 3syb, 3szv, 3v8x, 3vzt, 4c00, 4e1s, 4gey, 4n75, 4pr7, 4q35, and 7ahl.

We also use NMR structures to character the intrinsic flexibility of βMPs and to evaluate our loop modeling. The PDB codes of the NMR structures used are 1g90 and 1ge4 for OmpA, 1orm, 1q9f, 1q9g and 2mnh for OmpX, 1mm4 and 1mm5 for PagP, 2jqy for OmpG, 2mhl for OmpW, 2lhf for OprH, and 2maf and 2mlh for Opa60.

2.2.2 Workflow of 3D–BMPP

βMPs have strong thermal and chemical resistance due to the well–knit H–bond network (37), in which each residue in the TM strand is H–bonded to residues on the adjacent TM strands (Figure 4). We use a physical model that accounts for strong H–bonds, weak H–bonds, and sidechain interactions between adjacent strands in the barrel domain (64; 63; 65; 34). In addition, we incorporate interstrand loop entropy, right handedness of the βMP, and medium–to–long range contacts predicted from sequence covariation information.

To predict structures of βMPs, we proceed in four steps: predicting strand registers (interstrand H–bond contacts) locally, optimizing strand registers globally predicting 3D coordinates of residues in barrel domains, and modeling non–barrel–domain residues (Figure 3).

2.2.3 Secondary structure determination

Existing computational approaches can successfully identify the location of β-strands (47; 48). However, to assess our 3D modeling approach without any short coming from the secondary structure prediction we use the β-strands from the DSSP program that uses PDB structure to calculate the location of the β-strands (66). Only l_{cut} number of resides from the periplasmic

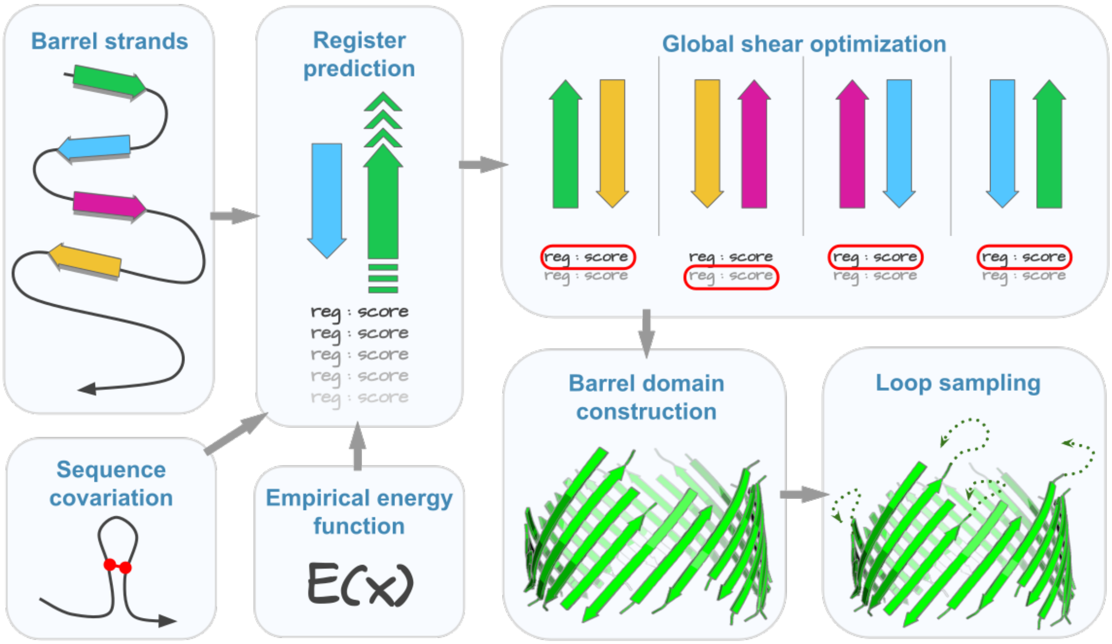


Figure 3: The flowchart of βMP structure prediction method 3D–BMPP. The strand registers are predicted using a combination of empirical energy function and sequence covariation information. Global shear optimization is then performed upon the predicted register candidates. The 3D coordinates of C_α atoms of TM and non-TM residues are then predicted using a parametric structural model. We also predict ensembles of loop conformations.

side of each DSSP strand are used for register prediction. We choose $l_{cut} = 12$ since the length of strands in the dataset has a mean of 12.7, mode of 12, and median of 12. For 3D structure construction, complete DSSP strands are used.

2.2.4 Strand Register Prediction

We use a discrete model of reduced states to represent the conformational space of the strands, in which the relative position between a pair of adjacent strands can adopt $L_1 + L_2 - 1$ different registers, where L_1 and L_2 are the lengths of the two strands (Figure 4) (34). In a strand pair, we fix one strand while sliding the other strand up and down to generate all possible conformations in the discrete state space, each of which has different interstrand residue contacts and thus the register (Figure 3).

To predict the register of a stand pair, we have developed a model incorporating both the empirical potential scores of physical interactions between strands from our previous study(61) and the sequence covariation information that can identify medium–to–large range residue contacts based on the concept that spatially close residues might coevolve. Our model gives a score for each register with Equation 2.1

$$E(r) = E_{emp}(r) + E_{sc}(r), \tag{2.1}$$

where r is a given register of the strand pair, $E_{emp}(r)$ is the empirical potential score of physical interstrand interactions, and E_{sc} is the score from sequence covariation analysis. The register with the lowest score is selected as the prediction.

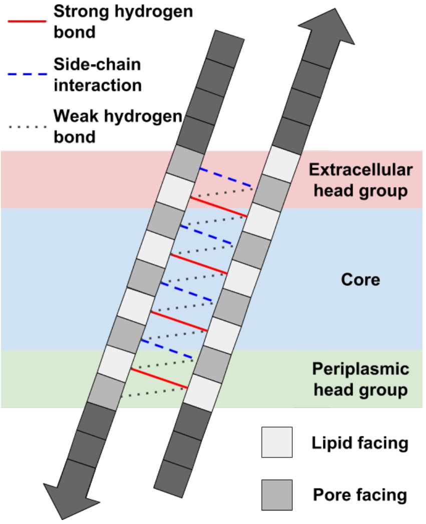


Figure 4: Model for interstrand interactions between adjacent strands.

2.2.4.1 Model for interstrand interactions

The model for physical interactions between a strand pair from Ref(61) is used in this study. Briefly, the model assumes that neighboring strands interact through canonical strong H–bonds, weak H–bonds, and non-H–bonds (sidechain interactions), which is based on the observation of the periodic dyad bonding repeat pattern of antiparallel β-sheets (63) (Figure 4). The entropy for unbonded regions and left/right handedness of the strand pair are considered as well. See

Ref(61) for more detailed description of the model. The total empirical score of certain register r of a given strand pair is calculated with the empirical scoring function

$$\begin{aligned} E_{emp}(r) =&\alpha \sum_{k_i} \sum_{k_{i+1}} E_{SH}(k_i, k_{i+1}; r) + \beta \sum_{k_i} \sum_{k_{i+1}} E_{WH}(k_i, k_{i+1}; r) \\ &\gamma \sum_{k_i} \sum_{k_{i+1}} E_{NH}(k_i, k_{i+1}; r) + \delta \ln(\frac{n_{ref} + \Delta L(r)}{n_{ref}}) + \varepsilon[LH(r)], \end{aligned} \tag{2.2}$$

where $E_{SH}(k_i, k_{i+1}; r)$, $E_{WH}(k_i, k_{i+1}; r)$, and $E_{NH}(k_i, k_{i+1}; r)$ are the empirical energies of strong, weak, and non-H–bonds between the residue k_i on strand i and the residue k_{i+1} on strand $i + 1$, respectively. $n_{ref} = 8.5$ is the average length of loops. $\Delta L(r)$ is related to the number of residues that do not share a H–bond with the adjacent strand in the register r, minus the difference in strand lengths. $LH(r)$ is the penalty for left handed twist ($r < 0$) since all β-sheets are right handed.

$$LH(r) = \begin{cases} r & r < 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

2.2.4.2 Model for sequence covariation

We use PSICOV(67) to calculate the sequence covariation scores of each residue pairs in TM regions. The score of certain register of a strand pair is calculated as the weighted summation of sequence covariation scores of residue pairs:

$$\begin{aligned} E_{sc} =&w_0 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 0) Q(k_i, k_{i+1}) + w_1 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 1) Q(k_i, k_{i+1}) \\ &+ w_2 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 2) Q(k_i, k_{i+1}) \end{aligned} \tag{2.4}$$

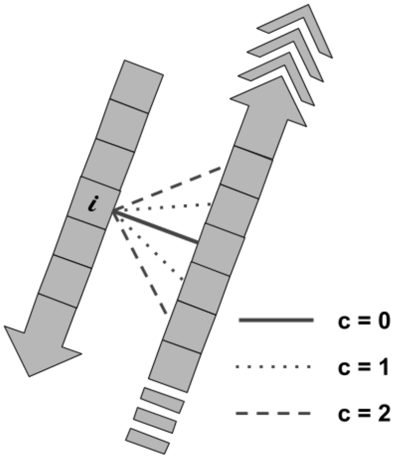


Figure 5: Model for calculating sequence covariation between adjacent strands.

where $Q(i, j)$ is the sequence covariation score of the residues k_i and k_{i+1}, $d_{k_i,k_{i+1}}$ is the distance between the two residues in the discretized conformational state space (Figure 5), w_c ($c = 0,\ 1,\ \text{or}\ 2$) is the weight of residue pair whose distance is c, and $\delta(d_{k_i,k_{i+1}}, c)$ is the Kronecker delta function which identifies if the distance of the residues k_i and k_{i+1} is c. All residue pairs with distance larger than 2 are ignored in the calculation, for they are unlikely to have any physical interaction.

2.2.4.3 Parameter determination and cross–validation

Based on the number of strands (or equivalently, number of residues) and the stability of the proteins(34), we divide the dataset into five subsets (Table I). All 59 βMPs are used to construct the empirical potential function, but predictions are only made for 51 proteins,

TABLE I: The groups of βMPs in this study. All the six groups are used in the construction of the empirical energy function. Structure predictions are made for only the first five groups.

Group	Description	PDB code
1	Small βMPs ($N < 16$) without in–plugs or out–clamps	1bxw, 1qj8, 1p4t, 2f1t, 1thq, 2erv, 2lhf, 2mlh, 3dzm, 1qd6, 2f1c, 1k24, 1i78, 2wjr, 4pr7
2	Small βMPs ($N < 16$) with in–plugs or out–clamps	1t16, 1uyn, 1tly, 3aeh, 3bs0, 3dwo, 3fid, 3kvn, 4e1s
3	Medium oligomeric βMPs ($16 \leq N < 20$)	2mpr, 1a0s, 2omf, 2por, 1prn, 1e54, 2o4v, 3vzt, 4n75
4	Medium monomeric βMPs ($16 \leq N < 20$)	2qdz, 2ynk, 3emn, 3rbh, 3syb, 3szv, 4c00, 4gey
5	Large βMPs ($N \geq 20$)	1fep, 2fcp, 1kmo, 1nqe, 1xkw, 2vqi, 3csl, 3rfz, 3v8x, 4q35
6	Multichain βMPs	1ek9, 1yc9, 2gr8, 2lme, 3pik, 3b07, 3o44, 7ahl

after excluding multichain–barrel βMPs to avoid over estimation of repeated interaction types (Table I).

We first fix the weights (w_0, w_1, and w_2) in the sequence covariation model. Since the sequence covariation analysis comes purely from sequences and needs no prior knowledge of the dataset, we neither use the leave–one–out scheme for the searching of these three weights, nor discriminate the groups of the dataset. The weights (w_0, w_1, and w_2) are determined by searching for the values such that the score E_{sc} alone can give a best prediction of the registers of the neighboring strand pairs in the dataset.

TABLE II: Values for α, β, γ, δ, ε, w_0, w_1, and w_2.

Group	α	β	γ	δ	ε	w_0	w_1	w_2
1	0.026	0.038	0.036	0.245	0.050	-0.500	-0.136	-0.364
2	0.055	0.100	0.075	0.450	0.120			
3	0.000	0.082	0.006	0.052	0.074			
4	0.045	0.020	0.024	0.290	0.100			
5	0.045	0.024	0.014	0.110	0.135			

Then the leave–one–out cross–validation (LOOCV) is used for searching the other undetermined weights (α, β, γ, δ, and ε) in Equation 2.2 so that the total scores calculated via Equation 2.1 give the best prediction. In LOOCV, we left one protein out of the data set while using the other proteins to construct the empirical potential function. The registers of the leave–out protein were predicted. This process was repeated for each protein to find the optimized values of the group–specific parameters (α, β, γ, δ, and ε), which gave the best register prediction accuracy for that group. The parameters (α, β, γ, δ, and ε) were optimized using an adaptive grid search. The final values used in the model is listed in Table II.

2.2.4.4 Sidechain direction prediction

The sidechain of a strand residue can be either lipid–facing or lumen–facing. Ignoring β-bulge, sidechain directions of a strand follow an alternative lipid–facing–lumen–facing pattern Hence, we only predict the sidechain direction of the first residue on the periplasmic side of each strand, and sidechain direction of all the other residues can be obtained accordingly.

In the original reduced state space (RSS) model, there are 2, 5, and 2 residues in the extracellular headgroup, the core, and the periplasmic headgroup regions, respectively (Figure 4). However, it is known that membrane could become either thinner or thicker around TM proteins adaptively. So, we use a variant of RSS where the number of the resides in each of these three regions can vary by 1 from the original RSS while the total thickness of these three regions is restricted to 7-11 residues.

We enumerated the combination of the sidechain directions and the legit conformations in the RSS variant aforementioned, and used the single body potential (34) derived from our βMP dataset to calculate the energy of each combination. The sidechain directions that give the lowest energy within the enumeration of each strand are selected as predictions, which gives a 98% accuracy.

2.2.4.5 Global register optimization

Strand register prediction considers H–bonds contact two adjacent strands at a time. However, global H–bond pattern is better represented by the shear number of the protein.

The shear number is the displacement of the relative positions in the TM strands if one starts to follow the backbone H–bond between strands, beginning from strand 1 and returning after a full circle to the same strand (see Figure 6, more examples can be found in Ref(68)). The shear number of a βMP also equals to the sum of the strand registers. When these registers are not known, the shear number can be estimated reliably from the number of TM strands (61) as the most common shear number S_{com} of the βMPs of the same strand number (Table III).

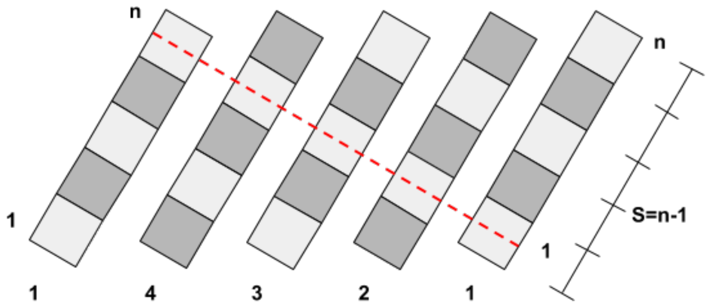


Figure 6: Shear number is the displacement of the relative positions in the TM strands if one starts to follow the backbone H–bond between strands, starting from strand 1 and returning after a full circle to the same strand. For example, the shear number for the 4 strand β-barrel shown above is $n-1$.

The predicted shear number S of a βMP can be calculated as the sum of the predicted strand registers:

$$S = \sum_{i=1}^{N} r_i, , \tag{2.5}$$

where r_i is the predicted strand register of the i-th strand. N is the total number of strands.

We optimize strand register prediction so that the predicted shear number S can be as close as possible to the most common shear number S_{com}. For each strand, two register candidates with lowest scores in the register prediction step are kept. The summation of the first register candidate of each strand gives the predicted shear number before optimization. This selection also gives the total score for the predicted protein conformation by summing up the score of each predicted register. The global shear optimization attempts to replace the first candidate with the second one of each strand in the final selection so that the predicted shear number is

TABLE III: Shear number of βMPs.

PDB code	N	S	PDB code	N	S	PDB code	N	S	PDB code	N	S
1qj8	8	8	3aeh	12	14	2o4v	16	20	3szv	18	22
1bxw	8	10	3fid	12	14	2omf	16	20	3emn	19	20
1p4t	8	10	3kvn	12	14	2por	16	20	1fep	22	24
1thq	8	10	4e1s	12	14	2qdz	16	20	1kmo	22	24
2erv	8	10	4pr7	12	14	3vzt	16	20	1nqe	22	24
2f1t	8	10	1qd6	12	16	4c00	16	20	1xkw	22	24
2lhf	8	10	1tly	12	16	4gey	16	20	2fcp	22	24
2mlh	8	10	1t16	14	14	4n75	16	22	3csl	22	24
3dzm	8	10	3bs0	14	14	2ynk	18	20	3v8x	22	24
1i78	10	12	3dwo	14	14	1a0s	18	20	2vqi	24	26
1k24	10	12	2f1c	14	16	2mpr	18	22	3rfz	24	26
1uyn	12	14	1e54	16	20	3rbh	18	22	4q35	26	30
2wjr	12	14	1prn	16	20	3syb	18	22			

as close to the target shear as possible while keeping the total score for the protein as close to the minimum score as possible.

The register candidates are first filtered according to the predicted sidechain directions of the first periplasmic residues of that strand and of its sequential neighbor: If the first residues of the i–th strand and of the $(i+1)$-th strand have the same sidechain direction, only the candidate(s) of the i–th strand with even register number is kept; otherwise, the odd one(s). This criteria is based on the fact that H–bonded residues on adjacent strands always have the same sidechain direction. When neither of the candidates satisfy this criteria, both are kept.

Subsequently, the strands are sorted in ascending order according to the difference between the scores of the two candidates of each strand. The difference is 0 if only one candidate was kept in the previous step. We scan the strands in this order and make the final selection for each strand. For the top two strands, the second candidate will be selected if it can bring the predicted shear number S closer to the target S_{com}. For the remaining strands, the second candidate will be selected only when it can keep the predicted shear number S in same parity with the target shear number $S_{com}$and can also reduce the shear number difference $|S - S_{com}|$ between prediction and target.

After optimization, the error in predicted shear numbers is decreased from -0.69 ± 3.63 to 0.12 ± 1.34. The improved global shear accuracy will lead to overall more accurate 3D structure prediction of βMPs.

2.2.5 Parametric model for 3D structures of barrel domains.

Parametric models have had recent successes in modeling and designing structures of α-helical proteins (69; 70). In this work, we have developed a novel parametric structural model, named *intertwined zigzag coil model* to generate 3D structures of βMPs from predicted strand registers (Figure 8). Following previous studies (61; 71), we model the overall shape of the β-barrel as a ideal cylinder. The C_α trace of each strand is described as a coiled zigzag wrapping around the hypothetical cylinder. This model captures the zigzag nature of a polypeptide in the βMP and the varied distance between C_α atoms on adjacent strands (Figure 7), which improves positioning of C_α atoms.

2.2.5.1 C_α trace construction

An intertwined coil model was used in one previous study (61), in which the C_α trace of a βMP was generated, followed by backbone generation, sidechain generation, and molecular dynamics (MD) minimization. If we look closely at the C_α trace of a βMP structure, however, we find that the intertwined coil model is not able to capture the following geometric properties: 1) the C_α trace of a strand is not as smooth as a coil, but is zigzag–like (Figure 7a). The C_α atoms of lipid–facing residues are farther away from the vertical axis of the barrel compared to those of lumen–facing residues; and 2) the C_α atoms on two adjacent strands are not equidistant as depicted by the intertwined coil model. The distance between of C_α atoms of residues sharing strong H–bonds are larger than those sharing non-H–bonds (Figure 7b and Figure 7c).

To capture these geometric properties, we developed a parametric structural model of intertwined zigzag coils, in which the C_α trace of each strand is depicted by a zigzag coil that wraps around a hypothetical cylinder. To calculate the C_α position of a strand, we first build a coil basis for the strand (Figure 8a).

The tilt angle θ of coil basis with respect to the vertical cylinder axis and the radius r of the cylinder are calculated using Equation 2.6 following McLachlan(71):

$$\begin{aligned} \theta &= \arctan(\frac{SA}{NB}), \\ r &= \frac{B}{2\sin(\frac{\pi}{N})\cos\theta}, \end{aligned} \tag{2.6}$$

where A is the distance between projections of consecutive C_α atoms on the same coil basis, and B is the distance between projections of C_α atoms sharing a strong or non–H–bond on adjacent strands. Note that A and B here are not the intra- and the inter–strand C_α distances. N is the number of β-strands, and S is the shear number for the βMP.

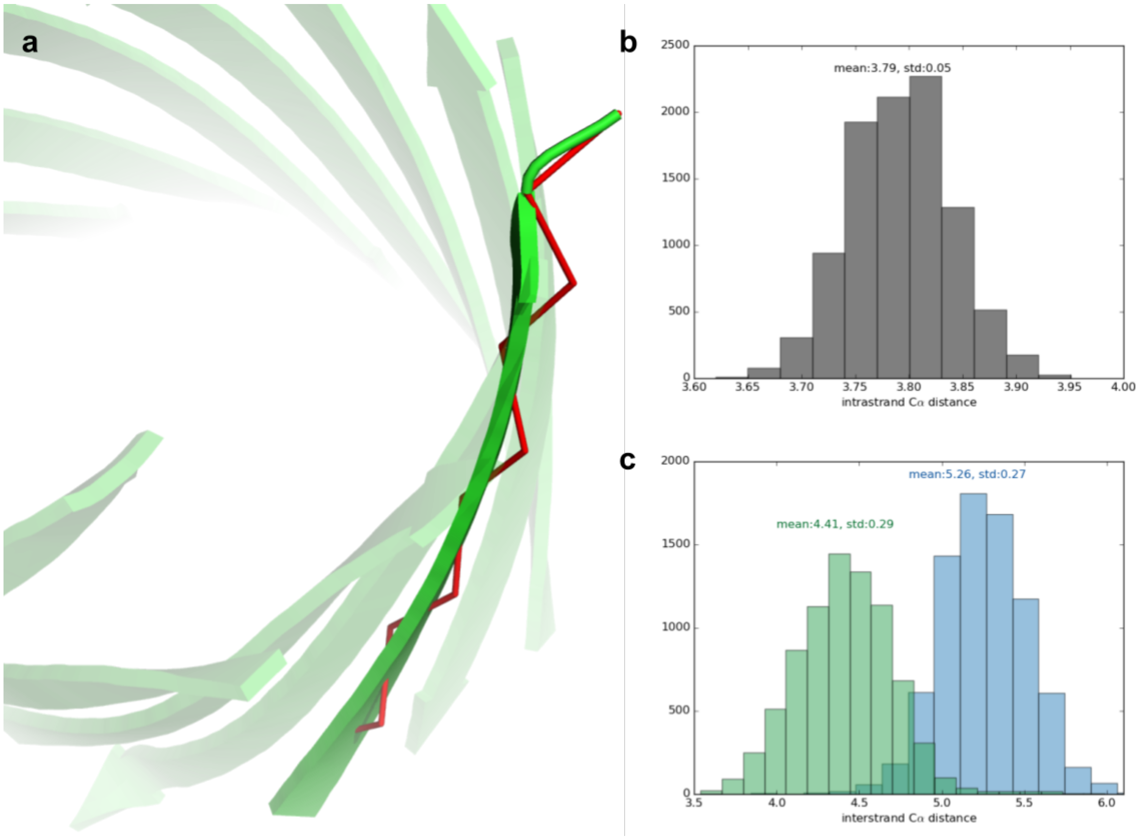


Figure 7: Geometric properties of βMPs. (a) The C_α trace of a β-strand shows a zigzag pattern (red). The structure used here is TodX (PDB code:3bs0). (b) Distribution of distance between consecutive C_α atoms on the same strand. (c) Distribution of distance between C_α atoms of residues sharing a non–H–bond (green) and of those sharing a strong H–bond (blue) on adjacent strands.

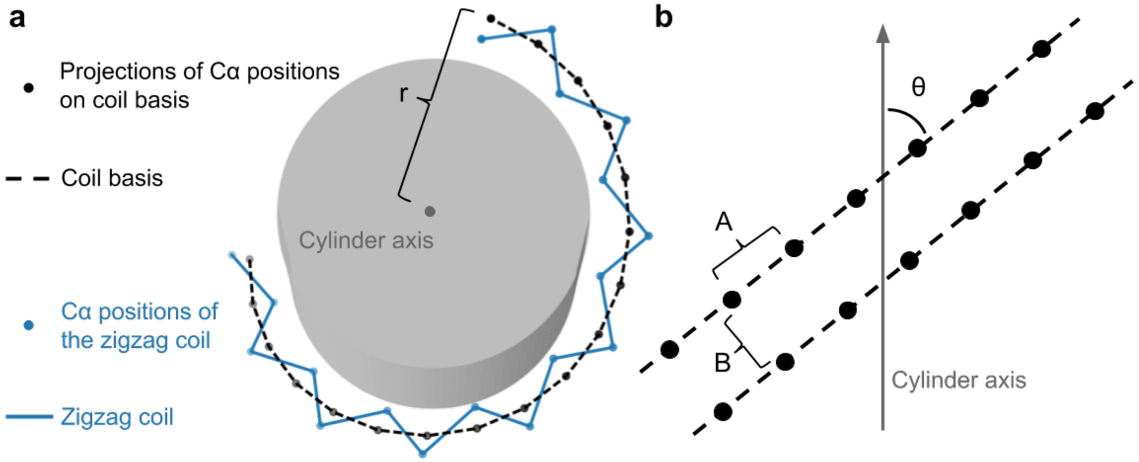

a
Projections of Cα positions on coil basis
Coil basis
Cα positions of the zigzag coil
Zigzag coil
r
Cylinder axis
b
θ
A
B
Cylinder axis

Figure 8: The parametric structural model of intertwined zigzag coils. (a) One zigzag coil (blue) and the corresponding coil basis (black) wrap around the hypothetical cylinder (grey). (b) The relative position and the corresponding parameters of coil bases are shown after unwrapping the coil bases onto a plane.

Using time curves from differential geometry (72), each position j of C$_\alpha$ projection on coil basis i is represented by a parametric curve represented by Equation 2.7.

$$\begin{aligned} c_i(t_{ij}) &= \left(r\cos(t_{ij} - \frac{2\pi i}{N}),\ r\sin(t_{ij} - \frac{2\pi i}{N}),\ bt_{ij} \right), \\ b &= \frac{r}{\tan\theta}, \end{aligned} \tag{2.7}$$

where $c_i(\cdot)$ is the parametric curve of the i-th coil basis. Let $V_r(t_{ij})$ be the vector from position j of coil basis i to position j of coil basis $i+1$, and $T_r(t_{ij})$ the tangent vector at position j of coil basis i. Given that the vector between two C$_\alpha$ atoms sharing a strong or non–H–bond on adjacent strands is roughly perpendicular to the strands, the inner product of the two vectors should be 0:

$$\begin{aligned}
&V_r(t_{ij}) \cdot T_r(t_{ij}) = 0, \\
&\qquad V_r(t_{ij}) = c_{i+1}(t_{i+1,j}) - c_i(t_{ij}), \\
&\qquad T_r(t_{ij}) = \left(-\frac{r}{c}\sin(t_{ij}),\ \frac{r}{c}\cos(t_{ij}),\ \frac{b}{c}\right), \\
&\qquad\qquad c = \sqrt{r^2 + b^2}.
\end{aligned} \tag{2.8}$$

By solving Equation 2.8, t_{ij} can be written as

$$\begin{aligned}
t_{ij} &= \frac{s_{ij}}{c}, \\
s_{ij} &= (j - \sum_{k=1}^{i-1} R_k)A + i\frac{2\pi r^2}{cN},
\end{aligned} \tag{2.9}$$

where R_k is the register of the k-th strand. Here is where we can feed the register prediction results to the intertwined zigzag coil model

Using different radii for the lipid–facing and the lumen–facing residues, $\tilde{c}_i(t_{ij})$, the zigzag pattern of the C$_\alpha$ trace (Figure 8a) can be taken into account by Equation 2.10.

$$\begin{aligned}
\tilde{c}_i(t_{ij}) &= \left(r'\cos(t_{ij} - \frac{2\pi i}{N}),\ r'\sin(t_{ij} - \frac{2\pi i}{N}),\ bt_{ij}\right), \\
b &= \frac{r}{\tan\theta}, \\
r' &= \begin{cases} r + \Delta r, & \text{if the position is lipid–facing} \\ r - \Delta r, & \text{if the position is lumen–facing} \end{cases}.
\end{aligned} \tag{2.10}$$

Considering that the distance between C_α atoms of residues sharing a strong H–bond is different from the distance between those sharing a non–H–bond (Figure 7c), the 3D coordinates $\mathbb{C}_i(t_{ij})$ of C_α atoms in the intertwined coiled zigzag model can be written as Equation 2.11

$$\mathbb{C}_i(t_{ij}) = \begin{cases} \tilde{c}_i(t_{ij}) + \Delta w \frac{V_{r'}(t_{ij})}{\|V_{r'}(t_{ij})\|}, & \text{if } i \text{ is odd and the position is lipid–facing,} \\ & \text{or } i \text{ is even and the position is lumen–facing} \cdot \\ \tilde{c}_i(t_{ij}) - \Delta w \frac{V_{r'}(t_{i-1,j})}{\|V_{r'}(t_{i-1,j})\|}, & \text{otherwise} \end{cases} \tag{2.11}$$

2.2.5.2 Parameter estimation

The intrastrand C_α distance has very little variance from its mean value of 3.79 Å, while the interstrand C_α distance of residues sharing a strong or a non–H–bond have different means (5.26 Å and 4.41 Å, respectively) and relatively larger variances (Figure 7c). We used $B = 4.83$Å, which is the mean value of interstrand C_α distance of residues sharing a strong or a non–H–bond, and did a grid search for the values of A, Δd, and Δw that satisfy the following criteria:

1. Any value that makes the intrastrand C_α distance out of the range [3.79±0.02] is rejected,

2. The average interstrand C_α distances of residues sharing a strong H–bond and of residues sharing a non–H–bond are as close to 5.26Å and to 4.41Å as possible,

The best parameters we found are $A = 3.345$Å, $\Delta r = 0.83$Å, and $\Delta w = 0.22$Å. The intertwined zigzag coil model using these parameters give intrastrand C_α distances of 3.77±0.06, interstrand C_α distances of residues sharing a strong H–bond of 5.28 ± 0.19 and interstrand C_α distances of residues not sharing a strong H–bond of 4.43 ± 0.18, which agree well with the experimentally solved structures (Figure 7b and Figure 7c)

As for a βMP with N strands, an approximation for the shear number S is

$$\begin{cases} S &= N, \quad N = 14, \\ S &= N + 4, \quad N = 16 \text{ or } 18, \\ S &= N + 2, \quad \text{otherwise.} \end{cases} \tag{2.12}$$

which is correct for all βMP structures with the exception of OmpX, OmpLA, Tsx, OmpG, Wzi, LptD and VDAC in our data set.

2.2.5.3 Construction of backbones and sidechains

Based on the predicted C_α trace, we used Gront *et al.*'s BBQ algorithm (73) to construct the backbone of the barrel. This algorithm constructs a four–residue fragments using three internal coordinates, namely, the three distances between C_α atoms. Positions of C, O, and N atoms are then determined based on information from known PDB structures (73). As loop region is ignored in our model at this stage, the strands are disconnected with each other. Directly applying BBQ to these disconnected strands tends to make mistake at the ends of the strands. Therefore, we constructed two additional pseudo C_α atoms to both extracellular end and periplasmic end of each strand using the same formula (Equation 2.11) The backbone obtained from BBQ was then fed to the Scwrl4 program(74) for sidechain generation. The pseudo C_α atoms and the corresponding backbone and sidechain atoms were removed after that, which leaves the 3D structure of the barrel domain.

2.2.6 Predict structures of βMP loops

Loops are the most flexible regions of βMPs, and are important for their functions (30). NMR structures of βMPs show that these loops adopt multiple conformations (31; 32), which likely contribute to the challenges in predicting binding affinity of βMP–ligand interactions (33). We model loops by investigating a large ensemble of loop conformations generated using an improved version of the m-DiSGro algorithm (75) with better volume avoidance control that guarantees clash–free conformations of the sampled loops. For each of the 7 βMPs with available NMR structures, once the structure of the barrel domain is predicted, we sampled 3×10^4 to 3×10^5 multi–loop conformations, with the specific number of conformations dictated by the number and the lengths of loops. We then perform clustering to generate an ensemble of ~ 400 multi–loop conformations as prediction for each protein.

2.3 Results

2.3.1 Register prediction

The results of strand register prediction for 51 βMPs show that overall 655 out of 771 registers are predicted correctly, representing an accuracy of $\sim 85\%$ (see Table XVI for details). This is a significant improvement over previous βMP register prediction methods of Jackups and Liang ($\sim 46\%$) (63), Randall *et al.* ($\sim 48\%$) (55), Naveed *et al.* ($\sim 73\%$) (61) and Hayat *et al.* ($\sim 44\%$) (62). It is also important to note that the dataset used is much larger than those used in the previous studies (Table IV). For 8 βMPs (OpA60, autotransporter Hbp, TodX,

TABLE IV: Comparison of different methods for strand register and 3D structure prediction for TM regions of βMPs. 3D-BMPP can predict strand registers with an accuracy of ~85%, and 3D structures of TM regions with an average mainchain RMSD of 3.48Å and average all atom RMSD of 4.26Å for a much bigger dataset (51 vs 14 – 23 βMPs).

Method	Num of βMPs	Num of strands	Register accuracy	Avg mainchain TM-RMSD	Avg all atom TM-RMSD
Jackups and Liang, 2005 (63)	19	256	46%	—	—
TMBpro–server, 2008 (55)	14	214	48%	—	7.3Å
3D-SPoT, 2012 (61)	23	324	73%	4.12Å	5.6Å
EVfold_bb, 2015 (62)	17	265	44%	6.66Å	—
3D-BMPP, 2018 (76)	**51**	**771**	**85%**	**3.48Å**	**4.26Å**

EstA, FhuA, FecA, FptA, and HasR that contain 8, 12, 14, 12, 22, 22, 22, and 22 strands respectively), we are able to predict all the strand registers correctly.

To assess the contribution of the sequence covariation information and the patterns of hydrogen–bonds and sidechain interactions (HSC), we predicted the strand registers using sequence covariation data and a reduced state space (SC+RSS). The strand register prediction accuracy with SC+RSS was found to be 52%, representing significant deterioration from the accuracy of 69% (61) using HSC+RSS. This result indicates that patterns of H–bonds and sidechain interactions derived from structural data can predict local strand registers more accurately than sequence covariation information. This conclusion is consistent with that of Hayat *et al.*, in which machine learning and sequence covariation were used to predict the strand register at an accuracy of 44% (62), suggesting that the reduced state space model also contributes significantly to the improved the accuracy of strand register prediction.

The sidechain orientation of the TM residues is an important determinant of the structure of βMPs. A residue can be either lipid–facing or lumen–facing, with consecutive residues in the TM region taking alternating orientations. Pore–facing residues are predominantly responsible for protein function (e.g., flux control of metabolites and ion–sensing), while lipid–facing residues are mostly responsible for protein insertion and stability. Residues on adjacent strands have the same sidechain orientation when they share strong H–bonds or sidechain interactions. Incorrect strand register can lead to erroneous sidechain orientation prediction. The correct prediction of strand register is therefore an important requirement in structure prediction of βMPs and is well recognized in literature (55). Our method can predict strand register at 85% accuracy. In contrast, the criteria was relaxed to allow +1 or -1 difference in strand register in a previous study (62). While this relaxation made the register prediction results more presentable (65% after relaxation vs 44% before relaxation), it is problematic, as it would lead to prediction of TM residues to adopt erroneous orientation opposite to that of the native structures. Such incorrect TM residue orientations would imply completely different properties of the barrel interior and exterior. Here we report correct prediction only when we are able to exactly match the register with the experimentally resolved structure.

2.3.2 Predicted 3D structures of TM regions of βMPs

Feeding the predicted registers of each βMP into the intertwined zigzag coil model, we constructed the 3D structures of TM regions of βMPs. Figure 9A depicts the predicted structures (green) of the TM regions of protein OmpA, TodX, Porin, BamA, OpdO and HasR, which are shown superimposed on experimentally determined structures (cyan). The RMSDs of the main-

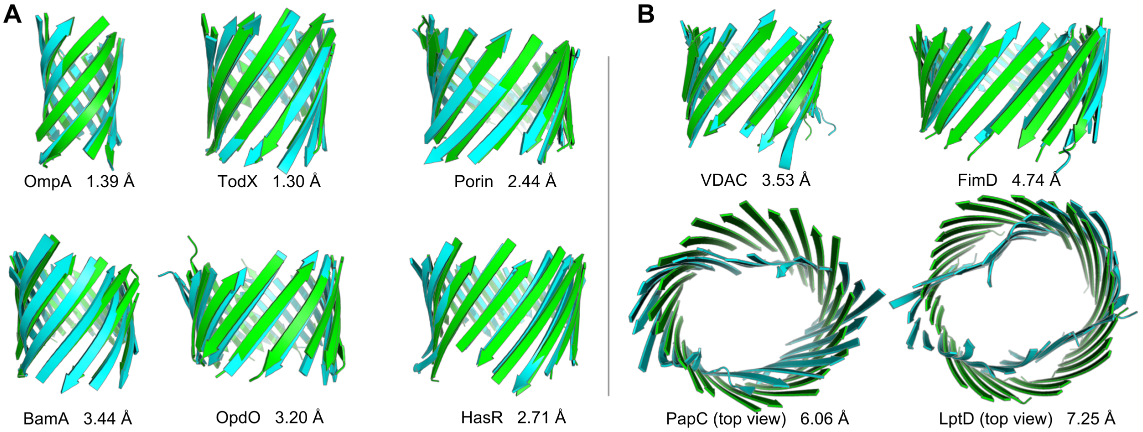


Figure 9: Structure prediction of TM regions. (A) Predicted structures of the TM regions (green) superimposed on experimentally determined structures (cyan): OmpA (1bxw), TodX (3bs0), Porin (1prn), BamA (4n75), OpdO (3szv), and HasR (3csl). (B) Predicted structures of the TM regions of proteins with novel folds (green) superimposed on experimentally determined structures (cyan): VDAC (3emn), FimD (3rfz), PapC (2vqi), and LptD (4q35). PapC and LptD are shown in top view.

chain atoms between the computed and experimentally resolved structures are 1.39Å, 1.30Å, 2.44Å, 3.44Å, 3.20Å and 2.71Å for OmpA, TodX, Porin, BamA, OpdO and HasR, respectively. The structures of the TM regions of 51 βMPs are predicted with an average RMSD of 3.48Å for mainchain atoms and 4.26Å for all atoms (see Table XVI and Figure 24 for details).

2.3.2.1 Intrinsic flexibility

TM regions of βMPs have considerable intrinsic flexibility: the NMR structures have an average mutual C_α-RMSD of 2.11 ± 0.79Å for the 7 βMPs with known NMR data (Table V, Column 2). The difference between the NMR and X-ray structures is more pronounced, with an average C_α-RMSD of 3.18 ± 1.16Å (Table V, Column 3). In contrast, the average C_α-RMSDs

TABLE V: Flexibility of TM regions of βMPs and the accuracy of the prediction of 3D-BMPP. $D^{\mathrm{TM}}_{s_1,s_2}$ is the average of the mutual C_α-RMSD between structures s_1 and s_2. * As no X-ray structures for these proteins are available, we used the first model of the NMR data.

PDB code	$D^{\mathrm{TM}}_{\mathrm{nmr,nmr}}$	$D^{\mathrm{TM}}_{\mathrm{nmr,X\text{-}ray}}$	$D^{\mathrm{TM}}_{\mathrm{pred,nmr}}$	$D^{\mathrm{TM}}_{\mathrm{pred,X\text{-}ray}}$
1bxw	1.41±0.42	1.99±0.31	1.83±0.15	1.36
1qj8	2.50±0.74	2.48±0.80	3.11±0.46	2.65
1thq	1.99±0.58	4.53±0.38	5.30±0.42	3.32
2f1c	2.42±0.37	2.80±0.21	3.93±0.21	3.06
2f1t	2.13±0.35	4.30±0.11	4.08±0.14	3.12
2lhf	0.82±0.22	No X-ray	1.60±0.08	1.48*
2mlh	1.48±0.28	No X-ray	1.49±0.14	1.44*
Mean	2.11±0.79	3.18±1.16	3.09±1.39	2.35±0.82

of our predicted structures against NMR and X-ray structures are 3.09 ± 1.39 and 2.35 ± 0.82, respectively (Table V, Column 4, 5). These differences are similar to the structural differences originating from the intrinsic flexibility of the proteins, suggesting that our prediction of TM regions of βMPs has excellent accuracy comparable to NMR structures.

2.3.2.2 Predicted structures of βMPs with novel folds.

It is challenging to predict the structures of βMPs with novel folds. βMPs were considered to have even numbers of strands from 8 to 22 (77). A βMP is considered to have a novel fold when its number of strands has not been observed in other experimentally determined structures. For example, VDAC in mitochondria has an odd number (19) of strands (25); PapC, FimD, and LptD all have more than 22 strands (24, 24, and 26, respectively). Predicting structures of a number of βMPs including VDAC, FimD, and LptD with reasonable accuracy was not possible in a recent study (62), likely due to inaccurate residue contact predictions and limitations in machine learning based procedure. Template–based prediction methods either fail to build any model or generate very poor structures. With the improved modeling procedure of 3D-BMPP, we are able to model the TM regions of the VDAC, FimD, PapC and LptD proteins with a mainchain RMSD of 3.53Å, 4.74Å, 6.06Å and 7.25Å, respectively (Figure 9B). While the structure of VDAC was previously predicted with an accuracy of 3.9Å (61) and 7.41Å (62), to the best of our knowledge the structures of FimD, PapC, and LptD have not been successfully predicted prior to this study. The large RMSDs of predicted structures of PapC and LptD show that our current idealized cylindrical structural model cannot yet model deformed barrels effectively.

2.3.2.3 Predicted structures of non-TM regions of βMPs

We also model the structures of the non-TM regions of βMPs, including the extended β-sheets (extended barrels) and loops connecting adjacent strands. The extended barrels have overall similar structures to those of the TM barrels. Including the extended barrel in our prediction increases the coverage of the modeled structures by 20% when measured by the average number of residues modeled in the 51 structures (159 in TM regions vs 191 in whole barrel regions, with the largest modeled barrel structure containing 350 residues), with little deterioration in the average mainchain RMSD (3.48Å vs 3.80Å).

2.3.2.4 Detailed prediction results of the barrel domains of all 51 βMPs

The set of 51 βMP structures are listed in Table XVI, along with the PDB codes, the organism for the protein, the number of TM strands, and the RMSD values between the TM region and the TM+extended barrel regions of real and modeled structures for mainchain and all atom models. It also lists the number of strands for which the strand register is correctly predicted before and after global shear optimization. The TM-regions of the predicted structures superimposed on experimentally determined structures are shown in Figure 24. A plot showing the RMSD against the size of the proteins can be seen in Figure 10.

2.3.2.5 Comparison with a previous study.

The accuracy of structure prediction is not sensitive to size of βMPs. For example, the prediction of a large βMP Iron(III) dicitrate transport protein FecA protein (237 TM residues) has a 2.71Å RMSD. This is in contrast to other prediction methods, where there is considerable deterioration in the quality of predicted structures (Table VI and Figure 10). The average TM-score of our predicted structures also compare favorably with those of a recent study (0.73 vs 0.54) (62). Furthermore, our results are over a much bigger dataset (51 vs 17 proteins). Thus, these results represent a very significant improvement. Moreover, the parametric structural model of intertwined zigzag coils improves accuracy of sidechains, as the all atom RMSD has improved by more than 1.30Å (4.26Å vs 5.60Å) compared to a previous study (61).

In a recent study, structures for 17 proteins (compared to the 51 proteins in this study) were predicted with an RMSD of 6.66 Å(62), as the number of sequences available for the remaining proteins is insufficient for computing sequence covariation. Our results show that this limitation can be removed by combining patterns of H–bond and sidechain interactions derived from experimentally resolved 3D structures with the sequence covariation information. Figure 11 shows that even when the available sequences are insufficient for sequence covariation analysis alone (accuracy $\sim 30\%$), our model model can make accurate strand register prediction ($\sim 70\%$). Our improved modeling methodology can predict the 3D structures of 51 βMPs with an average RMSD of 3.48 Å. Moreover, in Ref(62), TM-align was employed to assess accuracy of predicted structures, which does not give the appropriate assessment of prediction accuracy. TM-align is used when the correspondence or residue–residue mappings between two structures are not known, as it will decide which portions of the sub–structures are sufficiently similar for RMSD/TM-score calculation. In the case of computing the RMSD of a predicted structure and a known PDB structure, direct mappings of all TM residues between the two structures are already known and a straightforward direct RMSD calculation is required. We have carried out a direct measurement of RMSD using predicted structures of Hayat *et al.* The RMSD calculated using this approach is 6.66 Å, as compared to the reported 4.45 Å, which is the average RMSD of a subset of TM residues selected by TM-align. In addition, the authors of Ref(62) inflated their accuracy in strand register prediction by considering the predictions that were off by ±1 register as correct. As there is a direct relationship between the sidechain orientation and the functions of the proteins, this relaxed definition of "correct" registration implies erroneous sidechain orientation and thus incorrect functional regions of the proteins.

2.3.3 Structures of loops.

The predicted loop conformations are diverse (Figure 12A), and represent the broad conformational space that is accessible to loops (78). Examples of predicted loops are shown in Figure 12.

TABLE VI: Protein size and average mainchain RMSD using different prediction methods. Proteins with the number of strands ≤ 14 are grouped into the small dataset, those with > 14 and ≤ 20 strands are grouped into the medium dataset, and proteins with > 20 strands are grouped into the large dataset. In contrast to the other prediction methods, the quality of prediction of our methods, 3D-BMPP, does not deteriorate for large–sized proteins.

Method	Small	Medium	Large
TMBpro–server, 2008	6.0	6.3	11.8
3D-SPoT, 2012	3.9	4.5	**4.0**
EVfold_bb, 2015	4.9	7.7	9.3
3D-BMPP, 2018	**3.0**	**3.9**	**4.0**

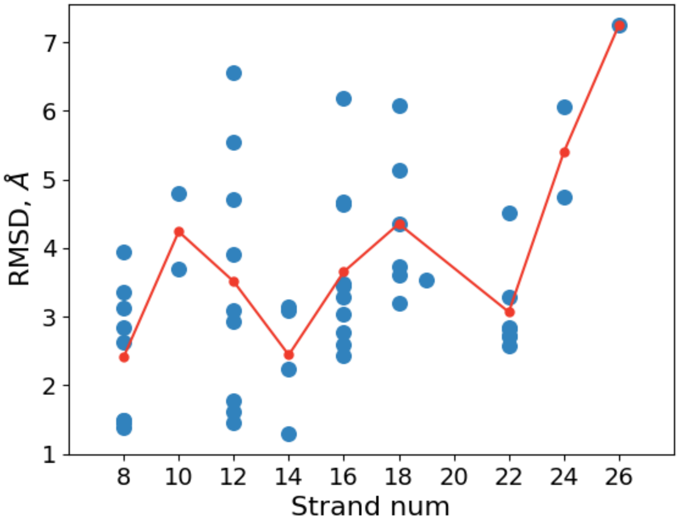


Figure 10: RMSDs of our prediction against the size of the proteins. Each blue dot represents one of the 51 predicted structures, while each red dot shows the average RMSD of predicted structures with the same corresponding strand number.

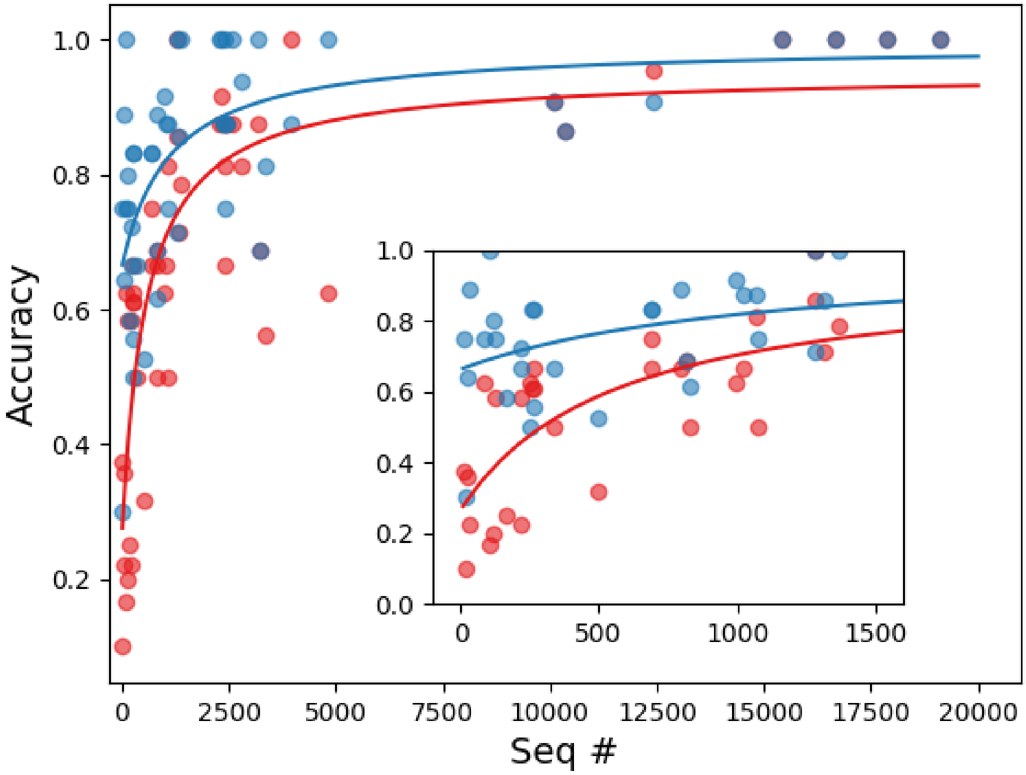


Figure 11: Our method on register prediction does not suffer from the limitation of requiring a large number (~ 1000) of available sequences for sequence covariation analysis. This figure shows how the number of available sequences affect register prediction accuracy. The numbers of sequences are found by HHblits (5). Each blue dot represents the register prediction for one protein using our model, while each red dot represents the prediction made by the sequence covariation analysis results alone (using Equation 2.4 and Figure 5). The blue and red curves are fitted from the corresponding dots, respectively. The inset shows the details of proteins when the available number of sequences is limited. In these cases, our model can still make accurate prediction (accuracy at $\sim 0.7 = 70\%$) while the prediction made using covariation analysis along is not reliable ($\sim 0.3 = 30\%$)

TABLE VII: Comparison of the accuracy of loop prediction for βMPs. We are able to sample most of the loop conformations seen in the NMR structures with <3Å deterioration in C_α-RMSD.

PDB code	$D^{\text{barrel}}_{\text{nmr,nmr}}$	$\Delta D^{\text{loop}}_{\text{nmr,nmr}}$	$D^{\text{barrel}}_{\text{nmr,pred}}$	$\Delta D^{\text{loop}}_{\text{nmr,pred}}$
1bxw	2.78±0.72	3.83±1.25	3.35±0.51	3.00±0.55
1qj8	3.31±0.80	0.61±0.26	4.14±0.57	0.67±0.27
1thq	1.99±0.58	0.79±0.35	5.30±0.42	0.52±0.21
2f1c	3.33±0.61	3.76±0.94	5.29±0.50	2.78±0.48
2f1t	2.58±0.54	1.01±0.55	4.35±0.15	0.45±0.20
2lhf	0.85±0.24	1.94±0.60	1.63±0.09	2.05±0.27
2mlh	1.48±0.28	1.51±0.64	1.49±0.14	0.99±0.26
Mean	3.65±1.21	1.03±0.89	3.64±1.46	1.12±0.89

To assess the quality of the predicted loop conformations, we define a metric $\Delta D^{\text{loop}}_{s_1,s_2}$ that measures how C_α-RMSD between structures s_1 and s_2 is changed upon incorporation of the loop regions: $\Delta D^{\text{loop}}_{s_1,s_2} = D^{\text{whole}}_{s_1,s_2} - D^{\text{barrel}}_{s_1,s_2}$, where $D^{\text{whole}}_{s_1,s_2}$ is the C_α-RMSD between the structures s_1 and s_2 including both the barrel and loop regions, and $D^{\text{barrel}}_{s_1,s_2}$ the C_α-RMSD between the barrel domains only. Since the number M of available NMR structures for each protein is limited compared to our predictions ($\sim 10 - 20$ vs ~ 400), we selected M predicted conformations closest to the NMR structures by $\Delta D^{\text{loop}}_{\text{nmr,pred}}$ from the modeled ensemble for each protein. The resulting $\Delta D^{\text{loop}}_{\text{nmr,pred}}$'s calculated using these structures are < 3Å, with an average of 1.12 ± 0.89 (Table VII Column 5), which is on par with the values of $\Delta D^{\text{loop}}_{\text{nmr,nmr}}$ (Table VII Column 3), suggesting that we are able to sample the loop conformations observed in the NMR structures accurately.

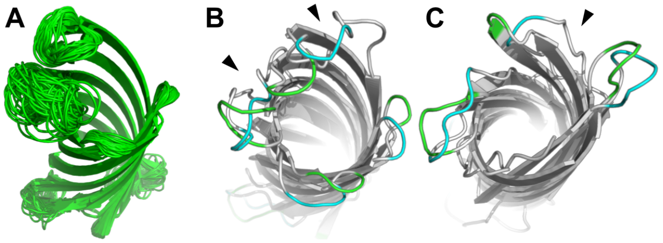


Figure 12: Structure prediction of loop regions. (A) Ensemble of predicted loop structures of OmpX (1qj8). (B & C) Examples of predicted loops on extracellular side (b, green) and on periplasmic side (c, green) superimposed on the corresponding NMR structure (cyan) (6). The black arrows indicate the big fluctuations in the barrel region.

2.4 Discussion

Due to the difficulties in experimental determination of membrane protein structures, there are a limited number of structures of nonhomologous βMPs. However, it is estimated that there are 15,000 βMPs across 600 different gram–negative chromosomes (79). Computational modeling has the promise to provide working 3D models for these sequences, enabling novel applications in nanopore engineering, drug design/delivery, as well as furthering understanding of the structural basis of function and mechanism of these βMPs. We have developed a method for predicting structures of βMPs, which combines a statistical mechanical model, sequence covariation information, and global register optimization, with a parametric structural model of intertwined zigzag coils. The results show that we can accurately predict structures of βMPs with a significantly expanded coverage of extended β-sheets and loops.

The incorporation of global register optimization increases the accuracy of the predicted structures by 0.24Å on average, suggesting that global H–bond network cannot be approximated accurately using local strand register alone. As an example, for the βMPs OmpA (PDB code: 1bxw), hypothetical protein HB27 (PDB code: 3dzm), and PagL (PDB code: 2erv), the strand registers were predicted correctly for 6 out of 8 strands before global register optimization, with an error in shear number of -4, -6, -6 respectively. After global register optimization, the strand register was predicted correctly for 8, 6, 4 strands respectively, and the error in shear number becomes 0 in all three cases. Moreover, the mainchain RMSD of these predicted structures is improved by 2.7Å, 2.5Å and 1.5Å , respectively.

Our parametric model of intertwined zigzag coils captures the zigzag nature of a polypeptide and the varied distance between C_α atoms of two adjacent strands, which depends on whether the respective residues share a mainchain H–bond. This results in significant improvement in RMSD for all atoms in general and sidechain atoms in particular. When we constructed structures of all 51 βMPs using our parametric model with true registers. The average mainchain RMSD of these structures was 2.5Å. Given our prediction accuracy of 3.48Å in this study, only $\sim$ 1Å error on average is due to incorrect register prediction, while the 2.5Å error is due to the structural deviation of βMPs from the ideal cylindrical shape.

Currently this ideal cylindrical model cannot capture ellipticity, twist, and curvature of local surface of the deformed barrel domains such as those observed in PapC and LptD (Figure 9B), and alternative hyperboloid models have been discussed in literature (80; 81). However, as current understanding of the physical factors determining these geometric properties is incomplete, further investigation of the heterogeneity of interactions in the TM region is required to develop more accurate geometric model that can account for deformed barrel domain.

In a recent study, structures for only 17 proteins (compared to 51 proteins in this study) were predicted (62), as the number of sequences available for the remaining proteins were insufficient to analyze sequence covariation. Here, we show that this limitation can be removed by combining patterns of H–bond and sidechain interactions derived from experimentally determined 3D structures with the sequence covariation information (Figure 11). Our method predicts the 3D structures of 51 βMPs with an average RMSD of 3.48Å, which compares favorably over the recent study that has an average RMSD of 6.66Å (62).

Our method revealed basic organizational principles of βMPs and requires no template structures. In addition, TM regions of βMPs with novel fold can also be modeled effectively, as evidenced by the predicted structures of VDAC and FimD. Furthermore, non-TM regions including both extended β-sheets and loops can be predicted accurately for the first time. Overall, our method opens the possibility of structural studies of many βMPs, including those in eukaryotic mitochondria and chloroplasts.

CHAPTER 3

EFFICIENT COMPUTATION OF TRANSFER FREE ENERGIES OF AMINO ACIDS IN BETA–BARREL MEMBRANE PROTEINS

Adapted from Tian, W., Lin, M., Naveed, H., and Liang, J.: Efficient computation of transfer free energies of amino acids in beta–barrel membrane proteins. Bioinformatics, 33(11):1664–1671, 2017.

3.1 Introduction

The transfer free energies (TFEs) of amino acid sidechains from aqueous environment to lipid bilayers provide the fundamental energetic contribution to the thermodynamic stability of βMPs (38; 82). The TFEs have been measured experimentally using different systems. Wimley and White used a set of peptides as the host and measured the water–to–octanol TFEs (4). Hessa *et al.* further measured TFEs by investigating the degree of insertion of a set of polypeptides through the translocon–meditated pathway into the endoplasmic reticulum (ER) membrane (39). A more recent significant development was reported by Moon and Fleming, who directly measured water–to–bilayer TFEs of amino acid residues in the context of a native transmembrane (TM) protein and a phospholipid bilayer for the first time (1).

In the study of Moon and Fleming (1), the protein outer membrane phospholipase A (OmpLA) was used as the TM scaffold, and the position of residue 210 which is close to the midplane of OmpLA was chosen as the host position Figure 13. The TFE of a residue from aqueous en-

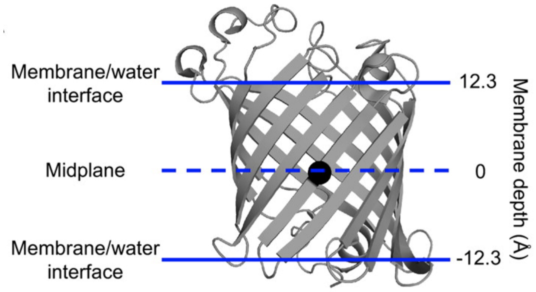


Figure 13: OmpLA and the residue 210. Figure adapted from Ref (1)

vironment to this position on the midplane of the lipid bilayer was measured, which is taken as the difference between the free energies of the spontaneous insertion of the scaffolds with that residue and with Ala at position 210. The TFEs of Leu and Arg to depths other than the midplane of the bilayer were also reported (1). This study provided direct measurements of the TFEs in the context of a whole protein. However, given the heterogeneity of the membrane along the bilayer normal direction, the reported TFEs are specific to positions near the midplane of lipid bilayers, except those of Leu and Arg. Moreover, experimental studies are costly and limited to a handful of client proteins due to the technical difficulties in establishing conditions for reversible folding (83; 84). It is therefore desirable to develop computational methods that allow rapidly computation of TFEs of residues at different depths of any βMP.

Several knowledge–based methods can approximate the TFE with statistical potentials calculated from depth–dependent propensities of amino acid residues (85; 86; 87; 88). However, this approach neglects important physical interactions between residues that are known to be important (42). In addition, these methods can only estimate averaged TFEs, but do not account for the specific local environment of a residue. TFEs have also been derived using molecular dynamics (MD) simulations (89; 90). However, the choice of the reference state remains a challenging problem, as reproducing TFEs of different residues requires different reference states (89).

An *ab initio* computational method has been recently developed, which can be used to compute the TFEs of TM residues in OmpLA (42). This method takes into account key physical interactions in the TM region and enumerates all conformations of the TM region in a reduced discrete conformational state space. While the computed TFEs of OmpLA are in an excellent agreement with the experimentally measured values, the application of this method is limited to βMPs with 14 or less TM strands, as the time cost of the enumeration of conformations grows rapidly with the number of strands. A computationally more efficient method is therefore necessary for larger βMPs.

Here we describe an approximation method to compute the TFEs of TM residues. Our approximation method is based on the divide–and–conquer strategy for conformational state enumeration, and allows rapid and accurate calculation of TFEs. We applied this new method to OmpLA, and the computed TFEs are in excellent agreement with the experimentally measured values. Our method can be used to derive the depth–dependent TFE profiles of all βMPs currently known, including the largest one consists of 26 strands.

TABLE VIII: The PDB codes of the βMPs used in this study.

Strand #	PDB code
8	1bxw, 1p4t, 1qj8, 1thq, 2erv, 2f1t, 2lhf, 2mlh, 3dzm
10	1i78, 1k24
12	1qd6, 1tly, 1uyn, 2wjr, 3aeh, 3fid, 3kvn, 4e1s, 4pr7
14	1t16, 2f1c, 3bs0, 3dwo
16	1e54, 1prn, 2o4v, 2omf, 2por, 2qdz, 3vzt, 4c00, 4gey, 4k3c
18	1a0s, 2mpr, 2ynk, 3rbh, 3syb, 3szv
22	1fep, 1kmo, 1nqe, 1xkw, 2fcp, 3csl, 3v8x
24	2vqi, 3rfz
26	4q35

3.2 Methods and Materials

3.2.1 Dataset

We use 50 non–homologous βMPs (resolution 1.45Å– 3.2Å) with less than 30% pairwise sequence identity for this study. The PDB codes of these proteins and the number of strands for each βMPs are listed in Table VIII

3.2.2 Reduced state space

Bacterial βMPs currently with known structures have even number of anti–parallel β–strands. Each strand of βMPs interacts with two neighbor strands via periodic repeating dyad bonds, which are characterized as strong H–bonds, weak H–bonds, and non–H–bonds (Van der Waals forces between sidechains) (91; 64; 3). We use a previously developed model to describe the state space of conformations of the TM region of a βMP (34; 42). In this model, the TM region of a βMP has n strands, each with a length of L. The model uses $L = 16$ based on the earlier study (3), but can take βMPs with TM strands of other length as input. A residue on one strand interacts with residues on the two neighbor strands via the three different types of interactions. The i-th strand can slide d_i residues away from its canonical central position, which is set as $d_i = 0$ (Figure 14A). The canonical central position of a strand is determined from the OPM database (92). A specific conformation of the TM region is therefore represented by a n-dimensional vector $\mathbf{d}$ in the state space:

$$\mathbf{d} = (d_1, \cdots, d_n) \in \mathbb{Z}^n.$$

We limit the state space by constraining the sliding window of each strand to $(-l, \cdots, l)$. Thus, the reduced state space Ω for the conformations of the TM region is :

$$\Omega = \{\mathbf{d} | \mathbf{d} = (d_1, \cdots, d_n) \in (-l, \cdots, l)^n\}.$$

The size of the state space is $|\Omega| = w^n$, where $w = 2l + 1$ is the width of the sliding window of the strands. In this study, we use $l = 3$, which makes Ω cover large enough conformational state space while remain tractable.

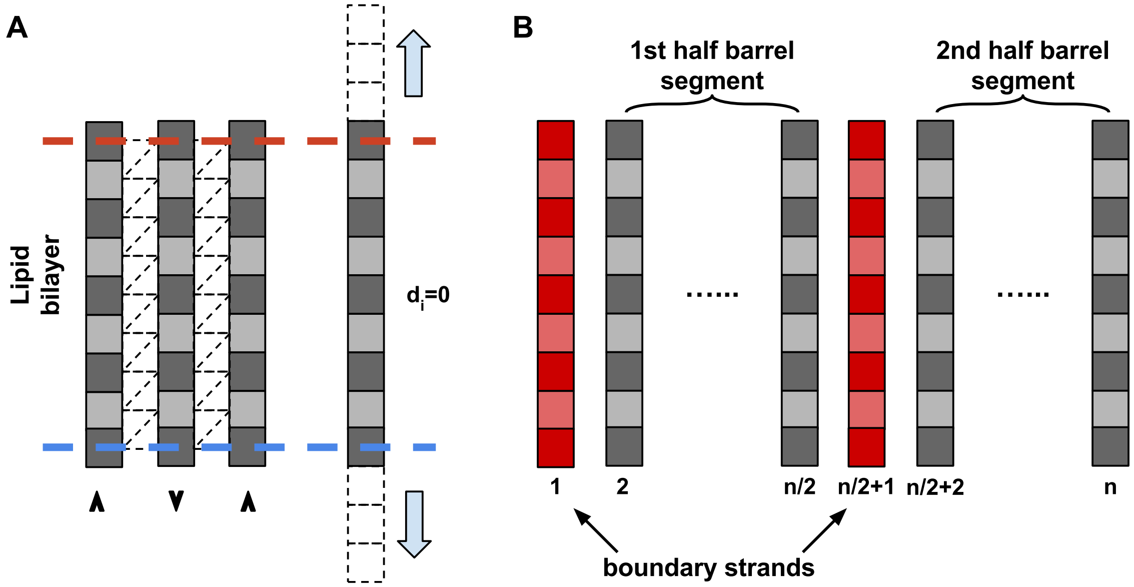


Figure 14: The reduced state space and the enumeration of conformational states . **A**. A strand interacts with neighbor strands through three different types of interactions (3). Each strand can slide up or down l positions. **B**. In approximation algorithms, the TM region of a βMP is divided into two half–barrel segments (grey) and two boundary strands (red).

3.2.3 Computing the TFE of an amino acid residue

The energy of the i-th strand of a specific conformation $\mathbf{d}$ is calculated using the empirical energy function

$$\begin{aligned} E(i;\mathbf{d}) &= w_{\mathrm{B}}E_{\mathrm{B}}(i,\mathbf{d}) + w_{\mathrm{Intra}}E_{\mathrm{Intra}}(i,\mathbf{d}) + w_{\mathrm{SH}}E_{\mathrm{SH}}(i,\mathbf{d}) \\ &\quad + w_{\mathrm{WH}}E_{\mathrm{WH}}(i,\mathbf{d}) + w_{\mathrm{NH}}E_{\mathrm{NH}}(i,\mathbf{d}), \end{aligned} \tag{3.1}$$

where E_{B} is the single residue burial energy, which depends on the residue location and the sidechain orientation; E_{intra} is the energy of intrastrand interactions among residues with the same sidechain orientation; E_{SH}, E_{WH}, and E_{NH} are the energies contributed by the interstrand strong H–bonds, the weak H–bonds, and the non–H–bonds, respectively. w_{B}, w_{Intra}, w_{SH}, w_{WH}, and w_{NH} are the corresponding weight coefficients. The derivation of the energy terms can be found in Ref (93; 3) and the values of the weights can be found in Ref (42). The total energy of the TM region of the βMP with a specific conformation $\mathbf{d}$ can then be computed as

$$E(\mathbf{d}) = \sum_{i}^{n} E(i;\mathbf{d}).$$

The partition function Z_{lip} of the TM conformational ensemble buried in the lipid bilayer can be calculated as

$$Z_{\mathrm{lip}} = \sum_{\mathrm{d}\in\Omega} \exp(-\frac{E(\mathbf{d})}{k_B T}). \tag{3.2}$$

The corresponding free energy is $G_{\text{lip}} = -k_B T \ln Z_{\text{lip}}$. The TFE of a residue in the TM region of the βMP from the aqueous environment to the lipid environment can be calculated as $\Delta G = G_{\text{lip}} - G_{\text{aq}}$, where G_{aq} is the free energy of the TM region of the βMP in the aqueous environment. For a specific amino acid residue in a given position of the TM region, its TFE with respect to an alanine at the same position can be calculated as $\Delta\Delta G_{\text{res}} = (G^{\text{res}}_{\text{lip}} - G^{\text{res}}_{\text{aq}}) - (G^{\text{ala}}_{\text{lip}} - G^{\text{ala}}_{\text{aq}})$. Assuming $G^{\text{ala}}_{\text{aq}} = G^{\text{res}}_{\text{aq}}$, we have

$$\Delta\Delta G_{\text{res}} = G^{\text{res}}_{\text{lip}} - G^{\text{ala}}_{\text{lip}}. \tag{3.3}$$

3.2.4 Exact algorithm of partition function calculation

3.2.4.1 Exact algorithm

The key step in calculating the TFEs is the computation of the partition function of the conformational ensemble of the TM region of a βMP. This can be achieved after enumerating all w^n number of conformations in the reduced state space Ω. The algorithm is listed as Algorithm 1.

We first enumerate the conformations of each strand–triplet (d_{i-1}, d_i, d_{i+1}) in the local state space $(-l, \cdots, l)^3$ (Algorithm 1, lines 1–5). The energy $E(i; d_{i-1}, d_i, d_{i+1})$ of the middle strand in a local conformation (d_{i-1}, d_i, d_{i+1}) is calculated using Equation 3.1, and the value is stored (Algorithm 1, line 3). The energy $E(\mathbf{d})$ of a specific conformation $\mathbf{d}$ of the whole TM region is then calculated by summing up the precomputed energies of the corresponding strands $E(\mathbf{d}) = \sum_{i=1}^{n} E(i; d_{i-1}, d_i, d_{i+1})$ (Algorithm 1, line 8). This is repeated for all the conformations of the TM region of the βMP in the reduced state space. The partition function can then be calculated using Equation 3.2 (Algorithm 1, lines 7–10). With this algorithm, every conformation of the TM region in the reduced state space is examined, and the partition function is computed exactly.

Algorithm 1: Exact algorithm of partition function calculation

```
   ▷ enumerate strand--triplet conformations and precompute energies of
     strands
1  for $i \leftarrow 1$ to $n$ do
2      foreach $(d_{i-1}, d_i, d_{i+1}) \in (-l, \cdots, l)^3$ do
3          $E(i; d_{i-1}, d_i, d_{i+1}) \leftarrow$ energy of the middle strand;
4      end
5  end
   ▷ compute the partition function
6  $Z \leftarrow 0$;
7  foreach $\mathbf{d} = (d_1, \cdots, d_n) \in (-l, \cdots, l)^n$ do
8      $E(\mathbf{d}) \leftarrow \sum_{i=1}^{n} E(i; d_{i-1}, d_i, d_{i+1})$;
9      $Z \leftarrow Z + \exp(-\beta E(\mathbf{d}))$;
10 end
11 return $Z$;
```

3.2.4.2 Time complexity

The time cost of the precomputation of strand energies (Algorithm 1, lines 1–5) is $O(nw^3)$. This is negligible compared to the enumeration of the whole TM region and the computation of the partition function (Algorithm 1, lines 7–10). To compute the energy $E(\mathbf{d})$ of a specific conformation $\mathbf{d}$, n number of additions (a) and 1 exponentiation (e) are required (Algorithm 1,

TABLE IX: Average running time using the exact algorithm (Algorithm 1) and the approximation algorithm with the histogram scheme (Algorithm 3) to calculate all 20 substitutions for a given host position. These values are recorded from programs running on a 2600MHZ CPU written in C++ programming language except those marked by *, which are extrapolated based on the complexity analysis (Equation 3.4).

# of strands	Exact Algorithm	Approx. Algorithm
8	5.07 sec	3.70 sec
10	4.26 min	1.03 min
12	3.58 hr	4.26 min
14	8.01 day	8.84 min
16	1.25 yr*	12.80 min
18	70.03 yr*	19.05 min
22	2.12×10^5 yr*	7.22 hr
24	1.15×10^7 yr*	2.09 day
26	6.22×10^8 yr*	15.22 day

lines 8–9). Since there are w^n number of different conformations in the reduced state space, the time complexity of computing the partition function is

$$O\left((an+e)\cdot w^n\right) = O\left(nw^n\right). \tag{3.4}$$

The running time of Algorithm 1 is only feasible when the number of strands is small ($n \leq 14$). Calculation of the partition function of a βMP with more strands would requires unrealistic amount of time as the time complexity is superexponential (Table IX).

3.2.5 Approximation algorithm of partition function calculation

3.2.5.1 Approximation algorithm

To compute the partition function of a βMP with a large number of strands ($n > 14$), we have developed an approximation algorithm based on the divide–and–conquer strategy, which is listed as Algorithm 2. We first divide the TM region of the βMP into four components. The first two components are the two half–barrel segments, consisting of strands from $i = 2$ to $\frac{n}{2}$ and from $i = \frac{n}{2} + 2$ to n, respectively. The other two components are the two boundary strands $i = 1$ and $i = \frac{n}{2} + 1$, separating the two half–barrel segments (Figure 14B).

For strands in the half–barrel segments, energies of strands are precomputed the same as in Algorithm 1 (Algorithm 2, line 4). For the boundary strands, interactions between the strand and the neighbor strands are ignored, and only the single strand energy $\hat{E}(i; \mathbf{d})$ is precomputed (Algorithm 2, line 5) as $\hat{E}(i; \mathbf{d}) = w_{\mathrm{B}} E_{\mathrm{B}}(i, \mathbf{d}) + w_{\mathrm{Intra}} E_{\mathrm{Intra}}(i, \mathbf{d})$, which is the summation of the first two terms of Equation 3.1. Conformations of each half–barrel segment are then enumerated, and energies of these conformations are calculated and stored (Algorithm 2, lines 13–16 and 17–20). The total energy of a given conformation $\mathbf{d}$ of the whole TM region is then calculated by combining the energies of the corresponding half–barrel segments and the boundary strands (Algorithm 2, line 23). Overall, the partition function can be computed by enumerating the positions of the boundary strands and the local conformations of the half–barrel segments.

The calculated free energy G of the TM region is underestimated, as the interactions between the boundary strands and the neighbor strands are ignored. We note that the TFE of a residue in the interior of a half–barrel segment is not affected much, given that the underestimation is systematic for both the free energy terms $G_{\text{lip}}^{\text{res}}$ and $G_{\text{lip}}^{\text{ala}}$ in Equation 3.3. In addition, the overall impact of neglecting strand–strand interaction for the boundary strands will decreases as the number of the strands increases, since the number of the neglected interactions over the number of the total interactions within the TM region decreases.

3.2.5.2 Strand reindexing

The TFE of a residue on a boundary strand or a neighbor strand may not be of sufficient accuracy. This can be solved by simply reindexing the strands of the βMP, so that the strand containing the residue of interest is no longer a boundary strand. Specifically, we set the index of the strand containing the residue of interest to $\lceil \frac{n}{4} \rceil + 1$, and change the indices of all the other strands accordingly: For instance, if the residue is located on the 1st strand of a βMP with 8 strands, the strands can be reindexed as $(3, 4, 5, \cdots, 8, 1, 2)$ instead of $(1, 2, 3, \cdots, 6, 7, 8)$. We then again use the 1st and the $(\frac{n}{2} + 1)$-th strands as the boundary strands with the new indices. The residue of interest will be located on the middle strand of a half–barrel segment after reindexing, which minimizes the accuracy loss out of the approximation.

3.2.5.3 Time complexity

In the approximation algorithm, w^2 number of combination of the positions of the boundary strands are enumerated (Algorithm 2, lines 9–10). For a given combination of the positions of the boundary strands, the conformations of each half–barrel segment are enumerated, and their contributions to the partition function are computed (Algorithm 2, lines 11–26). Each half–barrel segment has $\frac{n}{2} - 1$ number of strands, resulting in $w^{\frac{n}{2}-1}$ number of local conformational states. Thus, $(\frac{n}{2} - 2) \cdot w^{\frac{n}{2}-1}$ number of addition operations are required to compute the energies

Algorithm 2: Approximation algorithm of partition function calculation

```
reindex the strands if necessary;
▷ enumerate strand--triplet conformations and precompute energies of strands
for $i \leftarrow 1$ to $n$ do
    foreach $(d_{i-1}, d_i, d_{i+1}) \in (-l, \cdots, l)^3$ do
        $E(i; d_{i-1}, d_i, d_{i+1}) \leftarrow$ energy of the middle strand;
        $\hat{E}(i; d_i) \leftarrow$ single strand energy of the middle strand;
    end
end
$Z \leftarrow 0$;
▷ enumerate positions of the boundary strands
for $d_1 \leftarrow -l$ to $l$ do
    for $d_{n/2+1} \leftarrow -l$ to $l$ do
        $LE_1 \leftarrow$ new list;
        $LE_2 \leftarrow$ new list;
        ▷ compute energies of the 1st half--barrel segment
        foreach $(d_2, \cdots, d_{n/2}) \in (-l, \cdots, l)^{n/2-1}$ do
            $E_1 \leftarrow \sum_{i=2}^{n/2} E(i; d_{i-1}, d_i, d_{i+1})$;
            insert $E_1$ to $LE_1$;
        end
        ▷ compute energies of the 2nd half--barrel segment
        foreach $(d_{n/2+2}, \cdots, d_n) \in (-l, \cdots, l)^{n/2-1}$ do
            $E_2 \leftarrow \sum_{i=n/2+2}^{n} E(i; d_{i-1}, d_i, d_{i+1})$;
            insert $E_2$ to $LE_2$;
        end
        ▷ combine energies of the four components
        foreach $E_1$ in $LE_1$ do
            foreach $E_2$ in $LE_2$ do
                $E \leftarrow E_1 + E_2 + \hat{E}(i; d_i) + \hat{E}(n/2+1; d_{n/2+1})$;
                $Z \leftarrow Z + \exp(-\beta E)$;
            end
        end
    end
end
return $Z$;
```

of the local conformations. The time complexity of enumerating both segments (Algorithm 2, lines 13–20) is $O\left(2 \cdot a \cdot (\frac{n}{2} - 2) \cdot w^{\frac{n}{2}-1}\right) = O\left(a(n-4)w^{\frac{n}{2}-1}\right)$. Since the contribution of one half–barrel segment to the partition function is independent of the other segment, $(w^{\frac{n}{2}-1})^2$ times of combining both segments into global conformations are required (Algorithm 2, lines 21–22), where each combining operation needs 4 addition and 1 exponentiation operations (Algorithm 2, lines 23–24). Therefore, together with the w^2 number of positions of the boundary strands, the overall time complexity of this approximation algorithm is

$$\begin{aligned} & w^2 \cdot O\left(a(n-4)w^{\frac{n}{2}-1} + (w^{\frac{n}{2}-1})^2 \cdot (4a+e)\right) \\ & = O\left(nw^{\frac{n}{2}} + w^n\right). \end{aligned} \tag{3.5}$$

To summarize, the first term $O\left(nw^{\frac{n}{2}}\right)$ of Equation 3.5 comes from energy calculation of the two half–barrel segments, while the second term $O\left(w^n\right)$ comes from combining energies into partition function. The overall complexity shows an exponential running time when n is large.

3.2.6 Approximation algorithm with histogram scheme

3.2.6.1 Histogram scheme

To further improve the approximation algorithm, we developed a histogram scheme which reduces the time complexity of Equation 3.5. The approximation algorithm using this histogram scheme is listed as Algorithm 3. We record both the maximum and the minimum energies of each middle strands (Algorithm 3, lines 8–9) when enumerating the local conformation of strand–triplets, from which the upper and the lower bounds of the energies of the two half–

barrel segments can be estimated (Algorithm 3, lines 12–13). The range of energy for each half–barrel segment is then divided into small intervals, each associated with a bin to record the number of conformations of the half–barrel segment whose energy falls within this interval. When calculating the energies of the conformations of the two half–barrel segments, we only need to record the number of hits of each bin instead of storing the energy values (Algorithm 3, lines 21–22 and 26–27). Note that the histogram scheme can also be used in the exact algorithm to reduce the running time (Algorithm 4).

3.2.6.2 Time complexity

With this histogram scheme, the time complexity of combining the two half–barrel segments (Algorithm 3, lines 29–36) no longer grows exponentially with the number of strands (the second term in Equation 3.5), but depends only on the bin size and the ranges of the energies ($max_1 - min_1$ and $max_2 - min_2$). In practice, the number of bins for a half–barrel segment is typical less than 10^4, resulting in a much smaller cost to combine the energies of the two half–barrel segments, namely, 10^8 additions, instead of the $10^{10} \sim 10^{20}$ additions required in Algorithm 2. With the histogram scheme, the second term $O(w^n)$ in Equation 3.5 becomes roughly a constant for large n. Although the first term $O(nw^{\frac{n}{2}})$ remains superexponential, it is much smaller than $O(w^n)$ when $n > 14$. TFEs of the largest βMP currently known ($n = 26$) can be effectively computed using Algorithm 3.

Algorithm 3: Approximation enumeration with histogram scheme

```
reindex the strands if necessary;
▷ enumerate strand--triplet conformations and precompute energies of strands
for $i \leftarrow 1$ to $n$ do
    $\min_{E_i} \leftarrow$ +inf;
    $\max_{E_i} \leftarrow$ -inf;
    foreach $(d_{i-1}, d_i, d_{i+1}) \in (-l, \cdots, l)^3$ do
        $E(i; d_{i-1}, d_i, d_{i+1}) \leftarrow$ energy of the middle strand;
        $\hat{E}(i; d_i) \leftarrow$ single strand energy of the middle strand;
        $\min_{E_i} \leftarrow min(\min_{E_i}, E(i; d_{i-1}, d_i, d_{i+1}), \hat{E}(i; d_i))$;
        $\max_{E_i} \leftarrow max(\max_{E_i}, E(i; d_{i-1}, d_i, d_{i+1}), \hat{E}(i; d_i))$;
    end
end
▷ estimate the ranges of the energies of the half--barrel segments
$[\min_1, \max_1] \leftarrow [\sum_{i=2}^{n/2} \min_{E_i}, \sum_{i=2}^{n/2} \max_{E_i}]$;
$[\min_2, \max_2] \leftarrow [\sum_{i=n/2+2}^{n} \min_{E_i}, \sum_{i=n/2+2}^{n} \max_{E_i}]$;
$LB_l \leftarrow$ new bin list covers range $[min_1, max_1]$;
$LB_r \leftarrow$ new bin list covers range $[min_2, max_2]$;
$Z \leftarrow 0$;
▷ (continue on the next page) ;
```

Algorithm 3: (Cont'd) Approximation enumeration with histogram scheme

```
▷ enumerate positions of the boundary strands
for $d_1 \leftarrow -l$ to $l$ do
    for $d_{n/2+1} \leftarrow -l$ to $l$ do
        ▷ compute energies of the 1st half--barrel segment
        foreach $(d_2, \cdots, d_{n/2}) \in (-l, \cdots, l)^{n/2-1}$ do
            $E_1 \leftarrow \sum_{i=2}^{n/2} E(i; d_{i-1}, d_i, d_{i+1})$;
            from $LB_1$ get bin $b_1$ corresponding to $E_1$;
            $b_1 \leftarrow b_1 + 1$;
        end
        ▷ compute energies of the 2nd half--barrel segment
        foreach $(d_{n/2+2}, \cdots, d_n) \in (-l, \cdots, l)^{n/2-1}$ do
            $E_2 \leftarrow \sum_{i=n/2+2}^{n} E(i; d_{i-1}, d_i, d_{i+1})$;
            from $LB_2$ get bin $b_2$ corresponding to $E_2$;
            $b_2 \leftarrow b_2 + 1$;
        end
        ▷ combine energies of the four components
        foreach $b_1 \in LB_1$ do
            foreach $b_2 \in LB_2$ do
                $E_1 \leftarrow$ the energy value corresponding to bin $b_1$;
                $E_2 \leftarrow$ the energy value corresponding to bin $b_2$;
                $E \leftarrow E_1 + E_2 + \hat{E}(i; d_i) + \hat{E}(n/2+1; d_{n/2+1})$;
                $Z \leftarrow Z + b_1 \cdot b_2 \cdot \exp(-\beta E)$;
            end
        end
    end
end
return $Z$;
```

3.3 Results

3.3.1 Accuracy of the approximation algorithm

The exact algorithm was previously used to calculate the TFEs of the 20 amino acid residues at the position 210 of OmpLA, and the results are in excellent agreement with experimental measurements (42). To evaluate the accuracy of the approximation algorithms, we compare results computed using the approximation algorithms on a collection of βMPs with strand number $n \leq 12$ with results computed using the exact algorithm. Specifically, for each lipid–facing host position in the TM region of a βMP, the TFEs of all 20 amino acid substitutions are calculated using the approximation algorithm (AA), the approximation algorithm with reindexing (AAR), and the exact algorithm. For each host position, each method generates a 20 dimensional vector of the TFEs of the 20 amino acid substitution. The root–mean–square error (RMSE) between the vectors computed using the approximation algorithms and using the exact algorithm is then calculated to assess the accuracy of the approximation algorithms. For instance, an RMSE value of 0.1 $kcal/mol$ indicates that the error of the TFEs of a given residue calculated by the approximation algorithm is on average 0.1 $kcal/mol$ compared to the exact algorithm, and this level of accuracy is sufficient. Examples of the 20 dimensional vectors and of the RMSEs can be found in Table XVII.

Table X summarized the accuracies of the approximation algorithms. For AA, the calculated TFEs residues around boundary strands (BD) are less accurate than those of the residues in interior of half–barrel segments. When strand reindexing is applied (AAR), the accuracy is improved significantly.

TABLE X: The average RMSEs between the results of the approximation algorithms and of the exact algorithm. The column BD shows the RMSEs of the TFEs of the residues on or neighboring a boundary strand, while the column non–BD shows the RMSEs of the residues in the interior of a half–barrel segment. Small RMSE values correspond to high accuracy.

	AA			AAR
# of strands	BD	non–BD	overall	
8	0.3436	0.1603	0.2743	0.1519
10	0.4773	0.1863	0.3350	0.0950
12	0.3430	0.0875	0.1932	0.0461

As expected, the accuracy improves as the number of strands increases for both AA and AAR. The accuracy of AAR is already adequate when the number of strand is as small as 12, indicating that AAR will give accurate results for larger βMPs.

Bin size for the histogram scheme

To identify the appropriate size of the small energy intervals, or the bin size, used in the histogram scheme, we assess the accuracy of results calculated using a version of the exact algorithm where the histogram scheme is used (Algorithm 4) by comparing with results calculated with the original exact algorithm (Algorithm 1). Different bin sizes are tested. The accuracy improves as the bin size decreases (Table XI), and is already sufficiently accurate when the bin size is set to 0.01 $kcal/mol$. Indeed, using the approximation algorithm with reindexing and the histogram scheme at the bin size of 0.01 $kcal/mol$, the calculated TFEs of the host position 210 in OmpLA are in excellent agreement with the experimental data (Figure 15) as the exact algorithm (42). There is a deviation of our result of Pro from that of the experimental measure-

TABLE XI: The RMSEs between results of exact algorithms with and without the histogram scheme. Different bin sizes (BS) are tested. The unit of the bin sizes is $kcal/mol$.

# of strands	BS=1	BS=0.1	BS=0.01	BS=0.001
8	0.2000	0.0385	0.0042	0.0004
10	0.2375	0.0422	0.0052	0.0005
12	0.1660	0.0318	0.0203	0.0003

TABLE XII: Comparison between the method in this study and the other computational methods. The R^2 of the results of each computational method is calculated against experimentally measured values (1). The *original* column is calculated with all 20 amino acids, while the *filtered* column is calculated with Cys, Met, Thr, Trp, and Tyr excluded.
[1]: Data of Cys not reported.
[2]: Data of Trp and Tyr not reported.
[3]: Data of Cys, Met, and Thr not reported.

R^2	Original	Filtered
Adamian *et al.*(85)	0.73	0.73
Slusky *et al.*[1](88)	NA	0.32
Schramm *et al.*[2](86)	NA	0.53
Hsieh *et al.*[3](87)	NA	0.66
This study	0.81	0.80

ment. We note that the TFE of Pro is known to have wide discrepancy among different scales (4; 39; 1). Comparison between the results of our method and other computational studies is shown in Table XII.

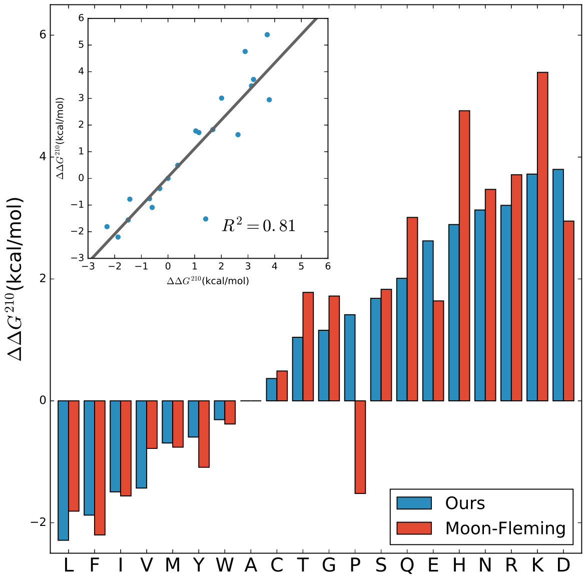


Figure 15: TFE $\Delta\Delta G^{210}$s at the host position 210 of OmpLA. Computed TFEs are shown in blue bars. Experimentally measured values are shown in red bars. The results of the approximation method correlate well with the experimental data with an R^2 of 0.81 (inset).

3.3.2 Efficiency of the approximation algorithms

Compared with the exact algorithm, the approximation algorithm using the histogram scheme has significantly improved computing efficiency (Table IX). For all βMPs we tested, including the largest βMP currently known ($n = 26$), the computation of TFEs at any host position can be completed in a realistic amount of time.

3.3.3 Transfer free energy profile of a large βMP

With guaranteed accuracy and efficiency of the approximation algorithm, we are able to extend the calculation of TFEs beyond small βMPs to medium and large βMPs. We calculated TFEs for residues in the TM region of lipopolysaccharide transport proteins D (LptD, PDB code: 4q35) (57), which has 26 TM strands. The full set of 116 lipid–facing host positions in the TM region are systematically substituted to all the 20 amino acids, and the corresponding TFEs calculated. By averaging the TFEs of the same amino acid residues in the same depth of lipid bilayer following ref (42), we obtained the depth–dependent TFE profile of this βMP (Figure 16). As the outer membrane is highly heterogeneous, the energy cost of transferring one residue into positions of different depth in the lipid bilayer is different, which is consistent with the profile of OmpLA previously reported (42). Moreover, comparison of these two TFE profiles suggest that they are highly correlated ($R^2 = 0.94$) despite their differences in size, assembly state, and function.

3.4 Discussion

The free energies of transferring amino acid sidechains from an aqueous environment to lipid bilayers quantify the fundamental energetic contributions to the thermodynamic stability of the

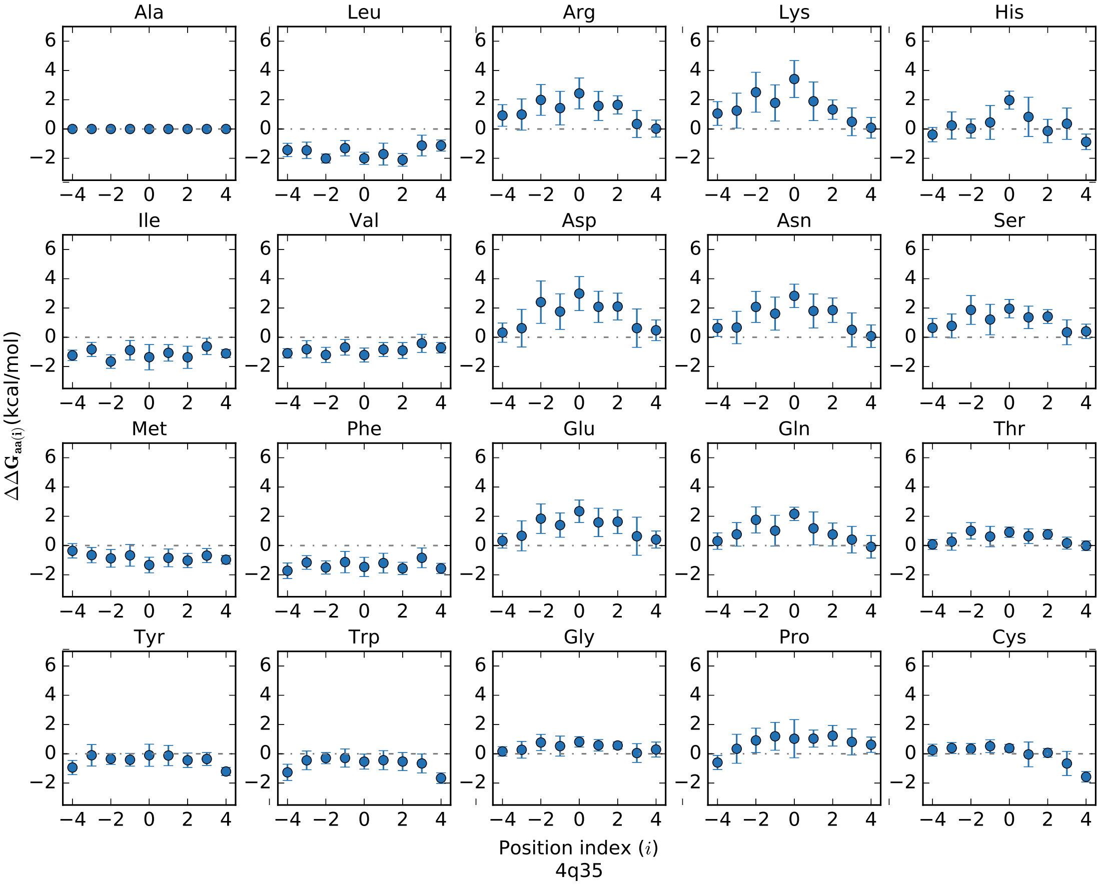


Figure 16: The depth–dependent TFE profile of lipopolysaccharide transport proteins D (PDB code: 4q35). The profile was calculated by systematically substituting each residue at all lipid–facing host positions in the TM region to the other 19 amino acids. The averages and the standard deviations of the calculated TFEs of the same amino acid substitution at host positions of the same depth are plotted. The position index of the depth is 0 at the midplane, and -4 – -1 on the periplasmic leaflet and 1 – 4 on the extracellular leaflet.

TM regions of TM proteins. A previously reported *ab initio* method can successfully calculate the TFEs of the lipid–facing residues of OmpLA, a βMP with 12 strands (42). However, the method was not applicable to large βMPs since it requires an unrealistic amount of computing time. We have developed an efficient approximation method based on the divide–and–conquer strategy and a histogram scheme. The new algorithm enables us to compute TFEs of residues for all βMPs currently with known structures.

In the new approximation method, we first divide the TM region of a βMP into two half–barrel segments, and enumerate conformations of each half–barrel segment. The energies of the half–barrel segments are then combined into the partition function of the whole TM region. In general, the TM region of a βMP can be divided into k number of partial–barrel segments, and conformation enumeration and partition function combination can be carried out in a way similar to that of dividing the TM region into just two half–barrel segments. Without the histogram scheme, the time complexity of an algorithm with k partial–barrel segments will be $O\left((n-2k)w^{\frac{n}{k}+k-1}+w^{n}\right)$, where the first term comes from the conformation numeration and the energy calculation of the partial–barrel segments, and the second from combining these energies into the partition function. Since the second term dominates the time complexity, increasing k will not reduce the running time significantly while reducing the accuracy. Therefore, we chose $k=2$ in this study, and introduced a histogram scheme which further reduces the second term of the time cost to roughly a constant.

Our current study focuses on the βMPs located in the bacterial outer membrane. However, other βMPs, such as the VDAC protein in mitochondria and the beta–barrel toxins such as α–

hemolysin and γ–hemolysin can also be studied using our methods. While the bacterial outer membrane is asymmetric, as the outer leaflet consists of lipopolysaccharides, the membrane environments for VDAC and the beta–barrel toxins are symmetric. This difference can be taken into account with minor modification to the empirical energy function used in the computation as demonstrated in (42). Another difference with VDAC is that it has an odd number of strands, resulting in the parallel N– and C–terminal strands instead of the anti–parallel N– and C–terminal strands in all the other βMPs. This can be accounted for by using the appropriate H–bond pattern for this unique strand pair.

Furthermore, our method does not require knowledge of experimentally solved or computationally predicted structures, as long as the sequences of the TM region can be determined. Several methods predicting TM segments from the sequence can be used to identify TM strands (47; 94; 48). The absolute accuracy in strand prediction is not required. For example, deviations of the midplane position will have limited effects in the computed TFEs, as the correct position will be included during enumeration, and those conformations with significant deviations will have higher energy and contribute little in their Boltzmann factors. As an example, we calculated the TFEs of the position 210 of OmpLA using the TM segments predicted by BOCTOPUS2 (48), and the results agree well with the experimental results ($R^2 = 0.76$).

βMPs are drawing increasing attention because of their potential applications in bionanotechnology, including protein profiling (18), DNA sequencing (19), and small molecule detection (20). With our new methods, we can derive the depth–dependent TFE profile of each βMPs, which may help in understanding the general folding principles and membrane insertion processes of βMPs, as well as in delineating the structure–function relationship of a specific βMP. Such knowledges may also help in tailoring natural βMPs or designing artificial βMP channels with desired stability.

CHAPTER 4

GETFEP: A GENERAL TRANSFER FREE ENERGY PROFILE OF TRANSMEMBRANE PROTEINS

4.1 Introduction

A widely used measure to estimate the stabilities of membrane proteins is the transfer free energies (TFEs), which quantify the free energies of transferring amino acid residues from aqueous environment into lipid bilayers. The computational method described in the previous chapter enables us in efficient and reliable calculation of TFEs of any lipid–facing TM resides in βMPs (see Chapter 2 and Ref (53) for details). However, it is still useful if a general TFE profile applicable for all βMPs can be derived, which will not only help us in understanding βMP folding process and structure–funciton relationship, but also facilitate efficient evaluation of future engineering and design of βMP nanopores.

Although experiments can measure TFEs of specific residues in certain systems(4; 36; 39; 1), the technical challenges and high cost restrain large scale measurements. Complementing experimentally measured TFEs, several hydrophobicity scales have been derived computationally, which can aid in our understanding of the governing principles of membrane protein folding (95; 96; 97). The $E_Z\alpha$ and $E_Z\beta$ empirical potentials are knowledge–based hydrophobicity scales. They have been successfully applied in predicting the positioning of membrane proteins in the lipid bilayer, in discriminating sidechain decoys, and in identifying protein–lipid interfaces (98; 87). However, these scales obtained from statistical analysis do not consider the physical interactions either between residues from neighboring helices/strands or within the same helix/strand, which are known to be important for membrane protein folding(99; 100). There have also been studies based on molecular dynamics (MD) simulations to calculate TFEs (101; 89; 102), although the choice of the reference state before membrane insertion remains a challenging task (89).

In this study, we use the our new TFE calculation method to compute the depth–dependent TFE profile of each βMP in a non–redundant set of 58 βMPs. After examining their overall patterns, we found that there exists a general TFE profile applicable to all βMPs, which we call the <u>Ge</u>neral <u>T</u>ransfer <u>F</u>ree <u>E</u>nergy <u>P</u>rofile (GeTFEP). The GeTFEP agrees well with previously measured and computed TFEs. Analysis based on GeTFEP shows that residues in different regions of the TM segment have different roles during the membrane insertion process. Our results further reveal the importance of the sequence pattern of TM segments in stabilizing βMPs in the membrane environment. In addition, we also show that GeTFEP can be used to predict positioning and orientation of βMPs when embedded in the membrane, with overall results in good agreement with experimental data. Furthermore, we show that the GeTFEP can be used to locate structurally or functionally important sites of βMPs. In addition, TM segments of α–helical membrane proteins (αMPs) can also be accurately predicted using the GeTFEP, suggesting that the GeTFEP captures fundamental thermodynamic properties of amino acid residues inside membrane, and has general applicability in studying membrane protein.

4.2 Materials and methods

4.2.1 Dataset

We use 58 non–homologous β–barrel membrane proteins with less than 30% pairwise sequence identity for this study. The PDB codes are: 1a0s, 1bxw, 1e54, 1ek9, 1fep, 1i78, 1k24, 1kmo, 1nqe, 1p4t, 1prn, 1qd6, 1qj8, 1t16, 1thq, 1tly, 1uyn, 1xkw, 1yc9, 2erv, 2f1c, 2f1t, 2fcp, 2gr8, 2lhf, 2lme, 2mlh, 2mpr, 2o4v, 2omf, 2por, 2qdz, 2vqi, 2wjr, 2ynk, 3aeh, 3bs0, 3csl, 3dwo, 3dzm, 3fid, 3kvn, 3pik, 3rbh, 3rfz, 3syb, 3szv, 3v8x, 3vzt, 4c00, 4e1s, 4gey, 4k3c, 4pr7, 4q35, 7ahl, 3b07, 3o44.

4.3 Results

4.3.1 GeTFEP: General Transfer Free Energy Profile

4.3.1.1 Computation of TFE profiles of βMPs

Using the computational method to calculate TFEs (see Chapter 2 and Ref (53) for details), we calculate the depth–dependent TFE profiles for each βMP in the dataset. Briefly, for each βMP, we substituted each lipid–facing residue in the TM region to the other 19 amino acids. We calculated the TFEs of each amino acid substitution using Ala as the reference. The TFE profile of the protein was then obtained by taking average of the TFE values of the same amino acid type at the same depth position in the membrane. As an example, Figure 16 shows the computed TFE profile of the protein LptD, the largest βMP with known structure (PDB code: 4q35).

4.3.1.2 Derivation of the general TFE profiles of βMPs by clustering analysis.

Although the 58 βMPs are in different oligomerization states, have different sizes (strand numbers) of TM segments, and come from different organisms, their TFE profiles are remarkably similar. To see if the profiles have similar patterns, we clustered the 58 TFE profiles using hierarchical clustering. We used euclidean distance between the TFE profiles of the βMPs and single linkage in the hierarchical clustering, and evaluated the clustering using the silhouette score, In practice, a > 0.5 silhouette score indicates a good clustering Our results show that it is not reasonable to cluster the βMPs into two groups, since the silhouette score is < 0.5 at 2 clusters (Figure 17A). When we increase the number of clusters, the silhouette score keeps decreasing. Therefore, we conclude that only one group exists for our βMP dataset. Figure 17B visualizes the cluster result after the profiles of βMPs are reduced to a 3D space using the Principal Component Analysis (PCA).

In the hierarchical clustering, we also tried other parameter settings with correlation distance and/or other reasonable linkages (eg. average linkage or weighted linkage), and the conclusion remains the same.

Results of clustering analysis show that the 58 βMPs can be grouped into only one group (with 56 βMPs) and two outliers: α– and γ–hemolysins (PDB codes: 7ahl and 3b07). Unlike the other βMPs, the TM regions of both α– and γ–hemolysins are formed by repeated β–hairpin (Figure 25B), which make their TFE profiles highly sensitive to the composition of the β–hairpin and the local interactions of residues within the hairpin (Figure 25 C and D). Accordingly, we further investigate whether α– and γ–hemolysins have truly different thermodynamic properties

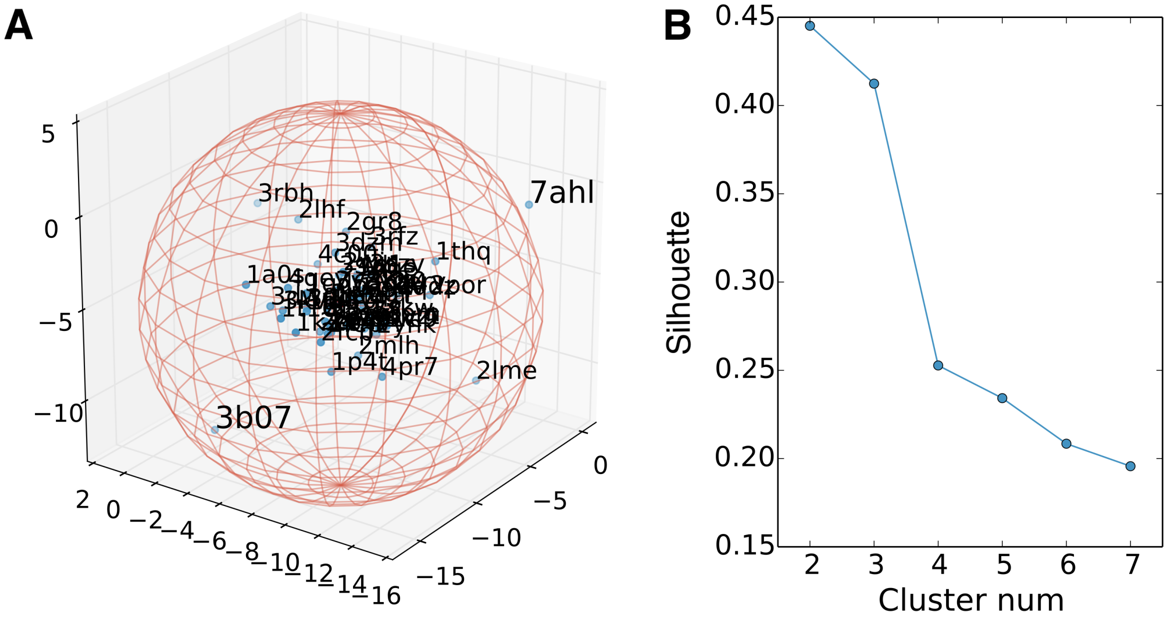


Figure 17: **A.** Visualization of the βMP TFE profiles after reduced to a 3D space via PCA. **B.** The silhouette scores for different cluster numbers of the βMP TFE profiles.

than the other βMPs, or their outlier status is due to the special architecture of repeated β–hairpins.

We first computed the TFE profiles of artificially generated hemolysin–like βMPs constructed by repeating each β–hairpin in our βMP set. Altogether, we computed TFE profiles for 778 artificial hemolysin–like βMPs. We then sampled from these profiles with replacement, and computed the distribution of the distance from each sampled profile to the average profile of all sampled artificial βMPs. The distances from the TFE profiles of both α– and γ– hemolysins to the average profile are at the 80th percentile in the distance distribution (Figure 18B), indicating that α– and γ–hemolysins are not fundamentally different in their thermodynamic properties from other βMPs. Therefore, we conclude that a general TFE profile exists and is applicable to all βMPs, including α– and γ–hemolysins. We derive the Ge̲neral T̲ransfer F̲ree E̲nergy P̲rofile (GeTFEP) by averaging the TFEs of a specific amino acid at the same lipid bilayer depth position for all 58 βMPs (Figure 18C).

4.3.1.3 Comparison with other hydrophobicity scales

We then examine how GeTFEP compares with other hydrophobicity scales. Since most experimentally measured scales are not depth–dependent, we first compare the scale of the TFEs at the hydrocarbon core position of depth 0 in the GeTFEP with other hydrophobicity scales. We refer this hydrophobicity scale as the mid–GeTFEP scale. The mid–GeTFEP scale correlates well with the experimentally measured hydrophobicity scales, having Pearson correlation coefficients $r = 0.83$ with the WW–scale, and $r = 0.92$ with the Bio–scale. It also correlates well with the computational βMP OmpLA scale(42; 53), with $r = 0.90$ (Figure 26).

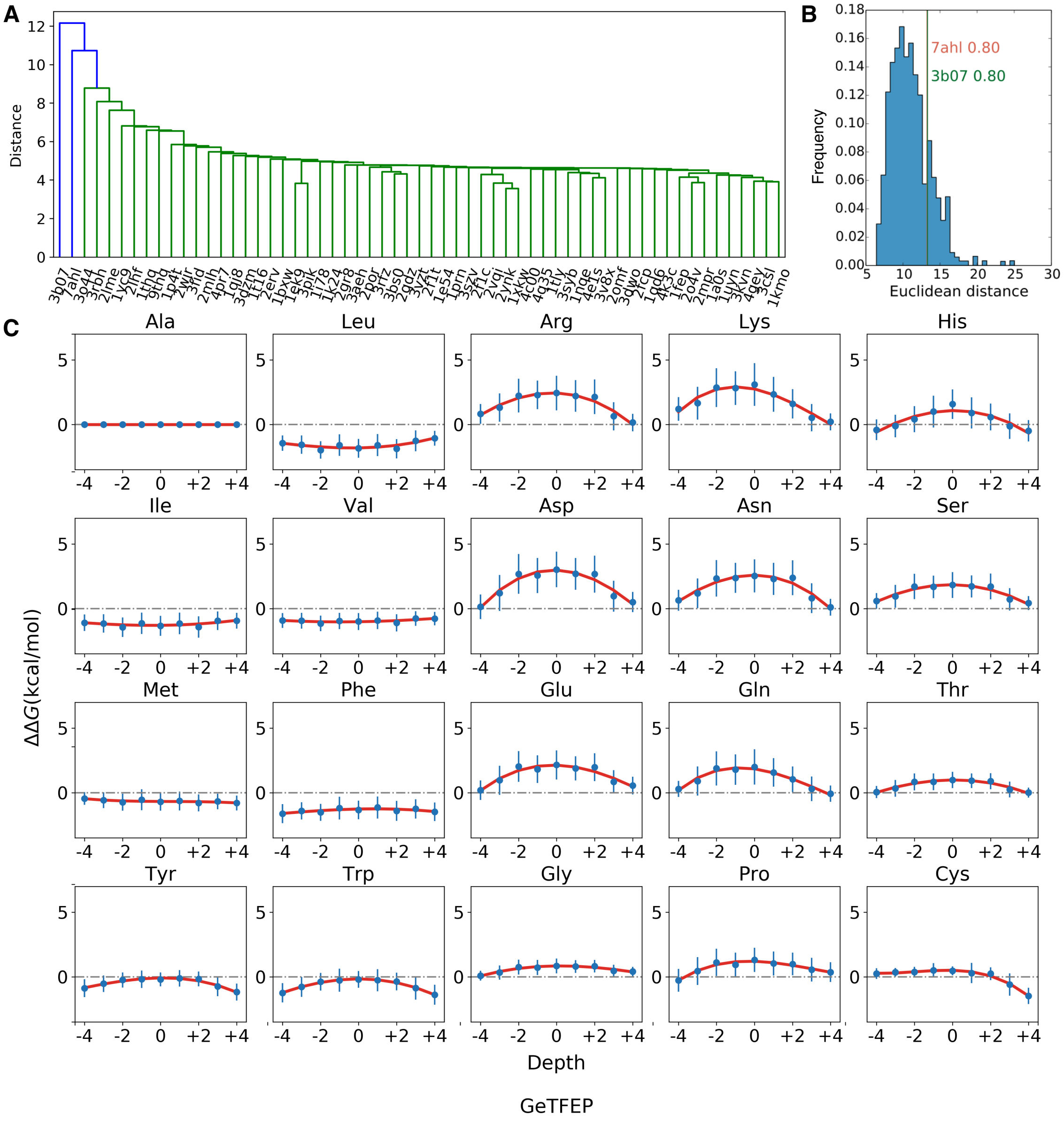


Figure 18: Derivation of the GeTFEP. **A.** Results of hierarchical clustering shows that all βMPs in the dataset can be group into one cluster, except α– and γ–hemolysins (7ahl and 3b07). **B.** The distribution of distance between the sampled TFE profiles of the artificially constructed hemolysin–like βMPs and their average TFE profile. The distances of both α– and γ–hemolysins are at the 80th percentile of the distribution. **C.** The General Transfer Free Energy Profile (GeTFEP) of each residues (blue), and the corresponding curves fitted by 3rd degree polynomials (red).

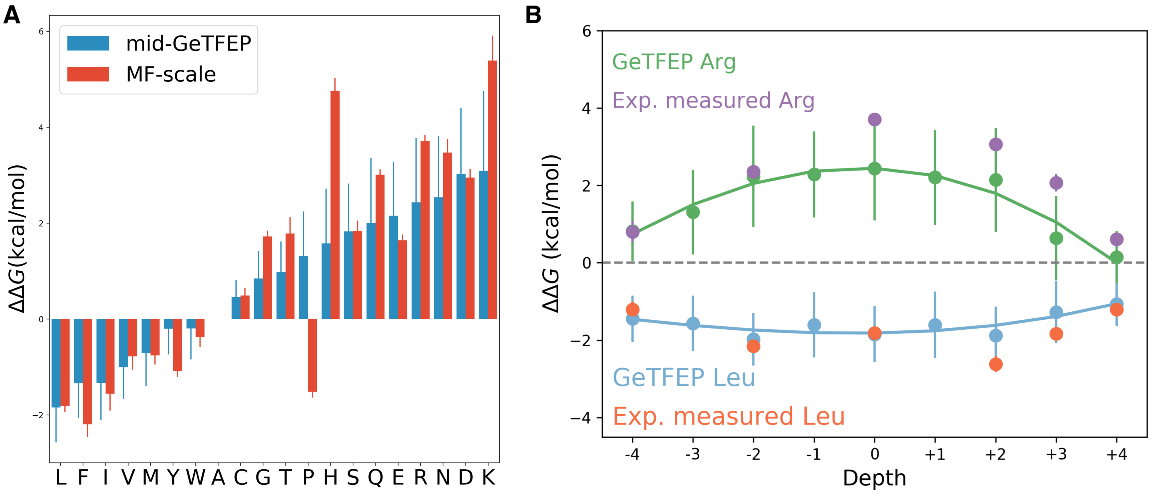


Figure 19: Comparison between the GeTFEP and the experimentally measured MF–scale. **A.** The mid–GeTFEP scale agrees well with the MF–scale, except Pro and His. **B.** The depth–dependent TFEs of Arg and Leu of GeTFEP also agree well with the experimental measurements (1).

When compared with the experimentally measured MF–scale of the βMP OmpLA mid–GeTFEP has a correlation of $r = 0.87$. One noticeable difference between mid–GeTFEP and the MF–scale is that the TFE value of His is less unfavorable in mid–GeTFEP (Figure 19A). This is expected since the MF–scale was measured in acidic condition at pH=3.8, where His was fully protonated (1). The different value in mid–GeTFEP likely reflects the property of His in physiological conditions of the outer membrane.

Another notable difference is Pro. It is found that Pro is unfavorable in the membrane environment according to the mid–GeTFEP scale, while it is found to be favorable according to the MF–scale (Figure 19A). Pro tends to disrupt the structures of both α–helix and β–sheet, and is thermodynamically unfavorable in the non–polar core of the membrane(103). The value of Pro in the GetFEP–mid scale reflects the general situation.

We then examined the depth–dependency of the GeTFEP of Arg and Leu, whose experimental results are available (1). Their TFEs at different depth positions of the membrane are in good agreement with the experimentally measured values, with $r = 0.87$ for Arg and $r = 0.75$ for Leu (Figure 19B), suggesting the GeTFEP captures the depth–dependency of TFEs of amino acids.

4.3.2 Insertion of βMPs into membrane

4.3.2.1 βMP insertion as a thermodynamically driven spontaneous process

Upon synthesis in the cytoplasm, βMPs need to be transported across the periplasm and then folded into the outer membrane. As there is no energy source such as ATP in the periplasm, it was suggested that the free energies of βMP folding provide an adequate source to ensure successful periplasm translocation (82). A computational study showed that the TFE of lipid–facing residues of the hydrophobic core regions are indeed the main driving force for membrane insertion (42). Analysis also showed that lipid–facing residues in the TM regions of of βMPs have clear patterns of amino acid composition (3). However, it is still unclear whether the insertion of βMPs into the membrane is primarily due to the extensive property of the hydrophobicity of lipid–facing residues, or the specific pattern of amino acid composition also plays important roles.

To investigate this question, we employed a simplified βMP insertion model based on the concerted folding mechanism proposed in Ref (7). We ignore the effects of non–TM loops and discretizes the insertion process into 17 steps (Figure 20A). We take the position recorded in the widely–used Orientations of Proteins in Membranes (OPM) database(92) as the fully inserted position of each βMP. This position is denoted as the reference position 0, and the other positions are indexed accordingly from -8 to $+8$. βMPs start the insertion process at position -8 from periplasmic side and become fully inserted into the membrane at position 0. From position 0 to $+8$, βMPs would translocate across the membrane. We assume that the stability of the TM region of a βMP can be approximated by summarizing TFEs of all lipid–facing residues in the membrane region. The stability of the βMP at each position was then calculated using the GeTFEP following this additive model. As an example, Figure 20B shows stability of the protein OmpA (PDB code: 1bxw) at different insertion positions. Overall, results of all βMPs show a funnel–like pattern of insertion energy (Figure 20C). Most βMPs (52 of 58) have the minimum free energy when they are fully inserted into membranes (position 0. See Table XVIII and Table XIX for details). The funnel–like pattern indicates that the insertion of βMPs into outer membranes is indeed a spontaneous process. βMPs become energetically trapped after being fully inserted.

For the βMPs (6 of 58) which are not most stable when fully inserted, the mismatch could come either from wrong fully inserted positions or insufficiency of the additive model. However, the most stable steps of all these βMPs are close to step 0 (steps 1 or -1), and the minimum TFEs are close to the TFEs of step 0 as well (Table XIX). Nevertheless, we use only the 52 βMPs with energetic minimum in step 0 for the further tests in the next section.

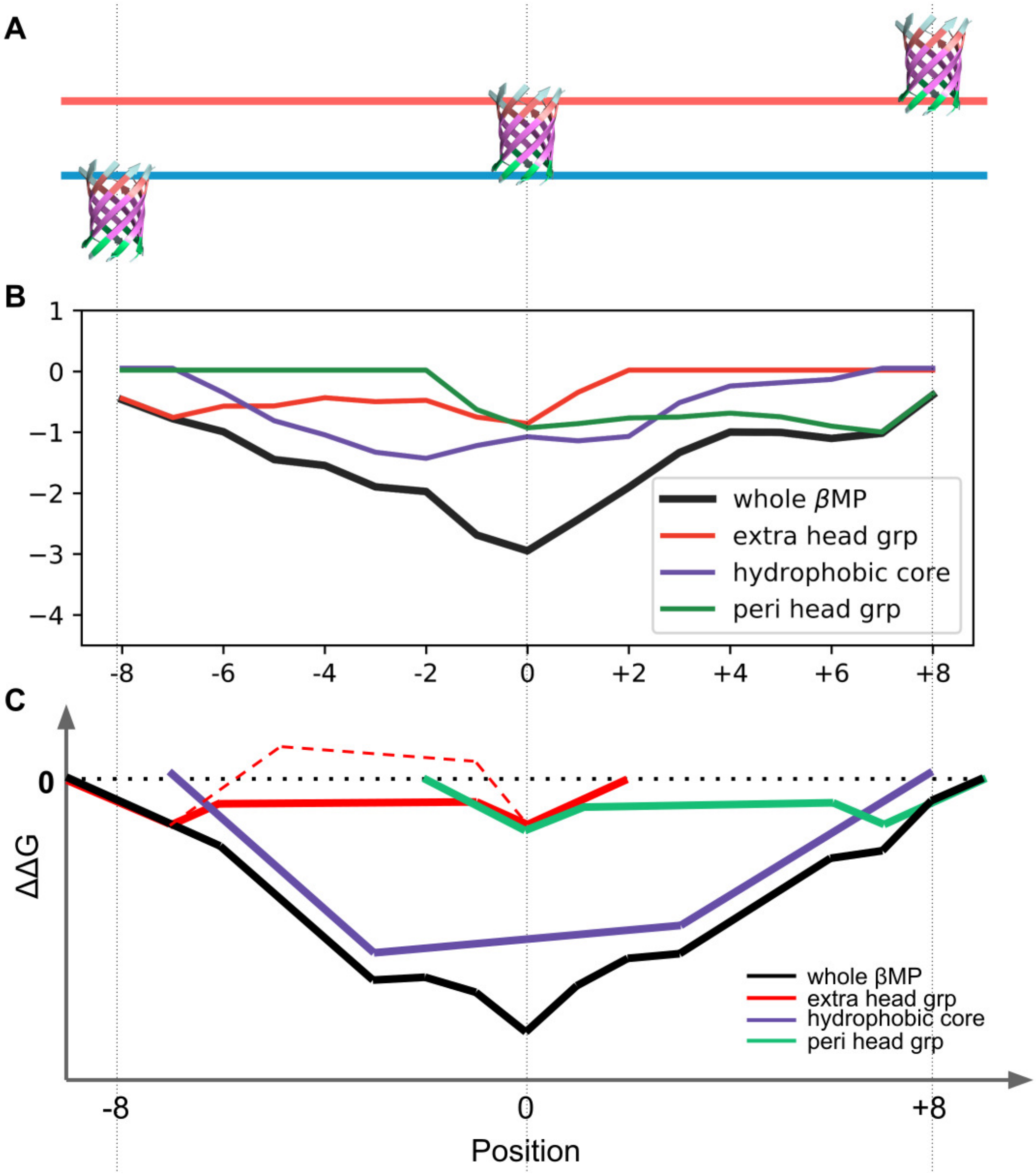


Figure 20: Illustration of the membrane insertion process of βMPs. **A.** The simplified βMP insertion model following Ref (7). **B.** The computed insertion energies of the OmpA protein at different depth positions. Contribution of different regions are color coded. **C.** Illustration on how free energies change with the position of βMP in the membrane. The dashed red segments show that lipid–facing residues in extracellular head group region sometimes become energetically unfavorable.

4.3.2.2 Importance of patterns of TM lipid–facing residues in membrane insertion

We then examine if the funnel–like insertion energy pattern arises from the extensive property of the TFEs of the hydrophobic residues alone. We considered only the 52 βMPs whose minimum free energies are at the fully inserted position. We first shuffled the sequences of the β–strands within the TM segment of each βMP. While the sidechain direction as well as the interstrand H–bond pairing at each residue position in β–strands are maintained, all TM residues are permuted. Each βMP is shuffled 2,000 times. We found that it is highly unfavorable to insert the shuffled βMPs into the membrane. This is expected, since the shuffling changes hydrophobicity of TM segments βMPs. Before the shuffling, the ionizable/polar residues were enriched among lumen–facing residues of βMPs, while lipid–facing residues were mostly apolar. After the shuffling, they were much evenly distributed.

We then investigate how insertion energy is affected if only the lipid–facing residues are shuffled. While the insertion of the shuffled βMPs remains energetically favorable (see Figure 21 for an example), shuffled βMPs are less stable compared to the original βMPs at the fully inserted position for 50 out of 52 βMPs: The insertion energy for the shuffled βMPs is on average 6.36 kcal/mol higher (Table XVIII). In addition, the fully inserted position (position 0) is no longer the most stable position for 17.4% of the shuffled βMPs (Table XVIII). These results indicate that the locational patterns of lipid–facing residues (22) in the TM region are optimized for βMPs to gain stability in the membrane environment.

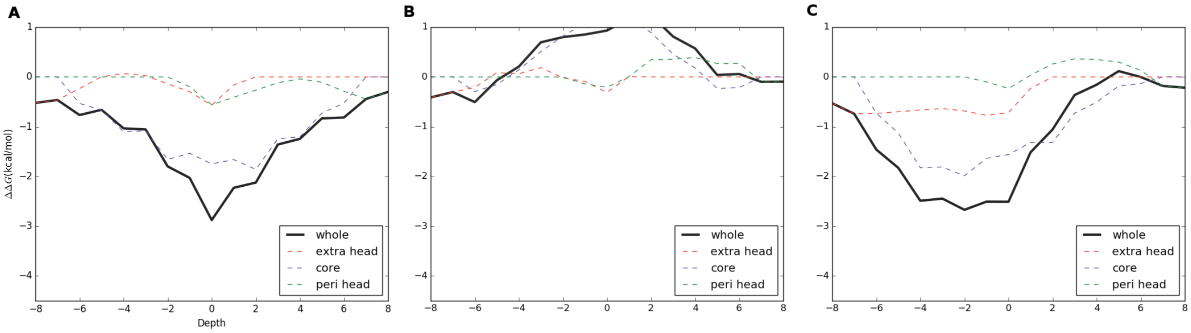


Figure 21: An example of the insertion TFEs (Omp32, PDB code:1e54). **A.** The insertion TFEs of the intact Omp32 shows a funnel pattern. **B.** The insertion TFEs of the residue–shuffled Omp32 regardless of sidechain directions. **C.** The insertion TFEs of the lipid–facing–residue–shuffled Omp32

4.3.2.3 Roles of residues in different TM regions during membrane insertion

The TM segment of a βMP can be divided into three regions, namely, the periplasmic headgroup region, the hydrophobic core region, and the extracellular headgroup region(3). We investigate how these regions contribute to the insertion energy of the βMP. We found that residues in the same regions across all 52 βMPs shared similar patterns in their insertion free energy profile (Figure 20C), indicating that they play similar roles in the insertion process. Among these, lipid–facing residues of the extracellular headgroup region facilitate the initialization of the insertion process, as they are energetically favorable in the interfacial region on the periplasmic side (position -8 and -7). As insertion proceeds, these residues become less favorable and occasionally unfavorable when they become more embedded in the membrane. At this time, lipid–facing residues of the hydrophobic core region start to be inserted in the membrane, and strongly drive the insertion process (position -6 to -2). When lipid–facing residues of the extracellular headgroup region approach the interfacial region of the extracellular side, they become energetically favorable again. At the same time, lipid–facing residues of the periplasmic headgroup region become inserted (position -1 and 0), and the TFE of the whole βMP reaches its minimum at position 0.

Although lipid–facing residues of the hydrophobic core region are known to provide the main driving force for membrane insertion of βMPs (42), we found that the TFEs of hydrophobic core region do not reach their minimum when βMPs are fully inserted at position 0 for all 52 βMPs. Upon incorporation of contributions from other regions, the overall TFEs of the whole βMPs indeed reach the minimum at the fully inserted position. The "W" shape of the free energy curves of the two head group regions (the red and green curve in Figure 20C) suggests that lipid–facing residues in these regions act like "energetic latches" to lock βMPs into their fully inserted position.

4.3.2.4 Prediction of βMP positioning and orientation in the membrane

GeTFEP can be used to predict positioning and orientation of βMPs in the membrane, similarly to previous studies (98; 87). Here, the membrane is idealized as an infinite slab with a thickness of h. Each βMP is initially positioned in the membrane with its center of mass of the barrel domain at the midplane of the membrane and its barrel axis aligned with the Normal direction (z–axis) of the membrane (Figure 22A). The protein can be rotated around the x– and y–axes with angles θ_x and θ_y, respectively. The two rotation angles together determine the tilt angle of the protein. The protein can also be translated with a displacement d_z. This

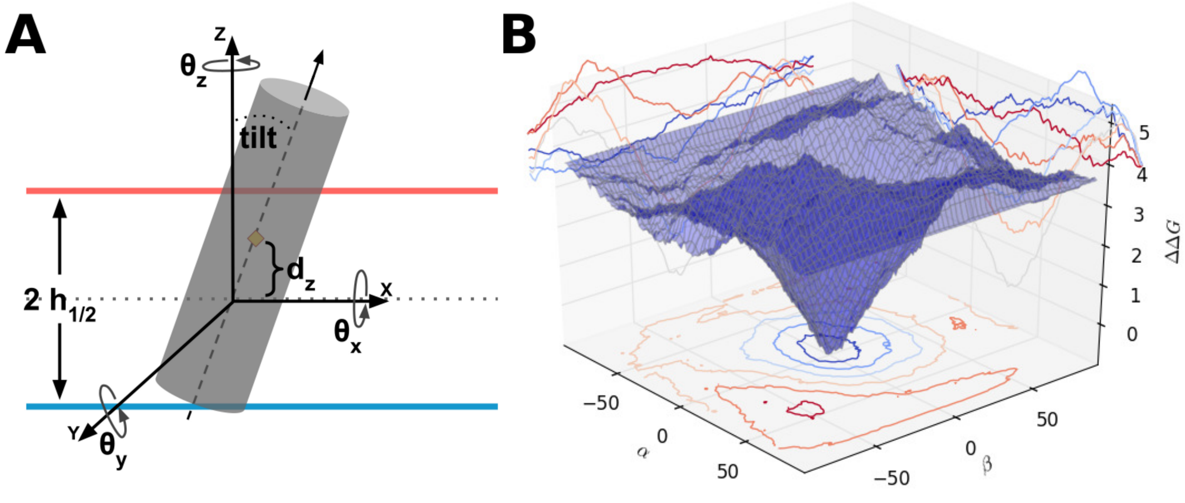


Figure 22: Prediction of positioning and orientation of βMPs. **A.** The positioning and orientation of the βMP inside the membrane are determine by the rotation angles θ_x and θ_y, the translation displacement d_z, and the membrane thickness h. **B.** The funnel–like landscape of the stability of the BtuB protein. It shows how rotation angles affect the stability of BtuB when d_z and h are fixed.

displacement and the membrane thickness determine the TM segment of the protein. When embedded in the membrane, the lipid–facing residues of the TM region and the loop residues are used to calculate the total energy of the βMP using the GeTFEP. As an example, Figure 22B shows how rotation angles θ_x and θ_y affect the stability of the protein BtuB (PDB code: 1nqe) when the displacement d_z and the membrane thickness h are fixed.

We systematically examine the parameter combination of θ_x, θ_y, d_z, and h. A βMP is predicted to take the position and the orientation when the lowest free energy is reached. The

TABLE XIII: Comparison between the predicted positioning and orientation and experimental results (2) of βMPs.

	Protein	PDB code	Experiment	GeTFEP	OPM
TM tilt (°)	FhuA	2fcp	46.0*	38.2	38.3
	OmpA	1bxw	44.5*	40.2	38.7
Membrane thickness (Å)	FhuA	1fep	≥ 23.1	23.5	24.3
	OmpF	2omf	~ 21.0	22.8	25.2
	BtuB	1nqe	≥ 20.2	23.0	23.4

* The experimentally measured tilt angles are the upper bounds of the actual values (2).

predicted protein tilt angles of all 58 βMPs correlate well ($r = 0.76$) with OPM records(92). The average protein tilt angle of 7.3° is consistent with that of $6.2 \pm 1.8°$ recorded in the OPM. The strand tilt angles and the membrane thickness predicted are again in good agreement with experimentally determined results (Table XIII).

4.3.3 Prediction of structurally and functionally important sites of βMPs

While overall the computed TFEs of lipid–facing residues of βMPs follow the general pattern of the GeTFEP, the TFE values of a specific residue in a particular βMP can deviate significantly from values in the general profile. For a lipid–facing residue in a βMP, we calculate the z–score of its TFE by $z = \frac{\text{TFE}-\mu}{\sigma}$, where μ and σ are respectively the mean and the standard deviation values in GeTFEP of the same amino acid in the same depth. We consider the deviation to be significant when $z > 1.64$ or < -1.64 (which correspond to 5% and 95% in the Normal distribution).

Among all 3,500 lipid–facing residues in the TM segments of all 58 βMPs, we find that 305 or 8.7% of the residues have TFE values deviate significantly from the GeTFEP. Since lipid–facing residues are overall the major contributors to the stability of βMPs as discussed above, the deviation from the general profile indicate that the residue is likely to have important roles other than providing stability. To understand the origin of these deviations, we examined three proteins in details, namely, OmpLA, PagP, and PagL, which have sufficient experimental information. We found that most deviant residues either have functional roles or have local structures quite different from residues in the canonical model of beta barrels (Table XIV).

Among the deviant residues in OmpLA, 142H and 156N are both in the catalytic triad (104; 105) that are essential for its phospholipase activities; 40L and 92Y are the sites where substrates bind (28); Furthermore, the deviant residue 116P interacts with 92Y and 142H through H–bonds. Among the deviant residues in PagP, 69L interacts with the out–clamp α–helix of PagP (106); 27I and 125L are both at the lateral routes where β–hydrogen bonding is absent (Figure 23), which ensure that substrates can access the protein interior so that PagP can carry out its enzymatic functions (11). In PagL, the deviant residue 108I is in the ligand binding site (107), and 126H is part of the catalytic triad of its enzymatic site (108).

As the calculation of TFEs does not require knowledge of 3D structures of βMPs, our results suggest that deviation analysis can help to discover functional sites and/or structurally anomalous sites using sequence information only. While our analysis is restricted to three proteins due to the limited nature of experiment data, we believe overall deviant residues play

TABLE XIV: Predicted important sites in OmpLA, PagP and PagL by deviation analysis

Protein	Residue	Notes
OmpLA(1qd6)	38N	
	40L	Substrate binding (28)
	92Y	Substrate binding (28)
	116P	Interstand neighbor of 92Y and 142H
	120L	
	142H	Catalytic site (104; 105)
	156N	Catalytic site (104; 105)
	237L	
PagP(1thq)	27I	Lateral route from membrane to protein interior (106)
	69L	Interact with the out–clamp α–helix (106)
	125L	Lateral route from membrane to protein interior (106)
	131L	
PagL(2erv)	108I	Ligand binding site (107)
	126H	Catalytic site (108)

special roles in either performing biological function or in maintaining the unique structural form of βMPs.

4.3.4 GeTFEP can predict TM region of α–helical membrane proteins

Although the MF–scale was measured in the βMP system, it was suggested that the scale is also applicable to TM region of αMPs, since the MF–scale has a strong correlation with the nonpolar solvent accessible surface areas of the residues (1). We hypothesize that the GeTFEP may also reflects fundamental thermodynamic properties of transferring sidechains of amino acids to the membrane environment, regardless whether the residue is in a β–barrel or a α–helical membrane protein. We carried out the standard hydropathy analysis (40) using the

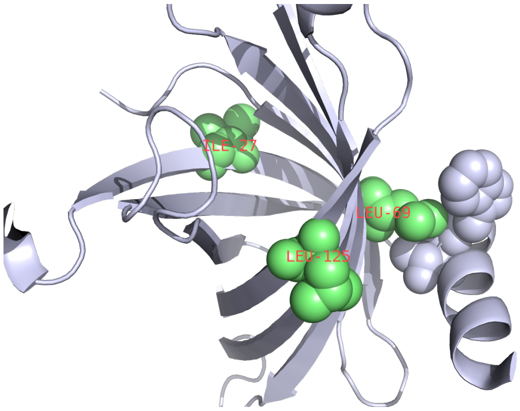


Figure 23: Predicted important sites of PagP. Residues 27I and 125L are at the sites where the H–bonds between the β–strands are disrupt. 69L has interaction with the out–clamp α–helix of PagP.

Membrane Protein Explorer (MPEx) program (8) on 131 αMPs obtained from the MPTopo database(109). Since MPEx uses depth–independent hydrophobicity scales, we used the mid–GeTFEP scale for our calculation.

The results show that this simple analysis correctly predicts both the TM regions and the numbers of the TM segments for 90 or ∼69% of the 131 αMPs in the dataset (see Figure 27A for an example). This compares favorably to other hydrophobicity scales, including those measured or derived from αMPs (Table XV). For most of the remaining 41 proteins, GeTFEP correctly predicted the TM regions, but predicted the numbers of the TM segments incorrectly due to the ambiguity in assignment of whether two consecutive TM segments should be considered as one TM segment (see Figure 27B for an example). Examination of the number of TM residues correctly predicted by the mid–GeTFEP scale show that we achieves a precision of ∼85% and a recall of ∼71%, which compares favorably to other hydrophobicity scales (Table XV). These results suggest that the GeTFEP reflects fundamental thermodynamic properties of amino acid residues inside membrane, and can be used to study the general stability of both α–helical and β–barrel membrane proteins.

4.3.4.1 The validity of transfer free energy value of Pro in the GeTFEP

We further examine the TFE value of Pro in the mid-GeTFEP scale, which is qualitatively different from that in the MF–scale. We swapped the value of Pro from MF–scale into the mid–GeTFEP scale, and used this Pro–swapped scale in the hydropathy analysis. This is reasonable as the mid–GeTFEP scale is strongly correlated with the MF–scale, and has comparable values. However, we found that the precision of predicting TM residues deteriorates significantly from 85% using the mid–GeTFEP scale to 72% using the Pro–swapped scale (Table XV). This result suggests that Pro is more likely to be membrane unfavorable as characterized by the mid–GeTFEP scale rather than membrane favorable as characterized by the MF–scale.

4.4 Conclusions and discussion

In this study, we derived the General Transfer Free Energy Profile (GeTFEP) from a non–redundant set of 58 βMPs. We showed that the GeTFEP agrees well with previous experimentally measured and computationally derived TFEs. The GeTFEP reveals fundamental thermodynamic properties of amino acid residues inside membrane environment, and it is useful in analysis of stability and function of membrane proteins (110).

As the lipid membrane bilayer is anisotropic along the bilayer normal (111), a residue at different depth of the membrane will have different interaction with lipid molecules in the

TABLE XV: Prediction of TM segments and residues αMPs. The mid-GeTFEP scale performs better than the other hydrophobicity scales. The first three scales are measured or derived in α–helical systems, the others in βMPs.

Hydrophobicity scale	αMPs % (#) with TM segs. correctly predicted	TM res. precision	TM res. recall	TM res. F–measure
WW–scale	50%(66)	73%	75%	0.74
Bio–scale	22%(29)	95%	21%	0.34
$E_Z\alpha$	49%(64)	71%	77%	0.74
MF–scale	48%(63)	77%	65%	0.70
Pro–swapped	49%(64)	72%	74%	0.71
mid–GeTFEP	69%(90)	85%	71%	0.78

environment, resulting in the depth–dependency of TFEs. However, there are few experimental measurements of TFEs at different depth positions other than the hydrophobic core, except Arg and Leu (1). Comparison between the GeTFEP and the experimentally measured values of Arg and Leu shows that the GeTFEP captures this depth–dependency well.

In addition, the GeTFEP exhibits asymmetric values between TFEs of residues in the membrane inner leaflet (depth -4 to 0) and in the outer leaflet (depth 0 to +4, Figure 18C). Most βMPs in our dataset resides in the bacterial outer membrane, whose outer leaflet contains additional complex lipolysaccharides in contrast to its inner leaflet of phospholipids. This asymmetry in membrane composition results in the asymmetry of the TFEs in the GeTFEP. To understand membrane proteins in an environment of symmetric membrane leaflets, we also derived a symmetric TFE profile, named sym–GeTFEP, by mirroring the TFE values of the inner leaflet side of the GeTFEP (Figure 28). In this study, the sym–GeTFEP was used to analyze the non–outer–membrane βMPs, e.g. α– and γ–hemolysins and vibrio cholerae cytolysin.

We explored the energetic contribution of different regions of βMPs during the membrane insertion process. Our analysis showed that the stability of βMPs does not come alone from the extensive property of the hydrophobicity of lipid–facing residues in the TM segment. Rather, the pattern of the amino acid residues in the TM segment also play significant roles. Results from analysis of sequence shuffling show that the patterns and location of amino acid residues are optimized to stabilize βMPs in the membrane environment. Using the GeTFEP, we are also able to predict membrane positioning and orientations of βMPs.

The GeTFEP can also be used to detect structurally or functionally important residues in βMPs. This can be achieved by examination of residues whose TFEs deviate significantly from the GeTFEP. As calculation of TFEs of residues of a specific βMP only requires rough estimation of relative positions between adjacent β–strands, which can be reliably predicted from the protein sequence (61; 76), computing the TFE deviation therefore requires only sequence information. The GeTFEP–deviation analysis can aid in discovery of functional sites or structurally important sites in novel βMPs, without requiring knowledge of their 3D structures. In addition, GeTFEP–based analysis can aid in design and engineering of novel βMPs.

Furthermore, we demonstrated that GeTFEP can be used to predict TM residues of αMPs. Results showed that GeTFEP performs better than the hydrophobicity scales measured/calculated in αMP systems, suggesting that the GeTFEP reflects fundamental thermodynamic properties of amino acid residues inside membrane, and can be used to study the general stability of both α–helical and β–barrel membrane proteins.

CHAPTER 5

CONCLUSION

In this thesis, we have developed computational methods to predict 3D structures of βMPs and to calculate transfer free energies of residues in βMPs. In addition, we have dervied a general transfer free energy profile based on the systematical calculation of TFEs of βMPs.

5.1 3D structure prediction of βMPs

We predicted interstrand interaction between β–strands of βMPs using a combination of an empirical energy function and information from sequence covariation analysis. We are able to predict the strand registers at an accuray of 85%, which is a big improvement from previous studies. We then introduced a global shear optimization scheme to adjust the registers predicted from local information. This optimization step help to improve the accuracy of further 3D structure construction. We also developed a parametric structural template named *intertwined zigzag coil model*, which captures major geometric properties of the barrel domains of βMPs, to construct their 3D structures. In a blind test of 51 nonhomologous βMPs, including proteins for which no prediction has been attempted before, our method generates accurate 3D structures of TM regions with an average main-chain rmsd of 3.48Å. which is a significant improvement over previous studies. In addition, predictions are expanded to include non-TM regions, including both extended β–sheets and loops, resulting in over 30% increase in the coverage of residues compared with previous methods.

5.2 Transfer free energy computation

In this study, we improved a method for calculation of transfer free energy of βMPs (42). Although the original method can accurately calculate TFEs, its application is limited to only small βMPs due to its computational complexity. We have introduced several approximation schemes, resulting in large reducing in time cost of TFE computation with little loss of the computational accuracy. The new method is efficient and applicable to all bacterial TMBs regardless of the size of the proteins.

5.3 GeTFEP: General Transfer Free Energy Profile for membrane proteins

In this study, we derived a TFE profile named General Transfer Free Energy Profile (GeTFEP) based on systematical computation of the TFEs of 58 βMPs. The GeTFEP agrees well with experimentally measured and computationally derived TFEs. Analysis based on the GeTFEP shows that residues in different regions of the TM segments of βMPs have different roles during the membrane insertion process. Results further reveal the importance of the sequence pattern of transmembrane strands in stabilizing βMPs in the membrane environment. In addition, we show that GeTFEP can be used to predict the positioning and the orientation of βMPs in the membrane. We also show that GeTFEP can be used to identify structurally or functionally important amino acid residue sites of βMPs. Furthermore, the TM segments of α–helical membrane proteins can be accurately predicted with GeTFEP, suggesting that the GeTFEP captures fundamental thermodynamic properties of amino acid residues inside membrane, and is of general applicability in studying membrane proteins.

APPENDICES

TABLE XVI: Dataset and prediction results. Strand register prediction and RMSD between the TM region and the TM+extended barrel regions of real and modeled structures of 51 non–homologous βMPs. (Continue on the next page)

Protein/PDB	Organism	Strands #	Correct register # before/after optimization	Shear # before/after optimization	TM domian C_α-RMSD		Barrel domain C_α-RMSD	
					Main chain	All atom	Main chain	All atom
OmpA/1bxw	*E. coli*	8	6/8	6/10/10	1.39	2.83	1.46	2.86
NspA/1p4t	*N. meningitidis*	8	7/8	8/10/10	1.45	2.50	1.83	2.95
OmpX/1qj8	*E. coli*	8	7/7	6/10/8	2.63	3.47	3.01	3.94
PagP/1thq	*E. coli*	8	5/6	5/10/10	3.35	4.25	3.35	4.25
PagL/2erv	*P. aeruginosa*	8	6/4	4/10/10	3.94	4.47	3.94	4.47
OmpW/2f1t	*E. coli*	8	7/6	12/10/10	3.12	4.00	3.20	4.22
OprH/2lhf	*P. aeruginosa*	8	7/8	12/10/10	1.49	2.43	1.49	2.42
Opa60/2mlh	*N. gonorrhoeae*	8	8/8	10/10/10	1.49	2.69	1.49	2.69
HB27/3dzm	*T. thermophilus*	8	6/6	4/10/10	2.85	3.41	3.00	3.65
OmpT/1i78	*E. coli*	10	9/8	14/12/12	3.69	4.53	4.84	5.86
OpcA/1k24	*N. meningitidis*	10	6/3	18/12/12	4.79	5.31	4.84	5.49
OmpLA/1qd6	*E. coli*	12	9/8	16/14/16	5.55	6.68	5.71	6.86
Txs/1tly	*E. coli*	12	7/7	18/14/16	4.71	5.57	4.72	5.59
NalP/1uyn	*N. meningitidis*	12	11/12	12/14/14	1.45	2.88	1.56	3.06
NanC/2wjr	*E. coli*	12	11/10	12/14/14	2.94	3.54	2.94	3.54
Hbp/3aeh	*E. coli*	12	12/12	14/14/14	1.78	3.02	1.80	2.97
LpxR/3fid	*S. enterica*	12	8/8	14/14/14	6.56	6.93	6.58	7.09
EstA/3kvn	*P. aeruginosa*	12	12/12	14/14/14	1.61	2.90	2.03	3.24
intimin/4e1s	*E. coli*	12	9/10	16/14/14	3.10	3.90	3.10	3.89
KdgM/4pr7	*D. dadantii*	12	9/9	14/14/14	3.91	4.52	3.99	4.55
FadL/1t16	*E. coli*	14	13/12	12/14/14	2.23	3.10	2.84	3.73
OmpG/2f1c	*E. coli*	14	10/9	12/14/16	3.09	3.96	3.15	4.10
TodX/3bs0	*P. putida*	14	14/14	14/14/14	1.30	2.13	2.01	2.97
FadL/3dwo	*P. aeruginosa*	14	12/10	10/14/14	3.15	3.76	3.82	4.42
Omp32/1e54	*C. acidovorans*	16	14/14	16/20/20	3.03	3.63	3.10	3.67
Porin/1prn	*R. balistica*	16	14/14	20/20/20	2.44	3.28	2.44	3.27
Porin P/2o4v	*P. aeruginosa*	16	15/14	18/20/20	3.28	4.26	3.58	4.44
OmpF/2omf	*E. coli*	16	15/16	18/20/20	2.60	3.90	2.79	4.01
Porin/2por	*R. capsulatus*	16	13/12	18/20/20	2.77	3.36	2.81	3.39

TABLE XVI: (Cont'd) Dataset and prediction results. Strand register prediction and RMSD between the TM region and the TM+extended barrel regions of real and modeled structures of 51 non–homologous βMPs.

Protein/PDB	Organism	Strands #	Correct register # before/after optimization	Shear # before/after optimization	TM domian C$_\alpha$-RMSD		Barrel domain C$_\alpha$-RMSD	
					Main chain	All atom	Main chain	All atom
FhaC/2qdz	*B. pertussis*	16	12/11	23/21/20	6.18	6.68	5.97	6.49
PorB/3vzt	*N. meningitidis*	16	14/14	16/20/20	3.48	4.08	3.75	4.52
TamA/4c00	*E. coli*	16	12/13	18/20/20	4.68	5.12	4.78	5.21
OprB/4gey	*P. putida*	16	12/11	10/20/20	4.64	5.22	4.69	5.31
BamA/4n75	*E. coli*	16	14/15	18/20/22	3.44	4.07	3.53	4.23
ScrY/1a0s	*S. typhimurium*	18	15/15	25/25/20	5.14	5.60	5.17	5.64
LamB/2mpr	*S. typhimurium*	18	13/10	32/22/22	6.08	6.86	7.65	8.27
Wzi/2ynk	*E. coli*	18	14/13	20/22/20	3.61	4.42	3.92	4.70
AlgE/3rbh	*P. aeruginosa*	18	13/16	16/22/22	4.36	5.30	4.30	5.20
OpdP/3syb	*P. aeruginosa*	18	16/15	18/19/22	3.73	4.41	3.72	4.46
OpdO/3szv	*P. aeruginosa*	18	17/16	20/22/22	3.20	3.89	3.17	3.87
VADC1/3emn	*M. musculus*	19	10/10	19/19/20	3.53	4.34	3.53	4.34
FepA/1fep	*E. coli*	22	21/19	29/25/24	4.51	4.96	5.14	5.67
FecA/1kmo	*E. coli*	22	22/22	24/24/24	2.71	3.47	3.19	3.94
BtuB/1nqe	*E. coli*	22	21/20	22/24/24	2.84	3.60	3.53	4.26
FptA/1xkw	*P. aeruginosa*	22	22/22	24/24/24	3.29	3.84	3.88	4.34
FhuA/2fcp	*E. coli*	22	22/22	24/24/24	2.83	3.41	5.20	5.57
HasR/3csl	*S. marcescens*	22	22/22	24/24/24	2.71	3.35	3.16	3.89
TbpA/3v8x	*N. meningitidis*	22	21/20	26/24/24	2.58	3.68	4.96	5.74
PapC/2vqi	*E. coli*	24	22/22	23/26/26	6.06	6.62	6.42	7.02
FimD/3rfz	*E. coli*	24	20/21	30/31/26	4.74	5.47	4.79	5.60
LptD/4q35	*S. flexneri*	26	18/16	37/32/30	7.25	7.67	7.53	7.96

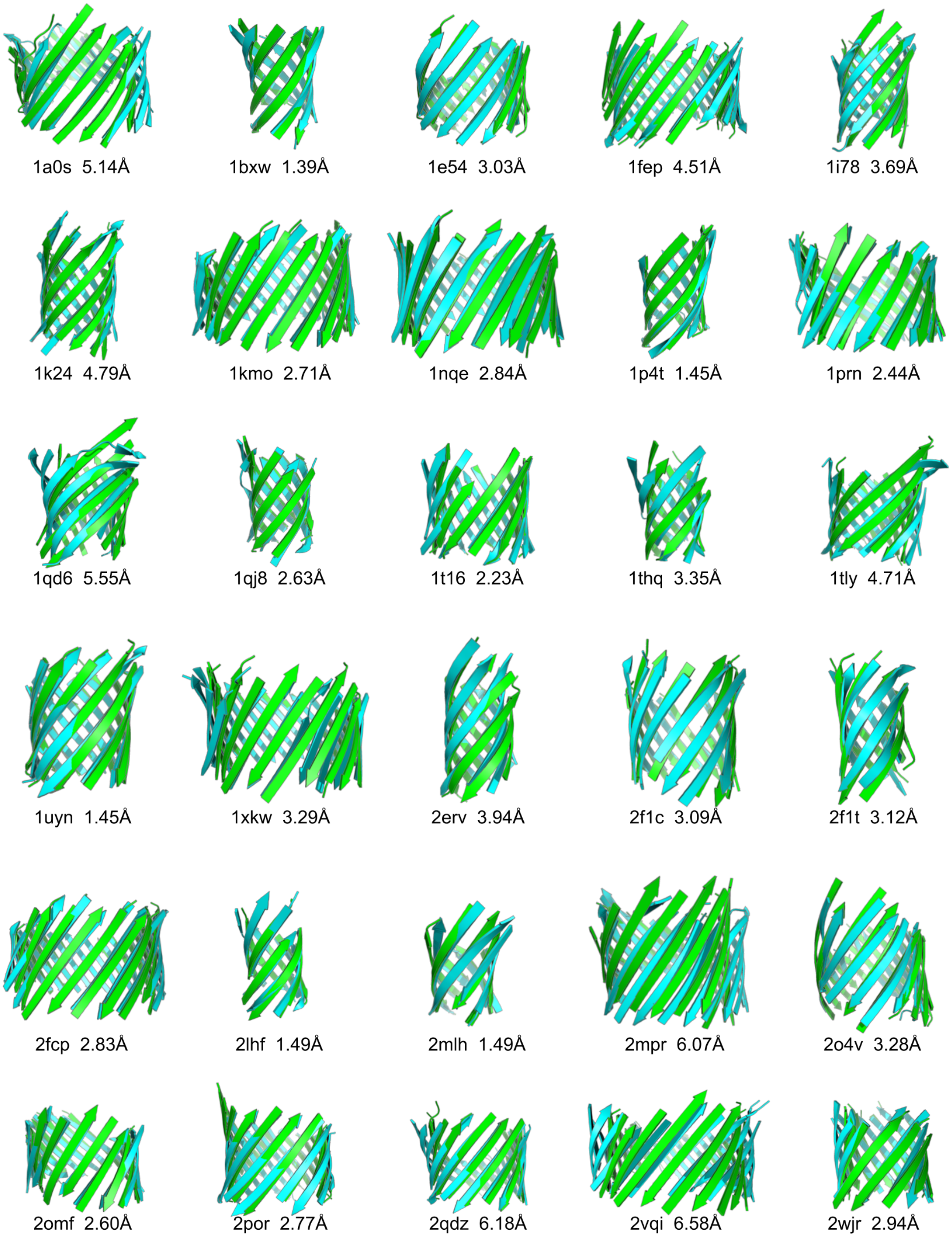


Figure 24: Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan). (Continue on the next page)

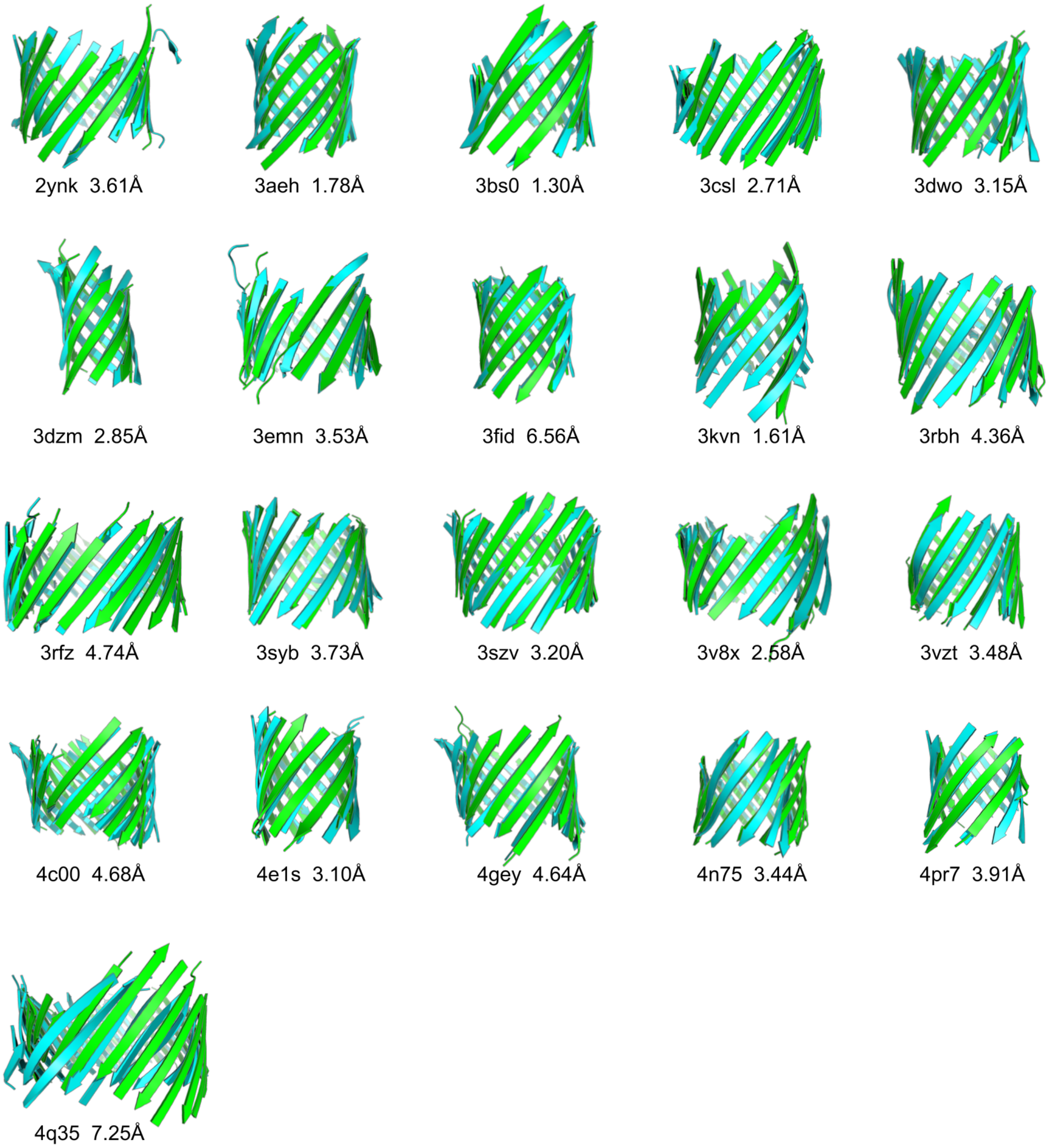


Figure 24: (Cont'd) Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan).

Algorithm 4: Exact algorithm with the histogram scheme. This algorithm is compared with the exact algorithm without the histogram scheme (Algorithm 1 in the main text) to determine the appropriate bin size.

```
▷ enumerate strand--triplet conformations and precompute energies of strands
for $i \leftarrow 1$ to $n$ do
    $\min_{E_i} \leftarrow +inf$;
    $\max_{E_i} \leftarrow -inf$;
    foreach $(d_{i-1}, d_i, d_{i+1}) \in (-l, \cdots, l)^3$ do
        $E(i; d_{i-1}, d_i, d_{i+1}) \leftarrow$ energy of the middle strand;
        $\min_{E_i} \leftarrow min(\min_{E_i}, E(i; d_{i-1}, d_i, d_{i+1}))$;
        $\max_{E_i} \leftarrow max(\max_{E_i}, E(i; d_{i-1}, d_i, d_{i+1}))$;
    end
end
▷ estimate the ranges of the energies of the TM region
$[\min_E, \max_E] \leftarrow [\sum_{i=1}^{n} \min_{E_i}, \sum_{i=1}^{n} \max_{E_i}]$;
$LB \leftarrow$ new bin list covers range $[min_E, max_E]$;
$Z \leftarrow 0$;
▷ compute the partition function
foreach $\mathbf{d} = (d_1, \cdots, d_n) \in (-l, \cdots, l)^n$ do
    $E \leftarrow \sum_{i=1}^{n} E(i; d_{i-1}, d_i, d_{i+1})$;
    from $LB$ get bin $b$ corresponding to $E$;
    $b \leftarrow b + 1$;
end
foreach $b \in LB$ do
    $E \leftarrow$ the energy value corresponding to bin $b$;
    $Z \leftarrow Z + b \cdot \exp(-\beta E)$;
end
return $Z$;
```

TABLE XVII: Examples of the 20D TFE vectors calculated using our methods. The RMSEs between the results calculated from the approximation algorithm (Algorithm 3) and the exact algorithm (Algorithm 1) are listed in the last row. The average RMSEs in Table X and Table XI in the main text are computed from this kind of vectors of corresponding residues.

seqid	210		120	
Algorithm	1	3	1	3
A	0.0000	0.0000	0.0000	0.0000
R	3.2799	3.2081	2.0733	2.0717
N	3.2325	3.1320	1.8049	1.8016
D	4.1038	3.7979	1.6017	1.6030
C	0.3648	0.3667	-0.5673	-0.5664
Q	2.0878	2.0105	1.6194	1.6192
E	2.7305	2.6250	1.6726	1.6726
G	1.1782	1.1580	0.2463	0.2467
H	3.5356	2.8931	-0.3928	-0.3904
I	-1.4800	-1.4935	-2.3433	-2.3410
L	-2.2877	-2.2880	-3.1131	-3.1105
K	3.8210	3.7209	4.0928	4.0737
M	-0.6982	-0.6919	-0.9090	-0.9246
F	-1.8795	-1.8751	-1.5421	-1.5400
P	1.5798	1.4134	2.9298	2.9030
S	1.6882	1.6822	1.6991	1.6942
T	1.0420	1.0432	1.0402	1.0412
W	-0.2927	-0.3090	0.3138	0.3139
Y	-0.6041	-0.5941	-0.3928	-0.3922
V	-1.4205	-1.4311	-1.6605	-1.6590
RMSE		0.1699		0.0083

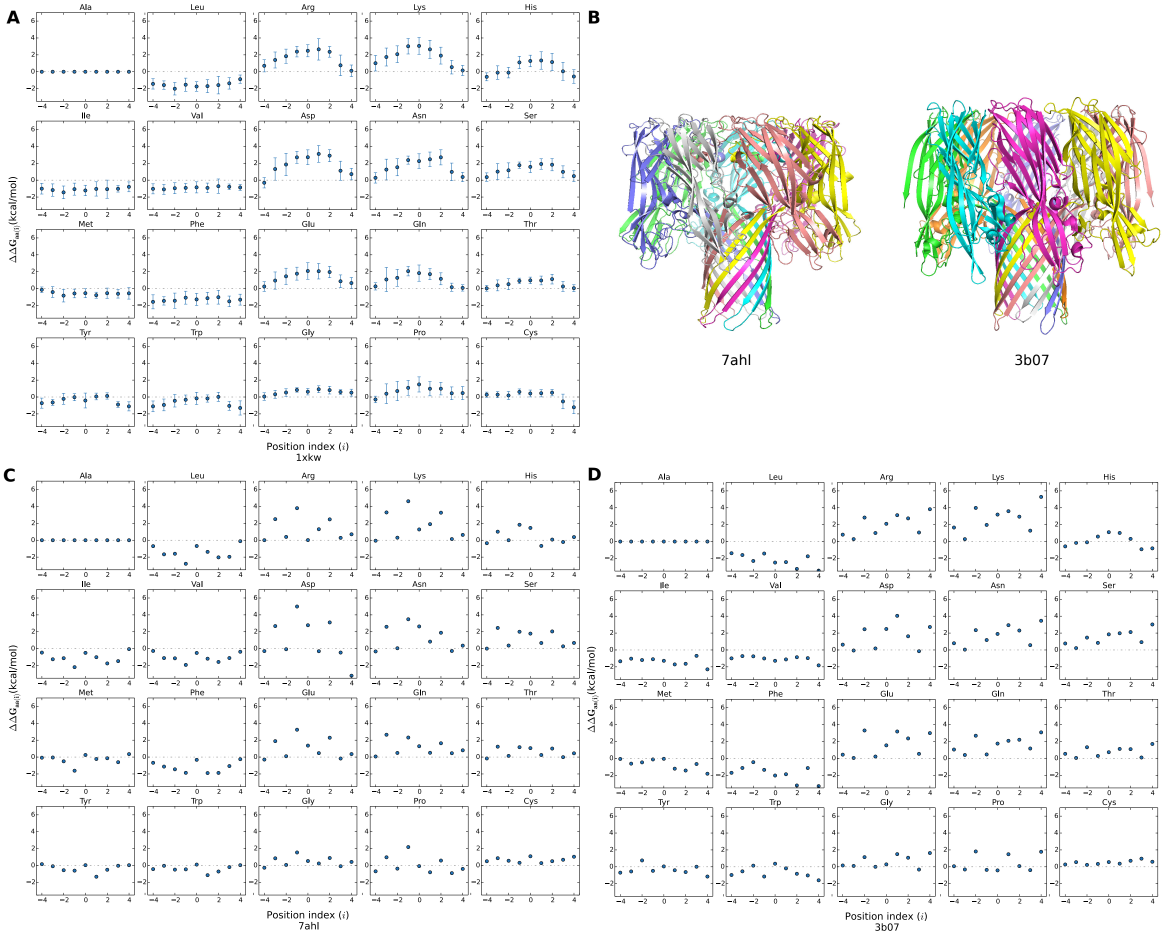


Figure 25: Structures and TFE profiles of α– and γ–hemolysin. **A.** A typical TFE profile of βMPs (FptA, PDB code: 1xkw). **B.** The structures of α–hemolysin (PDB code: 7ahl) and γ–hemolysin (PDB code: 3b07) Both TM segments are constructed with repeated β–hairpins. **C.** The TFE profile of α–hemolysin. **D.** The TFE profile of γ–hemolysin. Since the structures are both repeated hairpin, there is only one data point for each amino acid residue in every depth of their profiles.

TABLE XVIII: The insertion TFEs of WT and lipid–facing residue–shuffled βMPs calculated with GeTFEP. The $\Delta\Delta G$ shows the differences between TFEs of the WT βMPs at step 0 and the average of the minimum TFEs of the lipid–facing residue–shuffled βMPs. (Continue on the next page)

PDB code	Position w/ min $\Delta\Delta G$	min $\Delta\Delta G$	mis–insertion # (of 2,000)	$\Delta\Delta\Delta G$
1bxw	0	-23.58	616	1.63
1e54	0	-46.03	66	7.90
1ek9	0	-58.23	403	10.35
1fep	0	-61.58	35	8.53
1i78	0	-30.88	216	4.58
1k24	0	-33.41	163	1.99
1kmo	0	-60.55	16	7.95
1nqe	0	-49.83	91	6.18
1p4t	0	-25.68	282	3.61
1prn	0	-43.02	959	9.74
1qd6	0	-34.92	559	7.02
1qj8	0	-21.53	196	4.15
1t16	0	-45.64	08	9.88
1thq	0	-23.10	253	5.52
1tly	0	-28.58	415	2.42
1uyn	0	-33.95	128	4.81
1xkw	0	-59.91	133	9.54
2erv	0	-24.56	398	4.54
2f1c	0	-43.85	473	0.14
2f1t	0	-23.14	280	3.00
2fcp	0	-52.14	717	6.73
2lhf	0	-16.19	757	2.89
2lme	0	-34.17	525	-2.64
2mlh	0	-26.15	676	1.55
2mpr	0	-36.30	688	9.95
2o4v	0	-48.16	190	7.11
2omf	0	-37.71	98	8.65
2por	0	-32.70	994	8.12
2qdz	0	-52.23	289	10.44
2vqi	0	-43.92	441	15.28
2wjr	0	-33.58	72	5.30
2ynk	0	-44.68	141	7.75
3aeh	0	-25.59	471	6.23
3b07	0	-5.14	1670	0.67
3bs0	0	-47.60	95	6.18

TABLE XVIII: (Cont'd) The insertion TFEs of WT and lipid–facing residue–shuffled βMPs calculated with GeTFEP. The $\Delta\Delta G$ shows the differences between TFEs of the WT βMPs at step 0 and the average of the minimum TFEs of the lipid–facing residue–shuffled βMPs.

PDB code	Position w/ min $\Delta\Delta G$	min $\Delta\Delta G$	mis–insertion # (of 2,000)	$\Delta\Delta\Delta G$
3csl	0	-70.94	11	9.55
3dwo	0	-48.32	41	9.17
3dzm	0	-33.27	86	1.26
3kvn	0	-36.38	116	6.60
3pik	0	-36.54	779	4.83
3rbh	0	-41.60	162	6.78
3syb	0	-45.97	236	10.70
3szv	0	-55.25	129	12.65
3v8x	0	-58.94	68	13.24
3vzt	0	-25.56	1018	4.71
4c00	0	-36.32	147	6.76
4e1s	0	-36.20	112	7.20
4gey	0	-49.42	164	8.25
4k3c	0	-48.53	92	6.36
4pr7	0	-38.13	104	3.15
4q35	0	-55.56	97	12.52
7ahl	0	-3.93	1246	-0.49
Summary			17.4%	6.36 ± 3.70

TABLE XIX: **T**he computed insertion TFEs of the 6 βMPs that do not have the minimum energy at position 0. However, their most stable position is close to 0, and the minimum TFEs are close to their TFEs at position 0. Nonetheless, we exclude these 6 βMPs in our other analysis of membrane insertion stability.

PDB code	Position w/ min $\Delta\Delta G$	min $\Delta\Delta G$	Position 0 $\Delta\Delta G$
1a0s	1	-29.77	-30.92
1yc9	-1	-48.40	-49.94
2gr8	-1	-22.74	-24.37
3fid	-1	-35.96	-36.37
3o44	1	-23.77	-29.70
3rfz	-1	-43.08	-43.46

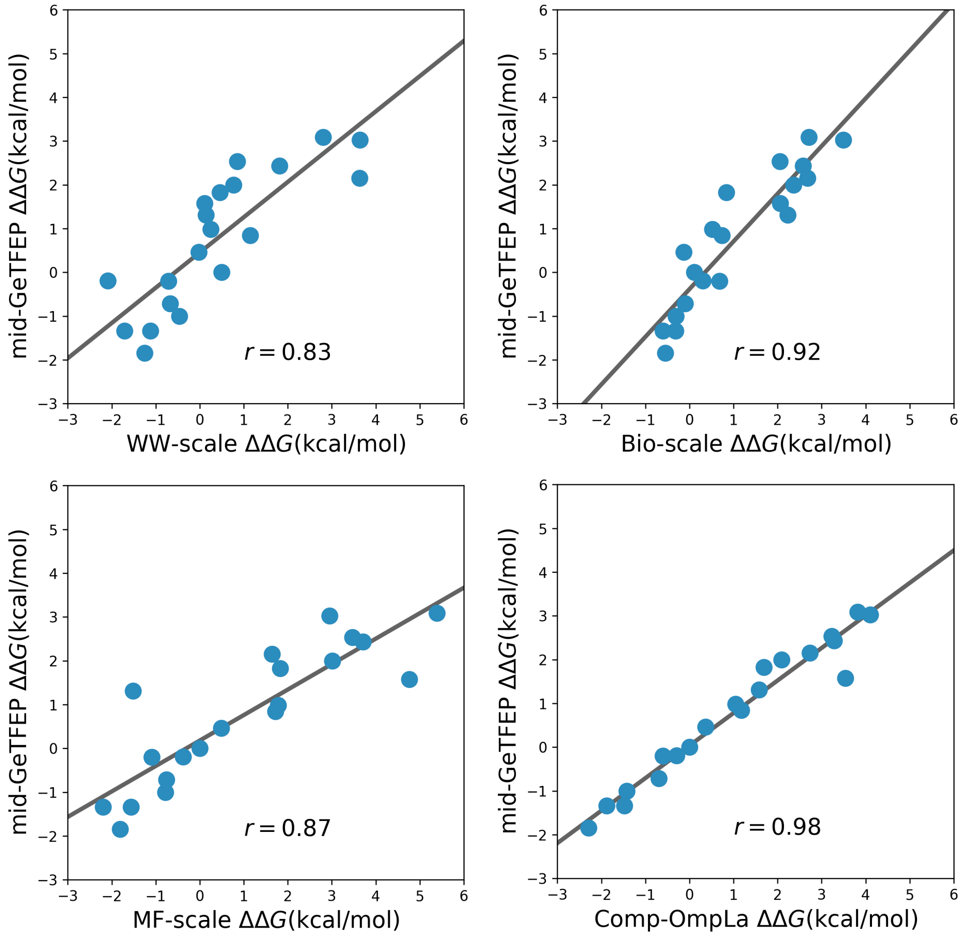


Figure 26: Comparison between the mid–GeTFEP scale and other hydrophobicity scales. The mid-GeTFEP scale agrees well with previously measured or derived hydrophobicity scales.

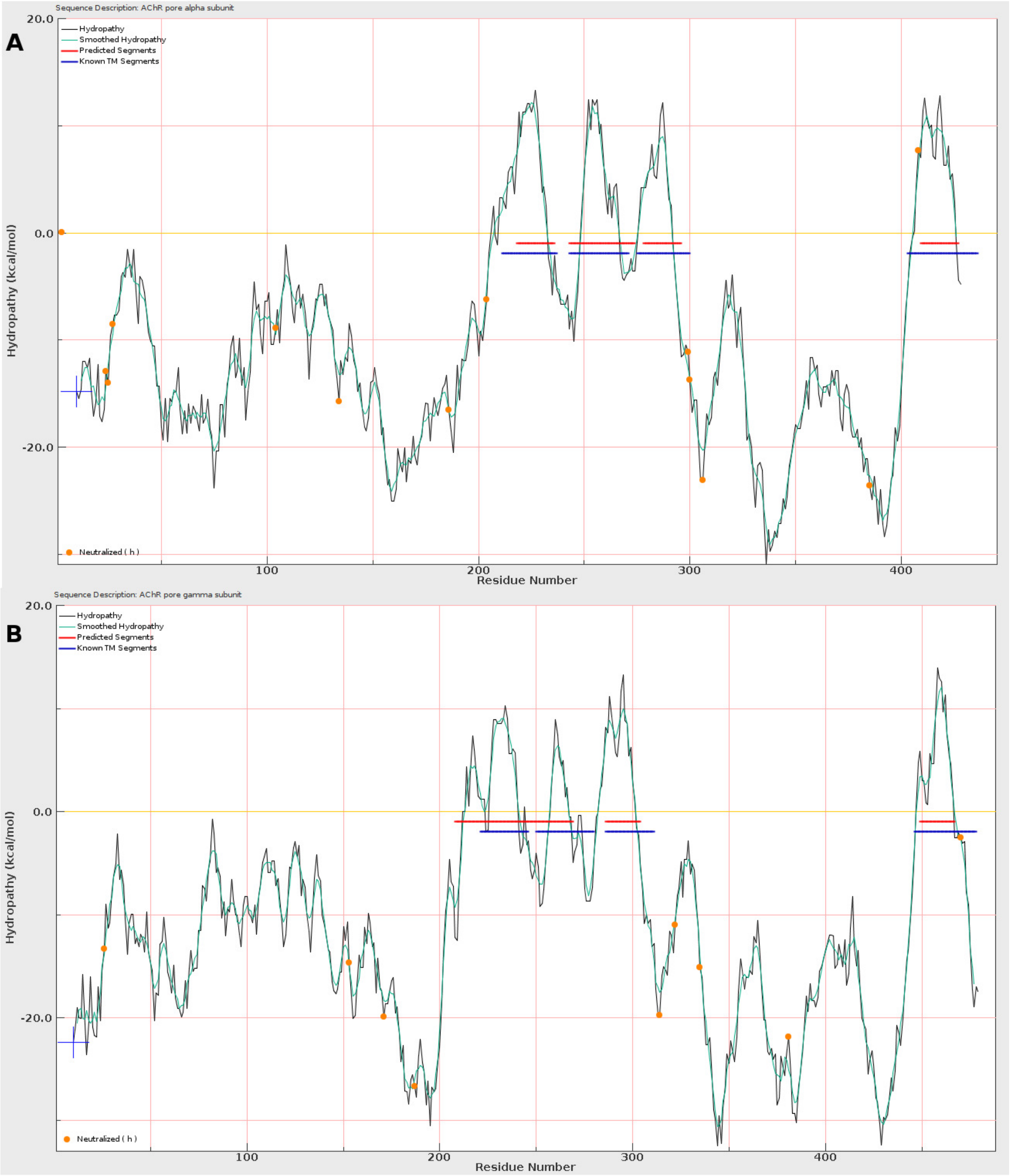


Figure 27: Hydropathy analysis with mid–GeTFEP. The blue segments are the known TM segments, while the red ones are predicted by the hydropathy analysis. The analysis was carried out using Membrane Protein Explorer (MPEx) (8) **A**. An example (AChR pore α subunit) shows both the TM region and the number of the TM segments are correctly predicted. **B**. An example (AChR pore γ subunit) shows the predicted number of the TM segments are wrong, though the TM regions are correctly predicted.

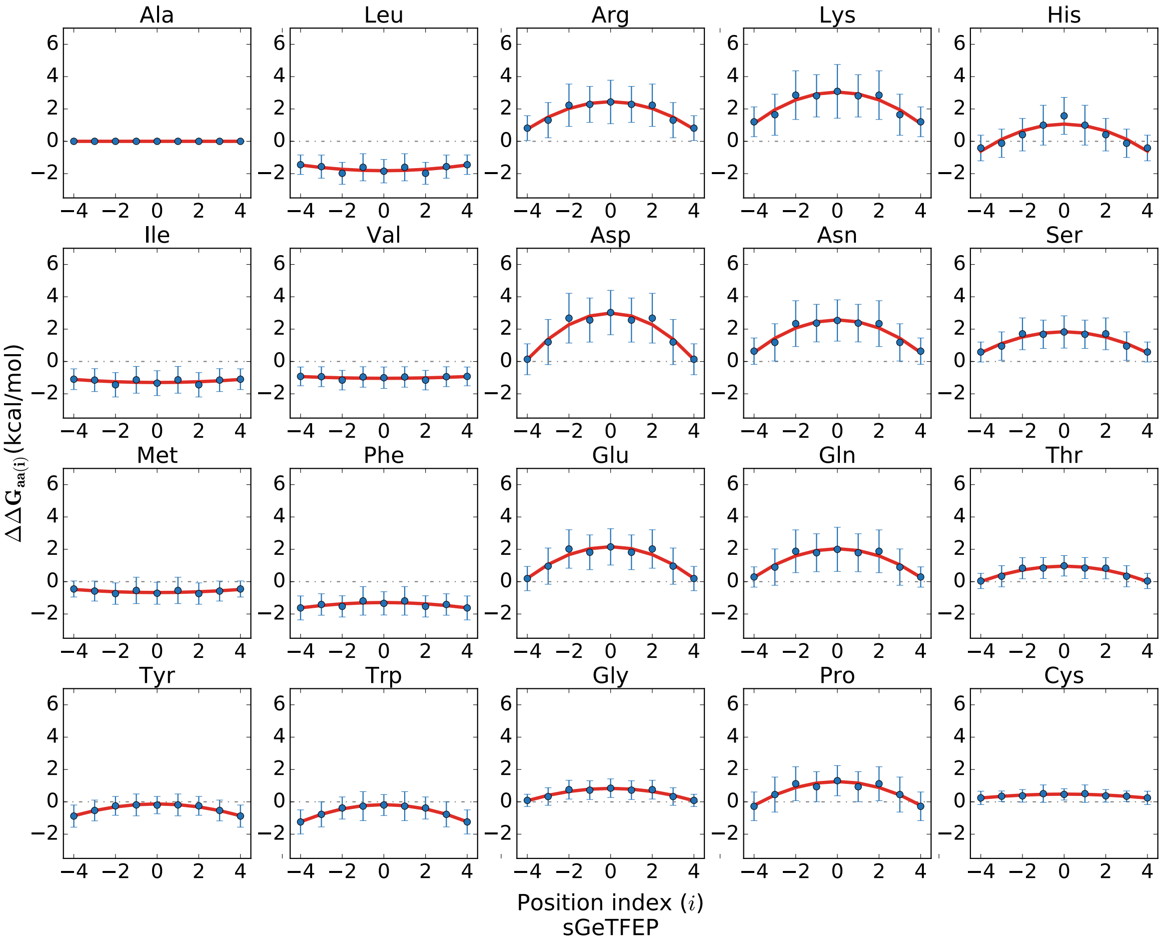
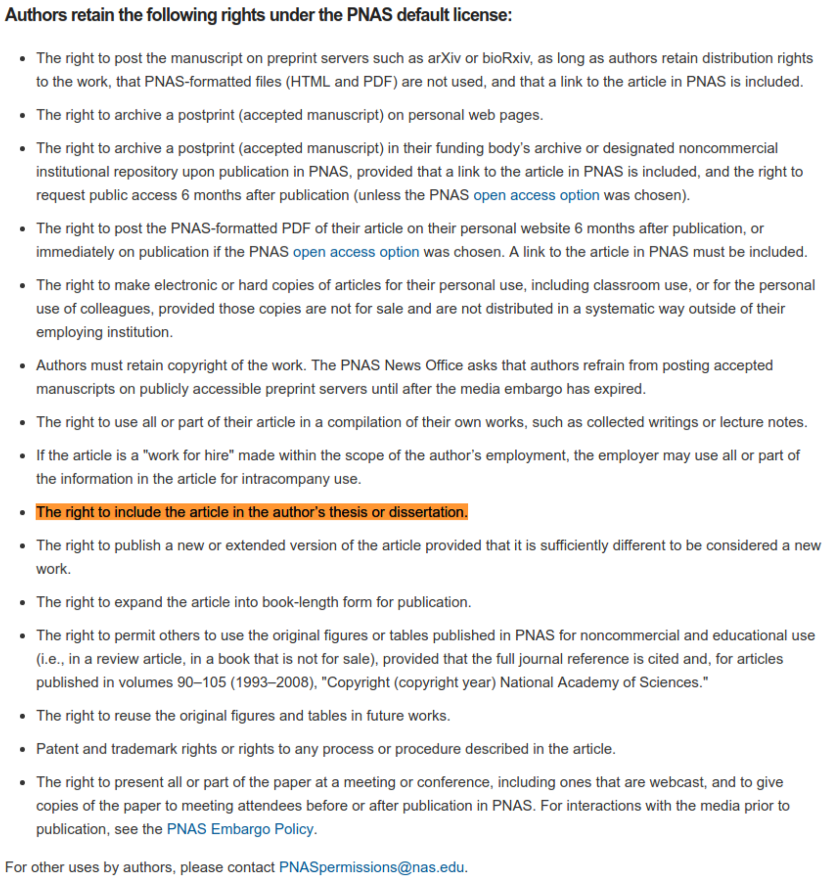


Figure 28: The sym–GeTFEP for symmetric membranes. This profile is derived by mirroring the left part (depth -4 to -1) of the original GeTFEP.

Following is the proof of permission for this publication:

Tian, W., Lin, M., Tang, K., Liang, J., and Naveed, H.: High–resolution structure prediction of β–barrel membrane proteins. Proceedings of the National Academy of Sciences 115(7):1511–1516. doi:10.1073/pnas.1716817115, 2018.



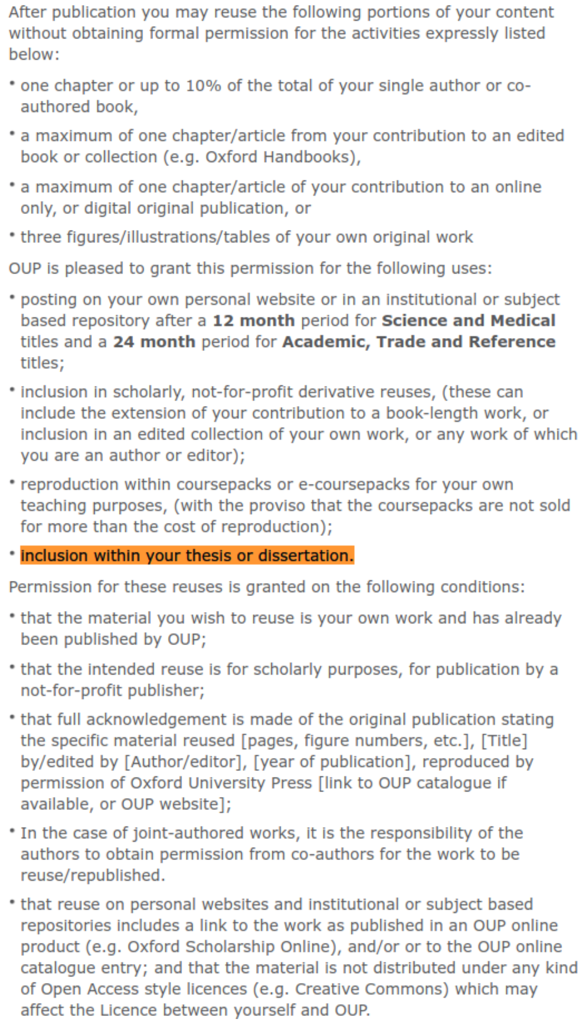
Authors retain the following rights under the PNAS default license:

- The right to post the manuscript on preprint servers such as arXiv or bioRxiv, as long as authors retain distribution rights to the work, that PNAS-formatted files (HTML and PDF) are not used, and that a link to the article in PNAS is included.
- The right to archive a postprint (accepted manuscript) on personal web pages.
- The right to archive a postprint (accepted manuscript) in their funding body's archive or designated noncommercial institutional repository upon publication in PNAS, provided that a link to the article in PNAS is included, and the right to request public access 6 months after publication (unless the PNAS open access option was chosen).
- The right to post the PNAS-formatted PDF of their article on their personal website 6 months after publication, or immediately on publication if the PNAS open access option was chosen. A link to the article in PNAS must be included.
- The right to make electronic or hard copies of articles for their personal use, including classroom use, or for the personal use of colleagues, provided those copies are not for sale and are not distributed in a systematic way outside of their employing institution.
- Authors must retain copyright of the work. The PNAS News Office asks that authors refrain from posting accepted manuscripts on publicly accessible preprint servers until after the media embargo has expired.
- The right to use all or part of their article in a compilation of their own works, such as collected writings or lecture notes.
- If the article is a "work for hire" made within the scope of the author's employment, the employer may use all or part of the information in the article for intracompany use.
- The right to include the article in the author's thesis or dissertation.
- The right to publish a new or extended version of the article provided that it is sufficiently different to be considered a new work.
- The right to expand the article into book-length form for publication.
- The right to permit others to use the original figures or tables published in PNAS for noncommercial and educational use (i.e., in a review article, in a book that is not for sale), provided that the full journal reference is cited and, for articles published in volumes 90–105 (1993–2008), "Copyright (copyright year) National Academy of Sciences."
- The right to reuse the original figures and tables in future works.
- Patent and trademark rights or rights to any process or procedure described in the article.
- The right to present all or part of the paper at a meeting or conference, including ones that are webcast, and to give copies of the paper to meeting attendees before or after publication in PNAS. For interactions with the media prior to publication, see the PNAS Embargo Policy.

For other uses by authors, please contact PNASpermissions@nas.edu.

Following is the proof of permission for this publication:

Tian, W., Lin, M., Naveed, H., and Liang, J.: Efficient computation of transfer free energies of amino acids in beta–barrel membrane proteins. Bioinformatics 33(11):1664–1671. doi:10.1093/bioinformatics/btx053, 2017.



After publication you may reuse the following portions of your content without obtaining formal permission for the activities expressly listed below:

- one chapter or up to 10% of the total of your single author or co-authored book,
- a maximum of one chapter/article from your contribution to an edited book or collection (e.g. Oxford Handbooks),
- a maximum of one chapter/article of your contribution to an online only, or digital original publication, or
- three figures/illustrations/tables of your own original work

OUP is pleased to grant this permission for the following uses:

- posting on your own personal website or in an institutional or subject based repository after a **12 month** period for **Science and Medical** titles and a **24 month** period for **Academic, Trade and Reference** titles;
- inclusion in scholarly, not-for-profit derivative reuses, (these can include the extension of your contribution to a book-length work, or inclusion in an edited collection of your own work, or any work of which you are an author or editor);
- reproduction within coursepacks or e-coursepacks for your own teaching purposes, (with the proviso that the coursepacks are not sold for more than the cost of reproduction);
- inclusion within your thesis or dissertation.

Permission for these reuses is granted on the following conditions:

- that the material you wish to reuse is your own work and has already been published by OUP;
- that the intended reuse is for scholarly purposes, for publication by a not-for-profit publisher;
- that full acknowledgement is made of the original publication stating the specific material reused [pages, figure numbers, etc.], [Title] by/edited by [Author/editor], [year of publication], reproduced by permission of Oxford University Press [link to OUP catalogue if available, or OUP website];
- In the case of joint-authored works, it is the responsibility of the authors to obtain permission from co-authors for the work to be reuse/republished.
- that reuse on personal websites and institutional or subject based repositories includes a link to the work as published in an OUP online product (e.g. Oxford Scholarship Online), and/or or to the OUP online catalogue entry; and that the material is not distributed under any kind of Open Access style licences (e.g. Creative Commons) which may affect the Licence between yourself and OUP.

CITED LITERATURE

1. Moon, C. P. and Fleming, K. G.: Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. Proceedings of the National Academy of Sciences of the United States of America, 108(25):10174–7, June 2011.

2. Lomize, A. L., Pogozheva, I. D., Lomize, M. a., and Mosberg, H. I.: Positioning of proteins in membranes: A computational approach. Protein Science, 15(6):1318–1333, 2006.

3. Jackups, R. and Liang, J.: Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. Journal of molecular biology, 354(4):979–93, dec 2005.

4. Wimley, W. C. and White, S. H.: Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nature structural biology, 3:842–848, 1996.

5. Remmert, M., Biegert, A., Hauser, A., and Söding, J.: HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2):173–175, 2012.

6. Fernandez, C., Hilty, C., Wider, G., Guntert, P., and Wuthrich, K.: NMR structure of the integral membrane protein OmpX. J Mol Biol, 336(5):1211–1221, 2004.

7. Kleinschmidt, J. H., den Blaauwen, T., Driessen, A. J., and Tamm, L. K.: Outer membrane protein A of Escherichia coli inserts and folds into lipid bilayers by a concerted mechanism. Biochemistry, 38(16):5006–16, apr 1999.

8. Snider, C., Jayasinghe, S., Hristova, K., and White, S. H.: MPEx: A tool for exploring membrane proteins. Protein Science, 18(12):2624–2628, 2009.

9. Wimley, W. C.: The versatile β-barrel membrane protein. Current Opinion in Structural Biology, 13(4):404–411, August 2003.

10. Delcour, A.: Structure and function of pore-forming beta-barrels from bacteria. J Mol Microbiol Biotechnol, 4(1):1–10, Jan 2002.

11. Bishop, R. E.: Structural biology of membrane-intrinsic β-barrel enzymes: Sentinels of the bacterial outer membrane. Biochimica et Biophysica Acta - Biomembranes, 1778(9):1881–1896, 2008.

12. Song, L., Hobaugh, M., Shustak, C., Cheley, S., Bayley, H., and Gouaux, J.: Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science, 274(5294):1859–1866, Dec 1996.

13. Noinaj, N., Kuszak, A., Gumbart, J., Lukacik, P., Chang, H., Easley, N., Lithgow, T., and Buchanan, S.: Structural insight into the biogenesis of beta-barrel membrane proteins. Nature, 501(7467):385–390, Sep 2013.

14. Koebnik, R., Locher, K. P., and Van Gelder, P.: Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Molecular microbiology, 37(2):239–253, 2000.

15. Bajaj, H., Tran, Q. T., Mahendran, K. R., Nasrallah, C., Colletier, J. P., Davin-Regli, A., Bolla, J. M., Pagès, J. M., and Winterhalter, M.: Antibiotic uptake through membrane channels: Role of Providencia stuartii omppst1 porin in carbapenem resistance. Biochemistry, 51:10244–10249, 2012.

16. Manczak, M. and Reddy, P. H.: Abnormal interaction of VDAC1 with amyloid beta and phosphorylated tau causes mitochondrial dysfunction in Alzheimer's disease. Human Molecular Genetics, 21:5131–5146, 2012.

17. Bender, A., Desplats, P., Spencer, B., Rockenstein, E., Adame, A., Elstner, M., Laub, C., Mueller, S., Koob, A. O., Mante, M., Pham, E., Klopstock, T., and Masliah, E.: TOM40 Mediates Mitochondrial Dysfunction Induced by α-Synuclein Accumulation in Parkinson's Disease. PLoS ONE, 8, 2013.

18. Oukhaled, A., Bacri, L., Pastoriza-Gallego, M., Betton, J. M., and Pelta, J.: Sensing proteins through nanopores: Fundamental to applications. ACS Chemical Biology, 7(12):1935–1949, 2012.

19. Farimani, A. B., Heiranian, M., and Aluru, N. R.: Electromechanical signatures for DNA sequencing through a mechanosensitive nanopore. Journal of Physical Chemistry Letters, 6(4):650–657, 2015.

20. Campos, E., McVey, C. E., Carney, R. P., Stellacci, F., Astier, Y., and Yates, J.: Sensing single mixed-monolayer protected gold nanoparticles by the α-hemolysin nanopore. Analytical Chemistry, 85(21):10149–10158, 2013.

21. Hong, H., Joh, N., J.U., B., and L.K., T.: Methods for measuring the thermodynamic stability of membrane proteins. Methods in Enzymology, 455:213–36, 2009.

22. Jackups, R., Cheng, S., and Liang, J.: Sequence motifs and antimotifs in beta-barrel membrane proteins from a genome-wide analysis: the Ala-Tyr dichotomy and chaperone binding motifs. Journal of molecular biology, 363(2):611–23, oct 2006.

23. Hu, J., Worrall, L. J., Hong, C., Vuckovic, M., Atkinson, C. E., Caveney, N., Yu, Z., and Strynadka, N. C.: Cryo-EM analysis of the T3S injectisome reveals the structure of the needle and open secretin. Nature Communications, 2018.

24. Ruan, J., Xia, S., Liu, X., Lieberman, J., and Wu, H.: Cryo-EM structure of the gasdermin A3 membrane pore. Nature, 2018.

25. Ujwal, R., Cascio, D., Colletier, J., Faham, S., Zhang, J., Toro, L., Ping, P., and Abramson, J.: The crystal structure of mouse VDAC1 at 2.3 a resolution reveals mechanistic insights into metabolite gating. Proc Natl Acad Sci U S A, 105(46):17742–17747, Nov 2008.

26. Cowan, S. W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R. A., Jansonius, J. N., and Rosenbusch, J. P.: Crystal structures explain functional properties of two E. coli porins. Nature, 1992.

27. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., and Hughes, C.: Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. Nature, 2000.

28. Snijder, H. J., Ubarretxena-Belandia, I., Blaauw, M., Kalk, K. H., Verheij, H. M., Egmond, M. R., Dekker, N., and Dijkstra, B. W.: Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. Nature, 401(6754):717–721, 1999.

29. Schulz, G. E.: The structure of bacterial outer membrane proteins. Biochim Biophys Acta., 1565:308–317, 2002.

30. Koebnik, R.: Structural and functional roles of the surface-exposed loops of the beta-barrel membrane protein ompa from escherichia coli. J Bacteriol, 181(12):3688–3694, Jun 1999.

31. Arora, A., Abildgaard, F., Bushweller, J., and Tamm, L.: Structure of outer membrane protein a transmembrane domain by NMR spectroscopy. Nat Struct Biol, 8(4):334–338, Apr 2001.

32. Cierpicki, T., Liang, B., Tamm, L., and Bushweller, J.: Increasing the accuracy of solution NMR structures of membrane proteins by application of residual dipolar couplings. high-resolution structure of outer membrane protein a. J Am Chem Soc, 128(21):6947–6951, May 2006.

33. Koehler Leman, J., Ulmschneider, M., and Gray, J.: Computational modeling of membrane proteins. Proteins, 83(1):1–24, Jan 2015.

34. Naveed, H., Jackups, Jr., R., and Liang, J.: Predicting weakly stable regions, oligomerization state, and protein-protein interfaces in transmembrane domains of outer membrane proteins. Proc Natl Acad Sci U S A, 106(31):12735–12740, Aug 2009.

35. Naveed, H., Jimenez-Morales, D., Tian, J., Pasupuleti, V., Kenney, L. J., and Liang, J.: Engineered oligomerization state of OmpF protein through computational design decouples oligomer dissociation from unfolding. Journal of Molecular Biology, 2012.

36. White, S. H. and Wimley, W. C.: Membrane protein folding and stability: physical principles. Annual review of biophysics and biomolecular structure, 28:319–365, 1999.

37. Gessmann, D., Mager, F., Naveed, H., Arnold, T., Weirich, S., Linke, D., Liang, J., and Nussberger, S.: Improving the resistance of a eukaryotic beta-barrel protein to thermal and chemical perturbations. J Mol Biol, 413(1):150–161, Oct 2011.

38. Tanford, C.: The hydrophobic effect and the organization of living matter. Science (New York, N.Y.), 200:1012–1018, 1978.

39. Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H., and von Heijne, G.: Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature, 433(7024):377–81, jan 2005.

40. Kyte, J. and Doolittle, R. F.: A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1):105–132, 1982.

41. Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., and Elofsson, A.: Prediction of membrane-protein topology from first principles. Proceedings of the National Academy of Sciences, 2008.

42. Lin, M., Gessmann, D., Naveed, H., and Liang, J.: Outer Membrane Protein Folding and Topology from a Computational Transfer Free Energy Scale. Journal of the American Chemical Society, 138(8):2592–2601, mar 2016.

43. Tian, W., Naveed, H., Lin, M., and Liang, J.: GeTFEP: a general transfer free energy profile for transmembrane proteins. preprint on webpage at `https://doi.org/10.1101/191650`, oct 2017.

44. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P.: The protein data bank. Nucleic Acids Res, 28(1):235–242, Jan 2000.

45. Ou, Y., Gromiha, M., Chen, S., and Suwa, M.: TMBETADISC-RBF: Discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. Comput Biol Chem, 32(3):227–231, Jun 2008.

46. Freeman, Jr., T. and Wimley, W.: A highly accurate statistical approach for the prediction of transmembrane beta-barrels. Bioinformatics, 26(16):1965–1974, Aug 2010.

47. Ou, Y., Chen, S., and Gromiha, M.: Prediction of membrane spanning segments and topology in beta-barrel membrane proteins at better accuracy. J Comput Chem, 31(1):217–223, Jan 2010.

48. Hayat, S., Peters, C., Shu, N., Tsirigos, K., and Elofsson, A.: Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. Bioinformatics, Jan 2016.

49. Jackups, Jr., R. and Liang, J.: Combinatorial analysis for sequence and spatial motif discovery in short sequence fragments. IEEE/ACM Trans Comput Biol Bioinform, 7(3):524–536, Jul 2010.

50. Naveed, H. and Liang, J.: Weakly stable regions and protein-protein interactions in beta-barrel membrane proteins. Curr Pharm Des, 20(8):1268–1273, 2014.

51. Geula, S., Naveed, H., Liang, J., and Shoshan-Barmatz, V.: Structure-based analysis of vdac1: Defining oligomer contact sites. Journal of Biological Chemistry, 2011.

52. Naveed, H., Jimenez-Morales, D., Tian, J., Pasupuleti, V., Kenney, L., and Liang, J.: Engineered oligomerization state of ompf protein through computational design decouples oligomer dissociation from unfolding. J Mol Biol, 419(1-2):89–101, May 2012.

53. Tian, W., Lin, M., Naveed, H., and Liang, J.: Efficient computation of transfer free energies of amino acids in beta-barrel membrane proteins. Bioinformatics (Oxford, England), 33(11):1664–1671, jun 2017.

54. Yang, J., Zhang, W., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., Virtanen, J., Xue, Z., Shen, H.-B., and Zhang, Y.: Template-based protein structure prediction in casp11 and retrospect of i-tasser in the last decade. Proteins: Structure, Function, and Bioinformatics, 84:233–246, 2016.

55. Randall, A., Cheng, J., Sweredoski, M., and Baldi, P.: TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. Bioinformatics, 24(4):513–520, Feb 2008.

56. Remaut, H., Tang, C., Henderson, N. S., Pinkner, J. S., Wang, T., Hultgren, S. J., Thanassi, D. G., Waksman, G., and Li, H.: Fiber Formation across the Bacterial Outer Membrane by the Chaperone/Usher Pathway. Cell, 133:640–652, 2008.

57. Qiao, S., Luo, Q., Zhao, Y., Zhang, X., and Huang, Y.: Structural basis for lipopolysaccharide insertion in the bacterial outer membrane. Nature, 511(7507):108–111, Jul 2014.

58. Murzin, A., Lesk, A., and Chothia, C.: Principles determining the structure of beta-sheet barrels in proteins. i. a theoretical analysis. J Mol Biol, 236(5):1369–1381, Mar 1994.

59. Murzin, A., Lesk, A., and Chothia, C.: Principles determining the structure of beta-sheet barrels in proteins. ii. the observed structures. J Mol Biol, 236(5):1382–1400, Mar 1994.

60. Albrecht, R., Schutz, M., Oberhettinger, P., Faulstich, M., Bermejo, I., Rudel, T., Diederichs, K., and Zeth, K.: Structure of bama, an essential factor in outer

membrane protein biogenesis. Acta Crystallogr D Biol Crystallogr, 70(Pt 6):1779–1789, Jun 2014.

61. Naveed, H., Xu, Y., Jackups, Jr., R., and Liang, J.: Predicting three-dimensional structures of transmembrane domains of beta-barrel membrane proteins. J Am Chem Soc, 134(3):1775–1781, Jan 2012.

62. Hayat, S., Sander, C., Marks, D., and Elofsson, A.: All-atom 3d structure prediction of transmembrane beta-barrel proteins from sequences. Proc Natl Acad Sci U S A, 112(17):5413–5418, Apr 2015.

63. Jackups Jr, R. and Liang, J.: Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol, 354(4):979–93, 2005.

64. Ho, B. and Curmi, P.: Twist and shear in beta-sheets and beta-ribbons. J Mol Biol, 317(2):291–308, Mar 2002.

65. Jackups, Jr., R. and Liang, J.: Combinatorial model for sequence and spatial motif discovery in short sequence fragments: examples from beta-barrel membrane proteins. Conf Proc IEEE Eng Med Biol Soc, 1:3470–3473, 2006.

66. Kabsch, W. and Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12):2577–2637, Dec 1983.

67. Jones, D., Buchan, D., Cozzetto, D., and Pontil, M.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, 28(2):184–190, Jan 2012.

68. Liu, W.: Shear numbers of protein beta-barrels: definition refinements and statistics. J Mol Biol, 275(4):541–545, Jan 1998.

69. Schmidt, N., Grigoryan, G., and DeGrado, W.: The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils. application to understand signaling in histidine kinases. Protein Sci, Dec 2016.

70. Huang, P., Oberdorfer, G., Xu, C., Pei, X., Nannenga, B., Rogers, J., DiMaio, F., Gonen, T., Luisi, B., and Baker, D.: High thermodynamic stability of parametrically designed helical bundles. Science, 346(6208):481–485, Oct 2014.

71. McLachlan, A.: Gene duplications in the structural evolution of chymotrypsin. J Mol Biol, 128(1):49–79, Feb 1979.

72. O'Neill, B.: Elementary Differential Geometry.. London UK, Academic Press Inc., 1966.

73. Gront, D., Kmiecik, S., and Kolinski, A.: Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J Comput Chem, 28(9):1593–1597, Jul 2007.

74. Krivov, G., Shapovalov, M., and Dunbrack, Jr., R.: Improved prediction of protein side-chain conformations with SCWRL4. Proteins, 77(4):778–795, Dec 2009.

75. Tang, K., Wong, S., Liu, J., Zhang, J., and Liang, J.: Conformational sampling and structure prediction of multiple interacting loops in soluble and beta-barrel membrane proteins using multi-loop distance-guided chain-growth monte carlo method. Bioinformatics, 31(16):2646–2652, Aug 2015.

76. Tian, W., Lin, M., Tang, K., Liang, J., and Naveed, H.: High-resolution structure prediction of β–barrel membrane proteins. Proceedings of the National Academy of Sciences, page 201716817, 2018.

77. Fioroni, M., Dworeck, T., and Rodriguez-Ropero, F.: Beta-Barrel Channel Proteins As Tools in Nanotechnology: Biology, Basic Science and Advanced Applications. Springer Publishing Company, Incorporated, 2013.

78. Tang, K., Zhang, J., and Liang, J.: Distance-guided forward and backward chain-growth Monte Carlo method for conformational sampling and structural prediction of antibody CDR-H3 loops. Journal of Chemical Theory and Computation, 13(1):380–388, 2017.

79. Freeman, Jr., T. and Wimley, W.: TMBB-DB: a transmembrane beta-barrel proteome database. Bioinformatics, 28(19):2425–2430, Oct 2012.

80. Novotný, J., Bruccoleri, R. E., and Newell, J.: Twisted hyperboloid (strophoid) as a model of β-barrels in proteins. Journal of Molecular Biology, 177(3):567–573, 1984.

81. Lasters, I., Wodak, S. J., Alard, P., and van Cutsem, E.: Structural principles of parallel beta-barrels in proteins. Proceedings of the National Academy of Sciences of the United States of America, 85(10):3338–42, 1988.

82. Moon, C. P., Zaccai, N. R., Fleming, P. J., Gessmann, D., and Fleming, K. G.: Membrane protein thermodynamic stability may serve as the energy sink for sorting in the periplasm. Proceedings of the National Academy of Sciences of the United States of America, 110(11):4285–90, March 2013.

83. Moon, C. P., Kwon, S., and Fleming, K. G.: Overcoming hysteresis to attain reversible equilibrium folding for outer membrane phospholipase A in phospholipid bilayers. Journal of Molecular Biology, 413(2):484–494, 2011.

84. Otzen, D. E. and Andersen, K. K.: Folding of outer membrane proteins. Archives of Biochemistry and Biophysics, 531(1-2):34–43, 2013.

85. Adamian, L., Nanda, V., DeGrado, W. F., and Liang, J.: Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. Proteins, 59(3):496–509, May 2005.

86. Schramm, C. A., Hannigan, B. T., Donald, J. E., Keasar, C., Saven, J. G., Degrado, W. F., and Samish, I.: Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. Structure (London, England : 1993), 20(5):924–35, May 2012.

87. Hsieh, D., Davis, A., and Nanda, V.: A knowledge-based potential highlights unique features of membrane α-helical and β-barrel protein insertion and folding. Protein science : a publication of the Protein Society, 21(1):50–62, jan 2012.

88. Slusky, J. S. G. and Dunbrack, R. L.: Charge asymmetry in the proteins of the outer membrane. Bioinformatics, 29(17):2122–2128, 2013.

89. Gumbart, J. and Roux, B.: Determination of membrane-insertion free energies by molecular dynamics simulations. Biophysical journal, 102(4):795–801, February 2012.

90. Ulmschneider, J. P., Andersson, M., and Ulmschneider, M. B.: Determining peptide partitioning properties via computer simulation. The Journal of membrane biology, 239(1-2):15–26, January 2011.

91. Wouters, M. A. and Curmi, P. M. G.: An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: The differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. Proteins: Structure, Function and Genetics, 22:119–131, 1995.

92. Lomize, M. A., Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I.: OPM: Orientations of proteins in membranes database. Bioinformatics, 22(5):623–625, 2006.

93. Miyazawa, S. and Jernigan, R. L.: Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of molecular biology, 256:623–644, 1996.

94. Tsirigos, K., Elofsson, A., and Bagos, P.: PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. Bioinformatics, 32(17):i665–i671, Sep 2016.

95. Liang, J.: Experimental and computational studies of determinants of membrane-protein folding. 6(6):878–884, 2002.

96. Adamian, L. and Liang, J.: Prediction of buried helices in multispan alpha helical membrane proteins. Proteins, 63(1):1–5, apr 2006.

97. Liang, J., Naveed, H., Jimenez-Morales, D., Adamian, L., and Lin, M.: Computational studies of membrane proteins: Models and predictions for biological understanding. 1818(4):927–941, 2012.

98. Senes, A., Chadi, D. C., Law, P. B., Walters, R. F. S., Nanda, V., and DeGrado, W. F.: Ez, a Depth-dependent Potential for Assessing the Energies of Insertion of Amino Acid Side-chains into Membranes: Derivation and Applications to Determining the Orientation of Transmembrane and Interfacial Helices. Journal of Molecular Biology, 366(2):436–448, 2007.

99. Moore, D. T., Berger, B. W., and DeGrado, W. F.: Protein-Protein Interactions in the Membrane: Sequence, Structural, and Biological Motifs. 16(7):991–1001, 2008.

100. Hedin, L. E., Öjemalm, K., Bernsel, A., Hennerdal, A., Illergård, K., Enquist, K., Kauko, A., Cristobal, S., von Heijne, G., Lerch-Bader, M., Nilsson, I., and Elofsson, A.: Membrane Insertion of Marginally Hydrophobic Transmembrane Helices Depends on Sequence Context. Journal of Molecular Biology, 396(1):221–229, 2010.

101. MacCallum, J. L., Bennett, W. F. D., and Tieleman, D. P.: Partitioning of amino acid side chains into lipid bilayers: results from computer simulations and comparison to experiment. The Journal of general physiology, 129(5):371–377, 2007.

102. Ulmschneider, M. B., Ulmschneider, J. P., Schiller, N., Wallace, B. a., von Heijne, G., and White, S. H.: Spontaneous transmembrane helix insertion thermodynamically mimics translocon-guided insertion. Nature Communications, 5:4863, 2014.

103. Senes, A., Engel, D. E., and DeGrado, W. F.: Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Current opinion in structural biology, 14(4):465–79, 2004.

104. Dekker, N.: Outer-membrane phospholipase A: Known structure, unknown biological function. Molecular Microbiology, 35(4):711–717, 2000.

105. Kingma, R. L., Fragiathaki, M., Snijder, H. J., Dijkstra, B. W., Verheij, H. M., Dekker, N., and Egmond, M. R.: Unusual catalytic triad of Escherichia coli outer membrane phospholipase A. Biochemistry, 38(4):10017–10022, 2000.

106. PDB structure of PagP. doi:10.2210/pdb1thq/pdb.

107. PDB structure of PagL. doi:10.2210/pdb2erv/pdb.

108. Rutten, L., Geurtsen, J., Lambert, W., Smolenaers, J. J. M., Bonvin, A. M., de Haan, A., van der Ley, P., Egmond, M. R., Gros, P., and Tommassen, J.: Crystal structure and catalytic mechanism of the LPS 3-O-deacylase PagL from Pseudomonas aeruginosa. Proceedings of the National Academy of Sciences of the United States of America, 103(18):7071–7076, 2006.

109. Jayasinghe, S., Hristova, K., and White, S. H.: MPtopo: A database of membrane protein topology. Protein Science, 10(2):455–458, 2001.

110. Yang, Y., Guo, R., Gaffney, K., Kim, M., Muhammednazaar, S., Tian, W., Wang, B., Liang, J., and Hong, H.: Folding-Degradation Relationship of a Membrane Protein Mediated by the Universally Conserved ATP-Dependent Protease FtsH. Journal of the American Chemical Society, 140(13):4656–4665, 2018.

111. Pogozheva, I. D., Mosberg, H. I., and Lomize, A. L.: Life at the border: Adaptation of proteins to anisotropic membrane environment. Protein Science, 23(9):1165–1196, sep 2014.

112. Sanders, C. R. and Nagy, J. K.: Misfolding of membrane proteins in health and disease: The lady or the tiger? 10(4):438–442, 2000.

113. Fairman, J. W., Noinaj, N., and Buchanan, S. K.: The structural biology of β-barrel membrane proteins: A summary of recent reports. 21(4):523–531, 2011.

114. Deber, C. M., Wang, C., Liu, L. P., Prior, a. S., Agrawal, S., Muskat, B. L., and Cuticchia, a. J.: TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. Protein science : a publication of the Protein Society, 10(1):212–219, 2001.

115. MacCallum, J. L. and Tieleman, D. P.: Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions. Trends in biochemical sciences, 36(12):653–62, dec 2011.

116. Derrington, I. M., Butler, T. Z., Collins, M. D., Manrao, E., Pavlenok, M., Niederweis, M., and Gundlach, J. H.: Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A, 107(37):16060–16065, 2010.

117. Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., and Bayley, H.: Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proceedings of the National Academy of Sciences of the United States of America, 106(19):7702–7707, 2009.

118. Ulmschneider, M. B. and Sansom, M. S. P.: Amino acid distributions in integral membrane protein structures. Biochim Biophys Acta, 1512(1):1–14, 2001.

119. Fiedler, S., Broecker, J., and Keller, S.: Protein folding in membranes., 2010.

120. Kaufman, L. and Rousseeuw, P. J.: Finding groups in data: an introduction to cluster analysis. New York Wiley, 1990. Bibliography: p. 320-331. "A Wiley-Interscience publication." Includes index.

121. Freeman, Jr., T., Landry, S., and Wimley, W.: The prediction and characterization of ysha, an unknown outer-membrane protein from salmonella typhimurium. Biochim Biophys Acta, 1808(1):287–297, Jan 2011.

122. Fahie, M., Yang, B., Mullis, M., Holden, M., and Chen, M.: Selective detection of protein homologues in serum using an ompg nanopore. Anal Chem, 87(21):11143–11149, Nov 2015.

123. Fahie, M., Chisholm, C., and Chen, M.: Resolved single-molecule detection of individual species within a mixture of anti-biotin antibodies using an engineered monomeric nanopore. ACS Nano, 9(2):1089–1098, Feb 2015.

124. Ayub, M., Stoddart, D., and Bayley, H.: Nucleobase recognition by truncated alpha-hemolysin pores. ACS Nano, 9(8):7895–7903, Aug 2015.

125. Panchal, R., Cusack, E., Cheley, S., and Bayley, H.: Tumor protease-activated, pore-forming toxins from a combinatorial library. Nat Biotechnol, 14(7):852–856, Jul 1996.

126. Tamm, L., Arora, A., and Kleinschmidt, J.: Structure and assembly of β-barrel membrane proteins. J Biol Chem, 276:32399–32402, 2001.

127. Tamm, L. K., Hong, H., and Liang, B.: Folding and assembly of β-barrel membrane proteins. Biochim Biophys Acta, 1666:250–263, 2004.

128. Burgess, N., Dao, T., Stanley, A., and Fleming, K.: Beta-barrel proteins that reside in the Escherichia coli outer membrane in vivo demonstrate varied folding behavior in vitro. J Biol Chem, 283(39):26748–58, 2008.

129. Wimley, W. C.: Toward genomic identification of β-barrel membrane proteins: composition and architecture of known structures. Protein Sci., 11:301–312, 2002.

130. Bonhivers, M., Desmadril, M., Moeck, G., Boulanger, P., Colomer-Pallas, A., and Letellier, L.: Stability studies of fhua, a two-domain outer membrane protein from escherichia coli. Biochemistry, 40(8):2606–2613, Feb 2001.

131. Bagos, P., Liakopoulos, T., Spyropoulos, I., and Hamodrakas, S.: PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. Nucleic Acids Res, 32(Web Server issue):W400–4, Jul 2004.

132. Bigelow, H. and Rost, B.: PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res, 34(Web Server issue):W186–8, Jul 2006.

133. Fariselli, P., Martelli, P., and Casadio, R.: A new decoding algorithm for hidden markov models improves the prediction of the topology of all-beta membrane proteins. BMC Bioinformatics, 6 Suppl 4:S12, Dec 2005.

134. Gromiha, M., Majumdar, R., and Ponnuswamy, P.: Identification of membrane spanning beta strands in bacterial porins. Protein Eng, 10(5):497–500, May 1997.

135. Gromiha, M., Ahmad, S., and Suwa, M.: Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. J Comput Chem, 25(5):762–767, Apr 2004.

136. Gromiha, M., Ahmad, S., and Suwa, M.: TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in outer membrane proteins. Nucleic Acids Res, 33(Web Server issue):W164–7, Jul 2005.

137. Liu, Q., Zhu, Y., Wang, B., and Li, Y.: A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. Comput Biol Chem, 27(1):69–76, Feb 2003.

138. Jacoboni, I., Martelli, P., Fariselli, P., De Pinto, V., and Casadio, R.: Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. Protein Sci, 10(4):779–787, Apr 2001.

139. Natt, N., Kaur, H., and Raghava, G.: Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. Proteins, 56(1):11–18, Jul 2004.

140. Waldispuhl, J., Berger, B., Clote, P., and Steyaert, J.: transfold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res, 34(Web Server issue):W189–93, Jul 2006.

141. Tress, M. and Valencia, A.: Predicted residue-residue contacts can help the scoring of 3d models. Proteins, 78(8):1980–1991, Jun 2010.

142. Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J., and Levy, Y.: Assessment of CASP8 structure predictions for template free targets. Proteins, 77 Suppl 9:50–65, 2009.

143. Wu, S., Skolnick, J., and Zhang, Y.: Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol, 5:17, 2007.

144. Kaufmann, K., Lemmon, G., Deluca, S., Sheehan, J., and Meiler, J.: Practically useful: what the rosetta protein modeling suite can do for you. Biochemistry, 49(14):2987–2998, Apr 2010.

145. Venclovas, C., Ginalski, K., and Fidelis, K.: Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins. Proteins, Suppl 3:73–80, 1999.

146. Valavanis, I., Bagos, P., and Emiris, I.: beta-barrel transmembrane proteins: Geometric modelling, detection of transmembrane region, and structural properties. Comput Biol Chem, 30(6):416–424, Dec 2006.

147. Chou, K., Nemethy, G., and Scheraga, H.: Role of interchain interactions in the stabilization of the right-handed twist of beta-sheets. J Mol Biol, 168(2):389–407, Aug 1983.

148. Chou, K. and Scheraga, H.: Origin of the right-handed twist of beta-sheets of poly(LVal) chains. Proc Natl Acad Sci U S A, 79(22):7047–7051, Nov 1982.

149. Chou, K., Pottle, M., Nemethy, G., Ueda, Y., and Scheraga, H.: Structure of beta-sheets. origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets. J Mol Biol, 162(1):89–112, Nov 1982.

150. Wang, L., O'Connell, T., Tropsha, A., and Hermans, J.: Molecular simulations of beta-sheet twisting. J Mol Biol, 262(2):283–293, Sep 1996.

151. Wang, L., Rivera, E., Benavides-Garcia, M., and Nall, B.: Loop entropy and cytochrome c stability. J Mol Biol, 353(3):719–729, Oct 2005.

152. Humphrey, W., Dalke, A., and Schulten, K.: VMD: visual molecular dynamics. J Mol Graph, 14(1):33–8, 27–8, Feb 1996.

153. Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L., and Schulten, K.: Scalable molecular dynamics with NAMD. J Comput Chem, 26(16):1781–1802, Dec 2005.

154. Srinivas, N., Jetter, P., Ueberbacher, B., Werneburg, M., Zerbe, K., Steinmann, J., Van der Meijden, B., Bernardini, F., Lederer, A., Dias, R., Misson, P., Henze, H., Zumbrunn, J., Gombert, F., Obrecht, D., Hunziker, P., Schauer, S., Ziegler, U., Kach, A., Eberl, L., Riedel, K., DeMarco, S., and Robinson, J.: Peptidomimetic antibiotics target outer-membrane biogenesis in pseudomonas aeruginosa. Science, 327(5968):1010–1013, Feb 2010.

155. Yildirim, M., Goh, K., Cusick, M., Barabasi, A., and Vidal, M.: Drug-target network. Nat Biotechnol, 25(10):1119–1126, Oct 2007.

156. Ferguson, A., Hofmann, E., Coulton, J., Diederichs, K., and Welte, W.: Siderophore-mediated iron transport: crystal structure of fhua with bound lipopolysaccharide. Science, 282(5397):2215–2220, Dec 1998.

157. Locher, K. and Rosenbusch, J.: Oligomeric states and siderophore binding of the ligand-gated fhua protein that forms channels across escherichia coli outer membranes. Eur J Biochem, 247(3):770–775, Aug 1997.

158. Van Gelder, P. and Tommassen, J.: Demonstration of a folded monomeric form of porin phoe of escherichia coli in vivo. J Bacteriol, 178(17):5320–5322, Sep 1996.

159. Barth, P., Wallner, B., and Baker, D.: Prediction of membrane protein structures with complex topologies using limited constraints. Proc Natl Acad Sci U S A, 106(5):1409–1414, Feb 2009.

160. Vogt, J. and Schulz, G.: The structure of the outer membrane protein ompx from escherichia coli reveals possible mechanisms of virulence. Structure, 7(10):1301–1309, Oct 1999.

161. Furini, S., Domene, C., Rossi, M., Tartagni, M., and Cavalcanti, S.: Model-based prediction of the alpha-hemolysin structure in the hexameric state. Biophys J, 95(5):2265–2274, Sep 2008.

162. Hearn, E., Patel, D., Lepore, B., Indic, M., and van den Berg, B.: Transmembrane passage of hydrophobic compounds through a protein channel wall. Nature, 458(7236):367–370, Mar 2009.

163. Karginov, V., Nestorovich, E., Schmidtmann, F., Robinson, T., Yohannes, A., Fahmi, N., Bezrukov, S., and Hecht, S.: Inhibition of s. aureus alpha-hemolysin and b. anthracis lethal toxin by beta-cyclodextrin derivatives. Bioorg Med Chem, 15(16):5424–5431, Aug 2007.

164. Banerjee, A., Mikhailova, E., Cheley, S., Gu, L., Montoya, M., Nagaoka, Y., Gouaux, E., and Bayley, H.: Molecular bases of cyclodextrin adapter interactions with engineered protein nanopores. Proc Natl Acad Sci U S A, 107(18):8165–8170, May 2010.

165. Adiga, S., Jin, C., Curtiss, L., Monteiro-Riviere, N., and Narayan, R.: Nanoporous membranes for medical and biological applications. Wiley Interdiscip Rev Nanomed Nanobiotechnol, 1(5):568–581, Sep 2009.

166. Czajkowsky, D., Sheng, S., and Shao, Z.: Staphylococcal alpha-hemolysin can form hexamers in phospholipid bilayers. J Mol Biol, 276(2):325–330, Feb 1998.

167. Zhu, P., Klutch, M., Derrick, J., Prince, S., Tsang, R., and Tsai, C.: Identification of opca gene in neisseria polysaccharea: interspecies diversity of opc protein family. Gene, 307:31–40, Mar 2003.

168. Gabriel, K., Egan, B., and Lithgow, T.: Tom40, the import channel of the mitochondrial outer membrane, plays an active role in sorting imported proteins. EMBO J, 22(10):2380–2386, May 2003.

169. Gabriel, K., Buchanan, S., and Lithgow, T.: The alpha and the beta: protein translocation across mitochondrial and plastid outer membranes. Trends Biochem Sci, 26(1):36–40, Jan 2001.

170. Neupert, W.: Protein import into mitochondria. Annu Rev Biochem, 66:863–917, 1997.

171. Voos, W., Martin, H., Krimmer, T., and Pfanner, N.: Mechanisms of protein translocation into mitochondria. Biochim Biophys Acta, 1422(3):235–254, Nov 1999.

172. Endo, T. and Yamano, K.: Transport of proteins across or into the mitochondrial outer membrane. Biochim Biophys Acta, 1803(6):706–714, Jun 2010.

173. Andres, C., Agne, B., and Kessler, F.: The TOC complex: preprotein gateway to the chloroplast. Biochim Biophys Acta, 1803(6):715–723, Jun 2010.

174. Li, H. and Chiu, C.: Protein transport into chloroplasts. Annu Rev Plant Biol, 61:157–180, Jun 2010.

175. Bryson, K., McGuffin, L., Marsden, R., Ward, J., Sodhi, J., and Jones, D.: Protein structure prediction servers at university college london. Nucleic Acids Res, 33(Web Server issue):W36–8, Jul 2005.

176. Meier, W., Nardin, C., and Winterhalter, M.: Reconstitution of channel proteins in (polymerized) ABA triblock copolymer membranes this work was supported by the swiss national science foundation. we thank dr. t. hirt and dr. j. leukel for the

synthesis of the triblock copolymer, dr. p. van gelder and dr. f. dumas for bright and helpful discussions, and t. haefele for his contribution to the experimental part. Angew Chem Int Ed Engl, 39(24):4599–4602, Dec 2000.

177. Choi, H., Germain, J., and Montemagno, C.: Effects of different reconstitution procedures on membrane protein activities in proteopolymersomes. Nanotechnology, 17:1825–1830, 2006.

178. Muhammad, N., Dworeck, T., Fioroni, M., and Schwaneberg, U.: Engineering of the e. coli outer membrane protein fhua to overcome the hydrophobic mismatch in thick polymeric membranes. J Nanobiotechnology, 9(1):8, 2011.

179. Chen, M., Khalid, S., Sansom, M., and Bayley, H.: Outer membrane protein g: Engineering a quiet pore for biosensing. Proc Natl Acad Sci U S A, 105(17):6272–6277, Apr 2008.

180. Waldispuhl, J., O'Donnell, C., Devadas, S., Clote, P., and Berger, B.: Modeling ensembles of transmembrane beta-barrel proteins. Proteins, 71(3):1097–1112, May 2008.

181. Adamian, L., Naveed, H., and Liang, J.: Lipid-binding surfaces of membrane proteins: Evidence from evolutionary and structural analysis. Biochim Biophys Acta, 1808(4):1092–1102, Apr 2011.

182. Langosch, D. and Arkin, I.: Interaction and conformational dynamics of membrane-spanning protein helices. Protein Sci, 18(7):1343–1358, Jul 2009.

183. Elofsson, A. and von Heijne, G.: Membrane protein structure: prediction versus reality. Annu Rev Biochem, 76:125–140, 2007.

184. Pautsch, A. and Schulz, G.: Structure of the outer membrane protein a transmembrane domain. Nat Struct Biol, 5(11):1013–1017, Nov 1998.

185. Wallin, E. and von Heijne, G.: Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci, 7(4):1029–1038, Apr 1998.

186. Arkin, I. and Brunger, A.: Statistical analysis of predicted transmembrane alpha-helices. Biochim Biophys Acta, 1429(1):113–128, Dec 1998.

187. Arnold, T., Poynor, M., Nussberger, S., Lupas, A., and Linke, D.: Gene duplication of the eight-stranded beta-barrel ompx produces a functional pore: a scenario for the evolution of transmembrane beta-barrels. J Mol Biol, 366(4):1174–1184, Mar 2007.

188. Zeth, K.: Structure and evolution of mitochondrial outer membrane proteins of beta-barrel topology. Biochim Biophys Acta, 1797(6-7):1292–1299, Jun 2010.

189. Butler, T., Pavlenok, M., Derrington, I., Niederweis, M., and Gundlach, J.: Single-molecule DNA detection with an engineered mspa protein nanopore. Proc Natl Acad Sci U S A, 105(52):20647–20652, Dec 2008.

190. Ayalew, S., Confer, A., Hartson, S., and Shrestha, B.: Immunoproteomic analyses of outer membrane proteins of mannheimia haemolytica and identification of potential vaccine candidates. Proteomics, 10(11):2151–2164, Jun 2010.

191. Pinne, M. and Haake, D.: A comprehensive approach to identification of surface-exposed, outer membrane-spanning proteins of leptospira interrogans. PLoS One, 4(6):e6071, 2009.

192. Boyce, J., Cullen, P., Nguyen, V., Wilkie, I., and Adler, B.: Analysis of the pasteurella multocida outer membrane sub-proteome and its response to the in vivo environment of the natural host. Proteomics, 6(3):870–880, Feb 2006.

193. Jimenez-Morales, D., Adamian, L., and Liang, J.: 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, 2008.

194. Jimenez-Morales, D. and Liang, J.: Pattern of amino acid substitutions in transmembrane domains of beta-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria. PLoS One, 6(11):e26400, 2011.

195. Shoshan-Barmatz, V., De Pinto, V., Zweckstetter, M., Raviv, Z., Keinan, N., and Arbel, N.: VDAC, a multi-functional mitochondrial protein regulating cell life and death. Mol Aspects Med, 31(3):227–285, Jun 2010.

196. Shoshan-Barmatz, V., Israelson, A., Brdiczka, D., and Sheu, S.: The voltage-dependent anion channel (VDAC): function in intracellular signalling, cell life and cell death. Curr Pharm Des, 12(18):2249–2270, 2006.

197. Lemasters, J. and Holmuhamedov, E.: Voltage-dependent anion channel (VDAC) as mitochondrial governor–thinking outside the box. Biochim Biophys Acta, 1762(2):181–190, Feb 2006.

198. Werhahna, W., Jnschb, L., and Braun, H.: Identification of novel subunits of the tom complex from arabidopsis thaliana. Plant Physiology and Biochemistry, 41(5):407 – 416, 2003.

199. Kauko, A., Hedin, L., Thebaud, E., Cristobal, S., Elofsson, A., and von Heijne, G.: Repositioning of transmembrane alpha-helices during membrane protein folding. J Mol Biol, 397(1):190–201, Mar 2010.

200. Lin, M., Zhang, J., Lu, H., Chen, R., and Liang, J.: Constrained proper sampling of conformations of transition state ensemble of protein folding. J Chem Phys, 134(7):075103, Feb 2011.

201. PDB: Pdb-101. http://www.pdb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/resolution.html.

202. Baker, M.: Making membrane proteins for structures: a trillion tiny tweaks. Nat Methods, 7(6):429–434, Jun 2010.

203. Monastyrskyy, B., Fidelis, K., Tramontano, A., and Kryshtafovych, A.: Evaluation of residue-residue contact predictions in CASP9. Proteins, Aug 2011.

204. Ou, Y., Chen, S., and Gromiha, M.: Prediction of membrane spanning segments and topology in beta-barrel membrane proteins at better accuracy. J Comput Chem, 31(1):217–223, Jan 2010.

205. Sela, M., White, Jr., F., and Anfinsen, C.: Reductive cleavage of disulfide bridges in ribonuclease. Science, 125(3250):691–692, Apr 1957.

206. Anfinsen, C.: Principles that govern the folding of protein chains. Science, 181(96):223–230, Jul 1973.

207. Dill, K.: Polymer principles and protein folding. Protein Sci, 8(6):1166–1180, Jun 1999.

208. Dill, K., Bromberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P., and Chan, H.: Principles of protein folding–a perspective from simple exact models. Protein Sci, 4(4):561–602, Apr 1995.

209. Brooks, 3rd, C., Gruebele, M., Onuchic, J., and Wolynes, P.: Chemical physics of protein folding. Proc Natl Acad Sci U S A, 95(19):11037–11038, Sep 1998.

210. Onuchic, J. and Wolynes, P.: Theory of protein folding. Curr Opin Struct Biol, 14(1):70–75, Feb 2004.

211. Ni, D., Wang, Y., Yang, X., Zhou, H., Hou, X., Cao, B., Lu, Z., Zhao, X., Yang, K., and Huang, Y.: Structural and functional analysis of the beta-barrel domain of bama from escherichia coli. FASEB J, 28(6):2677–2685, Jun 2014.

212. Shapovalov, M. and Dunbrack, Jr., R.: A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure, 19(6):844–858, Jun 2011.

213. Evanics, F., Hwang, P., Cheng, Y., Kay, L., and Prosser, R.: Topology of an outer-membrane enzyme: Measuring oxygen and water contacts in solution NMR studies of pagp. J Am Chem Soc, 128(25):8256–8264, Jun 2006.

214. Tang, K., Zhang, J., and Liang, J.: Fast Protein Loop Sampling and Structure Prediction Using Distance-Guided Sequential Chain-Growth Monte Carlo Method. PLoS Computational Biology, 10(4), 2014.

VITA

Wei Tian

Education

- 2019 PhD (Bioinformatics) University of Illinois at Chicago (UIC), USA
- 2009 BS (Computer Science) Shanghai Jiao Tong University (SJTU), China

Experience

- 2011–2019 Research assistant, Bioengineering, UIC
- 2011 Teaching assistant, Bioengineering, UIC
- 2018–2019 Predoctoral scientist, Institute of Science and Technology Austria
- 2015 Research assistant, Toyota Technological Institute at Chicago
- 2009–2011 Student researcher, SJTU
- 2008 Student intern, Microsoft Research Asia

Publications

Co–first authorship marked by *

- **W Tian**, M Lin, K Tang, J Liang, H Naveed. High–resolution structure prediction of β–barrel membrane proteins. *Proceedings of the National Academy of Sciences*, 115 (7), 1511–1516, 2018
- **W Tian**, C Chen, X Lei, J Zhao, J Liang. CASTp 3.0: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Research*, 46(W1), W363W367, 2018

- Y Yang, R Guo, K Gaffney, M Kim, S Muhammednazaar, **W Tian**, B Wang, J Liang, H Hong. Folding–Degradation Relationship of a Membrane Protein Mediated by the Universally Conserved ATP–dependent Protease FtsH, *Journal of the American Chemical Society*, 140(13):4656–4665, 2018
- **W Tian**, M Lin, H Naveed, J Liang. Efficient computation of transfer free energies of amino acids in beta–barrel membrane proteins. *Bioinformatics*, 33 (11), 1664–1671, 2017
- **W Tian***, X Lei*, LH Kauffman, J Liang. A knot polynomial invariant for analysis of Topology of RNA Stems and Protein Disulfide Bonds. *Molecular Based Mathematical Biology*, 5 (1), 21–30, 2017
- A Ismael, **W Tian**, N Waszczak, X Wang, Y Cao, D Suchkov, E Bar, M Metodiev, J Liang, R Arkowitz, D Stone. Gβ promotes pheromone receptor polarization and yeast chemotropism by inhibiting receptor phosphorylation. *Science Signaling*, 9 (423), ra38–ra38, 2016
- X Lei, **W Tian**, H Zhu, T Chen, P Ao. Biological Sources of Intrinsic and Extrinsic Noise in cI Expression of Lysogenic Phage Lambda. *Scientific Reports*, 5, 13597–13597, 2015
- **W Tian**, Y Cao, A Ismael, D Stone, J Liang. Roles of regulated internalization in the polarization of cell surface receptors. *Conf Proc IEEE Eng Med Biol Soc.*, 1166–1169, 2014
- J Zhao, H Naveed, S Kachalo, Y Cao, **W Tian**, J Liang. Dynamic mechanical finite element model of biological cells for studying cellular pattern formation. *Conf Proc IEEE Eng Med Biol Soc.*, 4517–4520, 2013
- H Zhu, T Chen, X Lei, **W Tian**, Y Cao, P Ao. Understand the noise of CI expression in phage λ lysogen. *31st Chinese Control Conference* (CCC), 7432–7436, 2012
- **W Tian**, H Zhu, X Lei, P Ao. Extrinsic vs. intrinsic noises in Phage lambda genetic switch. *Systems Biology (ISB) 2011 IEEE International Conference*, 67–71, 2011