Department of Computer Science
George Mason University
Technical Report Series

4400 University Drive MS#4A5
Fairfax, VA 22030-4444 USA
http://cs.gmu.edu/
703-993-1530

An Information Theoretic Approach for the Analysis of RNA and DNA Binding Sites.

Sheng Li
sli8@gmu.edu

Huzefa Rangwala
rangwala@cs.gmu.edu

Technical Report GMU-CS-TR-2010-7

Abstract

Proteins perform several critical biological processes by interacting with other macromolecules (DNA, RNA) and small molecules. Several computational approaches have been developed to determine the protein interaction sites using sequence and structure features. Instead of building another adhoc prediction algorithm, the purpose of this study is to understand the contribution of a protein's residue in a RNA-binding event and compare it with the DNA-binding process. We evaluate several sequence and structure-based features using mutual information theory. We show that solvent accessibility and profile-based features can be used for developing good protein-RNA binding site determination algorithms. We also recommend features that could discriminate between RNA and DNA binding sites. This work can be extended to understand protein-protein and protein-ligand interactions as well.

1 Introduction

Proteins govern several processes within the cell by interacting with other proteins, DNA, RNA and small molecules. In fact proteins are ubiquitous as well as promiscuous in terms of their interaction partners. Protein-RNA and protein-DNA interactions play a crucial role in biological processes that includes regulation of gene expression and protein synthesis [19, 15].

The number of experimentally available protein-RNA complexes in the Protein Data Bank (PDB) [4] are relatively few and biased towards certain families of complexes [20]. As such several computational methods have been developed over the years to determine which proteins interact with RNA [20] molecules, and specifically which protein residues are involved in the interaction. These interacting residues involved are interchangeably referred to as binding sites or contact sites in this paper. Similar computational methods have been developed for determining DNA binding proteins and interacting residues [5, 23, 1, 2].

These methods determine binding sites using geometric approaches if three-dimensional structure of a protein is available or use a statistical learning based approach to predict the binding sites using information derived from the protein sequence only. For example, RNABindR [20] uses a naive Bayesian classifier to predict the RNA binding sites from sequence only. BindN [22] is a support vector machine (SVM) based classifier that uses physicochemical properties of residues to predict DNA-binding and RNA-binding residues. KYG [13] uses structure information along with evolutionary information to determine RNA-binding sites for proteins whose structures are known but have not been co-crystallized. TCBRP [8] is a useful web service that parses the PDB to find complexes and determines the binding residues by computing the atomic distance between the protein atoms and the interacting partner's atoms.

In this paper we take an information theoretic approach to understand the contribution of a residue's contribution in a RNA-binding event. Using mutual information we evaluate various sequence-based and structure-based features derived for a protein residues to determine the likelihood of specific features in determining RNA-binding sites. We compare the MI values obtained for single or a combination of paired features to a previous study on protein-DNA interaction analysis [12].

We determine features that could be used for developing effective prediction algorithms to determine protein-

RNA binding sites. We also provide details on specific features like solvent accessibility that could be used for discriminating between protein-RNA and protein-DNA interactions.

2 Methods

2.1 Mutual Information (MI)

In this study we evaluate the strengths of different sequence and local structure features for predicting protein-RNA interaction sites. As done previously in the analysis of protein-DNA interaction sites, [12] we use mutual information (MI) to capture the inter-dependence between two variables: (i) features being evaluated and (ii) protein-RNA interaction sites. Mutual information for two discrete random variables X and Y is given by

$$MI(X,Y) = \sum_{(x \in X, y \in Y)} p(x,y) * \log_2(\frac{p(x,y)}{p(x)p(y)}), \quad (1)$$

where x and y are the discrete values taken for the random variables X and Y, respectively. $p(x)$ is the probability of taking the discrete value x and $p(x,y)$ is the probability that x and y occur together. Mutual information takes on the unit of bits due to the base-two logarithm.

We use the variable X to represent whether the residue is an interaction site (protein-RNA binding) or not. Thus X is always binary whereas random variable Y is used for representing the various features that are evaluate in this study. The MI between the RNA interaction feature and the other features provides us with an estimate of the usefulness of the features in predicting the interacting sites. We also compare the MI values for the same features obtained from the protein-DNA interaction data. This allows us to compare the protein-RNA interactions with protein-DNA interactions. We also perform an experiment (See Section 3.3) where the protein-RNA and protein-DNA datasets are pooled together. In this case the variable X can take three values.

2.2 Defining Binding Sites

To compute the mutual information we need to determine for the variable X whether the residue is a binding (interaction/contacting) site or not. For determining the RNA-binding residues we compute the distance between each atom of a residue in the protein and each atom in the RNA structures of the protein-RNA complex file. The smallest distance is taken as the distance between the protein residue and RNA macromolecule. Based on a distance cutoff we define a residue to be RNA binding or not. We increment the distance cutoff (starting at 0 Angstroms) in steps of 0.25 and compute the mutual information with all the different features at every step. This allows us to plot a curve for every feature showing the characteristics of the signal separating the RNA binding residues from the non-binding residues. Mutual information is set to zero when a combination of values for the feature does not exist. This can happen when the feature can take on large values or when the distance cutoff is very small or very high. A similar procedure for defining the DNA-binding sites was followed by Kauffman et. al. [12].

2.3 Feature Description

Table 1 summarizes the various sequence-based and structure-based features that were evaluated in this study. Some of these features are discrete in nature, whereas some are continuous or vectors. To compute the mutual information (Equation 1) we need to convert the continuous and vector valued features into discrete variables. Continuous valued features can be broken into discrete bins by defining boundaries. To compare our work with the protein-DNA interaction study [12], for the continuous valued features we choose the same boundary definitions as done in that study. We verified on a smaller sample dataset that the boundary definitions were generally the same if chosen to maximize the mutual information. The vector-valued features can be discretized by using a clustering algorithm (described below).

2.3.1 Sequence and Profile-based Features

We use the amino acid composition (AAC) for a residue. This feature is inherently discrete in nature (has twenty different values that a natural amino acid can take).

Profiles capture evolutionary information that has shown to be useful in a wide range of protein sequence prediction problems like remote homology detection [17] and local structure prediction problems [18]. The profile of a protein X is derived by computing a multiple sequence alignment of X with a set of sequences $\{Y_1, \ldots, Y_m\}$ that have a statistically significant sequence similarity with X (i.e., they are sequence homologs).

We obtain the profiles using PSI-BLAST [3] as it combines both steps, is very fast, and has been shown to produce reasonably good results. The profile of a sequence X of length n is represented by two $n \times 20$ matrices. The first is its position-specific scoring matrix PSSM that is computed directly by PSI-BLAST using the scheme described in [3]. The rows of this matrix correspond to the various positions in X and the columns correspond to the 20 distinct amino acids. The second matrix is its position-specific *frequency* matrix PSFM that contains the frequencies used by PSI-BLAST to derive PSSM. We collectively refer to the two matrices as “Profiles” in the study.

However the PSSM and Profile features are vectors of length 20 and 40 per residue, respectively. We use a clustering based approach to discretize these vector-based features. Specifically, we use CLUTO [11] (version 2.1.2) with default options to create various number of clusters. Each cluster serves as a discrete value for the vector-valued feature. This clustering based approach was used in the analysis of protein-DNA study as well [12].

The PSI-BLAST output also provides a measure of the sequence diversity per column of the profile known as information per position (IPP). Low values indicate less diversity which shows a strong preference for particular amino acids in that position. We discretize the continuous values into either two, three or four bins with the cutoff values shown in Table 1. These boundary definitions were determined in the protein-DNA interaction study [12]. We found these boundary definitions to produce the maximum MI value after a grid search on small held out protein-RNA interaction dataset.

Since the functional and structural properties of residues (in this case binding) are highly dependent on the local sequential residues [18] as well as local spatial residues (structurally close) we evaluate the features using the sequentially neighboring residues. For a residue x_i at position i we define a $(2w+1)$-length subsequence called a $w\mathtt{mer}$ consisting of x_i and w residues immediately to the left and right of the residue x_i. To evaluate the PSSM features we concatenate the 20-length feature vectors obtained for each residue within the $w\mathtt{mer}$. This feature vector is denoted as WPSSM and produces a vector of length $(2w+1)\times 20$. Using the clustering approach described earlier we discretize the vectors into clusters of size 5, 10 or 20.

2.3.2 Local Structure-based Features

Protein residues involved in binding often taken on particular local structural shapes or configurations. In this paper we compared the local structural properties for residues involved in binding with DNA and RNA macromolecules.

Even the secondary structure (SS) definition of the protein derived using the program DSSP [10] produces three discrete classes denoting the most commonly occurring local topological structures, namely the alpha helix, beta sheets, and coil regions [10].

A universal definition of local structure is the secondary structure which captures recurring, locally occurring shapes in the PDB. Researchers have developed several sequence-based prediction programs like PSIPRED [9] to accurately predict the secondary structure of residues upto 80% accuracy. Using the DSSP program [10] we parse the three-dimension structure files to generate the three secondary structure classes: (i) alpha helix, (ii) beta sheets, and (iii) coil regions. We use the three discrete class labels as secondary structure features (SS) to evaluate their relationship to RNA and DNA binding residues.

From the DSSP program we also extract the solvent accessibility surface area (SASA) which provides the surface area of a residue accessible to solvent (water) molecules. The SASA values are normalized based on the maximum SASA of a residue in Gly-X-Gly calculated using the values of Miller et. al. [14]. Both the SASA and SS features though determined using three-dimensional structure of a protein can be predicted accurately from sequence information using tools like svmPRAT [18].

We evaluate two other features: (i) the amino acid composition and (ii) PSSMs (captures evolutionary information) by using the spatially proximal residues to the residues of interest. We determine the structural neighborhood of a particular residue by using a 14 Angstrom cutoff distance between the C_α atoms. The amino acid composition feature is aggregated across the neighborhood residues and is denoted as StrN in this study. We also aggregate a twenty length PSSM vector and denote the same as StrN-PSSM. Residues that are sequentially less than 3 amino acids apart are neglected for evaluation.

2.4 Joint Features

The binding properties will generally be determined by not one single features but a gamut of features combined together. As such, prediction algorithms developed use a range of features together to achieve the best accuracy [13]. We combine pairs of features to compute the mutual information. The paired features take values that are every possible combination of the values taken by the individual features. As such the size of joint feature space is the product of the sizes of the individual features. This makes the problem intractable beyond two features. For example, combining the amino acid composition (AAC) with 20 labels and the PSSM features with 20 cluster labels leads to 400 values. In Table 3 we show the joint features that were studied in this paper along with the mutual information values.

2.5 Datasets

2.5.1 RNA Interaction Dataset

The RNA interaction dataset used in this study was extracted from the web-based binding site detection tool TCBRP [8]. TCBRP provides us a list of 546 protein-RNA co-crystal PDB files. Each co-crystal file may contain several chains with identical sequence which can

Table 1: Residue Features Considered for Mutual Information with RNA-contacting classes.

Feature	**Notation**	**Description**	**Discretization**
Amino-Acid Composition	AAC	Amino acid residue type	20 values
Information Per Position	IPP	PSI-BLAST [3] produced column that computes sequence divergence per position. Lower value indicates stronger preference for certain amino acids	2-values: 0.0-1.15, >1.15 3-values: 0.0-0.65,0.65-1.15, >1.15 4-values: 0-0.25,0.25-0.65, 0.65-1.15,>1.15
Secondary Structure	SS	Secondary structure assigned to a residue by DSSP	3 values: alpha-helix, beta-sheet and coil/others
Position Specific Score Matrix	PSSM	Only the PSSM from the PSI-BLAST profile	5, 10, and 20 clusters
Concatenated PSSM	WPSSM	The PSSMs of residues within a sliding window of size 5 concatenated together	5, 10 and 20 clusters
Profiles	Profiles	Combination of the PSI-BLAST (3 iteration against the NR database) derived Position Specific Scoring Matrix (PSSM) and the Position Specific Frequency Matrix (PSFM)	5, 10 and 20 clusters
Solvent Access Surface Area	SASA	DSSP computed values of a residue's exposure to solvent molecules.	2-values: 0.00-0.20, >0.20 3-values: 0.00-0.07, 0.07-0.20, >0.20 4-values: 0.0-0.01, 0.01-0.07, 0.07-0.20, >0.20
Structural Neighbors	StrN	Sum of amino acid types within a 14 Angstrom sphere. The distance is between the center of mass of two residues	5, 10 and 20 clusters
Structural Neighbors PSSM	StrN-PSSM	Sum of PSSM for all residues within 14 Angstrom radius	5, 10 and 20 clusters

lead to unfair bias when computing the mutual information values. We use the PISCES culling server [21] so as to create a non-redundant dataset such that no pairs of protein chains within the dataset have greater than 30% sequence identity. This results in a dataset of 143 protein chains with 40,925 protein residues. In Figure 1 we show the percentage of RNA contacting residues according to the different distance cutoff values (plot in green).

2.5.2 DNA Interaction Dataset

For the DNA interaction dataset we use the protein lists provided by Kauffman et. al. [12]. The dataset consists of 246 different chains and 51,268 residues culled using PISCES [21] as described above to not include any pairs having greater than 30% identity. In Figure 1 we show the percentage of DNA contacting residues based on the increasing distance cutoff values (plot in red).

3 Results And Discussion

3.1 Single Features

The mutual information values obtained for the individual features is in the order of hundreths of bits. This range is consistent with the values obtained for the previous protein-DNA interaction study [12], pairwise contact potentials [6] and sequence-structure correlations [7]. As observed in the protein-DNA interaction study [12] increasing the number of clusters for vector valued features leads to an increase in the mutual information values. We experiment with 20, 10, and 5 clusters for the vector-based PSSM, Profiles, WPSSM, StrN and StrN-PSSM features.

In Table 2 we report the maximum obtained MI score along with the distance cutoff for the features evaluated individually on the protein-RNA study. These results are sorted in decreasing values of MI obtained. We also report the corresponding results from the previous protein-DNA interaction study [12]. Figure 2 shows how mutual information for some of the representative features varies as the protein-RNA distance cutoff defining the contacting residues is increased. The results are shown in decreasing MI values for the protein-RNA dataset.

We observe that a combination of structural features (SASA and StrN, StrN-PSSM) along with profile-based sequence features (PSSM and Profiles) show the highest bit scores. This is different from the protein-DNA interaction dataset where sequence-based and sequence-derived features had higher MI values in comparison to the structure-based features. The highest MI obtained for the protein-RNA interaction data is for the solvent accessibility surface area feature (SASA with four discrete labels) is 4.44×10^{-2} bits at a distance cutoff of 4.25 Angstroms, whereas for the protein-DNA interaction dataset it is only 1.41×10^{-2} at a cutoff of 3.77 Angstroms. This suggests that the two types of interactions vary based on the number of residues that are accessible on the surface. For both the interaction datasets the secondary structure shows the lowest MI values suggesting no particular preference for a particular local topo-

Table 2: Mutual Information for Single Features on RNA-binding and DNA-binding datasets.

		protein-RNA		protein-DNA	
Feature	N_{val}	MI	DC	MI	DC
SASA	4	**4.44** $\times 10^{-2}$	4.25	1.41 $\times 10^{-2}$	3.77
PSSM	20	**3.94** $\times 10^{-2}$	5.25	**3.23** $\times 10^{-2}$	4.97
Profiles	20	**3.82** $\times 10^{-2}$	4.25	**3.18** $\times 10^{-2}$	4.97
StrN	20	3.57 $\times 10^{-2}$	5.5	2.26 $\times 10^{-2}$	8.57
StrN-PSSM	20	3.11 $\times 10^{-2}$	6	2.67 $\times 10^{-2}$	10.17
PSSM	10	2.87 $\times 10^{-2}$	5	2.43 $\times 10^{-2}$	4.97
StrN	10	2.83 $\times 10^{-2}$	6.25	1.82 $\times 10^{-2}$	7.17
Profiles	10	2.73 $\times 10^{-2}$	5.75	2.67 $\times 10^{-2}$	4.77
AAC	20	2.56 $\times 10^{-2}$	3.5	**2.91** $\times 10^{-2}$	3.37
PSSM	5	2.43 $\times 10^{-2}$	4.75	1.98 $\times 10^{-2}$	4.97
StrN-PSSM	10	2.41 $\times 10^{-2}$	5.25	1.96 $\times 10^{-2}$	9.57
IPP	4	2.19 $\times 10^{-2}$	13.25	1.26 $\times 10^{-2}$	9.57
Profiles	5	2.04 $\times 10^{-2}$	5.25	1.29 $\times 10^{-2}$	3.97
StrN	5	1.95 $\times 10^{-2}$	6	1.48 $\times 10^{-2}$	6.97
StrN-PSSM	5	1.91 $\times 10^{-2}$	7	1.27 $\times 10^{-2}$	9.57
WPSSM	20	1.83 $\times 10^{-2}$	5.5	1.69 $\times 10^{-2}$	5.17
WPSSM	10	1.69 $\times 10^{-2}$	5.25	1.57 $\times 10^{-2}$	4.97
WPSSM	5	1.03 $\times 10^{-2}$	5.25	1.11 $\times 10^{-2}$	4.97
SS	3	4.40 $\times 10^{-3}$	6.25	2.50 $\times 10^{-3}$	5.77

MI and DC denotes Mutual Information (in bits) and Distance Cutoff (in Angstrom), respectively. N_{val} denotes the total number of discrete values for the feature. We report the largest MI value obtained for a distance cutoff that denotes a residue to be in the contacting or non-contacting class. We ran the protein-DNA results ourselves and obtain the reported results in the study [12]. The three largest MI values for both protein-RNA and protein-DNA datasets are highlighted in bold

Table 3: Mutual Information for Joint Features on RNA-binding and DNA-binding datasets.

Joint Features					protein-RNA		protein-DNA	
Feature 1	N_{val1}	Feature 2	N_{val2}	N_{tot}	MI	DC	MI	DC
AAC	20	SASA	4	80	**6.44** $\times 10^{-2}$	4	4.037 $\times 10^{-2}$	3.77
PSSM	20	SASA	4	80	**6.09** $\times 10^{-2}$	4.25	4.389 $\times 10^{-2}$	3.97
PSSM	20	StrN	20	400	**5.86** $\times 10^{-2}$	6	**5.278** $\times 10^{-2}$	5.77
PSSM	20	StrN	10	200	5.68 $\times 10^{-2}$	5.5	**4.756** $\times 10^{-2}$	5.37
Profile	20	StrN	10	200	5.61 $\times 10^{-2}$	6	**4.691** $\times 10^{-2}$	6.57
Profile	20	SASA	4	80	5.41 $\times 10^{-2}$	4.5	4.464 $\times 10^{-2}$	4.17
Profile	10	StrN	20	200	4.85 $\times 10^{-2}$	6.25	4.555 $\times 10^{-2}$	5.97
Profile	10	IPP	4	40	4.59 $\times 10^{-2}$	6.75	4.239 $\times 10^{-2}$	5.37
PSSM	20	IPP	4	80	4.28 $\times 10^{-2}$	5.75	4.495 $\times 10^{-2}$	4.97
PSSM	10	StrN	20	200	4.25 $\times 10^{-2}$	11	4.558 $\times 10^{-2}$	5.97
Profile	20	StrN-PSSM	5	100	3.68 $\times 10^{-2}$	5.75	3.951 $\times 10^{-2}$	5.37
AAC	20	SS	3	60	3.02 $\times 10^{-2}$	3.5	3.234 $\times 10^{-2}$	3.57
AAC	20	IPP	4	80	2.85 $\times 10^{-2}$	5.75	4.263 $\times 10^{-2}$	3.57

MI and DC denotes Mutual Information (in bits) and Distance Cutoff (in Angstrom), respectively. N_{val1}, N_{val2} and N_{tot} denotes the total number of discrete values for the Feature 1, Feature 2 and the joint feature, respectively. We report the largest MI value obtained for a distance cutoff that denotes a residue to be in the contacting or non-contacting class. We ran the protein-DNA results ourselves and obtain the reported results in the study [12]. The three largest MI values for both protein-RNA and protein-DNA datasets are highlighted in bold.

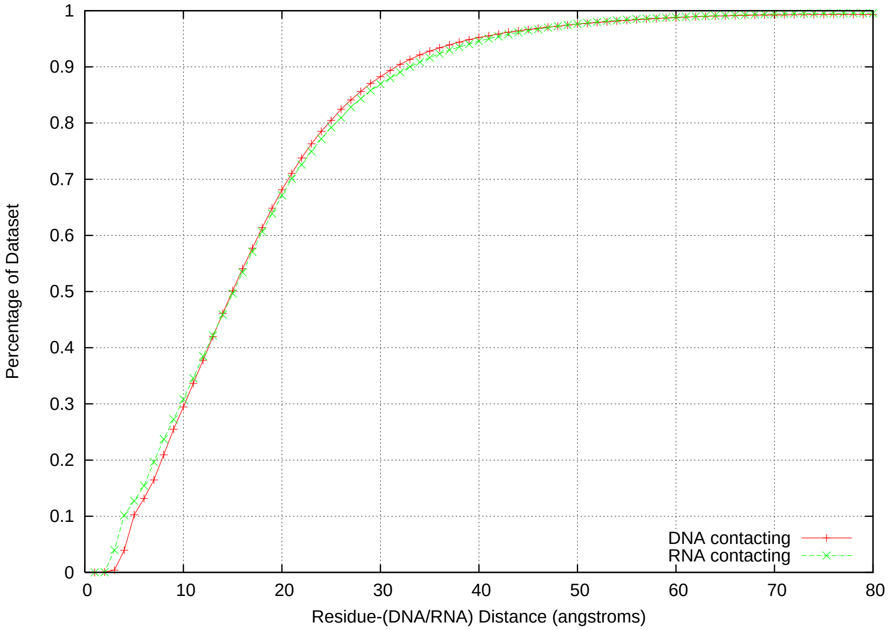

Figure 1: Percentage of Contacting Residues vs. Distance Cutoff in Angstrom.

logical shape when the protein residues are involved in binding.

A interesting result is the low MI value for the $w$mer based concatenated PSSM in comparison to the single residue based PSSM. Several prediction algorithms [18] use a $w$mer based subsequence window to capture information for a residue from its sequential neighbors to predict local structure and functional properties like disorder, ligand-binding and secondary structure prediction. We experimented with several values of w and found setting the window size as 5 to produce the largest MI value. The current method of discretization could be a reason for the low MI values for the WPSSM features but needs further investigation.

3.2 Joint Features

For the joint features we summarize a sample of the various combinations that were tested. In Table 3 we report the largest MI value obtained from the combination of two features for the protein-RNA and protein-DNA interaction study along with the distance cutoff. The combination of SASA (highest MI amongst single valued in Table 2) along with PSSM and AAC features (total 80 features) leads to the highest joint MI values for the protein-RNA interaction study of 0.0644 and 0.0609 bits, respectively. In comparison for the protein-DNA interaction study we notice the highest joint MI value of 0.05278 for the combination of PSSM and StrN (total 400 features).

Figure 3 shows the trend of the MI value for the different joint features on the protein-RNA interaction dataset with increasing distance cutoff values. The combination of information per position along with PSSMs and Profiles leads to an increased MI value. Combining the secondary structure (SS) with other features does not lead to a large increase in the MI values.

3.3 Pooling RNA and DNA Binding Datasets

We also performed an experiment where we pooled the RNA and DNA interaction datasets. We can characterize how the features varied across the two different types of interactions. In the pooled experiment for computing the mutual information we defined three different discrete values for the variable X in Equation 1: (i) RNA contacting residue, (ii) DNA contacting residue and (iii) non-contacting residue. As done in the previous experiments we compute the MI value for the different single

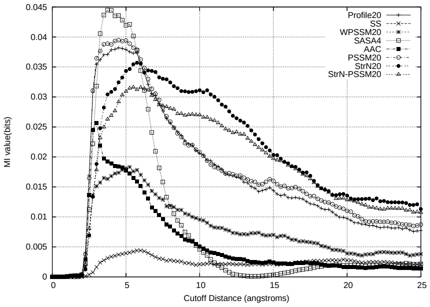

Figure 2: RNA-contact distance cutoff (unit is Angstrom) vs. mutual information (unit is bits) for single features. The cutoff distance which defines RNA-contacting versus non-contacting residues is incremented to characterize individual features and their MI with RNA-contacting classes.

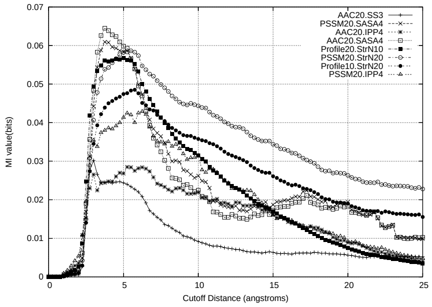

Figure 3: RNA-contact distance cutoff (unit is Angstrom) vs. mutual information (unit is bits) for joint features. The cutoff distance which defines RNA-contacting versus non-contacting residues is incremented to characterize paired combination of individual features and their MI with RNA-contacting classes.

and joint features with increasing distance cutoff values. It was ensured that there was no overlap in the chains involved in the RNA and DNA interaction i.e., we restrict the analysis to residues involved only in one type of interaction. The pooled results for the single and joint features are shown in Figures 4 and 5, respectively. The MI values for this pooled experiment can be compared to the results obtained for the individual protein-RNA and protein-DNA datasets.

As expected the solvent accessibility feature (SASA) stands out individually (peak MI value of 0.048 bits) and in combination with the PSSM feature produces the highest MI values (MI value of 0.092 bits) . This suggests that SASA can be important in distinguishing between the RNA and DNA binding properties of residues. From a pure sequence based perspective we can use prediction methods like svmPRAT [18] or ACCPRO [16] to predict SASA features from sequence and use the predicted values to discriminate between the RNA and DNA binding residues.

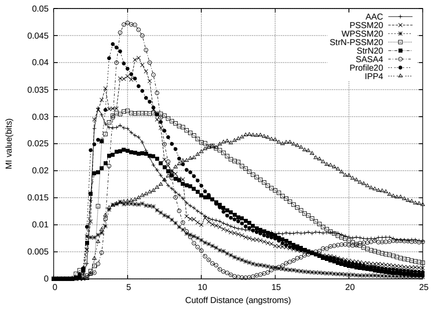

Figure 4: Contact distance cutoff (unit is Angstrom) vs. mutual information (unit is bits) for single features on a pooled RNA and DNA binding dataset.

4 Conclusions and Future Directions

Using the mutual information as a measure of dependence we determine specific properties of protein residues that are involved in RNA-binding and DNA-binding events. We observe the strong relationship between the relative solvent accessibility surface area and protein-RNA binding site. Comparing these features to the protein-DNA binding sites also helps us characterize the nature of interaction and difficulty of predicting binding sites from sequence. Since the high scoring features are structural in nature, it may lead to the conclusion that protein-RNA interaction sites may be hard to predict using just sequence information. However, we could predict features like contact order (i.e., number of protein residues in close proximity to another residue) as well as solvent accessibility if we were to build a sequence-based predictive model. Using the information learned from this study we would like to build a machine learning classifier to predict protein-RNA binding sites.

The results of the concatenated PSSM were surprising since the use of subsequence or local sequence information always improves results for predicting local structure [18]. In the future we would like to use this approach to study protein-protein interactions as well as

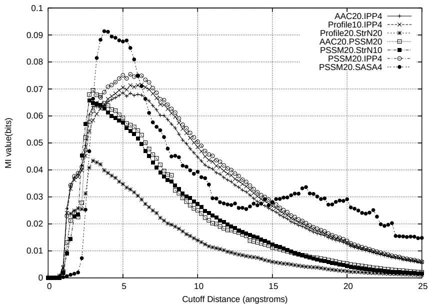

Figure 5: Contact distance cutoff (unit is Angstrom) vs. mutual information (unit is bits) for joint features on a pooled RNA and DNA binding dataset.

protein and small molecule interaction.

References

[1] S. Ahmad, M. M. Gromiha, and A. Sarai. Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–486, Mar 2004.

[2] S. Ahmad and A. Sarai. Pssm-based prediction of dna binding sites in proteins. *BMC Bioinformatics*, 6:33, 2005.

[3] S. F. Altschul, L. T. Madden, A. A. Schffer, 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*, 25(17):3389–402, 1997.

[4] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. G. H. Weissig, and J. Westbrook. The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, 7:957–959, November 2000.

[5] N. Bhardwaj and H. Lu. Residue-level prediction of dna-binding sites and its application on dna-binding protein predictions. *FEBS Letters*, 581:1058–1066, Mar 2007.

[6] M. S. Cline, K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers, and D. Haussler. Information-theoretic dissection of pairwise contact potentials. *Proteins*, 49(1):7–14, Oct 2002.

[7] G. E. Crooks, J. Wolfe, and S. E. Brenner. Measurements of protein sequence-structure correlations. *Proteins: Structure, Function, and Bioinformatics*, 57:804–810, 2004.

[8] J. Hu and C. Yan. A tool for calculating binding-site residues on proteins from pdb structures. *BMC Structural Biology*, 9(1):52, 2009.

[9] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matricies. *J. Mol. Biol.*, 292:195–202, 1999.

[10] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637, 1983.

[11] G. Karypis. Cluto: A clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. (http://www.cs.umn.edu/˜cluto).

[12] C. Kauffman and G. Karypis. An analysis of information content present in protein-dna interactions. In *Pacific Symposium on Biocomputing*, Hawai, 2008. (in press).

[13] O. T. P. Kim, K. Yura, and N. Go. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucl. Acids Res.*, 34(22):6450–6460, 2006.

[14] S. Miller, J. Janin, A. M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, 196:641–656, Aug 1987.

[15] H. F. Noller. RNA Structure: Reading the Ribosome. *Science*, 309(5740):1508–1514, 2005.

[16] G. Pollastri, P. Baldi, P. Farselli, and R. Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins: Structure, Function, and Genetics*, 47:142–153, 2002.

[17] H. Rangwala and G. Karypis. Profile based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21(23):4239–4247, 2005.

[18] H. Rangwala, C. Kauffman, and G. Karypis. svmprat: Svm-based protein residue annotation toolkit. *BMC Bioinformatics*, 10(1):439, 2009.

[19] J. M. S. and M. M. J. Pre-mrna splicing: awash in a sea of proteins. *Molecular Cell*, 12:5–14, 2003.

[20] M. Terribilini, J. D. Sander, J.-H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucl. Acids Res.*, 35(suppl_2):W578–584, 2007.

[21] G. Wang and J. Dunbrack, Roland L. PISCES: recent improvements to a PDB sequence culling server. *Nucl. Acids Res.*, 33(suppl_2):W94–98, 2005.

[22] L. Wang and S. J. Brown. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucl. Acids Res.*, 34(suppl_2):W243–248, 2006.

[23] L. Wang, M. Yang, and J. Yang. Prediction of dna-binding residues from protein sequence information using random forests. *BMC Genomics*, 10(Suppl 1):S1, 2009.