Wang et al. BMC Bioinformatics 2019, 20(Suppl 25):699
https://doi.org/10.1186/s12859-019-3274-7

BMC Bioinformatics

RESEARCH Open Access

Semi-supervised prediction of protein interaction sites from unlabeled sample information

Ye Wang[1†], Changqing Mei[1†], Yuming Zhou[1], Yan Wang[1], Chunhou Zheng[2], Xiao Zhen[3], Yan Xiong[4], Peng Chen[5*], Jun Zhang[6] and Bing Wang[1,2*]

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: The recognition of protein interaction sites is of great significance in many biological processes, signaling pathways and drug designs. However, most sites on protein sequences cannot be defined as interface or non-interface sites because only a small part of protein interactions had been identified, which will cause the lack of prediction accuracy and generalization ability of predictors in protein interaction sites prediction. Therefore, it is necessary to effectively improve prediction performance of protein interaction sites using large amounts of unlabeled data together with small amounts of labeled data and background knowledge today.

Results: In this work, three semi-supervised support vector machine–based methods are proposed to improve the performance in the protein interaction sites prediction, in which the information of unlabeled protein sites can be involved. Herein, five features related with the evolutionary conservation of amino acids are extracted from HSSP database and Consurf Sever, i.e., residue spatial sequence spectrum, residue sequence information entropy and relative entropy, residue sequence conserved weight and residual Base evolution rate, to represent the residues within the protein sequence. Then three predictors are built for identifying the interface residues from protein surface using three types of semi-supervised support vector machine algorithms.

Conclusion: The experimental results demonstrated that the semi-supervised approaches can effectively improve prediction performance of protein interaction sites when unlabeled information is involved into the predictors and one of them can achieve the best prediction performance, i.e., the accuracy of 70.7%, the sensitivity of 62.67% and the specificity of 78.72%, respectively. With comparison to the existing studies, the semi-supervised models show the improvement of the predication performance.

Keywords: Protein interaction site, Unlabeled information, Conservative feature, Semi-supervised support vector machine

* Correspondence: pchen@ahu.edu.cn; wangbing@ustc.edu
[5]Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China
[1]School of Electrical and Information Engineering, Anhui University of Technology, Maanshan 243002, Anhui, China
Full list of author information is available at the end of the article

Background

Protein-protein interactions (PPIs) are involved in various life activities, such as metabolism and signal transduction, gene transcription, protein translation, modification and localization, and are also closely related to disease production [1–9]. However, PPI varies from cell to cell and from time to time, which poses a challenge to the studies of them.

Due to the rapid development of machine learning methods, many classical methods, such as Bayesian, support vector machine (SVM), and artificial neural networks, have been used to predict protein interaction sites [10–19]. Sprinzak et al. used the correlated sequence-signatures as identifiers for the interacting protein which can significantly reduce the search space and implement a directional experiment interactive screen and achieved high quality experimental results [5]. Bock et al. proposed a phylogenetic bootstrapping algorithm which suggests traversal of a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of protein interactions in genetically-similar organisms [6]. Enright et al. developed the hydrophobic free energy functions with the fusion detection based on sequence analysis and trained a SVM learning system to recognize and predict interactions based solely on primary structure and associated physicochemical properties, and the overall performance of the classifier has been significantly improved [20]. Chen et al. proposed a radial basis function neural networks optimized by the particle swarm optimization algorithm to predict protein interaction sites [2]. Wang et al. presented a SVM based algorithm to identify protein-protein interactions sites on the residues level by incorporating residues spatial sequence profile and evolution rate [21]. Wang et al. in another work implemented a dataset reconstruction strategy by using manifold learning under a hypothesis that the interaction and non-interaction sites have different inherent structure manifolds [13, 22]. Although these methods have driven advances in PPI research, there is still a problem that a lot of interactions cannot be tagged from experiments, and only a small part of labeled samples can be used for model training in the prediction of PPI sites, which will make it difficult for the well-trained learning systems to have strong generalization ability [23].

Therefore, this paper proposed three semi-supervised machine learning-based computational models to address the problem that the information of a large number of unlabeled samples can be utilized effectively to improve the performance of protein interaction site prediction when only a few of labeled samples can be available. Firstly, five evolutionary conserved features of amino acids based on multiple sequence alignments are extracted, i.e., the spatial sequence spectrum of residues, sequence information entropy, relative entropy, conservative weight and residue evolution rate. Then three semi-supervised learning methods are proposed, i.e., the self-balancing semi-supervised support vector machine based on multi-core learning (Means3vm-mkl), the iterative-based label average self-training semi-supervised support vector machine (Means3vm-iter) and the safe semi-supervised support vector machine (S4VM), to build the prediction model for identification of protein interaction sites [24–26]. The experimental results demonstrated the superiority of our proposed methods, such as the prediction accuracy of 0.707 for S4VM model, with comparison to the existing supervised and other approaches.

Results

In this work, three semi-supervised SVM algorithms, i.e., Means3vm-mkl, Means3vm-mkl, and S4VM, have been applied for the prediction of protein interaction sites from protein sequences. Compared to the traditional supervised SVM, semi-supervised models can effectively use the information from both of labeled and unlabeled samples. A popular software of support vector classification, Libsvm, is adopted in this work, where the empirically optimal parameters are used, such as C_1 is 1, C_2 equals 0.1, and the maximum number of generations is 50. To validate the effectiveness of the proposed models, a 5-fold cross-validation technique, and an original residue data set with 91 protein chains are used to evaluate the prediction performance of the proposed models. Herein, 2299 interface residues drawn from the definition of interaction sites can be used for the construction of the three semi-supervised SVM models.

It can be seen from Fig. 1 that the proposed three semi-supervised methods can classify the protein interaction and non-interaction sites on protein sequences. Means3vm-iter predictor can get good prediction measures, i.e., 0.636 of accuracy, 0.589 of F-measure, 0.67 of precision, 0.745 of specificity, 0.526 of sensitivity, and 0.278 of MCC, from the original protein residue data set D. Compared to Means3vm-iter method, the Means3vm-mkl based-model shows better prediction performance, where the accuracy rate and F-measure are increased by 3%. The reason is that multi-core learning can utilizes the feature mapping capabilities of each basic kernel, and the data is better expressed in the combined feature space constructed by multiple feature spaces, which can significantly improve the classification accuracy. But the process of multi-core learning is complex, and the time required is relatively longer than the method based on iterative optimization. Means3vm-iter transforms the optimization problem into a quadratic programming which and is, thus, easily and quickly solved by standard programs, although it may fall into local minimum, and the classification accuracy is slightly lower than the Means3vm-mkl based-model.

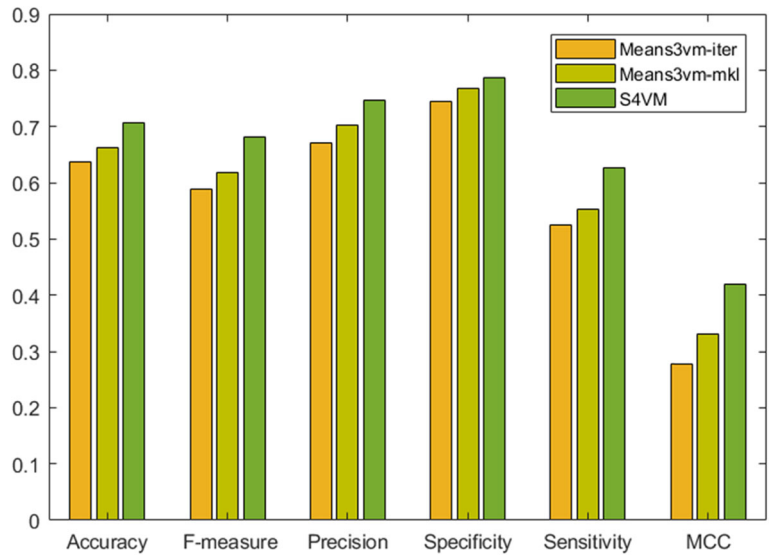

Fig. 1 Classification performance evaluation of three Semi-supervised methods on datasets

It also can be found that the S4VM method can achieve the best prediction performance among the three semi-supervised models in all of the seven predictor measures, i.e., 0.707 of accuracy, 0.787 of specificity, 0.627 of sensitivity, 0.746 of precision, 0.681 of F measure and 0.419 of MCC. The overall prediction performance of S4VM is improved by more than 4% compared to Means3vm-iter and Means3vm-mkl methods. S4VM attempts to consider all possible low-density boundaries to effectively prevent performance degradation, and therefore it can deduce the false negative rate and false positive rate in prediction, which can be confirmed by the relatively high values of 0.787 and 0.627 for specificity and sensitivity, respectively.

To assess the generalization ability of prediction models, a five cross-validation strategy is adopted within the predictors' construction. For each performance indicator, the differences between its value in each run and mean value in all of five cross-validation times are calculated. It can be observed from Fig. 2 that the three semi-supervised algorithms-based predictors are robust, and

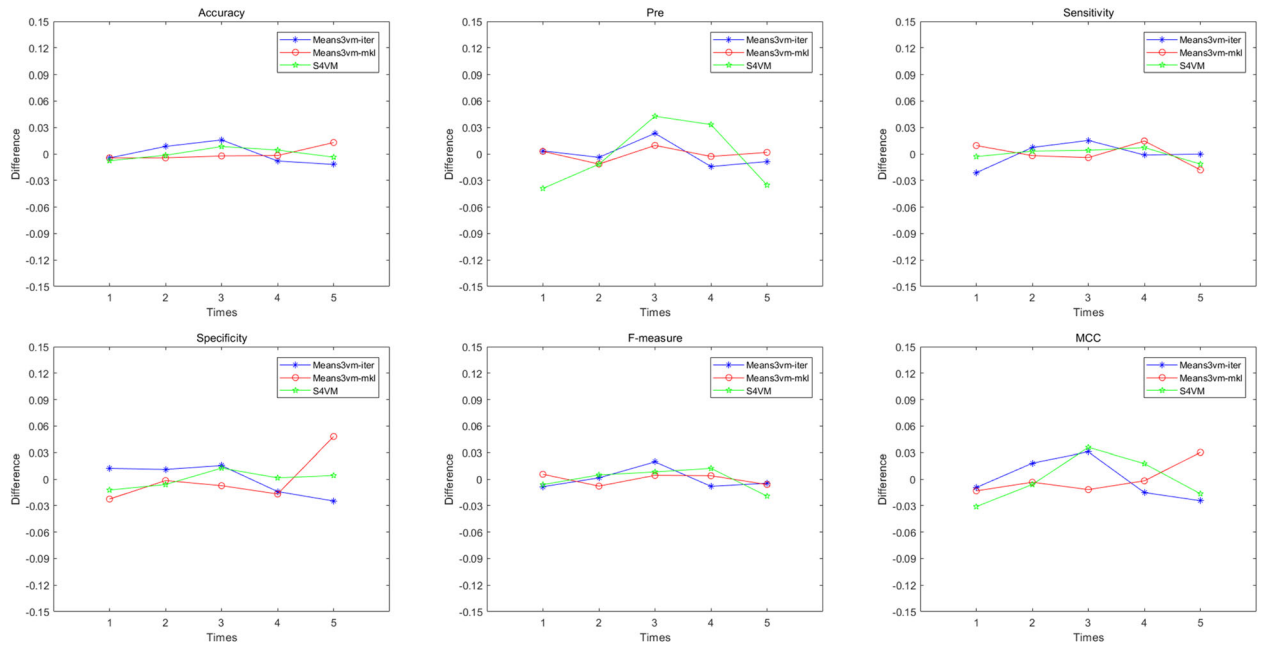

Fig. 2 Prediction performance measures in 5 repetitions of cross-validation

most of their fluctuation range is less than 0.03, which indicates that the proposed models have excellent generalization ability when new samples are introduced. Among them, the biggest difference of accuracy is Means3vm-iter model which has value of 0.015, precision in S4VM model with 0.043, sensitivity in Means3vm-iter model with 0.021, F-measure in Means3vm-iter model with 0.02, MCC in S4VM model with 0.036 and only Means3vm-mkl has a specificity of 0.048. Among the three predictors, S4VM performs best, and the mean value of difference is only 0.005 for accuracy, 0.032 for precision, 0.006 for sensitivity, 0.007 for specificity, 0.01 for F-measure and 0.021 for MCC.

Discussion

It can be found that the three semi-supervised models proposed in this work can predict protein-protein interaction sites based on the features. To further evaluate the effectiveness of these models, the results of some previous works had been used to compare prediction performance.

Prediction performance comparison between supervised and semi-supervised SVM

Most studies adopted supervised machine learning algorithms to predict interaction sites from protein sequences or structures in previous works, and some of them have to use data sampling technologies to balance the number of positives and negatives to void the prediction bias [27–30]. Instead of semi-supervised methods where all of samples in model training are labeled, semi-supervised machine learning approaches are trying to learning from labeled and unlabeled samples, which can effectively make use more information for learning, which is very significant for the studies of protein interaction where many of them are still unknown.

To evaluate the value of unlabeled sample information in protein interface residues, the traditional supervised SVM algorithm is also directly used to make predictions of protein interaction sites, and its result is shown in Fig. 3. Based on the same dataset, it can be seen that the predictive performance of supervised SVM-based predictor is much lower than that of semi-supervised based one, and the accuracy is only 0.586, which is 0.12 lower than that of S4VM. On other measures, the proposed semi-supervised models also outperform the supervised predictor, i.e., the F-measure is only 0.56, MCC is only0.23, precision is only 0.598, sensitivity is only 0.529 and specificity is only 0.643.These results suggest that unlabeled sample information, when used in conjunction with a small data set of labeled data, can get much improvement in learning accuracy, which is important for the current situation that many protein interactions are not identified by experiments.

Comparison with other approaches

In this work, five features related to amino acid conservation are extracted from protein sequence for identification of protein interface residues from protein surface. To validate the effectiveness of the extracted features in discrimination between interface and non-interface residues, the comparison with a previous study based on the same data set by Li and Kuo has been taken. Compared to the evolutionary conservation features used in this work, they predicted protein interaction sites using five sequence features. Figure 4 shows that both evolutionary and sequence features can successfully identify protein

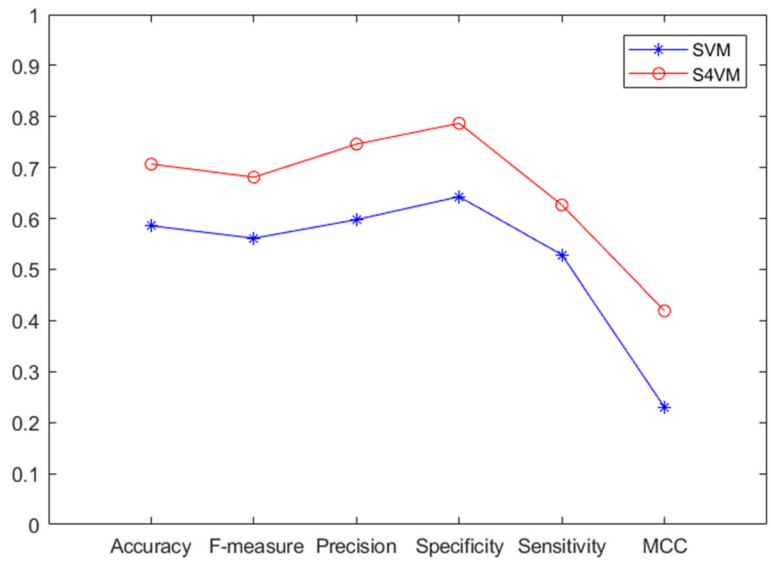

Fig. 3 Comparison of experimental results between SVM and S4VM

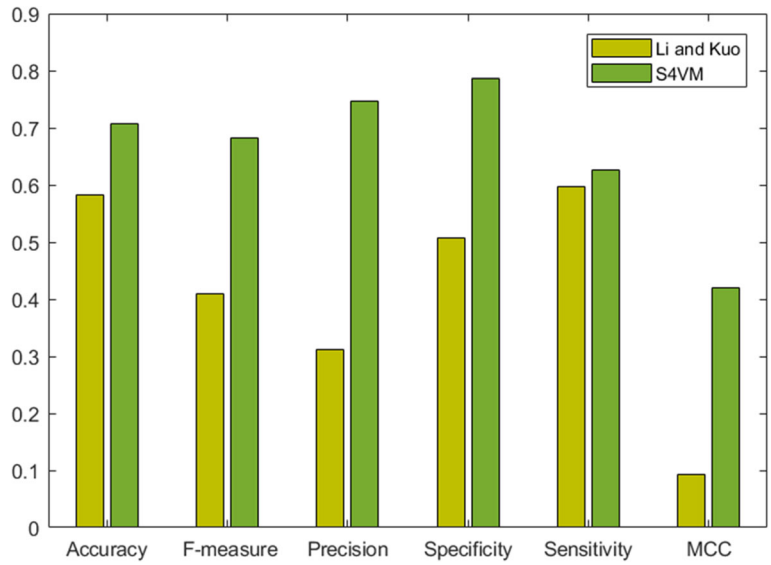

Fig. 4 Compared with the former's evaluation performance

interaction sites, but evolutionarily conserved features show stronger classification capacity than that of sequence features. It can be found that our proposed method can produce more accurate prediction than sequence features-based model did, i.e., 0.124 higher in accuracy, 0.03 in sensitivity, 0.279 in specificity, 0.272 in F-measure and 0.326 in MCC. Especially, the value of precision measure has a 0.435 higher than that in Li and Kuo's work, which means the false positive rate of prediction deduced dramatically, and the features in this work are really sound in discrimination between protein interaction and non-interaction sites.

Visualization of experimental results

To further validate the predictions achieved by our proposed semi-supervised model, a test on one chain of protein complex 1A4Y was taken as an example. We use the molecular visualization tool - Pymol to show our predictions. Figure 5 shows the protein chain 1A4Y_A data set and the results obtained under three semi-supervised models. In D, E, and F, there are 218 balls that represent the surface residues involved in the prediction. Green balls, red balls, yellow balls and blue balls represent the number of TP, TN, FP and FN, respectively. The numbers in details on the complex 1A4Y can be found in Table 1. Our approach improves overall predictive performance and reduces false positives, and S4VM methods perform well. Only 5.2% of the interface residue in the S4VM method was not predicted.

Conclusion

This paper proposed a semi-supervised learning strategy for protein interaction site prediction. Firstly, a non-redundancy dataset with 91 protein chains were selected, and five evolutionary conserved features were extracted for the vectorization of each amino acid residue from the common databases and servers. Then three semi-supervised learning methods, Means3vm-mkl, Means3vm-iter and S4VM are proposed to identity interaction sites from surfaces of protein complexes. The experimental results show that the Means3vm-mkl and S4VM methods have excellent classification results a five-fold cross-validation is used, and the accuracy is 0.662 and 0.707, respectively. The S4VM method can achieve the best overall performance, such as the highest value 0.419 for MCC and 0.681 for F-measured value. By comparison with supervised model and other studies, this work produced much improvement in protein interaction sites prediction, which suggests that the effectiveness of the proposed semi-supervised strategies in discrimination of protein interaction and non-interaction sites. Furthermore, the experimental results demonstrated that residues in protein surface, even its interface label cannot be tagged yet, contain a lot of information of protein interactions, which is important for understanding cellular activity and drug design.

Methods

Dataset

The dataset used in this work are come from a previous work investigated by Ansari and Helms et al., where 170 pairs of transient protein interactions has been collected [31]. Protein chains of less than 50 residues and some outdated small family protein chains are discarded to make the data more representative. If there is a plurality of interacting partners for the same chain, the partner chain with most interfacial residues is represented. The

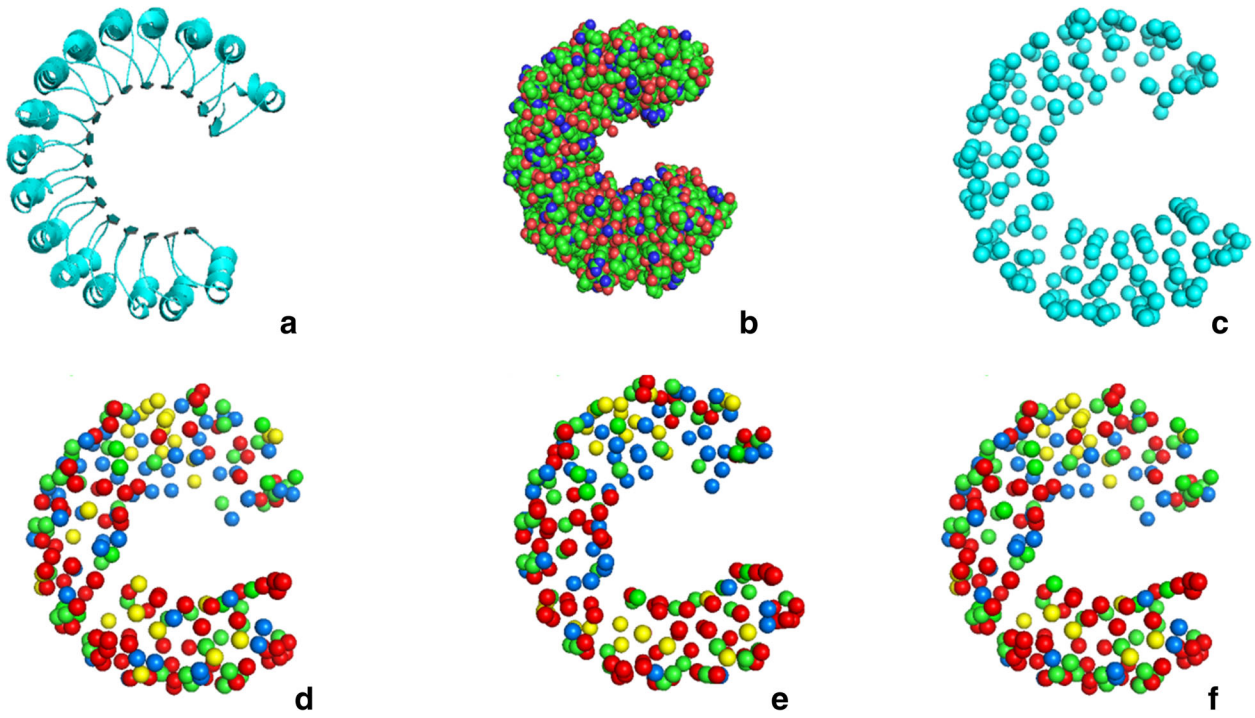

Fig. 5 Experimental visualization results. **a** represents the protein chain 1A4Y_A(**a** and **b**) is its spherical representation. **c** is the 1A4Y protein chain after extraction of surface residues. **d**, **e** and **f** show the predicted results of Means3vm-mkl, Means3vm-iter and S4VM, and the green balls, red balls, yellow balls and blue balls represent the number of TP, TN, FP and FN, respectively

BLASTCLUST program was used to exclude protein chains with sequence similarity greater than 30%, and finally 91 non-redundant protein chains has been remained for this study, which can be found in Table 2 [15, 32, 33].

The definitions of the residues are same as what Fariselli et al. did in their work [30]. Surface residues are defined if the relative accessible surface area residue is bigger than 16% of the maximum accessible surface area for each type of amino acid. Among the surface residues, a residue can be defined as interface residues if the distance between its alpha carbon atom and that of any residues in the interaction chain is less than 1.2 nm, otherwise it will be categorized as non-interface residues. Based on the above definitions, the original residue data set D is composed of 2299 interface residues and 8131 non-interface residues which are obtained from the 91 protein chains used in this work.

Table 1 The number of predictions in TP, TN, FP and FN

	Samples	Results			
		TP	TN	FP	FN
Means3vm-iter	218	57	82	28	51
Means3vm-mkl		60	84	26	48
S4VM		68	87	23	40

Feature extraction

There are many properties of amino acids had been used for protein interactions or interaction sites prediction, among them evolutionary conservation analyses have been widely applied to characterize functionally/structurally important residues because these amino acids in a protein sequence are conserved through selective evolutionary pressure [20, 34–36]. In this work, five evolutionary conservation relevant features are extracted for protein interaction sites prediction, where residue spatial sequence, sequence information entropy, relative entropy and residue sequence weight are extracted from the HSSP database, and evolutionary rate residues are extracted from Consurf Serve.

The spatial sequence profile of amino acid residues, a feature widely used in protein related studies, represents the frequency of various amino acids at a given residue position in the primary structure of proteins. Protein residue sequence entropy is based on Shannon's information theory to estimate the conservation score of sequence variability. Relative entropy is the normalized sequence information entropy. The conserved weight of the residue sequence is a calculation of position conservativeness of the protein sequence. The evolution rate of residues can be traded off from a statistical point of view, considering the linkages generated by the system in the stochastic process of sequence and evolution and

Table 2 The protein chains used in this work

1AY7_A	1B6C_A	1B7Y_B	1AZS_B	1B7Y_A	1AVG_H	1AZS_C	1B6C_B
1UDI_E	1UGH_E	1ZBD_A	1UEA_A	1UUZ_A	1TCO_A	3TGI_I	1WQ1_G
1HLU_P	1IRA_Y	1KKL_A	1HWH_B	1JSU_C	1HLU_A	1IRA_X	1ITB_A
1BDJ_B	1BMQ_A	1BRB_I	1BGX_T	1BP3_A	1BDJ_A	1BI7_A	1BMQ_B
1QBK_B	1SMP_A	7CEI_A	1QBK_C	1STF_E	1PYT_B	1SGP_E	1SMP_I
1FLT_Y	1GLA_F	1HJA_C	1GFW_A	4SGB_I	1FLT_V	1GFW_B	1GLA_G
1ABR_A	1AHW_C	1ATN_D	1ABR_B	1AK4_D	1A4Y_A	1ACB_I	1AK4_A
1BVK_A	1CA0_B	1D4V_B	1BVK_C	1D4V_A	1BRS_A	1BVN_P	1CXZ_B
2KAI_B	2SIC_I	3SGB_I	2PCC_A	2TEC_E	1ZBD_B	2PCC_B	2SNI_I
1DAN_U	1E9H_B	1FAP_B	1DFJ_E	1ETH_A	1DAN_L	1E96_A	1EFU_B
1L0Y_A	1NOC_B	1PYT_A	1L0Y_B	1PDK_B	1KKL_H	1MAH_A	1PDK_A
1GUA_B	1STF_I	1UEA_B					

the maximum likelihood estimate of the evolution rate can be accessed using the Rate4Site algorithm to calculate the conservation of each amino acid position [37]. For each residue, 20 dimensions for protein sequence profile and one dimension for each of other four features can be extracted in this work.

As many previous studies did, a slide-window strategy is also adopted in this work to consider the interface information, which is formed by the target residues centered with 10 spatially closest ones. Therefore, each target residue can be represented by a 264-dimensional vector and used for subsequent prediction construction.

Semi-supervised models

In semi-supervised learning, the labeled sample set is $\{x_1, \ldots, x_l\}\{x_1, \ldots, x_l\}$, and the unlabeled sample set is $\{x_{l+1}, \ldots, x_{l+u}\}$, where l and u are the number of labeled and unlabeled samples, respectively, $y_i = \{+1, -1\}$. The labels of labeled and unlabeled sample set can noted as $I_l = \{1, \ldots, l\}$, and $I_u = \{l+1, l+2, \ldots, l+u\}$. As one of the most popular semi-supervised learning methods, Semi-supervised support vector machines (S3VM) attempts to standardize and adjust decision boundaries by exploring unlabeled data based on clustering assumptions, whose illustration can be found in Fig. 6 [26, 38]. The meanS3VM, a fast S3VM algorithm, estimates the category average of the unlabeled data, so that the classification performance is very similar to the supervised SVM. The core idea of meanS3VM algorithm is to maximize the interval between the class averages of the two categories of samples, and thus the goal is to find the decision function $f(x) = w' \varnothing(x) + b$ to minimize.

$$\min_{d\in\Delta} \min_{w,b,\rho,\xi} \frac{1}{2}\|w\|^2 + c_1\sum_{i=1}^{l}\xi_i - c_2\rho \tag{1}$$

$$s.t. y_i\left(w'\varnothing(x_i) + b\right) \geq 1-\xi_i, \xi_i \geq 0, i = 1, \ldots, l,$$

$$\frac{1}{u_+}\left(w'\sum_{j=l+1}^{l+u} d_{j-l}\varnothing(x_j)\right) + b \geq \rho$$

$$\frac{1}{u_-}\left(w'\sum_{j=l+1}^{l+u}(1-d_{j-l})\varnothing(x_j)\right) + b \leq -\rho$$

$$\sum_{i\in I_u} sgn\left(w'\varnothing(x_i) + b\right) = r$$

Herein, the last constraint is an equilibrium constraint that avoids assigning all unlabeled samples to the same category, r is a user-defined parameter, and $\Delta = \{d | d_i \in \{0, 1\}, \sum_{i=1}^{u} d_i = u_+\}$, $u_+ = \frac{r+u}{2}$, $u_- = \frac{-r+u}{2}$. Since the bilinear constrains between w and b, this formula is non-convex, and the two algorithms can be used to solve it. The first one is based on multiple kernels learning (MeanS3VM_mkl), while the second one is based on alternating optimization (MeanS3VM_iter) [24].

1) MeanS3VM_mkl

Mathematically, the goal of S3VM can be solved with the dual form:

$$min_{d\in\Delta} max_{\alpha\in A} \alpha' \tilde{l} - \frac{1}{2}(\alpha \bullet \tilde{y})' K^d (\alpha \bullet \tilde{y}) \tag{2}$$

which can be expressed in the form of a multicore learning optimization problem:

$$min_{\mu\in M} max_{\alpha\in A} \alpha' \tilde{I} - \frac{1}{2}(\alpha \bullet \tilde{y})' \left(\Sigma_{t:dt\in\Delta} \mu t K^{dt}\right)(\alpha \bullet \tilde{y}) \tag{3}$$

where $M = \{\mu | \Sigma u_t = 1, u_t \geq 0\}$,

$$\alpha = [\alpha_1, \ldots, \alpha_{l+2}]' \epsilon R^{l+2}$$

$$\tilde{I} = [I_1, \ldots, I_t, 0, 0]' \epsilon R^{l+2}$$

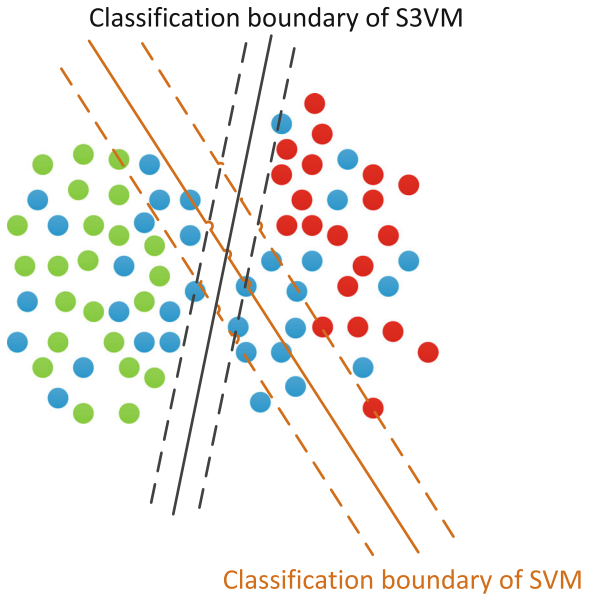

Fig. 6 Illustration of different classification boundaries of SVM which considers only labeled data, and S3VM which considers labeled and unlabeled data, where green balls denote positives, red ones are negatives, and blue ones are unlabeled samples

$$\tilde{y} = [y_1, \ldots, y_t, 1, -1]' \in R^{l+2}$$

$$A = \{\alpha \mid \sum_{i=1}^{l+2} \alpha_i \widetilde{y_i} = 0, \sum_{i=1}^{l+2} \alpha_i = c_2; 0 \le \alpha_i \le c_1,$$

$$\forall i = 1, \ldots, l; 0 \le \alpha_l + 1, \alpha_l + 2 \le c_2\}$$

The nuclear matrix $K^d \in R^{(l+2)*(l+2)}$, the element is $K_{ij}^d = (\varnothing_i^d)'(\varnothing_i^d)$.

$$\varnothing_i^d = \begin{cases} \frac{1}{u_+} \sum_{j=l+1}^{l+u} d_j - l\varnothing(x_i) i = l+1 \\ \varnothing(x_i) i = 1, \ldots, l \\ \frac{1}{u_-} \sum_{-j=l+1}^{l+u} (1 - d_j - l)\varnothing(x_i) \; i = l+2 \end{cases} \quad (4)$$

Since all $d_t \in \Delta$ are to be minimized and there will be a large number of reasonable d_t, the cut plane algorithm is adopted to solve the above problem and find the optimal label d vector of the unlabeled samples, whose details can be found in references [17, 24].

2) *Means3vm-iter*

Another way to solve the problem of MeanS3VM is to alternate optimization, which can be abbreviated as:

$$\max_{d \in \Delta, \rho} \rho \quad (5)$$

$$s.t. \frac{1}{u_+} \left(w' \sum_{j=l+1}^{l+u} d_j - l\varnothing(x_j) \right) + b \ge \rho$$

$$\frac{1}{u_-} \left(w' \sum_{j=l+1}^{l+u} (1 - d_j - l)\varnothing(x_j) \right) + b \le -\rho$$

In the optimization process, if $f(x_i) > f(x_j)$, $\forall i, j \in I_u$, then $d_{i-1} \ge d_{j-1}$, which has been proved [11]. Assigning labels to unlabeled samples based on predicted values using this theorem, which can ensure that the label d vector obtained each time is better than the previous one [17, 24].

S4VM

Given a large amount of unlabeled samples in data set, there may be multiple "intervals" of low-density boundaries, and it is difficult to determine which one is the best. Although these low-density boundaries and the number of labeled samples are limited, due to the large differences, there will be a large loss if the selection is wrong, resulting in performance degradation, even worse than using only labeled samples, which limits the use of semi-supervised learning methods in certain key areas.

S4VM has been improved on traditional S3VM. The difference between S4VM and S3VM is that S3VM tries to focus on the best low-density boundary, while S4VM focuses on multiple possible low-density boundaries. The main idea is to optimize without giving many different "interval" boundaries. Class division of labeled samples. This maximizes the performance improvement of the support vector machine over the worst case labeled samples. The specific practices are as follows:

$h(f, \hat{y})$ is the objective function to be optimized by S3VM.

$$h(f, \hat{y}) = \frac{\|f\|_H}{2} + C_1 \sum_{i=1}^{l} l(y_i, f(x_i)) + C_2 \sum_{j=1}^{u} l\left(\widehat{y_j}, f\left(\widehat{x_j}\right)\right) \quad (6)$$

The goal is to find multiple low-density boundary lines $\{f_t\}_{t=1}^{T}$ with "intervals" and the corresponding category division $\{\widehat{y_t}\}_{t=1}^{T}$ so that the following functions are minimized.

$$\min_{\left\{f_t, \widehat{y_t} \in \beta\right\}_{t=1}^{T}} \sum_{t=1}^{T} h(f_t, \widehat{y_t}) + M\Omega\left(\{\widehat{y_t}\}_{t=1}^{T}\right) \quad (7)$$

where T is the number of dividing lines, and Ω is a penalty function that measures the differentiation of the dividing line. Various functions can be used in the implementation. M is a large constant used to ensure the difference. Obviously, minimizing (7) can ensure the difference of the boundary line and the large interval.

Without loss of generality, we assume that f is a linear function, f(x) = w ∅ (x) + b. The optimization problem that needs to be solved is expressed as.

$$\min_{\left\{w_t,b_t,\widehat{y_t}\in\beta\right\}_{t=1}^{T}} \sum_{t=1}^{T}\left(\frac{1}{2}\|w_t\|^2 + c_1\sum_{i=1}^{l}\xi_i + c_2\sum_{j=1}^{u}\widehat{\xi}_j\right) + M\Omega\left(\{\widehat{y}_t\}_{t=1}^{T}\right) \tag{8}$$

$$s.t.y_i\left(w_t'\varnothing(x_i) + b_t\right) \geq 1-\xi_i, \xi_i \geq 0$$

$$\widehat{y_{t,j}}\left(w_t'\varnothing(\widehat{x_j}) + b_t\right) \geq 1-\widehat{\xi}_j, \widehat{\xi}_j \geq 0$$

$$\forall i = 1, \ldots, l, \forall j = 1, \ldots, u, \forall t = 1, \ldots, T$$

Then use an efficient sampling search strategy to solve (8). First, through the local search, find multiple large margin low-density separators. The k-means clustering algorithm is then used to identify representative splitters with a large variety of diversity [16].

Evaluation criteria

In addition to the accuracy, precision, sensitivity, and specificity which often used to evaluate predicted performance, F-measure and Mathew's Correlation Coefficient (MCC) values are introduced. F-measure is a weighted harmonic average of recalls and precision, often used to evaluate classification models, and MCC is an effective measure in imbalanced data classification.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$Specificity = \frac{TN}{FP + TN} \tag{14}$$

$$F\text{–}measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{15}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{16}$$

where TP, FP, TN and FN represent the number of true positives (correctly predicted interface residues), the number of false positives (incorrectly predicted interface residues), the number of true negatives (correctly predicted non- interface residues) and the number of false negatives (incorrectly predicted non- interface residues), respectively.

Abbreviations

HSSP: Homology-derived Secondary Structure of Proteins; MCC: Mathew's Correlation Coefficient; PPI: Protein-protein interaction; S3VM: Semi-supervised support vector machine; S4VM: Safe semi-supervised support vector machine

Acknowledgments

None.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 25, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-25.

Authors' contributions

YW (YeWang) and CM conceived of the study; XZ, YW (Yan Wang), YZ, JZ and YX participated in the experiment design; YW (Ye Wang), PC and BW carried it out and drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 61472282, 61672035 and 61872004), Key Program for Educational Commission of Anhui Province of China (No. KJ2019ZD05, KJ2017A041), Co-Innovation Center for Information Supply & Assurance Technology in AHU (ADXXBZ201705), Anhui Province Funds for Excellent Youth Scholars in Colleges (gxyqZD2016068), and Anhui Scientific Research Foundation for Returned Scholars.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Electrical and Information Engineering, Anhui University of Technology, Maanshan 243002, Anhui, China. [2]Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, Anhui, China. [3]School of Computer Science and Technology, Anhui University of Technology, Maanshan 243002, Anhui, China. [4]School of Computer Science and Technology, University of Science & Technology, Hefei 230026, Anhui, China. [5]Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China. [6]College of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China.

Published: 24 December 2019

References

1. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437(7062):1173–8.
2. Chen Y, Xu J, Yang B, Zhao Y, He W. A novel method for prediction of protein interaction sites based on integrated RBF neural networks. Comput Biol Med. 2012;42(4):402–7.
3. Liu Q, Chen P, Wang B, Zhang J. Li J: dbMPIKT: a database of kinetic and thermodynamic mutant protein interactions. BMC Bioinformatics. 2018;19(1):455.
4. Ji Z, Wang B, Yan K, Dong L, Meng G, Shi L. A linear programming computational framework integrates phosphor-proteomics and prior knowledge to predict drug efficacy. BMC Syst Biol. 2017;11(Suppl 7):127.
5. Zhu M, Song X, Chen P, Wang W. Wang B: dbHDPLS: a database of human disease-related protein-ligand structures. Comput Biol Chem. 2019;78:353–8.

6. Yang C, Ge SG. Zheng CH: ndmaSNF: cancer subtype discovery based on integrative framework assisted by network diffusion model. Oncotarget. 2017;8(51):89021–32.
7. Ge SG, Xia J, Sha W, Zheng CH. Cancer subtype discovery based on integrative model of multigenomic data. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(5):1115–21.
8. Chen P, Han K, Li X, Huang DS. Predicting key long-range interaction sites by B-factors. Protein Pept lett. 2008;15(5):478–83.
9. Shen Z, Bao W, Huang DS. Recurrent neural network for predicting transcription factor binding sites. Sci Rep. 2018;8(1):15270.
10. Pan XY, Zhang YN, Shen HB. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. J Proteome Res. 2010;9(10):4992–5001.
11. Xia JF, Wang SL, Lei YK. Computational methods for the prediction of protein-protein interactions. Protein Pept Lett. 2010;17(9):1069.
12. Zhang YN, Pan XY, Huang Y, Shen HB. Adaptive compressive learning for prediction of protein-protein interactions from primary sequence. J Theor Biol. 2011;283(1):44–52.
13. Wang B, Huang DS, Jiang C. A new strategy for protein interface identification using manifold learning method. IEEE Trans Nanobioscience. 2014;13(2):118–23.
14. Jiang J, Wang N, Chen P, Zheng C, Wang B. Prediction of Protein Hotspots from Whole Protein Sequences by a Random Projection Ensemble System. Int J Mol Sci. 2017;18(7):1453.
15. Wang B, Chen P, Wang P, Zhao G, Zhang X. Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes. Protein Pept Lett. 2010;17(9):1111–6.
16. Ji ZW, Wang B, Yan K, Dong LG, Meng GM, Shi L. A linear programming computational framework integrates phosphor-proteomics and prior knowledge to predict drug efficacy. BMC Syst Biol. 2017;11(S 7):127.
17. Hu SS, Chen P, Wang B, Li J. Protein binding hot spots prediction from sequence only by a new ensemble learning method. Amino Acids. 2017; 49(10):1773–85.
18. Zhu L, Deng SP, You ZH, Huang DS. Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding. Ieee Acm T Comput Bi. 2017;14(2):345–52.
19. Zhu L, You ZH, Huang DS, Wang B. LSE: A Novel Robust Geometric Approach for Modeling Protein-Protein Interaction Networks. PLoS One. 2013;8(4):e58368.
20. Liu Q, Chen P, Wang B, Zhang J, Li J. Hot spot prediction in protein-protein interactions by an ensemble system. BMC Syst Biol. 2018;12(Suppl 9):132.
21. Wang B, Chen P, Huang D-S, Li J-J, Lok T-M, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett. 2006;580(2):380–4.
22. Wang B, Huang DS. Dataset reconstruction for protein interface identification using manifold learning method. In: IEEE International Conference on Bioinformatics and Biomedicine; 2014. p. 398–403.
23. Zhu L, Deng SP, You ZH, Huang DS. Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(2):345–52.
24. Li Y-F, Kwok JT, Zhou Z-H. Semi-supervised learning using label mean. In: International Conference on Machine Learning; 2009. p. 633–40.
25. Li Y-F, Zhou Z-H. S4VM: Safe Semi-Supervised Support Vector Machine. In: Computing Research Repository; 2010. abs/1005.1001.
26. Bennett K, Demiriz A. Semi-supervised support vector machines. Adv Neural Inf Proces Syst. 1999;11:368–74.
27. Iqbal M, Freitas AA, Johnson CG. A Hybrid Rule-Induction/Likelihood-Ratio Based Approach for Predicting Protein-Protein Interactions; 2009.
28. Liu L, Cai Y, Lu W, Feng K, Peng C, Niu B. Prediction of protein–protein interactions based on PseAA composition and hybrid feature selection. Biochem Biophys Res Commun. 2009;380(2):318–22.
29. Oh M, Joo KJ. Protein-binding site prediction based on three-dimensional protein modeling. Proteins Structure Function & Bioinformatics. 2009;77(S9):152.
30. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of proteinâ protein interaction sites in heterocomplexes with neural networks . FEBS J. 2010;269(5):1356–61.
31. Ansari S, Helms V. Statistical analysis of predominantly transient protein–protein interfaces. Proteins Struct Funct Bioinform. 2010;61(2):344–55.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
33. Chen P, Hu SS, Zhang J, Gao X, Li JY, Xia JF, Wang B. A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. Ieee Acm T Comput Bi. 2016;13(5):901–12.
34. Choi YS, Han SK, Kim J, Yang JS, Jeon J, Ryu SH, Kim S. ConPlex: a server for the evolutionary conservation analysis of protein complex structures. Nucleic Acids Res. 2010;38(Web Server issue):W450–6.
35. Wei PJ, Zhang D, Li HT, Xia J, Zheng CH, Wei PJ, Zhang D, Li HT, Xia J, Zheng CH. DriverFinder: a gene length-based network method to identify Cancer driver genes. Complexity. 2017;2017(99):1–10.
36. Wei PJ, Zhang D, Xia J, Zheng CH. LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network. Bmc Bioinformatics. 2016;17(Suppl 17):467.
37. Glaser F, Pupko T, Paz I, Bell RE, 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):163–4.
38. Zhang X, Tian Y, Cheng R, Jin Y. A Decision Variable Clustering Based Evolutionary Algorithm for Large-scale Many-objective Optimization. IEEE Trans Evol Comput. 2018;22(1):97–112.

Publisher's Note