Dna binding protein prediction english language essay

Linguistics, English

egin{figure}[! htbp]centeringincludegraphics[width= 0. 75extwidth]{chap3/figures/aim_1_CMYK. eps}caption[What we want to learn from nucleic acid-binding protein prediction]{A) Which of these proteins binds DNA? B) Given that a particular protein binds DNA, which residues are involved in binding? DNA: url{http://www. dk. co. uk/static/clipart/uk/dk/future/image_future007. jpg}.}label{fig: DNA-bp_prediction_goals}end{figure}section{Project Goals}We approach the DNA-binding prediction project in two ways, illustrated in ef{fig: DNA-bp_prediction_goals}. We asked two questions: 1) Given a set of proteins, which of the proteins in this set binds DNA? 2) Given that a particular protein binds DNA, which residues are involved in binding? Though these are two very different questions, the prediction strategy used during the implementation of these ideas is very similar. section{Prediction Strategy}Our prediction strategy starts with either a protein structure or protein sequence. We then gather features for either the entire protein or each residue depending on what type of prediction we are making. Next, we follow one of two paths: 1) calculate structure-based features, which are attributes of 3-dimensional structures acquired using X-ray crystallography or NMR, or 2) calculate sequence-based features. These features are collected and possibly normalized or converted in some way depending on the algorithm. We then train models using this attribute information and subsequently test our models with new data (for which we calculate the same features) in order to predict either binding or non-binding.egin{figure}[! htbp]centeringincludegraphics[width= extwidth]{chap3/figures/prediction_strategy_CMYK. eps}caption[NA-binding prediction strategy]{DNA-binding prediction strategy}label{fig: prediction_strategy}end{figure}section{Protein-level Prediction}subsection{Data Set}Our data set for DNA-binding prediction included two classes of proteins; those that bind DNA (positive class) and those not known to bind DNA (negative class). It consisted of 75 DNA-binding proteins and 214 others not known to bind DNA including membrane-binding proteins, chaperones, and enzymes. This negative set is a subset of one used by Stawiski, Gregoret, and Mandel-Gutfreund cite{Stawiski2003}. These sets were culled using the PISCES web server cite{Wang2003}, and only structures with a sequence identity of $leq$20\\% and a resolution of $leq$3{AA} were included in our experiments. subsection{Calculated Attributes}For protein-level binding prediction, we used a set of 42 features related to protein structure and sequence. The sequence-based features included the amino acid composition (20 features) and the net charge calculated using the CHARMM cite{Brooks1983} force field (1 feature). Structure-based features included surface amino acid composition derived from DSSP cite{Kabsch1983} (20 features) and the size of the largest positively charged patch cite{Bhardwaj2005, Bhardwaj2006} (1 feature). subsection{Results}We performed DNA-binding protein classification using sequence- and structure-based features with several algorithms (ef{tab: DNA-BP_structure_pred}). We varied the size of the data sets via 2- and 5-fold cross validation. These results demonstrate our ability to distinguish between DNA-binding and non-DNA-binding proteins. AdaTree performed the best with 88. 5\\% accuracy, 66. 7\\% sensitivity, 96. 3\\% specificity, and an AUC of 88. 7\\%. These results can be interpreted in the following way. Given a random data set of proteins with the same class distribution, AdaTree should correctly assign $approx$88\\% of these proteins to the correct class. If we know that a protein binds DNA, our classifier will correctly categorize it $approx$66\\% of the time. Similarly, if we have prior knowledge that a protein that does not bind DNA, we can correctly predict this $approx$96\\% of the time. All of these metrics are dependent on class distribution with the exception of the AUC, and because our data set is imbalanced we place more confidence in the AUC. One interesting finding is that the results for SVM (a very slow algorithm to train) and the second fastest (AdaStump) are fairly close. This tells us that we can use a fast tree algorithm such as this and not sacrifice much accuracy. This is useful because the SVM algorithm ran at a rate $approx$25 times slower than AdaStump.egin{table}egin{center}egin{tabular}{| c| c| c| c| c| c| c|}hlinemulticolumn{7}{| c|}{DNA-binding: The Performance of Five Classifiers} \\ hline%multicolumn{7}{c}{} \\ hlineCV&Metric&AdaTree&AdaC4. 5&AdaStump&SVM&C4. 5\\ hlinemultirow{4}{*}{2-fold}&ACC&86. 5&86. 3&83. 6&84. 5&79. 4\\ cline{2-7}&SEN&61. 4&61. 5&60. 5&57. 5&59. 8\\ cline{2-7}&SPE&95. 3&95. 0&91. 7&94. 0&86. 3\\ cline{2-7}&AUC&88. 0&88. 8&81. 6&84. 4&61. 3\\ hlinemultirow{4}{*}{5-fold}&ACC&87. 2&86. 6&84. 1&85. 7&79. 4\\ cline{2-7}&SEN&63. 8&61. 4&61. 2&59. 1&59. 9\\ cline{2-7}&SPE&95. 4&95. 4&92. 2&95. 0&86. 3\\ cline{2-7}&AUC&88. 4&89. 6&82. 7&85. 9&54. 1\\ hlinemultirow{4}{*}{LOO}&ACC&88. 5&86. 5&85. 1&86. 3&80. 0\\ cline{2-7}&SEN&66. 7&61. 3&62. 7&62. 7&65. 3\\ cline{2-7}&SPE&96. 3&95. 3&93. 0&93. 9&85. 0\\ cline{2-7}&AUC&88. 7&89. 8&84. 6&86. 3&74.

0\\ hlineend{tabular}caption[DNA-binding: The Performance of Five Classifiers]{Comparison of four metrics and three different cross validation techniques using five different classifiers for protein-level prediction over the protein-DNA data set.}label{tab: DNA-BP_structure_pred}end{center}end{table}% From Lanlois 2007%The learning algorithms comprise four tree-based algorithms and SVM. Specifically, the C4. 5 decision tree algorithm forms the weak learning algorithm for the AdaBoost procedure; its results demonstrate the effectiveness of boosting. Next, the boosted C4. 5 algorithm (AdaC4. 5) serves as a baseline to compare our custom decision tree implementation, which forms the weak learners in AdaTree and AdaStump. Finally, the odd man out, SVM, provides a connection to our previous study; 6 however, in this study we choose to maximize the accuracy rather than find a more balanced prediction.% From Lanlois 2007%The results in Table 1 demonstrate that our method is effective in discriminating DNA-binding proteins. That is, given a large random set of proteins (with the same distribution as our data set) the best classifier, AdaTree, should correctly assign on average about 88 of 100 proteins to the appropriate category. Likewise, given a protein that binds DNA, this classifier will assign 66 of 100 correctly to that category. Finally, given a protein that does not bind DNA, about 96 of 100 will be correctly assigned to this category. Indeed, this is an unbalanced result originating from both an unbalanced data set and a set of classifiers that minimizethe overall error. In other words, each of these metrics depends on the distribution of the data set. The area under the ROC curve (AUC) furnishes a metric independent of the data set distribution. It also gives some indication of the tradeoff between sensitivity and specificity when varying the threshold. Specifically, the AdaC4. 5 learning algorithm achieves almost a 90\\% AUC; that is, about 90\\% of the predictions are ordered correctly. This ordering is important both for achieving good results on other distributions of the data set and allowing the learning algorithm to produce a meaningful confidence in its prediction. Likewise, there are several more important trends in Table 1. For example, one interesting result stems from the comparison of sensitivities for each classifier. That is, none of the observed sensitivities vary much from the C4. 5 algorithm. In fact, in each superior learning algorithm (to C4. 5), the increase in accuracy corresponds to a proportional increase in specificity. However, the better learning algorithms also have a larger AUC. Indeed, a larger AUC indicates that trading sensitivity for specificity will most likely have less effect on the overall accuracy over a larger range. Another interesting result in Table 1 originates from the relative independence of each classifier over each metric for different sizes of the training set. That is, for the first four algorithms, accuracy and sensitivity show the greatest change with training set size, yet this change is limited to only a few percent. If 2-fold crossvalidation is a pessimistic estimate and leave-one-out an optimistic estimate, then the results of 5-fold crossvalidation can be considered reliable and probably will not change much on a larger data set. Note that only the AUC for the C4. 5 algorithm improves dramatically with the increase in training examples. Finally, it is interesting to note that the results of the slowest algorithm (SVM) and the second fastest (AdaStump) match relatively well, i. e., on this data set the speed ratio between AdaStump and SVM was on average about

1: 25, respectively. section{Residue-level Prediction}subsection{Data Sets}The proteins comprising our data sets were extracted from the PDB database (url{http://www. rcsb. org}) and culled using the PISCES web server cite{Wang2003} with a sequence identity of leq 25\\%. All structures were determined by X-ray diffraction and had a resolution of leq 3. 0{AA}. Our sequence-based DNA-binding residue data set consisted of 54 proteins and 14780 residues, 2083 of which were identified as DNA-binding and 12697 considered non-binding based on distance from the DNA molecule in the bound structure (class ratio of $approx$1/6). The RNA-binding residue data set used for sequence-based prediction contained 84 proteins and 60, 016 residues, 5, 934 classified as RNA-binding and 54082 as non-binding (class ratio of $approx$1/9). subsection{Definition of Binding Residues}Because we have formulated residue prediction in this case as a binary classification problem, each residue in the data set must be defined as DNA-binding or non-DNA-binding. As with previous studies cite{Ahmad2004, Ahmad2005, Kuznetsov2006}, we based this class distinction on a residue's distance from the DNA molecule in the complex. A residue was defined as binding if any heavy atom (carbon, nitrogen, oxygen, or sulfur) belonging to the residue fell within a distance of 4. 5{AA} of any atom in the DNA molecule. In agreement with Kuznetsov, Gou, Li, and Hwang cite{Kuznetsov2006}, we found that this distance provided the best accuracy for predictions. Any residues without atomic coordinates in the PDB file were not included in the data set. subsection{Calculated Attributes}subsubsection{Residue ID}A twenty-dimensional feature vector representing the 20 common amino acids is used to identify each residue, where a single non-zero entry indicates the current residue. subsubsection{Residue Charge}Since DNA molecules are negatively charged, positively charged, basic amino acid residues can play an important role in nucleic acid binding. Accordingly, we include a charge attribute for each residue. Arginine and lysine residues are assigned a charge of +1, histidines +0. 5, and all others 0. subsubsection{Measures of Evolutionary Conservation}In order to consider the level of evolutionary conservation of each residue and its sequence neighbors, we create a position-specific scoring matrix (PSSM) for each residue in the test protein. Along with the NCBI-NR90 database cite{Ahmad2005}, which contains leq 90\\% sequence identity between any two proteins, PSI-BLAST cite{Altschul1997} is used to create a matrix representing the distribution of all 20 amino acids at each position in the protein sequence. A 7-residue sliding window, which represents the distribution of amino acid residues at the positions occupied by three sequence neighbors on either side of the central residue, is subsequently created. This results in a 140-element feature vector for each residue. A similar 7-residue window is created using the BLOSUM62 matrix cite{Henikoff1992} in order to capture non-position-specific evolutionary conservation information for the sequence neighborhood of each residue, resulting in another 140-element feature vector. subsection{Comparisons with Other Algorithms}We evaluated the performance of our models against five other classification algorithms (SVM, Alternating Decision Tree, WillowBoost, C4. 5 with Adaptive Boosting, and C4. 5 with bootstrap aggregation). We built two models for each using sequence-based features: one for DNA-binding proteins and one for RNA-binding proteins. ef{fig:

ROC_DNA_RNA} describes the results for this comparison and shows the performance of each algorithm in terms of accuracy, sensitivity, specificity, precision, Matthews correlation coefficient (MCC), and the area under the Receiver Operating Characteristic curve (AUC). The AUC provides a measure of a model's ability to separate positive and negative examples and is generated from a plot of the true positive rate versus the false positive rate for each example in the data set ef{fig: ROC_DNA_RNA}. A perfect model would have an AUC of 1, while a random model would have an AUC of 0. 5.egin{figure}[! htbp]centeringincludegraphics[width= extwidth]{chap3/figures/ROC_DNA_RNA_larger_table. eps}caption[ROC curves and metrics for DNA- and RNA-binding residue prediction]{Receiver Operating Characteristic (ROC) curves comparing six classifiers for A) DNA-binding residue prediction models and B) RNA-binding residue prediction models. C) Results of a 10-fold cross validation over the binding residue data sets. Six metrics describe the performance of each of the 12 classifiers: ACC = accuracy, SEN = sensitivity, SPE = specificity, PRE = precision, AUC = area under the ROC curve, MCC = Matthews Correlation Coefficient. Algorithms: SVM (Support Vector Machines), ADTree (alternating decision tree), WillowBoost (in-house-developed tree algorithm w/ boosting), C4. 5ADA (C4. 5 decision tree w/ AdaBoost), C4. 5BAG (C4. 5 decision tree w/ bootstrap aggregation), C4. 5BAGCST (C4. 5 decision tree w/ bootstrap aggregation and costing). The highlighted classifiers (C4. 5BAGCST) are those used for the NAPS web server.}label{fig: ROC_DNA_RNA}end{figure}egin{figure}[! htbp]centeringincludegraphics[width= extwidth]{chap3/figures/DNA-bp_previous_data_sets_rotated_CMYK. eps}caption[Classifiers built from previous NA-binding data sets]{The results shown are from 10-fold CV using a C4. 5BAGCST classifier (unless otherwise indicated) built using each of the previous data sets for DNA- and RNA-binding residue prediction. White rows indicate the performance of our classifier over a particular data set, gray rows indicate the performance of the classifiers reported in the original publication. Gaps identify metrics which were not reported. str: sequence- + structure-based predictor, seq: sequence-based predictor, b: balanced data set, ub: unbalanced data set.}label{fig: DNA-bp_previous_data_sets}end{figure}subsection{Evaluation Using Previous Data Sets}In order to demonstrate the stability of our classifiers, we built models using previously compiled data sets for both DNA- and RNA-binding residue predictions. ef{fig: DNA-bp_previous_data_sets} shows the comparisons between the original classifier and ours using two previously compiled DNA-binding protein data sets and one RNA-binding protein data set used in seven publications cite{Ahmad2005, Kumar2008, Kuznetsov2006, Ofran2007, Terribilini2006, Wang2006, Wang2008}. The classifiers were created using 10-fold cross-validation for both selection and validation. For the costing algorithm, the weight assigned to each class was equal to the class distribution and 200 costing iterations were run. Net accuracy was used to find the best model. The prediction metrics from previous works shown are either those reported as the best results from the publications, or if the author??? s intended best result is unclear, the results with the best accuracy or MCC. Overall we found that, based on the metrics reported in these previous publications, we were able to improve on those results over each of three previously compiled data sets. First, we built our own classifier on the PDNA-62 data set, which was originally compiled by Selvaraj, Kono, and Sarai cite{Selvaraj2002} and used for binding residue prediction in three subsequent publications cite{Ahmad2005, Kuznetsov2006, Wang2006}. Our model (C. 45 with bagging and costing) achieved $approx$78\\% accuracy, $approx$80\\% sensitivity, $approx$77\\% specificity, $approx$86\\% AUC, and an MCC of 0. 57, which is an improvement of +0. 12 in the MCC for the best previous result cite{Kuznetsov2006}. The second data set we tested was compiled and used by Ofran, Mysore, and Rost cite{Ofran2007} and consisted of 274 proteins. Our classifier reached $approx$86\\% accuracy, $approx$85\\% sensitivity, $approx$88\\% specificity, $approx$93\\% AUC, and an MCC of 0. 725. The only directly comparable metric reported in this previous work is accuracy. While our accuracy is slightly lower than that reported by Ofran cite{Ofran2007}, we believe that our model actually offers a more reliable result. In their work, they used sequence to derive evolutionary profiles, sequence neighborhood, and predicted structural features. Their SVM classifier gave its best performance at 89\\% accuracy. However, their `positive accuracy' (precision) and `positive coverage' (sensitivity) were imbalanced. For example, at a sensitivity rate of $approx$80\\% (the number of true positive examples correctly classified), the precision rate is quite low ($approx$55\\%), which indicates that the classifier has low confidence that the predicted positive examples are actually positive. Finally, we tested 109 RNA-binding protein chains originally collected by Terribilini et al. cite{Terribilini2006} and used in three works cite{Kumar2008, Terribilini2006, Wang2008}. Our model achieved $approx$76\\% accuracy,

$approx$75\\% sensitivity, $approx$77\\% specificity, $approx$83\\% AUC, and an MCC of 0. 523 over this set, which is an improvement of +0. 07 in the MCC over the best result cite{Wang2008}. The sequence-based feature sets used in the previous publications varied between works, as did the type of classifier used for prediction and the type of validation performed. While comparisons of this type are not ideal, they do demonstrate that, toward the goal of distinguishing binding from non-binding residues, each of the classifiers we have built using C4. 5 with bagging and costing provides consistent results in terms of overall accuracy when trained over various data sets, thus increasing our confidence in this ensemble method.egin{figure}[! htbp]centeringincludegraphics[width= extwidth]{chap3/figures/GP16_CMYK. eps}caption[Predicted binding residues of GP16]{Predicted binding residues of GP16. Yellow boxes indicate regions with potential DNA-binding residues as identified by our collaborator. Pink boxes indicate binding residues predicted by our classifier. Light blue boxes indicate regions which align to the DNA-binding region of the FtsK protein (PDB ID: 2IUS, cite{Massey2006}) in emph{E. coli}. At the time of this work, FtsK was the most closely-related protein to GP16 having an experimentally-determined 3-D structure in the Protein Data Bank (PDB). FtsK is a multifunctional protein that acts in cell division and chromosome segregation. $approx$55\\% of predicted binding residues lie within the DNA-binding region of FtsK. Red boxes indicate regions of GP16 having $geq$ 50\\% sequence conservation among the 181 sequences in the ABC{\\ _}ATPase superfamily.}label{fig: GP16}end{figure}% Global sequence ID = 13. 3\\%. The highlighted sequence indicates the ABC (ATP-binding cassette) transporter domain of FtsK. subsection{A Prediction Test Case: GP16}In an attempt to validate our method, we predicted the binding residues of the gene-16 protein (GP16), a DNA-packing motor protein in Bacillus phage phi29, for one of our collaborators. This protein contains an ABC transporter nucleotide-binding domain and is known to bind ds-DNA. However, the DNA-binding residues for GP-16 are unknown, and there are no highly-related crystallized protein structures available. Our collaborator had some prior evidence that pointed toward two particular regions of interest in the protein. Using our methods, we were able to focus the costly experimental validation of nucleic acid binding residues on a few key locations in the sequence. We predicted the binding residues of this protein using our sequence-based DNA-binding classifier based on a Platt-calibrated version of the cost-sensitive method described above, which was built using residue charge, identity, and sequence homology information. ef{fig: GP16} shows that our predicted binding residues overlap significantly with the collaborator??? s residues of interest. subsection{The Nucleic Acid Prediction Server: NAPS}The NAPS web server (url{http://proteomics. bioengr. uic. edu/NAPS}) takes a DNA- or RNA-binding protein sequence as input and returns a list of residues, the predicted class (binding or non-binding), and a score indicating the classifier??? s confidence in the decision (ef{fig: NAPSfig}). The model classifier assigns a confidence score between 0 and 1 for each residue in the test protein. This score reflects the level of certainty in the assigned class with 0. 5 as the threshold. Residues with a confidence score between 0 and 0. 5 are classified as non-binding residues; those with a score between 0. 5 and 1 are classified as binding residues (ef{fig: NAPSfig}). A table of calculated statistics, including the total number of residues binned by confidence score, the number of binding and non-binding residues in the protein, the percentage of each class, and the mean confidence value, is also returned. The server calculates a total of 301 sequence-based attributes for each residue in the test protein. We consider a ??? sequence-based attribute??? to be any residue feature that can be calculated without the use of a crystal structure (i. e., only protein sequence). The descriptors are described in more detail below.egin{figure}[! htbp]centeringincludegraphics[width= 0. 75extwidth]{chap3/figures/composite_NAPS_with_conf_CMYK. eps}caption[Screenshot of the extbf{N}ucleic extbf{A}cid extbf{P}rediction extbf{S}erver]{NAPS screen shot. A) The NAPS home page, B) an example output page from the DNA-binding prediction of the protein GP16, C) an illustration of the confidence score used by NAPS to determine residue class.}label{fig: NAPSfig}end{figure}