Bioinformatics, 34(15), 2018, 2598–2604
doi: 10.1093/bioinformatics/bty162
Advance Access Publication Date: 14 March 2018
Original Paper

Structural bioinformatics

CoABind: a novel algorithm for Coenzyme A (CoA)- and CoA derivatives-binding residues prediction

Qiaozhen Meng[1], Zhenling Peng[1,*] and Jianyi Yang[2,*]

[1]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China and [2]School of Mathematical Sciences, Nankai University, Tianjin 300071, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 26, 2017; revised on February 26, 2018; editorial decision on March 12, 2018; accepted on March 13, 2018

Abstract

Motivation: Coenzyme A (CoA)-protein binding plays an important role in various cellular functions and metabolic pathways. However, no computational methods can be employed for CoA-binding residues prediction.

Results: We developed three methods for the prediction of CoA- and CoA derivatives-binding residues, including an *ab initio* method SVMpred, a template-based method TemPred and a consensus-based method CoABind. In SVMpred, a comprehensive set of features are designed from two complementary sequence profiles and the predicted secondary structure and solvent accessibility. The engine for classification in SVMpred is selected as the support vector machine. For TemPred, the prediction is transferred from homologous templates in the training set, which are detected by the program HHsearch. The assessment on an independent test set consisting of 73 proteins shows that SVMpred and TemPred achieve Matthews correlation coefficient (MCC) of 0.438 and 0.481, respectively. Analysis on the predictions by SVMpred and TemPred shows that these two methods are complementary to each other. Therefore, we combined them together, forming the third method CoABind, which further improves the MCC to 0.489 on the same set. Experiments demonstrate that the proposed methods significantly outperform the state-of-the-art general-purpose ligand-binding residues prediction algorithm COACH. As the first-of-its-kind method, we anticipate CoABind to be helpful for studying CoA-protein interaction.

Availability and implementation: http://yanglab.nankai.edu.cn/CoABind

Contact: zhenling@tju.edu.cn or yangjy@nankai.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Coenzyme A (CoA) is 'an important catalytic substance involved in the cellular conversion of food into energy' (Shampo and Kyle, 2000). CoA was first discovered more than 70 years ago by Lipmann (Lipmann, 1945), who was awarded the Nobel Prize owing to his contribution to the discovery and other research of CoA (Shampo and Kyle, 2000). Due to its prominent role in the metabolism of carboxylic acids, including short- and long-chain fatty acid, substantial experimental studies have been performed on CoA and CoA derivatives (e.g., acyl-CoA) (Sibon and Strauss, 2016). For example, in the latest version of the Protein Data Bank (PDB) (Rose *et al.*, 2017), about 500 structures of CoA-binding proteins have been determined. However, it is often time-consuming and costly to carry out wet-lab experiments. And it is necessary to make full use of the existing experimental data to develop corresponding computational algorithms for CoA-related studies, such as evolutionary analysis of CoA-binding proteins and CoA-binding residues prediction. For example, Engel and Wierenga compared the topologies of the structures of CoA-binding proteins in PDB and found that CoA-binding proteins had a diversity of folds (Engel and Wierenga, 1996). Burton *et al.* (2005) investigated the distribution and conservation of acyl-CoA-binding proteins using sequence alignment-based evolutionary analysis.

In this work, we proposed three computational algorithms to predict CoA- and CoA derivatives-binding residues in proteins. Specifically, we first designed an *ab initio* predictor SVMpred by using the evolution-enriched information as the inputs of support vector machine (SVM). We also trained a template-based method TemPred, using homologous templates recognized by the profile-profile alignment program HHsearch (Soding, 2005). Our assessment suggests that these two methods are complementary to each other. Therefore, the third method CoABind was developed, by combining SVMpred and TemPred. CoABind was shown to have the highest accuracy among these three methods. In addition, the proposed methods significantly outperform the state-of-the-art general purpose ligand-binding residues prediction algorithm COACH.

2 Materials and methods

2.1 Benchmark dataset

The benchmark dataset was constructed from the BioLiP database (Yang *et al.*, 2013b), which is a resource for biologically relevant ligand–protein interaction. To obtain enough data for training and test, we selected not just the CoA ligand with ID ‘COA’ but also other CoA derivatives. In total, there are 62 CoA derivatives, which share ≥ 0.8 similarity (measured by the Tanimoto coefficient of ligand fingerprints) with CoA. Please refer to the Supplementary Table S1 for the summary of these ligands and the corresponding number of binding protein chains.

In total, we extracted 1972 protein chains in BioLiP that bind any of the 63 selected ligands. Two steps are used to process these data. The first is to enrich experimental annotations and the second is to remove redundancy. (i) We used CD-HIT (Huang *et al.*, 2010) to cluster the chain sequences with 100% sequence identity. The longest sequence in a cluster is selected as the representative of the cluster, resulting to 776 sequences. To enrich experimental binding data, we transferred the CoA-binding residues of other sequences in a given cluster to the corresponding representative sequence based on the pairwise sequence alignment. This procedure was motivated by a recent work to assess the predictive quality of DNA/RNA-binding residues (Yan *et al.*, 2016). By such procedure, the number of CoA-binding residues was enriched by about 14%, compared with the original binding residues. (ii) We removed the redundancy of the 776 sequences from the first step at <25% sequence identity, by utilizing PSI-CD-HIT (Huang *et al.*, 2010). Finally, we got 219 sequences with 4024 CoA-binding residues. This dataset was then split into two non-overlapping subsets (2/3 for training and 1/3 for test) at random, including the training set with 146 proteins and 2679 binding residues (TR146), and the test dataset with 73 proteins and 1345 binding residues (TE73). The above benchmark datasets are available for download at: http://yanglab.nankai.edu.cn/CoABind/benchmark.

2.2 Architecture of the CoABind method

As shown in Figure 1, the proposed predictor CoABind is a combination of the *ab initio* predictor SVMpred and the template-based predictor TemPred. In CoABind, the propensity score p_c for each residue is calculated as the average of the two corresponding propensity scores from SVMpred and TemPred, respectively. A residue is then predicted as a binding residue if p_c is higher than a predetermined cutoff, which is tuned using the training set.

2.2.1 Template-based method TemPred

Template-based modeling is a leading approach to protein structure and function prediction. Therefore, we first proposed a template-based predictor TemPred, by following the flow chart shown in the darker gray panel of Figure 1. The query sequence is aligned to templates in the training set by the profile-profile alignment program HHsearch (Soding, 2005). The profile of a sequence is represented in the form of a hidden Markov model (HMM) generated by the program HHblits (Remmert *et al.*, 2012). Default parameters are used to run HHsearch. The top-ranked templates with *e*-value $\leq$ 0.001 are used to transfer the binding annotations to the query based on the template-query alignment. For the *i*th residue in the query sequence, the propensity score $p_t(i)$ for binding is calculated as:

$$p_t(i) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{Prob}_j \times \delta_j(i) \tag{1}$$

where N is the total number of templates with *e*-value $\leq$ 0.001; Prob_j is the probability value from HHsearch, measuring the quality of the alignment between the *j*th template and the query; and $\delta_j(i)$ is an indicator function, which equals to 1 if the *i*th residue in the query sequence is aligned to a binding residue in the *j*th template, and 0 otherwise. When there is no template with *e*-value $\leq$ 0.001, the one ranked at the top is used instead. The binary prediction by TemPred is then obtained by setting a cutoff for the propensity score: a residue is predicted as a binding residue if the propensity score is higher than the cutoff. This cutoff will be determined later using the training set. It is apparent that TemPred depends on the availability of homologous templates. For hard targets that HHsearch does not find any templates, no prediction can be made, and the corresponding metrics are set to 0 for assessment.

2.2.2 *Ab initio* predictor SVMpred

As mentioned earlier, the template-based predictor TemPred does not work for queries lacking homologous templates. To solve this problem, we developed the *ab-initio* predictor SVMpred that does not rely on templates. As shown in the lighter gray panel of Figure 1, the query is submitted to three programs PSI-BLAST (Altschul *et al.*, 1997), HHblits (Remmert *et al.*, 2012) and SPIDER3 (Heffernan *et al.*, 2017), to generate two sequence profiles and predict its secondary structure and solvent accessibility, respectively. They are believed to contain rich evolution and structure information and have been widely used in the field of protein structure prediction (Xia *et al.*, 2017). Here, we utilize them to encode each amino acid in the query for CoA-binding residues prediction. Specifically, each amino acid is represented by a 81D feature vector, which are next normalized to the range of [0, 1] and fed into SVM for classification. A residue is predicted as a binding residue if the probability score from SVM is higher than the specified threshold. These 81 features were selected from a total of 195 features extracted from the neighboring residues in a sliding window. To reduce the time used for training, the same window size was applied to all feature groups introduced later.

PSI-BLAST-based features. Evolutionary conservation is suggested to be a powerful indicator for the functionally important residues, which are usually more conserved than others. The residue conservation can be inferred from a multiple sequence alignment

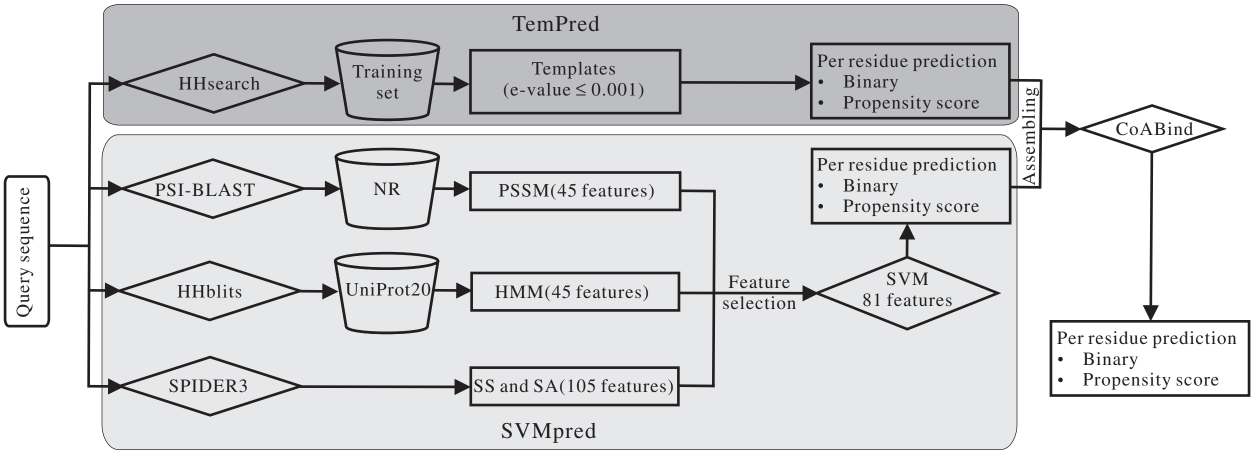

Fig. 1. The architecture of the consensus-based method CoABind, by assembling the template-based predictor TemPred and the *ab initio* predictor SVMpred

(MSA). To generate a MSA for each query sequence, PSI-BLAST is used to search its homologous sequences from the NCBI's non-redundant (NR) dataset with three iterations and the *e*-value threshold of 0.001 ('-j 3 -h 0.001'). A position-specific scoring matrix (PSSM) and a probability matrix are then calculated from the MSA. The conservation of each amino acid is represented by the 20D features from the PSSM, the relative entropy (RE) and the close neighbors correlation coefficient (CNCC) (Taherzadeh *et al.*, 2016) from the probability matrix.

$$\mathrm{RE}_i = \sum_{k=1}^{20} p_{ik} \times \log_2 \frac{p_{ik}}{b_k} \qquad (2)$$

$$\mathrm{CNCC}_{ij} = \frac{P_i \cdot P_j}{|P_i||P_j|} = \frac{\sum_{k=1}^{20} p_{ik}p_{jk}}{\sqrt{\sum_{k=1}^{20} p_{ik}^2 \sum_{k=1}^{20} p_{jk}^2}}, i \neq j \qquad (3)$$

where $i, j = 1, \ldots, L$, are the *i*th and the *j*th residue, respectively, in the query sequence with L amino acids; k represents one of the 20 standard amino acids; b_k is the Robinson background frequency of the *k*th amino acid (Robinson and Robinson, 1991); p_{ik} is the probability of the *k*th amino acid appearing at the *i*th row of the MSA (corresponding to the *i*th position of the query); P_i is the 20D probability vector, corresponding to the *i*th row in the probability matrix. Note that, we used a sliding window of size w to calculate the CNCC for each residue in the query. Therefore, we extracted 20 + w features in total from PSI-BLAST.

HHblits-based features. Besides PSI-BLAST, the program HHblits is used to generate another sequence profile, which is represented in the form of a HMM (Remmert, *et al.*, 2012). The HMM profile is obtained by scanning the query sequence through the database uniprot20_2015_06 with parameters '-n 3 -maxfilt 500 000 -diff inf -id 99 -cov 60'. Each line in this profile comprises the emission frequencies (EFs) for the 20 standard amino acids, 7 transition probabilities and 3 local diversities. The EF is defined by the equation:

$$\mathrm{EF}_{ik} = -1000 \times \log_2 p_{ik} \qquad (4)$$

where $i = 1, \ldots, L$ is the *i*-th residue in the query; k represents a standard residue; Based on this equation, each EF_{ik} is converted into probability p_{ik}, which is 0 when the EF is denoted by a '*'. Similar to the PSI-BLAST-based features, the recovered probability matrix is then used to measure the residue conservation, by including the 20D vector for this matrix, the RE, and the CNCC as well. Here, we used the window size of w for the calculation of HMM-based CNCC values. Therefore, the resulting HHblits-based feature set contains 20 + w features.

SPIDER3-based features. The predicted secondary structure (SS) and solvent accessibility (SA) have been applied to the detection of RNA-/DNA- and peptide-binding residues (Peng and Kurgan, 2015; Taherzadeh, *et al.*, 2016; Zhang, *et al.*, 2010). We employed the program SPIDER3 (Heffernan *et al.*, 2017) to predict the SS and SA for each sequence. Here, the SS profile comprises the most likely SS state and the corresponding probability in each of the three states, which are α-helix, β-strand and random coil. The SA profile gives the predicted solvent accessible surface area for each residue. We normalized it into a relative SA profile (RSA) by the corresponding maximum possible solvent accessible surface area for each residue. A residue is regarded as exposed if its RAS value is >0.5; otherwise, it is regarded as buried. For each residue, a sliding window of size w centered at this residue is used to extract features. For the SS profile, we calculated the fraction of each SS state (three features) and collected the probability values for all residues inside the window (3*w features). For the RSA profile, to represent the status of a residue (exposed or buried), two features are used (0, 1; or 1, 0). In addition, the RSA values for all residues inside the sliding window of size w are also used (w features). To summarize, each residue is converted into a (5 + 4*w)-dimensional feature vector from the SPIDER3 profile.

Support vector machine. For a window of size w, the total number of features extracted is (45 + 6*w). These features are fed into SVM for training and test. This was motivated by the fact that SVM is one of the state-of-the-art algorithms with wide applications in classification problems, such as protein fold classification (Chen and Kurgan, 2007; Xia *et al.*, 2017) and peptide-binding residues prediction (Taherzadeh *et al.*, 2016). SVM has several basic kernel functions, including linear, polynomial, radial basis function (RBF) and sigmoid. We used the RBF kernel here since it provided higher accuracy in our concerned classification problem. We need to optimize the SVM regularization factor C and the RBF kernel parameter γ, which were implemented by the strategy of grid search. That is to say, each pair of tested values for parameters C and γ is represented by a point in a 6 × 5 grid, where the values for C and γ *are* $[2^0, 2^1, \ldots, 2^5]$ and $[2^{-1}, \ldots, 2^{-5}]$, respectively. The parameter optimization is conducted to maximize the average Matthews correlation coefficient (MCC) evaluated by 5-fold cross validation on the training set. The LIBSVM package (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used for the implementation of SVM.

2.3 Evaluation measures

The predictions by the proposed methods contain both binary values and propensity scores. The binary value indicates each residue in the query sequence to be a CoA-binding residue or not; and the propensity score quantifies its probability of being a CoA-binding residue. Since the data are highly imbalanced (positive to negative ratio is about 1:15), the binary prediction is assessed by the following three metrics, Precision (Pre), Recall (Rec) and MCC:

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{6}$$

where TP (true positive) is the number of correctly predicted binding residues, TN (true negative) is the number of correctly predicted non-binding residues, FP (false positive) is the number of non-binding residues predicted to be binding residues, and FN (false negative) is the number of binding residues predicted to be non-binding residues. The value of a metric equals zero when the denominator is zero. The higher the above metrics are, the better the prediction is. MCC ranges from −1 to 1 and it is especially suitable for assessing data with imbalanced distribution.

The predictions with propensity scores are evaluated by the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). For a cutoff p (from 0 to 1), the residues with propensity score $\geq p$ are set as positives (i.e. binding residues); otherwise, the residues are set as negatives (non-binding residues). Thus one point (FP-rate, TP-rate) in a plane is obtained from each cutoff, where TP-rate = TP/(TP + FN), and FP-rate = FP/(FP + TN). The ROC curve is then generated by connecting all points. The AUC implies the quality of the prediction. The AUC value is between 0 and 1 and the higher AUC value, the better the prediction is. An AUC value of around 0.5 suggests a random prediction. Because higher FP-rates will lead to a massive over-prediction of binding residues while this part of the curve dominates the AUC value, another metrics ($\mathrm{AUC_L}$) is also calculated by limiting the region to the low FP-rate from 0 to 0.1. The ratio (R) of $\mathrm{AUC_L}$ over the corresponding $\mathrm{AUC_L}$ for random prediction is also reported to reflect the accuracy of the prediction. These two metrics were proposed in (Meng and Kurgan, 2016; Yan and Kurgan, 2017) and we mainly use them for the comparisons between different methods. It is worthwhile to mention that all metrics defined above were computed per protein and the averages over all proteins in a dataset are reported.

3 Results

3.1 Parameter optimization and feature selection

There are four parameters to be determined: the window size w in SVMpred and the three cutoffs for making binary predictions in SVMpred, TemPred and CoABind. The parameters are selected to optimize the MCC on the training set TR146 based on 5-fold cross validation. When tuning the window size, the cutoff in SVMpred was set to the default one in LIBSVM (i.e. 0.5). The Supplementary Figure S1 shows that the maximum MCC was achieved at the window size of 25. At this window size, the total number of features is 195 (=45 + 6*25). These features may be redundant and thus feature selection was applied to select a subset of non-redundant and key features. To this end, the correlation-based feature subset selection algorithm implemented in Weka (Hall *et al.*, 2009) was applied with default parameters. A total of 81 features were selected, at which the MCC was similar to that with all features (0.32 versus 0.33). With these 81 features, we further trained the cutoff for making binary predictions to maximize the MCC. The Supplementary Figure S2 shows that the MCC is the highest when the cutoff is 0.14. The cutoffs for TemPred and CoABind were adjusted similarly and the optimal values are 0.3 and 0.31, respectively.

3.2 Contribution of features in SVMpred

The contribution of the 81 selected features to the SVMpred prediction was investigated as below. We found these features are from different profiles: 28 from the PSI-BLAST profile, 16 from the HHblits profile and 37 from the SPIDER3 profile. Thus these features are divided into three groups correspondingly. We first investigate the contribution of each individual feature group to the prediction. This analysis is performed on the training set based on 5-fold cross validation. The predictive quality is presented by the average MCC and AUC values, which is summarized in Figure 2. We can see that all the three groups of features are powerful indicators of CoA-binding residues, based on the fact that the MCC and AUC values for the SVM models built with individual group of features are above 0.15 and 0.69, respectively. The SVM model with the SPIDER3-based features has the highest MCC value of 0.289 and AUC value of 0.728. When combining any two groups of features, we found that the combinations lead to better SVM models with 15–121% improvement in MCC and 3–8% increase in AUC (the gray bars to the white bars in Fig. 2). This observation suggests that the three feature groups are complementary to each other. We note that PSSM feature group is different from the HMM feature group, even if both of them are derived from MSA. This difference may result from different algorithms for alignment (sequence-profile alignment in PSI-BLAST, and profile-profile alignment in HHblits). When combining all these features together to build the model, it makes 36–160% higher MCC and 7–11% higher AUC than any individual SVM models (the black bar to the white bars in Fig. 2). It also has at least 6% more MCC and 2% more AUC than the combinations with two feature groups (the black bar to the gray bars in Fig. 2). This further supports the conclusion that these features are

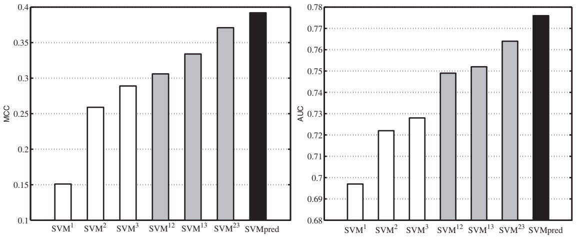

Fig. 2. Evaluation on the SVM models built with individual feature group and combinations of feature groups. The white, gray, black bars show the average MCC (the left panel) and AUC (the right panel) values for the corresponding SVM models by 5-fold cross validation on the training set. SVM^i (i = 1, 2, 3) denotes the SVM model built with the feature group i, where 1, 2 and 3 represent the HHblits-, PSI-BLAST- and SPIDER3-based feature set, respectively. SVM^{ij} ($i \neq j$, i, j = 1, 2, 3) represents the SVM model by using the feature set i together with the feature set j. SVMpred is the model by combining all features

complementary to each other. Therefore, the SVM model with all features is finally implemented as the *ab initio* predictor SVMpred.

3.3 Performance of the proposed methods on the independent test set

We next assessed the performance of the proposed methods on the independent test set TE73. The results are shown in Table 1, which shows that SVMpred achieves satisfactory MCC and AUC, which are 0.438 and 0.817, respectively. The MCC for TemPred is higher than that for SVMpred but with lower AUC and $\mathrm{AUC_L}$. In addition, the Pre for SVMpred is 3% higher than TemPred while the Rec for TemPred is about 18% higher than SVMpred. These data show that these two methods are complementary and the consensus of them by CoABind is possible to improve the prediction. This is in fact true as reflected in Table 1. It shows that CoABind achieves higher values for all metrics over both methods except for the Pre, which is comparable to SVMpred (0.510 versus 0.513).

Statistical tests were performed on the difference between the methods in Table 1 based on the MCC values, similar to the procedure used in (Meng and Kurgan, 2016; Yan and Kurgan, 2017). We randomly drew a half number of proteins from the test set TE73 and then computed their average MCC for each pair of methods. This experiment was repeated 10 times to generate 10-paired results. The Anderson–Darling test was first used to test whether the data follow a normal distribution or not at 0.05 significance level. The paired *t*-test is applied for a normal distribution. Otherwise, the nonparametric Wilcoxon signed-rank test is utilized. The *P*-value returned from the test indicates the significance level of the difference between two compared methods. The results for the all-against-all tests are shown in the Supplementary Table S2 (the data with gray background). It indicates that TemPred outperforms SVMpred significantly at $P\text{-value} < 10^{-4}$, while CoABind outperforms both SVMpred and TemPred significantly at $P\text{-value} < 0.01$.

Note that the maximum sequence identity between the test and the training proteins is 25%, which is a global metric. It would be interesting to measure the similarity with a local metric. To this end, by searching each test protein against the training set with PSI-BLAST, we found that there are 41/32 proteins with/without homologous templates (at e-value ≤ 0.001). Thus, we divided the whole test set into two subsets TE41 and TE32, on which the methods were further compared. The comparison results are summarized in the Supplementary Table S3. For all methods, the predictions for proteins in TE41 are significantly more accurate than TE32. This is understandable as both SVMpred and TemPred rely on the local similarity to the training proteins to make predictions. On both subsets, SVMpred has lower MCC but higher $\mathrm{AUC_L}$ values than TemPred, showing these two methods are complementary again.

From Supplementary Table S3, on the new test set TE32, we can see that the MCC for TemPred is still satisfactory and higher than SVMpred (0.32 versus 0.287), though the similarity between TE32 and the training set TR146 is small (both globally and locally). This is probably because the alignment tool HHsearch (used in TemPred) is more sensitive than PSI-BLAST (used for defining local similarity between test and training proteins). When we re-define the local similarity based on HHsearch, the number of test proteins without significant local similarity (at *e*-value ≤ 0.001) to the training proteins was reduced to 12. On these proteins, the average MCCs for SVMpred, TemPred and CoABind are very low, which are 0.061, 0.044 and 0.084, respectively.

Table 1. The predictive quality of the proposed methods on the independent test set TE73

Methods	MCC	Pre	Rec	AUC	$\mathrm{AUC_L}$	R
SVMpred	0.438	**0.513**	0.453	0.817	0.047	9.067
TemPred	0.481	0.498	0.535	0.794	0.039	7.573
CoABind	**0.489**	0.510	**0.571**	**0.849**	**0.055**	**10.516**

The best results are highlighted in bold type.

3.4 Case study

We further investigate the relationship between the predictions of the proposed methods with a specific example. The example protein is '*spermidine N-acetyltransferase from Vibrio cholerae in complex with acetyl-CoA*' (PDB ID: 4R57). There are 26 CoA-binding residues for the chain A of this protein. SVMpred predicts 24 binding residues and 23 of them are true positives, resulting to 0.907 MCC. TemPred predicts 22 binding residues based on 40 homologous templates. Among the predicted binding residues, 21 and 1 are true positives and false positives, respectively. This represents to the MCC, Pre and Rec of 0.859, 0.955 and 0.808, respectively. Further investigation on the predicted binding residues by SVMpred and TemPred shows that the former has two correctly predicted binding residues (30Y and 31W) that are missed by the latter (see the Supplementary Fig. S3 or compare Fig. 3A and B). The consensus of both methods by CoABind increases the number of true positives to 25, improving the MCC and recall to 0.933 and 0.962, respectively.

3.5 Comparison with random prediction and S-SITE

Since no specific tool is available for the identification of CoA-binding residues, the proposed methods are first compared with random predictions. By definition, the average MCC and AUC values for random prediction are around 0 and 0.5, respectively, which are much worse than the corresponding results of the proposed methods.

In addition, we compared our methods with COACH, one of the state-of-the-art general-purpose algorithms for protein-ligand binding residues prediction (Yang *et al.*, 2013a). The standalone version of COACH is available in the I-TASSER Suite (Yang *et al.*, 2015). Here, the S-SITE program was compared because it is the only sequence-based method in the package and was reported to have comparable performance to the other two structure-based programs (TM-SITE and COFACTOR). The same condition was applied when running S-SITE, i.e. by excluding templates with 25% sequence identity to the query sequence. As S-SITE does not differentiate between ligand types, the predictions for ligands of CoA and CoA-derivatives were used in the comparison.

Figure 4 shows the comparison based on the MCC and AUC values on the test set TE73. The results for other metrics are available in the Supplementary Table S4. It shows that the MCC and AUC

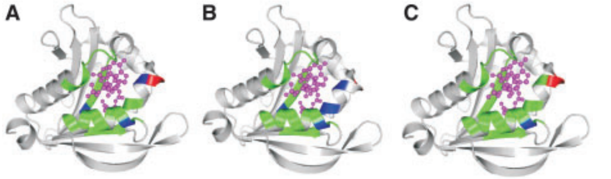

Fig. 3. An example of the predicted CoA-binding residues by (**A**) SVMpred, (**B**) TemPred, (**C**) CoABind. The protein structure is shown in gray cartoon. The ligand structure is shown in magenta ball-and-stick. True positives, false positives, and false negatives are shown in green, red, and blue cartoon, respectively

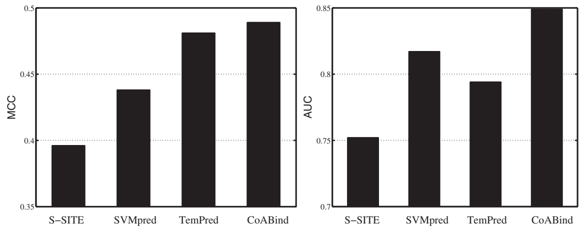

Fig. 4. Comparison of the proposed methods with S-SITE on the test set TE73

values for the proposed methods are all significantly higher than S-SITE. Interestingly, the MCC for TemPred is higher than S-SITE by 21.5%, though both methods make predictions based on templates. This is probably because S-SITE is a general-purpose algorithm while TemPred is specially designed for CoA and CoA-derivatives. Statistical tests were performed to compare the significance level of the improvement over S-SITE. The *P*-values are listed in the Supplementary Table S2 (the data with yellow background), which are all smaller than 10^{-4}, showing that the improvements made by the proposed methods over S-SITE are very significant.

4 Discussion

The performance of the proposed methods may be affected by the following factors: the template identification algorithm, the enriched binding annotation, the inclusion of CoA-derivatives, the size of training set and the random division for training and test sets. As it is time-consuming to re-train SVMpred, we discuss the influence of these factors to the method TemPred only.

4.1 Impact of template identification algorithm

We have applied the sequence alignment algorithm HHsearch to build the TemPred method. We replaced HHsearch by PSI-BLAST to test the impact of template identification algorithm to TemPred. Similar to HHsearch, the prediction of binding residues from the PSI-BLAST search was also based on Equation (1). As PSI-BLAST does not return the probability value Prob_j for the *j*th template, we calculated it based on the corresponding *e*-value. It equals to 1 when the *e*-value is < 0.001 and 0.001/*e*-value, otherwise. The results are shown in the Supplementary Table S5 (the row $\text{TemPred}^{\text{P}}$), which shows that HHsearch can find homologies for 26% more queries than PSI-BLAST with *e*-value ≤ 0.001. For all metrics, HHsearch obtains higher values than PSI-BLAST. The statistical test shows that HHsearch outperforms PSI-BLAST significantly at *P*-value $< 0^{-7}$. These results are in consistent with the fact that HHsearch is more sensitive for the homology detection (Soding, 2005). We thus employed HHsearch to develop the template-based method TemPred.

4.2 Impact of the enriched-binding annotation

When designing the benchmark dataset, we transferred the CoA-binding residues from identical sequences to enrich experimental annotations. We test if this transferal contributes to the prediction accuracy. To this end, we assessed TemPred on the original dataset without such transferal. The results are listed in the Supplementary Table S5 (the row TemPred*), which shows that the accuracy for the dataset with transferal is only slightly higher than the original dataset. The *P*-value (0.055) from the statistical test also indicates that the improvement with transferal of binding annotation is not significant at significance level of 0.05. There may be two reasons for this. The first is the number of increased binding annotations may not be enough (about 14%), as 100% sequence identity has been required to make the transferal. The second is the number of proteins used for training is relatively small (146).

4.3 Impact of the CoA-derivatives and training set size

As the total number of CoA-binding proteins is too small, we have included the CoA-derivatives to enlarge the benchmark dataset. We tested how much this contributes to the prediction. For comparison, we removed the proteins with CoA-derivatives from the training and the test sets. This results to the new training and test sets of 80 and 42 proteins, respectively. The TemPred results on this 42 proteins are available in the Supplementary Table S5 (the row $\text{TemPred}^{\text{T}}$), which are all lower than on TE73. For example, the MCC is 0.419, which is 14.8% lower than that on TE73. This suggests that inclusion of related ligand types does contribute to the prediction accuracy. Another reason for the difference is the size of training set. For the one with CoA-derivatives, the number of training proteins is 146, which is 1.8 times of the one without CoA-derivatives (80).

4.4 Impact of the random division for training and test sets

As we can see from the data in Figure 2 and Table 1, the MCC for SVMpred on the training set is lower than the test set (0.392 versus 0.438). It is also similar for TemPred, which has MCC of 0.447 and 0.481 on the training and test set, respectively. One question one may come up with is if the test set happens to be easier than the training set. We tested this by trying 10 more random divisions of the dataset. For each division, we ran TemPred and collected the average accuracy. The results over the 10 divisions are presented in the Supplementary Table S6, which are very similar to the one presented in Table 1. It shows that the average MCC on the training sets is lower than the test sets (0.423 versus 0.485). This suggests that the higher accuracy on the test set is not because the test set is not easier than the training set. Note that the accuracy on the training set was obtained based on 5-fold cross validation. As a result, for each protein in the training set, only 4/5 proteins (i.e. 116) are in the template library. However, for each protein in the test set, the number of proteins in the template library is 146 (i.e. the size of the training set). Therefore, we conclude that it is because more templates are used for the test set than for the training set.

5 Conclusions

CoA-protein binding plays an important role in various cellular functions and metabolic pathways. We have developed three methods for the prediction of CoA- and CoA derivatives-binding residues. The first is an accurate template-based method TemPred by employing the sensitive profile–profile alignment algorithm HHsearch for template detection. The second is an *ab-initio* predictor SVMpred with high predictive quality. The success of SVMpred is attributed to the design of a comprehensive set of features from two sequence profiles and the predicted secondary structure and solvent accessibility. SVMpred is experimentally shown to be complementary to TemPred. We thus combined them together and developed the third consensus-based method CoABind. It significantly outperforms both SVMpred and TemPred, achieving MCC and AUC of 0.489 and 0.849, respectively, on an independent test set of 73 proteins. Experiments demonstrate that the proposed methods significantly outperform the state-of-the-art general-purpose ligand-binding residues prediction algorithm COACH. A web server implementing the proposed methods is freely available at: http://yanglab.nankai.edu.cn/CoABind.

Funding

The work was supported in part by National Natural Science Foundation of China [NSFC 11501407 and 11501306], Fok Ying-Tong Education Foundation [161003] and the Thousand Youth Talents Plan of China.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.

Burton,M. *et al.* (2005) Evolution of the acyl-CoA binding protein (ACBP). *Biochem. J.*, **392**, 299–307.

Chen,K. and Kurgan,L. (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, **23**, 2843–2850.

Engel,C. and Wierenga,R. (1996) The diverse world of coenzyme A binding proteins. *Curr. Opin. Struct. Biol.*, **6**, 790–797.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsletter*, **11**, 10–18.

Heffernan,R. *et al.* (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, **33**, 2842–2849.

Huang,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Lipmann,F. (1945) Acetylation of sulfanilamide by liver homogenates and extracts. *J. Biol. Chem.*, **160**, 173–190.

Meng,F. and Kurgan,L. (2016) DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, **32**, i341–i350.

Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Robinson,A.B., and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. USA*, **88**, 8880–8884.

Rose,P.W. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.

Shampo,M.A. and Kyle,R.A. (2000) Fritz Lipmann–Nobel Prize in discovery of coenzyme A. *Mayo Clin. Proc.*, **75**, 30.

Sibon,O.C. and Strauss,E. (2016) Coenzyme A: to make it or uptake it? *Nat. Rev. Mol. Cell Biol.*, **17**, 605–606.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Taherzadeh,G. *et al.* (2016) Sequence-based prediction of protein-peptide binding sites using support vector machine. *J. Comput. Chem.*, **37**, 1223–1229.

Xia,J. *et al.* (2017) An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. *Bioinformatics*, **33**, 863–870.

Yan,J. *et al.* (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief. Bioinformatics*, **17**, 88–105.

Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.*, **45**, e84.

Yang,J. *et al.* (2013a) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

Yang,J. *et al.* (2013b) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.

Yang,J. *et al.* (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.

Zhang,T. *et al.* (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.*, **11**, 609–628.