Predicting Protein-RNA Binding Sites Using Structural Information

Cornelia Caragea, Michael Terribilini, Jivko Sinapov, Jae-Hyung Lee, Fadi Towfic,
Drena Dobbs and Vasant Honavar

RNA molecules play diverse functional and structural roles in cells: they function as messengers for transferring genetic information from DNA to proteins, as the primary genetic material in many viruses, as enzymes important for protein synthesis and RNA processing, and as essential and ubiquitous regulators of gene expression in living organisms. All of these functions depend on precisely orchestrated interactions between RNA molecules and specific proteins in cells. Understanding the molecular mechanisms by which proteins recognize and bind RNA is essential for comprehending the functional implications of these interactions, but the recognition "code" that mediates interactions between proteins and RNA is not yet understood [1].

In this study, we use machine learning algorithms for training classifiers to predict protein-RNA interfaces using information derived from the sequence. We develop a two-stage classifier, called *Struct-SVM* that takes into account structural information: in the first stage, the instances that correspond to the surface target residues (i.e., target residues that are on the surface) are separated from those that correspond to the non-surface target residues; in the second stage, if the target residue is on the surface, the classifier returns a probability that this residue is an interface residue given the sequence features as input to the classifier; otherwise, if the target residue is not on the surface, the classifier assigns this residue as a non-interface residue.

We compare the performance of *Support Vector Machines* (*SVM*) [2] with that of *Struct-SVM* classifiers on a 181 protein-RNA dataset using sequence features as input to the classifiers. The results of our experiments show that *Struct-SVM* outperforms *SVM* on the problem of predicting the protein-RNA binding interfaces in a protein sequence in terms of a range of standard measures for comparing the performance of classifiers (Fig.1, Table 1).

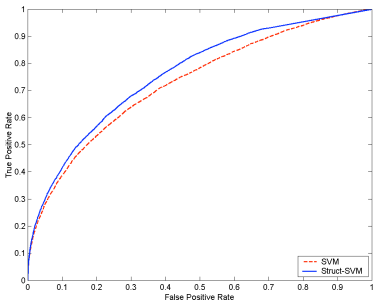

Fig. 1. Receiver Operaring Characteristi (ROC) Curves for SVM and Struct-SVM Classifiers on the protein-RNA dataset

Classifier/ PerfMeasure	SVM	Struct-SVM
Accuracy	0.68	**0.74**
Correlation Coefficient	0.25	**0.30**
Area Under ROC Curve	0.73	**0.76**

Table 1. Accuracy, Correlation Coefficient and Area Under the ROC Curves for SVM and Struct-SVM

References

[1] Chen, Y., Varani, G. (2005). Protein families and RNA recognition. Febs J 272:2088-2097.
[2] Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2:121–167, 1998.