JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

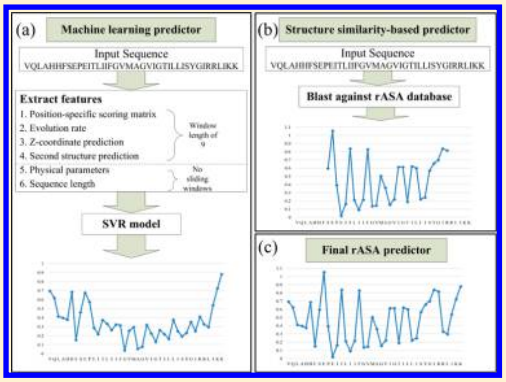
Prediction Enhancement of Residue Real-Value Relative Accessible Surface Area in Transmembrane Helical Proteins by Solving the Output Preference Problem of Machine Learning-Based Predictors

Feng Xiao and Hong-Bin Shen*

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

Supporting Information

ABSTRACT: The α-helical transmembrane proteins constitute 25% of the entire human proteome space and are difficult targets in high-resolution wet-lab structural studies, calling for accurate computational predictors. We present a novel sequence-based method called MemBrain-Rasa to predict relative solvent accessibility surface area (rASA) from primary sequences. MemBrain-Rasa features by an ensemble prediction protocol composed of a statistical machine-learning engine, which is trained in the sequential feature space, and a segment template similarity-based engine, which is constructed with solved structures and sequence alignment. We locally constructed a comprehensive database of residue relative solvent accessibility surface area from the solved protein 3D structures in the PDB database. It is searched against for segment templates that are expected to be structurally similar to the query sequence's segments. The segment template-based prediction is then fused with the support vector regression outputs using knowledge rules.



Our experiments show that pure machine learning output cannot cover the entire rASA solution space and will have a serious prediction preference problem due to the relatively small size of membrane protein structures that can be used as the training samples. The template-based engine solves this problem very well, resulting in significant improvement of the prediction performance. MemBrain-Rasa achieves a Pearson correlation coefficient of 0.733 and mean absolute error of 13.593 on the benchmark dataset, which are 26.4% and 26.1% better than existing predictors. MemBrain-Rasa represents a new progress in structure modeling of α-helical transmembrane proteins. MemBrain-Rasa is available at www.csbio.sjtu.edu.cn/bioinf/MemBrain/.

1. INTRODUCTION

Membrane proteins (MPs) are mostly located in cell membranes and are encoded by 20−35% of genes.[1−3] They have many special and complex physiological functions and take part in various vital activities of cells, including molecular transport, cell recognition and immune response, and signal transduction and energy transfer. MPs are important drug targets and constitute approximately 60% of known drug targets. A transmembrane protein (TMP) is a type of membrane protein (MP) which spans the entire biological membrane to which it is permanently attached. There are two basic types of TMPs: α-helical (TMH) and beta-barrels (TMB) according to transmembrane segment classification of secondary structure. The membrane-embedded α-helical, polytopic proteins constitute the majority of ion channels, transporters, and receptors in living organisms.[4]

TMH proteins are infamously difficult targets for high-resolution structural studies. Because of 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. Although helical membrane proteins present a high degree of experimental difficulty, their conformation is in many ways predictable.[5−8] For instance, the transmembrane helices must span the hydrophobic layer of membrane or membrane-mimetic detergent micelles, whereas the amphipathic helices or loops are either associated with the headgroup region or exposed to bulk solvent.[9] These conditions effectively speed up the progress of computational model construction. To date, many pseudo-2D space topological structure predictors have been developed, such as the TMH segment prediction tools MemBrain,[9] Memsat,[10] Memsat-SVM,[11] TMHMM,[12] a large-scale benchmark on existing TMH predictors,[13] and the TMH-TMH contact map prediction.[14]

Progress in 3D structure prediction for membrane proteins is much slower than that for soluble proteins. Only a few membrane-specific protein 3D structure predictors are available, e.g., single chain predictors Membrane-Rosetta[15] and FILM3[16] and several multichain complex modeling systems such as BCL::MP-Fold.[17] However, for soluble proteins, there are now many well-established software packages that can be used to predict their 3D structures, e.g., I-TASSER,[18] Rosetta,[19]

Received: April 28, 2015
Published: October 10, 2015

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

Table 1. Comparison among Existing Methods for Predicting Protein Residue rASA

method	algorithm	application scope	binary/real value	no. of training samples	reported correlation coefficient
SABLE	neural networks	globular protein	real	603	0.66
ASAP	SVR	transmembrane	real	59	0.65
TMX	SVM	transmembrane	binary	43	–
TM expo	SVR	transmembrane	real	21	0.66
MPRAP	SVR	whole chain of membrane protein	real	80	0.58

TASSER,[20] etc. This is due to the rapid increase in the number of solved soluble protein 3D structures in the PDB (Protein Data Bank).[21] According to the recent CASP (Critical Assessment of protein Structure Prediction) competitions, the top ranked predictors are generally the template-based algorithms.[22] However, when compared with soluble proteins, MPs' structures are hard to solve due to the chemical complexity of the biological membrane phospholipid bilayer. For example, in the PDB (released on 2014/05/16), there are in total 103,015 solved structure records, yet only 2,131 proteins are TMPs, of which 1,840 (~1.8%) are α-helical, and 283 (~0.3%) are beta-barrel membrane proteins. This situation is not expected to improve dramatically in the near future. The lack of high-resolution membrane protein 3D structures and the reality that the membrane protein structure is far more complex than previously thought significantly slow down the progress of developing powerful predictors.

Previous studies have demonstrated that accurately predicting the structural features in a pseudo-2D space from the amino acid sequence is an important step to generate a reliable model in the 3D conformational space. In soluble proteins, predictions of residue secondary structure (SS) and relative accessible surface area (rASA) are well studied.[23] Many methods have been developed to predict these features, such as PSIPRED,[24] SSpro,[25] and Spine-X[26] for SS and SABLE[27] for SA (solvent accessibility). In α-helical TMPs, there are also several methods to predict SS accurately, such as Memsat,[10] Memsat-SVM,[11] MemBrain,[9] and TMHMM[12] for TMH segments prediction. During the past decade, along with many TMH protein structures being solved, a few methods for predicting membrane protein residue solvent accessibility have also been developed. These can be generally grouped into two classes according to their application area: (1) methods for predicting the residue exposure feature within the membrane, e.g., ASAP,[28] TMX,[29] and TMexpo,[30] and (2) methods that can be applied to the entire membrane protein sequence, both within and outside of the cell membrane regions, e.g. MPRAP.[31] Table 1 has briefly summarized the features and reported performances of these existing predictors.

All of the methods for predicting membrane protein residue rASA mentioned above are machine-learning-based predictors. In these approaches, each residue is encoded using a discrete feature vector, which is further fed into a trained statistical machine-learning model for prediction, e.g., support vector machine (SVM) or support vector regression (SVR). It is worth pointing out that the prediction of rASA on the full sequence is more challenging than those that only apply to residues inside the cell membrane (TMH regions). To the best of our knowledge, the MPRAP is the only predictor that can be applied to the entire membrane protein chain to date. The reason can be due to the fact either that the loop region has more flexibility or that a single statistical learning model cannot catch the different features of TMH and loop residues. In this article, we report a more powerful whole-sequence-oriented computational approach (MemBrain-Rasa) to predict real-value rASA of the α-helical TMPs. The new algorithm features by a consensus protocol that combines both statistical machine learning-based prediction and segment template-based structural similarity estimates. Benefiting from recent progress in structural biology, more 3D protein structures are deposited into the PDB, which significantly increases the possibility to find similar segment templates for a given amino acid sequence.

In MemBrain-Rasa, the machine-learning engine is achieved by a SVR model which will generate real-value outputs. The input features to SVR include 6 types: position specific scoring matrix (PSSM), predicted evolution rate by rate4site (R4S), predicted residue Z-coordinate, predicted secondary structure (SS) information, representative physical parameters (PP), and sequence length. To fully exploit experimentally solved protein 3D structure information and solve the output preference problem of SVR, we designed a complementary structure template-based estimation engine motivated by the SSPro for soluble proteins.[25] We locally constructed a comprehensive protein rASA-orientated database with the corresponding amino-acid sequences from all of the known 3D structures in the PDB. Then, we used the BLAST software to detect the segment templates in the constructed rASA database for the query membrane protein sequence. The annotated rASA of the homology structural segments will be used to estimate the rASA of their aligned segments on the query sequence. In the case where no structural templates can be found, the final prediction is dependent on the SVR outputs. This new fusion protocol enables MemBrain-Rasa to generate a great improvement on the rASA predictions for the whole chain. On the same benchmark data set as MPRAP, MemBrain-Rasa achieves a correlation coefficient (CC) of 0.733 and a MAE (mean absolute error) of 13.593, which are much higher than the CC of 0.58 and the MAE of 18.4 achieved by MPRAP.

2. MATERIALS AND METHODS

2.1. Membrane Protein Benchmark Data Set. In order to fairly evaluate the developed predictor and concurrently remain capable of predicting the solvent accessibility of both single- and multiple-chain membrane proteins, we used the same benchmark data set and the way for calculating the residue rASA values as originally used in MPRAP.[31] In this data set, the sequence identity cutoff was set to 20%, and the length cutoff (which specifies the length of coverage) was set to 0.9. Fragments, low-resolution structures, and structures with secondary structure or membrane boundary problems were excluded. Finally, this data set consists of 52 membrane protein complexes composed of 80 chains. This data set was also divided into five folds in advance to avoid high homology in different folds, where chains from the same superfamily were put in the same fold. All of the following experiments of this study are performed on the five folds with a 5-fold cross-validation.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

2.2. Calculation of rASA. The residue solvent accessibility surface area (ASA) is defined as the sum of all atomic surface accessibility in the residue. Several software programs are capable of calculating ASA, such as MSMS,[32] Naccess,[33] and Dictionary of Protein Secondary Structure (DSSP).[34] A residue's ASA is divided by the corresponding standard accessibilities to generate the relative solvent accessibility surface area (rASA).[35] Note that different programs can possibly differ on how to choose the standard, resulting in slight differences in their outputs. For instance, in MSMS, the standard accessibility for each residue is a GLY-X-GLY tripeptide with extended conformation; while in Naccess, an extended ALA-X-ALA tripeptide is used. In this study, the rASA of each residue was calculated by Naccess 2.1.1 (http://www.bioinf.man.ac.uk/naccess), the same as that used in MPRAP.[31] In Naccess, the probe size 1.4 Å mimics water molecules, 2.0 Å mimics the CH2-group, and the combination of these two values is used to calculate the rASA of transmembrane proteins. It is worth noting that for making the developed predictor able to predict residue solvent accessibility of both single- and multiple-chain proteins, when in the case of a complex composed by multiple chains, the residues' rASA values are calculated in its complex state, i.e., all chains in the complex are considered simultaneously in their nature complex state. This is different from the single chain calculation mode, where the interchain interactions are not considered, which will affect the rASA values on the interaction interface.

2.3. Machine-Learning-Based Prediction Engine.

2.3.1. Feature Extraction. We extracted six types of features from a given amino acid sequence, including position specific scoring matrix (PSSM), evolution rate predicted by rate4site (R4S), Z-coordinate score, predicted secondary structure (SS) information, representative physical parameters (PP), and sequence length. These features can represent both global and local characteristics of the query protein amino acid sequence.

2.3.1.1. Position Specific Scoring Matrix. Position specific scoring matrix (PSSM) contains evolutionary information calculated from an aligned set of sequences. In this study, it is generated by PSI-BLAST to search against the UniRef90 database with three iterations and an E-value cutoff of 10^{-5}. The PSSM for each chain is represented by an $L \times 20$ matrix, where L is the sequence length. The raw PSSM scores were scaled to the range $[0, 1]$ using the standardized logistic function

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

where a is the original score.

2.3.1.2. Evolution Rate. Rate4Site (R4S) is used to detect conserved amino-acid residues by computing the relative evolutionary rate for sites in the multiple sequence alignment (MSA). According to previous work,[36] buried residues will evolve slowly and are considered to be conserved, whereas exposed residues evolve rapidly and are considered to be the opposite. At last, the relative substitution rate was found almost linearly related to the solvent accessibility in membrane protein complexes. In our experiments, the CC between the conserved score alone obtained by Rate4Site, and the relative solvent accessibility was 0.44. In addition, the conserved score was found to be the most important in all features mentioned in this article according to some feature-selection algorithms, i.e., mRMR[37] and stepwise discriminant analysis (SDA).[38] Considering its importance, the evolution rate score is also incorporated into MemBrain-Rasa. To restrict the computational time for computing the residue evolution rate, at most the top 50 selected sequences per structure were allowed. If the number of aligned sequences is less than 50, all aligned sequences are included for computation.

In Rate4Site, the rate p_{ij} of residue i in protein j is first normalized by subtracting the average conservation score $\overline{p}_j$ and dividing by the standard deviation δ_j of the protein j.

$$q_{ij} = (p_{ij} - \overline{p}_j)/\delta_j \tag{2}$$

where q_{ij} is the normalized value of p_{ij}. The raw score of q_{ij} is then further normalized among all the sequences using the following equation:

$$q'_{ij} = \frac{q_{ij} - \min(q_j)}{\max(q_j) - \min(q_j)} \tag{3}$$

2.3.1.3. Z-coordinate. Z-coordinate is an important constituent in the field of membrane protein structure prediction and is used to detect the relative position of a residue with respect to the membrane. A decade ago, Nugent et al. used FILM to study folding in the lipid membrane region and took the z-axis as perpendicular to the bilayer surface to calculate the energy potential.[16] Researchers used Membrane-Rosetta[15] and BCL::MP-Fold[17] with the Z-coordinate to improve their structure prediction of membrane proteins. Zpred[39] was developed to predict the residue Z-coordinate, and its results represent the relative position of each residue in the membrane. For this article, we also constructed MemBrain-Rasa with the predicted Z-coordinate by Zpred to encode a residue. According to our statistics, almost all of the predicted absolute Z-coordinate scores were no more than 25 Å, and then the absolute values were normalized by dividing by 25, which is used in the final system.

2.3.1.4. Secondary Structure. Secondary structure information has proved effective in the field of solvent accessibility and torsion angle prediction of soluble proteins[26,40,41] and is also expected to be useful for this study's membrane proteins. In this article, the secondary structure of each residue was predicted by PSIPRED to be one of three classes (coil (C), helix (H), and strand (E)). For each residue, we take its predicted three possibilities directly as the input features.

2.3.1.5. Representative Residue Physical Parameters. Tremendous statistics studies have demonstrated that different amino acids have different properties, resulting in different propensities for forming specific 3D spatial conformations. For instance, the TMH segments in membrane proteins are dominated by hydrophobic amino acids. According to our local tests, we found that the following 12 representative amino-acid physical parameters can improve the system performance and hence were incorporated into the prediction system: a steric parameter, hydrophobicity, volume, polarity, isoelectric point, helix probability, strand probability, average accessible surface area (ASA), charge, acidity, the probability of occurrence, and average mass of 20 common amino acids (Supporting Information). Note that all of these parameters were normalized to range from 0 to 1.

2.3.1.6. Sequence Length. In addition to the features mentioned above, sequence length as a global feature is also used in this study. A one-dimensional vector was used to encode sequence length. In the MemBrain-Rasa system, if a

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

query sequence length was less than 900, then its sequence length score was $0.1*\lceil N/100 \rceil$; otherwise, the score was set to 1.

2.3.2. Using a Sliding Window Approach to Include Neighborhood Information. We then used a sliding window to include more useful residue neighborhood information to improve the prediction accuracy. This makes sense when considering that the status change of neighboring residues is usually continuous. In this study, we used sliding windows to cover neighborhood information in four types of features: (1) PSSM, (2) evolution rate, (3) Z-coordinate, and (4) predicted secondary structure information. In our work, we found that the window size of 9 was optimal. Thus, for each residue, we obtained $(20 + 1 + 1 + 3) \times 9 = 225$ encoded components by applying the sliding window to the four aforementioned features. Furthermore, by combining with two other features (representative residue physical parameters and sequence length), each residue was encoded into a 238-D vector (225 + 12 + 1). The constructed feature vectors were used to train a SVR model as described next.

2.3.3. Support Vector Regression. SVM is a type of supervised learning algorithm, which has been proved to be successful in dealing with complex biological data.[42] Support vector machines include support vector classification (SVC) and SVR. SVC has been demonstrated to work very well for binary classification, whereas SVR is better for the real value prediction. In our study, SVR was applied since we aim to predict the real-value rASA of α-helical TMPs. In SVR, there are three popular kernels that can be used: linear, polynomial, or radial basis functions. In our experiments, the radial basis kernel was used due to its better performance. The parameters in SVR were optimized using a grid-search method. SVM-light packages are applied to the model construction, which is available at http://svmlight.joachims.org/.

2.4. Segment Template-Based Prediction. In order to improve the prediction of rASA, we not only used machine learning but also developed a segment template-based engine in order to use the experimentally solved 3D structure knowledge for prediction.[25] We constructed a local residue rASA database using the following steps. We first retrieved all solved structures downloaded from http://www.pdb.org/ then removed the duplicate structures by CD-HIT,[43] resulting in 111,648 structure records. Then, the corresponding rASA index table was constructed by calculating rASA for each structure using the Naccess program. In this way, we have constructed a comprehensive database covering all the recent solved structures.

In the prediction step, we developed a homology-segment-based prediction protocol. Given a query sequence, we first try to identify the homology sequences from the constructed database using the BLAST program. A homology segment is found if the following five conditions are satisfied: (1) the minimum length of aligned structures is 20, (2) gap is not allowed, (3) the maximum E-value is 10^{-9}, (4) the minimum sequence identity is 70%, and (5) the minimum positive is 75%. To provide a completely fair evaluation, we excluded searched protein chains that are identical to any chain in the benchmark data set in our following experiments. If segments that meet the above five requirements are identified, they will be used to estimate the residue rASA for the query sequence with the hypothesis that homology sequences will have similar structures. We summed the rASA values of the corresponding positions in the sequences from the structure templates and divided by the total number of templates. The concrete procedure can be seen in Figure 1.

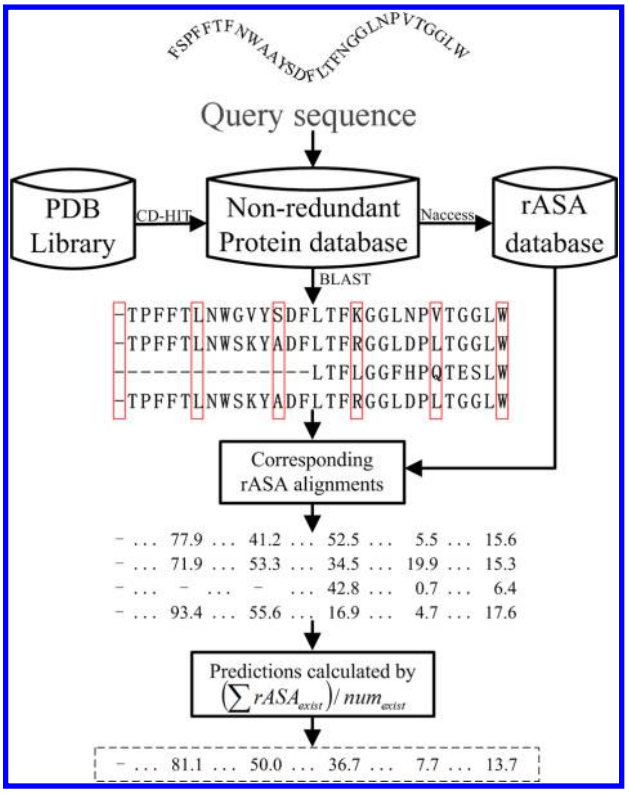


Figure 1. Protocol of structure segment template similarity-based rASA prediction.

2.5. Hierarchical MemBrain-Rasa Consensus Prediction System. The final system contains both machine-learning SVR-based prediction and the structure similarity-based prediction. To realize a proper combination of the two estimates, we used the following simple knowledge-guided approach as shown in Figure 2. The input to the final

1. Given a query sequence

2. Generate prediction output using the trained SVR model (denoted as ML)

3. Search the query sequence against the local rASA database to obtain the profile

For each residue in the query sequence,

IF Not exist residues from homolog structure alignments

Then rASA=ML

Else Generate prediction output using the searched homology segments (denoted as Hom)

Else if (|ML-Hom|>Threshold)

Then rASA=Hom

Else rASA=ML

Figure 2. Knowledge-guided approach to hierarchically combine machine learning and structure similarity-based predictions. Threshold is used to denote the deviation from the two prediction engines, and 5 is used in this study.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

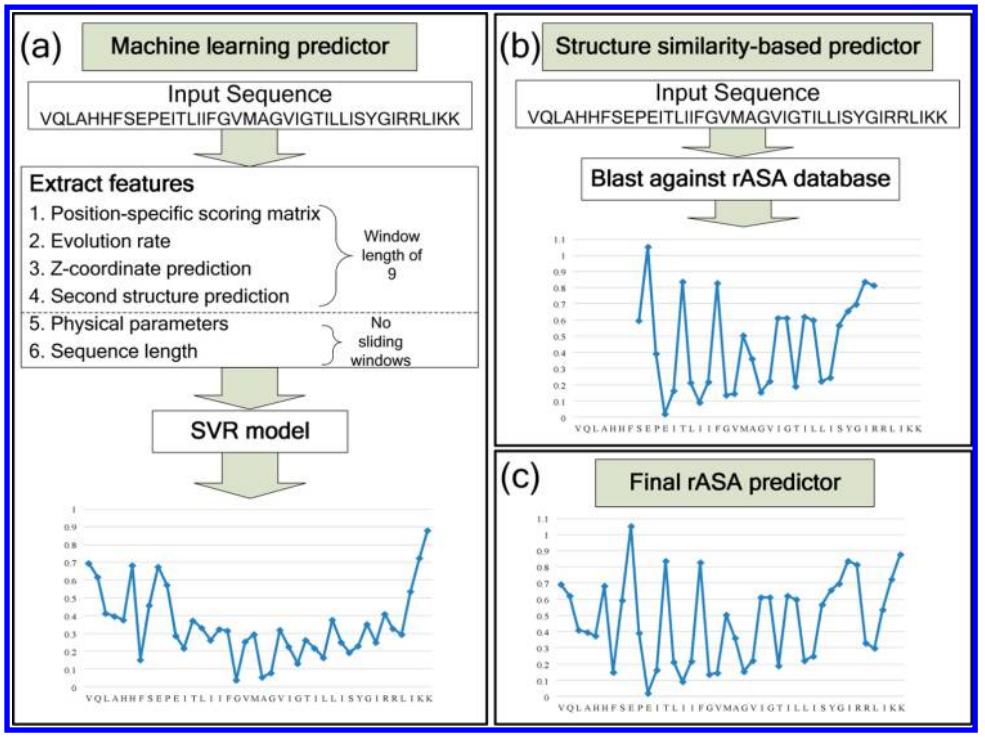


Figure 3. Flowchart of the three-step approach in MemBrain-Rasa. (a) The procedure of a machine learning predictor. (b) The procedure of segment structure similarity-based predictor. (c) The final rASA predictor by combining both machine learning and structure similarity-based predictors.

MemBrain-Rasa system is the primary sequence, and the output is the predicted rASA for each residue, which can be summarized in three steps as follows (Figure 3). In the first step, we extract six types of features and use a trained SVR for predictions. In the second step, we use the BLAST program to search against the local rASA database and obtain a prediction for those residues that have homology segments in the rASA database. In the third and final step, we combine both of the two predictions together by using the IF THEN knowledge rules. The threshold used in the fusion process of Figure 2 is 5 according to the preliminary local experiments.

2.6. Performance Evaluation. All of the results shown in this article are from 5-fold cross-validation. To quantitatively evaluate the real value predictions of relative solvent accessibility prediction, mean absolute error (MAE) and Pearson correlation coefficients (CC) were used. CC is a measure of the linear correlation between predictions and real values. CC value ranges in $[-1, 1]$, where -1 represents total negative correlation, 0 no correlation, and 1 total positive correlation. The definition of CC is

$$CC = \frac{\sum_{i=1}^{L}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{L}(x_i - \bar{x})^2][\sum_{i=1}^{L}(y_i - \bar{y})^2]}} \quad (4)$$

where x_i and y_i represent the i-th residue's observed and predicted values respectively, $\bar{x}$ and $\bar{y}$ represent the corresponding mean values. L represents the total number of residues.

The mean absolute error (MAE) measures how close predictions are to their real values and is defined as the average difference between the predicted and observed rASA values of all residues, i.e.,

$$MAE = \frac{1}{L}\sum_{i=1}^{L}|y_i - x_i| \quad (5)$$

3. RESULTS

In the final MemBrain-Rasa system, there are two subengines. One is the machine learning SVR-based engine, and the other is the homology segment structure similarity-based engine. Our experimental results show that these two parts complement each other very well, resulting in a significant improvement when compared with the traditional pure machine learning-based predictors.

3.1. Machine Learning Prediction Engine. *3.1.1. Parameter Selection.* Parameters are one of the key factors of a machine-learning algorithm, and different parameters can result in very different results. In this part, there are a total of three parameters: window length for encoding the features and the two parameters of "C" and "g" in the SVR algorithm. For the window length, different sizes of a symmetric windows were tested, and we found 9 to be optimal and used that in this study, which is the same as that in MPRAP.[31] In the SVR algorithm, "C" represents the trade-off between training error and margin, and "g" represents parameter gamma in the RBF kernel function. These two parameters were optimized through a grid search approach with 5-fold cross-validation, where the parameters corresponding to the best prediction results are used. In the final version of the SVR model, we selected a radial basis kernel function (RBF) with g = 0.04 and C = 18.

3.1.2. Feature Fusion Improves the Prediction Accuracy. A basic rule for the statistical machine-learning algorithm is "good input, good output." In order to make the SVR learn effective rules from the data set, the input features to SVR must be useful. However, not all the features that can be calculated are

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

useful for the residue rASA prediction. To test the six features discussed above, we have tested both independent features and their combinations. Table 2 shows the results.

Table 2. Prediction Performance of Membrane Protein Residue rASA Using the SVR Predictors Based on Individual Input Features and Their Various Combinations

input features[a]	MAE	CC
PSSM	23.00	0.485
R4S	24.78	0.374
SS	26.52	0.284
Zpred	27.32	0.197
PSSM+R4S	22.140	0.525
PSSM+R4S+Zpred	18.202	0.583
PSSM+R4S+Zpred+SS	18.159	0.589
PSSM+R4S+Zpred+SS+length	18.136	0.593
PSSM+R4S+Zpred+SS+length+PP	**18.013**	**0.599**

[a]PSSM means position specific scoring matrix, R4S means evolution rate calculated by rate4site, SS means predicted second structure, Zpred means z-coordinate prediction, length means sequence length, and PP means representative physical parameters.

All of the four independent features (PSSM, R4S, SS, and Zpred) contain some useful information for predicting rASA by themselves as is shown in Table 2. Among these independent inputs, PSSM achieves the best overall results (MAE = 23.00 and CC = 0.485). This result suggests that PSSM is important in rASA prediction, possibly due to the existence of some residue evolutionary knowledge. R4S is tested as the second-best feature for predicting the residue rASA values. The success of the R4S feature is also due to its strong correlation with residue evolution rates and is consistent with previous research.[36]

When combining multiple features, we find that the best CC values always accompany the least MAE, indicating that the two evaluation matrices are consistent, as was also shown by a previous study that the CC values are almost linearly related to the MAE values.[28] Since our independent feature test has shown that PSSM and R4S have the best results on rASA prediction, we added these two features together first. The sequence encoding scheme of the combination of "PSSM+R4S" increases the prediction accuracy to CC = 0.525, which is a significant improvement compared to CC = 0.485 (PSSM alone). Afterward, the encoding scheme "PSSM+R4S+Zpred," which adds the Zpred information, gives another obvious improvement. Compared to "PSSM+R4S," the MAE value improves by almost 4%, and the CC value reaches 0.58 for the new input. This is likely due to the fact that the Z-coordinate predicted by Zpred is a different point of view to the target residue from the PSSM and R4S and contains the information that PSSM and R4S do not have, resulting in a much better prediction model. The performance of the "PSSM+R4S +Zpred" model is very comparable to those reported in the literature of MPRAP (MAE = 18.4 and CC = 0.58). To further test whether other features also contribute, we continued to add SS, sequence length, and PP in turn to the feature set. Our results show that we obtain the best overall result when combining all of the independent features together, with CC = 0.599 and MAE = 18.013. Through this stepwise feature test process, we finally selected the encoding scheme of "PSSM +R4S+Zpred+SS+length+PP" to generate a SVR prediction model.

3.1.3. Neither Dimension Reduction nor the Two-Stage SVR Model Considerably Help to Improve Performance. In order to improve prediction performance, in addition to adding new features, we have made many other attempts. We first considered that there could be redundant components in the feature set; thus, we applied two feature-selection methods which have been proven effective to solve classification problems in the field of machine learning and bioinformatics.[38,44] The two tested feature-selection methods are the minimum redundancy and maximum relevance feature-selection (mRMR) method, which is based on mutual information calculations,[37] and the stepwise discriminant analysis (SDA) method. They use statistical methods to exclude redundant features. However, the results are not as expected, whereby the two feature selection methods did not result in obvious improvement. This may indicate that the distribution of the information for the prediction is not in a compact space, making the problem more difficult.

Previous studies have also indicated that a two-stage model performed well for predicting solvent accessibility in water-soluble proteins.[41,45] We also tried to use a two-stage SVR model rather than a simple SVR on the current problem to see whether we can improve the performance. However, in our work, a two-stage SVR model is found to only improve the mean absolute error (MAE) but decrease the CC index simultaneously. The result difference on membrane and water-soluble proteins may be due to smaller sample size in membrane proteins compared to that of the globular proteins.

3.2. Structure Segment Template-Based Prediction Engine. In the segment structural similarity prediction part, the

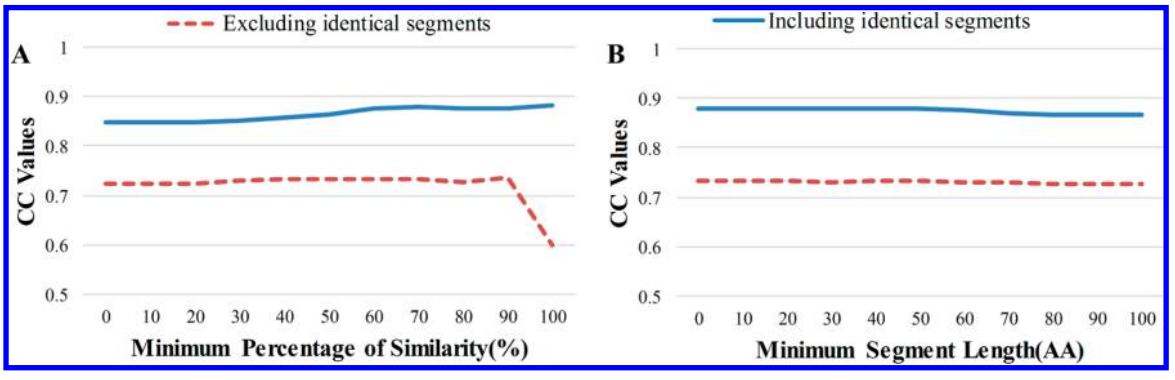


Figure 4. (A) CC values between the observed and predicted rASA with different minimum percentages of sequence identity. (B) CC values between the observed and predicted rASA with different homology segment lengths.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

key idea is to find the structure templates and use them to help predict the rASA of the query sequence. There are two important parameters during this homology-segment-based prediction process. The first is how to judge whether two sequences are close enough when using the BLAST tool to search the query sequence against the local rASA database; the other is how to remove noise in the alignment by selecting a proper segment length. In order to get the optimal combination of the above two parameters, we used a modified grid-search strategy. First, we set that a valid segment length needed to be at least 20 (aligned sequences whose lengths are shorter than 20 were excluded from further consideration); next, we determined which sequence similarity will generate the best result by searching from 0 to 100% in steps of 10%.

The results of MemBrain-Rasa at different sequence identity cutoff thresholds are shown in Figure 4A. The difference between the two sets of data in this figure is the inclusion of identical segments. Obviously, the results of including the identical segments are considerably better than those where identical segments are excluded. To get a fair and strict result, we mainly evaluate the performance when identical segments are excluded. It is clear that when we count only those searched segments that are of 100% sequence identity to the segments on the query sequence, the final results are solely from the SVR predictions, which are CC = 0.599 (Table 2). It is interesting to observe that the prediction results are very stable on tested thresholds. The reason could be that we have set a strict homology segment screening condition (the minimum length of aligned segments should be at least 20 amino acids long, a gap is not allowed, and the minimum E-value is 10^{-9}), which gives us relatively stable searched result sets on all of the thresholds. Finally, in the MemBrain-Rasa system, the minimum sequence identity for screening a homology segment is set to 70%.

For the second parameter of the homology segment length, we fixed the minimum sequence identity to be 70% and then searched the sequence length from 0 to 100 in steps of 10. We found that prediction performance changes very little when the sequence length is lower than 60 (Figure 4B). Finally, we decided that a sequence length of 20 would be used in the MemBrain-Rasa.

3.3. Consensus Model Helps to Solve the Preference Prediction Problem of SVR. The SVR algorithm is a typical statistical machine-learning-based predictor, which learns the regression parameters from a precollected data set. The final trained model is greatly affected by the data set distribution. Considering the general problem of incompleteness in the precollected training data set, the trained SVR model will not be a perfect predictor. Thus, the SVR-based predictor usually has some preferences to make a prediction in specific regions due to having learned more rules from them in the training data set. The incompleteness of a training data set is more serious in this article due to the relatively small number of solved membrane protein 3D structures.

Figure 5 illustrates the distribution of 10 subsets of the benchmark residue rASA value, with the pure SVR-based predictions and the consensus MemBrain-Rasa predictions also listed. As can be seen, the distribution of the benchmark residue rASA is not balanced, where 8,300 residues lie in the group of [0, 10], whereas only ~1,000 residues are located in the three groups of [70, 80], [80, 90], and [90, 100]. The effect of this imbalanced data set distribution on the trained SVR predictor is obvious, resulting in bad performance of the trained SVR on

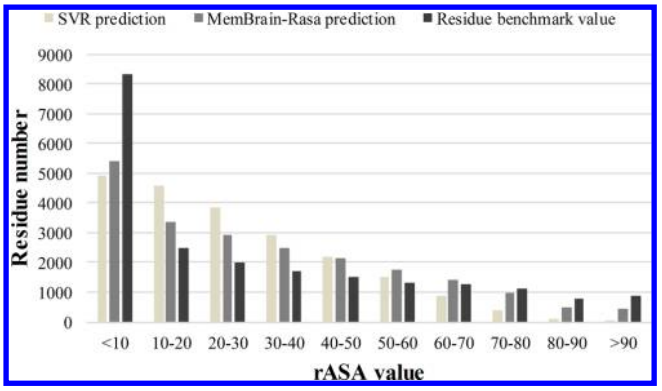


Figure 5. Distribution of SVR and MemBrain-Rasa predictions, as well as the residue benchmark rASA value.

those three groups. Especially in the groups [80, 90] and [90, 100], where almost no predictions are made.

In the proposed MemBrain-Rasa predictor, the combination of structure template-based engine with the machine learning prediction is capable of resolving the effects of imbalanced data sets on the machine learning predictors. As can be seen from Figure 5, the distribution of MemBrain-Rasa's outputs is much closer to the benchmark as compared to the pure SVR method. This significant improvement implies that we can now predict residues that lie in region of [70, 100] more accurately, especially in region [70, 80]. The reason for this improvement is that the homology segment-based prediction does not need the training process.

In order to further demonstrate the superiority of the proposed MemBrain-Rasa protocol over the pure SVR-based predictor, Figure 6 illustrates the CC and MAE distribution between the prediction and benchmark on 80 α-helical TMPs in the benchmark data set. We can see that when only using the

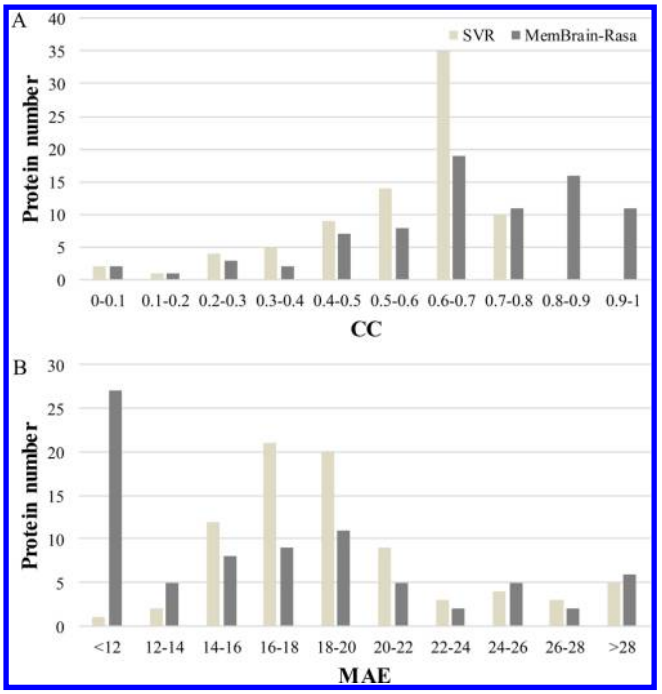


Figure 6. (A) Distributions of CC of rASA predictions for 80 sequences in the benchmark data set. (B) Distributions of MAE of rASA predictions for 80 sequences in the benchmark data set.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

SVR model, 59 (59/80 = 73.75%) chains have a predicted CC over 0.5. However, no chain has a CC greater than 0.8. These results suggest that the SVR-based predictor cannot generate very accurate predictions.

In the developed MemBrain-Rasa predictor, we combined the homology segment structural similarity-based predictor with the SVR-based predictor. Our results show that there are 49 proteins (with a total of 13,306 residues) that can find some homology segments in the local constructed rASA database out of all 80 proteins (21,418 residues). Among the 13,306 residues, 8,091 residues need to be modified from the original SVR predictions to the segment template-based predictions because the absolute difference between these two predictions is more than 5 (see the combination rule in Figure 2). The great difference between SVR's output and MemBrain-Rasa's is that there are 27 (27/80 = 33.8%) chains with CC values over 0.8 in MemBrain-Rasa but none in SVR. The MAE distribution in Figure 6B is very similar to the CC distribution. There are 27 chains with MAE < 12 in our MemBrain-Rasa method; however, there is only 1 in SVR. These results demonstrate that the two prediction subengines complement each other very well and that their combination significantly enhanced the prediction quality.

3.4. Performance Comparison of MemBrain-Rasa with Existing Methods. There are two existing methods that can be applied to predict residue rASA in α-helical TMPs: MPRAP[31] and ASAP.[28] Both of them are pure machine-learning-based predictors. The MPRAP was constructed based on the SVR algorithm, and the inputs to the SVR predictor include PSSM generated by PSI-BLAST, R4S predicted by Rate4Site, and Z-coordinate predicted by Zpred. Comparable to MPRAP, our SVR-based engine added 3 more features, i.e., SS, PP, and sequence length. Different from the MPRAP, where only one single prediction model was constructed for the whole chain, the ASAP trains two SVR models to predict residue solvent accessibility in membrane ($ASAP_{mem}$) and soluble ($ASAP_{glob}$) regions separately.

Table 3 compares MemBrain-Rasa with the above two methods on the same benchmark data set in the entire chain and in the transmembrane region. For a baseline test to see whether the predictor designed for the soluble proteins could also be applied to the membrane proteins, we included SABLE[27] for comparison, which uses neural-network-based regression to predict real-value rASA in soluble proteins. The inputs to SABLE include PSSM, SS predictions, and neighborhood information.

Table 3. Performance Comparison of rASA Prediction between Different Methods

method	MAE	CC
Comparison between Different Methods in the Whole Region		
SABLE	21.9	0.40
$ASAP_{glob}$	21.6	0.41
MPRAP	18.4	0.58
MemBrain-Rasa-SVR	18.013	0.599
MemBrain-Rasa	13.593	0.733
Comparison between Different Methods in the Transmembrane Region		
$ASAP_{mem}$	24.3	0.18
MPRAP	–	–
MemBrain-Rasa-SVR	16.104	0.575
MemBrain-Rasa	13.336	0.683

It is clear that our method outperforms other methods by achieving the highest CC and the lowest MAE. If we only consider the SVR subengine in the MemBrain-Rasa (MemBrain-Rasa-SVR), it is very comparable to the MPRAP. However, when we combined the SVR prediction with the new structure segment template-based method, the MAE and CC greatly improve to 13.593 and 0.733 respectively. In addition, because the SABLE and $ASAP_{glob}$ are both rASA predictors for soluble proteins, they perform comparably on membrane proteins and achieve MAE and CC of approximately 21 and 0.40, respectively. This result indicates that because the membrane proteins have very distinct characteristics from soluble proteins, tools designed for soluble proteins will not work well for membrane proteins.

The results in the transmembrane region also demonstrate that MemBrain-Rasa is the best. For this test, the transmembrane region is defined by membrane definitions from the OPM database.[46] On the benchmark data set, the $ASAP_{mem}$, which was developed for predicting rASA of transmembrane residues, achieves a CC of 0.18 with MAE = 24.3. The poor results for $ASAP_{mem}$ on the current benchmark data set are probably due to the fact that (1) it was trained on a very small data set, e.g., 28 proteins in which transmembrane residues were manually annotated, leading to limited generalization ability, and that (2) it used a simple feature set of PSI-BLAST profile, which did not include some additional useful features, such as the Z-score.

When only considering the SVR engine, MemBrain-Rasa-SVR achieves MAE and CC of 16.104 and 0.575, respectively. In the final system of MemBrain-Rasa, the CC in the membrane region is 0.683, which is lower than 0.733 on the whole chain; whereas, the MAE in the transmembrane region is 13.336, which is comparable to its value (13.593) on the whole chain. This result indicates that the residues in the transmembrane region and the loop region are also different.

3.5. Case Studies. To understand the normal procedure of predicting a rASA value, we present a representative example to show the performance between SVR and MemBrain-Rasa methods. This independent test example is a membrane protein named 2bs2:C[47] (PDB ID, 2bs2; chain, C) with 254 residues and 5 transmembrane helices.

In the first step, rASA values of protein chain 2bs2:C are predicted using SVR models with a CC of 0.63 and a MAE of 18.39, as shown in Figure 7A. After we added structural similarity predictions with the SVR predictions, we can achieve a CC of 0.95 and a MAE of 4.93 between the predicted and observed rASA values. The results are shown in Figure 7B. The benchmark proteins are listed in the Supporting Information.

4. DISCUSSION

Most, if not all, existing membrane protein residue real-value rASA predictors were constructed by only using statistical machine-learning methods, such as SVR, neural networks (NN), and others. These machine-learning algorithms have transformed rASA prediction to a many-to-one mapping problem, where the input is a high-dimensional feature vector, and the output is a 1-D real value. Although the idea is simple, the statistical learning algorithm has the following difficulties on this topic: (1) performance significantly depends on the training set size and quality; (2) performance is also affected by the different features that encode the residues; (3) machine-learning-based predictors seem to have an accuracy limit for this study. For instance, two different feature-selection methods

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

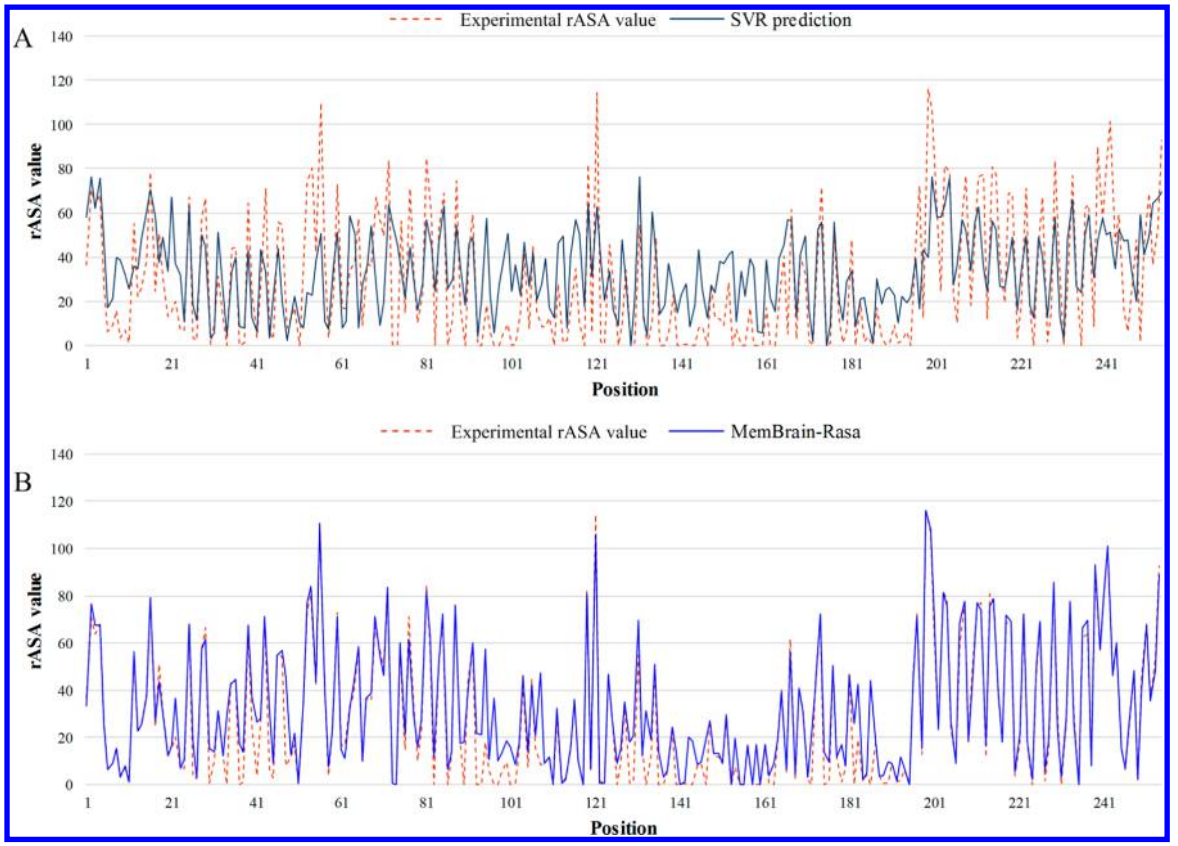


Figure 7. (A) rASA predictions of 2bs2:C using the SVR model (MAE = 18.39, CC = 0.63) (B) rASA predictions of 2bs2:C using the combination of SVR models and structure similarity-based methods (MAE = 4.93, CC = 0.95).

and a two-stage SVR method did not result in obvious improvement. The reason for a potential accuracy limit is that the "many-to-one" mapping in this study can be nonlinear, making the mapping rules difficult to learn.

Because of the rapid progress in structural biology and concomitant increase in available 3D structures in the PDB, our results demonstrate that the proposed segment template-based rASA predictions can help to enhance the machine-learning-based predictions greatly by solving the output preference problem in the machine learning engine. During our experiments, the proposed MemBrain-Rasa consensus predictor is able to improve the rASA value predictions to CC = 0.733 and MAE = 13.593, which is the best performance on membrane proteins ever reported to the best of our knowledge.

Compared to globular proteins, predicting residue rASA values for membrane proteins is more challenging. The first reason is that, in nature, there are both hydrophobic and hydrophilic environments in membrane proteins. For the transmembrane residues, they typically have hydrophobic preference, and our previous study has shown that TMH-TMH contact mode is also different from residue contacts in the loop region.[14] This could be a reason for the existing ASAP predictor to develop two different subpredictors for residues in the membrane and loop regions separately.

The other difficulty in developing a membrane protein-specific residue rASA predictor is that the number of experimentally solved 3D structures is small, resulting in a relatively small sample-size problem in the machine-learning algorithm. Although our proposed homology segment-based engine can partly solve this problem, we did find some cases that are also limited by the small number of available membrane protein 3D structures. Even in the final MemBrain-Rasa predictor, there are three chains with their predicted CC values lower than 0.2 (Figure 6). The PDB IDs of these three protein chains are 1zza:A, 2pil:A, and 2axt:H. We then looked closely at these three cases to see what happens.

The tested protein of 1zza:A has the lowest CC value and the maximum MAE (44.5) in all of our tested proteins. 1zza:A is a chain of solution NMR structures of the membrane protein stannin with one α-helical transmembrane subunit (N-terminus intermembrane space). When we used only the SVR engine, the prediction of this protein is also the worst. When we tried to search the homology segments for this protein, we found that there are no homology segments that can be found in the local rASA database. This probably indicates that the 1zza:A structure has a completely different folding mode from the available structures, resulting in poor performance by both the SVR model and the homology segments-based engine.

The second hard target is the protein of 2pil:A with a predicted CC of 0.043 and a MAE of 34.15. 2pil:A is a chain of crystallographic structures of phosphorylated pilin from *Neisseria*. When searching its homologue segments against the local rASA database, except for three identical segments, there were only 12 short segments (length <20) of sequence identities lower than 70%. According to our above protocol, these identical and short segments will not be used in the homology segment-based prediction, meaning that the final prediction for 2pil:A is also solely dependent on the SVR-predictor. The third hard case is 2axt:H with a predicted CC of 0.188 and MAE of 27.46, which is a crystal structure of the photosystem from *Thermosynechococcus elongates*. We also did not find any applicable homology segments for this protein.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474

A common thing for the above three hard targets is that they did not find any segment templates, and the homology segment structural similarity-based engine cannot be applied. It is expected that with the further expansion of experimentally solved protein 3D structures and the regular update of the proposed MemBrain-Rasa protocol, we can make accurate predictions for these hard targets. Furthermore, the update of segment structure similarity-based engine in the MemBrain-Rasa is very convenient since we only need to regularly add the newly solved PDB structures to update the rASA table (Figure 1).

In the future, we will also try to answer the question whether improvement of the residue solvent accessibility prediction will improve the ab-inito α-helical TMPs 3-D structure predictions. However, in soluble protein 3D structure predictions predicted solvent accessibility has been widely used, e.g., QUARK[48] and I-TASSER.[18] To the best of our knowledge, there have been no predictors that add rASA predictions directly into the energy function to predict 3D structures of α-helical TMPs to date. The first version of FILM used a rough solvent energy prediction which was related to the coarse prediction of the Z-coordinate. A decade later, FILM3[16] used the residue contact predictions to predict 3D structures of large membrane proteins. BCL::MP-Fold[17] used Bayes' theorem to score MP-specific environments instead of a real residue rASA value. One potential reason for not incorporating the predicted residue rASA into the membrane protein 3D structure prediction is that the predicted accuracy is too low to be effective. The work of this article represents an accurate residue rASA prediction for membrane proteins, which is expected to play an important role in generating accurate 3D structure models of α-helical TMPs. It is also expected that with the enhancement of ab-initio predicted 3D structures of α-helical TMPs by incorporating the residue rASA values, the whole topological structure knowledge of the target membrane proteins (such as the TMH spanning segments and their interactions) can also be improved.

In this study, we mainly focus on the TMH proteins. Compared to the large number of TMH proteins, the size of TMB proteins is rather small (2−3% in the whole genome).[49] Because of the smaller number of solved TMB protein structures, which may cause the so-called "small-sample" problem in statistics and result in low generalization ability of the prediction model, we do not include TMB proteins in our current predictor. Along with more TMB proteins' 3D structures being solved, we will update our MemBrain-Rasa for this type of transmembrane proteins in our future work.

5. CONCLUSIONS

We have developed a novel method (MemBrain-Rasa) for the sequence-based real-value prediction of residue rASA values in α-helical membrane proteins. MemBrain-Rasa features by a combination of a machine-learning-based engine and a structure segment template-based engine. These two parts complement each other very well. Our experimental results demonstrate that the structure-similarity based prediction helps to solve the serious preference problem in machine learning engine outputs, resulting in a significant improvement of the final consensus system. MemBrain-Rasa represents a new potential for structure modeling of transmembrane proteins by using the predicted residue rASA. MemBrain-Rasa is available at: http://www.csbio.sjtu.edu.cn/bioinf/MemBrain/.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00246.

Statistics of representative amino acid physical parameters used for feature extraction in MemBrain-Rasa (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +86-21-34205320. Fax: +86-21-34204022. E-mail: hbshen@sjtu.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to Sara Walker for reading through the manuscript. This work was supported by the National Natural Science Foundation of China (No. 61222306, 91130033, 61175024), Shanghai Science and Technology Commission (No. 11JC1404800), and a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (No. 201048).

■ REFERENCES

(1) Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **1998**, *7*, 1029−1038.

(2) Chou, K. C.; Cai, Y. D. Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Model.* **2005**, *45*, 407−413.

(3) Pieper, U.; Schlessinger, A.; Kloppmann, E.; Chang, G. A.; Chou, J. J.; Dumont, M. E.; Fox, B. G.; Fromme, P.; Hendrickson, W. A.; Malkowski, M. G.; Rees, D. C.; Stokes, D. L.; Stowell, M. H. B.; Wiener, M. C.; Rost, B.; Stroud, R. M.; Stevens, R. C.; Sali, A. Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nat. Struct. Mol. Biol.* **2013**, *20*, 135−138.

(4) Almén, M. S.; Nordström, K. J.; Fredriksson, R.; Schiöth, H. B. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.* **2009**, *7*, 50.

(5) Kneissl, B.; Mueller, S. C.; Tautermann, C. S.; Hildebrandt, A. String Kernels and High-Quality Data Set for Improved Prediction of Kinked Helices in alpha-Helical Membrane Proteins. *J. Chem. Inf. Model.* **2011**, *51*, 3017−3025.

(6) Langelaan, D. N.; Wieczorek, M.; Blouin, C.; Rainey, J. K. Improved Helix and Kink Characterization in Membrane Proteins Allows Evaluation of Kink Sequence Predictors. *J. Chem. Inf. Model.* **2010**, *50*, 2213−2220.

(7) Shen, H. B.; Chou, K. C. Using ensemble classifier to identify membrane protein types. *Amino Acids* **2007**, *32*, 483−488.

(8) Shen, H. B.; Yang, J.; Chou, K. C. Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. *J. Theor. Biol.* **2006**, *240*, 9−13.

(9) Shen, H.; Chou, J. J. MemBrain: improving the accuracy of predicting transmembrane helices. *PLoS One* **2008**, *3*, e2399.

(10) Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007**, *23*, 538−544.

(11) Nugent, T.; Jones, D. T. Transmembrane protein topology prediction using support vector machines. *BMC Bioinf.* **2009**, *10*, 159.

(12) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567−580.

(13) Rath, E. M.; Tessier, D.; Campbell, A. A.; Lee, H. C.; Werner, T.; Salam, N. K.; Lee, L. K.; Church, W. B. A benchmark server using high resolution protein structure data, and benchmark results for membrane helix predictions. *BMC Bioinf.* **2013**, *14*, 111.
(14) Yang, J.; Jang, R.; Zhang, Y.; Shen, H. B. High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics* **2013**, *29*, 2579−2587.
(15) Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1010−1025.
(16) Nugent, T.; Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1540−E1547.
(17) Weiner, B. E.; Woetzel, N.; Karakas, M.; Alexander, N.; Meiler, J. BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices. *Structure* **2013**, *21*, 1107−1117.
(18) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725−738.
(19) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209−225.
(20) Zhang, Y.; Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7594−7599.
(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(22) Monastyrskyy, B.; D'Andrea, D.; Fidelis, K.; Tramontano, A.; Kryshtafovych, A. Evaluation of residue-residue contact prediction in CASP10. *Proteins: Struct., Funct., Genet.* **2014**, *82* (Suppl 2), 138−153.
(23) Ahmad, S.; Gromiha, M. M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 629−635.
(24) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(25) Magnan, C. N.; Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014**, *30*, 2592−2597.
(26) Faraggi, E.; Zhang, T.; Yang, Y.; Kurgan, L.; Zhou, Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.* **2012**, *33*, 259−267.
(27) Adamczak, R.; Porollo, A.; Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 753−767.
(28) Yuan, Z.; Zhang, F.; Davis, M. J.; Boden, M.; Teasdale, R. D. Predicting the solvent accessibility of transmembrane residues from protein sequence. *J. Proteome Res.* **2006**, *5*, 1063−1070.
(29) Park, Y.; Hayat, S.; Helms, V. Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinf.* **2007**, *8*, 302.
(30) Lai, J. S.; Cheng, C. W.; Lo, A.; Sung, T. Y.; Hsu, W. L. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. *BMC Bioinf.* **2013**, *14*, 304.
(31) Illergard, K.; Callegari, S.; Elofsson, A. MPRAP: an accessibility predictor for a-helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinf.* **2010**, *11*, 333.
(32) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.
(33) Hubbard, S. J.; Thornton, J. M. *NACCESS*; Department of Biochemistry and Molecular Biology, University College London: London, U.K., 1993.
(34) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(35) Hubbard, S. J.; Campbell, S. F.; Thornton, J. M. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **1991**, *220*, 507−530.
(36) Kauko, A.; Illergard, K.; Elofsson, A. Coils in the membrane core are conserved and functionally important. *J. Mol. Biol.* **2008**, *380*, 170−180.
(37) Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226−1238.
(38) Xu, Y. Y.; Yang, F.; Zhang, Y.; Shen, H. B. An image-based multi-label human protein subcellular localization predictor (*i*Locator) reveals protein mislocalizations in cancer tissues. *Bioinformatics* **2013**, *29*, 2032−2040.
(39) Granseth, E.; Viklund, H.; Elofsson, A. ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. *Bioinformatics* **2006**, *22*, e191−e196.
(40) Wu, S.; Zhang, Y. ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS One* **2008**, *3*, e3400.
(41) Song, J.; Tan, H.; Wang, M.; Webb, G. I.; Akutsu, T. TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. *PLoS One* **2012**, *7*, e30361.
(42) Cherkassky, V. The nature of statistical learning theory~. *IEEE Trans. Neural Netw.* **1997**, *8*, 1564.
(43) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658−1659.
(44) Zhu, L.; Yang, J.; Song, J. N.; Chou, K. C.; Shen, H. B. Improving the accuracy of predicting disulfide connectivity by feature selection. *J. Comput. Chem.* **2010**, *31*, 1478−1485.
(45) Nguyen, M. N.; Rajapakse, J. C. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins: Struct., Funct., Genet.* **2006**, *63*, 542−550.
(46) Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **2006**, *22*, 623−625.
(47) Madej, M. G.; Nasiri, H. R.; Hilgendorff, N. S.; Schwalbe, H.; Lancaster, C. R. Evidence for transmembrane proton transfer in a dihaem-containing membrane protein complex. *EMBO J.* **2006**, *25*, 4963−4970.
(48) Xu, D.; Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 1715−1735.
(49) Gromiha, M. M.; Yabuki, Y.; Suwa, M. TMB finding pipeline: Novel approach for detecting ss-barrel membrane proteins in genomic sequences. *J. Chem. Inf. Model.* **2007**, *47*, 2456−2461.

DOI: 10.1021/acs.jcim.5b00246
J. Chem. Inf. Model. 2015, 55, 2464−2474