33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Structure-Based Prediction of Protein Activity Changes: Assessing the Impact of Single Residue Replacements

Majid Masso, *Member, IEEE*, and Iosif I. Vaisman

***Abstract*— A computational mutagenesis methodology founded upon a structure-dependent and knowledge-based four-body statistical potential is utilized in generating feature vectors that characterize over 8500 individual amino acid substitutions occurring in seven proteins, each mutant having been experimentally ascertained for its relative effect on native protein activity. The proteins are diverse with respect to host organism (viral, bacterial, human) and function (enzymatic, nucleic acid binding, signaling), the structures span all four major SCOP classifications, and the mutations occur at positions well distributed throughout the seven structures. Implementation of the random forest algorithm, for classifying mutant activity as either unaffected or affected relative to the native protein, yields 84% accuracy based on tenfold cross-validation. A freely available online server for obtaining predictions with the trained model, which also displays 84% accuracy on an independent test set of mutants, is available at http://proteins.gmu.edu/automute/AUTO-MUTE_Activity.html**

I. Introduction

THROUGH the introduction of single residue replacements, protein engineering experiments aim to modify an existing property, such as level of activity or degree of structural stability [1]. These new proteins may be employed directly in commercial applications, or they may serve to guide further research on such topics as protein function, the protein folding process, and disease association studies. Given the time and cost constraints associated with exhaustive traversal of the protein mutational landscape, availability of reliable predictions from computational models assists in prioritizing experiments. Ideally, a model is trained using a uniformly distributed sample of amino acid substitutions drawn from the population, whose effects with respect to a particular property (e.g., activity or stability changes relative to the native proteins) are already known based on previously published experimental data.

Each state-of-the-art predictive model of protein stability or functional change upon mutation described in the literature uniquely incorporates sequence, structure, and/or evolutionary information [2]. Recently, we implemented supervised classification and regression statistical machine learning techniques for developing accurate predictive models of relative stability changes upon mutation [3, 4]. Such models learn complex nonlinear functions based on training sets of mutants represented quantitatively as feature vectors that consist of components (i.e., input attributes or predictors of mutant effect) encoding information obtained from a computational mutagenesis technique we developed, with mutant relative stability change denoting the output attribute. Our *in silico* mutagenesis approach relies on a knowledge-based four-body statistical potential and utilizes Delaunay tessellation, a computational geometry tiling algorithm, to objectively identify quadruplets of neighboring residues in protein structures. For each single residue replacement in a protein, the procedure empirically quantifies environmental perturbations at the mutated residue position and at all tessellation-based structurally proximal positions that define its local neighborhood.

Here we implement the random forest (RF) supervised classification algorithm [5] to develop a predictive model of relative activity changes upon mutation. The model is trained on 8561 protein mutants, 5251 unaffected (U) and 3310 affected (A), retrieved from published data regarding the functional impact of introducing single residue substitutions into each of seven diverse proteins: HIV-1 protease (PR) [6, 7] and reverse transcriptase (RT) [8], bacteriophage T4 lysozyme (lys) [9], bacteriophage f1 gene V protein (GVP) [10], the *E. coli* proteins barnase (barn) [11] and lac repressor (lac) [12], and human interleukin-3 (IL-3) [13]. By generating mutant feature vectors in a manner identical to that outlined in the previous paragraph, our RF model achieves 84% accuracy as assessed both through tenfold cross-validation as well as by predictions on an independent test set of mutants. All datasets can be downloaded from http://proteins.gmu.edu/automute/AUTO-MUTE_Activity_Details.html.

II. Materials and Methods

A. Computational Mutagenesis

An outline of the procedure is provided below, and additional details with illustrative figures are available from Masso and Vaisman [3, 4]. Given a set of points $P = \{x_1, x_2, x_3, \ldots, x_N\}$ in 3D Euclidean space corresponding to the Cα coordinates of all constituent amino acid residues in a protein structure, the Delaunay tessellation algorithm generates a convex hull of space-filling, non-overlapping, irregular tetrahedra whose combined vertices coincide with all the elements of P. Each tetrahedron objectively

Manuscript received March 25, 2011.

Majid Masso is with the Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, Manassas, VA 20110 USA (phone: 703-257-5756; fax: 703-993-8976; e-mail: mmasso@gmu.edu).

Iosif I. Vaisman is with the Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, Manassas, VA 20110 USA (e-mail: ivaisman@gmu.edu).

identifies, through its four Cα vertices, a quadruplet of nearest neighbor residues in the protein; however, given that any two adjacent tetrahedral tiles in the tessellation may share a single vertex, a linear edge (two vertices in common), or a triangular face (three vertices in common), an amino acid in the protein generally participates in more than one such quadruplet of nearest neighbor residues. The *local structural neighborhood* of an amino acid consists of itself as well as all nearest neighbors defined by the residue quadruplets in which it participates (i.e., all residues with which the amino acid shares a tessellation edge). To ensure only biochemically feasible quadruplet interactions, all protein structure tessellations are modified by the removal of edges longer than 12 angstroms.

A total of 8855 distinct unordered 4-letter subsets (i.e., permutations excluded) can be enumerated by selecting with replacement from the standard 20-letter protein alphabet. In order to reliably calculate the observed relative frequency of occurrence for each quadruplet in protein structure space, a diverse dataset of 1417 high-resolution proteins with low sequence and structure similarity is selected for tessellation from the Protein Data Bank (PDB) [14]. Next, a multinomial distribution is used to obtain the rate expected by chance for each quadruplet. By applying the inverse Boltzmann principle [15], a knowledge-based four-body statistical potential is generated by calculating the logarithm of the ratio of observed to expected rates (i.e., a log-likelihood score) for each quadruplet.

Given the tessellation of any protein structure, the four-body statistical potential equips every constituent tetrahedral simplex with a score equivalent to that of the quadruplet of residues identified at its four vertices. Consequently for each amino acid position in the protein, we calculate a *residue environment score* by adding together scores of all tetrahedra that share as a vertex the Cα coordinate of the amino acid. Altering the amino acid identity at a particular vertex in the tessellation changes the scores of precisely all those tetrahedra that share the vertex. In turn, changes to residue environment scores also occur, specifically at all amino acid positions forming the local structural neighborhood of the mutated position, and subtracting the original residue environment scores from the new ones at these positions yields their *environmental change (EC) scores*. In particular, the EC score at the mutated position itself is termed the *residual score* of a mutant, a scalar that empirically quantifies overall change to protein sequence-structure compatibility upon mutation.

B. Dataset and Feature Vectors

For each of the seven proteins whose single residue replacements constitute our dataset, the respective numbers of mutants that experimentally display unaffected (U) and affected (A) activity levels relative to wild type are listed in Table I. Additionally, Table I provides PDB accession codes (structural coordinate files), SCOP structural classifications [16], and organism sources and biological functions of these proteins, which collectively reflect a well-distributed dataset.

TABLE I
DATASET CHARACTERISTICS

Protein	Source	Function	Mutant Data	PDB Code	SCOP Class
PR	HIV-1	proteinase	U: 218 A: 294	3phvA	all β
RT	HIV-1	transferase	U: 170 A: 196	1rtjA	α / β
lys	phage T4	hydrolase	U: 1364 A: 638	3lzmA	α + β
GVP	phage f1	DNA binding (replication)	U: 130 A: 221	1gvpA	all β
barn	*E. coli*	RNase	U: 643 A: 34	1bniA	α + β
lac	*E. coli*	DNA binding (regulation)	U: 2256 A: 1773	1efaB	all α
IL-3	human	signaling (growth factor)	U: 395 A: 229	1jliA	all α

Local structural neighborhoods are identified by tessellation of the protein structure and vary in size by residue position, yet each includes the mutated position along with no fewer than six nearest neighbors. For each single residue replacement in the dataset, the four-body statistical potential is used in conjunction with the computational mutagenesis technique from the previous section to calculate EC scores at the mutated position (i.e., the residual score) and at all of its local structural neighbors.

With a focus on common attributes among all mutants, the following are encoded into each mutant feature vector [3]: identities of the native and replacement amino acids at the mutated position; residual score; EC scores at the six local neighborhood positions that are closest in Euclidean distance to the mutated position (lengths of tessellation edges between vertices), ordered by neighbor proximity to the mutated position; ordered residue identities at the six closest positions; ordered differences in primary sequence numbers between each of the six closest local neighbors and the mutated position; mean volume and mean tetrahedrality of all tetrahedral simplices in the tessellation that utilize the mutated position as a vertex; tessellation defined depth (S, surface; U, undersurface; B, buried) at the mutated position; number of surface positions that share a tessellation edge with the mutated position; and secondary structure at the mutated position (H, helix; S, strand; T, turn; C, coil).

C. Random Forest Classification and Model Performance

We use the random forest (RF) algorithm as implemented in the Weka software [17]. Multiple bootstrap datasets are generated by the RF algorithm, each obtained by selecting mutants one at a time and with replacement from the original training set. The RF algorithm employs a technique referred to as *bagging* (bootstrap aggregating), whereby each bootstrap dataset is used to train an unpruned classification tree, and final mutant predictions are obtained from the ensemble of trees via majority vote [5]. Additionally, a small fixed size subset of the feature vector components is randomly selected to split at every node encountered in each of the growing trees, where subset size is a function of feature vector length. These combined properties allow the RF algorithm to generally perform better than most other supervised classification methods, and regardless of the number of trees in the forest, the algorithm does not suffer from overfitting [5]. The RF parameter values we use in this study include a forest of 100 trees as well as subsets of 5 random features selected for splitting at each tree node.

The expected effectiveness of the RF model is primarily based on the evaluation and analysis of mutant predictions obtained via stratified tenfold cross-validation (10-fold CV) testing, though the trained RF model is also ultimately evaluated on an independent test set of mutants. Prediction performance is measured by computing the following quantities. Letting P and N respectively refer to the unaffected (U) and affected (A) classes of mutants,

ACC = accuracy = (TP + TN) / (TP + FP + TN + FN),

where TP (TN) refer to the number of correct U (A) mutant predictions and FN (FP) are the respective counts of misclassifications. Class-specific measures include S(U) = sensitivity = TP / (TP + FN) and P(U) = precision = TP / (TP + FP), with S(A) and P(A) analogously defined. We also calculate the following values: balanced error rate (BER) and balanced accuracy rate (BAR), calculated as BER = 0.5 × [FN / (FN + TP) + FP / (FP + TN)] and BAR = 1 – BER; Matthews correlation coefficient, given by

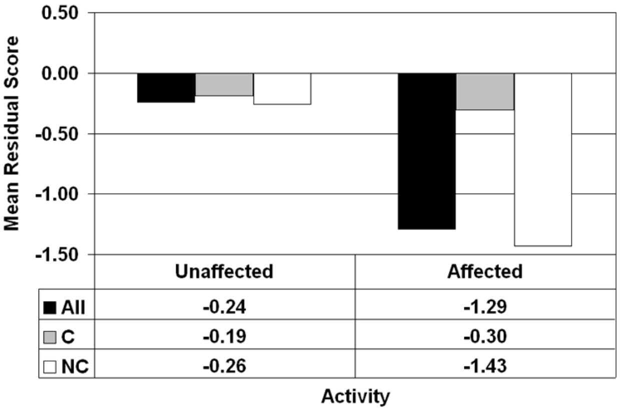
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}};$$

and area (AUC) under the receiver operating characteristic (ROC) curve, a plot of true positive rate (sensitivity) versus false positive rate (1 – specificity). An AUC of 1.0 indicates a perfect classifier whereas 0.5 suggests random guessing.

III. Results and Discussion

A. Structure-Function Relationships

We begin by examining the computed residual scores of the 8561 mutants, values that empirically quantify overall structural changes upon single residue replacements. A mean residual score is calculated for each activity class and reflects a clear trend (Fig. 1, All category), whereby a greater detrimental effect to structure (i.e., more negative mean residual score) is associated with increased functional impairment (i.e., A/affected class). Moreover, a statistically significant difference exists between the mean residual scores of the two activity change classes (*t*-test, $p < 0.0001$). The mutants within each activity class of Fig. 1 are further categorized according to whether they represent conservative (C) or non-conservative (NC) replacements of the respective native residues [18], revealing that NC substitutions drive the overall trend. On the other hand, C substitutions by definition minimally impact protein structure regardless of effects on activity, a physicochemical trait of amino acids captured by mean residual scores of C substitutions in Fig. 1.



	Unaffected	Affected
■ All	-0.24	-1.29
▨ C	-0.19	-0.30
□ NC	-0.26	-1.43

Fig. 1. Structure-function relationships. The residual scores of mutants elucidate a structure (mean residual score) – function (activity change) correlation (C / NC = conservative / non-conservative residue substitutions).

TABLE II
RF 10-fold CV Performance Measures

Data	ACC	S(U)	P(U)	S(A)	P(A)	BER	MCC	AUC
PR	0.83	0.74	0.83	0.89	0.82	0.18	0.64	0.89
RT	0.73	0.72	0.71	0.74	0.75	0.27	0.46	0.78
lys	0.82	0.88	0.87	0.71	0.73	0.21	0.59	0.89
GVP	0.74	0.72	0.62	0.75	0.82	0.27	0.45	0.78
barn	0.97	0.99	0.97	0.50	0.71	0.26	0.57	0.88
lac	0.84	0.86	0.85	0.81	0.82	0.16	0.67	0.92
IL-3	0.85	0.87	0.93	0.79	0.66	0.17	0.62	0.88
ALL	0.84	0.89	0.85	0.76	0.81	0.18	0.65	0.91

B. Random Forest Classification of Mutant Activity

The 10-fold CV prediction results reported in Table II are based on the evaluation of protein-specific mutant subsets as well as the combined set of all 8561 mutants. To assess statistical significance, we calculate 10-fold CV performance on 1000 control datasets, each derived from the combined set by randomly shuffling the U/A class labels among the mutants. Based on MCC and BAR permutation distributions generated from the controls (Fig. 2), the *p*-value for predictive power of the RF model trained with the combined set is 0.001. Next, plots of learning curves are used to evaluate the influence of dataset size on model performance (Fig. 3). We start by applying 10-fold CV to each of ten stratified random samples of 1000 mutants, where each subset is selected independently from the combined set, and mean performance values are calculated; subsequent iterations increment by 1000 the sizes of the sampled datasets. All curves of Fig. 3 reach plateaus, suggesting more training data may not improve performance. Third, considering protein-specific mutant subsets, our prediction results (AUTO-MUTE) reported in Table II consistently outperform those of the related methods SIFT [19], MAPP [20], and Pmut [21] (Table III). Finally as a practical test, we used the RF model trained with the combined set to obtain predictions for a diverse collection of 248 single residue substitutions, retrieved from the Protein Mutant Database [22], with available

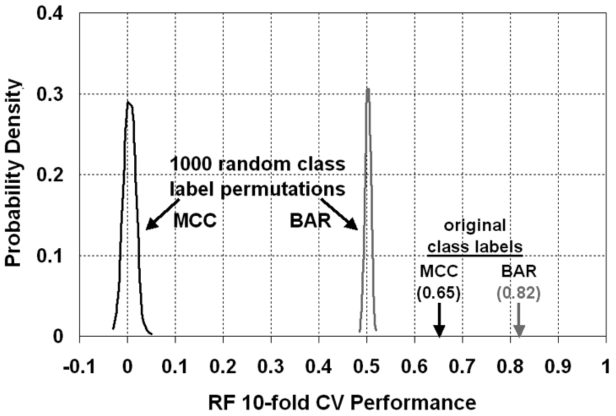


Fig. 2. Assessing the statistical significance of RF model predictions reported for the original combined set of mutants (Table II, ALL).

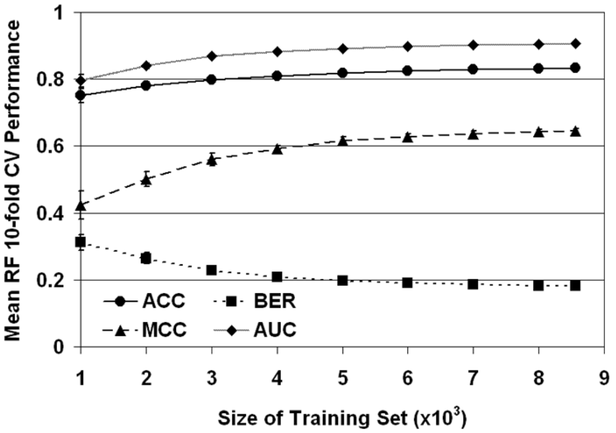


Fig. 3. Learning curves. Error bars represent ±1 standard deviation.

TABLE III
COMPARISONS WITH RELATED METHODS

Protein / Method	ACC	S(U)	P(U)	S(A)	P(A)	BER	MCC
PR							
AUTO-MUTE	0.83	0.74	0.83	0.89	0.82	0.18	0.64
SIFT	0.78	0.70	0.66	0.82	0.85	0.24	0.51
MAPP	0.76	0.62	0.89	0.92	0.68	0.23	0.55
Pmut	0.61	0.09	0.95	0.99	0.60	0.46	0.21
RT							
AUTO-MUTE	0.73	0.72	0.71	0.74	0.75	0.27	0.46
MAPP	0.64	0.85	0.44	0.56	0.90	0.30	0.37
Pmut	0.56	0.05	0.90	0.99	0.55	0.48	0.15
lys							
AUTO-MUTE	0.82	0.88	0.87	0.71	0.73	0.21	0.59
SIFT	0.63	0.59	0.82	0.72	0.45	0.35	0.29
MAPP	0.73	0.70	0.87	0.79	0.56	0.26	0.46
Pmut	0.52	0.42	0.77	0.74	0.37	0.42	0.15
lac							
AUTO-MUTE	0.84	0.86	0.85	0.81	0.82	0.16	0.67
SIFT	0.68	0.78	0.70	0.57	0.66	0.33	0.35
MAPP	0.69	0.72	0.72	0.66	0.66	0.31	0.38
Pmut	0.61	0.77	0.66	0.36	0.49	0.44	0.14

experimental activity change data. The mutations occur in 51 proteins with PDB structures, a necessary prerequisite to generate feature vectors with our *in silico* mutagenesis. Comparing predictions to experimental data yields ACC = 0.84, MCC = 0.54, and BER = 0.24.

REFERENCES

[1] J. Damborsky and J. Brezovsky, "Computational tools for designing and engineering biocatalysts," Curr Opin Chem Biol, vol. 13, 2009, pp. 26-34.

[2] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," Annu Rev Genomics Hum Genet, vol. 7, 2006, pp. 61-80.

[3] M. Masso and Vaisman, II, "AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements," Protein Eng Des Sel, vol. 23, 2010, pp. 683-687.

[4] M. Masso and Vaisman, II, "Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis," Bioinformatics, vol. 24, 2008, pp. 2002-2009.

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, 2001, pp. 5-32.

[6] D. D. Loeb, R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, et al., "Complete mutagenesis of the HIV-1 protease," Nature, vol. 340, 1989, pp. 397-400.

[7] R. Swanstrom, University of North Carolina, Chapel Hill, NC, private communication, April 2004.

[8] J. A. Wrobel, S. F. Chao, M. J. Conrad, J. D. Merker, R. Swanstrom, et al., "A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase," Proc Natl Acad Sci U S A, vol. 95, 1998, pp. 638-645.

[9] D. Rennell, S. E. Bouvier, L. W. Hardy, and A. R. Poteete, "Systematic mutation of bacteriophage T4 lysozyme," J Mol Biol, vol. 222, 1991, pp. 67-88.

[10] T. C. Terwilliger, H. B. Zabin, M. P. Horvath, W. S. Sandberg, and P. M. Schlunk, "In vivo characterization of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis," J Mol Biol, vol. 236, 1994, pp. 556-571.

[11] D. D. Axe, N. W. Foster, and A. R. Fersht, "A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease," Biochemistry, vol. 37, 1998, pp. 7157-7166.

[12] P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret, and J. H. Miller, "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence," J Mol Biol, vol. 240, 1994, pp. 421-433.

[13] P. O. Olins, S. C. Bauer, S. Braford-Goldberg, K. Sterbenz, J. O. Polazzi, et al., "Saturation mutagenesis of human interleukin-3," J Biol Chem, vol. 270, 1995, pp. 23754-23760.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, et al., "The Protein Data Bank," Nucleic Acids Res, vol. 28, 2000, pp. 235-242.

[15] M. J. Sippl, "Boltzmann's principle, knowledge-based mean fields and protein folding," Journal of Computer-Aided Molecular Design, vol. 7, 1993, pp. 473-501.

[16] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," J Mol Biol, vol. 247, 1995, pp. 536-540.

[17] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," Bioinformatics, vol. 20, 2004, pp. 2479-2481.

[18] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcut, "A model for evolutionary change in proteins," in Atlas of protein sequence and structure, vol. 5, M. O. Dayhoff, Ed. Washington D.C.: National Biomedical Research Foundation, 1978, pp. 345-352.

[19] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," Nucleic Acids Res, vol. 31, 2003, pp. 3812-3814.

[20] E. A. Stone and A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity," Genome Res, vol. 15, 2005, pp. 978-986.

[21] C. Ferrer-Costa, J. L. Gelpi, L. Zamakola, I. Parraga, X. de la Cruz, et al., "PMUT: a web-based tool for the annotation of pathological mutations on proteins," Bioinformatics, vol. 21, 2005, pp. 3176-3178.

[22] T. Kawabata, M. Ota, and K. Nishikawa, "The Protein Mutant Database," Nucleic Acids Res, vol. 27, 1999, pp. 355-357.