High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling

Jing Yang1,2, Richard Jang3, Yang Zhang3,4,* and Hong-Bin Shen1,2,3,*

1Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 2Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China, 3Department of Computational Medicine and Bioinformatics and 4Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

ABSTRACT

Motivation: Residue-residue contacts across the transmembrane helices dictate the three-dimensional topology of alpha-helical membrane proteins. However, contact determination through experiments is difficult because most transmembrane proteins are hard to crystallize.

Results: We present a novel method (MemBrain) to derive transmembrane inter-helix contacts from amino acid sequences by combining correlated mutations and multiple machine learning classifiers. Tested on 60 non-redundant polytopic proteins using a strict leave-one-out cross-validation protocol, MemBrain achieves an average accuracy of 62%, which is 12.5% higher than the current best method from the literature. When applied to 13 recently solved G protein-coupled receptors, the MemBrain contact predictions helped increase the TM-score of the I-TASSER models by 37% in the transmembrane region. The number of foldable cases (TM-score >0.5) increased by 100%, where all G protein-coupled receptor templates and homologous templates with sequence identity >30% were excluded. These results demonstrate significant progress in contact prediction and a potential for contact-driven structure modeling of transmembrane proteins.

Availability: www.csbio.sjtu.edu.cn/bioinf/MemBrain/

Contact: hbshen@sjtu.edu.cn or zhng@umich.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 10, 2013; revised on July 21, 2013; accepted on July 27, 2013

1 INTRODUCTION

Membrane proteins constitute ~30% of the proteins in both prokaryotic and eukaryotic genomes (Adamian and Liang, 2006), and they participate in various crucial cellular processes, from basic small-molecule transport to complicated signaling pathways (Elofsson and von Heijne, 2007). It has been shown that ~50% of current drug targets are membrane proteins (Hopkins and Groom, 2002), where the top four gene families of Food and Drug Administration-approved drugs are all membrane proteins; i.e. G protein-coupled receptor (GPCR), nuclear receptors, ligand-gated ion channels and voltage-gated ion channels (Overington et al., 2006). Membrane-embedded alpha-helical polytopic proteins constitute the majority of ion channels, transporters and receptors in living organisms. This class of proteins, which accounts for ~40% of all membrane proteins, is famously difficult for high-resolution structural studies. Due to the intrinsic structural plasticity associated with many of these proteins, the chance of obtaining crystals suitable for X-ray or electron diffraction studies is small (Yarov-Yarovoy et al., 2006).

Although helical membrane proteins have many experimental difficulties, their structural conformation has been demonstrated to be predictable in a number of ways, e.g. transmembrane helix (TMH) domain topology ( Krogh et al., 2001; Shen and Chou, 2008).

Structure prediction of the TMH bundle for alpha-helical membrane proteins can be fulfilled by first predicting TMHs and then applying helix-packing constraints (White, 2003). The first step of the TMH prediction has a long history, and the current methods can achieve an accuracy of ~90% (Nugent and Jones, 2009; Shen and Chou, 2008). In the second step, it has been widely acknowledged that residue-residue contact maps contain crucial constraints for ab initio assembly of protein structures. In the past 5 years, several methods have been proposed to predict inter-TMH residue contacts and TMH–TMH interactions from the primary sequence. These approaches generally can be classified into two categories: (i) correlated mutation analysis (CMA)-based approaches [e.g. HelixCorr (Fuchs et al., 2007)], and (ii) machine learning (ML)-based methods [e.g. TMHcon (Fuchs et al., 2009), TMHit ( Lo et al., 2009), MEMPACK (Nugent and Jones, 2010) and TMHhcp (Wang et al., 2011)]. CMA-based approaches identify co-evolving residue pairs that tend to be in contact from multiple sequence alignments (MSAs). ML-based methods predict inter-TMH residue contacts by training statistical models using various sequence-derived features. The predicted contacts are used to predict TMH–TMH interactions. Despite the progress, there is much space for improvement.

First, in the case of CMA-based approaches, the performance depends highly on the number of aligned sequences in MSAs. For example, the recent CMA-based residue contact prediction algorithm PSICOV ( Jones et al., 2012) performs poorly on the CASP hard targets, which tend to have few homologous sequences (Di Lena et al., 2012). This indicates that contact prediction based on CMA alone is insufficient. Second, from ML point of view, inter-TMH residue contact prediction is an imbalanced learning problem, where the number of samples in different classes (contact versus non-contact) differs significantly. Existing ML-based methods use random under-sampling
(He and Garcia, 2009) to select the negative samples. Except for random forest in TMhhcp (Wang et al., 2011), which used 100 decision trees, all the existing predictors were trained with a single model using a 1:1 or 1:4 ratio of contact and non-contact samples. There is no single best ML algorithm because each has a different mathematical hypothesis. Therefore, an ensemble classifier constructed by applying multiple algorithms on multiple training sets can combine the diversities among different predictors and yield better results (Shen and Chou, 2006). Third, in all the existing ML-based models, feature vectors of two residues are serially concatenated together. Although this serial combination is widely used, the doubled feature vector dimensions may decrease performance due to over-fitting effects of high-dimensional data. The parallel strategy is another feature fusion approach (Yang et al., 2003), which represents the two vectors as real and imaginary parts in a complex space. The benefit is that the dimension of the obtained complex vector is the same as the vector of each residue, which we show improves performance.

In this study, we present a novel method called MemBrain to predict inter-TMH residue contacts by merging an ML-based engine with the CMA-based approach. Here, we use the PSICOV (Jones et al., 2012) algorithm to calculate correlated mutation scores (CMs). The ML-based engine in the proposed protocol is implemented with an ensemble classifier. It is the fusion of five OET-KNN (Zouhal and Denoeux, 1998) classifiers and five SVM classifiers, where each independent classifier is trained with a different training set obtained from five independent under-samplings. By doing so, we can get the diversities from different algorithms and, at the same time, reduce the information loss via multiple samplings.

2 MATERIALS AND METHODS

2.1 Data sets

For fair comparison, we used the same two benchmark data sets from previous studies. The training data set was taken from TMHcon (Fuchs et al., 2009) consisting of 62 alpha-helical transmembrane (TMH) proteins. Here, we only used 60 of these proteins and discarded 117vA and 1vf5B because of either non-standard atomic coordinate records or too few positive contacts, as defined in the next section. All the 60 proteins have at least three TMHs with the pairwise sequence identity <40% among the training data set. TMH locations and their topologies were extracted from the databases of TOPDB (Tusnady et al., 2008), PDBTM (Tusnady et al., 2005) and OPM (Lomize et al., 2006). For extra validation, an independent data set was taken from TMhhcp (Wang et al., 2011), which contains 21 TMH proteins with the pairwise sequence identity <40% in itself and to the training data set.

2.2 Contact definition and evaluation criteria

We adopted the contact definition from TMHcon (Fuchs et al., 2009), MEMPACK (Nugent and Jones, 2010) and TMhhcp (Wang et al., 2011) for the convenience of direct comparisons with these methods. Briefly, two residues from different TMHs are considered to be in contact if the minimal distance of their side chain or backbone heavy atoms is <5.5 Å.

In this work, we evaluated our method using a leave-one-out jackknife cross-validation, which is the same as previous studies. It takes one protein sequence out for testing, while keeping the remaining protein sequences for training. This procedure will be terminated when all the proteins have been tested individually. The overall prediction performance was evaluated by averaging the performances on individual proteins. We also assessed our method using a 4-fold cross-validation, where the four equal-size subsets have roughly the same number distribution of TMHs. Additionally, we used an independent data set for further validation. In this case, the final model was trained based on the whole training data set.

For inter-TMH residue contact prediction, the top $L$-predicted contacts were selected for assessing the prediction performance and then used to determine TMH–TMH interactions, which is the same as previous reports (Fuchs et al., 2009). Here, $L$ is the length of the concatenated TMHs. Concretely, three measures are used to evaluate the performance, i.e., Accuracy, Coverage and Accuracy ($\beta = 4$). Accuracy is defined as the fraction of correctly predicted contacts with respect to all the predicted contacts. Coverage is defined as the percentage of correctly predicted contacts from the total observed contacts. Accuracy ($\beta = 4$) is derived from 6-analysis (Ortiz et al., 1999), which calculates the fraction of predicted contacts lying within one helix turn (four residues each side) around the observed contacts.

For TMH–TMH interaction prediction, interacting helix pairs were derived from the top $L$-predicted contacts. As defined in previous studies, a helix pair is considered to be interacting if it contains at least one contact. Applying this definition to the training/independent data set, we obtained 681/334 interacting helix pairs and 757/347 non-interacting helix pairs, respectively, where the ratios between interacting and non-interacting samples are similar to previous reports (Nugent and Jones, 2010). Based on the observed interactions, the performance of predicting TMH–TMH interactions can then be assessed. Four measures are used to evaluate the performance, i.e., Accuracy, Sensitivity, Specificity and MCC. Here, Accuracy is defined as the fraction of correctly predicted interactions with respect to all the predicted interactions.

2.3 Feature extraction

We extracted four types of sequence-based features, including position-specific scoring matrix (PSSM), residue position, relative position within TMH and sequence separation between two residues, for model training. TMH locations and their topologies were extracted from annotated databases. Details for each type of input features are described in the following text:

2.3.1 Position-specific scoring matrix

The PSSM of each protein was generated by using PSI-BLAST (Altschul et al., 1997) program to search against the UniRef90 database with three iterations and an E-value threshold of 0.001. PSSM can be represented by an $L$ by 20 matrix, where $L$ is the protein length. To consider neighboring residues, a sliding window centered on a target residue, with four residues on each side of that residue, was used to extract a 180-dimensional feature vector. The original score in each position was normalized by the following logistic function of Equation (1) (Fuchs et al., 2009):

$$f(x) = \frac{1}{1 + e^{-x}}$$

where $x$ is the original score.

2.3.2 Residue position

A 10-dimensional binary vector is used to encode the residue position along the TMH to reflect the contact site. The first seven vector elements correspond to seven equally split parts along the TMH. These elements were set to $0$, other than the element at index calculated by Equation (2) defined as follows:

$$index = \left \lfloor \frac{L - \beta}{N} \right \rfloor + 1$$

where $N$ denotes the length of the TMH containing the target residue, and $\beta$ denotes the position of the target residue between 1 and $N$ on the TMH. Here, the function $f$ returns the floor value. The vector element at index was set to 1 if the target residue is in the index-th part of the helix. The last three vector elements indicate which side of the TMH the residue lies on. We divided each TMH into three parts to encode whether the
2.5 Prediction model

The developed MemBrain predictor is composed of two engines: CMA-based and ML-based prediction modules. The flowchart of MemBrain is shown in Fig. 1. The merit of CMA-based engine is that it has a clear biophysical interpretation, but its performance is highly dependent on the number of homologous sequences. The ML-based engine is not easily interpretable, but it handles the small-sample problem better. These two engines complement each other in inter-TMH residue contact prediction to drive MemBrain as an accurate prediction model.

2.5.1 CMA-based prediction engine

Approaches that use CMs to detect co-evolving residue pairs work commonly through calculating the Pearson correlation coefficient (Gobel et al., 1994; Pollock and Taylor, 1997) or mutual information (Burger and van Nimwegen, 2010; Dunn et al., 2008) between columns in MSAs. PSICOV (Jones et al., 2012), a recent approach that is based on mutual information, improved significantly the accuracy of contact prediction by using a sparse inverse covariance estimation technique (Meinshausen and Buhlmann, 2006) to deal with indirect coupling effects.

We used PSICOV to calculate CMs, where the residue pairs that lie on the same TMH were removed. The initial MSAs were collected by using PSI-BLAST to search against the UniRef90 database with three iterations and an E-value threshold of 0.001. In the final alignments, all columns not belonging to any TMH regions were deleted and duplicate


\[ s(x) = \frac{x - \min}{\max - \min} \]  

where \( x \) is the raw score and \( \min \) and \( \max \) are minimal and maximal scores for a given protein.

2.5.2 ML-based prediction engine Applying the contact definition described earlier, we obtained 11,526 contact residue pairs (positive samples) and 532,531 non-contact residue pairs (negative samples). The ratio between the total contact pairs and all the non-contact pairs is \( \sim 1:46 \), resulting in the problem of imbalanced learning. In principle, we can randomly sample a subset from the negative class and incorporate it with all positive samples to form a relative balanced training set. However, the information loss induced by random under-sampling will weaken the prediction performance. In light of this problem, we applied ensemble learning to reduce the impact of under-sampling.

We randomly sampled five subsets from the negative class independently and incorporated each subset with all positive samples to generate five different training sets. These five training sets were used to train five separate OET-KNN and SVM classifiers. The ratio between the positives and the sampled negatives is important because it will serve as a base for the statistical algorithm to learn the distributions of different classes. According to our experiments, the ratio 1:4 between positive and negative samples for each protein gives better performance and hence is used in MemBrain. OET-KNN and SVM classifiers were trained with parallel fused feature sets and serial combined feature sets, respectively, because our experimental results show that this is a better choice, and meanwhile, strong diversities can be generated among the component classifiers, which are critical for ensemble learning performance (Zhou and Yu, 2005).

We have trained multiple classifiers consisting of five OET-KNN classifiers and five SVM classifiers, which were fused together with a linear combination scheme based on posteriori probabilities. To give a general definition, suppose that there exist \( M_1 \) OET-KNN classifiers, \( M_2 \) SVM classifiers and \( Q \) pattern classes, we then define the linear combination scheme as follows:

\[ P_k = \alpha \cdot \sum_{i=1}^{M_1} w_i \cdot P(k|O_i) + (1 - \alpha) \cdot \sum_{j=1}^{M_2} w_j \cdot P(k|S_j), \quad (k = 1, 2, \ldots, Q) \]  

where \( P(k|O_i) \) and \( P(k|S_j) \) denote the probability of class \( k \) by the \( i \)-th OET-KNN classifier and the \( j \)-th SVM classifier, respectively; \( P_c \) denotes the fused probability; \( w_i \) and \( w_j \) denote the weights of the \( i \)-th OET-KNN classifier and the \( j \)-th SVM classifier, respectively, which are set to the average value for all members of the same classifier type, i.e., \( w_i = 1/M_1 \) and \( w_j = 1/M_2 \). The weight \( \alpha \) is selected by searching the value from 0 to 1 with a step of 0.01 via a jackknife cross-validation. Obviously here, \( M_1 = M_2 = 5 \) and \( Q = 2 \). Note that the raw scores generated by each classifier for a given protein were scaled to the range \([0, 1]\) with Equation (10) before combination. This ensemble ML predictor is denoted as OSC (OET-KNN and SVM Classifier).

2.5.3 Final prediction model The combination of CMA-based engine PSICOV and ML-based ensemble engine OSC forms the final prediction model, which is implemented in MemBrain. The outputs of OSC and PSICOV are merged by using a linear combination defined as follows:

\[ P = \beta \cdot P_{OSC} + (1 - \beta) \cdot P_{PSICOV} \]  

where \( P_{OSC} \) and \( P_{PSICOV} \) are the predicted contact probabilities generated by OSC and PSICOV, respectively, and \( P \) denotes the final contact propensity. The weight \( \beta \) is selected using the same search strategy as the weight \( \alpha \).

3 RESULTS

3.1 Benchmark test of CMA-based approach We applied the CMA-based algorithm PSICOV (Jones et al., 2012) for inter-TMH residue contact prediction. It was performed on the concatenated sequence, which consists of the TMH regions. Residue pairs that lie on the same TMH were removed from the predictions. All the possible residue pairs were ranked according to the generated CMs. The top ranked L/5 residue pairs were selected as the predicted contacts, where \( L \) is the total number of residues in the TMHs.

PSICOV calculates CMs from MSAs by PSI-BLAST search through the UniRef90 database. The quality of MSAs is critically important for the final prediction accuracy. Supplementary Figure S1 illustrates the average performance of PSICOV on 60 TMH proteins in the training data set. As can be seen, the prediction performance is highly dependent on sufficient homology. When the number of homologous sequences increases from 250 to 5000, the prediction accuracy improves as well from 21.6 to 42.1%. Although the performance can be improved by including more homologous sequences, there is a limit. We have tried to increase this parameter to greater than 5000, but found that the prediction performance did not change much. For a balance, we set the parameter of -b to 5000 for PSI-BLAST in the following experiments. With this parameter, the prediction accuracy is 42.1% on the training data set.

To examine the effects of MSAs’ size on accuracy, we divided the training data set into five subsets based on the number of sequences in MSAs. The performance of residue contact prediction on these five subsets is shown in Supplementary Table S1. As expected, better performance was achieved for proteins with a larger set of MSAs (52.8% in Group 5 with more than 5000 aligned sequences), whereas poorer performance for those with smaller MSAs (12.8% in Group 1 with less than 250 aligned sequences). In the case of difficult targets that have few homologous sequences, the prediction of residue contacts is still a challenge to PSICOV due to the incorrect co-evolution values that are calculated from small MSAs. For instance, the average accuracy on the five proteins in Group 1 with no more than 250 homologous sequences found is only 12.8%, with a coverage rate as low as 3.4%. These results demonstrate the weakness of the CMA-based approach.

3.2 Benchmark test of ML-based engine We have trained five OET-KNN classifiers (denoted as OET1, OET2, OET3, OET4 and OET5) (see Supplementary Fig. S2) and five SVM classifiers (denoted as SVM1, SVM2, SVM3, SVM4 and SVM5), where each of them was trained with different training sets. As shown in Supplementary Table S2, the prediction accuracy of exact contacts by individual OET-KNN classifiers varies from 45.5 to 47.8% and that with a residue variation within one helix turn (\( \delta = 4 \)) ranges from 78.0 to 78.5%. For individual SVM classifiers, the prediction accuracy varies from 46.9 to 49.0% and that with \( \delta = 4 \) ranges from 83.1 to 85.2%. These results indicate that the prediction performance is unstable with respect to classification algorithm and the sampling. When we combined individual OET-KNN classifiers with equal weights, denoted as OETs, the prediction accuracy
increased to 48.2%. Similarly, we combined individual SVM classifiers, denoted as SVMs, the prediction accuracy improved to 50.7%. As shown in Supplementary Figure S3, the areas under the curve of the combined classifiers OETs and SVMs are 0.805 and 0.841, which are superior to those of the individual OET-KNN and SVM classifiers, respectively.

Both of the improvements by OETs and SVMs demonstrate that fusing multiple classifiers is an effective way to reduce the information loss in the under-sampling process. Thus, we constructed an ensemble predictor called OSC by combining OETs and SVMs to make full use of diversities from multiple training subsets and classification algorithms according to Equation (11). Supplementary Figure S4A illustrates the results.

The combination of OETs and SVMs indeed performs better on small weights as expected because SVMs performs better than OETs. At first as the weight \( \alpha \) increases, the prediction performance improves, and then it degrades to the performance of OETs. When \( \alpha \) increases to the value 0.24, the highest prediction accuracy of 52.8% is obtained, which is then adopted in the ensemble classifier. As shown in Supplementary Table S2, OSC performs better than both OETs and SVMs on all the three measures. The prediction accuracy of OSC is 4.6% higher than OETs and 2.1% higher than SVMs. The area under the curve of OSC is 0.846, which is higher than 0.805 of OETs and 0.841 of SVMs (Supplementary Fig. S3). The good performance of OSC is due to the complementation of individual classifiers.

### 3.3 Merging CMA-based approach with ML-based engine

As the predicted CMs indicate the potential of residue pairs to form contacts, this information can be used not only as features but also as decisions. In Supplementary Table S3, we show that decision-level fusion (regarding CMs as independent prediction) discussed in this article outperforms feature-level fusion (regarding CMs as an additional feature fed into the ML model) applied in existing methods in both cases of OET-KNN and SVM. For instance, in the case of OET-KNN, feature-level fusion does not help to improve the performance, whereas decision level fusion improves the accuracy from 47.8 to 57.0%. Thus, we constructed a consensus predictor to further improve the prediction performance by merging the outputs of ensemble classifier OSC and PSICOV (Jones et al., 2012), which is implemented as MemBrain according to Equation (12). As shown in Supplementary Figure S4B, the highest prediction accuracy was obtained when the weight \( \beta \) is 0.56. The prediction accuracy of the consensus predictor is 62.0%, as listed in Supplementary Table S2.

The combination of OSC and PSICOV significantly improves the prediction performance in terms of all the three measures. The complementation contributes 9.2% prediction accuracy and 5.1% accuracy (\( \delta = 4 \)) to OSC. The consensus predictor achieves 19.9% higher prediction accuracy than PSICOV alone. We also analyzed statistical significance of the differences of the three criteria between MemBrain and OSC/PSICOV using a paired t-test. If the resulting \( P \)-value is below a level (e.g. 0.05), the performance difference between two methods is considered to be statistically significant. The resulting \( p \)-values in terms of accuracy, coverage and accuracy (\( \delta = 4 \)) are 1.2e-4/4.2e-7, 3.4e-3/2.6e-6 and 6.5e-3/1.7e-7, respectively. These results indicate that MemBrain is statistically better than the other two independent engines.

### 3.4 Performance comparison of residue contact prediction with existing methods

In this section, we compared our predictor MemBrain with existing methods on both training and independent data sets, including three inter-TMH residue contact predictors [TMHcon (Fuchs et al., 2009), MEMPACK (Nugent and Jones, 2010) and TMhhcp (Wang et al., 2011)] and two globular protein residue contact predictors [SVMcon (Cheng and Baldi, 2007) and SVMSEQ (Wu and Zhang, 2008)]. On the training data set, the compared predictors for TMH proteins are evaluated on the same data set with jackknife cross-validation. On the independent data set, experimental results of TMHcon and MEMPACK were taken from TMhhcp. PSICOV (Jones et al., 2012) is also included for comparison. TMhhcp reported two types of results based on all the features and the selected features (denoted as TMhhcp1 and TMhhcp2), we list both of them for comparison. For globular protein contact predictors, we only ranked the residue pairs that lie on different TMHs.

As shown in Table 1, MemBrain performs substantially better than other methods, with a prediction accuracy of 62.0%/64.1% on the training/independent data set. Note that MEMPACK selected the predicted contacts based on scores generated by the SVM predictor rather than the top \( L/5 \) predictions, and the coverage is the fraction of the total correctly predicted contacts from all struts and classification algorithms according to Equation (11). Supplementary Figure S4B, the highest prediction accuracy was obtained when the weight \( \beta \) increases to 0.24, the highest prediction accuracy of 52.8% is obtained, which is then adopted in the ensemble classifier. As shown in Supplementary Table S2, OSC performs better than both OETs and SVMs on all the three measures. The prediction accuracy of OSC is 4.6% higher than OETs and 2.1% higher than SVMs. The area under the curve of OSC is 0.846, which is higher than 0.805 of OETs and 0.841 of SVMs (Supplementary Fig. S3). The good performance of OSC is due to the complementation of individual classifiers.

Supplementary Table S1 also shows the performance of MemBrain on five groups of proteins according to MSAs’ size. Comparing with PSICOV, we find that accuracies in all the five groups are improved (see Fig. 2). The largest improvement occurs in Group 2, where the accuracy is increased from 28.0 (PSICOV) to 73.5% (MemBrain). For the very difficult targets in Group 1, the accuracy is increased from 12.8 to 23.6%. These results show that the combination of OSC and PSICOV is helpful for fusing the merits of both ML-based and CMA-based methods to enhance the prediction accuracy. Supplementary Figure S5 elucidates how OSC and PSICOV complement each other by taking two TMH proteins 1bccC and 2nr9A as instances.
Comparison on the training data set

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Coverage (%)</th>
<th>Accuracy (δ = 4) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMHcon</td>
<td>25.9</td>
<td>3.5</td>
<td>78.5</td>
</tr>
<tr>
<td>TMhhcp1</td>
<td>49.5</td>
<td>8.2</td>
<td>83.9</td>
</tr>
<tr>
<td>TMhhcp2</td>
<td>45.8</td>
<td>7.4</td>
<td>83.8</td>
</tr>
<tr>
<td>SVMcon</td>
<td>8.4</td>
<td>1.5</td>
<td>55.1</td>
</tr>
<tr>
<td>SVMSEQ</td>
<td>13.0</td>
<td>2.8</td>
<td>60.9</td>
</tr>
<tr>
<td>PSICOV</td>
<td>42.1</td>
<td>6.7</td>
<td>74.7</td>
</tr>
<tr>
<td>MemBraina</td>
<td>61.4</td>
<td>10.1</td>
<td>89.1</td>
</tr>
<tr>
<td>MemBrainb</td>
<td>62.0</td>
<td>10.2</td>
<td>90.4</td>
</tr>
</tbody>
</table>

Comparison on the independent data set

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Coverage (%)</th>
<th>Accuracy (δ = 4) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMHcon</td>
<td>23.6</td>
<td>3.0</td>
<td>83.4</td>
</tr>
<tr>
<td>MEMPACK</td>
<td>36.2</td>
<td>10.4</td>
<td>63.0</td>
</tr>
<tr>
<td>TMhhcp1</td>
<td>48.1</td>
<td>6.1</td>
<td>84.4</td>
</tr>
<tr>
<td>TMhhcp2</td>
<td>48.6</td>
<td>6.1</td>
<td>81.8</td>
</tr>
<tr>
<td>SVMcon</td>
<td>10.4</td>
<td>1.6</td>
<td>68.0</td>
</tr>
<tr>
<td>SVMSEQ</td>
<td>17.7</td>
<td>2.4</td>
<td>66.0</td>
</tr>
<tr>
<td>PSICOV</td>
<td>50.9</td>
<td>6.6</td>
<td>81.1</td>
</tr>
<tr>
<td>MemBrain</td>
<td>64.1</td>
<td>8.3</td>
<td>89.6</td>
</tr>
</tbody>
</table>

Table 1. Performance comparison of residue contact prediction

Note: aResults obtained from 4-fold cross-validation.

bResults obtained from jackknife cross-validation.

3.5 Performance comparison of TMH–TMH interaction prediction with existing methods

One important use of the predicted contacts is to identify interacting TMHs. Using the observed interactions described previously, the performance of TMH–TMH interaction prediction can be assessed. The performance comparison with existing methods is shown in Supplementary Table S5. MemBrain achieves 90.1%/87.9% prediction accuracy and an MCC of 0.555/0.526 on the training/independent data set, which performs better than the existing methods. As TMH interaction prediction is derived from the predicted contacts, globular protein contact predictors perform poorly on inter-TMH residue contact prediction and thus are not as good as the TMH protein contact predictors for TMH–TMH interaction prediction. Also, as can be seen from Supplementary Figure S5, the predicted contacts obtained by PSICOV (Jones et al., 2012) are widespread. In other words, it predicted more interacting TMH pairs with higher sensitivity. Meanwhile, it predicted more spurious interacting TMH pairs as well with lower accuracy, which resulted in a relative lower MCC compared with MemBrain.

3.6 Benchmark test on MemBrain-guided GPCR structure modeling

To have a direct examination of the impact of the MemBrain contact predictions on three-dimensional (3D) GPCR modeling, we collected all 13 GPCR proteins with known structure in the PDB library (PDBID: 1u19A, 2rh1A, 2y00A, 273A, 3emlA, 3dua, 3pblA, 3rzeA, 3w7A, 4dajA, 4djhA, 4ea3A, 4grvA, see Table 2). We then generate GPCR models by I-TASSER (Roy et al., 2010; Xu et al., 2011) with MemBrain contacts and also models without using MemBrain contacts.

Although MemBrain contact predictions are only on the Cα atoms, we found that the best GPCR models can be generated when the same contact restraints are extended for all Cα, Cβ and side-chain center of mass in the I-TASSER simulations. Table 2 summarized the results of I-TASSER with and without using MemBrain contacts, where all GPCR templates and the homologous templates with a sequence identity >30% or detectable by PSI-BLAST are excluded. First, although no GPCR templates were used, the inherent I-TASSER fragment assembly simulation was able to assemble structures with an approximately correct TMH arrangement (TM-score >0.5) in 7 of 13 cases; in 5 cases, the models have a TM-score >0.6. The average TM-score and RMSD to the X-ray structure are 0.465 and 10.9 Å, respectively, in the TMH regions.

The average accuracy of the MemBrain contact predictions is 0.57 for the top L/5 predictions, whereas I-TASSER uses the top L contacts, which have an average accuracy of 0.35. As shown in Table 2, the incorporation of MemBrain contacts has significantly improved the quality of the GPCR models. For example, the number of cases with a TM-score >0.6 increases from 5 to 10 and the average TM-score increases by 37%. This TM-score improvement on GPCRs by MemBrain contacts is dramatically higher than that observed on the globular proteins by SVMSEQ (Wu and Zhang, 2008) contacts for the I-TASSER models, which have a TM-score increase by 4.6% (Wu et al., 2011). The major reason for the difference is the higher accuracy of the long-range contacts by MemBrain, which have been
specifically trained for TMHs. If we apply SVMSEQ on the 13 GPCRs, the average accuracy of contacts is only 0.16 for top 30% detectable by PSI-BLAST have been excluded during I-TASSER simulations.

As expected, there is an obvious correlation between the TM-score improvement and the accuracy of the MemBrain contacts. For the targets with the most significant improvement (e.g. 2t73A and 3rzeA), the accuracy of MemBrain for the top L contacts is 46 and 31%, respectively. For targets of relatively lower accuracy (3em1A with a contact accuracy 25%), the TM-score of the I-TASSER models is slightly improved but still in the unfoldable region (TM-score ≈ 0.315). There are some targets, e.g. 4djhA, where the MemBrain contact accuracy is high (0.57 for L/5 and 0.37 for L), but the TM-score improvement (12%) is lower than average because a high proportion of MemBrain contacts (78%) are already included in the I-TASSER models. Nevertheless, the remaining 22% of the novel contacts from MemBrain serve to refine the models.

In Figure 3, we present the superposition of MemBrain-guided I-TASSER models and X-ray structures for all 13 GPCRs. Most of the I-TASSER models indeed show a similar topology of TMH packing to the targets. However, there are a few targets (e.g. 2t73A, 3pblA and 4djhA) where the I-TASSER models have the TMH structure bended away from the axis at the middle of the helices. This is due to the fact that the I-TASSER simulations were fully automated and no TMH requirements were implemented. Thus, the bended helices from non-homologous templates resulted in the helix bending. If TMH predictions from MEMSTAT (Jones et al., 1994) were used, the TMH bending could be eliminated but the overall TM-score was not obviously improved (data not shown).

Although the MemBrain contacts are only in the TMH regions, the overall I-TASSER models along the entire chain are also improved. As shown in Supplementary Table S6 and Supplementary Figure S6, the average TM-score of the whole-chain models is increased by 35% and RMSD reduced by 2.6 Å. The number of targets with a correct fold (TM-score > 0.5) increases from 5 to 10, when the MemBrain contact predictions are used. It is worth mentioning that all GPCR templates and homologous templates with sequence identity > 30% were both excluded in aforementioned I-TASSER simulations. If we only excluded the homologous templates with sequence identity > 30% in I-TASSER simulations, in which case some GPCR templates can be kept, then the final average RMSD is as low as 1.8 Å with a TM-score 0.902, as demonstrated in Supplementary Table S7.

### 3.7 Case studies

The usefulness of MemBrain contact predictions has been demonstrated in the case of GPCR 3D structure modeling. To further demonstrate the effectiveness of MemBrain contact predictions for other TMH proteins besides GPCRs, we performed structure prediction for two newly solved membrane protein structures, i.e. heterodimeric ABC transporter (3q4A) (Hohl et al., 2012) and channelrhodopsin (ChR) chimera between ChR1 and ChR2
Although the predicted contacts did not have much effect on the first model, the contacts did reduce the number of native contacts for the two proteins, respectively. Their correlation coefficients derived from δ-analysis (δ = 4) are both 100%. These results are shown in Supplementary Figure S7.

Similar to the GPCR benchmark, homologous templates were removed. We first tried removing templates that have sequence identity >30%. In both proteins when contact predictions were not used, the RMSD in the TMH region was 4.3 and 3.0 Å for 3qf4A and 3ug9A, respectively. The results are given in Table 3. For both proteins, the accuracy without predicted contacts is already high. Although MemBrain predictions are good relative to current contact prediction methods, the probability of improving these structures further with predicted contacts is small, and the structures may even be made slightly worse. This was indeed the case for both proteins (Table 3), where no improvement occurred. Nevertheless, contact prediction can improve structures of mediocre quality, so we also tried removing templates that have sequence identity >15% to generate these structures for improvement.

For 3qf4A, in the 15% sequence identity case, the resulting model without MemBrain contacts remained similar at 4.2 Å in the TMH region, but the RMSD of the entire chain increased significantly by >19.0 Å relative to the 30% case. With MemBrain contacts, the RMSD in the TMH region also remained similar at 4.2 Å, and the whole-chain RMSD increased by 17.0 Å. Although the predicted contacts did not have much effect on the first model, the contacts did reduce the number of poorly predicted structures among the top 10 models. Without the predicted contacts, the two worst models were 9.9 and 20.4 Å; whereas with the contacts, the worst model was 5.5 Å, which makes the significant improvement in the top 10 models in terms of both the average RMSD and TM-score (Table 3).

Table 3. The top I-TASSER model (largest cluster) for 3qf4A and 3ug9A with or without using MemBrain contact predictions

<table>
<thead>
<tr>
<th>PDBID</th>
<th>Cutoff</th>
<th>RMSD/ TM&lt;sup&gt;a&lt;/sup&gt;</th>
<th>RMSD/ TM&lt;sup&gt;b&lt;/sup&gt;</th>
<th>RMSD/ TM&lt;sup&gt;c&lt;/sup&gt;</th>
<th>RMSD/ TM&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>3qf4A</td>
<td>30%</td>
<td>4.3/0.607</td>
<td>4.2/0.610</td>
<td>6.1/0.715</td>
<td>6.0/0.716</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>4.2/0.605</td>
<td>4.2/0.606</td>
<td>25.3/0.437</td>
<td>25.0/0.449</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>16.7/0.264</td>
<td>16.5/0.261</td>
<td>22.2/0.288</td>
<td>22.0/0.286</td>
</tr>
<tr>
<td>3ug9A</td>
<td>30%</td>
<td>3.0/0.795</td>
<td>3.2/0.765</td>
<td>15.0/0.649</td>
<td>14.9/0.642</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>5.6/0.524</td>
<td>4.2/0.599</td>
<td>12.5/0.450</td>
<td>11.9/0.500</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>10.9/0.369</td>
<td>5.3/0.600</td>
<td>19.9/0.232</td>
<td>16.0/0.250</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>15.0/0.219</td>
<td>6.9/0.564</td>
<td>17.4/0.231</td>
<td>16.3/0.472</td>
</tr>
</tbody>
</table>

Note: The RMSD and TM-score to native for different degrees of template structure removal are given.

<sup>a</sup>RMSD (Å) and TM-score in the TMH region using MemBrain predictions.

<sup>b</sup>RMSD (Å) and TM-score in the TMH region using MemBrain predictions.

<sup>c</sup>RMSD (Å) and TM-score of the whole chain using MemBrain predictions.

<sup>d</sup>RMSD (Å) and TM-score of the whole chain using MemBrain predictions.

| Templates whose sequence has >X% sequence identity to the protein were removed.

<sup>e</sup>Average results of the top 10 models, where templates whose sequence has >15% sequence identity to the protein were removed.

<sup>f</sup>Templates whose structure has TM-score >0.5 to the native structure were removed.

(3ug9A) (Kato et al., 2012). The pairwise sequence identities of the two proteins to the training data set are <30%. Using MemBrain, 64.5 and 80.7% of the predicted top L/5 contacts are native contacts for the two proteins, respectively. Their corresponding accuracies derived from δ-analysis (δ = 4) are both 100%. These results are shown in Supplementary Figure S7.

To challenge MemBrain further, templates with TM-score >0.5 to the X-ray structure were removed to increase the dependence on the predicted contacts rather than on the templates for protein folding. In this case, the improvement was significant for 3ug9A, where the change was from 15.0 to 6.9 Å in the TMH region. We tried the same >0.5 cutoff for 3qf4A, but the result was >16 Å in the TMH region for both cases because of the lack of inter-helical contact information in the intracellular domain.

4 DISCUSSION
Our results demonstrate that the weakness of CMA (sufficient homology) can be compensated for by combining it with ML methods, and the weakness of statistical ML (local condensed predictions) can also be complemented by combining with CMA-based approaches. It would be interesting to investigate whether the idea of combining ML-based methods with CMA-based approaches in MemBrain can also be used to enhance the contact prediction accuracy in globular proteins. To demonstrate this, we used PSI-BLAST to search against UniRef90 database for the 22 targets containing free modeling or free modeling/template-based modeling domains in CASP9. Supplementary Figure S8 shows the distributions of the number of homologous sequences on the CASP9 targets.

The results in Supplementary Figure S8 demonstrate: (i) Although some targets have very few or no homologous structures in the PDB, which causes difficulties for homology structure modeling, they have a large number of homologous sequences available. For example, four of these targets have more than 1000 homologous sequences in the UniRef90 database. This means we can get reliable predicted contacts for these four proteins (~61% average accuracy for top L/5 predictions on the training data set); (ii) In the other 18 targets, 9 of them contain less than 250 homologous sequences. According to our results, the accuracy of PSICOV in this group was very low (12.8% in this work), and the combination of ML-based method with PSICOV is expected to improve the accuracy (23.6% in this work). Furthermore, there are six proteins in the homology size region of (250 500). When directly applying PSICOV on this group, the results are also not satisfactory (28.0% in this work). However, the MemBrain prediction significantly improves the results in this group (73.5% in this work).

To further illustrate the potential merit of the combination of ML-based methods with CMA-based approaches on globular proteins, we replaced the ML-part of the MemBrain predictor with SVMSEQ. The predictions of SVMSEQ for the 22 targets...
were extracted from http://www.predictioncenter.org/download_area/CASP9/predictions/, which were merged with PSICOV outputs using a linear combination strategy. Supplementary Table S8 lists the detailed performance of top L/5 predictions for each range contact on the 22 targets. As can be seen, the performance improved as expected. From the results, it is very interesting to see the following: SVMSEQ performs better than PSICOV on the short-range contacts (33.6% versus 12.8%), and worse than PSICOV on the long-range contacts (17.1% versus 20.3%). This is consistent with our previous observation that the ML-based approach gives more local condensed predictions. The results also show that SVMSEQ and PSICOV predictions are largely complementary on long-range contacts, which makes the combination of them more accurate. These results suggest that a proper consensus of CMA-based approaches with ML-based methods is also promising for enhancing the ab initio contact predictions in globular proteins.

Besides inter-TMH residue contacts in TMHs (denoted as Type 1), there exist two other contact types in TMH proteins: contacts where one residue is in the membrane region and the other residue in the water-soluble loop region (denoted as Type 2), and contacts where both residues are in the water soluble region (denoted as Type 3). These two types of contacts are also important for 3D structure prediction, especially for the loop region. By observing the native contacts from known structures in the training and the independent data set at the whole-chain level, we found that 48.0 and 45.9% are Type 3 contacts, and 14.1 and 14.8% are Type 2 contacts. Hence, predicting these two types of contacts will be important future work. In this case, the predictors developed for predicting residue contacts in globular proteins are expected to contribute.

Funding: The National Natural Science Foundation of China (61222306, 91130033, 61175024), Shanghai Science and Technology Commission (11JC1404800), a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (201048), the National Science Foundation Career Award (DBI 0746198) and the National Institute of General Medical Sciences (GM083107, GM084222).

Conflict of Interest: none declared.

REFERENCES


