Prediction of RNA binding sites in proteins

Muhammad Usman Akram
Faculty of Science
University of Trento, Italy
muhammadusman.akram@studenti.unitn.it

***Abstract*—SVM based approach is used for predicting RNA binding residues in proteins.**

I. Introduction

In biological systems, one of the most common mean of communication among cells is small molecules and proteins. Identifying proteins binding sites with RNA can help us understand many of these communications and processes. In literature, we can find many methods with considerable accuracies of prediction, which are developed over time for predicting protein-protein and protein-DNA binding sites. But this work done with focus on predicting protein-RNA bindings, as they are very important in protein synthesis among many other processes.

II. Materials and Methods

Prediction of binding sites in RNA-Protein complex is done in various steps, including Database selection, feature extraction, normalization and classification using support vector machine (SVM). All these steps, along with tools and methods employed are discussed in detail in following sections.

A. Data Set

We have employed this methods over data set clustered by Terribilini et al. [1], with some changes. Used data set contains 160 non redundant protein chains with sequence homology less than 30%. They belong to 58 RNA-protein complexes. Atomic coordinated were downloaded from Protein Data Bank [**?**].

B. HBPlus: Hydrogen Bond Calculator

We employed HBPlus [2] for calculation of Hydrogen bonds. It calculates geometric of all hydrogen bond in given complex, while given positions of Hydrogen. It also deals with Hydrogens that can take more than one position.

C. DSSP: Dictionary of Protein Secondary Structure

DSSP [3] is a dictionary for all proteins in the Protein Data Bank. So, It can calculate secondary structure from pdb entries and also provides accessible surface area among other outputs.

D. PSI-BLAST Profiles

PSI-Blast [4] or position specific iterative - basic local alignment search tool profiles were used to create position specific scoring matrix (PSSM) which measures residue conservation at a specific position. PSSM scores are directly proportional to conservation at given position.

E. Support Vector Machine (SVM)

For purpose of classification we have employed SVM. Support Vector Machines belong to group of large margins classifiers. SVM tries to find a hyperplane separating two classes with maximum margin. SVM is the maximal margin hyperplane in feature space build using kernel function in gene space [5]. We have used LibSVM [6] implementation of SVM. We have used radial biases function (rbf) as kernel for SVM, which is given by,

$$\kappa\left(p,q\right) = exp\left(-\gamma\|p-q\|^2\right) \tag{1}$$

Where, γ is width of rbf.

F. Feature Extraction and Encoding

First of all, we extracted protide chains and bindings using HBPlus (3 of initially 61 selected protine complexes were dropped due to issues with HBPlus) and calculated PSSM using PSI-Blast, this makes major portion of feature set. For calculation accessible surface area and secondary structure we used DSSP. Residue-wise feature vectors were made using above described features. These feature vectors were extracted using sliding window technique, using window of size 15. As residue being binding or non binding is affected by its neighbors residues. In order to allows window to extend over terminals, null residues represented by all zeroes were used. Feature vector can be given by,

$$|F| = \text{Secondary Structure, Accessible Surface Area, ...} \\ PSSM_{\text{using siding window}}$$

All features were normalized to range [0,1] using,

$$f\left(x\right) = \frac{1}{1+exp^{-x}} \tag{2}$$

G. Classification

Classification involves grid search to find optimal parameters of classification with 10 fold cross validation over taring data using libSVM. Once we have optimized parameters, its training a testing of 7 fold cross validation with changing test set and calculation of evaluation metrics.

III. Results

Results of prediction of protein binding cites are presented in following sub-sections.

Classifier	Accuracy	Precision	Recall	F-Score	Specificity	Balanced Acuracy	MCC
SVM [7]	87.1%	55.9%	45.6%				0.432
SVM[1]	$96.60\% \pm 2.05$	$83.86\% \pm 2.27$	$56.93\% \pm 2.67$	$67.74\% \pm 1.37$	$98.98\% \pm 0.24$	$77.96\% \pm 1.24$	$0.68\% \pm 0.01$

TABLE I

COMPARISON OF CLASSIFIERS OVER 7 FOLD CROSS VALIDATION WITH $c = 256$ & $\gamma = 0.625$

A. Performance Metrics

Following performance metrics are use to evaluate and compare performance of employed classifiers.

1) Mean & Standard Deviation of Accuracy: Accuracy is calculated over 7 fold cross validation. It is given by,

$$A = \frac{TP + FP}{TP + TN + FP + FN} \tag{3}$$

Where,

- TP is number of true positives
- TN is number of true negatives
- FP is number of false positives
- FN is number of false negatives

2) Mean & Standard Deviation of Precision: Precision is calculated over 7 fold cross validation, It is given by,

$$P = \frac{TP}{TP + FP} \tag{4}$$

Where,

- TP is number of true positives
- FP is number of false positives

3) Mean & Standard Deviation of Recall: Recall (also known as Coverage or Sensitivity) is calculated over 7 fold cross validation. It is given by,

$$R = \frac{TP}{TP + FN} \tag{5}$$

Where,

- TP is number of true positives
- FN is number of false negatives

4) Mean & Standard Deviation of F-Score: F-Score is calculated over 7 fold cross validation. It is given by,

$$F_Score = 2 * \frac{P * R}{P + R} \tag{6}$$

Where,

- P is value of Precision
- R is value of Recall

5) Mean & Standard Deviation of Specificity: Specificity is calculated over 7 fold cross validation. It is given by,

$$S = \frac{TN}{TN + FP} \tag{7}$$

Where,

- TN is number of true negatives
- FP is number of false positives

6) Mean & Standard Deviation of Balanced Accuracy: Balanced Accuracy is calculated over 7 fold cross validation. It is given by,

$$BAC = \frac{R + S}{2} \tag{8}$$

Where,

- R is value of Recall
- S is value of Specificity

7) Mean & Standard Deviation of MCC: MCC is calculated over 7 fold cross validation. It is given by,

$$MCC = \frac{TP * TN + FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{9}$$

Where,

- TP is number of true positives
- TN is number of true negatives
- FP is number of false positives
- FN is number of false negatives

B. Classification Results

For training SVM, we selected rbf kernel and used grid search to find optimal arguments for kernel. Results gathered using optimized paraments show considerable improvement over original results in [7] accuracy of 87.1% and precision 55.9% only. While method presented when employed, resulted in mean accuracy of $96.60\%(\pm 2.05)$ and mean precision $83.86\%(\pm 2.27)$. These results also show high standard deviation in values of evaluated metrics in 7 fold cross validation, which means classification accuracy highly depends on distribution of data. Comparison of results is presented in Table I.

IV. CONCLUSION

Employed methods have considerable improved results over [7], even though both employ same methodology and feature set. This it can be conclude that with addition of new profiles in PSI-BLAST nr database has improved PSSM calculations, which makes up major portion of feature set. For future work, I would propose employing prototype based classifiers and data pruning as classes are really imbalanced in data set (as we can see from Recall values of just 56.93%, which is very low) and using Localized SVM.

References

[1] M. Terribilini, J. hyung Lee, C. Yan, and R. L. Jernigan, "Prediction of rna binding sites in proteins from amino acid sequence," *RNA*, vol. 12, pp. 1450–1462, 2006.

[2] I. K. McDonald and J. M. Thornton, "Satisfying hydrogen bonding potential in proteins," *Journal of Molecular Biology*, vol. 238, no. 5, pp. 777 – 793, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283684713349

[3] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983. [Online]. Available: http://dx.doi.org/10.1002/bip.360221211

[4] S. F. Altschul, T. L. Madden, A. A., J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997. [Online]. Available: http://dx.doi.org/10.1093/nar/25.17.3389

[5] F. Markowetz, "Classification by support vector machines," Practical DNA Microarray Analysis, 2003.

[6] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, 2011. [Online]. Available: http://dx.doi.org/10.1145/1961189.1961199

[7] Y. Wang, Z. Xue, G. Shen, and J. Xu, "Printr: Prediction of rna binding sites in proteins using svm and profiles," *Amino Acids*, vol. 35, pp. 295–302, 2008, 10.1007/s00726-007-0634-9. [Online]. Available: http://dx.doi.org/10.1007/s00726-007-0634-9