Multi-level machine learning prediction of protein-protein interactions in Saccharomyces cerevisiae

, , , , , Dariusz Plewczynski

Accurate identification of protein-protein interactions (PPI) is the key step in understanding proteins' biological functions, which are typically context-dependent. Many existing PPI predictors rely on aggregated features from protein sequences, however only a few methods exploit local information about specific residue contacts. In this work we present a two-stage machine learning approach for prediction of protein-protein interactions. We start with the carefully filtered data on protein complexes available for Saccharomyces cerevisiae in the Protein Data Bank (PDB) database. First, we build linear descriptions of interacting and noninteracting sequence segment pairs based on their inter-residue distances. Secondly, we train machine learning classifier to predict binary segment interactions for any two short sequence fragments. The final prediction of the protein-protein interaction is done using the 2D matrix representation of all-against-all possible interacting sequence segments of both analysed proteins. The level-I predictor achieves 0.88 AUC for micro-scale, i.e. residue-level prediction. The level-II predictor improves the results further by more complex learning paradigm. We perform 30-fold macro-scale, i.e. protein-level cross-validation experiment. The level-II predictor using PSIPRED-predicted secondary structure reaches 0.70 precision, 0.68 recall, and 0.70 AUC, whereas other popular methods provide results below 0.6 threshold (recall, precision, AUC). Our results demonstrate that multi-scale sequence features aggregation procedure is able to improve the machine learning results by more than 10% as compared to other sequence representations. Prepared datasets and source code for our experimental pipeline are freely available for download from URL provided from authors upon request (open source Python implementation, OS independent).

Multi-level machine learning prediction of protein-protein interactions in Saccharomyces cerevisiae

Julian Zubek[1,2], Marcin Tatjewski[1,2], Adam Boniecki[3], Maciej Mnich[4], Subhadip Basu[5], and Dariusz Plewczynski[1]

[1]**Centre of New Technologies, University of Warsaw, Warsaw, Poland**
[2]**Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland**
[3]**Faculty of Mathematics, Informatics and Mechanics, Warsaw University, Warsaw, Poland**
[4]**Faculty of Mathematics and Computer Science, Jagiellonian University, Cracow, Poland**
[5]**Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India**

ABSTRACT

Accurate identification of protein-protein interactions (PPI) is the key step in understanding proteins' biological functions, which are typically context-dependent. Many existing PPI predictors rely on aggregated features from protein sequences, however only a few methods exploit local information about specific residue contacts.

In this work we present a two-stage machine learning approach for prediction of protein-protein interactions. We start with the carefully filtered data on protein complexes available for *Saccharomyces cerevisiae* in the Protein Data Bank (PDB) database. First, we build linear descriptions of interacting and non-interacting sequence segment pairs based on their inter-residue distances. Secondly, we train machine learning classifier to predict binary segment interactions for any two short sequence fragments. The final prediction of the protein-protein interaction is done using the 2D matrix representation of all-against-all possible interacting sequence segments of both analysed proteins. The level-I predictor achieves 0.88 AUC for micro-scale, i.e. **residue-level** prediction. The level-II predictor improves the results further by more complex learning paradigm. We perform 30-fold macro-scale, i.e. **protein-level** cross-validation experiment. The level-II predictor using PSIPRED-predicted secondary structure reaches 0.70 precision, 0.68 recall, and 0.70 AUC, whereas other popular methods provide results below 0.6 threshold (recall, precision, AUC). Our results demonstrate that multi-scale sequence features aggregation procedure is able to improve the machine learning results by more than 10% as compared to other sequence representations.

Prepared datasets and source code for our experimental pipeline are freely available for download from URL provided from authors upon request (open source Python implementation, OS independent).

Keywords: protein-protein interactions, protein interaction networks, multi-scale models, protein sequence, machine learning, physico-chemical indices, interaction patches, sequence segments, local sequence-structure segments

INTRODUCTION

Systems biology and bioinformatics study interactions between various biocomponents of living cells that spans across multiple spatial and temporal scales. The goal is to understand how the complex phenomena arise given the properties of building blocks. Specifically, proteins are characterised in multiple scales: first in the microscale, by their local post-translational modifications; second, by the interactions with metabolites and small chemical molecules (inhibitors); third in the mesoscale, by the three-dimensional structure of active sites, or interaction interfaces; fourth in the macroscale, by the global 3D structure that comprises the macromolecular complexes; and finally in the time-scale, by their dynamical properties related to the changes of their structure, or physico-chemical properties upon participating in the given

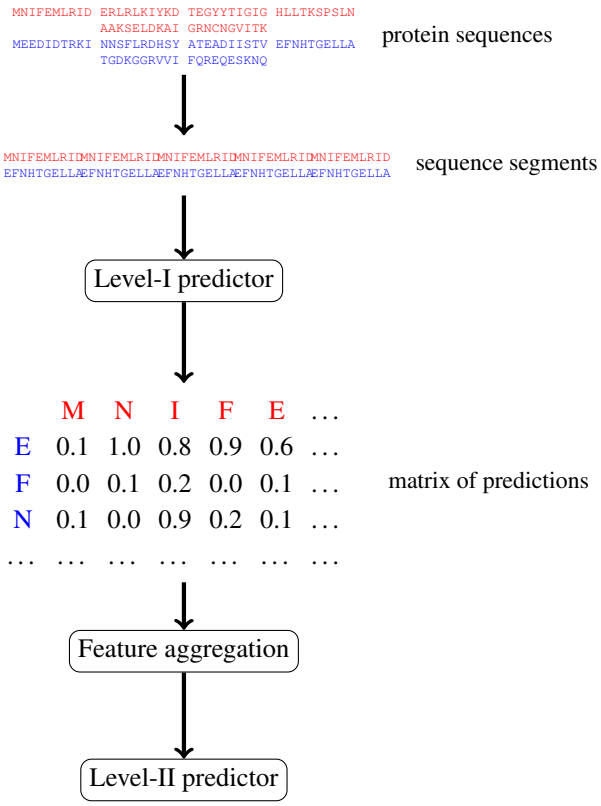


Figure 1. Schematic depiction of our two-stage ensemble method.

biophysical process. Such variety of scales, each linked with different biological function is rooted in their complex and spatio-temporal network of interactions with other smaller biomolecules (metabolites, ligands), comparable size proteins, RNA molecules, and much larger DNA macrochains. Starting from reliable information on single, binary interactions it is possible to reconstruct the whole interaction network, therefore providing further insight into proteins' biological functions on the whole-cell level.

In this paper, we focus on the protein-protein interactions. We develop ensemble learning method for identification of proteins binary interactions by predicting residue-residue interactions. Moreover, we demonstrate how to integrate sequence information from lower scales into higher scale machine learning predictor. In our approach, a binary interaction between two proteins is predicted by considering all possible interactions between their sequence segments using level-I predictor. An output of this phase is the matrix of scores with size corresponding to the proteins' lengths. Given the threshold on the likelihoods, this represent the whole network of all possible residue contacts that could be made during the complex formation. Later, we transform the scores matrix into a fixed length input vector suitable for further statistical data analysis (aggregated values over columns, diagonals, etc.), and we identify the network properties (e.g. sizes of connected components) using interaction graph. This data is used by the level-II predictor, which integrates information similarly to a human expert. Figure 1 presents this pipeline graphically.

Recently, several machine learning algorithms were applied to predicting protein interactions. Our study takes similar route as Yip et al. (2009), who predicted interactions of *Yeast* proteins at the level of residues, domains, and whole proteins. On the residue level, sequence and secondary structure was used; on the protein level, they used phylogenetic profiles, sub-cellular localisations, gene expression data and interactome-derived features; finally, on the domain level, the co-evolution was selected to characterize interacting proteins. They constructed classifier allowing for the information flow between the above three levels to improve final prediction. This approach was further developed by Saccà et al. (2014), who introduced a different model of knowledge integration, and demonstrated its superiority on the previously used benchmarking data. Reported AUC values reached 0.80 for residues, 0.96 for domains and 0.82

for proteins. However, these results are likely to be biased because in the testing phase the authors used different interactions but the same proteins as in the training phase. It is unknown how their method would perform on previously unseen proteins.

Our method differs from the previous approaches in several ways. First, we treat residue level predictions only as the input for identification of protein level interactions. The information flows only bottom-up: from the residue level to the protein level. In Yip et al. (2009) work information flow between levels was an additional technique to improve prediction, in our method it is the core of the predictor. Moreover, we use only sequence and secondary structure features for prediction; PDB data is only used in training, which makes it possible to apply our algorithm even if the three-dimensional structures of both proteins are not known. The secondary structure can be predicted with high accuracy by standard bioinformatics methods. Therefore, we can use our method even, when the high level properties of proteins remain unknown.

We compare our results with PPI predictors targeted at Yeast and exploiting global sequence properties (i.e. not considering local residue interactions). One of such methods was developed by Liu (2009). He constructed the feature vector by comparing values of the selected amino acid indices at selected distances in the sequence. The data contained 5926 interacting protein pairs from DIP database. His method achieved 0.87 precision and 0.90 recall. Another sequence-based method was proposed by Chang et al. (2010). Their data contained 691 protein interactions. The authors used sequence composition and the accessible surface area predicted by another algorithm (Chang et al., 2008) as basic features, which were averaged over the entire sequence. They observed 0.72 precision and 0.80 recall.

Here we would like to stress one of the most important outcomes of our study. Namely, the performance results reported in various publications are generally not comparable directly. They tend to differ in collected data methodology, definition of positives and negatives and evaluation procedures. In our work, we focused first on preparing the collection of high quality interactions data and remove any bias from the further evaluation procedure. To obtain a meaningful comparison between different methods we reimplemented several methods for aggregating features of protein sequences (including the method of Liu (2009)), and evaluated them on our benchmarking data set. Our method outperformed the others in terms of ROC AUC by the large margin (the smallest difference was 0.70 AUC for our method vs. 0.60 AUC for Liu's method).

The results obtained in our study are much lower than usually reported in the literature. We claim this is an effect of carefully balancing positives and negatives in our data sets and the rigorous evaluation strategy in which a predictor is always tested on proteins unseen during the training phase. What we are measuring is the ability to predict real protein compatibility and not just relative proteins' reactivity. This task is much harder, and we demonstrate that popular methods using sequence-derived features do not perform well in this context. Our result confirms previous methodological studies in this area (Park and Marcotte, 2012).

MATERIALS

We extracted the three dimensional (3D) structures of all *Yeast* protein-protein hetero-complexes from Protein Data Bank (PDB) (Berman et al., 2000), which were crystalised using X-RAY with the resolution below 3Å. Homologous structures were removed with 90% sequence identity threshold. We mapped the PDB complexes to Uniprot (Consortium, 2014) ids using SIFTS mapping (Velankar et al., 2013) tool. Secondary structure was extracted from PDB structures using DSSP software (Joosten et al., 2011; Kabsch and Sander, 1983). Because of the gaps present in PDB structures gaps in the extracted secondary structure also occurred. Such gaps were filled with the coil symbol. For evaluation purposes we also employed PSIPRED tool (Jones, 1999) to predict secondary structure from protein sequence.

On the residue level, from 3D structures it was possible to extract all interacting pairs of residues. We considered any two residues from two distinct proteins as interacting, if they were located in Euclidean space within a distance threshold of 4Å from each other. On the protein level, we identified two proteins as an interacting pair if there was at least one residue level microscopic interaction between them. We were interested only in heterodimers, i.e. interactions occurring between two different proteins. Such interactions can contribute a lot to our understanding and the reconstruction of the true protein interaction network (PIN).

Residue level positives and negatives

The first step of our procedure was to build the training dataset for level-I predictor. We employed a sliding window technique to extract fragments of protein sequence. Further in this work we refer to it as extraction window.

The positive examples of the training set were formed from pairs of fragments in which central residues interact, and additionally a certain number of other residues in a specified distance from the central one interact. We refer to the required number of interacting residues as interaction threshold and to the maximal distance from the central residue as the maximal neighbour interaction distance. By introducing this restriction we deliberately focused on strong interactions, filtering out the weaker ones, which could be just noise in the complex.

In the data preparation phase we fixed the maximal neighbour interaction distance to 10 residues. We have chosen this value following the studies of Youn et al. (2007) and Kauffman and Karypis (2010), who predicted binding residues from protein sequences and used the window of size 21 (1 central residue, 10 residues to the right, 10 to the left). As for interaction threshold, we benchmarked the values of 0, 5, 10, 15, and 20 interacting residues within the maximal interaction distance.

Let us observe that the values of maximal neighbour interaction distance and extraction window size are not necessary the same. One can imagine identifying residue level positives using larger interaction distance, and then encoding features of only few central residues using small extraction window, and vice versa. Indeed in our work we test different sizes of extraction window ranging from 3 to 31 independently of the fixed interaction distance.

In our data extracted from PDB we have no natural information on negative residue interactions. Pairing the sequence fragments totally randomly could result in lot noise and false negatives. Therefore, we decided to extract non-interacting pairs of fragments from interacting protein pairs. Pairs without any interacting residues or pairs in which one fragment has some interactions but the other has none were considered negatives. This way it was guaranteed that at least one of the fragments does not come from an interface region. The number of potential negatives was much larger than the number of positives, therefore we decided to keep the imbalance ratio at 3:1. The required amount of negatives was therefore sampled at random. This is a common practice in machine learning since most of the algorithms perform poorly on data sets with large class imbalance (see Chawla (2005) for review).

Protein level positives and negatives

The second step of our methodology is the data preparation for training level-II predictor. We used the same dataset of interacting protein pairs from PDB database as positives, and generated negative examples. The construction of the high quality negative examples is very difficult. Common methods for generating negatives include drawing random pairs of biomolecules from all known proteins found in a specific organism (Saha et al., 2014), or from the considered subset of the whole proteome, namely from proteins occurring in positive examples (Chang et al., 2010). We strongly believe that such methods have their inherent drawbacks, because they ignore network properties of the underlying protein interactome. We used the following procedure instead:

- Let G_1 be a graph representing positive examples. Denote $V = v_1, \ldots, v_n$ as the set of its vertices. Each vertex in V represents a protein and each edge v_i, v_j represents an interaction. Let $[Deg(v_1), \ldots, Deg(v_n)]$ be a vector containing degrees of vertices from V. Let G_2 be a graph of negative interactions. At first it has vertices identical to G_1 and no edges.

- While there exist v such that $Deg(v) > 0$:

    1. Find vertex v with the largest $Deg(v)$.

    2. Find vertex u if exist such that:

        (a) There is no edge (v,u) in G_1.

        (b) u has as large $Deg(v)$ as possible.

        (c) Distance $d(u,v)$ in G_1 is as large as possible.

    3. If u exist:

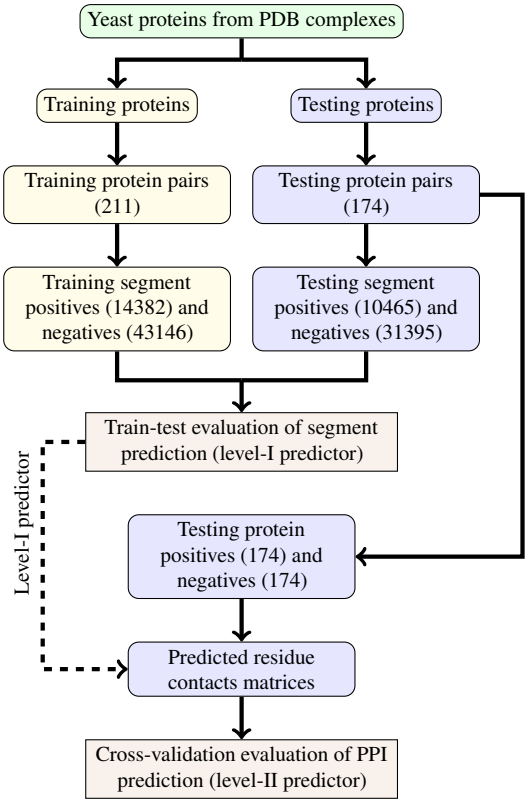


Figure 2. Schema of train-test split for evaluation of trained classifiers. Numbers of examples used are given in parentheses.

(a) Add edge (u,v) to G_2.

(b) $Deg(v) \leftarrow Deg(v) - 1$

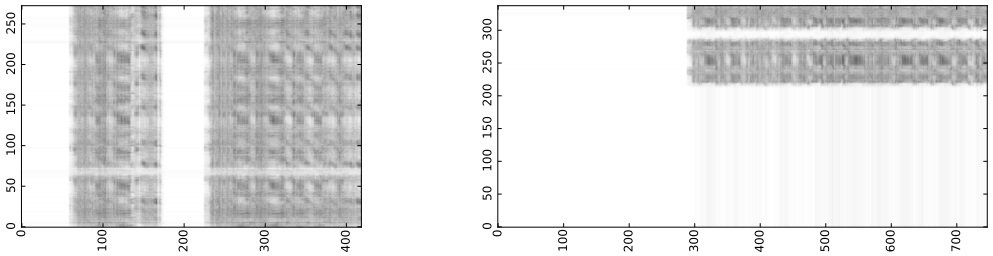
(c) $Deg(u) \leftarrow Deg(u) - 1$

4. else: $Deg(v) \leftarrow 0$

Such schema of constructing the negative set is unbiased, i.e. the protein composition of the positives and the negatives remains identical. Every single protein has the same number of positive and negative interactions. This forces the trained classifier to predict meaningful biophysical interactions rather than predicting general reactivity (the relative number of interactions) of a single protein. Otherwise, the best results would be achieved by the predictor, which predicts that the two proteins interact if each of them has a lot of interactions, regardless of their compatibility. What is also important, our algorithm favours protein pairs which are remote to each other in the interaction network, which reduces the risk of introducing false negatives.

Train-test split

The last step of data generation, is the splitting procedure into training and testing datasets. In order to truly evaluate our method in a realistic setup, we split the benchmarking dataset at the protein level, not at the residue level. This makes our goal more difficult, as compared to previous approaches that sometimes use residue-level splitting of benchmarking dataset. The schematic depiction of train-test split is given by Figure 2. Level I classifier is evaluated through train-test experiment with relatively large datasets. Level-II classifier is evaluated through cross-validation experiment which uses the test set only – information from the train set comes only in the form of trained level-I classifier. Such schema eliminates the risk of overoptimistic performance estimates caused by the same data appearing during training and testing phase.



(a) Interacting proteins **(b)** Non-interacting proteins

Figure 3. The level-I prediction matrices for two protein pairs. White colour corresponds to score 0.0, black colour corresponds to score 1.0.

METHODS

The level-I predictor is trained to recognise interacting pairs of fragments. This should be an equivalent of detecting compatible protein patches on the surface of a protein. For each possible pair of fragments from two different proteins a prediction is made, and the likelihood estimation for all against all pairs of fragments are stored in the interaction matrix. The level-II predictor uses the output of level-I predictor, predicting binary interactions between two proteins using the aggregated features, i.e. the complementarity between their surface patches.

Level-I predictor

We trained level-I predictor on interacting sequence fragments of proteins from the training set and tested it on the testing set. Input for level-I predictor consisted of pairs of sequence fragments of the length of extraction window. We compared different forms of feature encoding: raw sequence, sequence encoded with selected AAindex properties (HQI8 (Saha et al., 2012)) and secondary structure.

As the core classifier, we evaluated two popular machine learning methods: Random Forest and Support Vector Machine. Both algorithms are commonly used in bioinformatics and are considered best off-the-shelf classifiers (Yang et al., 2010). Their parameters values were chosen through a grid search. As the performance measure we have chosen ROC AUC (area under Receiver Operating Characteristic curve).

Multi-level feature aggregation

To infer a binary interaction between two proteins, we consider all possible interactions between their sequence segments as predicted by level-I predictor. An output of this phase is the matrix of likelihoods with the dimension equal to the multiplied proteins' lengths. Each prediction score is a real number between 0 and 1. Example matrices of scores for positive and negative case are presented in Figure 3.

To transform the 2D matrix into an input vector suitable for level-II predictor, we extracted the following features (numbers in parentheses denote the number of values in the final feature vector):

- the mean and variance of values over the matrix (2),
- the sums of values in 10 best rows and 10 best columns (20),
- the sums of values in 5 best diagonals of the original and transposed matrix (10),
- the sum of values on intersections of 10 best rows and 10 best columns (1),
- the histogram of scores distributed over 10 bins (10),
- graph features: fraction of nodes in the 3 largest connected components (3).

Graph features require further explanation. Predicted contacts between residues were represented as a bipartite graph. Nodes in the graph represented residues and edges represented predicted contact. To make the graph more realistic biologically, for each node we left only 3 strongest outgoing edges. We set the value of this threshold (3) following the observation that in our PDB structures the mean number of interactions of a single interacting residue is between 2 and 3. In such trimmed graph we calculated fractions of nodes contained in 3 largest connected components. These values were also appended to the feature vector.

Level-II predictor

The performance of level-II predictor was evaluated through a variant of stratified 30-fold cross-validation performed on the protein level. Each fold contained $\frac{1}{30}$ of positive protein pairs and $\frac{1}{30}$ of negative protein pairs from the testing set. There was no overlap between splits on the pair level, but there was still an overlap on the level of single proteins which constitute pairs. We observed that this introduced a huge bias into evaluation results (similar observation was previously made by Park and Marcotte (2012)). To fix our cross-validation scheme we applied the following procedure:

1. Let $O = \left\{(p_1^1, p_1^2), (p_2^1, p_2^2), \ldots, (p_n^1, p_n^2)\right\}$ be a set of all protein pairs.
2. For each fold $F \subset O$:
   (a) Build a set P composed of all proteins occurring in F:
   $$P = \left\{x : \exists_y \, (x,y) \in F \vee (y,x) \in F\right\}$$
   (b) Build a set $A \subset O$ composed only of pairs consisted of proteins occurring in P:
   $$A = \{(x,y) : x \in P \wedge y \in P\}$$
   (c) Build a set $B \subset O$ composed only of pairs consisted of protein not occurring in F but occurring in the testing set:
   $$B = \{(x,y) : x \notin P \wedge y \notin P \wedge (x,y) \in O\}$$
   (d) Train the classifier on B set, and test it on A.
3. Collect all the predictions for A-sets, and calculate performance metrics.

The above described procedure differs from the standard cross-validation, since the number of observations in constructed test sets vary slightly, but this variance is small, and does not influence the estimated performance. Such evaluation schema does not allow for any information leak: the data sets are always balanced, and the classifier is tested on previously unseen proteins.

As classification methods for level-II predictor we used Random Forest and Support Vector Machine with parameters tuned through a grid search.

Protein sequence feature aggregation

We compared our ensemble method with various sequence feature aggregation schemas that are commonly applied in machine learning methods for prediction of proteins interactions. To make the benchmarking results comparable between different algorithms, we used the same classification method (Random Forest) and evaluation procedure (30-fold cross-validation on the testing set) as for level-II predictor. We benchmarked the following feature aggregation schemas:

1. AAC – Amino Acid Composition (Nanni et al., 2014). Feature set is the set of frequencies of all amino acids in the sequence.
2. PseAAC – Pseudo Amino Acid Composition (Chou, 2001). Feature set consists of the standard AAC features with k-th tier correlation factors added. We calculate those correlations on HQI8 indices.
3. 2-grams (Nanni et al., 2014). Feature set comprises of frequencies of all 400 ordered pairs of amino acids in the sequence.
4. QRC – Quasiresidue Couples (Guo and Lin, 2005). A set of AAIndices is chosen. For each index d combined values of this property d for a given amino acid pair are summed up for all the pair's occurrences over the full protein sequence. Occurrences for pairs of residues separated from each other by $0, 1, 2 \ldots m$ residues. In effect, one obtains QRC^d vectors of length $400 \times m$. In this model we also use HQI8 indices.

Classifier	Features	AUC
RF	Raw sequence	0.65
	HQI8	0.71
	Secondary structure	**0.85**
SVM	Secondary structure	0.84

(a) Results for threshold 5.

Threshold	AUC
0	0.84
5	0.85
10	0.86
15	0.88
20	0.88

(b) Results for different thresholds for RF using secondary structure.

Table 1. ROC AUC scores of level-I predictor trained on different sets of features and different interaction thresholds. Extraction window size was set to 21. RF – Random Forest, 300 trees, maximum tree depth 15, SVM – Support Vector Machine, RBF kernel, $C = 2$, $\gamma = 0.048$.

5. Variation of Liu's protein pair features (Liu, 2009). The method starts from encoding each amino acid in protein sequence with 7 chosen physicochemical properties, thus obtaining 7 feature vectors for each sequence. For each feature vector its "deviation" is calculated:

$$\gamma_{dj} = \frac{1}{n-d} \sum_{i=1}^{n-d} x_{ij} \times x_{(d+i),j} \quad j = 1,\ldots,7 \;\; d = 1,\ldots,L$$

where x_{ij} is the value of descriptor j for amino acid at position i in sequence P, n is the length of protein sequence P, and d is the distance between residues in the sequence. For the purpose of comparison, we tested this method with the original 7 amino acid indices used by Liu, as well as with HQI8 features. We tested different values of L from 5 to 30 in a quick cross-validation experiment on our data and chose $L = 9$ as yielding the best results.

RESULTS AND DISCUSSION

We evaluated carefully all subsequent steps of our method to choose optimal features and parameter values. Then we compared performance of level-II predictor with popular sequence encoding schemas. We used DSSP-extracted secondary structure for training level-I predictor to facilitate optimally information from PDB complexes. For evaluation of level-II predictor, on the other hand, we used PSIPRED-predicted secondary structure to demonstrate that our method can be employed successfully when only protein sequences are known.

The first task was to decide on the set of optimal features for level-I predictor. Evaluation results of the predictor on different features for 5 interaction threshold are presented in Table 1a. The secondary structure performed better than sequence features. Including both secondary structure and HQI8 did not provide any improvement – secondary structure already contained the necessary information. Carefully tuned SVM did not perform better than Random Forest. In all further experiments we used secondary structure as the source of features for level-I predictor, together with Random Forest as the learning algorithm.

Table 1b reports ROC AUC values for different interaction thresholds. As each value concerns a bit different data set these results do not say which predictor is the best but rather which task is the easiest to solve. We can see that increasing interaction threshold indeed made level-I prediction task easier, but from these results it is not possible to say how it would affect the performance of level-II predictor.

In the next step we selected the optimal extraction window size. Figure 4 presents ROC AUC score against window size. After analysing the plot we decided to keep the size 21 as it provided good performance and it was previously used in other publications. In that way we fixed extraction window size to the same value as maximal interaction distance, used to define positive residue interactions.

After fixing the parameters of level-I predictor, we used it to construct a data set for level-II predictor from protein pairs contained in the testing set and generated negatives. Level-II predictor was compared with other algorithms using representations based on aggregated protein sequence. Results are presented in Table 2. Once again Random Forest proved to be more suitable for the task than Support Vector Machine. Interaction threshold 15 yielded best results. On this kind of data level-II predictor outperformed other methods significantly.

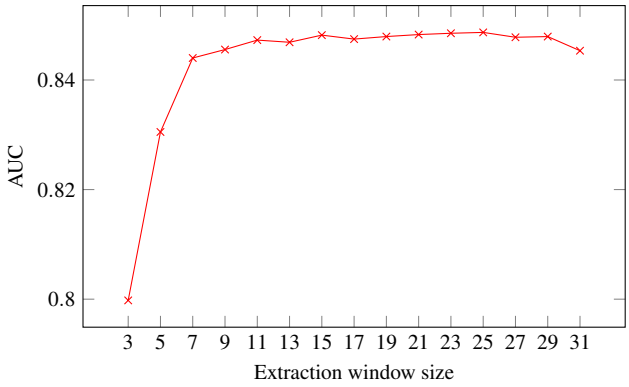


Figure 4. ROC AUC scores of level-I predictor trained on secondary structure for different extraction window sizes. Random Forest was used as the classifier.

Clf	Features	Accuracy	Precision	Recall	AUC
	Lvl-II pred ($t = 0$)	0.55	0.58	0.55	0.57
	Lvl-II pred ($t = 5$)	0.55	0.58	0.54	0.57
	Lvl-II pred ($t = 10$)	0.55	0.57	0.55	0.57
	Lvl-II pred ($t = 15$)	0.55	0.58	0.55	0.57
	Lvl-II pred ($t = 20$)	0.56	0.58	0.56	0.57
SVM	AAC	0.54	0.56	0.66	0.54
	PseAAC	0.54	0.55	0.61	0.55
	2grams	0.55	0.56	0.64	0.55
	QRC	0.51	0.53	0.59	0.53
	Liu's dev (HQI8)	0.55	0.57	0.60	0.56
	Liu's dev (original)	0.55	0.57	0.60	0.56
	Lvl-II pred ($t = 0$)	0.62	0.64	0.62	0.67
	Lvl-II pred ($t = 5$)	0.62	0.64	0.61	0.67
	Lvl-II pred ($t = 10$)	0.67	0.68	0.69	0.68
	Lvl-II pred ($t = 15$)	**0.68**	**0.70**	**0.68**	**0.70**
	Lvl-II pred ($t = 20$)	0.59	0.61	0.60	0.64
RF	AAC	0.54	0.57	0.54	0.56
	PseAAC	0.53	0.55	0.52	0.55
	2grams	0.53	0.56	0.49	0.55
	QRC	0.50	0.52	0.43	0.51
	Liu's dev (HQI8)	0.55	0.58	0.55	0.60
	Liu's dev (original)	0.56	0.59	0.57	0.59

Table 2. Performance scores. $t = x$ denotes interaction threshold of x interacting residues. Level-II predictor used secondary structure predicted by PSIPRED. RF – Random Forest, 300 trees, maximum tree depth 7, SVM – Support Vector Machine, RBF kernel, $C = 1$, $\gamma = 2$.

We draw reader's attention to the fact that the performance of popular protein representation strategies evaluated on our data was generally much lower than the results reported in the literature. One of the reasons may be relatively small size of our data set – it might contain not enough examples for a classifier to learn complex patterns. The other explanation is the way we constructed positives and negatives. In our case every protein occurred in the same number of positive and negative pairs. Moreover, we performed cross-validation with splits on protein level, which means that there were no proteins occurring simultaneously in training and testing set. In such conditions any method which makes a good prediction of general proteins' reactivity but does not consider their compatibility performs poorly. This observation is consistent with results obtained by Park and Marcotte (2012), who analysed the impact of performing splits on pair level instead of component level on cross-validation results. They reported AUC scores of popular protein interaction prediction methods dropping from 0.7–0.8 to 0.5–0.6. We plan to address these issues in our future work, and present a more detailed study of evaluation methods for this kind of predictors.

CONCLUSIONS

In this work we presented the method for constructing multi-level classifier for protein-protein interactions. We demonstrated that the information present at the lower level can be successfully propagated to the upper level to make the prediction. No additional features beside protein sequence and secondary structure predicted from sequence are required.

Our goal, i.e. predicting real compatibility between two proteins regardless of their relative reactivity, forced us to collect high quality data and develop a rigorous evaluation procedure. We have taken into account properties of protein interaction network to construct balanced negatives. During the evaluation we carefully separated training and testing proteins to avoid information leak. We demonstrated that our method is working under such conditions better than popular sequence feature aggregation schemas.

There is still the room for further improvements regarding classification accuracy. We plan to include additional features both at the residue level and at the protein level to see if our model can benefit from them. Another direction that we want to explore is expanding the model to include proteins from organism other than Yeast, and evaluating it on bigger data sets.

We hope that our work will inspire further discussion regarding evaluation strategies for protein interaction predictors. We believe that deeper understanding of these matters would allow to compare different methods in a more systematic manner, which would be beneficial for the research done in this area.

ACKNOWLEDGMENTS

The paper is founded by the European Union from financial resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", the 2013/09/B/NZ2/00121 grant from polish National Science Centre and and COST BM1405 EU action.

REFERENCES

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Chang, D. T., Syu, Y.-T., and Lin, P.-C. (2010). Predicting the protein-protein interactions using primary structures with predicted protein surface. *BMC Bioinformatics*, 11(Suppl 1):S3.

Chang, D. T.-H., Huang, H.-Y., Syu, Y.-T., and Wu, C.-P. (2008). Real value prediction of protein solvent accessibility using enhanced PSSM features. *BMC Bioinformatics*, 9(Suppl 12):S12.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer-Verlag, New York.

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, 43(3):246–255.

Consortium, T. U. (2014). Activities at the universal protein resource (UniProt). *Nucleic Acids Research*, 42(D1):D191–D198.

Guo, J. and Lin, Y. (2005). A novel method for protein subcellular localization: Combining residue-couple model and SVM. pages 117–129.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202.

Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39(Database issue):D411–D419.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kauffman, C. and Karypis, G. (2010). Ligand-binding residue prediction. In Rangwala, H. and Karypis, G., editors, *Introduction to Protein Structure Prediction*, pages 343–368. John Wiley & Sons, Inc.

Liu, H.-w. (2009). Protein-protein interaction detection by SVM from sequence. In *Information", The Third International Symposium on Optimization and Systems Biology*, pages 198–206.

Nanni, L., Lumini, A., and Brahnam, S. (2014). An empirical study of different approaches for protein classification. *The Scientific World Journal*, 2014:e236717.

Park, Y. and Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods*, 9(12):1134–1136.

Saccà, C., Teso, S., Diligenti, M., and Passerini, A. (2014). Improved multi-level protein¿ protein interaction prediction with semantic-based regularization. *BMC Bioinformatics*, 15(1):103.

Saha, I., Maulik, U., and Plewczynski, D. (2012). Application of high quality amino acid indices to AMS 3.0: A update note. In *BIC-TA (1)*, volume 201 of *Advances in Intelligent Systems and Computing*, pages 217–225. Springer.

Saha, I., Zubek, J., Klingström, T., Forsberg, S., Wikander, J., Kierczak, M., Maulik, U., and Plewczynski, D. (2014). Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Molecular BioSystems*, 10(4):820–830.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(Database issue):D483–489.

Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308.

Yip, K. Y., Kim, P. M., McDermott, D., and Gerstein, M. (2009). Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC Bioinformatics*, 10(1):241.

Youn, E., Peters, B., Radivojac, P., and Mooney, S. D. (2007). Evaluation of features for catalytic residue prediction in novel folds. *Protein Science : A Publication of the Protein Society*, 16(2):216–226.