
Struct-NB: predicting protein-RNA binding sites using structural features

Fadi Towfic*

Bioinformatics and Computational Biology Graduate Program,
Iowa State University,
Ames, IA 50011-1040, USA
E-mail: ftowfic@cs.iastate.edu
*Corresponding author

Cornelia Caragea

Department of Computer Science,
Iowa State University,
Ames, IA 50011-1040, USA
E-mail: cornelia@cs.iastate.edu

David C. Gemperline

Department of Biology, Department of Chemistry,
Carthage College,
2001 Alford Park Drive, Kenosha, WI 53140-1994, USA
E-mail: dcgemperline@gmail.com

Drena Dobbs

Department of Genetics, Development and Cell Biology,
Bioinformatics and Computational Biology Graduate Program,
Iowa State University,
Ames, IA 50011-1040, USA
E-mail: ddobbs@iastate.edu

Vasant Honavar

Department of Computer Science,
Bioinformatics and Computational Biology Graduate Program,
Iowa State University,
Ames, IA 50011-1040, USA
E-mail: honavar@cs.iastate.edu

Abstract: We analyse sequence and structural features of protein-RNA interfaces using RB-147, a non-redundant dataset of protein-RNA complexes extracted from the PDB. We train classifiers using machine learning algorithms to predict protein-RNA interfaces from sequence and structure-derived features of proteins. Our experiments show that Struct-NB, a Naïve Bayes classifier that exploits structural features, outperforms its counterparts that use only sequence features to predict protein-RNA binding residues.

Keywords: protein-RNA interactions; propensity; structural features.

Reference to this paper should be made as follows: Towfic, F., Caragea, C., Gemperline, D.C., Dobbs, D. and Honavar, V. (2010) 'Struct-NB: predicting protein-RNA binding sites using structural features', *Int. J. Data Mining and Bioinformatics*, Vol. 4, No. 1, pp.21–43.

Biographical notes: Fadi Towfic is a PhD student in the Bioinformatics and Computational Biology Program at Iowa State University. He received dual BS Degrees from the University of Iowa in Computer Science and Biology. His current research interests include construction and analysis of protein-protein interaction networks.

Cornelia Caragea is a PhD candidate in the Department of Computer Science at Iowa State University. She received her BS in Computer Science from the University of Bucharest, Romania. Her research focuses on development of probabilistic graphical models, relational learning, and learning predictive models from distributed data with emphasis on applications to problems in bioinformatics.

David C. Gemperline is a Senior Undergraduate at Carthage College. He participated in the 2007 Computational and Systems Biology Summer Institute (CSBSI) at Iowa State University.

Drena Dobbs received her PhD from the University of Oregon in 1983. She was an NIH Postdoctoral fellow in Molecular Biology at the University of California, Berkeley before joining Iowa State University as an Assistant Professor in 1987. From 1999 through 2001, she served as the Chair of the Bioinformatics and Computational Biology (BCB) PhD Program. She is currently a Professor in the Department of Genetics, Development, and Cell Biology at Iowa State University. Her research interests include macromolecular structure, interactions, and function.

Vasant Honavar received his PhD in Computer Science and Cognitive Science in 1990 from the University of Wisconsin (Madison). He joined the Faculty of Computer Science at Iowa State University (ISU) as an Assistant Professor in 1990 where he has been a full Professor since 2001. During 2003–2005, he served as the Chair of the Bioinformatics and Computational Biology (BCB) PhD program. He currently Directs the Center for Computational Intelligence, Learning and Discovery. His research interests include machine learning, bioinformatics, data mining, knowledge representation and inference, and semantic web. He has published over 200 papers in refereed journals and conferences.

1 Introduction

Protein-RNA interactions play a vital role in RNA splicing, translation, replication of many viruses as well as many other processes in the cell. The prediction of protein-RNA interfaces can aid in the design of drug-inhibitors for viruses, down-regulation of unwanted genes as well as contributing to our basic understanding

of the mechanisms involved in protein-RNA recognition (Freed and Mouland, 2006; Jurica and Moore, 2003; Moore, 2005; Noller, 2005). At least nine families of RNA-binding proteins have been identified using sequence-based analyses of the major groups of RNA-binding proteins, together with functional characterisation of mutations that affect the specificity or affinity of RNA binding (for review, see Chen and Varani, 2005). In contrast, the number of experimentally determined structures for protein-RNA complexes is still relatively small and heavily biased (ribosomal proteins represent 50% of all RNA binding proteins in the Protein Data Bank (PDB) (Berman et al., 2000).

Because of the importance of protein-RNA interactions in biological regulation and the considerable effort required to identify RNA binding residues through biophysical analyses of protein-RNA complexes or *in vitro* binding studies, there is an urgent need for computational methods to identify RNA binding sites given a protein's primary amino acid sequence, and when available, its 3-dimensional structure. Several recent studies have focused on the development of machine learning approaches to amino acid sequence-based prediction of RNA-binding residues in proteins (Terribilini et al., 2006b, 2007; Jeong et al., 2004; Jeong and Miyano, 2006). The predictions obtained using such methods have already contributed to the design of wet-lab experiments to decipher mechanisms of protein-RNA recognition (Terribilini et al., 2006a; Bechara et al., 2007). However, the machine learning approaches to prediction of RNA-binding residues of proteins have focused largely on the analysis of amino acid sequence as opposed to the structural features of the protein chain. Other analyses of protein-RNA interfaces (Jones et al., 2001; Treger and Westhof, 2001; Lejeune et al., 2005) have focused on the analysis of hydrogen bonds or van der Waals contacts in between the protein and the RNA. There has been relatively little attention paid to structural features of the interface (e.g., protrusion or roughness) rather than the atomic forces.

Against this background, it is natural to ask: *Do protein-RNA interfaces differ from non-interfaces in terms of their structural features? Do the structural features vary according to the type of the bound RNA (e.g., mRNA, siRNA etc.)? Can structural features be utilised to improve classification performance of protein-RNA interfaces relative to sequence-based classifiers?* If we find that the protein-RNA interfaces differ from noninterfaces in terms of their structural features, then the structural features can be exploited by machine learning approaches to predict protein-RNA interface residues when the structure of the protein is available but the structures of the complexes formed by the protein with RNA are not. If the different classes of protein-RNA interfaces significantly differ from each other with respect to their structural features, it might be possible to improve the specificity and sensitivity of protein-RNA interface residue prediction by training separate classifiers for each type of bound RNA.

We describe an analysis of the structural features of protein chains from RNA-binding proteins that explores this question using a non-redundant dataset of 147 protein chains from the RB147 dataset (Terribilini et al., 2007). We focus on two of the six structural properties of amino acid residues used in a recent analysis of protein-protein interfaces by Wu et al. (2007), namely, surface roughness (Lewis and Rees, 1985) and CX value (Pintar et al., 2002). Solid Angle (Connolly, 1986) was also used early in this study (see Towfic et al. (2007)). However, it was deemed unnecessary to include in this study since the results from Solid Angle overlap with those of Roughness (Lewis and Rees, 1985) with a correlation of 0.88 (roughness and CX overlap with a correlation of -0.56).

The results of our analysis show that protein-RNA interface residues tend to be protruding compared to non-interface residues. Furthermore, interface residues tend to have rough surfaces. Our analysis also shows that the protein chains in protein-RNA interfaces containing Viral-RNA and rRNA significantly differ from those that contain dsRNA, mRNA, siRNA, snRNA, SRP RNA and tRNA with respect to their CX values. We developed *Struct-NB* classifiers to demonstrate the utilisation of the structural features in predicting protein-RNA interface residues in a protein sequence.

The rest of the paper is organised as follows: Section 2 describes the RB147 dataset and each of two properties of amino acid residues examined in this study as well as the methods used in the construction and evaluation of the *Struct-NB* classifier. Section 3 presents the results of our analysis, comparing interface and noninterface residues based on these two properties and comparing the various *Naïve Bayes* (NB) and *Struct-NB* classifiers constructed using sequence and structural features. Section 4 concludes with a summary and an outline of some directions for further research.

2 Materials and methods

2.1 Dataset

The RB147 dataset (Terribilini et al., 2007) used in this study contains protein chains extracted from structures of protein-RNA complexes in the PDB solved by X-ray crystallography, after eliminating protein chains from structures with resolution worse than 3.5 Å and protein chains sharing a sequence identity greater than 30% with one or more other protein chains. The RB147 dataset contains 147 non-redundant protein chains and a total of 32,324 amino acids. The RNA-binding residues are defined as follows: an RNA-binding residue is an amino acid containing at least one atom within 5 Å of any atom in the bound RNA. According to this definition, RB147 contains a total of 6,157 RNA-binding residues and 26,167 non-binding residues.

2.2 Classification of the protein chains based on the type of the bound RNA

The protein chains in the RB147 dataset were classified into nine classes according to the type(s) of RNA that was found in the corresponding protein-RNA complex based on a taxonomy of RNA types used previously by Ellis et al. (2007): dsRNA, mRNA, rRNA, siRNA, snRNA, SRP RNA, tRNA, Viral RNA or 'other' (which denoted synthetic RNAs or pre-mRNAs or a class of RNAs not included in any of the other categories). The classification for each PDB id and chain in the dataset is shown in Table 1. Over half of the protein chains belong to complexes with rRNA, with tRNA, mRNA and viral RNA being the other dominant groups (in that order).

2.3 Analysis of structural properties

Each chain in the dataset was analysed in terms of its surface roughness (Lewis and Rees, 1985) and CX value (Pintar et al., 2002). The analysis was repeated on subsets of the dataset corresponding to the classification based on the type of the RNA found in the interface (see Table 1). We implemented a program in Java,

Structure-Analyser 1.0 (available at <http://www.public.iastate.edu/~ftowfic>) for this analysis. The Java package has an easy-to-use API to allow its use in other applications. The program generates a standard tab-delimited output file with the PDBID, chain name, residue name (three letter abbreviation), residue number, a + or - indicating whether or not the residue is part of the interface, a score derived from the structural property being examined (roughness and cx) and a + or - denoting whether or not the residue is part of the surface of the protein (the definition of a surface residue can be varied within the class as desired). In our analysis, surface residues are defined as residues that have a solvent accessible surface area that is at least 5% of their total surface area (Wu et al., 2007; Connolly, 1993).

2.4 Roughness calculation

The roughness value for a residue denotes the degree of irregularity of that point at the surface as outlined by Lewis and Rees (1985) and Lee and Richards (1971). The surface roughness value (D) is given by:

$$D = 2 - \frac{\partial \log A_s}{\partial \log R}.$$

The roughness calculation requires a molecule surface area (A_s), which is obtained by rolling a sphere with radius R against the protein and calculating the area of the resulting surface as implemented in the MSP software package (Connolly, 1993). The radius R of the sphere is varied from 0.2 to 4.0 Å in 0.1 increments. Thus, we get the surface area for each choice of the radius which in turn can be used to calculate the partial derivative $\frac{\partial \log A_s}{\partial \log R}$ for each residue. Thus, a final roughness value (D) is associated with each residue on the surface of the protein. For a perfectly smooth surface $D = 2$ whereas for a rough surface, $D > 2$.

2.5 CX value calculation

The CX value measures the ratio of the volume of atoms that occupy a 6 Å sphere compared to the empty volume within the sphere (Pintar et al., 2002). The CX score is calculated by centring a 6 Å sphere around an atom and calculating the CX value based on the volume of neighbouring atoms according to the formula:

$$CX = \frac{\text{Unoccupied volume of the 6 Å sphere}}{\text{Total volume of atoms that occupy the 6 Å sphere}}.$$

The CX value is calculated for each atom in the surface residues. The CX score for a residue can be calculated using three different methods: First, the CX score can be extracted from the alpha-carbon atom and that score is used for the whole residue (*alphacarbon* method). A second method is to simply average the CX score across all atoms for the residue and use the average score as the CX score for the residue (*averagecx* method). Finally, the CX scores for the atoms in the R -group of the residue can be averaged and that average can be used as the CX score for the residue (*rgroup* method). The CX score ranges from 0 (for residues that are non-protruding) to 15 (for residues that are highly protruding from the protein surface).

Table 1 Classification for each of the 147 protein chains in the dataset. The four letter PDB ids are subscripted by the chain. As can be seen from the table, over half (55.7%) of the RNAs are rRNAs

<i>RNA type</i>	<i>PDBIDs</i>
dsRNA	1DI2 _A , 1UVJ _A , 1YZ9 _A
mRNA	1AV6 _A , 1G2E _A , 1GTF _Q , 1KNZ _A , 1KQ2 _A , 1M8X _A , 1WPU _A , 1WSU _A , 2AIR _A , 2ASB _A
rRNA	1APG _A , 1DFU _P , 1FEU _A , 1FJG _B , 1FJG _C , 1FJG _D , 1FJG _E , 1FJG _G , 1FJG _I , 1FJG _J , 1FJG _K , 1FJG _L , 1FJG _M , 1FJG _N , 1FJG _P , 1FJG _Q , 1FJG _S , 1FJG _T , 1FJG _V , 1G1X _A , 1HRO _W , 1I6U _A , 1JBR _A , 1MZP _A , 1SDS _A , 1T0K _B , 1UN6 _B , 1VQO ₁ , 1VQO ₂ , 1VQO ₃ , 1VQO _A , 1VQO _B , 1VQO _C , 1VQO _D , 1VQO _E , 1VQO _G , 1VQO _H , 1VQO _I , 1VQO _J , 1VQO _K , 1VQO _L , 1VQO _M , 1VQO _N , 1VQO _P , 1VQO _Q , 1VQO _R , 1VQO _S , 1VQO _T , 1VQO _U , 1VQO _V , 1VQO _W , 1VQO _X , 1VQO _Y , 1VQO _Z , 1W2B ₅ , 1Y69 ₈ , 1Y69 _K , 1Y69 _U , 2AVY _F , 2AVY _U , 2AW4 ₀ , 2AW4 ₁ , 2AW4 ₂ , 2AW4 ₃ , 2AW4 _D , 2AW4 _E , 2AW4 _G , 2AW4 _H , 2AW4 _J , 2AW4 _L , 2AW4 _N , 2AW4 _P , 2AW4 _Q , 2AW4 _R , 2AW4 _S , 2AW4 _Y , 2AW4 _Z , 2BH2 _A , 2D3O ₁ , 2D3O _S , 1G1X _B , 1G1X _C
siRNA	1RPU _A , 1SI3 _A , 2BGG _A
snRNA	1A9N _A , 1EC6 _A , 1LNG _A , 1M8V _A , 1OOA _A
SRP RNA	1E8O _A , 1HQ1 _A
tRNA	1ASY _A , 1B23 _P , 1C0A _A , 1EIY _B , 1F7U _A , 1FFY _A , 1H3E _A , 1H4S _A , 1J1U _A , 1J2B _A , 1K8W _A , 1N78 _A , 1Q2S _A , 1QF6 _A , 1QTQ _A , 1R3E _A , 1SER _A , 1TFW _A , 1U0B _B , 1VFG _A , 1WZ2 _A , 2BTE _A , 2CT8 _A , 2FMT _A
Viral RNA	1A34 _A , 1DDL _A , 1H2C _A , 1LAJ _A , 1N35 _A , 1NB7 _A , 1PGL ₂ , 1RMV _A , 1WNE _A , 2AZ0 _A , 2BU1 _A
Other	1B2M _A , 1JID _A , 1M5O _C , 1YVP _A , 1ZH5 _A , 2A8V _A , 2BX2 _L

2.6 Interface propensity calculations

Consider a residue-based property (such as residue roughness) with k discrete values: (v_1, v_2, \dots, v_k) . Each surface residue is assigned to one of k disjoint subsets S_1, S_2, \dots, S_k based on the value of the residue property. Let I_i and N_i respectively be the *fractions* of interface residues and non-interface residues in the set S_i . Let I and N respectively denote the *fractions* of interface and non-interface residues in the entire dataset (over all of the score ranges). The log-propensity for the interface can then be expressed by:

$$IP_i = \log_2(\text{Propensity}_i) = \log_2 \left(\frac{\frac{I_i}{I}}{\frac{I_i + N_i}{I + N}} \right).$$

The interface propensity IP_i of the property at value v_i is a measure of the preference for the value (or a range of values) v_i among the interface residues (relative to the entire

set of surface residues). $IP_i > 0$ denotes that the specified property value (or a range of values) v_i tends to be more preferred among the interface residues relative to the surface residues. Similarly, $IP_i < 0$ denotes that the specified property value v_i tends to be less preferred among the interface residues relative to the entire set of surface residues.

2.7 Predicting protein-RNA interfaces using machine learning approaches

The problem of identifying protein-RNA interface residues in a protein sequence can be formulated as a binary classification problem: Given a sequence S of length N , $S = s_1s_2, \dots, s_N$ over the alphabet Σ of amino acids, $s_i \in \Sigma, i = 1, \dots, N$ and $S \in \Sigma^*$, the task is to predict whether or not a residue in the sequence is protein-RNA interface residue.

One particular challenge in training classifiers using standard machine learning algorithms is to capture predictive ‘features’ that result in accurate classification of new examples. Hence, in this study, we explored sequence and structural features as input to the classifiers trained to label each residue in a protein-RNA sequence.

2.7.1 Feature representations

Since many standard machine learning algorithms operate with a fixed number of input features, it is fairly common to use a ‘sliding window’ approach (Dietterich, 2002) to generate a collection of fixed length windows, where each window corresponds to the target amino acid and an equal number of its sequence neighbours on each side, $\mathbf{s}_i \rightarrow (s_{i-k}, s_{i-k+1}, \dots, s_{i-1}, s_i, s_{i+1}, \dots, s_{i+k-1}, s_{i+k})$ (sequence-based features). The classifier is trained to label the target residue with a ‘+’ or a ‘-’ label. The target residue has a ‘+’ label if it is a protein-RNA interface residue, and to a ‘-’ label otherwise. In addition to sequence-based features, we encoded each residue s_i in a sequence using the structural features, i.e., $\mathbf{s}_i \rightarrow (cxv_{i-k}, cxv_{i-k+1}, \dots, cxv_{i-1}, cxv_i, cxv_{i+1}, \dots, cxv_{i+k-1}, cxv_{i+k})$ where each $cxv_j, j = i - k, \dots, i + k$, represents *alphacarbon*, *averagecx*, *rgroup*, or *roughness* value respectively corresponding to the residue s_j in the sequence window.

2.7.2 Naïve Bayes and Gaussian Naïve Bayes classifiers

The NB classifier (Mitchell, 1997) is a generative model, in which the probabilities $p(\mathbf{s}_i | y)$ and $p(y)$ of the sequence window \mathbf{s}_i and the class label y (e.g., ‘+’ interface and ‘-’ non-interface residue labels) are estimated from the training data using maximum likelihood estimates. Typically the window \mathbf{s}_i is high-dimensional, represented as a tuple of nominal attribute values (amino acid residues), $\mathbf{s}_i = (s_{i-k}, s_{i-k+1}, \dots, s_{i-1}, s_i, s_{i+1}, \dots, s_{i+k-1}, s_{i+k})$, making it impossible to estimate $p(\mathbf{s}_i | y)$ for large values of the window length. However, the NB classifier makes the assumption that the attribute values are conditionally independent given the class. Therefore, training the NB classifier reduces to estimating probabilities $p(s_j | y)$, $j = i - k, \dots, i + k$, of the amino acids in the sequence window, and $p(y)$, of the class labels, from the training data. During classification, Bayes Rule is applied to compute $p(y | \mathbf{s}_{\text{test}})$ and the class label with the highest posterior probability is assigned to the new sequence window \mathbf{s}_{test} .

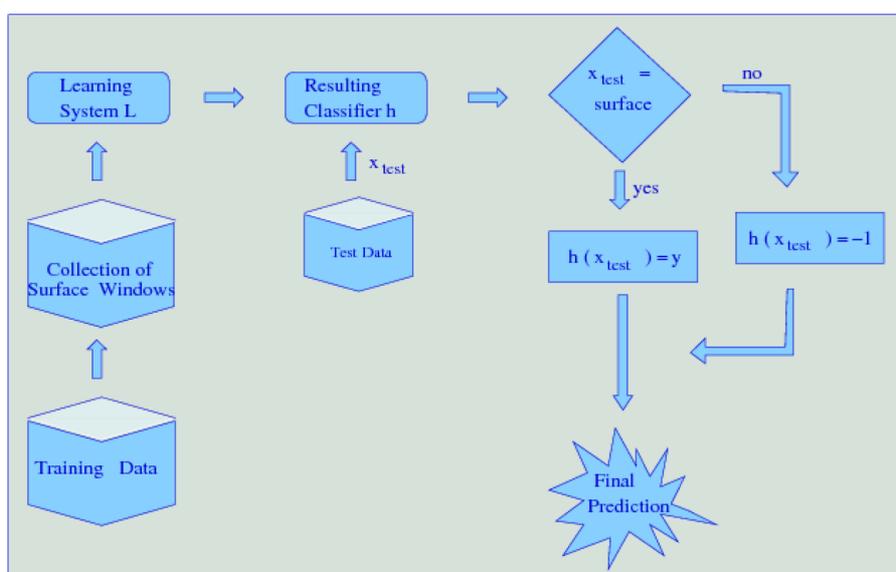
The *Gaussian Naïve Bayes* (GNB) classifier (Mitchell et al., 2004) is similar to the NB classifier, except that the attribute values are numerical, $\mathbf{s}_i = (cxv_{i-k}, \dots, cxv_{i-1}, cxv_i, cxv_{i+1}, \dots, cxv_{i+k})$. The estimated probability $p(cxv_j | y), j = i - k, \dots, i + k$ fits a univariate Gaussian distribution, using maximum likelihood estimates of the mean and variance obtained from the training data, while the probability $p(y)$ fits a Bernoulli distribution.

The NB classifier trained using sequence-based features is the same as the classifier used for predicting protein-RNA interfaces by RNABindR (Terribilini et al., 2007).

2.7.3 *Struct-NB classifier*

To take advantage of the structural information, we developed a two-stage classifier, called *Struct-NB*: in the first stage, the windows 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 or structural 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 (see Figure 1). One important advantage of this model is that it allows the use of existing machine learning methods. In this study, we used NB classifiers.

Figure 1 *Struct-NB*: a two-stage classifier that integrates domain knowledge (i.e., structural information) to improve classification performance. In the first stage, the windows corresponding to each residue are split according to their surface/non-surface label. In the second stage, if the current residue is a surface residue, the classifier returns a probability that this residue is an interface residue; otherwise, the classifier assigns this residue to be a non-interface residue (see online version for colours)



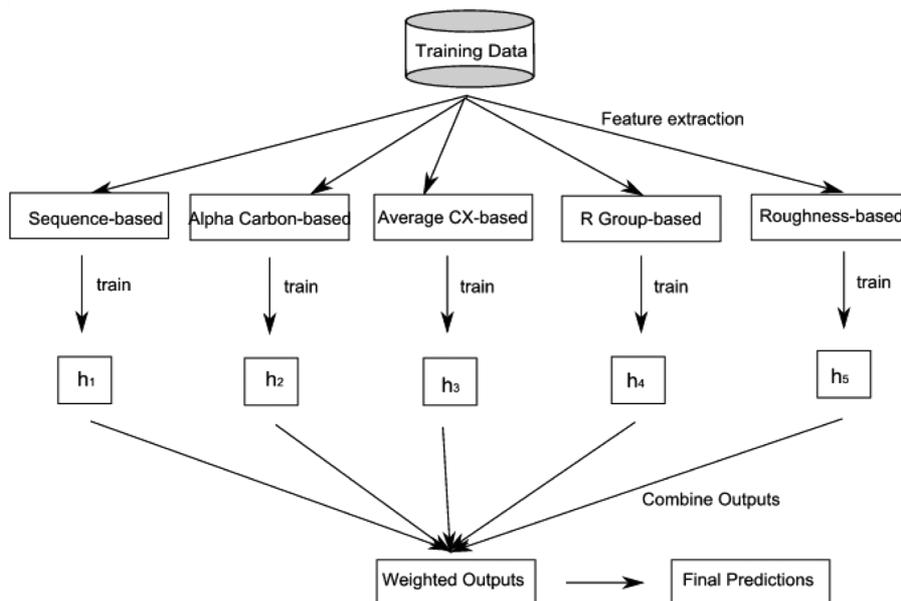
2.7.4 Ensemble of Naïve Bayes classifiers

An *ensemble of* NB classifiers is a collection of NB classifiers, each trained on a different feature representation of the training data (see Figure 2). The prediction of the *ensemble of* NB classifiers is computed from the predictions of the individual NB classifiers. That is, during classification, for a new instance \mathbf{x}_{test} , each individual NB classifier in the collection returns a probability $P_j(y_i | \mathbf{x}_{\text{test}})$, that \mathbf{x}_{test} belongs to a particular class y_i , where $j = 1, \dots, n$, and n is the number NB classifiers in the collection. The ensemble estimated probability, $P_{Ens}(y_i | \mathbf{x}_{\text{test}})$ is obtained by:

$$P_{Ens}(y_i | \mathbf{x}_{\text{test}}) = \frac{1}{n} \sum_j^n P_j(y_i | \mathbf{x}_{\text{test}}).$$

In our experiments, we trained 5 NB classifiers ($n = 5$), one for each feature representation: *sequence*-based, *alphacarbon*-based, *averagecx*-based, *rgroup*-based, and *roughness*-based feature representations, respectively.

Figure 2 A schematic of the Ensemble classifier used for combining information from the multiple feature representations. As can be seen from the figure, a classifier is trained using sequence and structural features. The prediction of the ensemble is then obtained from the predictions of individual classifiers



2.8 Performance evaluation of classifiers

Standard approaches to assessing the performance of classifiers rely on k -fold cross-validation wherein a dataset is partitioned into k disjoint subsets (folds). The performance measure of interest is estimated by averaging the measured performance of the classifier on k runs of a cross-validation experiment, each using a different choice of the $k - 1$ subsets for training and the remaining subset for

testing the classifier. In our experiments, we used sequence-based (as opposed to window-based) cross-validation, a procedure that guarantees that training and test sets are disjoint at the sequence level (Caragea et al., 2007).

To assess the performance of our classifiers we reported the following measures described in Baldi et al. (2000): Sensitivity +, and Specificity +, Receiver Operating Characteristic (ROC) curve, and Area Under the ROC Curve (AUC). If we denote true positives, false negatives, false positives, and true negatives by TP , FN , FP , and TN respectively, then the Sensitivity + and the Specificity + can be defined as follows:

$$\text{Sens+} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Spec+} = \frac{TN}{TN + FP}. \quad (2)$$

The ROC curve plots the proportion of correctly classified positive examples, True Positive Rate (TPR) as a function of the proportion of incorrectly classified negative examples, False Positive Rate (FPR) for different classification thresholds. In comparing two different classifiers using ROC curves, for the same FPR, the classifier with higher TPR gives better performance measures. Each point on the ROC curve represents a classification threshold θ and corresponds to particular values of TPR and FPR.

To evaluate how good a classifier is at discriminating between the positive and negative examples, we also report the AUC on the test set, which represents the probability of correct classification (Baldi et al., 2000). That is, an AUC of 0.5 indicates a random discrimination between positives and negatives (random classifier), while an AUC of 1 indicates a perfect discrimination (very good classifier).

3 Results

Now we proceed to explore the questions: *Do protein-RNA interfaces differ from non-interfaces in terms of their structural features? Do the structural features vary with the type of bound RNA? How useful can structural features be to reliably predicting protein-RNA interface residues?*

3.1 CX protrusion index

Figure 3 shows the relative CX score propensities of the interface residues, based on three different calculations of CX score for each residue: the average of the CX scores for all atoms (*averagecx*); use the average of the CX scores atoms in the R-group only (*rgroup*); or the CX score for the alpha carbon atom as the score for the corresponding residue (*alphacarbon*). The figure shows that averaging the CX score across all atoms in the residue produces similar results to averaging across the R-group atoms alone.

Figures 4–6 (respectively) show the residue propensities based on the types of the bound RNA using the *averagecx*, *alphacarbon* and *rgroup* methods respectively. The observed CX value residue propensities of interfaces involving different types of RNA appear to be sensitive to the method used to calculate the CX values.

Figure 3 Comparison of various methods to obtain CX values for residues on the protein surface. The figure shows that averaging the CX score across all atoms in the residue produces similar results to averaging across the *R*-group atoms alone (see online version for colours)

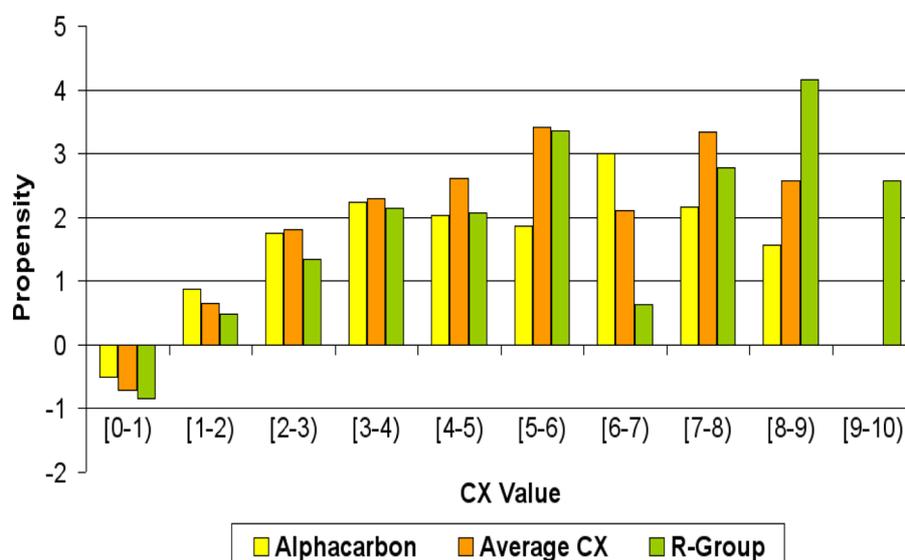
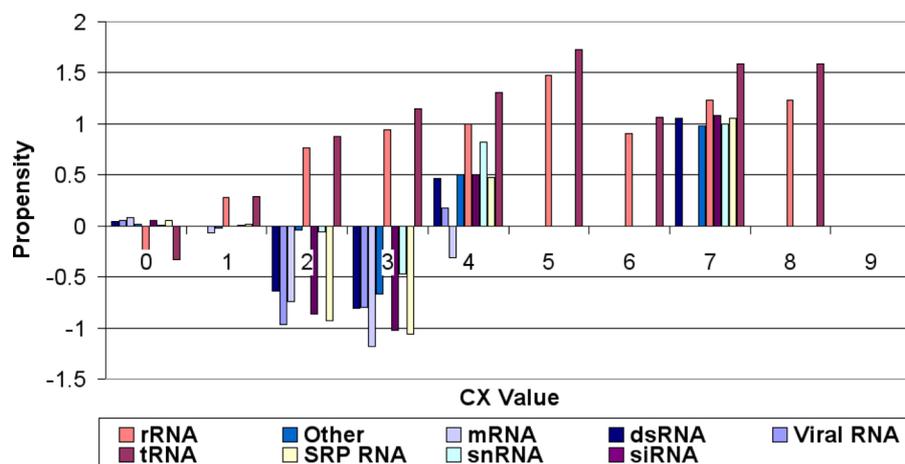


Figure 4 Propensity scores for CX values (*Y*-axis) calculated using the *averagecx* for different ranges of CX values (*X*-axis). Propensities for CX values 0–4 tend to vary across different RNA types as compared to propensities for higher CX scores. Different colours correspond to different RNA types (see online version for colours)



ANOVA analysis for the *alphacarbon* method (ANOVA p -value = 0.056, cutoff = 0.05) shows that tRNA and mRNA cluster together with variances around 0.2. SnRNA, rRNA, dsRNA and 'other' cluster together with variances around 0.3. Finally, siRNA, SRP RNA and Viral RNA cluster together with

variances around 0.65. ANOVA analysis for the *averagecx* method (ANOVA p -value = 0.00001, cutoff = 0.05) shows that Viral RNA, mRNA, ‘other’, and snRNA cluster together with variances around 0.2. DsRNA, rRNA siRNA and SRP RNA cluster together with variance around 0.35 and tRNA is the only RNA type with variance around 0.5. The results of ANOVA analysis for the *rgroup* method (ANOVA p -value = 0.0002, cutoff = 0.05) are similar to those of *averagecx* with mRNA, snRNA and ‘other’ clustering together with variances around 0.13, whereas dsRNA, SRP RNA, siRNA, rRNA and Viral RNA cluster together with variances around 0.37. Finally, the tRNA group is isolated with a variance of 0.5.

Figure 5 Propensity scores for CX values (Y-axis) calculated using the *alphacarbon* for different ranges of CX values (X-axis). Different colours correspond to different RNA types (see online version for colours)

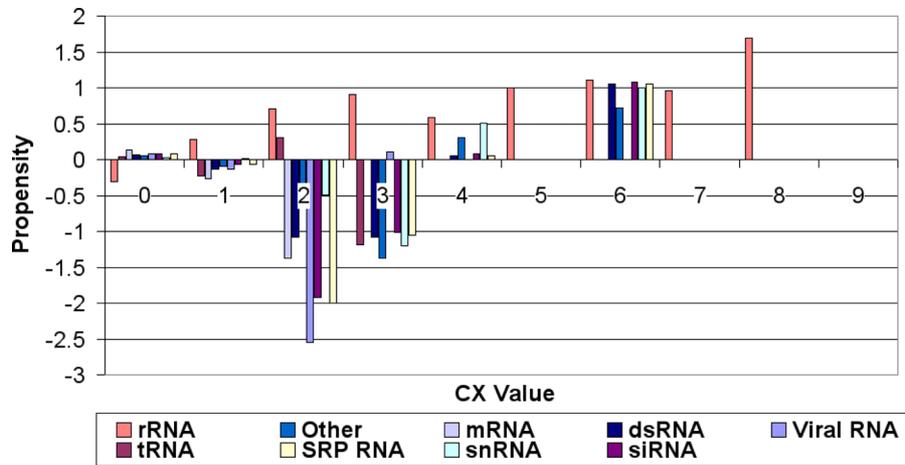


Figure 6 Propensity scores for CX values (Y-axis) calculated using the *rgroup* for different ranges of CX values (X-axis). Different colours correspond to different RNA types (see online version for colours)

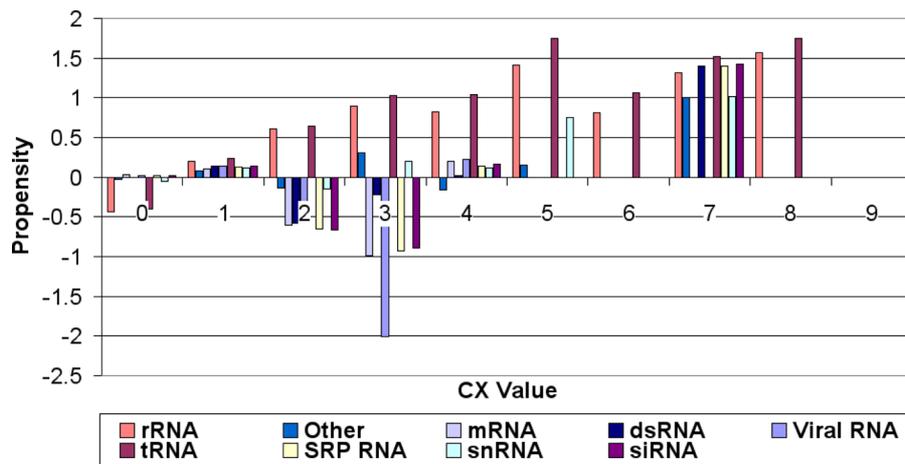


Figure 7 The propensity values for various roughness scores on the surface residues (see online version for colours)

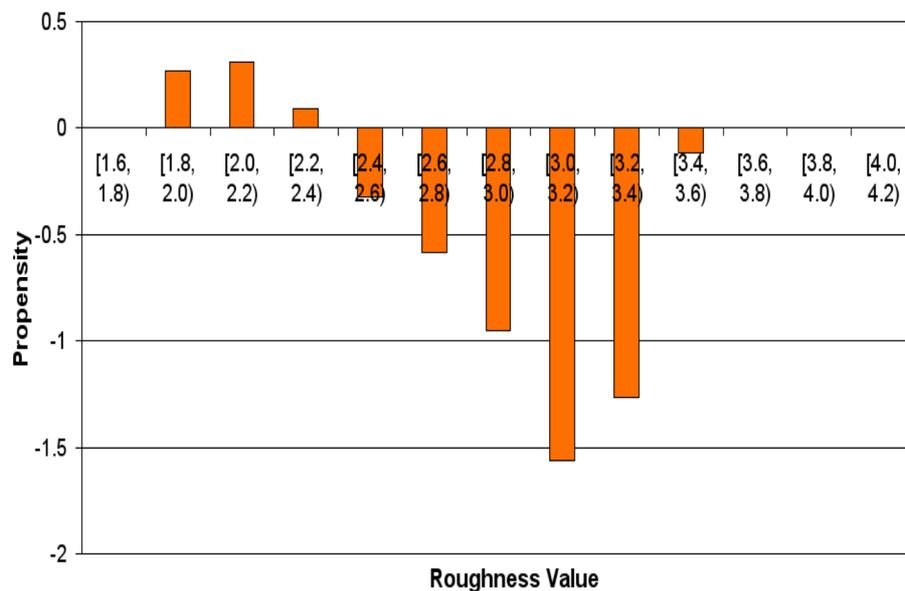
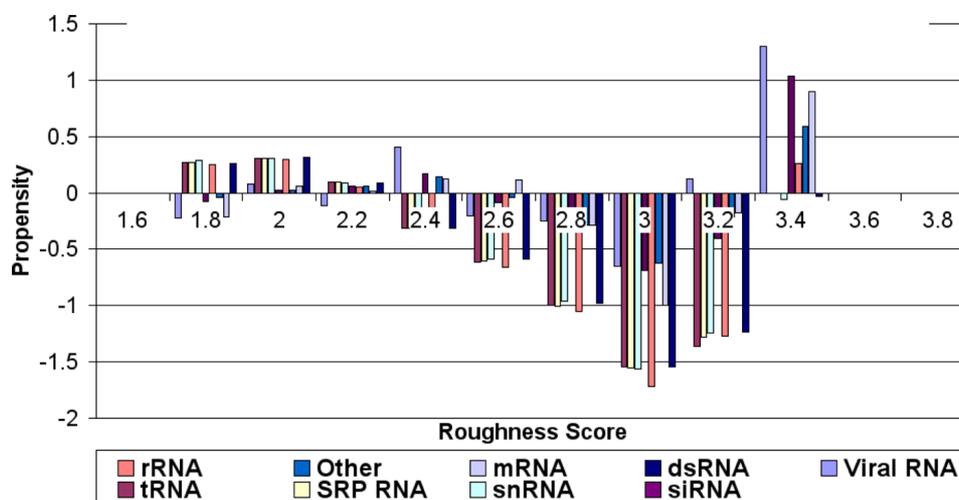


Figure 8 Propensity scores for roughness values classified by RNA type. Roughness propensity is similar across different types of RNA interfaces (see online version for colours)



Regardless of the method used to calculate the CX scores, we can observe some general trends: The rRNA and tRNA propensities are always negative at CX score range $[0, 1)$ and have an increasing propensity as the CX scores increase. However, all other types of RNA (mRNA, snRNA, ..., etc.) tend to have low (negative) propensity values from $[0, 4)$, and the propensities for all types then tend to rise after CX range $[4, 5)$. One interesting exception is the score range $[0, 1)$, which tends to

have slightly positive (albeit very small) interface propensities in the case of RNAs other than tRNA and rRNA, suggesting that in protein-RNA interfaces containing such RNAs non-protruding residues can be interface residues. The CX value residue propensities of interface residues bound to Viral RNAs and rRNAs appear to differ significantly from that of residues that bind to other types of RNAs Figures 4–6.

3.2 Roughness value

Figure 7 shows the propensities for the roughness score. Residues with rough surfaces are preferred in protein-RNA interfaces. Figure 8 shows the distribution of the propensities for each roughness score range classified by the type of RNA. The figure indicates that, unlike the CX scores, the roughness scores do not vary significantly among the different types of RNA at each score range as all RNA types behave almost identically for each roughness score range (ANOVA p -value = 0.32 with cutoff 0.05). Using the variances calculated by ANOVA, tRNA, SRP RNA, snRNA, rRNA and dsRNA seem to cluster well with each other (all have variances close to 0.2). The remaining RNA types (Viral RNA, siRNA, mRNA and ‘other’) have variances around 0.1.

3.3 Protein-RNA interface prediction using *struct-NB*

The main result of our study using machine learning approaches is that the *Struct-NB* classifiers described here outperform *sequence-based* NB and *structural-based* GNB classifiers on the problem of predicting RNA-binding sites in protein sequences.

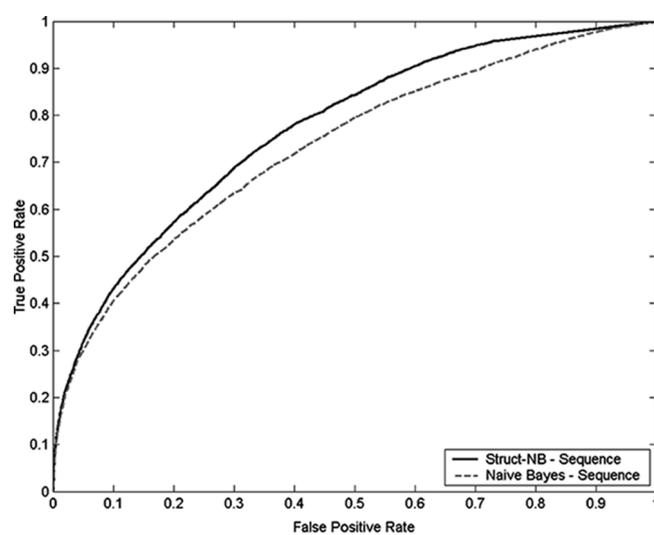
We trained *Struct-NB* classifiers, that take into account structural information, using sequence and structural features, NB using sequence features, and GNB using structural features. We compared the ROC curves of *sequence-based* NB and *structural-based* GNB with those of *Struct-NB* classifiers on the 147 protein-RNA dataset. The ROC curves show the tradeoff between TPR and FPR over their entire range of possible values. Hence, it is more informative to compare the ROC curves than to compare the performance of the classifiers for a particular choice of the tradeoff (which corresponds to a specific point θ on the ROC curve). We found that the ROC curve for *Struct-NB* dominates the ROC curve for *sequence-based* NB and *structural-based* GNB. That is, for any choice of FPR, *Struct-NB* offers a higher TPR than both *sequence-based* NB and *structural-based* GNB (Figures 9(a), (b) and 10(a)–(c)). In Figure 10(d) we compare the performance of the ensemble of *Struct-NB* classifiers, that combines the predictions of individual *Struct-NB* classifiers trained using sequence and structural features, with that of the NB using sequence features. As can be seen, for values of the false positive rate higher than 0.2, the ensemble of *Struct-NB* has a higher Sensitivity than that of the *sequence-based* NB.

In addition, we compared the ROC curves of NB using sequence features with that of GNB using structural features. We found that while the ROC curves for *sequence-based* NB dominates the ROC curves for *structural-based* GNB (see Figure 11(a)–(d)), the *structural-based* GNB has a higher sensitivity than *sequence-based* NB and *sequence-based* NB has a higher specificity than *structural-based* GNB.

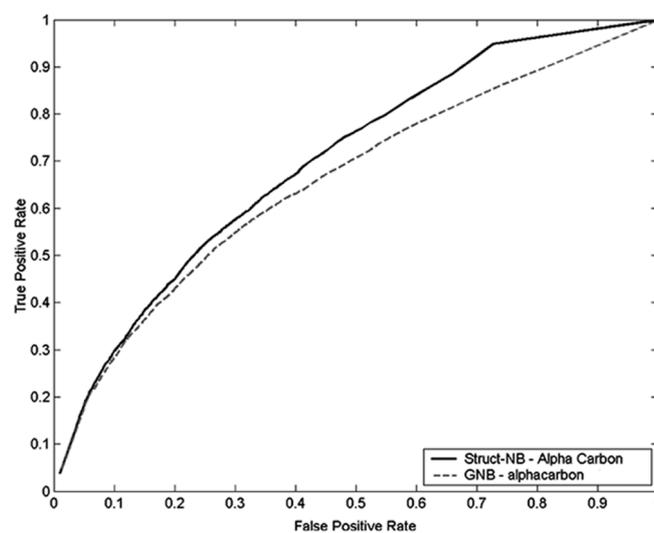
In Table 2, we show a comparison of the performance statistics for the GNB using alphacarbon, averagecx, rgroup, roughness features respectively, the NB using sequence features, and the ensemble of *Struct-NB*. The statistics

are reported for the same threshold $\theta = 0.5$ on the ROC curves. As can be seen, *sequence-based* NB achieves Sensitivity+ = 0.246, Specificity+ = 0.65, and AUC = 0.736, *structural-based GNB* using *roughness* features achieves Sensitivity+ = 0.349, Specificity+ = 0.469, and AUC = 0.696, and the ensemble of *Struct-NB* achieves Sensitivity+ = 0.359, Specificity+ = 0.443, and AUC = 0.752. Therefore, in applications that require a higher sensitivity, a better choice is to use structural feature as input to the *NB* classifiers. This is visually shown in Table 3.

Figure 9 Comparison of Struct-NB and Gaussian Naïve Bayes classifiers utilising sequence (a) and alphacarbon (b) as input to the classifiers: (a) comparison of the ROC curves for Struct-NB and Naïve Bayes classifiers using sequence features as input to the classifiers and (b) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using alphacarbon features as input to the classifiers



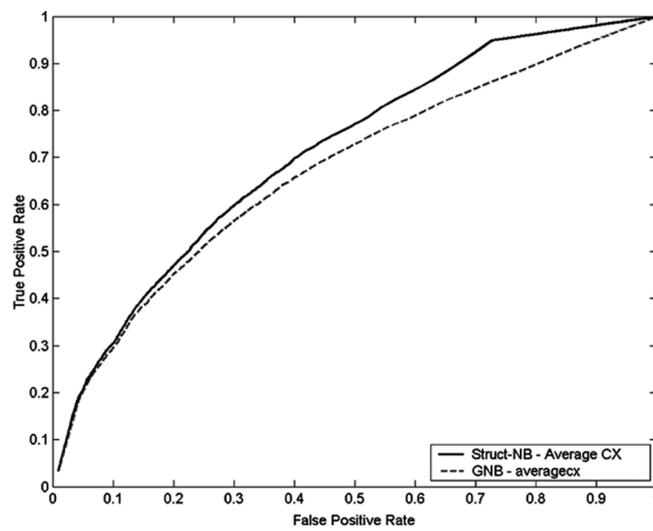
(a)



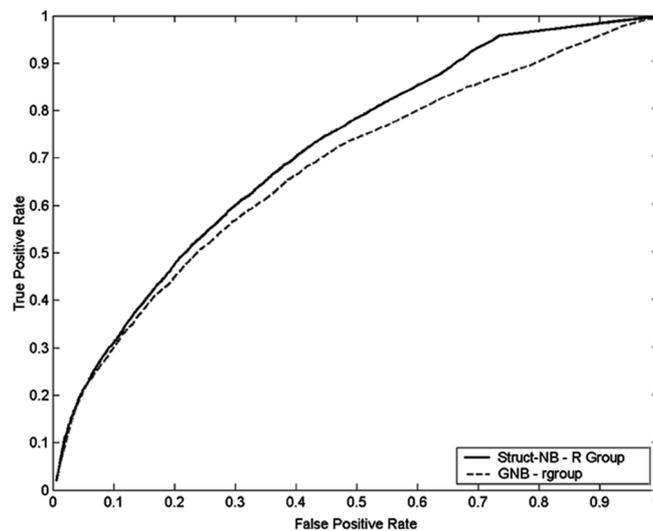
(b)

Figure 10 Comparison of Struct-NB and Gaussian Naïve Bayes classifiers utilising (a) averagecx, (b) rgroup and (c) roughness as input to the classifiers. Figure 10(d) shows a comparison of the ensemble Struct-NB classifier trained using sequence and structural features with a Naïve Bayes classifier using sequence features:

(a) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using averagecx features as input to the classifiers; (b) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using rgroup features as input to the classifiers; (c) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using roughness features as input to the classifiers and (d) comparison of the ROC curves for the ensemble of Struct-NB trained using sequence and structural features and Naïve Bayes classifiers using sequence features

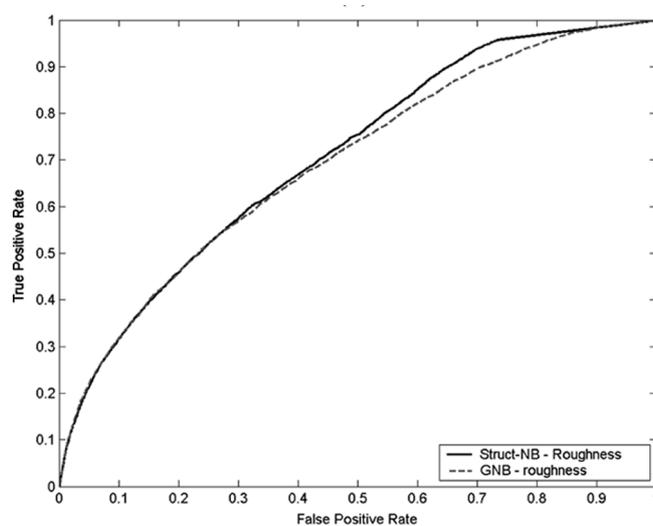


(a)

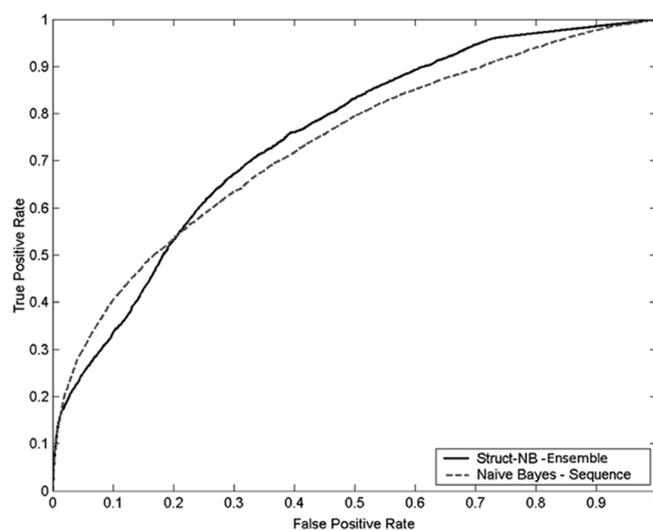


(b)

Figure 10 Comparison of Struct-NB and Gaussian Naïve Bayes classifiers utilising (a) averagecx, (b) rgroup and (c) roughness as input to the classifiers. Figure 10(d) shows a comparison of the ensemble Struct-NB classifier trained using sequence and structural features with a Naïve Bayes classifier using sequence features: (a) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using averagecx features as input to the classifiers; (b) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using rgroup features as input to the classifiers; (c) comparison of the ROC curves for Struct-NB and Gaussian Naïve Bayes classifiers using roughness features as input to the classifiers and (d) comparison of the ROC curves for the ensemble of Struct-NB trained using sequence and structural features and Naïve Bayes classifiers using sequence features (continued)

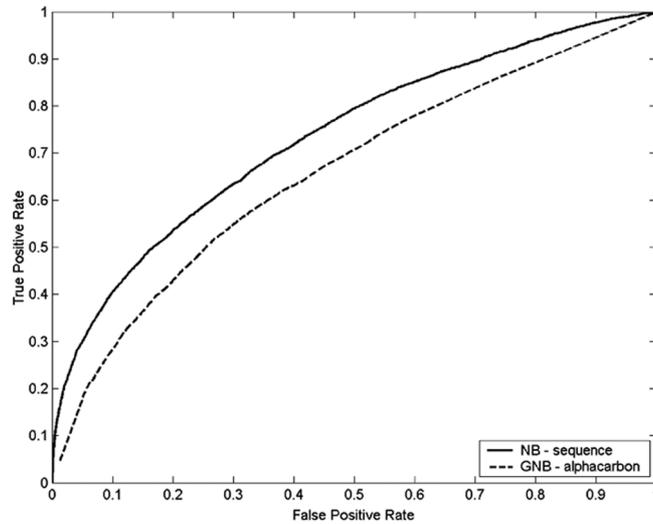


(c)

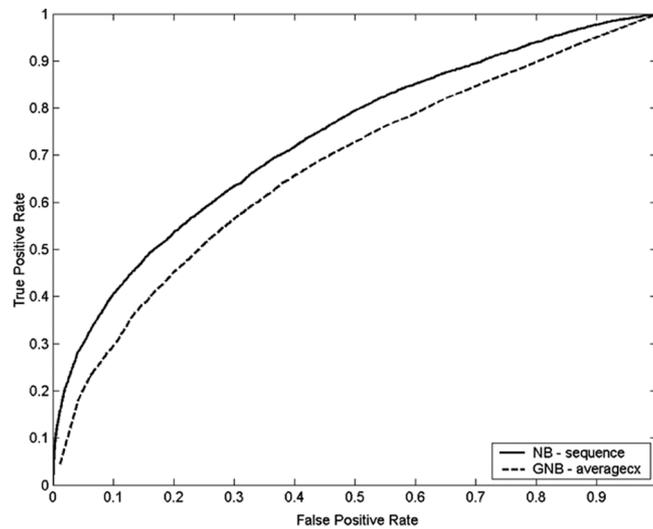


(d)

Figure 11 Comparison of Gaussian Naïve Bayes classifiers utilising (a) alphacarbon, (b) averagecx, (c) rgroup and (d) roughness as input to the classifiers with a Naïve Bayes classifier trained using sequence features: (a) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using alphacarbon features as input; (b) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using averagecx features as input; (c) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using rgroup features as input and (d) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using roughness features as input

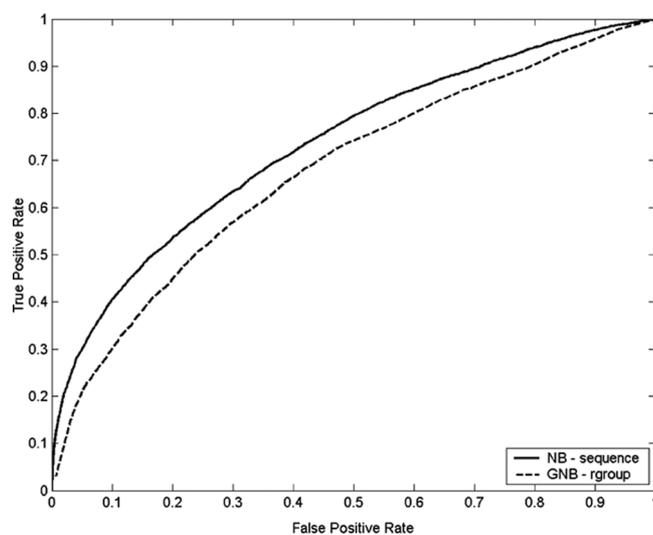


(a)

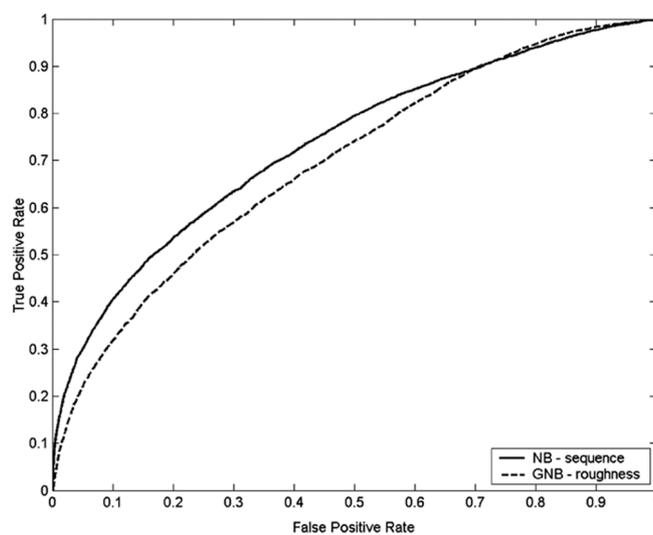


(b)

Figure 11 Comparison of Gaussian Naïve Bayes classifiers utilising (a) alphacarbon, (b) averagecx, (c) rgroup and (d) roughness as input to the classifiers with a Naïve Bayes classifier trained using sequence features: (a) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using alphacarbon features as input; (b) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using averagecx features as input; (c) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using rgroup features as input and (d) comparison of the ROC curves for Naïve Bayes using sequence features as input and Gaussian Naïve Bayes using roughness features as input (continued)



(c)



(d)

It is possible that the general trends observed across all RNA types is biased by rRNA-binding proteins, which make up over half of the protein-RNA complexes in PDB. One way to determine whether this is indeed the case is to repeat the protein-RNA interface analysis separately for each RNA-type. Also of concern is the relatively small size of the RB147 dataset protein-RNA dataset. Terribilini et al. (2007) have noted that PDB included only 198 protein-nucleic acid complexes in 1996, but by April 2007, this number had grown to 1734, of which 529 were protein-RNA complexes. The resulting availability of larger and more diverse datasets can be expected to significantly improve the quality and quantity of data available for performing the analysis of the sort reported here. Work in progress is aimed at:

- Assembling a comprehensive of Protein-RNA Interface Database (PRIDB), and associated services for querying, analysis, and visualisation of protein-RNA interfaces.
- Developing machine learning approaches for reliable identification of putative RNA-binding residues in proteins that improve upon the state-of-the-art sequence-based methods (Terribilini et al., 2006b) by taking advantage of structural and molecular dynamics simulations.
- Analysis of sequence and structural properties of protein-binding residues in RNA and the development of machine learning approaches for reliable identification of putative protein-binding RNA residues.
- Characterisation of the sequence and structural correlates of protein-RNA interfaces and the similarities and differences among different types of protein-RNA interfaces, and between protein-protein, protein-DNA, and protein-RNA interfaces.

Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (GM066387) to Vasant Honavar and Drena Dobbs, an Integrative Graduate Education and Research Training (IGERT) fellowship to Fadi Towfic, funded by the National Science Foundation grant (DGE 0504304) to Iowa State University, and a Bioengineering and Bioinformatics Summer Institute (BBSI) fellowship to David Gemperline, funded by a National Science Foundation award (EEC 0608769) to Iowa State University. This work has benefited from discussions with Dr. Robert Jernigan, Feihong Wu, Michael Terribilini and Yasser El-Manzalawy of Iowa State University.

References

- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. and Nielsen, H. (2000) 'Assessing the accuracy of prediction algorithms for classification: an overview', *Bioinformatics*, Vol. 16, No. 5, pp.412-424.
- Bechara, E., Davidovic, L., Melko, M., Bensaid, M., Tremblay, S., Grosgeorge, J., Khandjian, E.W., Lalli, E. and Bardoni, B. (2007) 'Fragile *x* related protein 1 isoforms differentially modulate the affinity of fragile *x* mental retardation protein for *g*-quartet RNA structure', *Nucleic Acids Research*, Vol. 35, pp.299-306.

- Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000) 'The protein data bank', *Nucleic Acids Res.*, Vol. 28, pp.235–242.
- Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D. and Honavar, V. (2007) 'Glycosylation site prediction using ensembles of support vector machine classifiers', *BMC Bioinformatics*, Vol. 8, No. 438, pp.438–450.
- Chen, Y. and Varani, G. (2005) 'Protein families and RNA recognition', *FEBS Journal*, Vol. 272, No. 9, pp.2088–2097.
- Connolly, M.L. (1986) 'Measurement of protein surface shape by solid angles', *Journal of Molecular Graphics*, Vol. 4, No. 1, pp.3–6.
- Connolly, M.L. (1993) 'The molecular surface package', *J. Mol. Graph.*, Vol. 11, No. 2, pp.139–141.
- Dietterich, T.G. (2002) *Machine learning for sequential data: A review. Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science, Vol. 2396, pp.15–30.
- Ellis, J., Broom, M. and Jones, S. (2007) 'Protein-RNA interactions: structural analysis and functional classes', *Proteins*, Vol. 66, No. 4, pp.903–911, <http://www.citeulike.org/user/jarasch/article/1205659>
- Freed, E.O. and Mouland, A.J. (2006) 'The cell biology of hiv-1 and other retroviruses', *Retrovirology*, Vol. 3, No. 77, pp.77–86, <http://www.retrovirology.com/content/3/1/77>
- Jeong, E., Chung, I-F. and Miyano, S. (2004) 'A neural network method for identification of RNA-interacting residues in protein', *Genome Informatics*, Vol. 15, No. 1, pp.105–116.
- Jeong, E. and Miyano, S. (2006) 'A weighted profile method for protein-RNA interaction prediction', *Trans. Comput. Syst. Biol.*, Vol. IV, pp.123–139.
- Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M. and Thornton, J.M. (2001) 'Protein-RNA interactions: a structural analysis', *Nucleic Acids Research*, Vol. 29, No. 4, pp.943–954.
- Jurica, M.S. and Moore, M.J. (2003) 'Pre-mRNA splicing: awash in a sea of proteins', *Mol. Cell*, Vol. 12, pp.5–14.
- Lee, B. and Richards, F.M. (1971) 'The interpretation of protein structures: estimation of static accessibility', *J. Mol. Biol.*, Vol. 55, pp.379–400.
- Lejeune, D., Delsaux, N., Charlotiaux, B., Thomas, A. and Brasseur, R. (2005) 'Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure', *Proteins*, Vol. 61, No. 2, pp.258–271, <http://www3.interscience.wiley.com/journal/111083051/abstract>
- Lewis, M. and Rees, D. (1985) 'Fractal surfaces of proteins', *Science*, Vol. 230, No. 4730, pp.1163–1165.
- Mitchell, T. (1997) *Machine Learning*, McGraw-Hill, Boston, MA.
- Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M. and Newman, S. (2004) 'Learning to decode cognitive states from brain images', *Machine Learning Journal*, Vol. 57, pp.145–175.
- Moore, M.J. (2005) 'From birth to death: the complex lives of eukaryotic mRNAs', *Science*, Vol. 309, pp.1514–1518.
- Noller, H.F. (2005) 'RNA structure: reading the ribosome', *Science*, Vol. 309, pp.1508–1514.
- Pintar, A., Carugo, O. and Pongor, S. (2002) 'Cx, an algorithm that identifies protruding atoms in proteins', *Bioinformatics*, Vol. 18, No. 7, pp.980–984.
- Terribilini, M., Lee, J-H., Yan, C., Jernigan, R.L., Carpenter, S., Honavar, V. and Dobbs, D. (2006a) 'Identifying interaction sites in 'recalcitrant' proteins: predicted protein and RNA binding sites in rev proteins of hiv-1 and eiav agree with experimental data', *Pacific Symposium on Biocomputing*, Vol. 11, pp.415–426, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2553683>

- Terribilini, M., Lee, J-H., Yan, C., Jernigan, R.L., Honavar, V. and Dobbs, D. (2006b) 'Prediction of RNA binding sites in proteins from amino acid sequence', *Bioinformatics*, Vol. 12, pp.1450–1462.
- Terribilini, M., Sander, J.D., Lee, J-H., Zaback, P., Jernigan, R.L., Honavar, V. and Dobbs, D. (2007) 'Rnabindr: a server for analyzing and predicting RNA-binding sites in proteins', *Nucleic Acids Research*, Vol. 35, No. 2, pp.1–7.
- Towfic, F., Gemperline, D.C., Caragea, C., Wu, F., Dobbs, D. and Caragea, C. (2007) 'Structural characterization of RNA-binding sites of proteins: preliminary results', *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp.60–66, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4425401
- Treger, M., and Westhof, E. (2001) 'Statistical analysis of atomic contacts at RNA-protein interfaces', *J. Mol. Recognit.*, Vol. 14, No. 4, pp.199–214.
- Wu, F., Towfic, F., Dobbs, D. and Honavar, V. (2007) 'Analysis of protein protein dimeric interfaces', *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*, pp.35–41.