

Automated Functional Inference of Enzyme Mutants Utilizing a Four-Body Statistical Potential

Majid Masso¹, Iosif I. Vaisman²

Keywords: Delaunay tessellation; protein mutagenesis; functional inference; supervised classification; receiver operating characteristics

Introduction

An important aspect of biochemical and biomedical research focuses on characterization of enzyme mutants. We describe a supervised learning approach for inferring activity levels of enzyme mutants generated by single residue substitutions throughout the primary sequence, given a training set consisting of a subset of these mutants with known activity. Each training set mutant belongs to one of a discrete number of activity classes. Models are trained using support vector machines (SVM), decision trees (DT), and neural networks (NN), and model performance is assessed via a sensitivity analysis that utilizes receiver operating characteristic (ROC) curves and the area under these curves (AUC). Measures of statistical significance for the number of correct predictions are also calculated.

Attribute vectors

The learning schemes recognize all mutants of an enzyme as vectors of the same dimension, such that the components form an ordered set of measurable attributes for each mutant. With the application of a four-body statistical potential based on Delaunay tessellation of protein structure [5], we have derived mutant attribute vectors (referred to as *residual profiles*) each with dimension equal to the length of the primary sequence of the enzyme. Specifically, the vector components of a mutant residual profile quantify the environmental changes from wt experienced at the corresponding residue positions in the enzyme due to the single point substitution that generated the mutant. We demonstrate that significant and sufficiently divergent signals are present in the residual profiles of mutants belonging to differing activity classes.

Experimental data

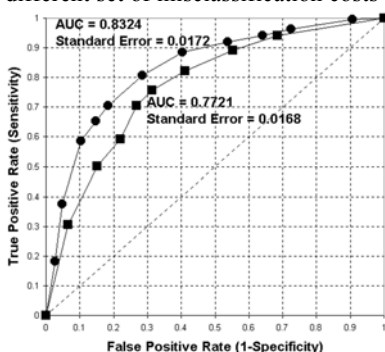
Our methodology is applied to HIV-1 protease and bacteriophage T4 lysozyme systems of enzyme mutants. Models are trained for the protease system by using a set of 536 single point mutants (out of 1881 total) each belonging to one of three activity classes (positive, intermediate, or negative) [3]; by labeling positive and intermediate mutants as active and negative mutants as inactive, two-class models are also investigated. For the T4 lysozyme system, the training set used to generate the models includes 2015 single point mutants (out of 3116 total) each belonging to one of four activity classes (high, medium, low, or negative) [4]. By following the analytical approach used by the researchers that experimentally obtained the T4 lysozyme data, we also develop two-class models by labeling high and medium mutants as active, and low and negative mutants as inactive.

¹ Bioinformatics and Computational Biology, School of Computational Sciences, George Mason University, 10900 University Blvd., MSN 5B3, Manassas, VA 20110. E-mail: mmasso@gmu.edu

² Bioinformatics and Computational Biology, School of Computational Sciences, George Mason University, 10900 University Blvd., MSN 5B3, Manassas, VA 20110. E-mail: ivaisman@gmu.edu

Results

The WEKA suite of machine learning tools is used to implement the supervised learning schemes [1]. ROC curves for both mutant systems, based on using two activity classes and decision tree learning, are shown in Figure 1. Each point reflects results obtained for models generated by using a different set of misclassification costs and applying tenfold cross-validation.



	Pos (1-against-1)	Int (1-against-1)	Neg (1-against-1)	Others Combined (1-against-all)
Pos	---	0.6522 (SVM) 0.5869 (DT) 0.6225 (NN)	0.8182 (SVM) 0.8414 (DT) 0.7877 (NN)	0.7389 (SVM) 0.7732 (DT) 0.7282 (NN)
Int		---	0.7558 (SVM) 0.7726 (DT) 0.7511 (NN)	0.6731 (SVM) 0.6632 (DT) 0.6814 (NN)
Neg			---	0.7810 (SVM) 0.8324 (DT) 0.7764 (NN)

Figure 1. ROC curves for HIV-1 protease mutants (circles) and T4 lysozyme mutants (squares) using two activity class labels (active vs. inactive) for each mutant system.

Table 1. Summary pairwise AUC values for HIV-1 protease ROC curves. The data suggest that signals from Pos and Neg mutants are most disparate, followed by Int and Neg mutants. Signals from Pos and Int mutants are difficult to distinguish.

Two approaches (1-against-all and 1-against-1) are used for obtaining overall AUC when the labels include more than two activity classes. The 1-against-all method yields a separate model for each class, where the reference class mutants form Class 1, and all the remaining mutants are pooled together to form Class 2. Overall AUC here is defined as a *weighted* average of the separate two-class AUCs based on the reference class frequency in the data set. With 1-against-1, the full training set is used to create multiple subsets, each containing mutants belonging to a pair of classes; individual two-class models are generated with each subset. Overall AUC here is the average of the activity class pair AUCs. All pairwise AUC values obtained in the process of performing the three-class analyses for HIV-1 protease are summarized in Table 1. Similar results have been obtained for T4 lysozyme.

Using two- and multi-activity class decision tree models generated with all of the training data described for each mutant system, predictions have been made regarding the activity classes of the uncharacterized 1345 protease mutants and 1101 T4 lysozyme mutants by preparing a test set containing the residual profiles of these mutants [2].

References

- [1] Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten I.H. 2004. Data mining in bioinformatics using WEKA. *Bioinformatics* **20**:2479-2481.
- [2] http://binf.gmu.edu/mmasso/protease_prediction.html; http://binf.gmu.edu/mmasso/lysozyme_prediction.html.
- [3] Loeb, D.D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S.E. and Hutchison III, C.A. 1989. Complete mutagenesis of the HIV-1 protease. *Nature* **340**:397-400. Also R. Swanstrom, personal communication.
- [4] Rennell, D., Bouvier, S.E., Hardy, L.W. and Poteete, A.R. 1991. Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology* **222**:67-88.
- [5] Singh, R.K., Tropsha, A. and Vaisman, I.I. 1996. Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues. *Journal of Computational Biology* **3**:213-222.