Bioinformatics Advance Access published September 28, 2009

LIBRUS: Combined Machine Learning and Homology Information for Sequence-based Ligand-Binding Residue Prediction

Chris Kauffman and George Karypis *

Department of Computer Science, University of Minnesota
117 Pleasant St SE, Room 464, Minneapolis, MN 55455

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation: Identifying residues that interact with ligands is useful as a first step to understanding protein function and as an aid to designing small molecules that target the protein for interaction. Several studies have shown sequence features are very informative for this type of prediction while structure features have also been useful when structure is available. We develop a sequence-based method, called LIBRUS, that combines homology-based transfer and direct prediction using machine learning and compare it to previous sequence-based work and current structure-based methods.

Results: Our analysis shows that homology-based transfer is slightly more discriminating than a support vector machine learner using profiles and predicted secondary structure. We combine these two approaches in a method called LIBRUS. On a benchmark of 885 sequence independent proteins, it achieves an area under the ROC curve (ROC) of 0.83 with 45% precision at 50% recall, a significant improvement over previous sequence-based efforts. On an independent benchmark set, a current method, FINDSITE, based on structure features achieves a 0.81 ROC with 54% precision at 50% recall while LIBRUS achieves a ROC of 0.82 with 39% precision at 50% recall at a smaller computational cost. When LIBRUS and FINDSITE predictions are combined, performance is increased beyond either reaching an ROC of 0.86 and 59% precision at 50% recall.

Availability: Software developed for this study is available at `http://bioinfo.cs.umn.edu/supplements/binf2009` along with supplemental data on the study.

Contact: kauffman@cs.umn.edu, karypis@cs.umn.edu

1 INTRODUCTION

Recent advances in high-throughput sequencing technologies have continued to increase the gap between the number of proteins whose function is well-characterized and the proteins for which there is no experimental functional data. As a result, life sciences researchers are becoming increasingly more dependent on computational methods to infer the function of proteins. To address this challenge, a number of novel and sophisticated methods have been developed within the field of computational biology which are designed to predict different aspects of a protein's function.

The focus of this paper is on methods that predict from the protein's primary sequence the residues that bind to small molecules, which are commonly referred to as *ligand-binding residues*. Being able to reliably identify these residues has applications to understanding the overall role and function of the protein by using them to subsequently predict the types of ligands that they bind to and, in the case of enzymes, the types of reactions that they catalyze. Moreover, knowing the residues that are involved in protein-ligand interactions has broad applications in drug discovery and chemical genetics as it can be used to better virtually screen large chemical compound libraries (Bock and Gough (2005)) and to aid the process of lead optimization (Bleicher *et al.* (2003); Weber *et al.* (2002)). In addition, the ligand-binding residues of a protein can be used to influence target-template sequence alignment in comparative protein modeling approaches which has been shown to improve the quality of the 3D models produced for the target's binding site (Kauffman *et al.* (2008)). These quality improvements in the binding site's 3D model is critical to docking-based approaches for virtual screening (Moitessier *et al.* (2007)).

Existing approaches for identifying ligand-binding residues can be broadly classified into two groups.

The first group applies machine learning methods that use a training set of residues with known ligand-binding information to learn a model of binding residues that takes into account residue-level features in order to make predictions. Various types of features and supervised learning methods have been explored. These include features such as sequence conservation measures and position specific scoring matrices and supervised learning methods based on Bayesian learning and support vector machines (Petrova and Wu (2006); Youn *et al.* (2007); Fischer *et al.* (2008)). The consensus of these studies has been that sequence profiles and conservation are the important features and support vector machines provide the best discrimination.

The second group of methods identify the ligand-binding residues of a protein (target) by aligning it to proteins with known ligands. These are referred to as *homology-based transfer methods* (HT) as properties of the target sequence are predicted by transferring them from homologous proteins (templates) (López *et al.* (2007)). Alignment of a target to the templates can be either sequence- or structure-based. The *firestar* algorithm of López *et al.* (2007) utilizes sequence profiles to align the target to templates and the

*to whom correspondence should be addressed

resulting multiple sequence alignment is used to transfer known binders from the templates to the target. The FINDSITE method by Brylinski and Skolnick (2008) identifies template by threading the target through candidate templates and retaining good scoring templates. The accumulated templates are then structurally aligned to the target structure. If the target structure is not available, it is predicted using one of several methods. The binding status of template residues is then transferred to target residues based on this structural alignment. This structure-based approach allows characteristics of potential binding molecules to be discerned but has the drawback of requiring the target structure. Predicting the structure of the target protein can be a computationally expensive proposition.

In this paper we present a new method for predicting the ligand-binding residues of a protein from its primary sequence. This method, called LIBRUS, combines elements from the above machine learning and homology-based transfer methods and achieves accuracy improvements over either one of them. LIBRUS uses support vector machines (SVM) (Vapnik (1995)) to build a prediction model based on features derived from (i) the protein's PSI-BLAST-computed position specific scoring matrix, (ii) its predicted secondary structure, and (iii) a homology-based transfer score that is computed for each residue by using a profile-based sequence alignment scoring method to align the protein to a database of protein sequences with known ligands. Experiments on a set of 885 sequence independent proteins show that LIBRUS achieves ROC and PR (area under ROC and precision-recall curves, Section 3.6) of 0.83 and 0.48, respectively. These are higher than those achieved by the next best-performing method (homology-based transfer), which reached corresponding scores of 0.78 and 0.45. Moreover, comparisons on a set of 564 proteins used in the evaluation of FINDSITE, show that LIBRUS performs comparably to FINDSITE at a fraction of FINDSITE's computational requirements. We also present a ligand-binding prediction method that combines the predictions made by LIBRUS and FINDSITE. This combined predictor outperforms LIBRUS and FINDSITE alone, achieving ROC and PR of 0.86 and 0.56, respectively. These improvements indicate that the sequence-based nature of LIBRUS exploits different but complementary types of signals than the structure-based nature of FINDSITE.

2 METHODS

The ligand-binding residue prediction method that we developed is a hybrid scheme that combines elements of an SVM-based machine learning method trained on sequence-derived features and features derived from a homology-based transfer method. In the subsequent sections we first describe the methods that we developed for predicting the ligand-binding residues of a protein using these two approaches in isolation and then proceed to describe how we combined them to derive LIBRUS.

Note that throughout the rest of this paper we utilize terminology common to the protein homology modeling field: a protein for which prediction is to be made is referred to as a *target* while proteins whose ligand-binding residues are known and are utilized to make the prediction are referred to as *templates*.

2.1 Prediction with Support Vector Machines

In this method, the prediction problem is treated as a supervised learning problem whose goal is to build a model that can predict whether a residue is ligand-binding or not (i.e., binary classification problem). In supervised learning, each object of interest is encoded by a feature vector and a model is learned that can predict the class based on those features.

Following recent research on building models for predicting various structural and functional properties of protein residues in Karypis (2006) and Rangwala *et al.* (2007), we utilized SVMs (Vapnik (1995)) on sequence features of each residue to classify the residue as a ligand-binder or non-binder. Our set of features was comprised of position specific scoring matrices and predicted secondary structure in a window around each residue. We used a sliding window of nine residues centered on the residue of interest and concatenated the PSSMs and SSEs of adjacent residues for a total of 207 features per residue ($9 \times (20+3)$). Window features which extended beyond the first or last residue of the sequence were assigned zero values.

Note that this feature representation is closest to that of Youn *et al.* (2007) where PSSMs in a sliding window of size 21 were employed in one of their methods for the related problem of predicting a protein's catalytic residues. We used a smaller window size, nine residues, as preliminary parameter searches indicated that increasing this to eleven residues did not improve ROC.

One important aspect of this combination was providing proper weights on the features as their numerical ranges varied greatly. We took the following approach. In our dataset, we computed the average norm of PSSM columns and SSE vectors. The average norm for SSEs was smaller than for PSSMs so we up-weighted SSEs to be of equal norm. Relations between the norms and weighting are given in Table 1. Properly weighting the combination of features significantly enhanced the performance of the final model.

Table 1. Average norms of residue features.

Statistic	PSSM	SSE	HTS
Average	13.53	2.00	0.07
Std. Dev.	3.88	0.53	0.11
Weight	1.00	6.75	207.00

Columns are position specific scoring matrices (PSSM), predicted secondary structure vector (SSE), and homology transfer scores (HTS). The bottom row is the weight used on these features in the combined SVMs of Section 2.1 and Section 2.3.

2.2 Prediction by Homology-based Transfer

In this method, given a target protein, a database of template sequences with known binding information is searched for high scoring profile-based alignments to the target. The templates in this database were determined by the experiment (see Section 3.2 and Section 3.4). Once the good templates are identified, a score was assigned to each residue in the target based on the number of template residues which aligned against it and are known to be ligand binders. We optimized the alignment and prediction along a number of dimensions including the profile-based scoring mechanism and the prediction score weighting. When not otherwise noted, parameters for the algorithms that are described below were selected based on performance of homology-based transfer during cross-validation.

2.2.1 Alignment Scoring The profile-based alignment scoring scheme that we used is derived from the work on PICASSO by Mittelman *et al.* (2003) which was shown to be very sensitive in subsequent studies by others (Heger and Holm (2001); Rangwala and Karypis (2007)). Briefly, we aligned sequences by computing an optimal alignment using an affine gap model with aligned residues i and j in sequences X and Y, respectively, scored using a combination of profile-to-profile scoring and secondary structure

matching. The score is given by

$$S(X_i, Y_j) = \sum_{k=1}^{20} PSSM_X(i,k) \times PSFM_Y(j,k) + \sum_{k=1}^{20} PSSM_Y(j,k) \times PSFM_X(i,k) + w_{SSE} \sum_{k=1}^{3} SSE_X(i,k) \times SSE_Y(j,k), \quad (1)$$

where $PSSM$ and $PSFM$ are the position specific scoring and frequency matrices, respectively and SSE is a position-specific matrix encoding the secondary structure (i.e., H, E, C) associated with each position. For a sequence with n residues, these matrices are of sizes $n \times 20$, $n \times 20$, and $n \times 3$, respectively. The parameter w_{SSE} is the relative weighting of the secondary structure score which is set to $w_{SSE} = 3$ based on our experience in Kauffman *et al.* (2008).

We obtained SSE matrices for target proteins by predicting their secondary structure using YASSPP (Karypis (2006)). For each position, the dimensions are the three-state predictions produced by YASSPP which measure the likelihood of that position to be in those states. For the template proteins, since their secondary structure is already known, a straightforward way of defining the SSE is for each position to assign 1 to the dimension corresponding to its true state and 0 to the other dimensions. However, to ensure that the secondary structure information utilized by the learner during training is similar to that used during prediction, the SSE information for the template proteins were encoded in the following way. First all the template proteins were predicted using YASSPP. For the template positions in a helix state, the average of the three-state YASSPP predictions over the helices in all the templates was computed and used as the values for the corresponding columns of the SSE. Identical steps were taken for strand and coil template residues.

In Kauffman *et al.* (2008), which explored homology models of protein-ligand interaction sites, we found gap open and gap extension costs of 3.0 and 1.5 to work well for modelling the binding site and thus re-used these alignment parameters.

We investigated three approaches for aligning the target's profile against the profiles of the templates, which are global, global end-space free, and local alignments. We found local alignments to be the best in terms of overall ROC on cross-validation. This is likely due to local alignments reporting only the best matching target-template subsequences which can increase the reliability of prediction.

We used the top-k scoring templates to make predictions on ligand-binding residues. We investigated $k \in \{5, 10, 20, 30, 40, 50\}$ and found that $k = 20$ had the best performance in terms of overall ROC on cross validation.

2.2.2 Prediction Score Weighting Since we used the top k alignments rather than all alignments above a threshold, properly weighting the contribution of alignments becomes important. Templates that match the target well should influence the prediction more than poor matches. We accomplished this by weighting the contributions of aligned template residues. The simplest way to weight sequences for the contribution to the prediction was equally. We also explored a global weighting based on the alignment score. Finally, local weighting assigned an individual weight to each template residue based on the alignment score (with gaps) between target and template in a window around the target residue of interest. In all weighting schemes, the weights of template residues associated with a target residue were normalized to sum to one. If a negative weight occurred, as is possible for alignment scores, all weights were shifted up so that the lowest weight was equal to one before normalization was performed. The target sequence binding predictions were made by summing the aligned positions in the templates. Positive template residues added their weight to this sum while negative residues and gaps added nothing. This results in a prediction score between zero and one for each residue for each type of weighting scheme. These are referred to as *homology-based transfer scores* (HTS). For the local scheme, we examined windows of width $w \in \{3, 5, 7\}$ and found $w = 7$ to be most effective. Out of the three weighting schemes, the local weighting scheme produced the best results.

2.3 LIBRUS: Combining SVM and Homology-based Transfer

Direct prediction by SVMs and prediction by homology-based transfer utilize training information in different ways to make their predictions. SVMs utilize intrinsic features of the residue represented as PSSMs and SSEs without any context for the residue within the whole protein nor any relation of the containing protein to other proteins in the training set. Conversely, homology-based transfer solely relies on the global context of the residue: where it is located in alignments of the containing protein against other proteins and how many ligand-binding residues align against it. The different characteristics of the information utilized by the two approaches suggests that their combination can lead to a better overall predictor.

To that end, we developed two methods for combining them. The first computes the overall prediction as the weighted linear combination of the predictions made by the two methods while the second approach couples the information that they utilize in a support vector machine.

For the linear combination of methods, we used a grid search to determine which weights, between 0.1 and 7.5, would optimize ROC on the training set. We found weighting the SVM by 0.5 and the HTS scores by 5.5 gave the best overall ROC.

For the SVM-based combination, we trained on the PSSMs and SSEs of the direct prediction method and the homology-based transfer scores of the HT method. The resulting hybrid predictor utilized a total of $9 \times (20 + 3 + 1) = 216$ features. As in Section 2.1, we weighted PSSM, SSE, and HTS features to have a equal average norms. The weights are shown in Table 1.

We will refer to the method that uses the SVM-based combination as LIBRUS and as the experiments reported later in Section 4 show, it achieves the best overall results.

3 MATERIALS

3.1 Data Sets

The methods were evaluated using two different datasets. The first dataset, referred to as DS1, consists of 885 protein chains (268,699 residues) that were derived from the RCSB Protein Data Bank in October of 2008 (PDB, Berman *et al.* (2000)). The set of proteins in DS1 were selected so that they satisfy the following constraints: (i) every structure has better than 2.5 Å resolution, (ii) is longer than 100 residues, (iii) has an unbroken backbone, (iv) has at least five residues in contact with a ligand with each ligand having at least eight heavy atoms, and (v) no two have above a 30% sequence identity according to NCBI's blastclust program.

The second dataset, referred to as DS2, consists of 564 proteins (136,316 residues) that were derived from the set of proteins used in the evaluation of FINDSITE after eliminating from it any proteins with 35% identity or better to any sequence in DS1 according to BLAST. These proteins were eliminated to ensure that we can fairly compare the performance of our methods on DS2 using DS1 as the training set.

Ligands in our datasets were small molecules in contact with proteins identified by scanning the PDB using the 'has ligand' search option. DNA, RNA, and other large proteins were excluded as candidate ligands as were ligands with fewer than eight heavy (non-hydrogen) atoms. We required proteins to have ligand-binding residues with a heavy atom within 5 Å of a ligand. By this definition, 8.6% of DS1 are the ligand-binding residues (positive class) and 9.2% of the DS2 are ligand-biding residues. In-house

software was developed to identify ligands and ligand-binding residues and is available from the web address mentioned in the abstract.

Protein sequences were derived directly from the structures using in-house software. When nonstandard amino acids appeared in the sequence, the three-letter to one-letter conversion table from Astral (Chandonia *et al.* (2002)) version 1.55 was used to generate the sequence. When multiple chains occurred in a PDB file, the chains were treated separately from one another. Profiles for each sequence were generated using PSI-BLAST version 2.2.13 (Altschul *et al.* (1997)) and the NCBI NR database (version 2.2.12 with 2.87 million sequence, downloaded August 2005). PSI-BLAST produces a position specific scoring matrix (PSSM) and position specific frequency matrix (PSFM) for a query protein, both of which are employed for our sequenced-based prediction and alignment methods. Three iterations were used in PSI-BLAST with the default 0.002 e-value threshold for inclusion in the profile and default 10.0 expectation value (options `-j 3 -h 2e-3 -e 10`).

Secondary structure was predicted for each protein of DS1 and DS2 using YASSPP (Karypis (2006)). YASSPP produces a vector of three scores, one for each of the three types of secondary structure, with high positive scores indicating confidence in that class. These scores were used as the secondary structure prediction features (SSE). On this data, YASSPP predicted the correct secondary structure for 83% of the residues in our data set.

3.2 Cross-Validation

The targets of DS1 were split into three sets, set one to three, of roughly equal size and a three-fold cross-validation was performed to assess how well the learner generalizes. In each step, two sets of the data were used to learn a model and predictions were made on the remaining set of targets. This generated a single prediction of binding/non-binding for each residue which was subsequently used in evaluation.

To generate homology-based transfer scores, all targets in set one used sets two and three as the template database and similarly for sets two and three. This amounts to having two thirds of the data as templates for training with remaining third as the test set. This allows us to directly compare the performance achieved by the three methods (direct SVM predictions, homology-based transfer, and LIBRUS) as all methods use identical training and testing data.

The same cross-validation approach was also used to compute the predictions for the method that uses a linear weighted combination of the predictions made by the direct SVM and the homology-based transfer approaches (Section 2.3). Specifically, to determine the predictions for set one, a grid search was performed to learn the combination weights that optimizes ROC on the training set (sets two and three). The combination parameters were then used to compute the predictions for set one. The same was done to predict sets two and three. In each case, the same weights of 0.5 on the direct SVM predictions and 5.5 on the homology-based transfer predictions were determined.

3.3 SVM Implementation

We chose the SVM*light* support vector machine implementation of Joachims (1999) to do prediction. For the kernel, we selected the radial basis function which has a parameter γ. In addition, SVMs have a parameter c representing the trade-off of training error to margin width which must be set. SVM*light* also allows for correction of skewed training data by allowing the cost of misclassifying positive examples to be different than for negative examples, accessed through a parameter j. After tuning parameters on a small subset of the targets, we selected $\gamma = 1 \times 10^{-6}, c = 10$ and $j = 10$ as good candidates for the full model.

3.4 FINDSITE Predictions

We compared our hybrid method, LIBRUS, to FINDSITE, a procedure developed by Brylinski and Skolnick (2008) to predict ligand-binding sites, ligand-binding residues, and other aspects of protein-ligand interactions. For our comparison, we employed predictions made by FINDSITE using entirely predicted structural information for the target as our method does not assume the true structure is available. These predictions were provided by the authors and were made according to their work in Brylinski and Skolnick (2008). Predictions based on the true structure of the target are superior to those based on predicted structures but cannot be made when the true structure is unavailable.

FINDSITE identifies a number of predicted binding sites with associated binding residues for each target. The prediction values for these correspond to the fraction of template structure residues that were identified as ligand binding and aligned against the target residue. We used up to the first five predicted binding sites in our comparison. Some residues appear as part of multiple binding sites in the FINDSITE predictions and have different scores associated with them in the different sites. In those cases, we used the score from the first binding site a residue occurred in as this was typically the largest and most well defined predicted binding site.

FINDSITE uses the nonredundant PDB for a template database. Templates with 35% or better identity to a given target are discarded. This still encompasses a very large number of potential templates. To compare, we used all 885 proteins in DS1 as both the training set for the SVM part of LIBRUS and the template database for the homology-based transfer part of LIBRUS.

Note that the definition of a binding residue in FINDSITE is produced by the LPC program of Sobolev *et al.* (1999). This is a slightly different definition than ours which was based solely on distance. However, there is 98% agreement between LPC and our distance definition on the class division of the DS2 dataset so we used the distanced-based definition for both our methods and FINDSITE in the comparison.

3.5 Combined LIBRUS and FINDSITE Predictions

We also investigated the performance gains that can be achieved by combining LIBRUS with FINDSITE. We did this by making predictions on the DS2 dataset using a linear combination of the prediction values of LIBRUS and FINDSITE. DS2 was split into three folds. To determine the combination weights for fold one, we used the predictions of LIBRUS and FINDSITE on folds two and three and performed a grid search to optimize ROC. The same was done to for folds two (which used one and three for tuning) and three (which used one and two). In each case, the weights scale the predictions from the two methods to a similar range and weight them appropriately for combination. The overall results are reported in Section 4.3.

3.6 Evaluation Metrics

We evaluated the performance of the different methods using the receiver operating characteristic (ROC) curve (Fawcett (2004)). This is obtained by varying the threshold at which residues are considered ligand binding or not according to value provided by the predictor. In the case of the SVM predictions, a continuous prediction value is produced which is the distance from a hyperplane optimized to separate the positive and negative classes. This is the threshold which is varied to produce the ROC curve. For homology-based transfer scores, the threshold to be assigned a ligand binding residue is varied to produce the ROC curve. The area under the this curve, referred to as ROC, is summarizes the predictor behaviour: a random predictor has $ROC = 0.5$ while a perfect predictor has $ROC = 1.0$ so that a larger ROC indicates better predictive power.

For any binary predictor, the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) determines standard classification statistics which we use later for comparison. These are

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ and} \tag{2}$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}. \tag{3}$$

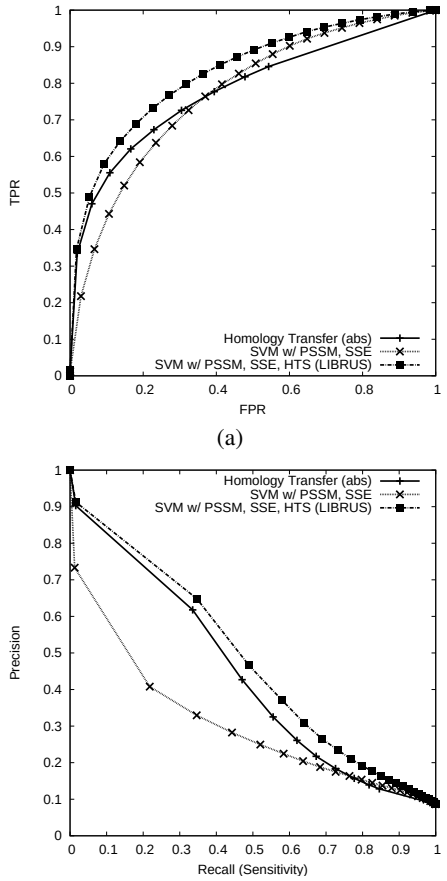


Fig. 1. Comparison of some of the sequence-only predictors developed in this work on the DS1 dataset. (a) ROC curves and (b) Precision vs. Recall.

Fischer *et al.* (2008) noted in their study of functional residue predictions that analyzing only an ROC curve can be misleading in terms of the performance of the predictor. As an alternative, they present precision vs. recall plots (called precision-sensitivity plots in their work, referred to as PR curves here) as a means to compare performance. We provide this measure as well both graphically and summarized by the area under the PR curve (PR).

Welch's t-test is used to assess the statistical significance of performance. This test assumes the two populations are normally distributed with potentially unequal variance and calculates a p-value that the mean of one is higher than the other. In our case, this corresponds to one method outperforming another. Welch's t-test was used in favor of Student's t-test as the latter assumes equal variance of the populations which may not be the case for the methods under consideration.

4 RESULTS

4.1 Cross-Validation Results on the DS1 Dataset

We compared the three methods described in Section 2 on the DS1 dataset. Section 3.2 describes how the data was split to determine training and testing sets for cross-validation. The performance achieved by the methods is shown in Table 2. The ROC and PR curves of some of these methods are shown in Figure 1.

Comparing the best performance achieved by each of the three classes of methods, we see that the methods that combine sequence-derived features along with homology-based transfer information achieve the best overall results. Among the two methods that fall in that category, we see that LIBRUS, which uses SVM to combine this information, achieves the best overall results. Specifically, it achieves an overall $ROC = 0.8334$, which is better than the ROCs of 0.7737 and 0.7849 that were obtained by the SVM and homology-based transfer methods, respectively. Its performance in terms of the overall PR is also better, achieving a $PR = 0.4807$ compared to the PRs of 0.2942 and 0.4516 achieved by the other two classes of methods. These relative performance gains also hold when the experiments are evaluated in terms of the average per-protein ROC and PR. Also, the performance of the other method within this class, which combines the individual predictions using a linear weighted combination also performs quite well, further re-enforcing the fact that coupling the two sources of information lead to a better overall predictor.

Comparing the other two classes of methods, we see that homology-based transfer outperforms the direct SVM-based approach that utilizes PSSM- and SSE-based features. The performance difference between these two schemes is more pronounced when the methods are evaluated in terms of their PR (both overall and per-protein). Finally, the results of Table 2 show that when predicted secondary structure information is used to augment the PSSM-based features, the performance of the SVM-based method improves. This fact is in agreement with a number of studies that have shown that the inclusion of this type of information helps the performance of supervised learning methods (Chen and Kurgan (2007); Ginalski *et al.* (2003)).

4.2 Comparison to FINDSITE

We compared FINDSITE and LIBRUS predictions on the proteins in dataset DS2. FINDSITE predictions were provided by its authors (Section 3.4) while LIBRUS was trained on all proteins in dataset DS1 which is sequence independent from DS2 (Section 3.1). Table 3 summarizes their performance while Figure 2 plots the ROC and PR curves obtained. Note that Tables 3–4 and Figure 2 also contain results for the scheme that combines the LIBRUS and FINDSITE predictions, which are discussed later in Section 4.3. Table 4 shows the results of a paired Welch's t-test comparing the methods. Comparisons on both ROC and PR are done in parts (a) and (b) of Table 4 respectively.

Examining the predictions of the various versions of FINDSITE and LIBRUS, in Table 3 we see that their overall prediction performance is quite close. The FINDSITE results using one site achieve the best PR (0.4955), whereas the FINDSITE results using three sites achieve the best ROC (0.8216). However, compared to the former method, LIBRUS achieves a better ROC (0.8169 vs 0.8088), whereas compared to the latter method, LIBRUS achieves a better PR (0.4565 vs 0.3760). The difference between FINDSITE and LIBRUS is somewhat more consistent when the per-protein results are considered, in which case the FINDSITE results using two sites lead to average ROC and PR (0.8043 and 0.4360) that are better than those produced by LIBRUS (0.7982 and 0.4165).

Table 4 (a) shows that there is no statistical difference between LIBRUS and FINDSITE in terms of ROC performance. This is seen in the LIB row and column of the table in which no small

Table 2. Cross validation results on the DS1 dataset.

Method	Overall		Per Protein			
	ROC	PR	μ_{ROC}	σ_{ROC}	μ_{PR}	σ_{PR}
SVM with PSSM	0.7545	0.2637	0.7487	0.1492	0.2930	0.1722
SVM with PSSM, SSE	0.7737	0.2942	0.7648	0.1532	0.3177	0.1886
Homology Transfer	0.7845	0.4516	0.7581	0.1811	0.4024	0.2971
Linearly Combined SVM and HTS	0.8259	0.4792	0.8030	0.1666	0.4342	0.2838
SVM with PSSM, SSE, HTS (LIBRUS)	0.8334	0.4807	0.8066	0.1686	0.4374	0.2809

Three-way cross validation was used on the set of 885 proteins of the DS1 dataset. The overall area under curve is given for ROC and precision/recall (PR) curves in the first two columns. The per protein averages, μ, and standard deviation, σ, for these two statistics are also given.

Table 3. Results on the DS2 dataset.

Method	Overall		Per Protein			
	ROC	PR	μ_{ROC}	σ_{ROC}	μ_{PR}	σ_{PR}
FINDSITE 1 Site	0.8088	0.4955	0.7981	0.2040	0.4841	0.2978
FINDSITE 2 Sites	0.8187	0.4258	0.8043	0.1935	0.4360	0.2697
FINDSITE 3 Sites	0.8216	0.3760	0.8034	0.1852	0.3957	0.2436
FINDSITE 4 Sites	0.8182	0.3370	0.7970	0.1808	0.3620	0.2228
FINDSITE 5 Sites	0.8155	0.3074	0.7918	0.1716	0.3340	0.2055
SVM with PSSM, SSE, HTS (LIBRUS)	0.8169	0.4565	0.7982	0.1600	0.4165	0.2550
Combined FINDSITE/LIBRUS prediction	0.8617	0.5618	0.8410	0.1741	0.5324	0.2991

The performance of FINDSITE considering the first 5 binding sites and the best SVM method, LIBRUS, are shown. The dataset comprised 564 proteins from the FINDSITE benchmark that were sequence independent from the DS1 dataset that was used to train LIBRUS. The last row shows the results obtained by linearly combining the predictions produced by LIBRUS and FINDSITE 1 Site. For column descriptions, see Table 2.

p-values occur. This lack of significance is interesting as it shows sequence and structure carry approximately equal amounts of information that may be used to identify ligand-binding residues. In terms of PR (Table 4 (b)), examining a single FINDSITE site outperforms LIBRUS at a statistically significant level ($p = 0.002$) while examining two FINDSITE sites is not significantly better than LIBRUS ($p = 0.106$). LIBRUS is nearly better than FINDSITE with three sites at a significant level ($p = 0.081$), and better than four and five sites ($p = 0.000$ for both).

Figure 2 shows the ROC and PR plots graphically. According to part (a), the strength of LIBRUS is at higher false positive rates where it exceeds the TPR of FINDSITE. At low FPR, FINDSITE dominates LIBRUS with the crossing point at FPR=0.35 and FPR=0.40 for one and two sites respectively. In part (b), LIBRUS is seen to have better precision at very low recall, but falls below FINDSITE at 11% recall for one site and at 34% recall for two sites. At 50% recall, LIBRUS achieves 40% precision while FINDSITE achieves 55% and 49% precision for one and two sites respectively.

One aspect that we have not touched on empirically so far is the time required to make predictions. According to communications with the FINDSITE authors, running their program for a protein takes from 30 minutes to several hours. This is not surprising as FINDSITE needs to initially predict the structure of the protein and also identify good templates from their database. The amount of time required by LIBRUS to predict the ligand-binding residues of a protein is much lower. Based on the average performance over many proteins, LIBRUS predictions can be made in under 10 minutes which encompasses profile generation, secondary structure prediction, alignment to the database, and final SVM prediction. A larger template database will lengthen this process somewhat, but we expect it to remain faster.

4.3 Combined LIBRUS and FINDSITE Results

While analyzing the nature of the predictions produced by FINDSITE and LIBRUS, we noticed that, though there is agreement on many of the residues they identified as being ligand-binding, there are enough differences to merit further inquiry. Figure 3 illustrates these differences by plotting the prediction scores produced by LIBRUS and FINDSITE (using one site) for the positive instances (ligand-binding residues) and the negative instances (non-binding residues). In Figure 3(a) (positive class) we see that there are two

Table 4. Statistical comparison of methods on the DS2 dataset.

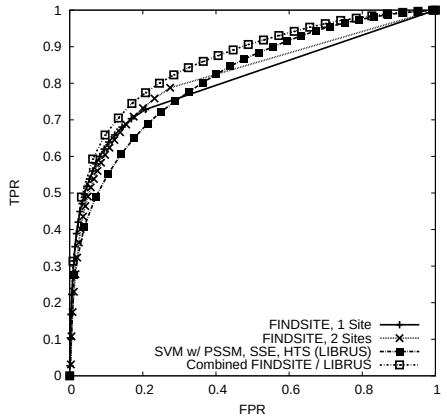
(a) Per Protein ROC p-values

	FS 1	FS 2	FS 3	FS 4	FS 5	LIB.	Comb.
FS 1	0.500	0.701	0.675	0.464	0.289	0.503	1.000
FS 2	0.299	0.500	0.466	0.257	0.126	0.281	1.000
FS 3	0.325	0.534	0.500	0.281	0.140	0.308	1.000
FS 4	0.536	0.743	0.719	0.500	0.310	0.545	1.000
FS 5	0.711	0.874	0.861	0.690	0.500	0.740	1.000
LIB.	0.496	0.719	0.692	0.455	0.260	0.500	1.000
Comb.	0.000	0.000	0.000	0.000	0.000	0.000	0.500

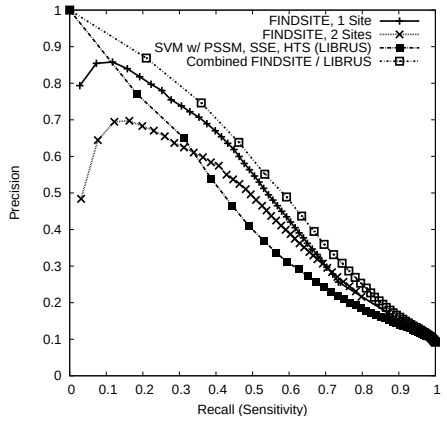
(b) Per Protein PR p-values

	FS 1	FS 2	FS 3	FS 4	FS 5	LIB.	Comb.
FS 1	0.500	0.002	0.000	0.000	0.000	0.000	0.997
FS 2	0.998	0.500	0.004	0.000	0.000	0.106	1.000
FS 3	1.000	0.996	0.500	0.008	0.000	0.919	1.000
FS 4	1.000	1.000	0.992	0.500	0.014	1.000	1.000
FS 5	1.000	1.000	1.000	0.986	0.500	1.000	1.000
LIB.	1.000	0.893	0.081	0.000	0.000	0.500	1.000
Comb.	0.003	0.000	0.000	0.000	0.000	0.000	0.500

Performance of the methods is compared via p-values on Welch's t-test. For the entry at row i, column j of the table, the alternate hypothesis that Method i has a higher mean than method j is tested as an alternative to the methods having equal means. A low p-value indicates that method i has better performance than method j. Part (a) of the table shows performance comparisons in terms of per protein ROC while part (b) shows per protein PR comparisons. FINDSITE for various number of sites are reported in the FS row/columns, LIBRUS in LIB, and the combined FINDSITE/LIBRUS predictor in Comb.



(a)



(b)

Fig. 2. Overall comparison of FINDSITE to the sequence-only SVM learner developed in this work on the 564 independent proteins from the FINDSITE benchmark. (a) ROC curves of FINDSITE based on the top binding sites, the SVM approach, and the combined predictor. (b) Precision vs. Recall of the methods.

clusters, one on the right and one on the left of the plot. The cluster on the right contains residues that FINDSITE predicts correctly, whereas the cluster on the left contains residues that FINDSITE mispredicts. The predictions produced by LIBRUS are, to a large extent, in agreement for the right cluster (even though LIBRUS mispredicts some of these residues) but are split for the left cluster. LIBRUS predicts correctly (i.e., positive SVM score) a noticeable fraction of the residues that are falsely predicted as negative by FINDSITE. Overall, the Pearson correlation coefficient between FINDSITE predictions and LIBRUS predictions is 0.48.

Figure 4(a) illustrates how the above trend carries over to the whole protein. It plots the per protein ROCs of LIBRUS and FINDSITE with one site on DS2 against one another. The greatest density lies in the upper right corner where both methods achieve high ROCs. Points below the main diagonal indicate LIBRUS outperforms FINDSITE while points above indicate the opposite. The large number of off-diagonal points shows that if information from both predictors can be exploited, overall predictions may be improved.

Motivated by the above differences, we developed a method that linearly combines the prediction scores of LIBRUS and FINDSITE. The results of this combined predictor are reported at the bottom of Table 3, and in Figure 2. The combined predictor achieves higher overall ROC and PR than either approach on its own. Also notable is the superior per protein prediction rate of both ROC and PR for the combined method which is statistically significant (Table 4, row/column Comb). This improvement is apparent in Figure 4 (b) in which the combined method achieves performance close to the maximum of both LIBRUS and FINDSITE.

4.4 Comparison to Other Methods

Fischer *et al.* (2008) noted that all methods they tested were below 30% precision at 50% recall. It can be seen from Figure 1(b) that this is the case for our SVM predictions using PSSM and SSE. However, at 50% recall, homology-based transfer on its own achieves 38% precision, whereas LIBRUS achieves a precision of 45% at 50% recall. We obtained evaluation data from Fischer *et al.* (2008) and calculated that their FRcons method achieved $ROC = 0.85$ and $PR = 0.32$ on their CSA-ligand dataset, the closest evaluated dataset to our own, whereas the corresponding values achieved by LIBRUS where $ROC = 0.83$ and $PR = 0.48$. Assuming the trend holds across the different data sets, the use of sequence-derived features and features based on homology-based transfer scores significantly boosts the precision/recall trade off at nearly the same ROC performance.

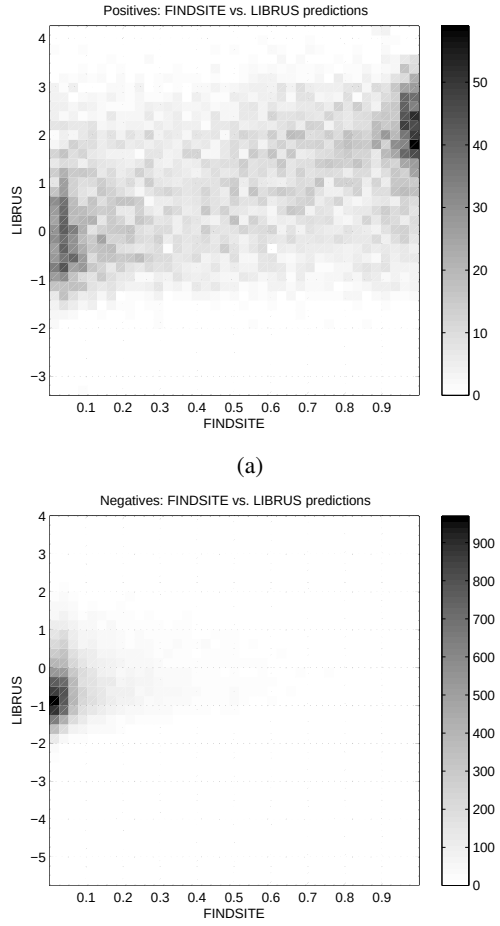


Fig. 3. Heatmap illustrating FINDSITE and LIBRUS values on the positive class (a) and the negative class (b). The positive LIBRUS predictions on some mispredicted FINDSITE residues indicates LIBRUS may provide additional information in some cases. The correlations between FINDSITE and LIBRUS are 0.52 on the positive class, 0.27 on the negative class, and 0.48 overall. Note that residues which had FINDSITE predictions of zero were eliminated as they dominate the nonzero predictions.

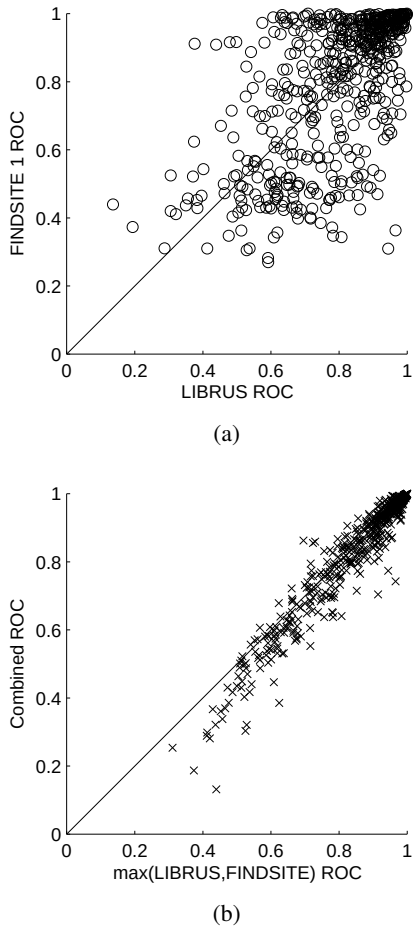


Fig. 4. (a): LIBRUS vs. FINDSITE. The abundance of off-diagonal entries indicate LIBRUS and FINDSITE outperform on another on certain proteins and must be exploiting different signals for those proteins. (b) The ROC of the combined method is plotted against the maximum of LIBRUS and FINDSITE and achieves nearly the same performance.

5 CONCLUSION

In this work we have shown that the combination of machine learning on protein sequence features (PSSMs and SSEs) and homology-based transfer scores results in a powerful binding-residue predictor. Previous efforts have explored these two approaches separately but here we find they provide complementary signals which we exploit in our sequence-only prediction method LIBRUS.

We compared LIBRUS to FINDSITE, a current method for binding residue identification which employs a large database of known structures in order to make predictions. LIBRUS predictions are quite competitive with FINDSITE despite using only sequence features. Additionally, LIBRUS has a comparatively short run time in part due to the focused nature of its prediction.

Combining LIBRUS and FINDSITE predictions achieved better predictive power than either method on its own. This superior performance indicates that the sequence- and structure-based methods are exploiting different types of signal and suggests future methods should focus on combining intrinsic sequence features, sequence homology information, and structural relationships. It is an open question whether this can be accomplished without incurring the computational cost present in FINDSITE. For the present, LIBRUS provides a good alternative to structure-based methods as it achieves comparable accuracy for a very modest runtime.

ACKNOWLEDGEMENTS

We would like to thank Michal Brylinski who graciously provided data from his FINDSITE studies for comparison and corresponded with us about the runtime of FINDSITE.

We would also like to thank Julia Fischer for providing evaluation data from her FRcons paper for comparison.

Funding: This work was supported by the National Institute of Health [T32GM008347, RLM008713A], the National Science Foundation [IIS-0431135], and the University of Minnesota Digital Technology Center.

Supplement: Supplementary materials for this work are available online at `http://bioinfo.cs.umn.edu/supplements/binf2009`.

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucl. Acids Res.*, **25**(17), 3389–3402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucl. Acids Res.*, **28**(1), 235–242.

Bleicher, K. H., Bohm, H.-J., Muller, K., and Alanine, A. I. (2003). Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov*, **2**, 369–378. 10.1038/nrd1086.

Bock, J. R. and Gough, D. A. (2005). Virtual screen for ligands of orphan g protein-coupled receptors. *Journal of Chemical Information and Modeling*, **45**(5), 1402–1414.

Brylinski, M. and Skolnick, J. (2008). A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A*, **105**(1), 129–134.

Chandonia, J.-M., Walker, N. S., Conte, L. L., Koehl, P., Levitt, M., and Brenner, S. E. (2002). Astral compendium enhancements. *Nucleic Acids Res*, **30**(1), 260–263.

Chen, K. and Kurgan, L. (2007). Pfres: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, **23**(21), 2843–2850.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers.

Fischer, J. D., Mayer, C. E., and Söding, J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*.

Ginalski, K., Pas, J., Wyrwicz, L. S., Grotthuss, M. v., Bujnicki, J. M., and Rychlewski, L. (2003). Orfeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucl. Acids Res.*, **31**(13), 3804–3807.

Heger, A. and Holm, L. (2001). Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**(3), 272–279.

Joachims, T. (1999). *Advances in Kernel Methods: Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press.

Karypis, G. (2006). Yasspp: Better kernels and coding schemes lead to improvements in svm-based secondary structure prediction. *Proteins: Structure, Function and Bioinformatics*, **64**(3), 575–586.

Kauffman, C., Rangwala, H., and Karypis, G. (2008). Improving homology models for protein-ligand binding sites. In *LSS Comput Syst Bioinformatics Conference*, San Francisco, CA. (in press). Available at www.cs.umn.edu/ karypis.

López, G., Valencia, A., and Tress, M. L. (2007). firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res*, **35**(Web Server issue), W573–W577.

Mittelman, D., Sadreyev, R., and Grishin, N. (2003). Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**(12), 1531–1539.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., and Corbeil, C. R. (2007). Towards the development of universal, fast and highly accurate docking//scoring methods: a long way to go. *Br J Pharmacol*, **153**(S1), S7–S26.

Petrova, N. V. and Wu, C. H. (2006). Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.

Rangwala, H. and Karypis, G. (2007). frmsdpred: predicting local rmsd between structural fragments using sequence information. *Comput Syst Bioinformatics Conf*, **6**, 311–322.

Rangwala, H., Kauffman, C., and Karypis, G. (2007). A generalized framework for protein sequence annotation. In *Proceedings of the NIPS Workshop on Machine Learning in Computational Biology*.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**(4), 327–332.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag.

Weber, D., Berger, C., Heinrich, T., Eickelmann, P., Antel, J., and Kessler, H. (2002). Systematic optimization of a lead-structure identities for a selective short peptide agonist for the human orphan receptor brs-3. *J Pept Sci*, **8**(8), 461–475.

Youn, E., Peters, B., Radivojac, P., and Mooney, S. D. (2007). Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci*, **16**(2), 216–226.