BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 5 2010, pages 617–624
doi:10.1093/bioinformatics/btq008

Structural bioinformatics

Advance Access publication January 14, 2010

Active site prediction using evolutionary and structural information

Sriram Sankararaman[1], Fei Sha[2], Jack F. Kirsch[3], Michael I. Jordan[1,4] and Kimmen Sjölander[5,6,*]

[1]Computer Science Division, University of California, Berkeley, [2]Computer Science Department, University of Southern California, [3]Department of Molecular and Cell Biology, [4]Department of Statistics, [5]Department of Bioengineering and [6]Department of Plant and Microbial Biology, University of California, Berkeley, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The identification of catalytic residues is a key step in understanding the function of enzymes. While a variety of computational methods have been developed for this task, accuracies have remained fairly low. The best existing method exploits information from sequence and structure to achieve a precision (the fraction of predicted catalytic residues that are catalytic) of 18.5% at a corresponding recall (the fraction of catalytic residues identified) of 57% on a standard benchmark. Here we present a new method, DISCERN, which provides a significant improvement over the state-of-the-art through the use of statistical techniques to derive a model with a small set of features that are jointly predictive of enzyme active sites.

Results: In cross-validation experiments on two benchmark datasets from the Catalytic Site Atlas and CATRES resources containing a total of 437 manually curated enzymes spanning 487 SCOP families, DISCERN increases catalytic site recall between 12% and 20% over methods that combine information from both sequence and structure, and by ≥50% over methods that make use of sequence conservation signal only. Controlled experiments show that DISCERN's improvement in catalytic residue prediction is derived from the combination of three ingredients: the use of the INTREPID phylogenomic method to extract conservation information; the use of 3D structure data, including features computed for residues that are proximal in the structure; and a statistical regularization procedure to prevent overfitting.

Contact: kimmen@berkeley.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 22, 2009; revised on January 1, 2010; accepted on January 4, 2010

1 INTRODUCTION

The prediction of protein function from limited data is an important challenge in the post-genomic era. Bioinformatics methods that provide clues to the roles of individual residues in a protein are used by biologists to prioritize site-directed mutagenesis experiments and to provide a more specific prediction of function than simple homology-based approaches (George *et al.*, 2005). In this work, we focus on the task of predicting catalytic residues in enzymes using information from sequence and structure.

The earliest methods for catalytic residue prediction relied on detecting conservation patterns across a family (Casari *et al.*, 1995; Landau *et al.*, 2005; Lichtarge *et al.*, 1996), followed by increasingly powerful sequence-based scoring functions (Aloy *et al.*, 2001; Mayrose *et al.*, 2004; Mihalek *et al.*, 2004; Sankararaman and Sjölander, 2008). Methods relying exclusively on information from solved 3D structures have been developed, analyzing features such as the geometric arrangements of residues (Fetrow and Skolnick, 1998), surface geometry (Peters *et al.*, 1996), electrostatics (Bate and Warwicker, 2004), energetics (Elcock, 2001; Laurie and Jackson, 2005) and chemical properties (Ondrechen *et al.*, 2001; Tong *et al.*, 2008). Other methods combine features derived from sequence and structure (Aloy *et al.*, 2001; Alterovitz *et al.*, 2009; Gutteridge *et al.*, 2003; Innis *et al.*, 2004; Landgraf *et al.*, 2001; Ota *et al.*, 2003; Pazos and Sternberg, 2004; Petrova and Wu, 2006; Youn *et al.*, 2007), or use sequence data in combination with predicted structure features to improve accuracy (Fischer *et al.*, 2008).

In this article, we present a new method for predicting catalytic residues, which we have named DISCERN. DISCERN is a statistical predictor that achieves a significant improvement in performance over other catalytic residue prediction methods. Previously, the best *recall* (the fraction of true catalytic residues that are predicted to be catalytic) reported on homology-reduced datasets is 57% at a *precision* (the fraction of predicted catalytic residues that are indeed catalytic) of 18.5% (Youn *et al.*, 2007). In comparison, at the same precision, DISCERN yields a recall of at least 69%, representing an improvement of ≥12% in recall over the best current methods for this task.

1.1 The DISCERN methodology for catalytic residue prediction

The statistical model underlying DISCERN is a binary logistic regression model (Hosmer and Lemeshow, 2000), which predicts catalytic residues based on a set of sequence and structure features describing a site. Features considered by DISCERN include evolutionary measures of positional conservation, relative and absolute solvent accessibility, presence in a cleft or pocket, secondary structure, polarity, charge and so on. Logistic regression takes a weighted linear combination of these features, where the weights are learned from a training set of experimentally

*To whom correspondence should be addressed.

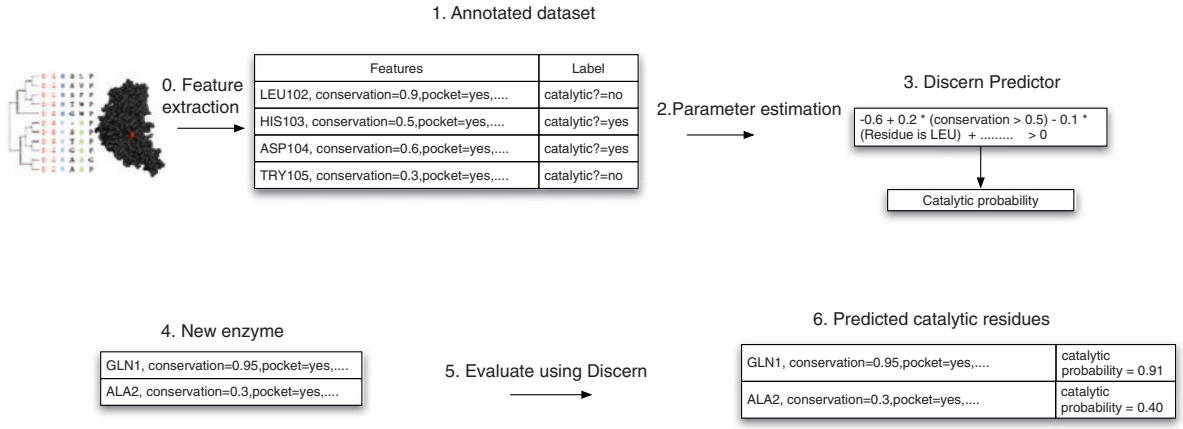

Fig. 1. Overview of the system for catalytic residue prediction: (0) Features are derived from the sequence and 3D structure of an enzyme and from homologs identified using PSI-BLAST. Many features are considered, including the identity of the amino acid, evolutionary conservation scores and presence in a pocket or cleft. (1) Annotated dataset (training data): a dataset of enzymes with labeled catalytic and non-catalytic residues, along with features derived for each residue. (2) We estimate the parameters of the logistic regression model from the training dataset (this is known as a *supervised learning* procedure) using L_1-regularized maximum likelihood. The parameters refer to the weights associated with the features. The L_1-regularization tends to set many of the parameters to zero, resulting in a sparse model. (3) The output of the training phase is a predictor. (4) To predict catalytic residues for a new enzyme, features are derived for the enzyme as in step 1 and the features are used by the logistic regression to classify each residue. (5) The predictor derived in step 3 is used to predict the probability that each residue is catalytic (step 6).

characterized enzymes, and then transforms the result to a probability scale (see Fig. 1 for an overview).

While statistical models making use of information from sequence and structure have been developed for catalytic residue prediction, and individual aspects of the DISCERN model have been used by other methods (Alterovitz *et al.*, 2009; Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007), DISCERN brings together three ideas that jointly differentiate it from existing predictors and which provide a dramatic improvement in prediction accuracy.

The first distinguishing aspect of the DISCERN model is the use of the INTREPID phylogenomic conservation score (Sankararaman and Sjölander, 2008). INTREPID uses Jensen–Shannon (JS) divergence and phylogenetic tree traversal to estimate the evolutionary conservation for each residue in a protein, computing this score at every node encountered on a path from the root of the tree to the leaf corresponding to the sequence of interest. The final score for each residue is the maximum JS divergence computed on that path. This procedure enables INTREPID to extract a conservation signal that may only appear at deeply nested subtrees in the superfamily phylogeny, and allows it to be applied to highly divergent datasets.

The second critical aspect of DISCERN is its use of structure information, in particular, the inclusion of features for structurally proximal residues in the feature vector describing a site. For instance, it is known that enzyme active sites are structurally conserved across distant homologs (Baker and Sali, 2001). This structural conservation is reflected by correspondingly high levels of sequence conservation in the vicinity of catalytic residues. Catalytic residues have other structural features, e.g. they are typically polar or charged, found in clefts or pockets, and at least somewhat solvent accessible (Bartlett *et al.*, 2002). The DISCERN predictor represents these fundamental characteristics of active sites by including features for the individual site whose catalycity is being predicted and also for its structural neighbors.

The inclusion of many features in the statistical model motivates the third critical aspect of DISCERN —the use of an L_1-regularization procedure to avoid model *overfitting*. Overfitting can result when a statistical model has many more parameters than the number of training data points, so that it can fit the training data very precisely but fail to generalize to new data (Hastie *et al.*, 2001). Our results show that regularization is essential for the considerable improvement in DISCERN prediction accuracy, and that performance degrades significantly without regularization (see Supplementary Materials for additional discussion of the overfitting problem). L_1-regularization addresses the problem of overfitting by maximizing the likelihood of the logistic regression model under a constraint on the sum of the absolute values of the model parameters; such a constrained estimation procedure yields a sparse model in which many parameters are set to zero and also derives appropriate weights for features that are highly correlated (or uninformative) (Tibshirani, 1996). L_1-regularization has been shown to yield models that are better predictors than those based on unregularized estimates (Greenshtein and Ritov, 2004; Hastie *et al.*, 2001; Tibshirani, 1996; van de Geer, 2008; Zhao and Yu, 2006), and has been used in a number of bioinformatics applications including gene expression microarray analysis (Segal *et al.*, 2003; Shevade and Keerthi, 2003) and genome-wide association studies (Hoggart *et al.*, 2008).

2 MATERIALS AND METHODS

In this section, we describe the logistic regression model and the estimation procedure underlying DISCERN. See Supplementary Materials for additional details.

2.1 L_1-regularized logistic regression

Given an enzyme i with n_i amino acid residues, we denote by $\boldsymbol{x}_j^{(i)}$ the d-dimensional vector of residue-specific features at site j, $j=1,\ldots,n_i$, by

$\boldsymbol{X}^{(i)}$ the $d \times n_i$ matrix of all such features, and by $z_j^{(i)} \in \{+1, -1\}$ the catalytic label of residue j (whether the residue is catalytic or not). We denote the set of structural neighborhood features by a $dN \times n_i$ matrix $\boldsymbol{Y}^{(i)}$. Here, N refers to the number of structural neighbors of each residue. We model the conditional distribution of the random variable $Z_j^{(i)} \in \{+1, -1\}$ by a logistic regression model

$$\Pr(Z_j^{(i)} = 1 | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{1+\exp\left(-\left(b + \boldsymbol{w}_1{}' \boldsymbol{x}_j^{(i)} + \boldsymbol{w}_2{}' \boldsymbol{y}_j^{(i)}\right)\right)}.$$

The model has parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$; b is the intercept term which controls the trade-off between false positives and false negatives, $\boldsymbol{w}_1$ is the set of weights corresponding to the residue features, while $\boldsymbol{w}_2$ is the set of weights for the structural neighbor features. Given a training set of enzymes and their catalytic residue annotations, we estimate the parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$ using a regularized maximum likelihood approach in which we maximize the sum of the likelihood and an L_1 penalty term:

$$\max_{\boldsymbol{w}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \Pr(z_j^{(i)} | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}) - \lambda \|\boldsymbol{w}\|_1,$$

where $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $\|\boldsymbol{w}\|_1 = \sum_k |w_k|$ is the L_1 norm. The non-negative regularization parameter λ controls the sparsity of the estimate of $\boldsymbol{w}$; larger values of λ lead to estimates with increasing numbers of zero components. We chose the value of λ by a cross-validation procedure. The optimization problem is solved using an interior point method as implemented in Koh *et al.* (2007).

2.2 Features for catalytic residue prediction

The feature vector used in our logistic regression model consists of a total of 528 features—48 features at the residue of interest and at 10 neighboring residues. We provide a brief description of these features in this section as well as some of the options we considered; further details are provided in the Supplementary Materials.

2.2.1 Sequence conservation features We made use of three sequence conservation scores. The first, termed Global-JS, is the JS divergence (Lin and Wong, 1990) between the amino acid distribution over the family as a whole and a background distribution derived from the BLOCKS (Henikoff and Henikoff, 1992) database [with prior weight = 0.5 as in (Capra and Singh, 2007)]. The other two sequence conservation scores make explicit use of the phylogenetic tree topology using the INTREPID algorithm (Sankararaman and Sjölander, 2008). The two variants used the JS divergence (INTREPID-JS) and the log frequency of the modal amino acid (INTREPID-LO). See (Sankararaman and Sjölander, 2008) for additional details.

Sequence conservation scores for each position were derived based on multiple sequence alignments of homologs gathered from the UniProt database (Apweiler *et al.*, 2004) using PSI-BLAST (Altschul *et al.*, 1997). PSI-BLAST was run for four iterations with an E-value inclusion threshold of 1×10^{-4} from which a maximum of 1000 homologs were retrieved. A multiple sequence alignment was estimated using MUSCLE (Edgar, 2004) with MAXITERS set to 2, followed by the removal of identical sequences and the deletion of columns in which the seed had a gap. Phylogenomic conservation scores computed using INTREPID also made use of phylogenetic trees from each alignment. A neighbor-joining tree was built from each alignment using the PHYLIP package (Felsenstein, 1993), using midpoint rooting (placing the root at the midpoint of the longest span in the tree).

2.2.2 Amino acid properties Amino acids have varying catalytic propensities as noted in Bartlett *et al.* (2002). We use the amino acid types as features and also classify the amino acid into one of three categories—charged (D,E,H,K,R), polar (Q,T,S,N,C,Y) or hydrophobic (A,F,G,I,L,M,P,V,W). See Supplementary Materials for a description of this classification.

2.2.3 Structure-based features For each residue, we compute the residue centrality, the B-factor, solvent accessibility, presence in a cleft and secondary structure as follows. We compute the B-factor, a measure of thermal motion for each residue as the average of the B-factors of all its atoms. We compute a measure of centrality for each residue j as the inverse of the average distance from a residue to all other residues in the enzyme; i.e. $C_j = \frac{n-1}{\sum_{k \neq j} d(k,j)}$ where $d(k,j)$ is the distance from j to k along the contact map. A residue that is located in the center of the protein has smaller average distance to all other residues and hence a high centrality measure. We use the seven-state secondary structure representation output by DSSP (Kabsch and Sander, 1983). The area of a residue accessible to the solvent is obtained from NACCESS (Hubbard and Thornton, 1993). We use LigSite^{csc} (Huang and Schroeder, 2006) to detect the presence of a residue in one of the three largest pockets in the enzyme.

2.3 Benchmark datasets

We present results from two datasets of manually curated enzymes from the CATRES (Bartlett *et al.*, 2002) and Catalytic Site Atlas (CSA; Porter *et al.*, 2004) datasets. CSA and CATRES define a residue as catalytic if it has been shown to be involved in catalysis either directly or through other molecules, to stabilize an intermediate transition state, or to influence a cofactor or substrate that aids catalysis. The manually curated sections of CSA and CATRES contain enzymes with solved PDB structures for which experimental evidence for catalytic sites have been obtained from the literature.

Our primary benchmark dataset, termed CATRES-FAM, consists of 140 enzymes from the CATRES dataset, and was included to allow a direct comparison with Youn *et al.* and Gutteridge *et al.*. This dataset contains a total of 471 catalytic residues out of a total of 49 180 residues with a median of three catalytic residues per enzyme.

Our second dataset, termed CSA-Fischer, consists of 423 enzymes from the manually curated section of the CSA selected by Fischer and colleagues (2008) to benchmark their FRcons method, and used here to allow a direct comparison to FRcons.

Additional information on these benchmark datasets and results and details on two other datasets are reported in Supplementary Materials.

2.4 Performance measurements

We measure the precision and recall on the test set as follows: Precision = $\frac{\text{TP}}{\text{TP+FP}}$, Recall = $\frac{\text{TP}}{\text{TP+FN}}$, where a true positive (TP) is a predicted residue included in the benchmark dataset, a false positive (FP) is a predicted residue not listed in the benchmark and a false negative (FN) is a catalytic residue in the benchmark which has been missed by a method. The precision–recall curves were averaged over all the cross-validation folds using the code from Davis and Goadrich (2006). See Section S-4.2.1 in Supplementary Materials for more details.

For the CSA-Fischer dataset, we followed the protocol described in Fischer *et al.* (2008), i.e. we performed 2-fold cross-validation, ensuring that no domain from the same SCOP superfamily appeared in both the folds. For CATRES-FAM and other datasets (reported in the Supplementary Materials), we used 10-fold cross-validation.

2.5 ConSurf and Evolutionary Trace results

ConSurf results for CATRES-FAM were obtained from the database of precomputed results (http://consurfdb.tau.ac.il). Evolutionary Trace (ET) results were obtained from the precomputed results of the ET server at the Baylor College of Medicine (http://mammoth.bcm.tmc.edu/ETserver.html).

3 RESULTS

In this section, we report results of large-scale experiments on manually curated enzymes from the CSA (Porter *et al.*, 2004) and

CATRES (Bartlett *et al.*, 2002) datasets, and compare DISCERN with the best methods for catalytic residue prediction reported in the literature. Three of these methods make use of machine learning algorithms to combine sequence and structure information (or inferences): a neural network approach from Gutteridge *et al.* (2003), a support vector machine (SVM) method from Youn *et al.* (2007) and the FRcons method from Fischer *et al.* (2008). (Note that FRcons uses sequence information only, but predicts structural features to improve performance.) Three other methods tested make use of sequence conservation information only: ConSurf (Landau *et al.*, 2005), ET (Mihalek *et al.*, 2004) and INTREPID (Sankararaman and Sjölander, 2008). Web servers, software or precomputed results were available for ET, ConSurf and INTREPID making possible a head-to-head comparison with these methods.

We compared DISCERN against Gutteridge *et al.*, Youn *et al.* and FRcons based on precision and recall statistics reported by the authors. We also include a control method in these experiments designed to evaluate the contributions of the different ingredients of the DISCERN predictor. The control was trained identically to DISCERN, but did not include features for structural neighbors or the INTREPID phylogenomic conservation scores, nor was any attempt made to enforce model sparsity. Notably, the performance of the control is very similar to the results reported in Youn *et al.*, suggesting that the improved performance of DISCERN relative to Youn *et al.* is unlikely to be an artifact of differences between the CATRES-FAM dataset and the datasets used by these authors.

We used cross-validation on two benchmark datasets to evaluate DISCERN performance in catalytic site prediction, reporting the average recall and precision in the withheld test sets in each partition. The first dataset, CATRES-FAM, was designed to allow comparisons to methods developed by Youn *et al.* (2007) and Gutteridge *et al.* (2003). The dataset used by Youn *et al.* (2007) consists of a random subset of the domains present in ASTRAL 40v1.65 (Chandonia *et al.*, 2004). Since the domains that were finally selected were not recorded (E. Youn, personal communication), we could not evaluate DISCERN on their dataset. CATRES-FAM consists of 140 enzymes from CATRES filtered at the SCOP (Structural Classification of Proteins; Murzin *et al.*, 1995) family level (i.e. no pair were from the same SCOP family). The second dataset, CSA-Fischer, consists of 423 enzymes from the CSA selected by Fischer and colleagues (2008) to benchmark FRcons, and used in these experiments to allow a direct comparison with FRcons.

On the CATRES-FAM dataset, as shown in Figure 2, DISCERN recall is 12–20% higher than that of Gutteridge *et al.* and Youn *et al.* at the levels of precision reported by these authors. Relative to methods that are restricted to conservation signal only (INTREPID, ConSurf and ET), DISCERN has 50% greater recall: at a precision of 18%, DISCERN has 69% recall, while INTREPID and ET reach 19% and 2% recall, respectively (ConSurf does not attain a precision of 18% over the entire range of recalls). We also evaluated two prediction methods that make use of 3D structure information only, LigSite^{csc} (Huang and Schroeder, 2006) and PASS (Brady and Stouten, 2000), on this dataset. Since these methods do not provide scores for individual residues, we used the residues in the top three sites identified by each method as predicted active site residues. Using this criterion, PASS attained a recall of 29.7% for a corresponding precision of 3%, and LigSite obtained a recall of 10.6% at a corresponding precision of 1.2%.

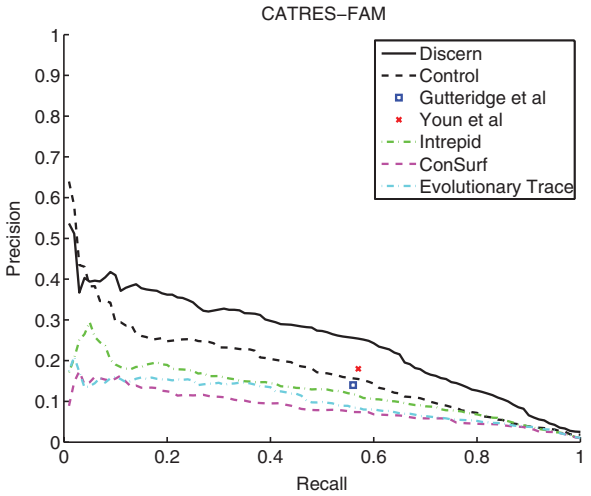

Fig. 2. Results on the CATRES-FAM benchmark dataset. Methods included in this analysis are a neural network method from Gutteridge *et al.* (2003), a Support Vector Machine approach from Youn *et al.* (2007), ConSurf (Landau *et al.*, 2005), INTREPID (Sankararaman and Sjölander, 2008), ET (Mihalek *et al.*, 2004) and the control method described in Table 1. The methods of Youn *et al.*, Gutteridge *et al.*and the control use information from both sequence and structure, while ConSurf, INTREPID and ET use conservation signal only to predict catalytic sites. In cross-validation experiments on this dataset, at 18% precision, DISCERN reaches 69% recall, corresponding to a gain in recall of 20% over Gutteridge *et al.* and 12% over Youn *et al.* (based on their reported performance at this level of precision on similar datasets). Relative to the methods that use conservation signal only, the difference is greater: at 18% precision, INTREPID reaches 19% recall, while ET reaches a recall of 2%. ConSurf does not reach 18% precision on this dataset.

On the CSA-Fischer dataset, DISCERN provides superior performance relative to FRcons for recall values >30%. At a precision of 18.5% [reported by (Youn *et al.*, 2007)], DISCERN achieves 15% higher recall than FRcons (DISCERN and FRcons achieve 65% and 50% recall, respectively). Analysis of the area under the precision–recall curve, termed PR-AUC, shows that the PR-AUC of FRcons is 0.1 compared with 0.23 for Discern. On this more extensive dataset, DISCERN recall is 14% higher than that of Youn *et al.* (2007) and 18% higher than that of Gutteridge *et al.* (2003) at the precision levels reported by these authors. See Supplementary Figure S-2 for details.

In addition to these large-scale experiments, we present in the next section a detailed case study of *Bovine* α-Chymotrypsin (PDB id:1acb). Additional experiments on datasets filtered to remove members from the same SCOP superfamily and a second case study on *Escherichia coli* Asparagine Synthetase (PDB id:12as) are reported in the Supplementary Materials (Section S-5.1).

3.1 *Bovine* α-Chymotrypsin (PDB id:1acb, E.C. number: 3.4.121.1)

Chymotrypsin (E.C. number 3.4.121.1) is the paradigmatic member of the so-called serine protease family of enzymes that are distinguished by having a catalytic triad of residues at the active site (H57, D102 and S195) (Hedstrom, 2002; Kraut, 1977; Polgar, 2005).

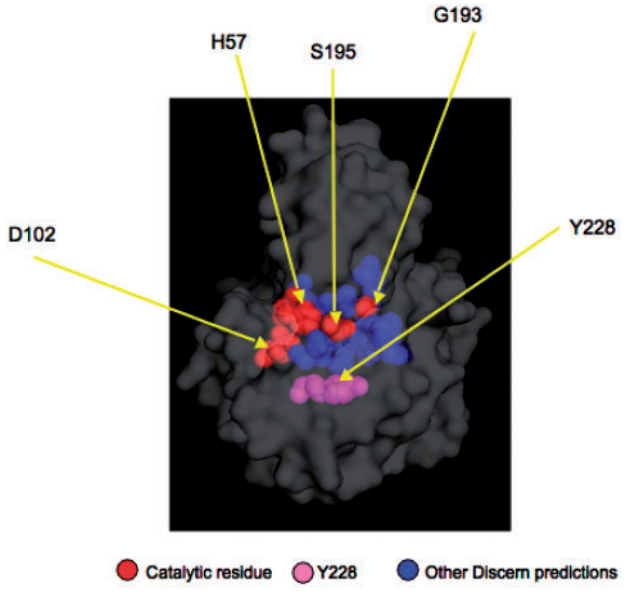

Fig. 3. DISCERN predictions on *Bovine* α-Chymotrypsin (PDB id:1acb). The top 15 DISCERN predictions are shown. DISCERN predicts the catalytic triad H57, D102 and S195, with ranks 6, 4 and 1, respectively. The catalytic glycine, G193, is predicted with rank 13. Y228 (DISCERN rank 10) is found in the S1 specificity pocket, but its functional role is unknown. The roles of Y228 and other residues (D194, C191, C42, C58, Q30, C220, S214, G197, H40 and G196) are described in Supplementary Table S-2.

To predict catalytic residues for the α-chymotrypsin structure 1acb, we estimated DISCERN parameters using a subset of the CATRES-FAM dataset, removing all enzymes in the same SCOP superfamily as 1acb. The top 15 residues predicted by DISCERN are shown in Figure 3, with additional details provided in the Supplementary Table S-2.

DISCERN gives the catalytic serine (S195) rank 1. The β-hydroxyl moiety of S195, aided by general base catalysis by the imidazole (NE2 nitrogen atom) group of H57 (rank 6), attacks the carbon atom of the scissile peptide or ester substrate to form a tetrahedral adduct, which, in turn, decomposes to form a covalent enzyme bound ester intermediate with concomitant release of the amino or hydroxyl portion of the peptide or ester substrate, respectively. The general base catalysis is assisted in a way that is not fully understood by the β-carboxylate of D102 (rank 4), whose β-carboxylate functionality makes a strong hydrogen bond (Frey *et al.*, 1994) with the second nitrogen atom (ND1) of H57. The covalent intermediate is subsequently hydrolyzed via H57/D102-mediated activation of the attacking water molecule to yield the carboxylate component of the substrate with regeneration of the enzyme. The transition state leading to the tetrahedral intermediate is stabilized by developing hydrogen bonds from the main chain NH groups of G193 (rank 13) and S195.

The pancreatic serine proteases are biosynthesized in the pancreas as inactive proenzymes, which are activated in the small intestine by proteolytic cleavage of a 15-member peptide from the N-termini. This results in a number of conformational changes with concomitant repositioning of hydrogen bonds involving several residues including Q30 (rank 8) (Kraut, 1977) and H40 (rank 14) (Berna *et al.*, 1997). D194 (rank 2) forms a salt bridge with the nascent I16 that forms the N-terminus of the active enzyme.

DISCERN identifies residues G197 (rank 12) and G196 (rank 15); these allow the peptide chain to form a distinct structural element called a β-bulge (Richardson *et al.*, 1978) which may be important for positioning the active site serine (S195). DISCERN also identifies C191 (rank 3) and C220 (rank 9); C191 and C220 form a disulfide bond which has been shown to be critical for enzymatic function (replacement of C191 and C220 with a pair of alanines resulted in a 100- to 1000-fold decrease in activity) (Vàrallyay *et al.*, 1997). Another pair of cysteine residues forming a disulfide bond are found in the top 15: C42 (rank 5) and C58 (rank 7). The C42–C58 disulfide is part of the binding site for the amino terminus of the scissile peptide bond (the P1′ site) (Kraut, 1977).

The roles of two remaining residues in DISCERN's top 15 predictions are unknown. The highly conserved S214 (rank 11) is in hydrogen bond contact with one of the β-carboxylate oxygen atoms of D102, and S214E and S214K mutants have been shown to disrupt function, but an S214A mutant is as active as wild-type enzyme in the hydrolysis of a tripeptide substrate (McGrath *et al.*, 1992). However mutation of this residue in thrombin, a closely related serine protease, does lead to increased K_m values for various substrates (Krem *et al.*, 2002). The proximity of this residue to the active site and the degree of conservation argue that it is important in function, although the role remains to be more precisely defined. Y228 (rank 10) is found in the S1 binding pocket (Hedstrom *et al.*, 1992), but its role is unknown.

In summary, of DISCERN's top 15 predicted residues, all but Y228 are known or proposed to have important roles in catalysis, substrate recognition, proenzyme activation or formation of key structural elements in chymotrypsin. Given the very high percentage of identification of important residues whose functions have been verified experimentally, the DISCERN results suggest that mutagenic probing of Y228 in particular might be illuminating, and that DISCERN can be generally useful in guiding experimental approaches to mechanistic investigations of enzymes that have been much less studied than chymotrypsin.

3.2 Aspects of the DISCERN predictor

DISCERN combines three ingredients in making a prediction—the use of phylogenomic scores, information from structure and features computed at structural neighbors, and a statistical regularization to control for overfitting. To investigate the relative importance of these three aspects of the predictor, we conducted a set of experiments in which subsets of these aspects were used. The results are shown in Table 1. We see that a performance gain is obtained by including phylogenomic scores. However, a decrease in performance is seen when structural neighborhood features are also included but the model is not regularized. This is presumably due to overfitting. Indeed, when the model is regularized, a significant performance gain is observed.

We investigated quantitative aspects of the full DISCERN predictor after it has been fit to the CATRES dataset (Fig. 4). Among the 528 candidate features considered, 157 had non-zero weights in the final model. Examining these weights provides insight into the ability of DISCERN to discriminate between catalytic and non-catalytic residues. The highest weights are associated with features identified by others as highly correlated with catalytic sites (e.g. high degrees

Table 1. Comparison of DISCERN to simplified logistic regression models

Method	Structural neighbors	Phylogenomic conservation scores	L_1-regularization	CATRES-FAM	
				Precision_{50} (%)	Recall_{18} (%)
Method 0 (Control)	–	–	–	17.00	48
Method 1	–	Y	–	20.45	55
Method 2	Y	Y	–	16.13	41
DISCERN	Y	Y	Y	27.30	69

We compare DISCERN to simplified models that make use of conservation signal across the family as a whole and structural features for the residue of interest, but do not include one or more of (1) features computed for structural neighbors, (2) INTREPID phylogenomic conservation scores and (3) L_1-regularization. Precision_{50} reports the precision at 50% recall, and Recall_{18} reports the recall at 18% precision (these precision and recall points were selected to allow direct comparison to the Youn *et al.* method). DISCERN provides an improvement over the control of 10.3% precision at 50% recall and an increase in recall of 21% at 18% precision. See Section S-5.3 and Figure S-6 in the Supplementary Materials for additional details on these experiments and full precision-recall curves.

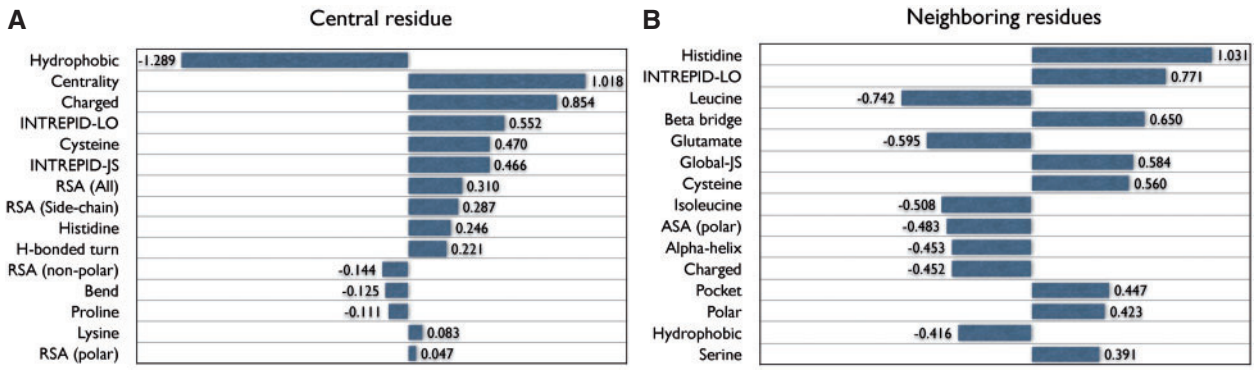

Fig. 4. Features selected by DISCERN. Shown here are the 15 features with the largest weights based on fitting the logistic regression to the entire CATRES-FAM dataset. Positive weights indicate positive correlation with putative catalytic residues; negative weights imply negative correlation. The magnitude of the weight is indicative of a feature's relative importance. (**A**) Features computed at the residue of interest. (**B**) Features summed over residues that are nearby in the 3D structure. See Supplementary Figure S-7 for additional details.

of sequence conservation across homologs, centrality in 3D structure and relative solvent accessibility), and the largest negative weights are those shown previously as anti-correlated (e.g. hydrophobicity) (Bartlett *et al.*, 2002).

A more subtle point is the fact that the DISCERN prediction is based on a combination of weighted features. For a residue to achieve a high rank (relative to other residues), a combination of features must be present (or absent, in the case of a feature with negative weights). For instance, while residue centrality has a strong positive weight, this alone will be insufficient to give a residue a high rank unless it is also highly conserved, polar or charged, and has some level of relative solvent accessibility.

L_1-regularization constrains the total weight allocated to a set of features, with the end result that some features receive zero weight. In many cases, these features are individually informative but are effectively redundant due to other features which are given non-zero weight (i.e. included in the final model). Location in a cleft or pocket is a case in point. We found that the explicit feature of presence in a cleft or pocket is given a weight of zero in our model, which is surprising given that presence in a cleft is known to be one of the hallmarks of catalytic residues (Bartlett *et al.*, 2002). However, residue centrality and relative solvent accessibility (features which were given positive weights) jointly encode for presence in a cleft; i.e. if a residue is both near the center of the molecule and exposed, it must be in a deep cleft. Thus, enforcing model sparsity using L_1-regularization resulted in dropping the feature of presence in a cleft or pocket, but retained residue centrality and solvent accessibility which allow this defining characteristic of active site residues to be recognized.

In summary, the features selected by the regularized logistic regression jointly describe highly conserved, charged, solvent-accessible residues that are found in clefts or pockets, and whose neighbors in the 3D structure are also highly conserved.

4 DISCUSSION

In this article, we have described a new approach to the prediction of active sites in proteins. Our results on benchmark datasets of manually curated enzymes from the CSA and CATRES resources show that DISCERN provides a significant improvement over the best methods that make use of information from sequence and/or structure to predict catalytic sites.

DISCERN is a statistical predictor that brings together three important ideas, the combination of which are needed in order to obtain the striking improvements in accuracy shown here. First, DISCERN uses an evolutionary modeling approach (specifically, the INTREPID phylogenomic method) to infer the degree to which residues are under selective pressure. Second, we incorporate

information from the structural neighborhood of a residue including features (such as sequence conservation, charge, solvent accessibility, etc.) computed for structurally proximal residues. Third, and critically, we use statistical sparsification methods (specifically, L_1-regularization) to cope with the fact that our statistical model is based on a large number of redundant, noisy features. Without such regularization, we find that our method overfits—in particular, the inclusion of information from structural neighbors leads to a decrease in accuracy. With regularization, we obtain a significant increase in accuracy. Regularization allows us to find a signal within the large set of candidate features that can be used to describe the structural and evolutionary neighborhood of an amino acid.

The parameters of the statistical model underlying DISCERN are the weights of various features that capture the evolutionary and structural context, computed both for the residue of interest and for its structural neighbors. The largest weights tend to be associated with features identified by others as highly correlated with catalytic sites (e.g. high degrees of sequence conservation across homologs, centrality in 3D structure and relative solvent accessibility), and the largest negative weights are those shown previously as anti-correlated (e.g. hydrophobicity). But the model is not restricted to such known features; it can create new features as linear combinations of the given features. Moreover, the model parameters act in concert: for a residue to achieve a high rank, a single feature is generally insufficient; multiple features must be present. The features selected by DISCERN jointly describe highly conserved, charged, solvent-accessible residues that are found in clefts or pockets, and whose neighbors in the 3D structure are also highly conserved.

While many catalytic site prediction methods exploit residue conservation as a primary source of signal (Gutteridge *et al.*, 2003; Youn *et al.*, 2007), most of these restrict homologs to only moderately divergent sequences, limiting the effective use of this signal. In contrast, DISCERN makes use of the INTREPID phylogenomic conservation score (Sankararaman and Sjölander, 2008), which is able to exploit the conservation information in highly divergent datasets.

DISCERN is not the only method to use information from structural neighbors for catalytic residue prediction, but there are a few differences between DISCERN and approaches used by others that may contribute to the improved performance. In particular, several methods use spatial clustering (Aloy *et al.*, 2001; Landgraf *et al.*, 2001; Panchenko *et al.*, 2004) as a post-processing step (Gutteridge *et al.*, 2003) based on classification of individual positions independently in an initial stage. In contrast, DISCERN uses features from structurally neighboring residues as an integral part of the model. Closer in spirit to DISCERN is the method proposed by Youn *et al.* (2007), which uses atom-level features (Bagley and Altman, 1995) in concentric shells (weighted equally within each shell) around the C_β atom of the residue of interest (Mooney *et al.*, 2005). As in DISCERN, this yields a rich set of features describing the neighborhood. Crucially, however, Youn *et al.* do not enforce a penalty that enforces sparsity of parameters in their model, and the poorer performance of Youn *et al.* (2007) relative to DISCERN may reflect the kind of overfitting that we observe in Table 1.

In this work, we have evaluated DISCERN on two large-scale datasets: the CATRES benchmark dataset (Bartlett *et al.*, 2002) and a homology-reduced subset of manually curated enzymes from the CSA (Porter *et al.*, 2004). While CATRES and CSA provide important resources to benchmark the accuracy of prediction methods, finite resources (e.g. a small number of biological curators entering data into the CSA) and the inevitable lag between publication and data entry can result in not all catalytic residues being included. As our case studies show, this can result in residues that are predicted by a method as catalytic being labeled as false positives even if they are, in fact, catalytic.

Finally, our case studies suggest that DISCERN can be effective at identifying general types of functionally important positions (such as ligand-binding residues), and is not restricted to catalytic residue identification *per se*. In fact, the general approach underlying DISCERN is extensible and general, and can be applied to model other types of functional residues such as binding pocket specificity determinants and interaction interfaces. Each of these application areas depends only on the availability of high-quality training data, such as that provided in the CSA.

Funding: Presidential Early Career Award for Scientists and Engineers (grant number 0238311 to K.S.) from the National Science Foundation; National Science Foundation (grant number 0732065 to K.S.); National Institutes of Health (grant number HG002769 to K.S.); Department of Energy (BER KP110201 to M.I.J.); NIH/NIGMS (R01 GM071749 to M.I.J.); National Institutes of Health (grant number GM35393 to J.F.K.).

Conflict of Interest: none declared.

REFERENCES

Aloy,P. *et al.* (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.

Alterovitz,R. *et al.* (2009) Resboost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinformatics*, **10**, 197.

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Bagley,S.C. and Altman,R.B. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci.*, **4**, 622–635.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Bartlett,G.J. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

Bate,P. and Warwicker,J. (2004) Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.*, **340**, 263–276.

Berna,P.P. *et al.* (1997) Residue accessibility, hydrogen bonding, and molecular recognition: metal-chelate probing of active site histidines in chymotrypsins. *Biochemistry*, **36**, 6896–6905.

Brady,G.P. and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **14**, 383–401.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Chandonia,J.M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**(Database issue), D189–D192.

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. In *ICML'06: Proceedings of the 23rd International Conference on Machine Learning*. ACM, New York, pp. 233–240.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

Felsenstein,J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. *Distributed by the author. Department of Genetics, University of Washington, Seattle.*

Fetrow,J. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **281**, 949–968.

Fischer,J.D. *et al.* (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.

Frey,P. *et al.* (1994) A low-barrier hydrogen bond in the catalytic triad of serine proteases. *Science*, **264**, 1927–1930.

George,R.A. *et al.* (2005) Effective function annotation through catalytic residue conservation. *Proc. Natl Acad. Sci. USA*, **102**, 12299–12304.

Greenshtein,E. and Ritov,Y. (2004) Persistence in high-dimensional predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 971–988.

Gutteridge,A. *et al.* (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.

Hastie,T. *et al.* (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hedstrom,L. *et al.* (1992) Converting trypsin to chymotrypsin: the role of surface loops. *Science*, **255**, 1249–1253.

Hedstrom,L. (2002) Serine protease mechanism and specificity. *Chem. Rev.*, **102**, 4501–4524.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hoggart,C.J. *et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*. John Wiley, New York.

Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.

Hubbard,S. and Thornton,J. (1993) A computer algorithm to calculate surface accessibility. Department of Biochemistry and Molecular Biology, University College, London.

Innis,C. *et al.* (2004) Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, **337**, 1053–1068.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Koh,K. *et al.* (2007) An interior-point method for large-scale L1-regularized logistic regression. *J. Mach. Learn. Res.*, **8**, 1519–1555.

Kraut,J. (1977) Serine proteases: structure and mechanism of catalysis. *Annu. Rev. Biochem.*, **46**, 331–358.

Krem,M.M. *et al.* (2002) Ser214 is crucial for substrate binding to serine proteases. *J. Biol. Chem.*, **277**, 40260–40264.

Landau,M. *et al.* (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**(Web Server issue), W299–W302.

Landgraf,R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lin,J. and Wong,S.K.M. (1990) A new directed divergence measure and its characterization. *Int. J. Gen. Syst.*, **17**, 73–81.

Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

McGrath,M.E. *et al.* (1992) Perturbing the polar environment of Asp102 in trypsin: consequences of replacing conserved Ser214. *Biochemistry*, **31**, 3059–3064.

Mihalek,I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

Mooney,S.D. *et al.* (2005) Structural characterization of proteins using residue environments. *Proteins Struct. Funct. Bioinform.*, **61**, 741–747.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Ondrechen,M.J. *et al.* (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.

Ota,M. *et al.* (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, **327**, 1053–1064.

Panchenko,A.R. *et al.* (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

Pazos,F. and Sternberg,M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.

Peters,K.P. *et al.* (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, **256**, 201–213.

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

Polgar,L. (2005) The catalytic triad of serine peptidases. *Cell. Mol. Life Sci.*, **62**, 2161–2172.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**(Database issue), D129–D133.

Richardson,J.S. *et al.* (1978) The beta bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.

Sankararaman,S. and Sjölander,K. (2008) INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics*, **24**, 2445–2452.

Segal,M. *et al.* (2003) Regression approaches for microarray data analysis. *J. Comput. Biol.*, **10**, 961–980.

Shevade,S. and Keerthi,S. (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, **19**, 2246–2253.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Stat. Meth.*, **58**, 267–288.

Tong,W. *et al.* (2008) Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Sci.*, **17**, 333–341.

van de Geer,S.A. (2008) High-dimensional generalized linear models and the lasso. *Ann. Stat.*, **36**, 614–645.

Vàrallyay,E. *et al.* (1997) The role of disulfide bond C191-C220 in trypsin and chymotrypsin. *Biochem. Biophys. Res. Commun.*, **230**, 592–596.

Youn,E. *et al.* (2007) Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, **16**, 216–226.

Zhao,P. and Yu,B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.