Prediction of protein-protein hetero interaction sites from local sequence information using Random Forest

Mile Šikić[1]. Kristian Vlahoviček[2] and Branko Jeren[1]

[1]University of Zagreb. Faculty of Electrical Engineering and Computing. Unska 3. HR-1000 Zagreb. Hrvatska.
mile.sikic@fer.hr and branko.jeren@fer.hr
[2] Bioinformatics Group. Department of Molecular Biology. Division of Biology. Faculty of Science. University of Zagreb. Horvatovac 102a. 10000 Zagreb. Croatia
kristian@bioinfo.hr

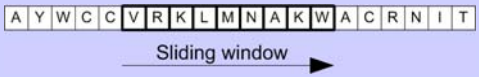
ABSTRACT

Identifying the interface between two interacting proteins provides important clues to the function of a protein, and is becoming increasing relevant to drug discovery. Many studies indicate that the compositions of contacting residues are unique. Here, we describe a method that identifies protein-protein interaction sites from sequence by using Random Forest algorithm. For prediction we use a non–redundant set of 333 protein complexes. The AUC for Random Forest classifier is 0.76. In the second step we used a data fusion method. When 77% of our predictions were right, we correctly predicted 43% of all interaction sites. In 88% of proteins chains we correctly predicted at least one interaction site. Furthermore, when in prediction we included residues that are up to 5 residues far from our predicted site, we covered 64% of all interaction sites. These results strongly indicate that prediction of interaction sites from sequence alone is possible and comparable with results obtained using structure information.

BACKGROUND AND METHODS

Prediction of protein interactions using only information that can be obtained from sequence is one of the most challenging tasks in bioinformatics. For successful prediction it is very important to have a non-redundant and large data set. In our work we used a data set of 333 protein complexes [2]. The protein interaction site is defined as a sliding window of nine residues that contains equal number or more contacts residues than specified by the threshold value. For the threshold we used values one to six. The contact residues are defined as residues that have at least one non hydrogen atom with distance less then 6 Å from a non-hydrogen atom of the neighbouring chain

For prediction we used Random Forest classifier. We chose this classifier because of its high accuracy and good performance in the case of imbalanced data. In this work ratio between negative and positives were around 4:1. For testing performance of the classifier we used 10 – fold cross-validation. In addition, we compared results with results obtained by random classifier by using method of randomized testing.



For further improvement of the prediction performance we used a data fusion method that used results obtained for all tested threshold cases. The data fusion method algorithm is based on a simple rule that we call Oracle. The rule is defined as follows: if any of classifiers (each uses different contact threshold value) predicts that particular window is interaction site that place is defined as interaction site. Since this is a new approach and we haven't implemented cross-validation in it, we used 2/3 of the set for training and 1/3 for testing.

RESULTS

$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{recall} = \frac{TP}{TP + FN} \qquad F - \text{measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

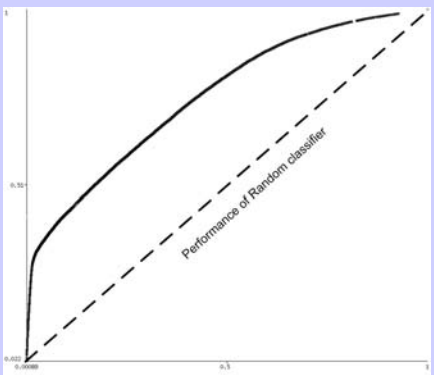


Figure 1 ROC curve in a case when an interface is defined with threshold of 5 contact residues in window of 9. The best performance is in bottom left corner. Since data is unbalanced this part of curve will be used for prediction.

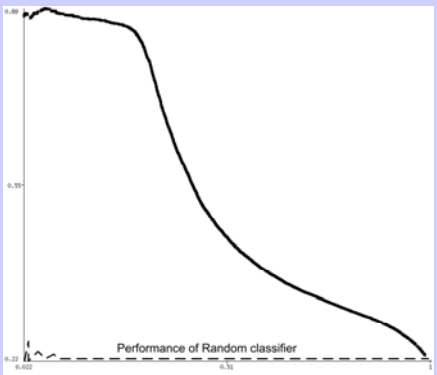


Figure 2 Precision – recall curve in a case when an interface is defined with threshold of 5 contact residues in window of 9. The performance in the part when recall is less then 0.35 is very good. The performance sharply decreases in region where recall is above this value.

Table 1 Number of interaction sites in non-redundant set and performance of Random Forest classifiers for different contact number thresholds in the window of nine residues. 10 – fold cross validation was used.

Threshold	Number of interaction sites in set	Percent	Precision	Recall	AUC
1	46087	100.00	0.67	0.39	0.76
2	45431	98.58	0.68	0.39	0.76
3	44072	95.63	0.68	0.49	0.76
4	41344	89.71	0.69	0.48	0.75
5	36970	80.22	0.75	0.34	0.76
6	30032	65.16	0.75	0.32	0.75

Table 2 Comparing performance for classifiers with different contact number thresholds in window of nine residues and data fusion with Oracle method. The 2/3 of chains are used for training and 1/3 for test. In the coverage we included all interaction sites that are up to 5 residues far from the predicted one.

Threshold	Precision	Recall	F - measure	Chains with at least one predicted interaction site		Coverage - included up to 5 residues far from predicted one.
				Number	Percent	
1	0.82	0.24	0.37	163	49.65	0.55
2	0.83	0.24	0.37	144	43.11	0.60
3	0.83	0.24	0.37	141	42.34	0.58
4	0.83	0.23	0.36	123	37.16	0.63
5	0.81	0.22	0.34	116	35.47	0.66
6	0.79	0.20	0.32	111	34.05	0.61
Oracle 1	0.77	0.43	0.55	295	88.06	0.64

CONCLUSION

In this study we present prediction of protein interaction sites from sequence using combination of Random Forest algorithm and a data fusion method. Obtained results are comparable with results obtained with methods based on structural information. Since, prediction of interaction sites using usual classifiers has very good precision, but unsatisfactory recall, we used a method that slightly decreases the precision with great gain in recall and F –measure.

REFERENCES

References:
[1] L. Breiman. "Random Forests." *Machine Learning.* vol. 45. no. 1. pp. 5--32. 2001.
[2] Y. Ofran. and B. Rost. "Predicted protein-protein interaction sites from local sequence information." *FEBS Lett.* vol. 544. no. 1-3. pp. 236-9. Jun 5. 2003.
[3] I. H. Witten. and E. Frank. *Data Mining: Practical machine learning tools and techniques.* 2nd. ed.. San Francisco: Morgan Kaufmann. 2005.
[4] G. Topic. and T. Smuc. "PARF - Parallel RF Algorithm." Institut Rudjer Boskovic. 2004.
[5] T. Fawcett. "An introduction to ROC analysis." *Pattern Recogn. Lett..* vol. 27. no. 8. pp. 861-874. 2006.