Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 541359, 10 pages
http://dx.doi.org/10.1155/2013/541359

Research Article

A Topology Structure Based Outer Membrane Proteins Segment Alignment Method

Han Wang,[1,2] Bo Liu,[3] Pingping Sun,[1,2,4] and Zhiqiang Ma[1,2]

[1] *School of Computer Science and Information Technology, Northeast Normal University, Changchun 130024, China*
[2] *Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China*
[3] *School of Physical Education, Northeast Normal University, Changchun 130117, China*
[4] *National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun 130024, China*

Correspondence should be addressed to Zhiqiang Ma; mazq@nenu.edu.cn

Received 15 July 2013; Accepted 7 September 2013

Academic Editor: William Guo

Outer membrane proteins (OMPs) are transmembrane proteins (TMPs) located in outer membranes. These proteins perform diverse biochemical functions and have immediate medical relevance, so that their spatial structures are important for studying. But the special physicochemical properties of OMP make it hard to obtain their structures experimentally. For the purpose of predicting OMP structures, discriminating OMPs and aligning their sequences to native structures are indispensable steps. We developed a novel method OMSA (Outer Membrane Segment Alignment), which implemented both steps in one program. OMSA integrates OMP-specific topology features to implement a sequence-to-structure alignment, for example, segment type and segment orientation, while a segment-dependent gap penalty model is employed to improve the alignment. Compared to peer top-leading methods, OMSA achieved higher accuracy in both OMP discrimination and alignment, which may further improve OMP structure studying.

1. Introduction

Outer membrane proteins (OMPs) are the transmembrane proteins (TMPs) found in outer membranes of cell, mitochondria, or plastids. They perform diverse biochemical functions such as active ion transport, passive nutrient uptake and intake, or enzymatic activity and structural anchoring [1]. Meanwhile, they also are the potential targets for antimicrobial drugs and vaccines [2–4]. Native structures of OMP are extremely scarce in Protein Data Bank (PDB) [5], where hundreds of OMPs account for no more than 2% of all solved structures [6]. However, there are thousands of OMPs estimated to be existing in the genomic databases currently [7], and the number is increasing by ongoing large-scale sequencing [8].

Computational structure prediction provides a practical approach to bridge the gap between sequences and structures. A few efforts have been made to predict the general structures of membrane proteins [9, 10], but the prediction accuracy remains to be further improved. Some methods that focused on the family of G-protein-coupled receptors (GPCRs) [11, 12] achieved better performances. However, those methods perform poorly on OMPs for lacking of homologous templates, predicting OMP structures is still a challengeable problem. For the purpose, the above all task is to accurately discriminating the OMP from sequences and aligning them to solved structures.

There are many methods that effort to discriminate the OMP from sequence databases; Gromiha and Suwa used OMP motifs for excluding and identifying OMP form globular proteins [13], and they further compared performances of several machine learning based methods; the highest accuracy is no more than 91% [14]. By contrast, few OMP alignment methods have been reported despite its importance for fold recognition [15]. The structure prediction of OMPs is estimated to obtain the accuracy as high as that of globular proteins under the condition that the alignment of OMPs achieves similar accuracy to its compeer [16], but alignment methods [17–21] for globular protein did not work well on OMPs, the reason being mainly because those

sequence-to-structure methods could not efficiently abstract the OMP-specific features used in the alignment process.

It is notable that OMPs have special physicochemical properties compared to globular proteins. They are composed of several beta strands which form a barrel to span outer membrane. These strands are enriched polar, charged, and hydrophobic residues, antiparallel crossing membrane alternatively, from one side of outer membrane to the other side. In addition, the number of transmembrane (TM) strands is mostly even. These remarkable properties can be clearly represented by topology structure of OMPs. Therefore, topology structure is considered helpful improving the OMP alignment as OMP-specific feature. Currently, several OMP topology structure predictors are available, such as TMBHMM [22], TMBETAPRED-RBF [23], and TMBpro [24]; these methods utilize many OMP-specific features to improve the prediction accuracy, including amino acid composition, alternating hydrophobicity pattern [25], or "positive-inside" rule [26, 27].

In this study, we developed a novel Outer Membrane Segment Alignment method (OMSA) to discriminate and align OMPs from sequences at the same time. The method for the first time used OMP-specific features abstracted from topology structure to improve the alignment, for example, TM segment type and TM segment orientation. Particular scoring function designed for OMP was applied to a local-global dynamic programming to optimize the alignment. Comprehensive OMP datasets have been used for training and testing, and the results represented that OMSA perform well on OMP discrimination and alignment compared to peer methods. Hence, our method will be useful for OMP structure studying.

2. Materials and Methods

2.1. Datasets. Orientations of Proteins in Membranes (OPM) database [28] was used in OMSA training and testing; it provides the most comprehensive collection of membrane proteins with calculated spatial arrangements. Differing to computational-based databases [29, 30], OPM database is more in agreement with the experimental data and further classifies the membrane proteins based on their main transmembrane domains by referencing SCOP [31] and TCDB [32]. In this database, 98 entries are classified to 26 superfamilies, and each of them is composed of one or more protein families. Here, proteins in the same superfamily are evolutionarily related and with superimposable tertiary structures, but in low sequence identity, while it is high among the proteins in the same family. We randomly picked two entries from each superfamily to comprise training and testing datasets, respectively. For those superfamilies which have only one entry, the entries were selected to training dataset. Finally, the training dataset is composed of 19 nonredundant entries, while testing dataset has 28 nonredundant entries (see Table S1 in supplementary material available online at http://dx.doi.org/10.1155/2013/541359).

For the purpose of benchmarking the performance of OMP discrimination, Gromiha and Suwa's dataset (GS-dataset) [13] is used, which includes 377 OMPs, 268 α-helical transmembrane proteins, and 674 globular protein chains. All these well-annotated transmembrane proteins included in the dataset were obtained from PSORT-B database [33], while those globular protein chains were obtained from the PDB40D_1.37 database of SCOP [34]. In this dataset, a few transmembrane proteins are homologous, and the globular proteins have sequence identity less than 30%.

2.2. Segment Type and Orientation. Segments on an OMP sequence are classified to TM segment (TMB) and non-TM segment according to their location relative to outer membrane. TM segment is the part of sequence that is across the outer membrane, and non-TM segments include outside segment and inside segment, where the outside segment locates outside the area surrendered by outer membrane, while the inside segment locates in the area. The segment orientation describes the direction of those segments across the outer membrane; it can be abstracted when segment types have been identified.

The segment type seg (i) at sequence position i is given as

$$\text{seg}(i) = \begin{cases} \text{B}, & \text{TMB}, \\ \text{I}, & \text{Inside}, \\ \text{O}, & \text{Outside}, \\ \text{U}, & \text{Unknown}. \end{cases} \tag{1}$$

The orientation of each TM segment is determinated according to the inside/outside it forwards to. TM segment orientation orn(i) at position i is valuated as follows:

$$\text{orn}(i) = \begin{cases} 1, & \text{from inside to outside}, \\ -1, & \text{from outside to inside}, \\ 0, & \text{else}. \end{cases} \tag{2}$$

Given an OMP as shown in Figure 1(a), segment types and orientations are clearly described by topology structure in Figure 1(b), and they have been labeled for each amino acid on its sequence using (1) and (2), respectively, shown in Figure 1(c). In each pairwise alignment, query sequences used predicted topologies, while the template sequences directly utilized native topology structures obtained from OPM database. We employed TMBHMM [22] to predict topology structures for its high performance and accuracy. The method integrates particular features to build a hidden Markov model (HMM) for topology structure predicting, for example, SASA (relative solvent-accessible surface area) feature [35] and frequency profile [36].

2.3. Sequence Profile. Almost all OMPs are homologous to each other [37], so that the sequence profile is especially conductive to identifying the compatibility within the same segment types by sequence patterns. This sequence-based profile generated by PSI-BLAST [36] is called Position Specific Scoring Matrix (PSSM), which is derived by aligning the amino acid sequence against NCBI's nonredundant sequence database (NR) with three iterations. The profiles present the evolutionary conservation of sequences by large-scale searching, whereas it has a significant impact on protein fold

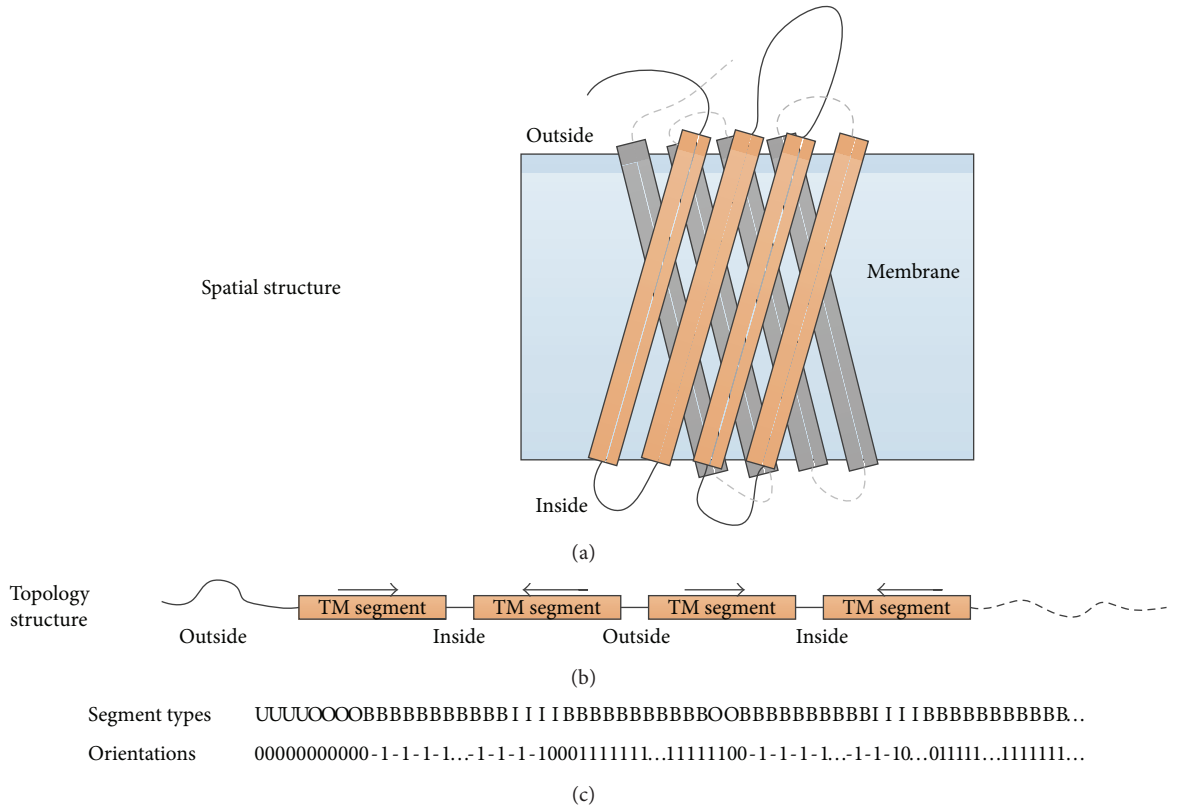

Figure 1: The special conformation of OMPs. (a) is the conformation of a sample OMP, and the rectangles represent the TM segments. Its N-terminus locates on the outside. (b) is the topology structure of the protein, which has been unfolded to sequence; the right arrow means that the orientation of the corresponding TM segment is from outside to inside and vice versa. Meanwhile, the non-TM segments alternatively locate on the outside and inside. (c) The protein sequences represented by segment types and orientations according to (1) and (2).

recognition [38]. A PSSM profile pm$[i, j]$ is a $n \times 20$ log-odds matrix, where the n represents the sequence length. Each element in pm$[i, j]$ negatively represents the frequency of the residue type j at position i.

2.4. Scoring Function. The scoring function is the kernel of dynamic programming (DP) which is used to derive the optimal path searching for alignment. It is composed of two major parts, fitness scoring and gap penalty model. The fitness scoring responds to measure the compatibility between any residue pair, while the gap penalty model affects the alignment accuracy by controlling the gap insertion. They coordinate with each other making a balance to achieve the best alignment accuracy. Tailing for the OMP, the scoring function used here is different from that of globular proteins, the fitness scoring adopts the OMP-specific features and correspondingly strategies, and the gap penalty is thereby particularly designed to support the OMP segment alignment. Here, the integrated three features, segment type, segment orientation, and sequence profile, compose a compact profile for the profile-to-profile scoring, where segment type guarantees the proteins are aligned by segment and segment orientation prevents the TM segments being incorrectly aligned, while sequence profile further improves the alignment accuracy between non-TM segments. Meanwhile, a segment-dependent gap model is employed correspondingly.

(1) Fitness Scoring. Differing to the sequence-to-sequence alignment method, the profile-to-profile alignment offers the more opportunities to align the proteins according to their overall properties. The so-called profile is the general conception of assemble of the selected features, and those features must be efficient to describe the properties of target proteins and less redundant to decrease the computational complex. The selected three features can overall describe the sequence patterns (by sequence profile) and conformation properties (by segment type and TM segment orientation) of OMPs.

The fitness score of ith position on query sequence and jth position on template sequence is calculated by the equation

$$\begin{aligned} \text{Fitness}(i, j) = {} & w_1 \text{PRO}(i, j) - w_2 \text{SEG}(i, j) \\ & - w_3 \text{ORN}(i, j) + w_{\text{shift}}, \end{aligned} \tag{3}$$

where PRO(i, j) is the fitness score of sequence profile, SEG(i, j) is the segment type fitness score, and ORN(i, j) is the segment orientation fitness score; w_1, w_2 and w_3 are the weights of the three fitness scores, while w_{shift} is a to-be-determinated constant that avoids the unrelated residues

aligned [39]. The segment type fitness score can be simply computed by

$$\mathrm{SEG}(i,j) = \begin{cases} 2, & \text{if } \mathrm{seg}(i) = \mathrm{seg}(j) = \text{“}B\text{”}, \\ 1, & \text{if } \mathrm{seg}(i) = \mathrm{seg}(j) \neq \text{“}B\text{”}, \\ -1, & \text{else.} \end{cases} \quad (4)$$

Segment orientation fitness score is calculated according to the equation

$$\mathrm{ORN}(i,j) = \begin{cases} 1, & \mathrm{orn}(i) = \mathrm{orn}(j) \text{ and } \mathrm{orn}(i) \neq 0 \text{ and } \mathrm{orn}(j) \neq 0, \\ -1, & \mathrm{orn}(i) \neq \mathrm{orn}(j) \text{ and } \mathrm{orn}(i) \neq 0 \text{ and } \mathrm{orn}(j) \neq 0, \\ 0, & \text{else.} \end{cases} \quad (5)$$

The evolution fitness score is calculated according to the equation

$$\mathrm{PRO}(i,j) = \sum_{k=0}^{20} \left(\mathrm{pm}_{\mathrm{query}}[i,k] \times \mathrm{pm}_{\mathrm{template}}[j,k] \right), \quad (6)$$

where $\mathrm{pm}_{\mathrm{query}}[i,k]$ is PSSM profile value of residue type k at position i on the target sequence, and $\mathrm{pm}_{\mathrm{template}}[j,k]$ follows the same meaning.

(2) *Gap Penalty*. Gap penalty is used to evaluate the cost of making an insertion (or deletion) option; the balance between fitness score and the gap penalty makes the decisions for the DP to trace the optimal aligning path. Various gap penalty models have been designed for globular protein alignment previously, for example, position-dependent gap penalty models used in [17, 39] or profile-based model [40]. Even a more complicate model [41] was used recently for low-homology protein threading. In this study, we employed a segment-dependent gap penalty model to satisfy the segment alignment, in which insertions in TM segments are more stricterly punished compared to non-TM segments, because these segments are more conserved. Considering the predictions accuracy of topology, we do not simply forbid the insertions as H. Zhou and Y. Zhou [42] did for globular protein alignment but allow the gaps insert to query sequence using open-gap-penalty $\mathrm{op}_{\mathrm{tm}}$ and one-time-penalty $\mathrm{ep}_{\mathrm{tm}}$, while they are forbidden to template sequences which use the native topology structures. Similarly, open-gap-penalty $\mathrm{op}_{\text{non-tm}}$ and one-time-penalty $\mathrm{ep}_{\text{non-tm}}$ are used for non-TM segments of query sequence. Here, the open-gap-penalty is used only when the gap opens, and the one-time-penalty is used for the continuous insertion after the gap opened.

2.5. Training Parameters. All parameters, w_1, w_2, w_3, w_{shift}, $\mathrm{op}_{\mathrm{tm}}$, $\mathrm{op}_{\text{non-tm}}$, $\mathrm{ep}_{\mathrm{tm}}$, and $\mathrm{ep}_{\text{non-tm}}$, used in the scoring function are trained using the same method as [42] on our training dataset. All the parameters are randomly assigned the start value and then optimized by grid search. Here, the gold standard TM-score [43] is used to supervise the searching. The higher TM-score derived by aligned sequences is considered the higher accuracy achieved. The iterations exit when the average TM-score does not increase any more.

2.6. Dynamic Programming. DP is the most popular paradigm in computational biology [44]. It is a method for solving complex problems by breaking them down into simpler subproblems and has been applied widely in sequence alignment or optimal searching. Decided by our scoring function, the DP process of OMSA is OMP specific, which can align the correct segments as much as possible. We use local-global algorithm to optimize the alignment path for requirement of segment alignment. By using the OMP-specific scoring function introduced above, the DP procedure of OMSA can find better path for an alignment. The segments with the same type are aligned preferentially, while different segment types are hard to match unless they are extremely compatible with the evolutionary conservation.

3. Results and Discussion

In this section, we will firstly show that the segment orientation significantly improves the alignment accuracy and then compare the OMSA to one of the top-leading generated profile-to-profile alignment methods, HHalign [45]. As the outputs of sequence-to-structure alignment, rawscores generated by OMSA negatively relate to the structural similarities. However, only the rawscores derived by the same query protein are comparable; they cannot be directly used as a criterion to evaluate the structure similarity between different alignment pairs. Thus, we transform the rawscores to rawscore correlation (RC), which bridges the rawscores and the structure similarity by normalizing the rawscores to interval (0, 1]. Finally, the RC is applied to OMP discrimination and the OMP fold recognition.

3.1. Topology Features Improve the Alignment Accuracy. OMSA improves alignment principally relied on topology features. We implemented method nOMSA that removed topology features for comparison, where the nOMSA used the secondary structure and sequence profile as features and trained using the same dataset. We made all-versus-all pairwise alignments on testing dataset using OMSA and nOMSA, respectively; average GDT_TS [46, 47] was used to evaluate the alignment accuracy. GDT_TS scores the structure similarity for two length-equal proteins using their 3D structures. The structure similarity is positively related with the GDT_TS when the score increases from 0 to 1. Therefore, for each alignment pair, the higher GDT_TS means the better alignment that has been made by the method, whereas the overall alignment accuracy of a query protein can be described using the average GDT_TS of all the templates.

Each query, there were aligned to 27 templates, and the average GDT_TS was statistic using the corresponding 27 GDT_TSs. The average GDT_TS of all the 28 queries derived by OMSA are shown using the solid polyline in Figure 2, while the corresponding nOMSA results are shown using dotted polyline, where the solid polyline lies obviously above the dotted polyline, which indicates that all the query proteins will obtain higher alignment accuracy when the topology features are involved.

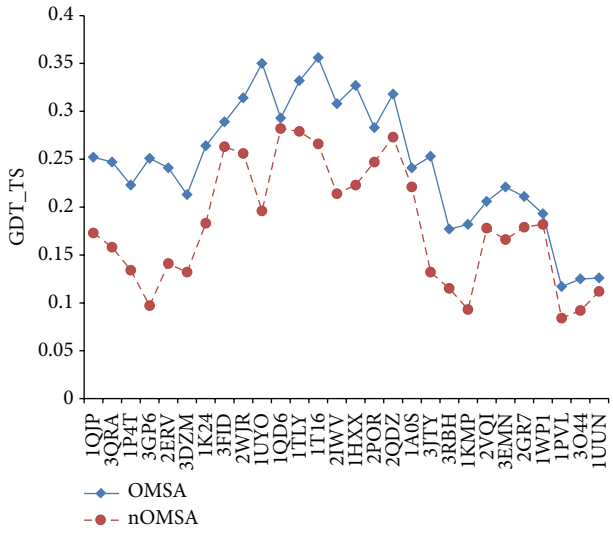

Figure 2: Feature of segment orientation improves the alignment accurate. The alignments of OMSA are more accuracy for all the testing proteins shown using solid polyline in blue than those of nOMSA (non-segment-orientation based). The results are derived based on the all-versus-all pairwise alignments on the testing dataset.

3.2. Performance of OMSA Alignment. Generally, comparison with the peer methods powerfully proves the improvements of the new methods; however, there is no such OMP pairwise alignment method currently available; thus, we chose top-leading general alignment method HHalign [45]. HHalign uses profile hidden Markov model (HMM) to make pairwise HMM-HMM (profile-HMM) alignments, and the confidence values and a full seven-state secondary structure are employed to improve the alignment quality; thus, it is a very sensitive repeat-identification tool.

To arrange the comparison, the profile-HMMs of 28 proteins of testing dataset were generated to make the all-versus-all pairwise alignment, in which the default parameters were used. Correspondently, the previous pairwise alignment results derived by OMSA are used for the comparison. Avoiding the pseudo increasing of alignment accuracy, those pairs which the queries align to themselves were excluded; finally, in total 28×27 pairs of alignments were used for comparison.

The alignment accuracy can be evaluated by two criterions: (1) calculating the percentage of correctly aligned positions [15] and (2) scoring the structural similarity between the aligned pairs [48]. For the first approach, each aligned residue pair in the alignment has to be verified whether it belongs to the correct alignment, whereas a golden standard of structural alignment is required, such as widely used method TM-align [49], since there is no unique solution that solves the problem finding the optimal structure alignment [50]. For the second one, structure similarity of alignment is scored directly by their native 3D structures of aligned parts, depicting whether they were correctly aligned or not; GDT_TS [46, 47] and TM-score [43] are commonly used for the purpose. Notably, TM-score is designed to be independent of protein lengths, and the structures with a score higher than 0.5 assume roughly the same fold [51], while it indicates that the proteins are unrelated when the score is below 0.20. To comprehensively show the performances, we adopted both approaches to exhibit the alignment results, where the alignment accuracy (ACC) was used according to approach (1) and TM-score and GDT_TS were used according to approach (2).

As shown in Table 1, OMSA derived the better alignment in average against the testing dataset compared to the HHalign. Here, the accuracy of the TM segments and the non-TM segments is counted separately, and the overall accuracy is also given. The alignment accuracy is obtained using the TM-align as the standard. The HHalign aligns both types of segments with almost the same accuracy, while the OMSA shows much difference in accuracy between them. There is more than eleven percentage margin that the OMSA surpasses the HHalign in the alignment accuracy of the TM parts, while the margin is just no more than three percent of the non-TM parts; for this reason, the OMSA achieves the 50.6% of overall alignment accuracy and surpasses its companion nearly eight percent. The result is consistent with the TM-score and the GDT_TS, where the alignments for TM segments are much better than non-TM segments. Correspondingly, the improvement of TM-score achieves the eight percent, as well nine percent improvement of GDT_TS.

Comparing the integrated features, sequence profile has been commonly used in both methods, but the HHalign addresses the predicted secondary structure in alignment, while OMSA uses topology-based features instead. Obviously, the utilization of topology structure based features improves alignment accuracy than secondary structure, especially for the segments alignment methods. However, it lacks to determinate the structures similarity for non-TM segments which have secondary structures and decreases the alignment accuracy for non-TM segments, but this shortness has been obscured by the less secondary structures existing outside TM segments. The results illuminate that the TMB orientation improves the OMP segment alignment by decreasing the misleading of the incorrect topology prediction and further distinguishing the TM segments and thereby improves the alignment performance.

3.3. Alignment Rawscore and Structural Similarity. The OMSA generates a rawscore for each alignment pair, and its value directly reflects the structure similarity between the proteins. The aligned pair with smaller rawscore illustrates that the two proteins are more likely having similar spatial conformations, which is exhibited in Figure 3 using the alignment results of Porin (PDB_ID:2POR) [52].

The left-top point represents the alignment result of 2POR with itself, so that it totally matched in structure which is presented by GDT_TS of value 1 and derived the smallest negative rawscore at the same time. The nearest point represents protein OmpF porin (PDB_ID:1HXX), which belongs to the same superfamily but not the same family with 2POR. The two proteins have the most similar conformations with each other within the testing dataset. In similar manner, the points in the left-top area are mostly the proteins that have similar

Table 1: Comparison of alignment accuracy with HHalign. The performances are compared separately within TM segments, non-TM segments, and overall sequence. OMSA achieved the best alignment accuracy in all the fields, and that within the TM segments has significantly surpassed the general alignment program HHalign. It indicates that the segment orientation further improves the alignment for OMPs.

Methods	Acc (%)			TM-score			GDT_TS		
	TM	Non-TM	Overall	TM	Non-TM	Overall	TM	Non-TM	Overall
OMSA	53.3	45.1	**50.6**	0.403	0.314	**0.322**	0.335	0.295	**0.307**
HHalign	42.9	42.6	42.7	0.217	0.256	0.242	0.209	0.226	0.216

TM: TM segments; Non-TM: non-TM segments; Acc: accuracy.

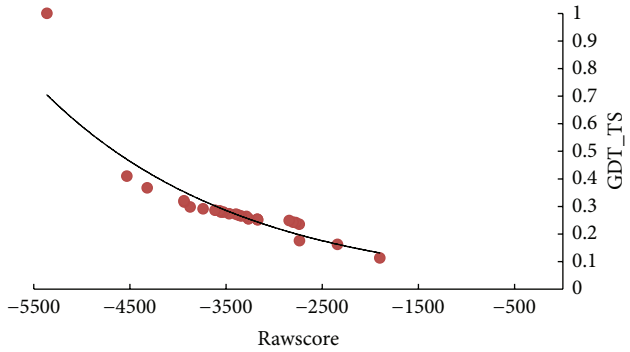

Figure 3: An example of 2POR shows rawscores negatively related to structure similarity. The structure similarity is represented using GDT_TS. The rawscores generated by OMSA negatively relate to the 3D structure similarity between the aligned proteins, where a smaller score indicates that the corresponding protein is similar to the query protein in spatial conformation.

structures with 2POR (16 TM segments each chain), such as Fatty Acid Transporter FadL (PDB_ID:1T16, 14 TM segments) [53], Porin OmpG (PDB_ID:2IWV, 14 TM segments) [54], and BenF-like porin (PDB_ID:3JTY, 18 TM segments) [55]. These proteins are similar to the topology structures, but it is not the only driving force for OMSA to recognize the proteins; even the difference among the same-type segments can be further distinguished, such as Alginate export protein (PDB_ID:3RBH) [56], which is reported recently with 18 TM segments, resembling to 2POR in topology structure but significantly differing in non-TM segments. Most proteins centralized in the middle area superimpose their 3D structures to 2POR partially, whereas they derived higher rawscores, while the other proteins fell to the right-bottom corner are totally different with 2POR, such as Leukocidin F (PDB_ID:1PVL) [57] and OprM (PDB_ID:1WP1) [58]. All the testing proteins showed similar rawscore distribution as 2POR; the results illuminate that the rawscore of OMSA negatively responds to the structure similarity.

3.4. Rawscore Correlations. Since the rawscores produced by OMSA are negatively related to the protein structural similarity, they are able to be used to discriminate the OMPs from the globular proteins. However, the comparison of the rawscores can be done only among those produced by aligning to the same query protein, otherwise, the rawscores are not comparable. The reason is easy to understand; assuming aligning two pairs of proteins at the same time, proteins in one pair have short sequences and very similar structures, while their companions have much longer sequences and less similar structures; even when the proteins in shorter pair are perfectly aligned, they still may not obtain the smaller rawscore; it is decided by the alignment between the other two proteins. Considering the homologous among the OMPs, two proteins that have similar sequence length tend to have more similar local structures, so that proteins in longer pair have more chance to derive the smaller rawscore. Therefore, the rawscore cannot be directly used for the purpose.

It has been noticed that each query protein will achieve the best rawscore aligning to itself; the other rawscores are all smaller than the value and decrease almost according to the structure similarity, so that rawscore correlation (RC) between query protein and templates can be described by the relative values of rawscores. The RC of query protein i and template protein j is calculated using the equation

$$\mathrm{RC}(i, j) = \frac{\mathrm{rs}(i, j)}{\mathrm{rs}(i, i)}, \tag{7}$$

where $\mathrm{rs}(i, j)$ and $\mathrm{rs}(i, i)$ are, respectively, the alignment rawscores of responding proteins. All the rawscores are negative numbers, so the RC is a positive number and belongs to $(0, 1]$. With the transformation, the rawscores are kind of normalized. The ranking of rawscores has not been changed for each query protein, but the structure similarity between the different query-template pairs becomes comparable.

To exhibit the correlation between rawscores and structures, all the testing pairwise alignment rawscores were transformed to RC value, each blue point in Figure 4 presents a RC value, and each column is composed of 28 points, which are the RCs of 28 templates (including itself). All the proteins are best matched with themselves, so the RC values are always 1 on the top of each column accordingly, and the others are smaller than 1. For the reason that we did not change the orders of rawscores, the RC values distribute similarly with rawscores, the primary function of the diagram is to present the commonness among the columns. It can be found that most RC values of each column concentrate between 0.45 and 0.8, and fewer points scatter above or under the region. The phenomenon indicate that most templates have some parts obviously similar with query protein in conformation, and fewer templates are significantly similar or different to that. Meanwhile, the last three columns which correspond to Leukocidin F (PDB_ID:1PVL) [57], Cytolysin and hemolysin HlyA Pore-forming toxin (PDB_ID:3O44) [59], and Porin

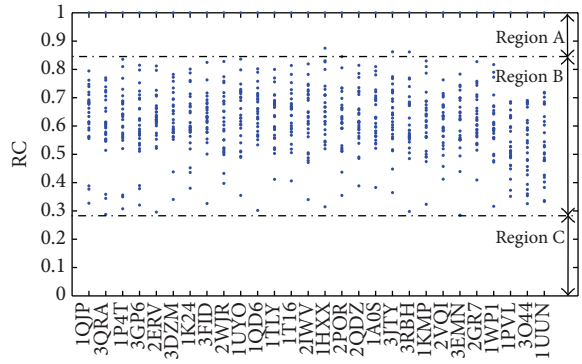

Figure 4: The rawscore correlations of testing proteins. All the RC values of the testing proteins are pointed in blue and shown in separate columns. Those RC = 1 points respond to the alignments of queries with themselves. According to the 3D structure similarity, RC values of testing proteins fall into three different regions: region A (proteins are in the same superfamily), region B (proteins are OMPs), and region C (proteins are not OMP). Therefore, RC can be further used to measure the similarity of 3D structures instead of rawscore.

MspA (PDB_ID:1UUN) [60] show the RC distributions slightly different that the average RC value smaller than the other proteins. The three proteins are all special OMPs that have particular conformations different from other proteins, so that the OMSA could not obtain the better alignment for them. It is also the partial reason that result the smaller RC values in other columns. The distribution illuminates the RC value responds to the structure similarity between the query and templates.

Although the structure differences exist among the OMPs, the RC values still remain bigger than 0.28 with each other, which indicates that they have common properties at least in some parts. Meanwhile, the scanty two superfamilies that have two proteins each are observed from Figure 4; all the four proteins derive RC values bigger than 0.84 and are the top 4 among the alignments between different query and template. According to the results, we divided the RC value into three intervals: region A: (0.84, 1], region B: (0.28, 0.84], and region C: (0, 0.28]. The proteins in region A are considered to share the same superfamily with corresponding query protein, those in region B can be regarded as OMPs, and those in region C are the non-OMPs. Because the OMSA effectively amplified the OMP-specific features, the non-OMPs are hard to achieve a bigger RC value than 0.28. The performance of discriminating OMPs is shown in the next section.

3.5. Improving OMP Discrimination. We adopted the GS-dataset to the benchmark the performance of OMSA to discriminate the OMPs, because the dataset was wildly used for the purpose as benchmark by the other OMP discrimination methods, such as DD [13], NN_AAC [14], and SVM_AAC_DPC [61]. As known, the testing dataset is nonredundant and covers all the OMP superfamilies; we used it as the OMP library. For each protein in GS-dataset, we calculated its RC values aligning to all the proteins of our OMP library; if the 26 of 28 RC values are bigger than 0.28, the corresponding protein is considered as the OMP.

Table 2: Performance of OMP discrimination methods.

Methods	MCC	Acc (%)	SN (%)	SP (%)
OMSA	0.896	95.4	86.4	99.8
DD	0.541	82.4	78.8	83.3
NN_ACC	0.716	91.0	79.3	93.8
SVM_ACC_DPC	0.816	93.9	90.9	94.7

The results are respectively cited from [13, 14, 61]. MCC: Matthews correlation coefficient; Acc: accuracy; SN: sensitivity, SP: specificity.

To compare the discrimination accuracy, four criterions were used to show the performance: Matthews correlation coefficient (MCC), accuracy (AC), sensitivity (SN), and specificity (SP). As shown in Table 2, OMSA achieves the higher accuracy than DD, NN_ACC, and SVM_ACC_DPC in MCC, AC, and SP. There were 2 false positives (FPs) and 59 false negatives (FNs) in the results of OMSA. It is obvious that our method is highly reliable to determinate OMPs with 95.4% of AC, accompanied by 0.896 of MCC, which is much higher than the second top one of 0.816. The FNs decreased the accuracy of OMSA and also resulted in that the SN was slightly lower than that of SVM_AAC_DPC, but that can be improved by further optimizing the discriminating threshold of RC. Therefore, OMSA has the potential abilities to achieve higher accuracy of OMP discrimination. Notably, it is hard to discriminate the transmembrane strands and beta water soluble proteins for most methods, because both of them share some common features, such as amphipathicity [62]. Benefited by the predicted topology structure, most all-beta water soluble proteins showed no TM segments, whereas the OMSA has been never confused by the two kinds of proteins.

3.6. Improving Fold Recognition. The pairwise alignment results of the testing dataset are used to show the fold recognition performance of our method. For the comparison, we applied top-leading general fold recognition program HHsearch [45] to the same dataset. As mentioned, the testing dataset is redundant but small, all the proteins included are at the fold level, there are only two four proteins at the superfamily level, and no proteins at the family level, whereas the comparison conducted here is slightly different from that among the globular proteins, where the top 3 recognized folds are enough to show the performance on such a small dataset, and the recognition accuracy is compared using the average TM-score. Differing to the usage in alignment accuracy, TM-scores are obtained here using the native structures of the query-template pairs, instead of aligned parts, to show the native structure similarity of recognized folds. To further show the performance, the accuracy of recognizing the best one is exhibited, respectively. As shown in Table 3, OMSA derived the best performance comparing to HHsearch; it recognized the most similar folds for the queries with accuracy of 57.1%, while HHsearch achieved the accuracy above ten percent less, and the results are consistent with native TM-scores. The accuracy with which best protein fold appears within the top 2 recognized folds increases to

Table 3: Comparison of the performance of fold recognition for OMPs with HHsearch.

Methods	Top 1		Top 2		Top 3	
	Acc. (%)	TM-score	Acc. (%)	TM-score	Acc. (%)	TM-score
OMSA	57.1	0.621	66.7	0.576	76.7	0.491
HHsearch	46.4	0.553	60.0	0.512	66.7	0.407

Acc.: accuracy.

66.7% of OMSA, and that remains ten percent more than HHsearch in the top 3.

4. Conclusions

This paper describes a novel segment OMP alignment method, OMSA, which is designed based on OMP-specific features. OMPs have distinct physicochemical properties compared to globular proteins, which is the biggest obstacle for other methods to predict their structures but provides opportunities for our method on the contrary. We extract the segment type and segment orientation as the features from topology structures, combining with sequence profile to comprise the aligning profile, and segment alignment is firstly employed to address the problem. Correspondingly, we reassign the scoring functions to satisfy the requirement, in which the residue-residue compatibility is scored using the OMP-specific features and a segment-dependent gap penalty model is employed. OMSA has been tested based on a nonredundant testing dataset. Compared to HHalign, our method performs well as a sequence-to-structure alignment method, where the feature of segment orientation indispensably improves the alignment accuracy, especially for TM segments. The alignment rawscore of OMSA is observed negatively relating to the structure similarity, so that the method surpassed HHsearch certain percentage of accuracy in fold recognition. Furthermore, RC value which is derived from the rawscore can be directly used to measure the structure similarity between query and template pairs. By using the RC value, OMSA achieved the best accuracy of OMP discrimination compared to other existing methods. For the same reason, the RC can also be applied to the fold recognition by further adjusting the threshold. This state-of-the-art technology will thereby facilitate the discriminating, structure recognizing, and function predicting for OMPs.

Conflict of Interests

The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence their work; there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, this paper.

References

[1] M. Remmert, D. Linke, A. N. Lupas, and J. Söding, "HHomp—prediction and classification of outer membrane proteins," *Nucleic Acids Research*, vol. 37, no. 2, pp. W446–W451, 2009.

[2] R. Jackups Jr. and J. Liang, "Interstrand pairing patterns in β-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction," *Journal of Molecular Biology*, vol. 354, no. 4, pp. 979–993, 2005.

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

[4] R. Pajón, D. Yero, A. Lage, A. Llanes, and C. J. Borroto, "Computational identification of beta-barrel outer-membrane proteins in Mycobacterium tuberculosis predicted proteomes as putative vaccine candidates," *Tuberculosis*, vol. 86, no. 3-4, pp. 290–302, 2006.

[5] H. M. Berman, "The Protein Data Bank: a historical perspective," *Acta Crystallographica A*, vol. 64, no. 1, pp. 88–95, 2007.

[6] H. M. Berman, T. N. Bhat, P. E. Bourne et al., "The Protein Data Bank and the challenge of structural genomics," *Nature Structural Biology*, vol. 7, pp. 957–959, 2000.

[7] W. C. Wimley, "The versatile β-barrel membrane protein," *Current Opinion in Structural Biology*, vol. 13, no. 4, pp. 404–411, 2003.

[8] S. Yooseph, G. Sutton, D. B. Rusch et al., "The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families," *PLoS Biology*, vol. 5, no. 3, article e16, 2007.

[9] S. Kelm, J. Shi, and C. M. Deane, "MEDELLER: homology-based coordinate generation for membrane proteins," *Bioinformatics*, vol. 26, no. 22, pp. 2833–2840, 2010.

[10] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, 2006.

[11] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., "A graph-theory algorithm for rapid protein side-chain prediction," *Protein Science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[12] M. Michino, J. Chen, R. C. Stevens, and C. L. Brooks III, "FoldGPCR: structure prediction protocol for the transmembrane domain of G proteincoupled receptors from class A," *Proteins*, vol. 78, no. 10, pp. 2189–2201, 2010.

[13] M. M. Gromiha and M. Suwa, "A simple statistical method for discriminating outer membrane proteins with better accuracy," *Bioinformatics*, vol. 21, no. 7, pp. 961–968, 2005.

[14] M. M. Gromiha and M. Suwa, "Discrimination of outer membrane proteins using machine learning algorithms," *Proteins*, vol. 63, no. 4, pp. 1031–1037, 2006.

[15] Y. Hu, X. Dong, A. Wu, Y. Cao, L. Tian, and T. Jiang, "Incorporation of local structural preference potential improves fold recognition," *PLoS ONE*, vol. 6, no. 2, Article ID e17215, 2011.

[16] L. R. Forrest, C. L. Tang, and B. Honig, "On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins," *Biophysical Journal*, vol. 91, no. 2, pp. 508–517, 2006.

[17] S. Liu, C. Zhang, S. Liang, and Y. Zhou, "Fold recognition by concurrent use of solvent accessibility and residue depth," *Proteins*, vol. 68, no. 3, pp. 636–645, 2007.

[18] K. Ellrott, J.-T. Guo, V. Olman, and Y. Xu, "Improvement in protein sequence-structure alignment using insertion/deletion frequency arrays," *Computational Systems Bioinformatics*, vol. 6, pp. 335–342, 2007.

[19] J.-Y. Yang and X. Chen, "Improving taxonomy-based protein fold recognition by using global and local features," *Proteins*, vol. 79, no. 7, pp. 2053–2064, 2011.

[20] D. Mittelman, R. Sadreyev, and N. Grishin, "Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments," *Bioinformatics*, vol. 19, no. 12, pp. 1531–1539, 2003.

[21] G. Wang and R. L. Dunbrack Jr., "Scoring profile-to-profile sequence alignments," *Protein Science*, vol. 13, no. 6, pp. 1612–1626, 2004.

[22] N. K. Singh, A. Goodman, P. Walter, V. Helms, and S. Hayat, "TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues," *Biochimica et Biophysica Acta*, vol. 1814, no. 5, pp. 664–670, 2011.

[23] Y.-Y. Ou, S.-A. Chen, and M. M. Gromiha, "Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy," *Journal of Computational Chemistry*, vol. 31, no. 1, pp. 217–223, 2010.

[24] A. Randall, J. Cheng, M. Sweredoski, and P. Baldi, "TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins," *Bioinformatics*, vol. 24, no. 4, pp. 513–520, 2008.

[25] M. Punta, L. R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, and B. Rost, "Membrane protein prediction methods," *Methods*, vol. 41, no. 4, pp. 460–474, 2007.

[26] G. Heijne, "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology," *The EMBO Journal*, vol. 5, pp. 3021–3027, 1986.

[27] G. Von Heijne, "Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule," *Journal of Molecular Biology*, vol. 225, no. 2, pp. 487–494, 1992.

[28] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, "OPM: orientations of proteins in membranes database," *Bioinformatics*, vol. 22, no. 5, pp. 623–625, 2006.

[29] G. E. Tusnády, Z. Dosztányi, and I. Simon, "PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank," *Nucleic Acids Research*, vol. 33, pp. D275–D278, 2005.

[30] G. E. Tusnády, L. Kalmár, and I. Simon, "TOPDB: topology data bank of transmembrane proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D234–D239, 2008.

[31] A. Andreeva, D. Howorth, J.-M. Chandonia et al., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. 1, pp. D419–D425, 2008.

[32] M. H. Saier Jr., M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan, "The transporter classification database: recent advances," *Nucleic Acids Research*, vol. 37, no. 1, pp. D274–D278, 2009.

[33] J. L. Gardy, C. Spencer, K. Wang et al., "PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3613–3617, 2003.

[34] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[35] Y. Park and V. Helms, "On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins," *Bioinformatics*, vol. 23, no. 6, pp. 701–708, 2007.

[36] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[37] T. Arnold, M. Poynor, S. Nussberger, A. N. Lupas, and D. Linke, "Gene duplication of the eight-stranded β-barrel OmpX produces a functional pore: a scenario for the evolution of transmembrane β-barrels," *Journal of Molecular Biology*, vol. 366, no. 4, pp. 1174–1184, 2007.

[38] G. Yona and M. Levitt, "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory," *Journal of Molecular Biology*, vol. 315, no. 5, pp. 1257–1275, 2002.

[39] S. Wu and Y. Zhang, "MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information," *Proteins*, vol. 72, no. 2, pp. 547–556, 2008.

[40] W. Zhang, S. Liu, and Y. Zhou, "SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model," *PLoS ONE*, vol. 3, no. 6, article e2325, 2008.

[41] J. Peng and J. Xu, "Low-homology protein threading," *Bioinformatics*, vol. 26, no. 12, pp. i294–i300, 2010.

[42] H. Zhou and Y. Zhou, "Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments," *Proteins*, vol. 58, no. 2, pp. 321–328, 2005.

[43] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, no. 4, pp. 702–710, 2004.

[44] R. Giegerich, "A systematic approach to dynamic programming in bioinformatics," *Bioinformatics*, vol. 16, no. 8, pp. 665–677, 2000.

[45] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[46] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis, "Processing and analysis of CASP3 protein structure predictions," *Proteins*, supplement 3, pp. 22–29, 1999.

[47] A. Zemla, Č. Venclovas, J. Moult, and K. Fidelis, "Processing and evaluation of predictions in CASP4," *Proteins*, vol. 45, no. 5, pp. 13–21, 2001.

[48] F. Teichert, J. Minning, U. Bastolla, and M. Porto, "High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABER-TOOTH," *BMC Bioinformatics*, vol. 11, article 251, 2010.

[49] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[50] A. Godzik, "The structural alignment between two proteins: is there a unique answer?" *Protein Science*, vol. 5, no. 7, pp. 1325–1338, 1996.

[51] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, Article ID btq066, pp. 889–895, 2010.

[52] M. S. Weiss and G. E. Schulz, "Structure of porin refined at 18 Å resolution," *Journal of Molecular Biology*, vol. 227, no. 2, pp. 493–509, 1992.

[53] B. Van Den Berg, P. N. Black, W. M. Clemons Jr., and T. A. Rapoport, "Crystal structure of the long-chain fatty acid transporter FadL," *Science*, vol. 304, no. 5676, pp. 1506–1509, 2004.

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

[55] P. Sampathkumar, F. Lu, X. Zhao et al., "Structure of a putative BenF-like porin from Pseudomonas fluorescens Pf-5 at 2.6 Å resolution," *Proteins*, vol. 78, no. 14, pp. 3056–3062, 2010.

[56] J. C. Whitney, I. D. Hay, C. Li et al., "Structural basis for alginate secretion across the bacterial outer membrane," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 32, pp. 13083–13088, 2011.

[57] J.-D. Pédelacq, L. Maveyraud, G. Prévost et al., "The structure of a Staphylococcus aureus leucocidin component (LukF-PV) reveals the fold of the water-soluble species of a family of transmembrane pore-forming toxins," *Structure*, vol. 7, no. 3, pp. 277–287, 1999.

[58] H. Akama, M. Kanemaki, M. Yoshimura et al., "Crystal structure of the drug discharge outer membrane protein, OprM, of Pseudomonas aeruginosa: dual modes of membrane anchoring and occluded cavity end," *Journal of Biological Chemistry*, vol. 279, no. 51, pp. 52816–52819, 2004.

[59] S. De and R. Olson, "Crystal structure of the Vibrio cholerae cytolysin heptamer reveals common features among disparate pore-forming toxins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 18, pp. 7385–7390, 2011.

[60] K. A. Bunting, S. M. Roe, and L. H. Pearl, "Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the β-clamp," *EMBO Journal*, vol. 22, no. 21, pp. 5883–5892, 2003.

[61] K.-J. Park, M. M. Gromiha, P. Horton, and M. Suwa, "Discrimination of outer membrane proteins using support vector machines," *Bioinformatics*, vol. 21, no. 23, pp. 4223–4229, 2005.

[62] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, "A Hidden Markov Model method, capable of predicting and discriminating β-barrel outer membrane proteins," *BMC Bioinformatics*, vol. 5, article 29, 2004.