BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 23 2007, pages 3155–3161
doi:10.1093/bioinformatics/btm509

Structural bioinformatics

Accurate prediction of enzyme mutant activity based on a multibody statistical potential

Majid Masso and Iosif I. Vaisman*

Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, 10900 University Boulevard, MSN 5B3, Manassas, VA 20110, USA

Received on April 9, 2007; revised on September 11, 2007; accepted on October 8, 2007

Advance Access publication October 31, 2007

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: An important area of research in biochemistry and molecular biology focuses on characterization of enzyme mutants. However, synthesis and analysis of experimental mutants is time consuming and expensive. We describe a machine-learning approach for inferring the activity levels of all unexplored single point mutants of an enzyme, based on a training set of such mutants with experimentally measured activity.

Results: Based on a Delaunay tessellation-derived four-body statistical potential function, a perturbation vector measuring environmental changes relative to wild type (wt) at every residue position uniquely characterizes each enzyme mutant for model development and prediction. First, a measure of model performance utilizing area (AUC) under the receiver operating characteristic (ROC) curve surpasses 0.83 and 0.77 for data sets of experimental HIV-1 protease and T4 lysozyme mutants, respectively. Additionally, a novel method is introduced for evaluating statistical significance associated with the number of correct test set predictions obtained from a trained model. Third, 100 stratified random splits of the protease and T4 lysozyme mutant data sets into training and test sets achieve 77.0% and 80.8% mean accuracy, respectively. Next, protease and T4 lysozyme models trained with experimental mutants are used to predict activity levels for all remaining mutants; a subsequent search for publications reporting on dozens of these test mutants reveals that experimental results are matched by 79% and 86% of predictions, respectively. Finally, learning curves for each mutant enzyme system indicate the influence of training set size on model performance.

Availability: Prediction databases at http://proteins.gmu.edu/automute/

Contact: ivaisman@gmu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Mutagenesis studies are one of the most widely used approaches for protein functional analysis. However, experimental mutagenesis techniques are expensive with regards to time, labor and cost. Computational efforts to infer protein function have utilized a variety of statistical and evolutionary methods (Bowers *et al.*, 2004; Dobson and Doig, 2005; Han *et al.*, 2005; Pazos and Sternberg, 2004; Sjolander, 2004; Tian *et al.*, 2004). Similarly, by analyzing sequence and structure information, success has been achieved in understanding the functional effects of coding non-synonymous single nucleotide polymorphisms (nsSNPs) (Chasman and Adams, 2001; Ng and Henikoff, 2001; Ramensky *et al.*, 2002; Saunders and Baker, 2002; Sunyaev *et al.*, 2001; Wang and Moult, 2001). A SNP is the result of a nucleotide variation at the same position in the genomic DNA of a given population, and a coding nsSNP leads to an amino acid substitution in the encoded protein sequence. Recent work has focused on using supervised learning techniques to predict the functionality of single site enzyme mutants resulting from coding nsSNPs (Karchin *et al.*, 2005; Krishnan and Westhead, 2003). Models developed with these latter methods are trained with a limited set of single site mutants, each belonging to one of a discrete set of activity classes based on experimental studies. Given that the models are expected to perform well at classifying mutants, they can subsequently be used to infer the activity classes to which the remaining uncharacterized enzyme mutants belong.

Supervised classification algorithms require that the mutants of an enzyme be represented as vectors of the same dimension with each vector component describing a particular attribute of the mutants. Model performance is significantly influenced by the strength of the signals embedded in these vectors, coupled with the degree of disparity of signals associated with differing classes. The attributes explored in the cited literature include information readily available from sequence data (e.g. physicochemical classes of wt and mutant residues, hydrophobicity difference and conservation score at the mutated residue position), and information directly predicted from protein structure (e.g. secondary structure, buried charge and solvent accessibility) (Karchin *et al.*, 2005; Krishnan and Westhead, 2003). Additionally, those studies place the training set nsSNP mutant enzymes into two classes (activity is either unaffected or affected by the mutation), regardless of the number of classes originally used by the investigators in their experimental studies.

Here we generalize the situation by including in our training and test sets all single site mutants of an enzyme, rather than focusing exclusively on mutants generated by coding nsSNPs.

*To whom correspondence should be addressed.

We also develop models both by using a two-class (active/inactive) labeling of the mutants, as well as by working with the larger number of mutant classes originally defined in the experimental studies. More importantly than these issues is the fact that each mutant attribute vector, derived using a four-body statistical potential function, consists of components that quantify the environmental changes from wt experienced at every residue position in the enzyme as a result of the specific amino acid substitution that generated the mutant (see Materials and Methods section). Hence, the dimensionality of every mutant attribute vector is equivalent to the number of amino acids in the primary sequence of the protein under consideration. The four-body statistical potential function is derived using the Delaunay tessellation of protein structures (Singh *et al*., 1996; Vaisman *et al*., 1998), and we refer to our mutant attribute vectors as *residual profiles*. Both sequence and structure information is embedded in the residual profiles of the mutants, and both contribute to the overall signal strength and interclass signal disparity. In particular, the non-zero components of a mutant residual profile correspond to all amino acid positions that participate in nearest-neighbor topological contacts with the mutated residue position, as well as the mutated position itself, based on the Delaunay tessellation of the protein structure. Additionally, the values at these non-zero components are a unique reflection of the type of residue replacement occurring at the mutated position.

The approach described above is applied to generating training and test sets for two enzyme systems of single site mutants: HIV-1 protease and bacteriophage T4 lysozyme. Since one monomer of the protease homodimer consists of 99 amino acids, there are $99 \times 19 = 1881$ possible single site mutants; similarly, the 164 amino acids forming the primary sequence of T4 lysozyme afford the possibility of $164 \times 19 = 3116$ mutants. Numerous mutagenesis experiments have been published analyzing both of these enzymes. However, the two most comprehensive studies experimentally measured activity levels for 536 mutants of HIV-1 protease (representing substitutions at all 99 positions), based on 336 published mutants (Loeb *et al*., 1989) and 200 additional mutants courtesy of R. Swanstrom, as well as 2015 mutants of T4 lysozyme obtained by introducing the same 13 residues as replacements at positions 2–164 and resulting in 104 additional non-mutant controls (Rennell *et al*., 1991). Residual profiles of mutants for which activity is known are used as a training set to build accurate inferential models for each enzyme, and these models are used to predict the activity levels of the remaining 1345 protease and 1101 T4 lysozyme mutants in their respective test sets based on the signals embedded in their residual profiles. According to the experimental measurements, the protease mutants each belong to one of three activity classes (positive, intermediate or negative), and the T4 lysozyme mutants each belong to one of four activity classes (high, medium, low or negative). Viewed as a two-class system, the protease mutants are either active (positive and intermediate classes combined) or inactive (negative class). Similarly, T4 lysozyme mutants in the high and medium classes are considered active, while mutants in the low and negative classes are inactive (Rennell *et al*., 1991). The supervised learning scheme used to illustrate our methodology is an implementation of Ross Quinlan's C4.5 decision tree algorithm (Quinlan, 1993) available with the Weka (Waikato Environment for Knowledge Analysis) suite of machine-learning tools (Frank *et al*., 2004).

2 MATERIALS AND METHODS

2.1 Delaunay tessellation and the four-body statistical potential

A training set of over 1300 high-resolution crystallographic protein structures with low primary sequence identity is selected from the Protein Data Bank (PDB) (Berman *et al*., 2000). Utilizing the PDB coordinates, each structure is represented as a discrete set of points in 3-dimensional space, corresponding to a weighted center of mass (CM) of the side-chain atoms of the constituent amino acid residues. Delaunay tessellation of each protein structure yields an aggregate of non-overlapping, space-filling, irregular tetrahedra (referred to as Delaunay simplices) whose vertices are the amino acid point representations (Singh *et al*., 1996; Vaisman *et al*., 1998). The Quickhull algorithm is used to perform the Delaunay tessellations (Barber *et al*., 1996), and an in-house suite of programs is used to perform pre- and post-processing as well as subsequent calculations and analyses.

Each simplex in a protein structure tessellation objectively defines a quadruplet of nearest-neighbor residues in the protein based on the identity of the four amino acids represented by the vertices of the simplex. Assuming order independence, theoretically there are 8855 distinct quadruplets that can be formed from the 20 amino acids naturally occurring in proteins (Singh *et al*., 1996; Vaisman *et al*., 1998). After individually tessellating each of the training set protein structures, the observed frequency of simplices representing each quadruplet type among the tessellations is calculated. The four-body statistical potential function is then defined as a set of log-likelihood scores that compares the observed normalized frequency to the expected chance of occurrence for every quadruplet (Singh *et al*., 1996; Vaisman *et al*., 1998). Specifically, the log-likelihood score for a quadruplet i, j, k, l of amino acids can be formulated as $q_{ijkl} = \log\,(f_{ijkl}/p_{ijkl})$. Here, f_{ijkl} is the observed normalized frequency of occurrence of simplices with vertices representing amino acids i, j, k, l among all the simplices formed by the tessellations of the training set proteins, and $p_{ijkl} = ca_ia_ja_ka_l$ is the expected rate of occurrence of the same quadruplet calculated from the multinomial distribution. In the formula for p_{ijkl}, a_r represents the normalized frequency of occurrence of the amino acid r among all of the training set proteins. If there are fewer than four distinct types of amino acids in the quadruplet, the formula will contain fewer than four a_r factors. Similarly, the number of factors in the denominator of the permutation factor $c = 4!/\Pi\ (t_r!)$ depends on the number of distinct residue types that form the quadruplet, where t_r represents the number of residues in the quadruplet that are of type r.

Given a CM point representation of a wt protein structure of interest, the total potential or topological score of the protein is calculated as the sum of the log-likelihood scores of all the simplices that form the Delaunay tessellation of the structure (Masso and Vaisman, 2003). The topological score of the same protein with a mutation is obtained by utilizing the tessellation of the wt structure while substituting only the amino acid label at the CM point representing the residue position of interest. This causes a change in the log-likelihood scores of all Delaunay simplices that use the point as a vertex, since one member of their respective quadruplets is mutated. Finally, the *residual score* of a mutant is defined as the difference between the mutant and wt protein topological scores (Masso *et al*., 2006). The Delaunay tessellations of HIV-1 protease and bacteriophage T4 lysozyme are based on the structural coordinates obtained from PDB accession files 3phv and 3lzm, respectively.

2.2 Residual profile vectors

The individual residue potential or residue environment score of each constituent amino acid position in a protein structure is obtained by summing the log-likelihood scores of only those simplices in the Delaunay tessellation for which the point representing the residue position participates as a vertex (Masso and Vaisman, 2003). Collectively, the vector of residue environment scores for all of the amino acids in a protein is referred to as the potential profile of the protein. A residue substitution at a single position in the protein, performed by relabeling the appropriate CM point in the tessellation of the wt structure as described in the previous section, only alters the environment scores of the mutated residue position as well as those residue positions whose respective points participate as vertices in simplices with the point representing the mutated residue position (Masso and Vaisman, 2003). The *residual profile* vector of a protein mutant is defined as the difference between the mutant and wt protein potential profile vectors, and the value of each component is referred to as an *environmental change (EC) score*. Hence, components with non-zero EC scores in the residual profile of a mutant identify the mutated position and all of its structural nearest-neighbors. In particular, the EC scores of the nearest-neighbor components signify the degree of environmental impact at those positions caused by the specific type of residue replacement at the mutated position, and the EC score of the component corresponding to the mutated position is identical to the residual score of the mutant.

2.3 Supervised learning with decision trees

Decision tree learning yields a classifier in the form of a rooted tree, such that a mutant is sorted down the tree by performing tests at each of the internal nodes. A decision is made at each internal node based on the EC score of a specific attribute (a component of the mutant residual profile vector), leading the mutant down a particular branch to the next node. Since the mutant attribute EC scores in this that are all numeric, the decisions that branch from an internal node are binary in nature and take the form $EC_N < a$ or $EC_N \geq a$, where N identifies the component number or position number in the primary sequence and a is a real number. The recursive process terminates once the mutant reaches a leaf node, where the mutant class is provided. A divide-and-conquer approach is employed during training, whereby at each stage starting from the root, an attribute is selected that best separates the classes (Witten and Frank, 2000). In order to avoid overfitting of the training data, which generally leads to poor model performance on independent test sets, the learned trees are typically pruned. We applied all of the default parameters associated with the Weka implementation of the C4.5 decision tree algorithm.

A stratified 10-fold cross-validation (10 CV) testing procedure is employed to generate ROC curves. The 10 CV approach entails a randomization of the original training set mutants into 10 equally sized subsets, and each subset is subsequently used as a test set after a decision tree classifier is trained with the remaining 9 mutant subsets combined. Stratification ensures that the activity class proportions in the original training set are maintained in each of the 10 subsets. In this way, an activity class prediction is obtained for every mutant in the original training set. With generic class labels P and N, a comparison of the actual and predicted classes for each mutant based on the outcome of 10 CV yields a summary 2×2 confusion matrix tabulating the number of correct predictions (TP, TN) and misclassifications (FN, FP). The following can also be computed: accuracy $= (TP + TN)/(TP + FN + TN + FP)$, true positive rate or $TPR = TP/(TP + FN) =$ sensitivity, and false positive rate or $FPR = FP/(FP + TN) = 1 -$ specificity. It is clear how sensitivity and specificity may also be evaluated for the negative class. An ROC curve is a plot of TPR versus FPR in the unit square joining (0,0) to (1,1), and each 10 CV run yields one point on the ROC curve. A default cost of 1 is assigned to each FN and FP misclassification during training; however, varying the FN/FP cost ratio prior to each 10 CV run generates multiple (FPR,TPR) points that span the ROC curve.

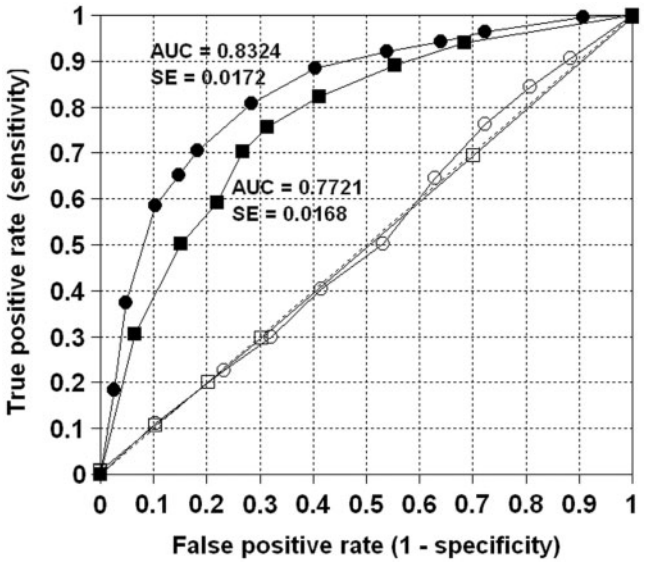

Fig. 1. ROC curves, and the associated AUC and standard error (SE) values, generated using decision tree learning in conjunction with residual profile data sets of 536 HIV-1 protease mutants (solid circles) and 2015 T4 lysozyme mutants (solid squares). Every mutant is labeled as either active or inactive for the purpose of obtaining these two-class ROC curves. Each point on the ROC curves is obtained via a 10 CV procedure using a specific ratio of misclassification costs. Control ROC curves (open circles and squares) that are also shown are based upon a random shuffling of the class labels among the mutants in each data set.

3 RESULTS AND DISCUSSION

3.1 ROC assessment of performance

Beginning with a two-class labeling of mutant activity, ROC curves were generated for the HIV-1 protease and T4 lysozyme mutant training sets by applying a 10 CV procedure (Fig. 1). Also displayed in Figure 1 are the AUC and standard error (SE) values for their respective ROC curves. The AUC is equivalent to the non-parametric Wilcoxon–Mann–Whitney test of ranks and provides a measure of performance that is insensitive to the distribution of the activity classes in test sets (Fawcett, 2003). In the extreme cases, AUC = 1.0 reveals perfect classification based on a model learned with the training set, while AUC = 0.5 indicates random guessing. We also utilized a conservative estimate for the SE of an ROC curve (Hanley and McNeil, 1982). With model performance of 0.8324 ± 0.0172 (AUC $\pm$ SE) for the HIV-1 protease system, and 0.7721 ± 0.0168 for the T4 lysozyme system, our residual profile vector representations of active and inactive mutants clearly encode significantly disparate signals exploitable by supervised classification. This observation is underscored by the fact that a model learned following a random shuffling of the class labels among the training set mutants is expected to perform no better

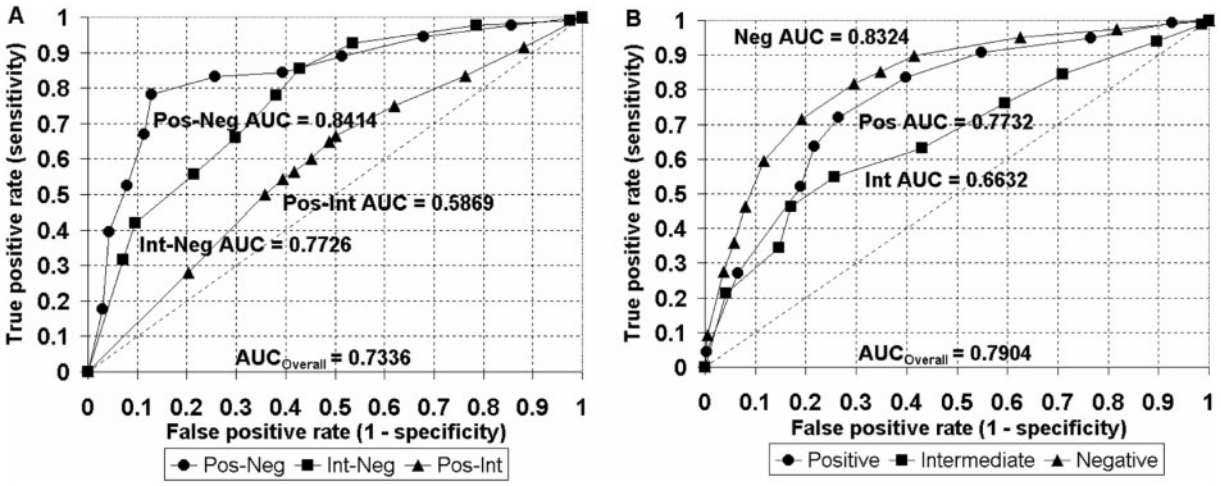

Fig. 2. ROC curves generated with 536 HIV-1 protease mutants by applying (**A**) the 1-against-1 and (**B**) the 1-against-all methods for handling multiple classes with decision trees. In the legend for (A), Pos-Neg refers to the subset consisting of mutants experimentally determined to belong only to the positive or negative pair of activity classes (intermediate mutants removed from the full set of 536 mutants); the remaining legend entries are similarly defined. The class pair AUC values reflect intuitive ideas based on biological principles, whereby the signals embedded in the residual profiles of the positive and negative class mutants are the most disparate, complementing the great structural and functional differences between these mutants. On the other hand, positive and intermediate mutants are more or less active, and their overlapping signals pose a challenge for accurate discrimination between these classes. In the legend for (B), Positive refers to the full set of 536 mutants, where positive is the reference class, and intermediate and negative mutants are combined into a single 'non-positive' class; the remaining legend entries are similarly defined. Note that the AUC value for the positive reference class ROC curve falls between the AUC values for the positive–negative and positive–intermediate class pair ROC curves; the same is true for the other reference classes.

than random guessing (Fig. 1; HIV-1 protease control AUC = 0.5059, and T4 lysozyme control AUC = 0.4967).

A more challenging classification problem entails use of the original experimental activity classes for the HIV-1 protease and T4 lysozyme mutants. We applied two methods for handling training sets consisting of mutants belonging to more than two activity classes. With the *1-against-1* approach, truncated subsets are created from the full mutant training set, each of which contains only residual profiles of mutants belonging to a pair of activity classes. The truncated subsets serve as training sets, from which two-class ROC curves are generated for every class pair. The overall AUC measure for the multi-class problem using this technique is defined as the average of the AUC values over all class pair ROC curves, expressed as

$$\text{Overall AUC} = \frac{2}{n(n-1)} \sum_{\{c_i,c_j\} \in C} \text{AUC}(c_i,c_j)$$

where n is the total number of mutant activity classes in the full training set, and AUC(c_i, c_j) is the area under the ROC curve generated by using a subset consisting of all the mutants belonging to only classes c_i and c_j (Fawcett, 2003; Hand and Till, 2001). Application of the *1-against-all* method utilizes the full training set by alternatively relabeling the mutant activity classes. In order to create the new training sets, a reference class is selected in the original training set, and all mutants not belonging to this class are relabeled as non-reference class mutants. A two-class ROC curve is generated for the reference class chosen, and the process is repeated so that each class serves as a reference. The overall AUC measure for the multi-class problem here is defined as a *weighted* average of the AUC values over all reference class ROC curves, where the weight on each AUC is the proportion of mutants in the full training set that belong to the reference class. Symbolically,

$$\text{Overall AUC} = \sum_{c_i \in C} \text{AUC}(c_i) \cdot p(c_i)$$

where AUC(c_i) is the area under the two-class ROC curve generated by using the class c_i as the reference class, and $p(c_i)$ is the proportion of class c_i mutants in the full training set (Fawcett, 2003; Provost and Domingos, 2001). For the HIV-1 protease (Fig. 2) and T4 lysozyme (Fig. 3) mutant systems, we plotted the ROC curves based on the 1-against-1 and 1-against-all methods, along with their respective AUC values as well as the Overall AUC measures. The AUC values mirror biological notions regarding the increasing structural and functional disparities between inactive mutants and those displaying higher levels of activity.

Similar ROC analyses based on application of support vector machine and neural network supervised classification machine-learning approaches yielded results analogous to those obtained by using the decision tree algorithm on the data sets of experimental HIV-1 protease (Fig. S5) and T4 lysozyme (Fig. S6) mutants. Although all three algorithms perform well on the two sets of mutant residual profile feature vectors, the best AUC values were generally obtained with the use of decision trees. Given the consistency of the results among a variety of techniques, it is clear that the observed performance is due primarily to the significance of the mutant residual profiles, with choice of algorithm providing an additional

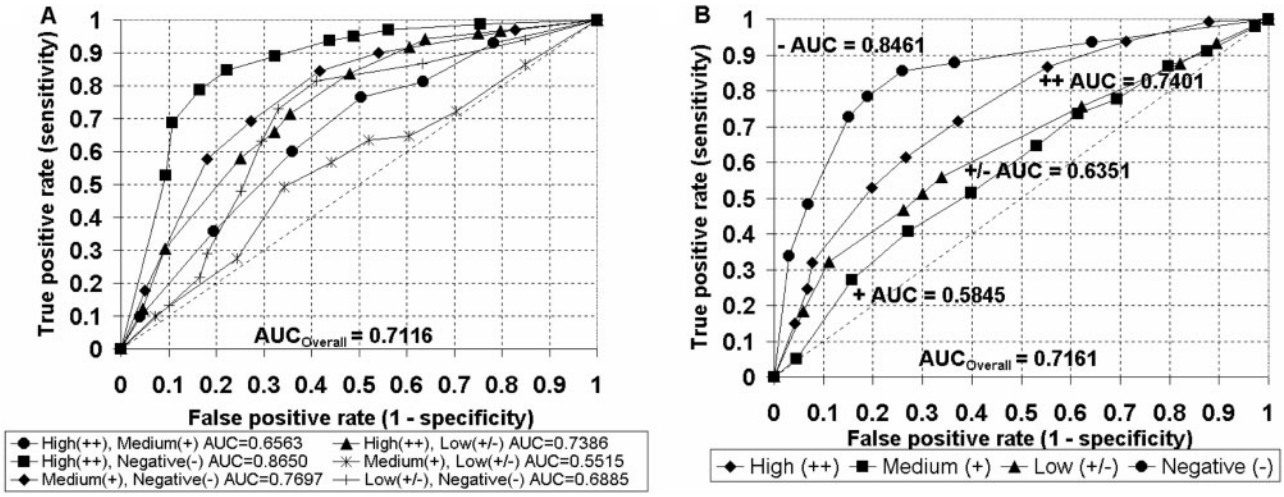

Fig. 3. ROC curves generated with 2015 T4 lysozyme mutants by applying (**A**) the 1-against-1 and (**B**) the 1-against-all methods for handling multiple classes with decision trees. Mutant activity pairs can be listed in order of decreasing AUC magnitude as high-negative, medium-negative, high-low, low-negative, high-medium and medium-low. The most divergent signals here are found between the two most active classes and the negative class. The order of the remaining pairs of mutants reflects issues concerning the mutant activity measurement technique and the experimental conditions. Mutant activity was measured qualitatively by visually inspecting the size of plaques formed on Petri dishes, and it was difficult for the investigators to assess and differentiate between medium and low activity mutants due to small plaque sizes, hazy plaque morphologies and other impediments (Rennell *et al*., 1991). As a result, their statistical analysis was based on the use of two classes: active (high and medium classes combined) and inactive (low and negative classes combined). This supports the observation that the signals associated with the high-low pair of classes are more disparate than those of the low-negative pair. Intuition suggests that the negative (no plaque formation) class contains the fewest experimental errors, followed by the high (large plaque) class due to possible mislabeling between the high and medium class mutants. This helps to explain why the low-negative class pair signals are more distinguishable from each other compared to those of the high-medium class pair. Finally, mutants belonging to the medium and low classes face the highest risk of having been mislabeled, yielding noisy signals and causing this class pair to have the lowest AUC value.

marginal benefit. Additionally, expanding each feature vector to include components that identify the wt and replacement amino acids of each mutant, as well as the residue position number, slightly enhanced the performance.

3.2 Novel methodology for computing the statistical significance of model accuracy

Using model accuracy (% correct) as the performance measure on a test set, we next illustrate an approach for computing the statistical significance associated with the number of correct predictions made by a model. A stratified sample of 100 HIV-1 protease mutants was randomly selected for testing, and a decision tree model was trained with the remaining 436 mutants. The breakdown of protease mutants consists of (121 positive, 66 intermediate, 249 negative) in the training set and (19, 18, 63) in the test set. Likewise, a 1815/200 split of the 2015 T4 lysozyme mutants resulted in (1277 high, 275 medium, 135 low, 128 negative) in the training set and (144, 28, 17, 11) in the test set. Based on an initial two-class (active/inactive) labeling of the mutants as described in the introduction and applying default misclassification costs, the trained HIV-1 protease and T4 lysozyme models correctly predicted the class memberships for 74/100 and 174/200 of the respective test set mutants. In order to assess the statistical significance of correctly predicting 74% of the protease test mutants, let $X = X_1 + X_2 + \ldots + X_{100}$, where each X_i is a Bernoulli random variable representing the outcome of a test set mutant activity prediction. The expected number of correct predictions on the test set by chance is

$$\mu = E(X) = 37 \times 187/436 + 63 \times 249/436 = 52,$$

and

$$\sigma^2 = \text{Variance}(X) = 100 \times (187/436) \times (249/436) = 24.5.$$

So $\sigma = 4.95$, and the P-value associated with this result is

$$P(X > 74; \mu = 52) = P\left(\frac{X-\mu}{\sigma} > \frac{74-52}{4.95}\right) = P(z > 4.44)$$
$$\approx 1 - \Phi(4.44) = 4.42 \times 10^{-6}$$

where Φ represents the cumulative distribution function for a standardized normal variable. Similar calculations for T4 lysozyme reveal that $\mu = 151$ and $\sigma = 4.98$, yielding a Z-score of 4.62 and a P-value of 1.94×10^{-6} for the 87% accuracy rate. These results are also significant when considering multi-class predictions based on 1-against-1 and 1-against-all decision tree models trained using the original mutant class labels (Table 1).

Unlike the AUC measure, the prediction accuracy used above is sensitive to class skew in test sets. For a test set consisting of a highly unbalanced pair of classes, a simple model that always predicts the majority class may yield high accuracy yet always misclassify the minority class. Hence for a stratified random split of the original data set into training and test sets, while the use of default costs is appropriate for learning a decision tree model that yields optimal sensitivity

Table 1. Summary of expected and actual numbers of correct test set mutant predictions and associated levels of statistical significance using default cost decision tree classifiers based on splits of (A) the HIV-1 protease mutants and (B) the T4 lysozyme mutants

A (436/100 split)	2 classes	3 classes (1-against-all)	3 classes (1-against-1)
Expected correct (std. dev.)	52 (4.95)	44 (4.64)	44 (4.64)
Actual correct	74	69	66
P-value	$P = 4.42 \times 10^{-6}$	$P = 3.57 \times 10^{-8}$	$P = 1.06 \times 10^{-6}$
B (1815/200 split)	**2 classes**	**4 classes (1-against-all)**	**4 classes (1-against-1)**
Expected correct (std. dev.)	151 (4.98)	107.6 (5.96)	107.6 (5.96)
Actual correct	174	134	145
P-value	$P = 1.94 \times 10^{-6}$	$P = 4.74 \times 10^{-6}$	$P = 1.76 \times 10^{-10}$

and specificity for the test set active and inactive classes of HIV-1 protease mutants (43% active and 57% inactive in the original data set), application of an increased cost for minority class misclassifications is required for a similarly optimized model for the highly skewed mutants of T4 lysozyme (86% active and 14% inactive in the original data set). In particular, assigning a cost to inactive T4 lysozyme mutant misclassifications that is 10 times higher than that of active mutants led to an appropriately trained model that performed well on both classes in the test set. Such cost-sensitive learning is equivalent to increasing the number inactive mutants in the training set by a factor of 10. We utilized this choice of costs during training to learn T4 lysozyme models and assess the statistical significance of their predictive accuracy, splitting the mutant data as before (1815 training/200 test mutants) while repeating the procedure over 100 stratified random splits. The mean prediction accuracy was 80.8%, with a standard deviation (std. dev.) of 2.6%, and the number of correct predictions obtained for each of the 100 models was highly statistically significant (minimum Z-score = 9.97, with corresponding P-value = 0). Using default costs and performing a similar analysis on the HIV-1 protease mutants (100 stratified random splits, 436 testing/100 test mutants) resulted in mean accuracy = 77.0% with std. dev. = 3.7%, where the minimum Z-score was 3.42 and corresponded to a P-value of 3.14×10^{-4}.

3.3 Predictive capacity of trained models: comparing predictions with experimental data

Next, we gauged the practical utility in using decision tree models trained with residual profile vectors for making functional predictions about protein mutants. Based on a two-class labeling of the set of 536 HIV-1 protease mutants, predictions were generated for the remaining 1345 protease mutants by including their residual profiles as attribute vectors in a test set. A thorough search of the protease literature identified experimental activity measurements for 47 single site mutants that were among those in the test set (Table S2). Based on a comparison of the experimental and predicted activity, we observed a match for 37 out of the 47 mutants (79%). Similarly, a cost-sensitive two-class decision tree model trained with the set of 2015 T4 lysozyme mutants was used to predict the class memberships of the remaining 1101 test set mutants. A search of the ProTherm database (Bava *et al*., 2004) identified experimental activity measurements for 35 T4 lysozyme mutants that were among those in the test set (Table S3). Comparison of the experimental and predicted activity of these mutants led to a match for 30 out of the 35 mutants (86%). To compare the performance of our algorithm with other prediction methods, we ran a series of tests using these methods and our data set. In all cases our models performed better, the results of these comparisons are summarized in Table S4.

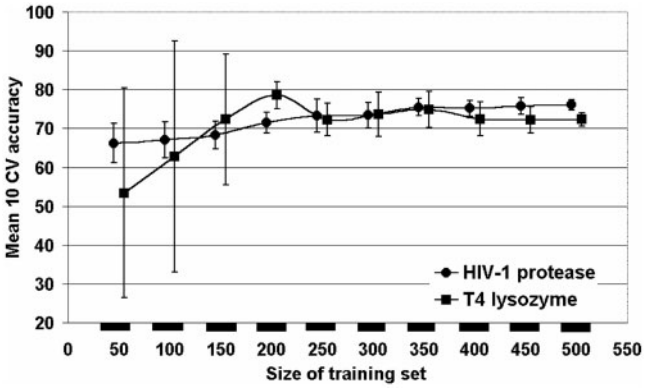

Fig. 4. Learning Curves for HIV-1 protease and T4 lysozyme mutants. The mean accuracy over 10 runs of 10 CV is used to evaluate decision tree classifier performance for each training set size. Error bars reflect ± 1 std. dev. from the mean.

3.4 Learning curves

Given the effort required to synthesize and analyze enzyme mutants, an important consideration for the wet-lab biologist is the number of training set mutants necessary to develop an accurate inferential model. We prepared learning curves using the two-class labeling of mutants in order to better understand the effect that training set size has on the accuracy of decision tree models trained using residual profile vectors. Training sets of increasing size were obtained through a stratified random sampling with replacement from each set of enzyme mutants, beginning with 50 mutants and incrementing by the same amount. For each training set, a 10 CV procedure was implemented ten times, and the reported accuracy rates were averaged. The plots (Fig. 4) suggest that approximately 300 mutants are adequate for use as training sets in order to develop decision tree classifiers displaying optimal accuracy. Ideally, two mutants at each residue position in T4 lysozyme should be chosen for the reduced training set (one from each activity class, if available), and three mutants at each position in HIV-1 protease should be selected.

4 CONCLUSIONS

It is encouraging to note that among the 26 residue positions in HIV-1 protease used for decision making at the nodes of the

two-class tree trained with all 536 experimental mutants, 14 of these positions are located within the most highly conserved regions (Loeb *et al*., 1989). Additionally, 7 of the 26 positions are known to have the ability to undergo mutations leading to HIV-1 protease inhibitor resistance (Kantor *et al*., 2004), ensuring that structurally and functionally important residues are well represented in the tree. An important consequence of such a tree, e.g. is its decision to predict as inactive all single residue substitutions within the D25-T26-G27 catalytic triad that were not part of the training set. Similarly, positions at the nodes of the two-class tree trained with the 2015 experimental mutants of T4 lysozyme include those important for stability and function. By combining the data for all experimentally known and predicted single point mutants, a protein mutational array can be constructed and analyzed, as shown in Figure S7 for the case of T4 lysozyme. The mutational array reveals an increased number of inactive predictions for point mutations that occur at the catalytic (E11, D20) and substrate-binding (G30, S117) sites and at nearby positions, complementing the existing experimental data. Notably, with the exception of N132C, all experimentally known mutants at this additional substrate-binding site in T4 lysozyme are active. Mirroring this fact, our model predicts as active all remaining mutants at N132, with the exception of two substitutions (M, V) that are similar to C in their degree of hydrophobicity.

Our current work focuses on trying to improve model performance and predictive capability by utilizing other supervised classification algorithms, and by including additional components to the mutant residual profiles that incorporate the sequence and structure information described in the previously mentioned reports (Karchin *et al*., 2005; Krishnan and Westhead, 2003). However, the key ingredient that will permit our method for mutant functional inference to enjoy a practical utility among experimentalists is the development of a standardized way for identifying a minimal training set of protein mutants. Such a tool will provide researchers interested in a comprehensive exploration of the single point mutation space with an idea of the preliminary work required to reap the benefits of reliable predictions for the remaining mutants.

ACKNOWLEDGEMENTS

We would like to thank Ron Swanstrom for providing unpublished data on HIV-1 protease mutant activity levels, Todd Taylor for help with the four-body statistical potential and Jitendra Ganju for fruitful discussions.

Conflict of Interest: none declared.

REFERENCES

Barber,C.B. *et al*. (1996) The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw*., **22**, 469–483.

Bava,K.A. *et al*. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res*., **32**, D120–D121.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Bowers,P.M. *et al*. (2004) Use of logic relationships to decipher protein network organization. *Science*, **306**, 2246–2249.

Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms. *J. Mol. Biol*., **307**, 683–706.

Dobson,P.D. and Doig,A.J. (2005) Predicting enzyme class from protein structure without alignments. *J. Mol. Biol*., **345**, 187–199.

Fawcett,T. (2003) ROC graphs: notes and practical considerations for researchers. *Technical report HPL-2003-4*. Hewlett-Packard Labs, Palo Alto..

Frank,E. *et al*. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Han,L.Y. *et al*. (2005) Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology*, **331**, 136–143.

Hand,D.J. and Till,R.J. (2001) A simple generalization of the area under the ROC curve to multiple class classification problems. *Mach. Learn*., **45**, 171–186.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Kantor,R. *et al*. (2004) Evolution of resistance to drugs in HIV-1-infected patients failing antiretroviral therapy. *AIDS*, **18**, 1503–1511.

Karchin,R. *et al*. (2005) Improving functional annotation of non-synonomous SNPs with information theory. *Pac. Symp. Biocomput*., 397–408.

Krishnan,V.G. and Westhead,D.R. (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.

Loeb,D.D. *et al*. (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

Masso,M. and Vaisman,I.I. (2003) Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem. Biophys. Res. Commun*., **305**, 322–326.

Masso,M. *et al*. (2006) Computational mutagenesis studies of protein structure-function correlations. *Proteins*, **64**, 234–245.

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*., **11**, 863–874.

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.

Provost,F. and Domingos,P. (2001) Well-trained PETs: improving probability estimation trees. In: *CeDER Technical report IS-00-04*. Stern School of Business, New York University, New York.

Quinlan,R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, San Mateo, CA.

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.

Rennell,D. *et al*. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol*., **222**, 67–88.

Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol*., **322**, 891–901.

Singh,R.K. *et al*. (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J. Comput. Biol*., **3**, 213–221.

Sjolander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.

Sunyaev,S. *et al*. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet*., **10**, 591–597.

Tian,W. *et al*. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res*., **32**, 6226–6239.

Vaisman,I.I. *et al*. (1998) Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. *Proc. IEEE Symp. Intell. Syst*., 163–168.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat*., **17**, 263–270.

Witten,I.H. and Frank,E. (2000) *Data Mining*. Morgan Kaufmann, San Francisco.