MPLs-Pred: Predicting Membrane Protein-Ligand Binding Sites Using Hybrid Sequence-based Features and ligand-specific models

Chang Lu[1,2], Zhe Liu[1,2], Enju Zhang[1,2], Fei He[1,2,*], Zhiqiang Ma[1,2,*] and Han Wang[1,2,*]

1 School of Information Science and Technology, Northeast Normal University, Changchun 130117, China.

2 Institute of Computational Biology, Northeast Normal University, Changchun,130117, China.

**Corresponding authors: Han Wang - wangh101@nenu.edu.cn;*

Abstract:

Membrane proteins (MPs) involve in many essential biomolecule mechanisms as a pivotal factor to enable the small molecular and signal transport between the two sides of the biological membrane, which is the reason that a large portion of modern medicinal drugs targets to MPs. Therefore, accurately identifying the Membrane Protein-Ligand Binding Sites (MPLs) will significantly improve drug discovering. In this study, we propose a sequence-based MPLs predictor called MPLs-Pred, where evolutionary profiles, topology structure, physicochemical properties, and primary sequence segment descriptors are combined as features apply to a random forest classifier, and an under-sampling scheme is used to enhance the classification capability with imbalance samples. As well, additional ligand-specific models are taken into consider refining the prediction. Corresponding experimental results presented our method achieved an appreciable performance, with 0.63 MCC (Mathew Correlation Coefficient) of the overall prediction precision, and that values are 0.604,0.7 and 0.692 respectively for the three main types of ligands: drugs, metal ions, and biomacromolecules. MPLs-Pred is freely accessible in http://icdtools.nenu.edu.cn/

Keywords: Membrane Protein, Binding Site Prediction, Protein-Ligand, Ligand-Specific Model

INTRODUCTION

Membrane protein (MP) is an important type of protein along with soluble globular protein, fibrous protein and disordered protein[1]. MPs involved in various crucial biological functions[2], such as

transportation through membranes, immune system molecule recognition, hormone reception, etc. That is the reason they have a strong potential to be the target of new drugs in the future since over half of modern therapeutic drugs target MPs[3]. Therefore, exploring the functions of membrane proteins, especially their binding capability, remains of profound importance on various fronts, not least of which includes drug discovery[4]. In Current, commercial drug targets discovering mainly rely on traditional approaches as high throughput screening or functional assays, etc. However, more efficient and economical approaches are required to detailly identify the MP-ligand binding sites(MPLs) aiming the modern medicine. Obviously, intelligent computation is one promising approach for the purpose.

During the past decades, many considerable efforts have been applied to predicting soluble protein-ligand interactions, classified into three major categories: structure-based, sequence-based and hybrid methods, which use both sequence and structure characteristics. Structure-based methods follow the assumption that ligands always interact with those proteins with similar structural properties in global or local. As a representative, identifying ligand binding pockets through the experimental protein structures was widely researched[5-8]. Regarding there are fewer protein structures available for extensive requirements, sequence-based methods raised to directly predict the residues who probably interact with particular ligand[9-13]. But the performance of these methods highly depends on whether the sequence-derived features could maximumly impact the spatial structural properties of the protein. Whereafter, Considerable attention has been paid to the hybrid method that combines structural and sequential information[14-17]. Previous studies demonstrated that hybrid methods are often superior to others because they inherit the advantages from both structure- and sequence-based methods.

Suresh et al[18] published the first sequence-based MPL predictor Tm-lig for membrane proteins, except this, few peer work was reported, while plentiful attention had been paid to their soluble partners. It is obvious that the challenges remain in the membrane protein MPL prediction, but opportunities are also existing currently.

The structure-based approach will not be applied to membrane proteins directly since their 3D structures required in homology comparative prediction are not abundant compared to soluble proteins, caused by the native environment[19-21], where exists the hydrophobic thickness of the lipid bilayer they

inserted to[22]. Consequently, sequence-based features are much more accessible than structure-based features for membrane proteins in many realistic scenarios. Thus, the sequence-based approach is preferred here rather than the structure-based prediction.

As many typical computational biology research, MP-ligand binding prediction will also exist the imbalanced learning problem, where the number of majority samples is significantly larger than that of minority samples. According to statistics, the number of the non-binding residues is about 150-200 times of that of the binding residues. Numerous studies have shown that using the traditional classifier algorithm directly to imbalanced problems often tend to bias to the larger classes[23-25]. Therefore, data imbalance phenomenon is an inevitable problem to be solved.

Most of the soluble protein-ligand binding sites prediction methods, as well as the only MPLs predict method, focused on predicting universal ligand binging site without considering the differences among various ligands. In fact, the significant distinction exists among the different types of ligands in their size, structure, function or other characteristics, and different types of ligands tend to attach with particular residues referenced the surrounding environment. many ligand-specific binding sites predictors have been developed recently[26-28], that are often superior to universal-purpose binding site predictors. Considering this, in addition to the universal ligand binding residue prediction model, we further build ligand-specific models to predict drug-like compound-binding, metal-binding, and biomacromolecule-binding residues respectively.

After all, we analyze the characteristics of MPLs and the contribution of different type of features in detail. The PSSM features are the most effective ones. At the same time, 4 features could be more sufficient with highest MCC value. From the biological explanation, PSSM, Topology, Evolutionary Profiles (PSSM), Topology Structure (TOPO), Physicochemical Properties (PCP), and Primary Sequence Segment Descriptors (SeqSeg) present the evolutionary information, the fundamental structure, the microenvironment, and the original sequence composition of the liagnd-binding sites, respectively.

In this study, we developed a membrane protein-ligand binding site predictor MPLs-Pred. The limitation of structure-based method motivates us to extract discriminative features of the MPLs from sequence information alone, i.e. Evolutionary profiles, topology structure, physicochemical properties,

and primary sequence segment descriptors. To tackle the serious impact of the data imbalance phenomenon, a Random Under-Sampling scheme is applied before using the Random Forest Classifier to predict MPLs. The universal MPLs-Pred derived 0.63 MCC value on the independent validation. Considering the distinction among the different type of ligand binding residues, we divided the ligand binding residues into 3 categories, including drug-like compound-binding, metal-binding, and biomacromolecule-binding residues. We build ligand-specific models to further improve the prediction performance and achieves considerable progress with 0.604, 0.7 and 0.692 MCC value on the drug-like compound, metal, and biomacromolecule ligand-specific predictors on independent validation respectively. MPLs-Pred is accessible freely in http://icdtools.nenu.edu.cn/

MATERIALS AND METHODS

Benchmark Datasets

A data set of membrane protein-ligand binding sites were extracted from the Protein Data Bank[29] by Suresh et al[18] in 2015. This data set contains 42 non-redundant protein sequences with 10657 residues and 1431 of them are identified as ligand binding residues. Considering the dramatically increasing number of membrane proteins in recent years, we construct a new benchmark data set. First of all, We analysis the annotations of all 102,429 manually annotated and reviewed membrane proteins from The Universal Protein Resource Databank (UniProt) released to date. Then, after removing the protein sequences less than 50 residues length and those with unknown residues such as X, we have 17, 590 sequences with exact ligand binding residues left. To reduce the influence of data redundancy and homology bias[30], these proteins were clustered by CD-HIT with 30% sequence identity cut-off and the representative sequence in each cluster is picked. After that, we have 2734 proteins with 10979 binding residues left. To evaluate the effectiveness of the proposed method, these proteins were divided randomly into training dataset with 2500 proteins, and independent validation dataset with 234 proteins.

The protein binding ligand could be roughly divided into major categories: drug-like compounds, metal ions, and biomacromolecules, among which biomacromolecules include proteins, fats, sugar,

nucleotides, and so on. Thus, we divided the dataset into 3 parts according to the type of ligand. Other types of ligands are ignored because the sample size is too small to be statistically significant. The details of the universal dataset and the ligand-specific datasets are illustrated in Table 1.

Table 1. The detailed composition of new built standard datasets.

Dataset	Training dataset			Testing dataset		
	Num of Proteins	Num of Residues[1]	Ratio[2]	Num of Protein	Num of Residues[1]	Ratio[2]
Universal	2500[3]	(10143, 1524372)	1:150	234	(836, 164792)	1:197
Drug-like compound	655	(1839, 386979)	1:210	45	(121,25193)	1:208
Metal	1375	(5734, 804610)	1:140	117	(503, 85437)	1:170
Biomacromolecule	857	(2505, 435298)	1:174	67	(161, 35022)	1:218

1. Figures in 2-tuple Num of Residues represent the number of positive samples (binding residues) and negative samples (non-binding residues);
2. Figure in Ratio represents the ratio of positive samples to negative samples;
3. The MP`s data set have overlap because some proteins interact with two or more type of ligands.

Selected Features

To distinguish ligand-binding and non-binding residues, we choose four kinds of features which can describe the characteristic of ligand binding residues in membrane proteins, who always has a specific residue composition, evolutionary conservation, physicochemical environment or topology characteristic. In this study, we employ the sliding window scheme to express the influence of the neighboring residue. Here we set the value of window size to 7 residues, in which 3 from upstream and 3 from downstream, after testing the different value of it.

A. Evolutionary Profiles (PSSM)

It has been proved that high conversed regions are always involved in basic cellular function, Research on the membrane[18] and non-membrane[11] protein-ligand interact residues indicates that the evolutionary information is useful. Position-Specific Scoring Matrix (PSSM) has been demonstrated to be an effective feature to encode the evolutionary information of protein sequence. It was widely used in many bioinformatics problems such as protein function prediction[31], protein-protein interaction sites prediction[32], protein secondary structure prediction[33], DNA-binding proteins prediction[34, 35], etc.

For a protein sequence with L residues, we obtain its PSSM by using the PSI-BLAST[36] to search the non-redundant database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.tar.gz) through 3 iterations and 0.001

E-value cutoff. An L×20 matrix was generated for each protein, Then, a sliding window is used to each residue to build its PSSM feature vector. Since the window size is 3, the dimensionality of the PSSM feature is $7 \times 20 = 140$ for each residue.

B. Topology Structure (TOPO)

Membrane protein spans or partly spans the lipid bilayer of the membrane, which is the major difference between MP and non-MP. Thus, the fundamental aspect of the structure of the transmembrane protein is the membrane topology, that is, the number of transmembrane segments, their position in the protein sequence and their orientation in the membrane[37, 38]. Many researchers pay attention to predicting the topology of the transmembrane protein and the newest predictor can achieve an accuracy value close to 80%.

In this work, the topology of transmembrane is predicted by TOPCONS[39], a reliable predictor marks each residue in the sequence as I, O, M or U, represent the residue located on the inside, outside or membrane region, or non-membrane region but location unknown. respectively. We further digitalize the topology descriptor with a 4-dimension vector. Each element in the vector represents the count of the corresponding topology in the sequence window. Hence, we get topology features with 4-dimension.

C. Physicochemical Properties (PCP)

Since residues are fundamental building blocks of protein, their physicochemical properties (PCP) influence the microenvironment of proteins, including energy, surface motions, dynamics and so on[40]. Previous studies have shown that PCP can be used in many predict methods such as Enzymatic proteins identification[41], protein lysine acetylation prediction[42], etc. Previous studies show that physicochemical properties play important roles in the success of soluble protein-ligand binding sites prediction. Since MPLs are always appear in the water-soluble regions in membrane proteins, PCP could also be used in the filed of MPLs prediction.

In this study, 15 PCPs are collected from AAindex[43] based on research experience, named Hydrophilicity value, Hydrophobicity, Net charge, Polarity, Size, Residue volume, Molecular weight, Diameter, Amino acid composition, Composition of amino acids in membrane proteins, Side chain interaction parameter, Solvation free energy, Transfer free energy, Average flexibility indices,

Accessible surface area. A further experiment shows 15 PCPs can significantly improve the performance of the predictor. Thus a $15 \times 7 = 105$ -D vector is formed to represent the physicochemical properties of each residue.

D. Primary Sequence Segment Descriptors (SeqSeg)

We believe that the original sequence is very important that can directly decide the structure and function of the protein, peptide with particular functions always shows special arrangement. According to the study of the composition difference between ligand-binding and non-ligand-binding residue, we build a 20-dimension vector to figuring the primary sequence segment around the target residue. The value of the element in the vector represents the number of the corresponding residue in the protein segment sliced by the window. Thus, we get a 20-D vector to represent the feature of primary sequence segment around the target residue.

Finally, the feature space of target residue is a $140 + 4 + 105 + 20 = 269$-dimension vector which contains four different kinds of sequence-derived features. The analysis of these features will be further described in the section "The Contribution of Features".

Random Under-Sampling (RUS)

The statistics in table 1 illustrate that MPLs prediction problem is a typical imbalanced learning problem, where the number of majority samples (non-binding residues) is significantly larger than that of minority samples (binding residues). The data imbalance phenomenon on MPLs prediction problem is really serious: the number of the non-binding residues is about 140-200 times of that of the binding residues. Numerous studies have shown that using the traditional classifier algorithm directly to imbalanced problems often tend to bias to the larger classes[23, 24, 44, 45]. Therefore, data imbalance phenomenon is an inevitable problem to be solved. In this study, a Random Under-Sampling (RUS) scheme is applied before training the predictor to reduce the negative influence of imbalanced data. Since the limit number of binding residues, we contain all binding residues and randomly select 30 times of non-binding residues. The impact of RUS scheme on the predictor will be further discussed in the section "Random Under-Sampling Scheme Influence the Predict Performance".

Random Forest Classifier

Random forests (RF) is a powerful classifier firstly proposed by Leo Breiman in 2001[46]. It is widely used for classification, regression, feature selection and other tasks in the field of bioinformatics[47, 48], such as prediction of protein-protein interaction sites[32], identification of membrane protein types[49], prediction of GPCR-drug interactions[50, 51], and so on. RF is an ensemble of decision trees that add an additional layer to bagging them together. Decision trees in RF are trained by a subset randomly selected from the primary feature set on data which is also randomly undersampling of the training dataset. After getting the forecast results, these decision trees vote on the class for the given input sample. In this study, the predictor shows best performance when the CART split scheme is used with the number of trees is 140 and the dimension of the candidate at each split is 20. Because the features and training samples of each decision tree are randomly selected, Random Forests can handle imbalance samples with high-dimension well without sampling or feature selection process.

RESULT AND DISCUSSIONS

Characteristics of Ligand Binding Residues

In this work, the ligand-binding residue is coded by the feature of evolutionary profiles, topology, physicochemical properties, and primary sequence segment descriptors. Before employee these characteristics as the feature spaces of the predictor, we demonstrate the effectiveness of them by the statistical and experimental method.

Figure 1 shows the relative composition of ligand binding residues based on the universal training dataset and ligand-specific binding residues on the corresponding dataset. The relative composition could reflect the enrichment and distribution of binding residues. It is observed that polar hydrophilic residues such as cysteine (C), histidine (H) and aspartic acid (D) are more likely to be binding sites for all kinds of ligand, especially for metal binding residues. Biomacromolecules prefer to interact with alkaline residues, such as histidine (H), arginine (R) and lysine (K), as well as the polar hydrophilic

residues. This phenomenon also exists in drug binding residues but is not as significant as others. The enrichment of metal and biomacromolecule binding residues is more obvious than that of drug binding residues. It is probably because the drugs are manmade chemical compounds, not natural ones. The result of long-term evolution makes the proteins to form a special region to recognize specific natural ligand and interact with it to perform specific functions. The phenomenon of the low specificity of drug binding residues leads to the interaction of chemical drugs and membrane proteins are not a perfectly one-to-one correspondence. That`s why drugs always have side effects. It is obvious that different type of ligand binding residues show a different preference for residue enrichment. Thus, the introduction of the ligand-specific predict strategy is explicable.

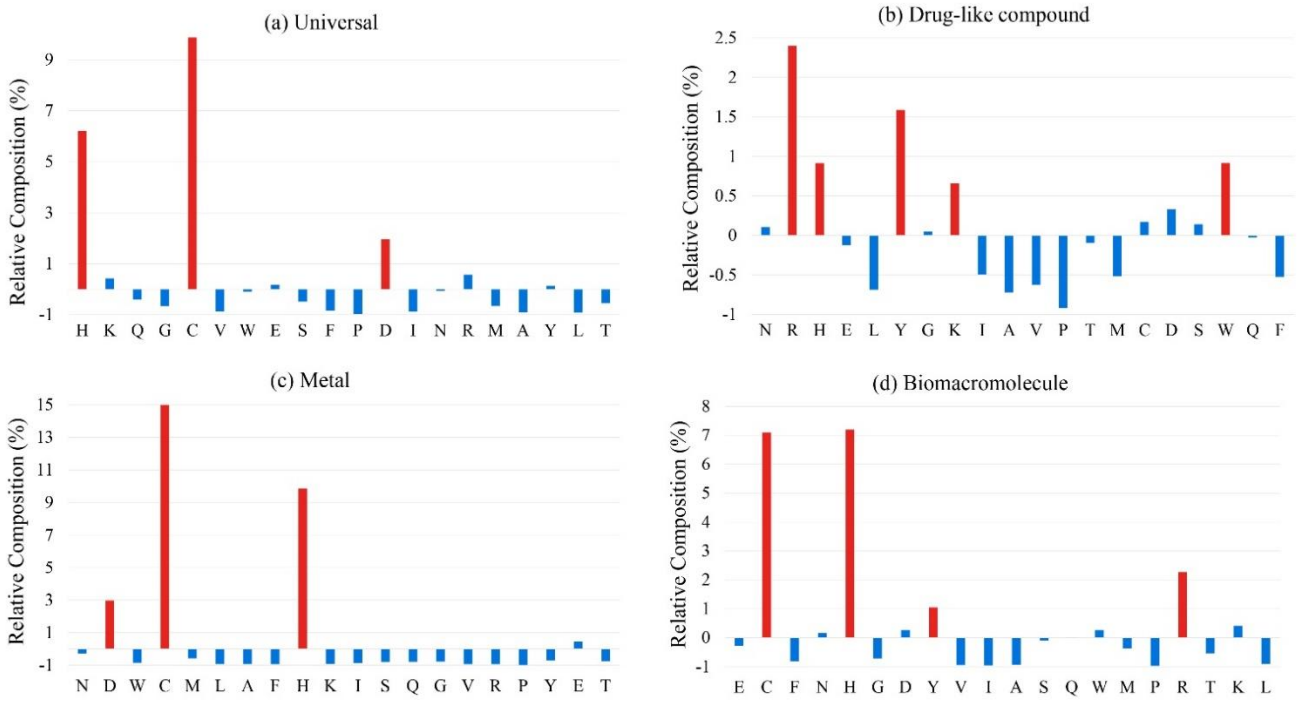

Figure 1. The relative composition of (a). universal ligand-binding residues, (b). drug-like compound binding residues, (c). metal binding residues and (d) biomacromolecule binding residues based on background distribution of all residues in corresponding datasets.

Furthermore, the residue preferences of the neighboring environment of target residues are investigated. Two-sample logos maps about universal ligand binding residue against corresponding non-binding residue are shown in Figure 2. According to the illustration, we can see that the enrichment phenomenon of neighboring residues of the target is not remarkable, it is probably a reflection of the

contribution limits of sequence neighbor residue during the interacting process. Thus, over introduce the information of neighbor residue may cause noisy. This phenomenon is different from the case of soluble proteins, the deeper reason is that we need to be further explored.

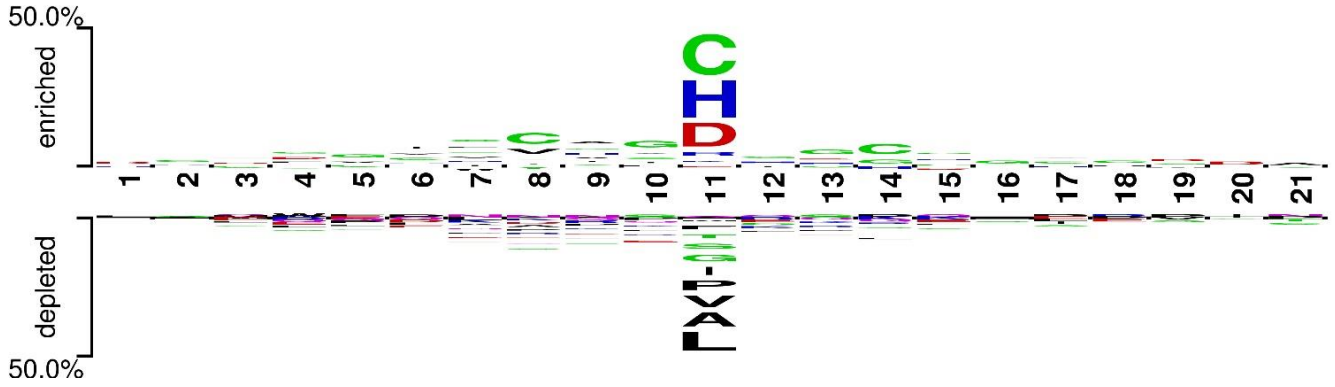

Figure 2. Two-sample logos of Universal ligand binding residues against non-binding residues.

We further investigate the topology distribution of binding residues on the training dataset. According to the predicted result of HMMTOP, among 10143 binding residues, 7978 residues (about 79%) located on the outer side of the membrane, 1788 residues (about 18%) located on the inner side of the membrane, only 377 residues (less than 4%) binding residues are located on the transmembrane region. The reason for this phenomenon is the special function of membrane proteins: transmitted ligand, signal, and energy inside and outside of the cell. The residues located between phospholipids are always stable to keep the channel structure of membrane proteins.

The Contribution of Features

As described in section “Selected Features”, we employee 4 kinds of features to construct the feature spaces of the predictor. To evaluate these features, we calculate the Pearson Correlation Coefficient between features and label on the universal and the ligand-specific training dataset. As shown by the heat map (Figure 3), the features of PSSM reveal the highest negative correlation with the label. The features of PCP also show significant correlations on all datasets. The features of TOPO and SeqSeg show a lower correlation with labels might because the feature vector is too sparse. Point-by-point comparing among universal and ligand-specific dataset, the linear correlation between features and labels is relatively lower on drug-like compound data and significantly higher on metal data. This phenomenon is consistent with the experimental results.

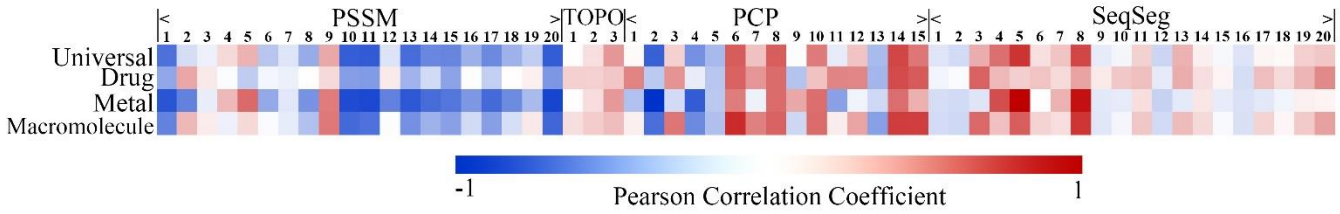

Figure 3. The heat map of the Pearson Correlation Coefficient between features and labels. Red represents the positive correlation and Blue represents the negative correlation. The deeper the color, the higher the correlation

We further analyze the predict contribution of different kinds of features. As shown in Table 2. The features of PSSM, which present the evolutionary information of protein sequence, are the most effective ones. The sequence conservation is obvious in the binding regions, but not all the conserved sequenced segments contain binding sites. Thus, this feature would be a necessary condition for the prediction (see from high ACC value), but not a sufficient condition (see from low Sen value). At the same time, 4 features could be more sufficient and balanced to exhibit the existence of binding sites, with similar high ACC, and the highest Sen, lead to the highest MCC. From the biological explanation, topology can present the fundamental aspect of the structure of the transmembrane protein; PCP can present the microenvironment of binding regions; SeqSeg can present the original sequence composition surrounding the binding sites.

Table 2. The Performance of different combination of features.

Feature Combination	ACC	Spe	Sen	MCC
PSSM	**0.984**	**0.999**	0.386	0.582
TOPO	0.548	0.309	0.788	0.113
PCP	0.973	0.998	0.216	0.413
SeqSeg	0.973	0.999	0.189	0.389
PSSM+TOPO	0.98	0.998	0.415	0.603
PSSM+PCP	0.979	0.998	0.421	0.599
PSSM+ SeqSeg	0.98	0.998	0.415	0.601
TOPO+PCP	0.973	0.998	0.22	0.416
TOPO+ SeqSeg	0.904	0.981	0.139	0.201
PCP+ SeqSeg	0.973	0.998	0.213	0.41
PSSM+TOPO+PCP	0.979	0.998	0.422	0.6
PSSM+TOPO+ SeqSeg	0.98	0.998	0.416	0.603
PSSM+PCP+ SeqSeg	0.979	0.998	0.421	0.601
TOPO+PCP+ SeqSeg	0.973	0.998	0.216	0.414
PSSM+TOPO+PCP+ SeqSeg	0.971	0.997	**0.464**	**0.627**

Random Under-Sampling Scheme Influence the Predict Performance

Membrane protein-ligand binding sites prediction is a typical imbalance problem. As illustrate in table 1, the negative samples are about 140-200 times more than positive ones and will cause considerable noise. In this study, we use the Random Under-Sampling scheme to reduce the negative influence of imbalanced data. Due to the limit number of the positive sample, we keep all of them and randomly select some negative samples to build a sub-training dataset. The ratio of negative and positive samples in the sub-training dataset is the most important parameter which seriously affects the performance of the predictor. Figure 5 shows the tendency of MCC value as Ratio changes on (a) training dataset and (b) independent testing dataset. We can see in Figure 4.(a) that the predictor achieves the best MCC value when ratio=1 and decreased rapidly as ratio increase and tends to stabilize when ratio>30. We put forward on inference about this situation. When ratio<30, selected negative samples are too little to describe the distribution of original negative samples and this may cause serious information lost. When ratio>=30, the MCC value tends to stabilize but decreased smoothly due to the increasing noise of redundant samples. The tendency of MCC value of independent validation shown in Figure 4.(b) further verifies our inference.

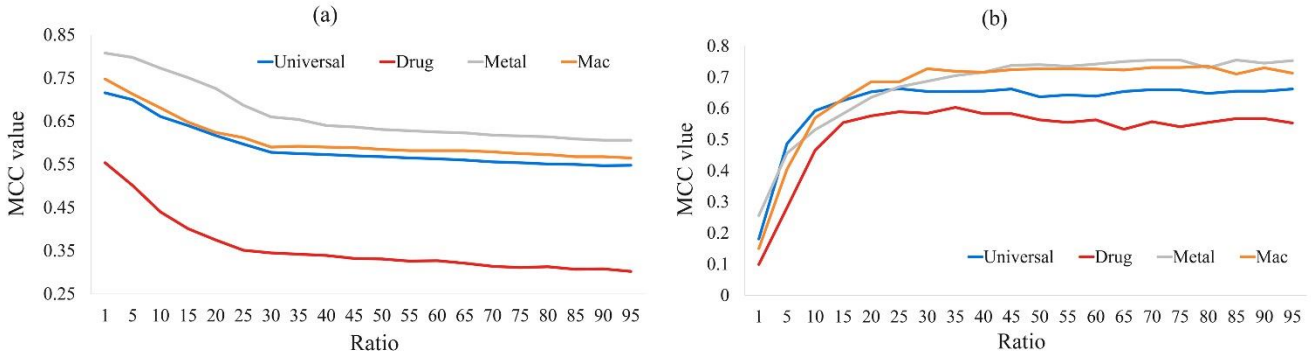

Figure 4. The tendency of MCC value as the ratio of non-binding residues and binding residues increase.

Compare with other Machine Learning Methods Over Cross Validation

In this section, we compare Random Forest Classifier with other classification methods on the training dataset. As illustrated in Table 3, it is obvious that ensemble classifier, including AdaBoost and RF, performs better than others, it might because of the serious imbalance between positive and negative

samples and ensemble method could reduce the negative influence of imbalanced data. Random Forest Classifier achieve best MCC value.

Table 3. Comparison RF with other classifiers

Method	ACC	Spe	Sen	MCC
SVM	0.9578	0.98	0.774	0.347
Naïve Bayes	0.844	0.85	0.67	0.246
AdaBoost*	0.974	0.995	0.334	0.472
RF	**0.971**	**0.997**	**0.464**	**0.627**

* Adaboost classifier employee decision tree as the basic classifier

Performance of MPLs-Pred

The performance of MPLs-Pred on the training datasets of 3 considered ligands over 10-fold cross validation are listed in Table 4. By observing the illustration of Table 4. Comparing with universal models, the ligand specific predictor can significantly improve the predictive effect for metal and biomacromolecule binding residues, especially for metal binding residues. But the prediction accuracy of drug-like compound binding residues is much worse than the universal model. We have more interest in why drug binding residues predicting perform worse than others. We speculate that drugs, manmade chemical compounds, are very different from natural ligands. According to the statistic of the previous section, The differences of drug-binding residues and non-drug-binding residues are not as significant as natural ligand`s. It is because the process of interaction of membrane protein and the ligand is interventional: membrane protein will not evolve a special region to recognize a particular drug and further binding with it. Thus, features derived from sequence could not commendably describe the characteristics of drug binding residues.

Table 4. Performance of MPLs-Pred on training dataset with universal model and ligand-specific models over 10-fold cross-validation

Model	ACC	Spe	Sen	MCC
Universal	0.971	0.997	0.464	0.627
Drug	0.973	1.0	0.153	0.366
Metal	0.984	0.997	0.589	0.704
Biomacromolecule	0.936	0.993	0.481	0.629

In order to prove the robustness of the predictor, we further compare the universal model and ligand-specific models on the independent testing dataset. The details are illustrated in Table 5 We find

that all the ligand specific predictors perform better than the universal one. The metal-ligand binding sites predictor achieves the best performance because of the highly significant characteristics of metal binding residues.

Table 5. Performance of MPLs-Pred on the independent testing dataset with universal model and ligand-specific models

Model	ACC	Spe	Sen	MCC
Universal	0.996	0.998	0.618	0.63
Drug	0.997	1.0	0.397	0.604
Metal	0.996	0.998	0.759	0.7
Biomacromolecule	0.997	0.999	0.596	0.692

Case Studies

To further demonstrate the effectiveness of the MPLs predictor on the universal model and ligand-specific models, we take a metal binding (UniProt ID: P00959, PDB ID: 1pfu) protein and a drug binding protein (UniProt ID: Q43133, PDB ID: 2j1p) in the testing dataset for case studies.

P00959 is a cytoplasm membrane protein of Escherichia Coli. It is the target protein of ATP, tRNA and Zinc ion, and also participate in aminoacyl-tRNA ligase activity. 4 Zinc binding sites are annotated in UniProt. The protein structure and Zinc ion binding residues are visualized in Figure 4.(a). It is obvious that the 4 amino acids that are not contiguous in the sequence are spatially clustered, forming a functional domain. Furthermore, the prediction results generated by MPLs-Pred with universal model and metal-specific model are also illustrated. The result shows that the metal-specific model outperforms the universal model. MPLs-Pred with metal-specific model correctly predict all 4 binding residues and universal model correctly identified 3 out of the 4 binding residues.

Q43133 is a chromoplast membrane protein of Sinapis Alba, It is the target protein of Magnesium, Isopentenyl diphosphate, and Dimethylallyl diphosphate. 6 Dimethylallyl diphosphate binding sites are annotated in UniProt. The details of the protein structure and drug-binding model are visualized in Figure 4. (b). Same as the previous one, 6 binding residues also forming a functional domain. MPLs-Pred also achieves considerable accuracy on this protein, universal model correctly predict 3 out of the 6 binding residues, and drug-specific model correctly predicts 5 out of the 6 binding residues.

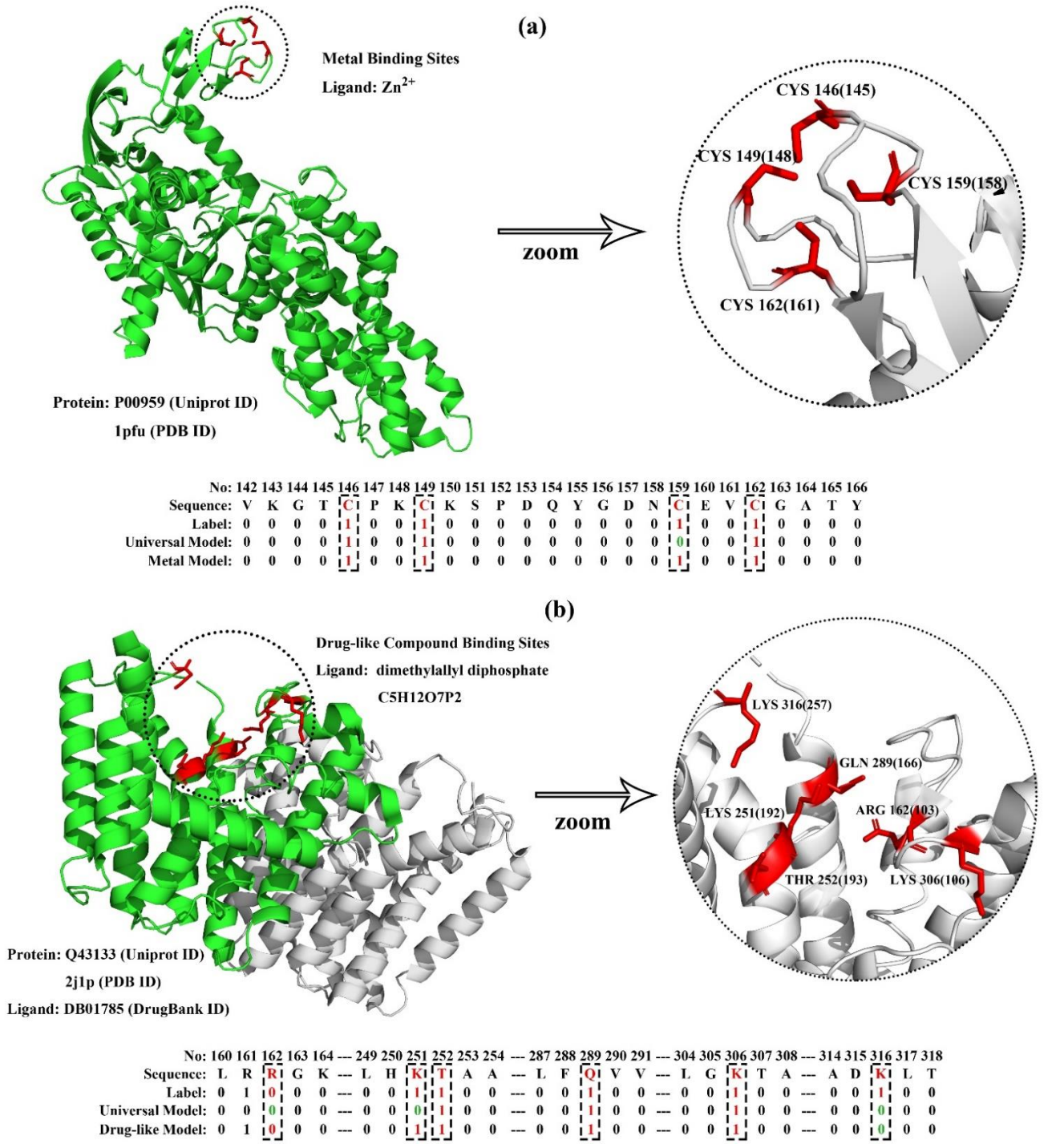

Figure 4. Visualization of (a). P00959: metal-binding MP with 4 Zinc ion binding residues and (b). Q43133: drug-binding MP with 6 dimethylallyl diphosphate binding residues. And their corresponding prediction result generated by MPLs-Pred with universal model and ligand-specific model.

Compare with Existing Methods on the Independent Testing Dataset

In this study, we proposed a novel method to predict ligand binding residues in membrane proteins named MPLs-Pred. In order to verify the generalization capability of MPLs-Pred, we compare it with existing predictor on the independent dataset.

There is only one previous study on the field of membrane protein-ligand binding residues prediction created by M.Xavirer Suresh and his team in 2015[18]. This predictor is named as tm-lig which is freely accessed from http://tmbeta-genome.cbrc.jp/tm-lig/tm-lig.html. tm-lig encode the residues by psi-blast generated PSSM profiles and employee Naïve Bayes classifier to predict the candidate ligand binding residues in membrane proteins. We compare MPLs-Pred predictor with tm-lig predictor on the independent dataset. The details are illustrated in Table 6.

The testing result proves that MPLs-Pred has better performance than tm-lig on all the evaluation indexes, especially on MCC value, which reflect the overall performance of the predictor.

Table 6. Comparison of MPLs-Pred with the previous study on the independent dataset

Method	ACC	Spe	Sen	MCC
Tm-lig	0.896	0.897	0.73	0.144
MPLs-Pred	**0.996**	**0.998**	**0.618**	**0.63**

Homo`s Membrane Protein-Ligand Interactions

We have much interest in homo`s MP-Ligand interactions. Here we build an exclusive predictor to identify ligand binding residues in homo membrane proteins and achieve considerable performance on 10-fold cross-validation. The ACC, Spe, Sen and MCC value is 0.992, 0.993, 0.705 and 0.486, respectively.

In this section, we analyze the Gene Ontology and Pathway of Homo`s drug binding membrane proteins to help to understand their functions.

Gene Ontology (GO) is an important initiative to unify the representation of gene and gene product attributes across all species. The enrichment analysis is to test whether a GO term is statistically enriched for given data. Among homo`s ligand binding membrane proteins. Figure 6 illustrates the GO analysis with up to 10 significantly enriched term in (a). Biological Process (BP), (b). Cell Component (CC) and Molecular Function (MF) respectively. For 337 homo proteins in the proposed dataset. 6100

biological processes are enriched and 4248 are statistically significant. Single-organism processes are the most important biological process for homo`s MP. 642 cell components are enriched and 401 are statistically significant. Intracellular, cytoplasm and organelle are top 3 CC enrichment but not much better than others, membrane proteins are distributed in almost all organelles without significant difference. 1153 molecular function are enriched and 561 are statistically significant. Binding with ligand in the most important function of membrane proteins.

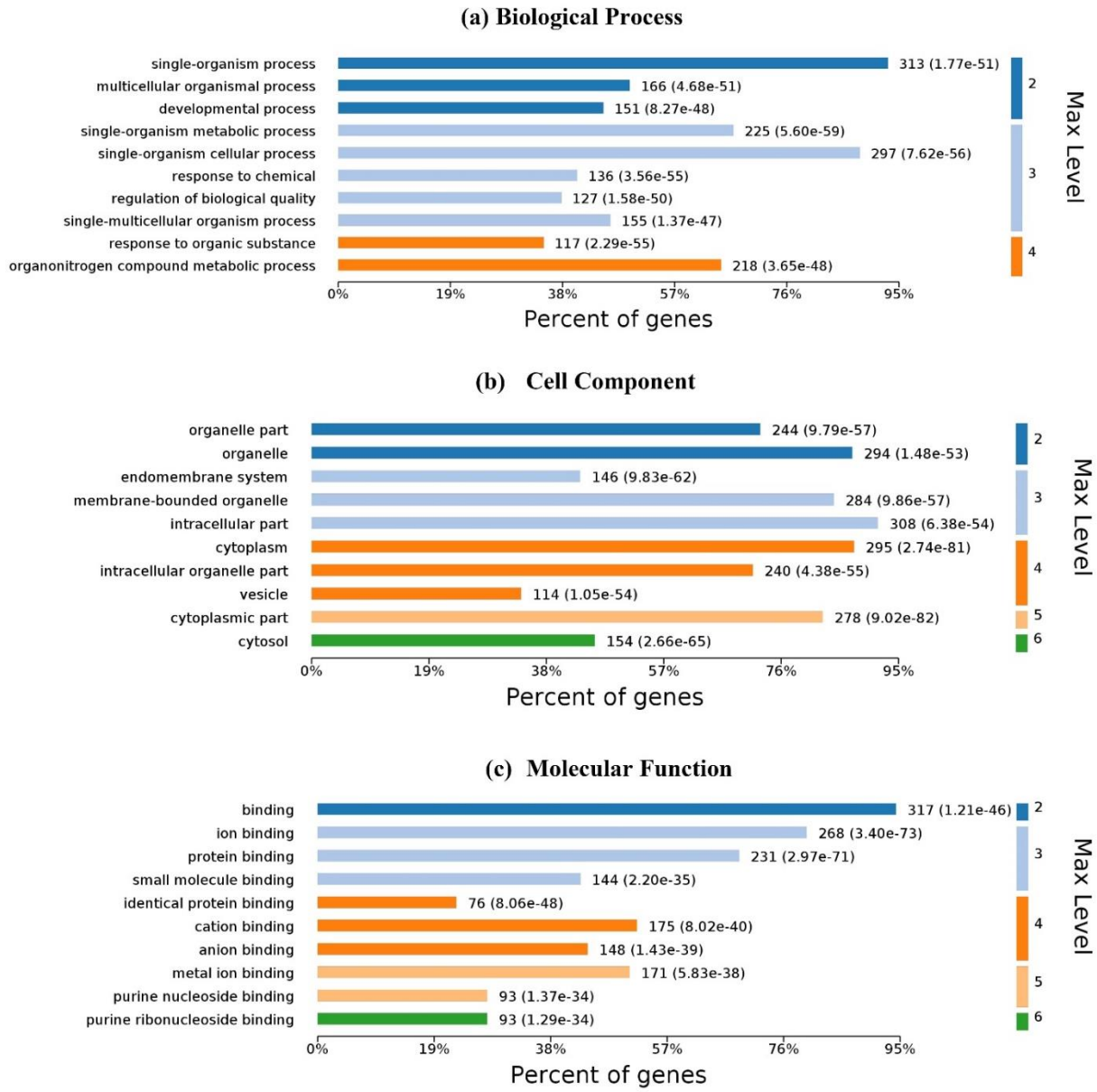

Figure 6. Gene Ontology Enrichment of homo`s ligand binding proteins: (a). Biological Process Enrichment; (b). Cell Component Enrichment; (c). Molecular Function Enrichment. The pictures were made with OmicsBean.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a manually curated pathway database for understanding high-level functions and utilities of the biological system. KEGG Pathway is a collection of manually drawn pathway maps. The enrichment analysis of KEGG Pathway helps researchers understanding the pathway a given set of proteins involved in. For homo`s ligand binding MPs, enriched processes are shown in Figure 7.

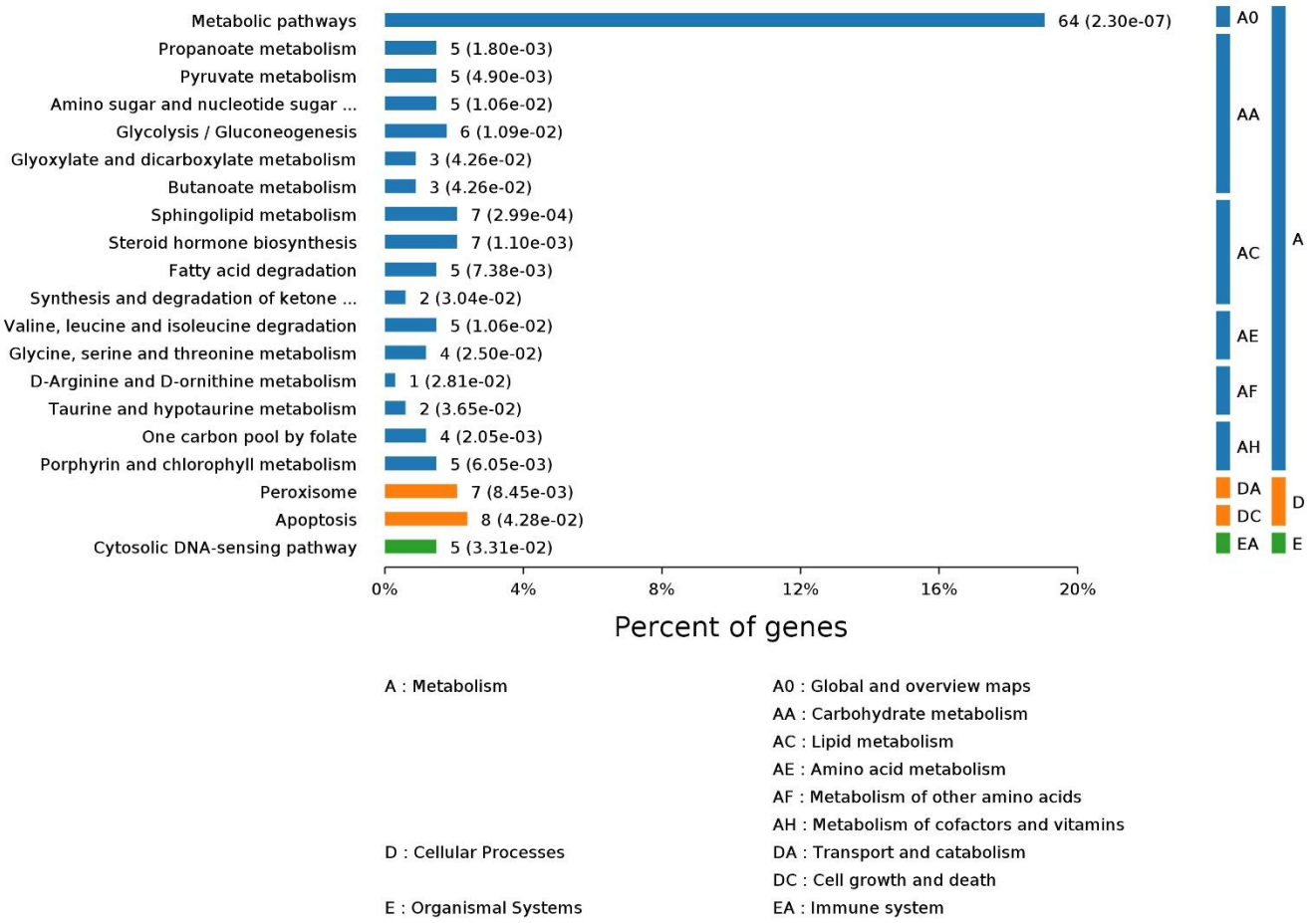

Figure 7. The KEGG Pathway Enrichment of homo`s ligand binding proteins. The pictures were made with OmicsBean.

Performance Evaluation

The proposed prediction model MPLs-Pred is first evaluated by ten-fold cross validation on the training dataset. First, positive samples and negative samples are randomly divided into 10 equal parts respectively. Then choose one part of the positive and negative subset to build the validation dataset and the remaining samples are used for training. This process would be repeated for ten times to build ten sub-predictors. The final performance is the average value of then sub-predictors. Then independent

validation is used to evaluate the generalization of the proposed method. Four metrics are employed to evaluate the performance of the predictor: Specificity (Spe), Sensitivity (Sen), Accuracy (ACC), and the Matthews Correlation Coefficients (MCC):

$$\text{Spe} = \frac{TN}{TN + FP}$$

$$\text{Sen} = \frac{TP}{TP + FN}$$

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Where TP, FP, TN, and FN represent true positive, false positive, true negative and false negative, respectively.

The prediction of membrane protein-ligand binding residues is a typical imbalanced learning problem, that is the number of non-binding residues is significantly more than that of binding residues. Thus, excessive pursuing of overall accuracy is one-sided. In the field of practical application, the researchers always expect that the predictor can provide more accuracy of ligand-binding residues instead of non-binding ones to get more candidate targets. In view of this, MCC value, which provides an overall measurement of performance of binary classification problems, is regarded as the most reliable evaluation index in the experiments of this paper. The performance of the predictor is positively correlated with the MCC value.

Conclusion

In this study, we proposed a novel membrane protein-ligand binding residues predictor named MPLs-Pred. We figure the target residue by 4 types of sequence derived features including evolution profiles, topology structure, physicochemical properties, and primary sequence segment descriptors. Than Random Forest classifier is employed to predict if a given residue is ligand binding sites or not. Experimental results showed that MPLs-Pred achieves considerable performance with MCC is 0.597 and 0.356 on cross-validation and independent validation, respectively. Above this, we propose a ligand specific models, classified ligand into drug, metal, and biomacromolecule, to further improve the

prediction accuracy. The ligand specific models significantly improve the performance compare with the universal model.

Acknowledgments

This work is supported by the National Natural Science Funds of China (No. 81671328, 61802057) the Jilin Scientific and Technological Development Program (20180414006GH, 20180520028JH, 20170520058JH), and The Science and Technology Research Project of the Education Department of Jilin Province under Grant No.JJKH20190290KJ, JJKH20191309KJ).

References

1. Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G., SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* **2014,** 42, (D1), D310-D314.
2. Almen, M. S.; Nordstrom, K. J. V.; Fredriksson, R.; Schioth, 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.
3. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L., Opinion - How many drug targets are there? *Nature Reviews Drug Discovery* **2006,** 5, (12), 993-996.
4. Phillips-Jones, M. K., Structural and biophysical characterisation of membrane protein-ligand binding Preface. *Bba-Biomembranes* **2014,** 1838, (1), 1-2.
5. Hernandez, M.; Ghersi, D.; Sanchez, R., SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res* **2009,** 37, (Web Server issue), W413-6.
6. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009,** 10, 168.
7. Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J., CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res* **2018,** 46, (W1), W363-W367.
8. Hendlich, M.; Rippmann, F.; Barnickel, G., LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* **1997,** 15, (6), 359-63, 389.
9. Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N., ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **2003,** 19, (1), 163-164.
10. Hu, J.; Li, Y.; Zhang, M.; Yang, X. B.; Shen, H. B.; Yu, D. J., Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs. *Ieee Acm T Comput Bi* **2017,** 14, (6), 1389-1398.
11. Yu, D. J.; Hu, J.; Yang, J.; Shen, H. B.; Tang, J.; Yang, J. Y., Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans Comput Biol Bioinform* **2013,** 10, (4), 994-1008.
12. Chen, K.; Mizianty, M. J.; Kurgan, L., Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* **2012,** 28, (3), 331-41.
13. Fathima, A. J.; Murugaboopathi, G.; Selvam, P., Pharmacophore Mapping of Ligand Based Virtual Screening, Molecular Docking and Molecular Dynamic Simulation Studies for Finding Potent NS2B/NS3 Protease Inhibitors as Potential Anti-dengue Drug Compounds. *Current Bioinformatics* **2018,** 13, (6), 606-616.
14. Hu, J.; Li, Y.; Zhang, Y.; Yu, D. J., ATPbind: Accurate Protein-ATP Binding Site Prediction by Combining Sequence-Profiling and Structure-Based Comparisons. *Journal of Chemical Information and Modeling* **2018,** 58, (2), 501-510.
15. Huang, B. D.; Schroeder, M., LIGSITE(csc): predicting ligand binding sites using the Connolly surface and degree of conservation. *Bmc Struct Biol* **2006,** 6, 19.
16. Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A., Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *Plos Comput Biol* **2009,** 5, (12), 12.
17. Yang, J. Y.; Roy, A.; Zhang, Y., Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013,** 29, (20), 2588-2595.

18. Suresh, M. X.; Gromiha, M. M.; Suwa, M., Development of a machine learning method to predict membrane protein-ligand binding residues using basic sequence information. *Adv Bioinformatics* **2015,** 2015, 843030.
19. Moraes, I.; Evans, G.; Sanchez-Weatherby, J.; Newstead, S.; Stewart, P. D. S., Membrane protein structure determination The next generation. *Bba-Biomembranes* **2014,** 1838, (1), 78-87.
20. Brown, D. A.; London, E., Functions of lipid rafts in biological membranes. *Annu Rev Cell Dev Biol* **1998,** 14, 111-36.
21. Hong, H.; Chang, Y. C.; Bowie, J. U., Measuring transmembrane helix interaction strengths in lipid bilayers using steric trapping. *Methods Mol Biol* **2013,** 1063, 37-56.
22. Alonso, M. A.; Millan, J., The role of lipid rafts in signalling and membrane trafficking in T lymphocytes. *J Cell Sci* **2001,** 114, (Pt 22), 3957-65.
23. Maldonado, S.; Lopez, J., Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recogn* **2014,** 47, (5), 2070-2079.
24. O'Brien, R.; Ishwaran, H., A random forests quantile classifier for class imbalanced data. *Pattern Recogn* **2019,** 90, 232-249.
25. Zou, Q.; Li, X.; Jiang, Y.; Zhao, Y.; Wang, G., BinMemPredict: a Web Server and Software for Predicting Membrane Protein Types. *Current Proteomics* **2013,** 10, (1), 2-9.
26. Zhang, J.; Chai, H.; Gao, B.; Yang, G.; Ma, Z., HEMEsPred: Structure-Based Ligand-Specific Heme Binding Residues Prediction by Using Fast-Adaptive Ensemble Learning Scheme. *IEEE/ACM Trans Comput Biol Bioinform* **2018,** 15, (1), 147-156.
27. Sodhi, J. S.; Bryson, K.; McGuffin, L. J.; Ward, J. J.; Wernisch, L.; Jones, D. T., Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* **2004,** 342, (1), 307-20.
28. Hu, J.; He, X.; Yu, D. J.; Yang, X. B.; Yang, J. Y.; Shen, H. B., A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. *PLoS One* **2014,** 9, (9), e107676.
29. 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, (1), 235-42.
30. Zou, Q.; Lin, G.; Jiang, X.; Liu, X.; Zeng, X., Sequence clustering in bioinformatics: an empirical study. *Briefings in Bioinformatics* **2019**, Doi: 10.1093/bib/bby090.
31. Jeong, J. C.; Lin, X.; Chen, X. W., On position-specific scoring matrix for protein function prediction. *IEEE/ACM Trans Comput Biol Bioinform* **2011,** 8, (2), 308-15.
32. Wei, Z. S.; Han, K.; Yang, J. Y.; Shen, H. B.; Yu, D. J., Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing* **2016,** 193, 201-212.
33. Zangooei, M. H.; Jalili, S., Protein secondary structure prediction using DWKF based on SVR-NSGAII. *Neurocomputing* **2012,** 94, 87-101.
34. Zhang, J.; Gao, B.; Chai, H. T.; Ma, Z. Q.; Yang, G. F., Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm. *Bmc Bioinformatics* **2016,** 17, 323.
35. Qu, K.; Wei, L.; Zou, Q., A Review of DNA-binding Proteins Prediction Methods. *Current Bioinformatics* **2019,** 14, (3), 246-254.
36. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** 25, (17), 3389-402.
37. von Heijne, G., Membrane-protein topology. *Nat Rev Mol Cell Bio* **2006,** 7, (12), 909-918.
38. Tsirigos, K. D.; Govindarajant, S.; Bassot, C.; Vastermark, A.; Lamb, J.; Shu, N. J.; Elofsson, A., Topology of membrane proteins - predictions, limitations and variations. *Curr Opin Struc Biol* **2018,** 50, 9-17.
39. Tsirigos, K. D.; Peters, C.; Shu, N.; Kall, L.; Elofsson, A., The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* **2015,** 43, (W1), W401-7.
40. Zhang, J.; Chai, H. T.; Yang, G. F.; Ma, Z. Q., Prediction of bioluminescent proteins by using sequence-derived features and lineage-specific scheme. *Bmc Bioinformatics* **2017,** 18, 294.
41. Chai, H. T.; Zhang, J., Identification of Mammalian Enzymatic Proteins Based on Sequence-Derived Features and Species-Specific Scheme. *Ieee Access* **2018,** 6, 8452-8458.
42. Suo, S. B.; Qiu, J. D.; Shi, S. P.; Sun, X. Y.; Huang, S. Y.; Chen, X.; Liang, R. P., Position-Specific Analysis and Prediction for Protein Lysine Acetylation Based on Multiple Features. *Plos One* **2012,** 7, (11), 11.
43. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M., AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* **2008,** 36, D202-D205.
44. Wan, S.; Duan, Y.; Zou, Q., HPSLPred: An Ensemble Multi-label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source. *Proteomics* **2017,** 17, 1700262.
45. Song, L.; Li, D.; Zeng, X.; Wu, Y.; Guo, L.; Zou, Q., nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *Bmc Bioinformatics* **2014,** 15, 298.
46. BREIMAN, L., Random Forests. *Machine Learning* **2001,** 45, 5-32.
47. Su, R.; Liu, X.; Wei, L.; Zou, Q., Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods (San Diego,*

Calif.) **2019**, Doi: 10.1016/j.ymeth.2019.02.009.
48. Zhao, X.; Zou, Q.; Liu, B.; Liu, X., Exploratory Predicting Protein Folding Model with Random Forest and Hybrid Features. *Current Proteomics* **2014,** 11, (4), 289-299.
49. Khan, M.; Hayat, M.; Khan, S. A.; Iqbal, N., Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. *J Theor Biol* **2017,** 415, 13-19.
50. Hu, J.; Li, Y.; Yang, J. Y.; Shen, H. B.; Yu, D. J., GPCR-drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. *Computational Biology and Chemistry* **2016,** 60, 59-71.
51. Liao, Z.; Ju, Y.; Zou, Q., Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest. *Scientifica* **2016,** 2016, 8309253.