TMDIM: an improved algorithm for the structure prediction of transmembrane domains of bitopic dimers

Han Cao · Marcus C. K. Ng · Siti Azma Jusoh · Hio Kuan Tai · Shirley W. I. Siu

Received: 28 April 2017 / Accepted: 17 August 2017 © Springer International Publishing AG 2017

Abstract α-Helical transmembrane proteins are the most important drug targets in rational drug development. However, solving the experimental structures of these proteins remains difficult, therefore computational methods to accurately and efficiently predict the structures are in great demand. We present an improved structure prediction method TMDIM based on Park et al. (Proteins 57:577–585, 2004) for predicting bitopic transmembrane protein dimers. Three major algorithmic improvements are introduction of the packing type classification, the multiple-condition decoy filtering, and the cluster-based candidate selection. In a test of predicting nine known bitopic dimers, approximately 78% of our predictions achieved a successful fit (RMSD <2.0 Å) and 78% of the cases are better predicted than the two other methods compared. Our method provides an alternative for modeling TM bitopic dimers of unknown structures for further computational studies. TMDIM is freely available on the web at https://cbbio.cis.umac.mo/TMDIM. Website is implemented in PHP, MySQL and Apache, with all major browsers supported.

Keywords Structure prediction · Bitopic dimer · Helix packing · Helix assembly · Scwrl4 · Bioinformatics

Introduction

Transmembrane (TM) proteins are crucial in diverse life processes such as signal transduction, membrane transport, cell recognition and energy generation. At present, they are also the most important drug targets and account for over 60% of the entire known drug targets in the drug design area [38]. Therefore, it is of great interest to determine the structures of the TM proteins and to understand their functions and mechanisms at the molecular level. Despite their biological and pharmaceutical significance, very few structures of TM proteins are determined to date. Because obtaining the experimental structures of TM proteins is technically very difficult [9], it is therefore no surprise that the number of solved TM structures in the Protein Data Bank (PDB) is dramatically underrepresented (in less than 1%).

Given this situation, computational approaches to predict the structure of TM proteins are highly demanded. Based on the widely accepted two-stage model proposed by Popot and Engelman [30], the prediction problem can be simplified into two main tasks: first, identification of the transmembrane segments and their secondary structures from the amino acid sequence; second, prediction of the optimal relative orientation of the α-helices (or β-sheets).

Based on the two-stage model, numerous TM protein structure prediction methods have been proposed to date. These methods can be classified into three categories according to [27]: sequence-based methods, conformational search-based methods and complementary of surface property-based methods.

Early efforts determined the patterns of amino acid sequences (motifs) which appear frequently in the dimerized interface, and then applied these motifs for prediction. Many motifs have been determined to date, such as the general GG4 motif (GxxxxG) [34] and the glycine zipper motif
(G,A,S)xxG and Gxx(G,S,T) [15]. Here, x refers to any kind of amino acid. These determined sequence motifs promoted the development of prediction algorithms for helix assembly which are based on statistical learning of contacting patterns in structural and sequence data.

However, as mentioned in [18], as more and more sequence motifs are explored, the incompleteness of sequence motif paradigm for predicting dimerization becomes increasingly clear. Even sequences with the recognizable motifs will not always result in helical contact in those motifs. With the recognition of the thermodynamic nature of helix dimerization, methods dependent on searching of low-energy structures on the large conformational space were developed. For these methods, the prediction problem is in fact an optimization problem. The algorithm will search for the optimal way of packing the helix pair into a dimer by exploring the conformational space in which the energy landscape is determined by a scoring function. In Kim’s method [14], the scoring function includes only the energy term for van der Waals (vdW) interactions of protein atoms. To improve the efficiency of scoring, Park et al. [26] developed the first residue-based vdW energy function as the scoring function which could produce a highly accurate prediction model of glycoporphin A. The scoring function was also applied in structure prediction of TM helix bundle proteins having modest numbers of TM helices [25]. Another recent method suggested that beside vdW interactions, optimizing the hydrogen-bonding between helix pairs could greatly improve structure predictions of GAS right dimers [23]. Beside energy-based methods, a non-energy-based scoring method was proposed by [12] that tried to maximize the number of contacts between residues that promote helix interactions and to penalize the burial of bulky amino acid residues. These residues, including ARG, HIS, LYS, MET, PHE, TRP and TYR, consist of large side chains and are assumed to accommodate conformers but not form steric hindrances. Common in the conformational search-based methods is that the lipid component was completely ignored and the structures were simulated in vacuum. Alternative methods include simplified membrane models such as [10, 36] to further increase the prediction accuracies.

Finally, the third category of methods evaluate the shape and physiochemical complementaries of the two helices in order to determine the location and the propensity of the contact interface. For example, Polyansky et al. [28] mapped the geometrical characteristics of the individual helix in a 3D structure onto a 2D energy map. Then, the most complementary regions in the two maps of the helix pair were located and their propensities in forming helix contacts were measured.

With the advance in computing power, several helix dimerization studies based on long molecular dynamics (MD) simulations using coarse-grained (CG) model [31, 32], united-atom (UA) model [37], or all-atom (AA) model [39] have been reported in recent years. From CG to AA models, the level of detail is increased but the computational cost is also multiplied. In these studies, membrane can be modeled explicitly to fully account for the environmental effect which cannot be easily considered in aforementioned methods. Besides, the helices are flexible (rather than rigid) and the entire assembly process can be investigated. When the process of helix assembly is not the focus of study, a fast prediction method can be firstly applied to obtain potential dimerized structures and then perform MD simulations to observe their stability or to optimize the structures.

In this paper, we present an accurate structure prediction method for α-helical TM bitopic dimers, named TMDIM, based on the framework of Park’s prediction method and its residue-scale scoring function [26] (for simplicity, we call it PARK in this paper). The predictive performance of TMDIM was estimated by testing against 9 helical homodimers whose NMR structures are available in the Protein Data Bank (PDB). We discuss in detail our computational results and compare them to a recently released prediction tool PREDDIMER [29].

Methods

The structure prediction algorithm

In PARK, the structure prediction problem of TM helical dimer is tackled by four main steps: dimer generation, side-chain optimization, scoring, and ranking. The first three steps are iterated systematically until the entire conformational space is explored, then the optimized structures are ranked by their energy scores where the lowest-energy structure is reported as the predicted model. The energy score is calculated as a pairwise sum of residue-residue interaction energies which are derived from atomic vdW interactions. The dimeric model of the two helices is generated by varying the five geometric parameters $(\alpha, \beta, \gamma, d, s)$: $\alpha$ and $\beta$, the rotational angle of the first and second helix around their helical axes; $\gamma$, the crossing angle between the two helical axes; $d$, the inter-helical (shortest) distance between the two helical axes; and $s$ the sliding distance along the respective helical axis from the helix’s centroid. PARK was developed for executing in single CPU and thus a complete run would take hours to days to finish depending on the length of the helices.

We modified PARK by proposing three major add-ons in the structure prediction algorithm, namely the packing type classification, the multiple-condition decoy filtering, and the cluster-based candidate selection. The range and step size of each dimerization parameter were re-defined in order to balance the computational cost and the search
resolution required to obtain structures which were less than 2 Å in RMSD to experiments. Besides, for the side-chain placement algorithm, we compared the recently published methods RASP [21] and SCRWL4 [17] to SCWRL3 which was employed for side-chain optimization in PARK. In addition, the newly revised scoring function by Park et al. [25] was adopted for scoring. In order to improve the time efficiency of a prediction run, the most computationally expensive scanning stage was ported to execute in multiple processors; each compute core was assigned to explore independently a different region of the conformational space and the result was aggregated for the final ranking and clustering. The flowchart of TMDIM is presented in Fig. 1. Parameters of TMDIM were largely based on PARK except that the inter-helical distance parameters were derived by exhaustive tests taking GpA (PDB 1afo) and ErbB3 (PDB 2l9u) as the model proteins for tight-packing case and non-tight packing case respectively. Finalized parameters were considered appropriate balancing accuracy and efficiency in predictions which are listed in Table 1.

The prediction workflow begins by giving an amino acid sequence as an input. The sequence is first examined to predict its packing type (namely tight or non-tight packing) from which the set of dimerization parameters is determined. Then, an initial dimeric structure is constructed by randomly placing two ideal right-handed α-helices together. In the scanning stage, all five dimerization parameters are systematically varied to generate decoys. Each of the decoys is structurally optimized and checked for validity. The valid decoy which passes the filtering step will be scored using the residue-based scoring function. Only decoys with energy less than 0 will be collected in the decoy set. After the scan is completed, the decoys are ranked and clustered based on their energies and structural similarities. Finally, the top k decoys identified from the k low energy clusters are reported. If there were no decoys collected in the first completed scan, the filtering conditions will be adjusted and the scan will be re-executed.

**Packing type classification**

By studying the experimental structures of TM helix pairs, we observed that when bulky residues are packed at the inter-helical packing interface, the inter-helical distance is relatively large compared to cases where no bulky residues are found at the interface. The former we is referred to as non-tight packing type and the latter tight packing type. Assuming that helix pairs are predominantly tightly packed, we try to classify the non-tightly packed cases from the tightly packed cases using a simple scheme based on the amino acid sequence:

**Fig. 1** Flowchart of the TMDIM method. The scanning stage is executed in N processors. Grey new or modified processes.
A dimeric structure is classified as non-tight packing type.

This scheme follows from a simple idea that in a helical protein four residues make up one helical turn; if there are three or more bulky residues inside one helical turn, then it is almost unavoidable to pack some of these bulky residues into the inter-helical interface and results in a looser packing.

After the packing type of the helix pair is determined, appropriate settings are applied in the subsequent scanning stage: 1) For tight-packing structures, the range of inter-helical distance $d_T$ of $[6.5, 7]$ Å is used and the bulky residue check filter is enabled. 2) For non-tight packing structures, the range of inter-helical distance $d_{NT}$ of $[8.5, 9]$ Å is used and the bulky residue check filter is disabled.

Multiple-condition decoy filtering

In the scanning stage, a decoy is checked against three conditions to decide if it is valid for further processing. The first condition is that the decoy should not contain steric clashes. Here, we used the vdW radii from the AMBER force field [8] (in Å: H: 0.6, N: 1.824, C: 1.908, O: 1.6612, S: 2.0) to calculate the minimum atomic distance $d_{ij}^{min}$ allowed between any two atoms:

$$d_{ij}^{min} = \sqrt{r_i + r_j}$$  \hspace{1cm} (1)

where $r_i$ and $r_j$ refer to the vdW radii of atoms $i$ and $j$ respectively.

The second condition is that there must exist inter-helical residue contacts between the two helices. For this, the contact definition from Fuchs et al. [13] is adopted. Namely, two residues are considered in contact if their minimal distance between side-chain or backbone heavy atoms of the two residues is <5.5 Å.

Finally, for tightly packed structures, bulky residues are less likely to be buried into the inter-helical interface due to the high cost of entropic loss [12]. Hence, the third condition which is checked when predicting tight packing cases is that the bulky residues (ARG, HIS, LYS, MET, PHE, TRP or TYR) from the two helices should not be in contact. This check is disabled when predicting non-tight packing cases. However, for tight packing cases, if no valid decoys are obtained at the end of the scanning stage, this bulky residue check will be disabled and the scan is repeated.

Cluster-based candidate selection

After the scanning stage is completed, the collected decoys in the decoy set are examined to decide for the most near-native conformation. According to Shortle et al. [35], when a reasonably good scoring function is used, a greater number of low-energy decoys will cluster near the native state than the lowest-energy non-native state. As a consequence, cluster centers are better prediction models than the lowest-energy models. Therefore, in TMDIM, we introduced a cluster-based candidate selection process to identify low-energy clusters and from them to select the models.

In our algorithm, we define a cluster to be a set of decoys and each of them is structurally similar to the center decoy with an RMSD less than a cutoff value. Our goal is to look for highly populated clusters of low energy decoys from the entire decoy set, which is of the order of $10^3$ to $10^5$. Instead of doing an exact clustering, we designed a heuristic search algorithm to locate highly populated region of decoys and a rapid scheme to identify decoys that are structurally similar to the center decoy without exhaustive RMSD computations. The search is based on a greedy approach that the direction to the most concentrated region within the current cluster is considered as the most promising direction to the populated region in the entire search space. Hence, starting from the lowest-energy decoy as the first center decoy, we compute the mean conformation of all current cluster members and select the one which has the smallest RMSD to the current center decoy to be the new center. The search is continued until there is no update of the center decoy and so the center decoy at the final iteration is considered as the candidate model.

To obtain top $k$ prediction models, the center decoy and associated cluster members in the final iteration are removed from the decoy set and the search procedure is restarted with the set of remaining decoys. The pseudocode of the cluster-based candidate selection process is presented in Algorithm 1.

| Table 1 | The dimerization parameters, ranges and step sizes |
|-------------------|-------------------|-------------------|
| Dimerization parameter | Notation | Range | Step size |
| Rotational angle | $a, \beta$ | $[0, 360]^\circ$ | $6^\circ$ |
| Crossing angle | $\gamma$ | $[-45, 45]^\circ$ | $6^\circ$ |
| Inter-helical distance for Tight packing type | $d_T$ | $[6.5, 7]$ Å | 0.5 Å |
| Non-tight packing type | $d_{NT}$ | $[8.5, 9]$ Å | 0.5 Å |
| Sliding distance | $s$ | $[-9, 9]$ Å | 1 Å |

 Springer
**Algorithm 1** The cluster-based candidate selection algorithm.

**Input:** $D$: decoy set; $k$: the number of prediction models

1. Let $c$ represents the current cluster center
2. Let $c'$ represents the new cluster center
3. $D_c ← sort D$ by ascending energy values
4. $R ← ∅$
5. for $i = 1; i ≤ k; i + +$
6. $c ← D_{[0]}$
7. while True do
8. $A ← select$ cluster members of $c$ from $D_c$
9. $\bar{A} ← calculate the mean conformation from all decoys in A$
10. $c' ← arg min_{a ∈ A} RMSD(a, \bar{A})$
11. if $c' == c$ then
12. $R[i] ← c'$
13. $D_c ← remove members of A from D_c$
14. break
15. else
16. $c ← c'$
17. end if
18. end while
19. end for

**Output:** $R$: top $k$ prediction models

Since each decoy was generated from a combination of dimerization parameter values during the scanning stage, our rapid scheme to identify cluster members is to calculate dimerization profiles of all potential decoys to the current center decoy. Four dimerization parameters $\alpha, \beta, \gamma, s$ are included in the comparison. Decoys whose parameter values are within the cutoff limits will be selected; the cutoff limits were decided empirically to be $[\alpha_c - 72^\circ, \alpha_c + 72^\circ], [\beta_c - 72^\circ, \beta_c + 72^\circ], [\gamma_c - 42^\circ, \gamma_c + 42^\circ]$ and $[s_c - 2 \AA, s_c + 2 \AA]$, where the suffix $c$ indicates the dimerization profile of the current center decoy.

### Data set

To evaluate the predictive performance of the proposed algorithm, we used a benchmark of nine $\alpha$-helical TM homodimers with known experimental structures. This set was selected from the list of membrane proteins classified under the Bitopic proteins superfamilies (Class 1.2) in the Orientations of Proteins in Membrane (OPM) database [19]. Out of 356 proteins, 26 right-handed transmembrane helical homodimers were identified. We performed all-vs-all pairwise global sequence alignments of these proteins in their transmembrane region using the FASTA program (ggsearch36). For sequences which have identity greater than 60% among themselves, only one of them will be included into the data set. Table 2 lists our benchmark data-set, PDB IDs, sequences and dimerization profiles. As we can see, the data set includes dimers with diverse conformations. Their inter-helical distances range from 7 to 10 Å, the sliding distances range from −8 Å to 9 Å, and their crossing angles span across $-50^\circ$ to $27^\circ$.

### Evaluation metrics

We will compare the predictive performance of our method TMDIM to PARK and PREDDIMER [29]. PARK is the method proposed by Park et al. [26] on which our method is based and PREDDIMER is a recently available online server by Polyansky et al. [29] for helical dimer prediction. All three methods accept an amino acid sequence as input and produce 3D structures as output. In TMDIM, the user can specify the desired number of prediction models as an input parameter and the result is a ranked list of best dimer conformations generated from the candidate selection process. Similarly, PARK returns a list of decoys ranked from the most favorable conformation to the least favorable.

### Table 2 Selected $\alpha$-helical TM homodimer experimental structures and their dimerization profiles

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB</th>
<th>Sequence</th>
<th>Residue number</th>
<th>Inter-helical distance (Å)</th>
<th>Sliding distance (Å)</th>
<th>Crossing angle (°)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNip3</td>
<td>2j5d</td>
<td>VFLPSLLLHSLAIGLGIYIG</td>
<td>164–184</td>
<td>7</td>
<td>−8</td>
<td>−39</td>
<td>[2]</td>
</tr>
<tr>
<td>GpA</td>
<td>1af0</td>
<td>ITLIIFVMAGVITLISYGI</td>
<td>73–95</td>
<td>7</td>
<td>4</td>
<td>−45</td>
<td>[20]</td>
</tr>
<tr>
<td>EphA1</td>
<td>2k1l</td>
<td>IVAIVFGLLGAALLGLYLF</td>
<td>548–568</td>
<td>7</td>
<td>5</td>
<td>−50</td>
<td>[3]</td>
</tr>
<tr>
<td>ErbB4</td>
<td>2Jt2</td>
<td>PLIAAVGGLFILVGLTFAVV</td>
<td>51–75</td>
<td>7</td>
<td>8</td>
<td>−45</td>
<td>[5]</td>
</tr>
<tr>
<td>EGFR</td>
<td>2m20</td>
<td>PSIATGLVGALLLVLVALGLGLF</td>
<td>3–27, 63–87</td>
<td>7</td>
<td>9</td>
<td>−45</td>
<td>[11]</td>
</tr>
<tr>
<td>ErbB2</td>
<td>2jwa</td>
<td>PTVISAVGILLVVLGVGFILI</td>
<td>51–75, 151–175</td>
<td>8</td>
<td>8</td>
<td>−45</td>
<td>[4]</td>
</tr>
<tr>
<td>PDGFR β</td>
<td>2i6w</td>
<td>VVVISAILVVTIIISLIIIML</td>
<td>8–31</td>
<td>8.5</td>
<td>8</td>
<td>25</td>
<td>[24]</td>
</tr>
<tr>
<td>FGFR3</td>
<td>2zl4</td>
<td>AGSVYAGILSYGVGFLFILVAAVTLC</td>
<td>369–396</td>
<td>9</td>
<td>−1</td>
<td>27</td>
<td>[6]</td>
</tr>
<tr>
<td>ErbB3</td>
<td>28u5</td>
<td>LTMALTVIAGLVVFMLGGFTLY</td>
<td>642–665</td>
<td>10</td>
<td>7</td>
<td>28</td>
<td>[22]</td>
</tr>
</tbody>
</table>

Dimerization values of experimental structures were measured using a modified code of our prediction program with a step size of 0.5 Å for inter-helical distance, 1 Å for sliding distance and 1° for crossing angle. The sampled structure with the lowest RMSD (<1 Å) is reported.
conformation according to the energy scores. For PREDDIMER, the number of model output depends on the number of matching regions analyzed by their algorithm. For example, it produced 3 models for the test case 1afo; 4 models for 2jwa, 2k1l, and 2l2t; 5 models for 2m20, 2lzl, and 2l6w; 7 models for 2l9u and 2j5d. Here, for completeness and fairness, we compared the top 3 prediction models from each of the methods.

The main performance metric we used to evaluate the accuracy of the prediction models is root-mean-squared deviation (RMSD). Note that only the backbone atoms are taken in the calculation and when comparing to the NMR structure, only the first experimental model in the PDB file (MODEL 0) is used. This is justified since the different NMR models are closely aligned at the transmembrane region where the helices dimerize. In our test, only the transmembrane segment of the helix is used as input to the prediction method (see Table 2). A prediction which has an RMSD ≤ 2 Å to experiment is considered a successful prediction. Statistical tests were performed using the non-parametric one-sample Wilcoxon signed-rank test implemented in the R statistics package[33].

The contacting interface of the two helices in the predicted dimer can be compared to that in the experimental dimer by their contact maps. A contact map is generated from a matrix of minimal distances between pairs of residues from the two helices. Short distance pairs of <8 Å are potential contacting residues and will be coded with corresponding color in the map.

Implementation

In TMDIM, the conformational scanning program was developed using MPI/C++ based on Park’s implementation in [26] utilizing the Biochemical Algorithms Library (BALL) [16]. The helix generation program and the candidate selection program were developed with Python 2.6.9 and the Protein Dynamics library (ProDy) [1]. All experiments were conducted on an Intel Xeon E5-2697V2 (3.5 GHz, 8.00GT/s) workstation with 12 CPU cores and 16 GB memory.

Results and discussion

Comparison of prediction methods

The prediction results by PARK, TMDIM and PREDDIMER in terms of backbone RMSD are shown in Table 3 and the lowest RMSD models are depicted in Fig. 2. On measuring the average predictive performance, for PARK the average RMSDs of the top 1, 2, and 3 predictions are 4.04, 3.92, 3.95 Å respectively; for TMDIM, they are 2.82, 3.49, 3.62 Å; and for PREDDIMER, they are 5.50, 3.93, 4.11 Å. In terms of success rate, PREDDIMER has 2 successful predictions out of 9 test cases, yielding a success rate of only 22%. For PARK, 4 cases are predicted successfully giving an improvement in success rate of 22% over PREDDIMER. Our algorithm TMDIM predicts 7 cases successfully that gives a success rate of 78%, a slight improvement over PARK (p value = 0.08986) but a significant improvement over PREDDIMER (p value = 0.01267). When considering only the rank-1 solution of the three methods, TMDIM predicts structures with 48.7% lower RMSD than PREDDIMER (p value = 0.01413) and 40% lower than PARK (p value = 0.04292).

Furthermore, if we consider predictions for each individual test case (the lowest RMSD prediction highlighted in Table 3 Backbone RMSD (in Å) of top 3 predictions by PARK, TMDIM, and PREDDIMER

<table>
<thead>
<tr>
<th>PDB</th>
<th>PARK</th>
<th>TMDIM</th>
<th>PREDDIMER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
</tr>
<tr>
<td>2j5d</td>
<td>6.97</td>
<td>6.95</td>
<td>6.99</td>
</tr>
<tr>
<td>1afo</td>
<td>0.69</td>
<td><strong>0.63</strong></td>
<td>0.74</td>
</tr>
<tr>
<td>2k1l</td>
<td>1.98</td>
<td>1.88</td>
<td>1.92</td>
</tr>
<tr>
<td>2l2t</td>
<td>1.51</td>
<td>1.43</td>
<td>1.48</td>
</tr>
<tr>
<td>2m20</td>
<td>6.31</td>
<td>6.32</td>
<td>6.33</td>
</tr>
<tr>
<td>2jwa</td>
<td>5.86</td>
<td>5.81</td>
<td>5.84</td>
</tr>
<tr>
<td>2l6w</td>
<td>6.65</td>
<td>6.65</td>
<td>6.67</td>
</tr>
<tr>
<td>2lzl</td>
<td>3.83</td>
<td>3.82</td>
<td>3.79</td>
</tr>
<tr>
<td>2l9u</td>
<td>2.50</td>
<td>1.77</td>
<td>1.77</td>
</tr>
<tr>
<td>Avg.</td>
<td>4.04</td>
<td>3.92</td>
<td>3.95</td>
</tr>
<tr>
<td>S.D.</td>
<td>2.45</td>
<td>2.54</td>
<td>2.53</td>
</tr>
<tr>
<td>Succ.</td>
<td>4 (44%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The best predictions are highlighted in bold

*aSuccessful cases are predictions with RMSD < 2.0 Å
bold in Table 3) among all methods, both PARK and PRED-DIMER have 1 best prediction whereas TMDIM has 7 best predictions, i.e. 78% of the test cases are better predicted by TMDIM than by the other two methods in this experiment.

How similar is the way that two helices dimerize in the predicted model to the experiment can be compared using their dimerization profiles, i.e. the inter-helical distance, the sliding distance, and the crossing angle. In Table 4, the dimerization profiles of the lowest RMSD models by TMDIM are presented and the differences (the residues) from experiments are indicated by values in brackets. In general, a successful prediction has a dimerization profile close to its experimental profile and the residue of a dimerization parameter is within 1–2 step size of the conformational scan.

The worst predicted cases in TMDIM are 2l6w and 2lzl. As shown in Fig. 2g, h, the orientation of the second helices (at the right-hand side) are completely different. For 2l6w, the predicted model has a large negative crossing angle of $-45^\circ$ while the experimental structure has a positive crossing angle of $25^\circ$. The bad prediction is presumably due to the misclassification of the packing type at the beginning of the prediction algorithm. Experimentally, this dimer has a helical distance of about 8.5 Å—a relatively loose packing structure, but because there is no bulky residue within a sliding window of four residues in the sequence, it was misclassified as the tight packing type. As for the predicted 2lzl, although the two helices are packed at the right inter-helical distance and at the right crossing angle, the sliding distance is shifted by 7 Å. All test cases were completed with one pass of the search loop (see Fig. 1) except 2j5d. The reason was that although this was a tight-packing dimer, a HIS was located at the packing interface which violated the bulky residue rule and ended up with no decoy found. In the second round, once the bulky residue check filter was switched off, a low energy RMSD structure was obtained.

The dimerization profile described above has given us only a partial picture of the helix dimerization because each helix can be rotated about its own helical axis independently.

![Fig. 2 Superposition of the lowest RMSD prediction (green) on the experimental structure (red) in nine helical dimers. For each case, the first helices at the left-hand side are fitted and so the contacting regions of the helix pair can be visually compared](image)

To give a full account of the helix dimerization, for each helical dimer, we calculated the minimal distances (heavy atoms only) of all residue-pairs of the adjacent helices and plotted them in a contour map in Fig. 3. In all test cases, the helix interactions are symmetrical which is indicated by the deepened color along the diagonal of the contour map. This is expected for a homodimer. All successful predictions (with $<$2 Å RMSD) can produce very similar residue-residue

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Packing type</th>
<th>Inter-helical distance (Å)</th>
<th>Sliding distance (Å)</th>
<th>Crossing angle (°)</th>
<th>#</th>
<th>RMSD (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2j5d</td>
<td>Tight</td>
<td>7</td>
<td>$-7 (-1)$</td>
<td>$-39$</td>
<td>1</td>
<td>1.01</td>
</tr>
<tr>
<td>1afo</td>
<td>Tight</td>
<td>7</td>
<td>$7 (+1)$</td>
<td>$-39 (-6)$</td>
<td>2</td>
<td>0.79</td>
</tr>
<tr>
<td>2k11</td>
<td>Tight</td>
<td>7</td>
<td>5</td>
<td>$-45 (-5)$</td>
<td>2</td>
<td>1.39</td>
</tr>
<tr>
<td>2l2t</td>
<td>Tight</td>
<td>7</td>
<td>8</td>
<td>$-45$</td>
<td>1</td>
<td>0.77</td>
</tr>
<tr>
<td>2m20</td>
<td>Tight</td>
<td>7</td>
<td>$8 (+1)$</td>
<td>$-45$</td>
<td>3</td>
<td>1.24</td>
</tr>
<tr>
<td>2jwa</td>
<td>Tight</td>
<td>7 $(+1)$</td>
<td>8</td>
<td>$-45 (+70)$</td>
<td>3</td>
<td>0.78</td>
</tr>
<tr>
<td>2l6w</td>
<td>Tight</td>
<td>7 $(+1.5)$</td>
<td>8</td>
<td>$-45 (+70)$</td>
<td>2</td>
<td>5.51</td>
</tr>
<tr>
<td>2lzl</td>
<td>Non-tight</td>
<td>9 $(-7)$</td>
<td>6 $(-7)$</td>
<td>27</td>
<td>1</td>
<td>3.81</td>
</tr>
<tr>
<td>2l9u</td>
<td>Non-tight</td>
<td>9 $(+1)$</td>
<td>6 $(+1)$</td>
<td>$33 (-5)$</td>
<td>3</td>
<td>1.42</td>
</tr>
</tbody>
</table>

Table 4 Dimerization profiles of the lowest RMSD predictions by TMDIM

Residues of the predicted dimerization values from experiments are shown in brackets.
contacts. In a few cases, for example in 2l9u, an contact at the C-terminal end of the helices is almost missing in the predicted model but instead enhanced interactions can be seen towards the N-terminal of the helices. The contact map also shows how different the predictions are from the experiments in the two badly predicted cases, 2l6w and 2lzl. In 2l6w, the predicted contacts are concentrated at the upper half of the helices while experiment shows that the two helices are aligned parallel to one another and interact all throughout the helical axis. Similarly, 2lzl has close contacts at the dimeric center as well as some longer contacts extended along the two ends of the helices; however, the predicted model has overall reduced interactions in both strength and range.

In terms of data volume in TMDIM, thanks to the multiple-condition decoy filtering, the amount of generated decoys in TMDIM is greatly reduced by 74–99.9% compared to PARK (see Table 5). Larger reductions are seen in tight packing cases as modeling residue side-chains within a limited volume will have a higher chance to steric clashes.

The use of different sidechain placement methods for structure prediction

In TMDIM, protein flexibility is incorporated by optimizing the side-chain conformation using a third-party side-chain placement program. This is a crucial step because obtaining an accurate structural model is the foundation to good energy prediction. Here, to see which side-chain placement program works best with TMDIM, we compared the predictive performance of TMDIM using SCWRL3 [7], SCWRL4 [17], and the recently published method RASP [21].

As shown in Fig. 4, models using SCWRL4, a new version of SCWRL3, always give the best or the second best prediction in all test cases, in particular in the prediction of 2jwa and 2l9u, the accuracies of the predicted models in

<table>
<thead>
<tr>
<th>PDB</th>
<th>(N_{gen})</th>
<th>(N_{remain})</th>
<th>Reduction percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2j5d</td>
<td>2188800</td>
<td>18554</td>
<td>99.2</td>
</tr>
<tr>
<td>1afo</td>
<td>2188800</td>
<td>793</td>
<td>99.9</td>
</tr>
<tr>
<td>2k11</td>
<td>2188800</td>
<td>179</td>
<td>99.9</td>
</tr>
<tr>
<td>2l2t</td>
<td>2188800</td>
<td>1789</td>
<td>99.9</td>
</tr>
<tr>
<td>2m20</td>
<td>2188800</td>
<td>8709</td>
<td>99.6</td>
</tr>
<tr>
<td>2jwa</td>
<td>2188800</td>
<td>4559</td>
<td>99.8</td>
</tr>
<tr>
<td>2l6w</td>
<td>2188800</td>
<td>2448</td>
<td>99.9</td>
</tr>
<tr>
<td>2lzl</td>
<td>2188800</td>
<td>486648</td>
<td>77.8</td>
</tr>
<tr>
<td>2l9u</td>
<td>2188800</td>
<td>560517</td>
<td>74.4</td>
</tr>
</tbody>
</table>

The column 2 and 3 show the number of conformations generated \((N_{gen})\) and the number of conformations remain \((N_{remain})\) after the filtering conditions were applied in predictions using SCWRL3 side-chain placement methods.
terms of backbone and side-chain RMSDs are improved by about 50% as compared to SCWRL3. While RASP’s predictions are close to SCWRL4, there are two cases where RASP are far off, i.e. 2j5d and 2m20. Nevertheless, the improved accuracy of SCWRL4 from SCWRL3 comes with a cost. As shown in Fig. 5, the total elapsed time for completing the conformational scan by SCWRL4 is increased by 70% on average.

Availability

A web-based server has been developed to provide easy-to-access of the TMDIM prediction method. The user only needs to give the helix sequences as input (currently only dimer prediction is supported). Other options such as parameters for conformational scan and side-chain placement algorithm are allowed to be adjusted. Once the submit button is pressed, the user request is entered into our queuing system and a job ID is returned to the user as a token to check back the result page. A screen snapshot of the web TMDIM server is shown in Fig. 6. The web system was implemented using PHP 5.5.24 and MySQL 5.6.24 and is available at https://cbbio.cis.umac.mo/TMDIM.

Conclusion

In this paper, we presented an accurate algorithm TMDIM to solve the structure prediction problem for bitopic transmembrane protein dimers. Based on the PARK method [26], we proposed three important modifications on the algorithm, namely packing type classification, multiple-condition decoy filtering, and cluster-based candidate selection, which leads to a significant improvement in the prediction accuracy in terms of RMSD and residue-residue contact. Furthermore, the program is implemented using the message passing interface (MPI) to parallelize the conformational scanning and results in a great speed up of the prediction task.

One important limitation of the current method is that the energy function which dictates the packing results assuming only vdW interactions between helices. For cases where electrostatics interactions (such as formation of a salt bridge) are important, the prediction would likely fail. In future, we consider to reparameterize the scoring function to include contributions from different interaction types (electrostatics, hydrogen bonding, etc.). While homodimers are the focus of the current method, we plan to extend its application also on heterodimers by refining the search algorithm to evaluate heterodimerization of helices. Regarding the computing efficiency of TMDIM, a significant part of the time has been spent in calling external programs SCWRL3 and SCWRL4 to perform sidechain prediction. To enhance the speed of TMDIM, we are developing an improved sidechain prediction method with knowledge-based scoring function to allow better sidechain prediction of transmembrane helices. A faster prediction method is desirable as our ultimate goal is to predict oligomerization of more transmembrane helices.
Acknowledgements  We thank Yungki Park and Volkhard Helms for granting the use of their program for the development of TMDIM. This work was performed in part at the High Performance Computing Cluster which is supported by Information and Communication Technology Office (ICTO) of the University of Macau.

Funding  The funding was provided by University of Macau (Grant Number MYRG2014-00104-FST).

References


Fig. 6 Screen snapshots of the TMDIM web system: (left) the input page and (right) the result page

© Springer