Protein Engineering, Design & Selection vol. 22 no. 11 pp. 665–671, 2009
Published online August 18, 2009 doi:10.1093/protein/gzp050

Modeling the functional consequences of single residue replacements in bacteriophage f1 gene V protein

Majid Masso[1], Ewy Mathe, Nida Parvez, Kahkeshan Hijazi and Iosif I. Vaisman

Laboratory for Structural Bioinformatics, Department of Bioinformatics and Computational Biology, George Mason University, 10900 University Blvd. MS 5B3, Manassas, VA 20110, USA

[1]To whom correspondence should be addressed.
E-mail: mmasso@gmu.edu

A computational mutagenesis methodology utilizing a four-body, knowledge-based, statistical contact potential is applied toward globally quantifying relative environmental perturbations (*residual scores*) in bacteriophage f1 gene V protein (GVP) due to single amino acid substitutions. We show that residual scores correlate well with experimentally measured relative changes in protein function upon mutation. Residual scores also distinguish between GVP amino acid positions grouped according to protein structural or functional roles or based on similarities in physicochemical characteristics. For each mutant, the *in silico* mutagenesis additionally yields local measures of environmental change (EC scores) occurring at every residue position (*residual profile*) relative to the native protein. Implementation of the random forest (RF) algorithm, utilizing experimental GVP mutants whose feature vector components include EC scores at the mutated position and at six structurally nearest neighbors, correctly classifies mutants based on function with up to 77% cross-validation accuracy while achieving 0.82 area under the receiver operating characteristic curve. A control experiment highlights the effectiveness of mutant feature vector signals, and a variety of learning curves are generated to analyze the impact of GVP mutant data set size on performance measures. An optimally trained RF model is subsequently used for inferring function for all the remaining unexplored GVP mutants.

Keywords: computational mutagenesis/Delaunay tessellation/knowledge-based statistical potential/random forest supervised classification/structure–function relationship

Introduction

Gene V protein (GVP) is a relatively small protein (87 amino acids), forming dimers that bind cooperatively to single-stranded DNA (ssDNA) intermediates during bacteriophage f1 replication for efficient ssDNA packaging into new phage particles (Terwilliger, 1995). The Ff filamentous phages f1, fd and M13 that infect *Escherichia coli* are very closely related, and the GVPs of these phages are identical (Skinner *et al*., 1994). When expressed at high levels, GVP also binds non-specifically to host ssDNA and ssRNA, leading to inhibition of *E.coli* growth by interfering with DNA replication or RNA translation (Terwilliger *et al*., 1994). The structure of GVP has been determined using both X-ray crystallography (Fig. 1A) (Su *et al*., 1997) and NMR (Folkers *et al*., 1994) techniques, making GVP an ideal model system for protein engineering experiments given its small size.

The analyses of experimental data obtained from large-scale mutagenesis studies on GVP have provided significant information about the structural and functional roles of the constituent amino acid residues, as well as the level of tolerance of each residue position to mutation (Terwilliger *et al*., 1994). In one study, a total of 371 single-point GVP mutants were synthesized and classified based on their degree of *E.coli* growth inhibition. A second investigation involved phenotypic classification of 138 single-point GVP mutants based on their ability to support phage f1 propagation. Each of these two sets of experiments considered a specific type of GVP function, and the phenotypic class assigned to each GVP mutant reflected the amount of functional change relative to the wild-type protein.

Since protein structure dictates function, it follows that appropriately quantified relative structural changes to GVP upon single residue replacements should correlate well with corresponding experimentally measured relative functional changes. We have developed a computational mutagenesis to compute these structural changes, which makes use of a four-body, knowledge-based, statistical contact potential (Masso and Vaisman, 2007, 2008). Underpinning these formulations is the representation of protein structures via Delaunay tessellation, a well-established computational geometry technique. For each GVP mutant, our methodology yields both a scalar *residual score* to quantify the overall relative change in sequence-structure compatibility and a vector *residual profile* to quantify relative environmental changes (EC scores) at every GVP residue position. As will be detailed in this manuscript, these quantities are useful both for elucidating structure–function relationships in GVP and for developing accurate classifiers of mutant GVP function.

Materials and methods

Experimental data

The collection of 371 GVP single-point mutants described in the literature, consisting of at least one residue substitution at each of positions 2–87, forms the principal data set for our computational studies (Terwilliger *et al*., 1994). Each mutant was overexpressed in an *E.coli* culture incubated at 37°C and functionally classified as fully active (strong inhibition of *E.coli* growth, 140 mutants), partially active (weak inhibition, 92 mutants) or inactive (no inhibition, 139 mutants).

In a second study, each of 138 GVP single-point mutants was assessed for sensitivity to temperature in their ability to support phage f1 propagation, as evidenced by the formation of plaques of differing sizes on lawns of *E.coli* at both 34°C and 40.5°C (Terwilliger *et al*., 1994). Mutants were classified as active, ts-1 (weakly temperature-sensitive, wild-type sized plaques at 34°C but slightly smaller at 40.5°C), ts-2 (much smaller plaques at 40.5°C), ts-3 (strongly

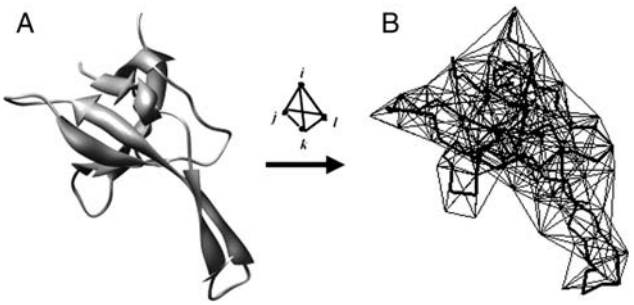

Fig. 1. (**A**) Ribbon diagram (Pettersen *et al*., 2004) of GVP based on the PDB coordinate file 1gvp. (**B**) Delaunay tessellation of GVP, using CM coordinates and a 12 Å edge-length filter, superimposed over a C-α backbone trace.

temperature-sensitive, no plaque formation at 40.5°C) or inactive. Our computational mutagenesis elucidates the correlation of structural and functional changes due to mutation, at a fixed temperature, and is not designed for simultaneously incorporating the dynamics of temperature variability on function. As such, we only considered the size of mutant GVP plaque formations at the 34°C baseline and functionally classified mutants as either active (112 mutants, including ts-1, ts-2 and ts-3) or inactive (26 mutants).

Delaunay tessellation and the four-body statistical potential

A diverse data set of 1375 high-resolution crystallographic protein structures was selected from the Protein Data Bank (PDB) (Berman *et al*., 2000). Each structure was represented as a discrete set of points in three-dimensional (3D) space, corresponding to a weighted center of mass (CM) of the side-chain atomic coordinates of the constituent amino acid residues. Delaunay tessellation was performed on each protein structure, whereby these points were utilized as vertices to generate an aggregate of non-overlapping, space-filling, irregular tetrahedral simplices (Fig. 1B) (Singh *et al*., 1996; Vaisman *et al*., 1998). The *qhull* implementation of the Quickhull algorithm (Barber *et al*., 1996) was used to tessellate each protein, and in-house programs were developed for data processing and analysis.

Each simplex in a protein tessellation objectively defines a quadruplet of nearest-neighbor residues at the vertices. For added assurance of biochemically feasible quadruplet interactions, we only considered simplices for which the lengths of all six edges were <12 Å. Excluding permutations, 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). For each quadruplet of amino acids, an observed frequency of occurrence was computed by identifying all simplices generated by the 1375 protein structure tessellations for which the quadruplet is represented by the vertices. A rate expected by chance was obtained for each quadruplet by using a multinomial reference distribution that relies on frequencies of the individual amino acids in the proteins. Modeled after the inverse Boltzmann principle, an empirical interaction potential was calculated for each quadruplet type by taking the logarithm of the ratio of observed to expected rates of occurrence, defining the four-body statistical potential (Singh *et al*., 1996; Vaisman *et al*., 1998).

Employing this potential, a score was assigned to each of the simplices in the tessellation of the GVP structure (PDB ID: 1gvp) based on the quadruplet represented by the vertices. A global *topological score* for GVP, defined by adding up the scores of all simplices in the tessellated protein, represents an overall measure of sequence-structure compatibility (Masso *et al*., 2006, 2008). A *residue environment score* was also calculated for each of the 87 amino acid positions in GVP by locally adding up only scores of simplices utilizing the corresponding CM coordinate as a vertex (Masso and Vaisman, 2007; Masso *et al*., 2008). A vector of residue environment scores, ordered by primary sequence position number, is referred to as a *3D–1D potential profile* (Fig. 2A) (Bowie *et al*., 1991; Masso and Vaisman, 2003).

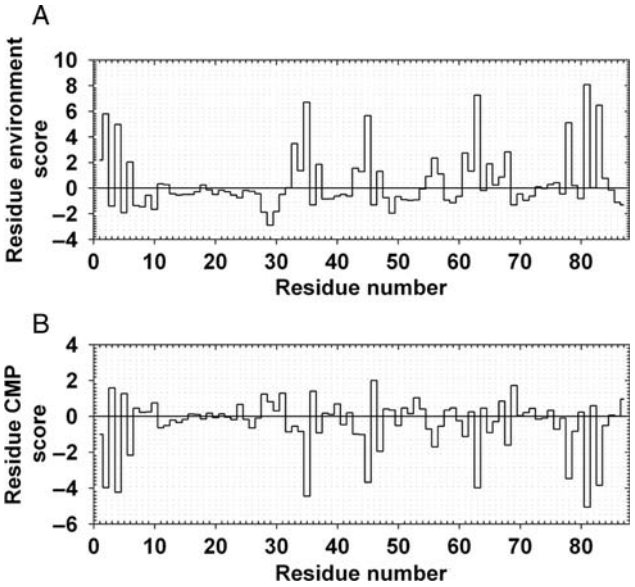

Fig. 2. (**A**) 3D–1D potential profile and (**B**) CMP profile for GVP.

Computational mutagenesis

A topological score was obtained for each single-point GVP mutant, by utilizing the tessellation of the wild-type protein structure (PDB ID: 1gvp) as a template, substituting the amino acid identity at the vertex corresponding to the position being mutated and recalculating simplex scores. The *residual score* of a GVP mutant is defined as the difference in topological scores between the mutant and wild-type protein, and provides a measure of the relative change in sequence-structure compatibility caused by the amino acid replacement (Masso *et al*., 2006; Masso and Vaisman, 2007). A *comprehensive mutational profile* (CMP) is defined by calculating, at each protein position, the mean of residual scores associated with all possible amino acid replacements (Fig. 2B) (Masso and Vaisman, 2003; Masso *et al*., 2006). Each CMP profile component is referred to as the *CMP score* of the corresponding position.

The use of a single native structural template for characterizing protein mutants is based on the following observations: most protein mutants have no corresponding solved structures; Delaunay tessellation is based on coarse-grained representation of protein structure at the residue level; and the tessellation is robust to small perturbations in the coordinates of the points representing the amino acids. Hence, tessellations for the few solved GVP structures with single residue replacements either perfectly overlap or are nearly

identical to the native structure tessellation. Additionally, in the case of mutants with solved structures for proteins in general, residual scores for the mutants (single residue substitutions in the native sequence) obtained by tessellating the native structure are comparable in magnitude but opposite in sign when compared with residual scores for the reverse mutants (single residue substitution back to the native sequence) obtained by tessellating the corresponding mutant structures (unpublished).

Replacing the amino acid identity at one vertex in the wild-type protein tessellation alters residue environment scores at this mutated position and at all nearest-neighbor positions defined by the simplices. The *residual profile* of a GVP mutant is defined as the difference in 3D−1D potential profiles between the mutant and wild-type protein, and the value of each residual profile component is referred to as an *EC score* (Masso and Vaisman, 2007, 2008). Mutant residual profiles contain implicit yet significant structure and sequence information, and the EC score at the component corresponding to the mutated position is identically equal to the residual score of the mutant.

Mutant attributes

A feature vector was generated for each single-point GVP mutant and contained as input attributes (independent variables or predictors) the identities of the native and replacement amino acids at the mutated position, the mutated position number, the residual score (EC score at the mutated position) and the EC scores at the six nearest neighbors to the mutated position, ordered nearest to farthest by Euclidean distance. Next, we included the ordered amino acid identities at the six nearest neighbors as well as their ordered primary sequence distances away from the mutated position (difference between neighbor and mutated position numbers). Finally, the following input attributes were added as feature vector components:

(1) a computed mean volume and mean tetrahedrality for the set of Delaunay simplices that utilize the mutated position as a vertex (Vaisman *et al*., 1998; Barenboim *et al*., 2008);
(2) the secondary structure {H, helix; S, strand; T, turn; C, coil} at the mutated position;
(3) depth {S, surface; U, undersurface; B, buried} at the mutated position (tessellation-based surface accessibility). Surface positions participate as one of three vertices defining a triangular facet for exactly one tetrahedron in the tessellation. Undersurface positions are defined as non-surface positions that share an edge with a surface position. All other positions are buried (Barenboim *et al*., 2008);
(4) a count of the number of simplex edges the mutated position shares with surface positions (zero by definition for buried positions).

The mutant GVP functional class defines the output attribute (dependent variable) associated with each feature vector.

Supervised learning for classification and prediction

The supervised classification scheme that we employed for this study is an implementation of Leo Breiman's random forest (RF) algorithm (Breiman, 2001), available as part of the Weka (Waikato environment for knowledge analysis) suite of machine learning tools (Frank *et al*., 2004). The RF algorithm incorporates a *bagging* (bootstrap aggregating) procedure, whereby bootstrapped data sets are used for training an ensemble of classification trees, from which predictions are obtained via majority vote (Breiman, 2001). Additionally, a fixed-size random subset of the predictor attributes is selected by the RF algorithm to split at every node encountered in each of the growing trees, all trees are unpruned and the algorithm does not overfit, regardless of the number of selected trees. The RF algorithm generally performs better than other supervised classification machine learning approaches (Qi *et al*., 2006; Bordner, 2008). We fixed adjustable RF parameters at 100 trees and five randomly selected input attributes for splitting at each tree node.

Performance of RF on data sets of GVP mutant feature vectors was evaluated by using stratified 10-fold cross-validation (10-fold CV), leave-one-out cross-validation (LOOCV) and stratified random split (66% of data set for model training and 34% for testing). Given a generic two-class training set consisting of 'positive' (P) and 'negative' (N) examples, $Q = \text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{TN})$ provides a simple measure of performance which is meaningful so long as class distributions are not highly skewed. Here, TP and TN represent the number of correct positive and negative predictions, respectively, and FP and FN are misclassifications. The balanced error rate (BER), calculated as $\text{BER} = 0.5 \times [\text{FN}/(\text{FN} + \text{TP}) + \text{FP}/(\text{FP} + \text{TN})]$, Matthew's correlation coefficient (MCC), given by

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

and area (AUC) under the receiver operating characteristic (ROC) curve provide alternative measures that are especially useful for highly unbalanced classes. A χ^2 test can be applied to assess MCC statistical significance, where the test statistic is given by $\chi^2 = N \times \text{MCC}^2$ (N = data set size) with one degree of freedom (Baldi *et al*., 2000).

The ROC curve is a plot of the true-positive rate (*sensitivity*) versus the false-positive rate (1-*specificity*), where $\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$ and $\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$, and the AUC is equivalent to the non-parametric Wilcoxon–Mann–Whitney test of ranks (Fawcett, 2003). An AUC value near 0.5 suggests that the trained model will perform no better than random guessing, while a value of 1.0 is indicative of a perfect classifier. For a data set consisting of examples that belong to more than two classes, we employ both *one-against-one* (all possible two-class subsets) and *one-against-all* (all possibilities for choosing one reference class and combining the others as non-reference) approaches. In the former case, an overall AUC for the multi-class data set is calculated as the mean of the AUC values that correspond to ROC curves for each of the two-class subsets (Hand and Till, 2001). In the latter case, the overall AUC is obtained by computing a weighted average of the AUC values that correspond to ROC curves for each of the reference/non-reference data sets, where each AUC weight equals the proportion of reference class examples (Provost and Domingos, 2001).

Results and discussion

GVP structure–function relationships

On the basis of the data set of 371 GVP mutants experimentally assessed for their ability to inhibit the growth of *E.coli*, we computed a mean residual score for the mutants in each class (Fig. 3A, 'All' category). A clear trend emerges whereby increasingly detrimental effects on structure due to mutation, as reflected by decreasing mean residual scores, are associated with higher levels of functional impairment. Furthermore, a *t*-test reveals that a statistically significant difference exists between mean residual scores for the most disparate class pair (full/inactive, $P < 0.001$). Within each class, mutants were also clustered based on whether they represented conservative (C) or non-conservative (NC) substitutions of the wild-type residue (Dayhoff *et al*., 1978), and we computed mean residual scores for each of these subgroups. With the 20 amino acids clustered into six groups as {(A,S,T,G,P), (D,E,N,Q), (R,K,H), (F,Y,W), (V,L,I,M), (C)} based on similarities in physicochemical properties, intraclass residue replacements are C whereas interclass substitutions are NC. Note that the overall trend is driven by NC mutations, since C substitutions generally have a minimal impact on sequence-structure compatibility regardless of the phenotype.

Classification models in the current computational mutagenesis literature are typically based on whether protein mutants are unaffected (e.g. full) or affected (e.g. partial and inactive combined) by their corresponding residue replacements (Krishnan and Westhead, 2003; Verzilli *et al*., 2005; Mathe *et al*., 2006; Ng and Henikoff, 2006; Bromberg and Rost, 2007). Additional justifications for such a two-class grouping of the 371 GVP mutants are discussed later when results concerning inferential models are presented, and a statistically significant difference exists between mean residual scores for this unaffected/affected class pair ($P < 0.001$).

Finally, mean residual scores were also used for elucidating the GVP structure–function relationship based on the ability of single-point GVP mutants to support phage f1 propagation at 34°C (Fig. 3B). Again, a statistically significant difference exists between mean residual scores for the active/inactive class pair ($P < 0.005$), and the observed trend is principally attributable to NC substitutions.

Classification of GVP residue positions

A strong inverse correlation ($R^2 = 0.86$) exists between the CMP profile of GVP, obtained by averaging the residual scores of all amino acid replacements at each position, and the 3D–1D potential profile of the protein, which provides an environment score for each position (Figs 2 and 4). By averaging residual scores of NC and C substitutions separately at each position, modified NC-CMP and C-CMP profiles showed that this correlation is due to the NC substitutions ($R^2 = 0.86$), without any contribution from the C substitutions ($R^2 = 0.02$). Similar observations based on this *in silico* application of our methodology have been made for HIV-1 protease (Masso and Vaisman, 2003; Masso *et al*., 2006), *lac* repressor (Masso *et al*., 2008) and a number of other proteins (unpublished), revealing a consistent pattern of residue clustering (hydrophobic, Quad 4; charged, Quad 2; polar, origin).

On the basis of annotations provided in the literature, 73 out of 87 GVP residue positions were each assigned to one of four groups according to structural locations and functional considerations. Table I provides a distribution of residue positions by group as well as by quadrant location (Fig. 4), and Fisher's exact test leads us to reject the null hypothesis that no association exists between the structural/functional groups and the quadrant locations ($P < 0.0001$). We also characterized each group based on both the mean of the residue environment scores (MRES) of the positions in the group and the mean of the mutant residual scores (All, C, NC) for all 19 residue replacements at all positions in the

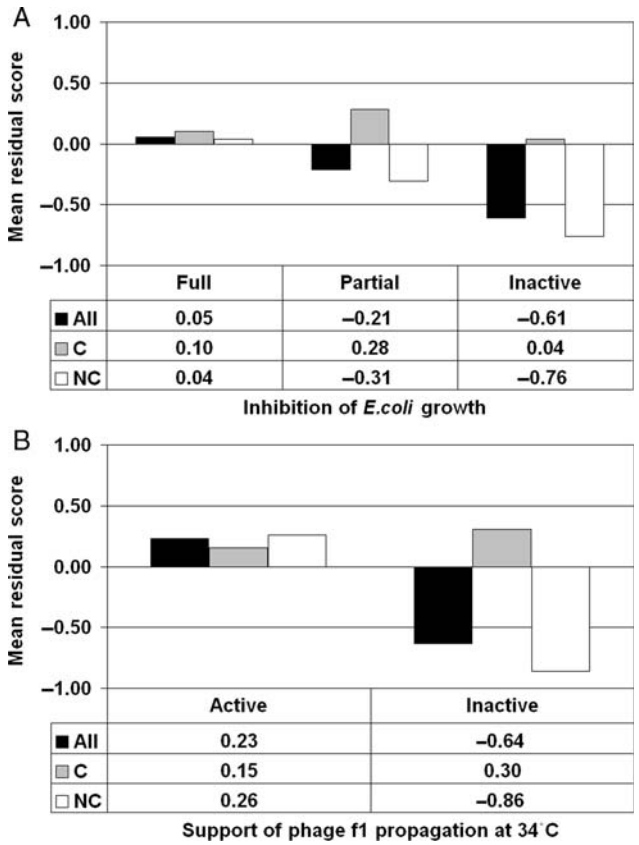

Fig. 3. GVP structure–function correlations (see text for C/NC mutant subsets).

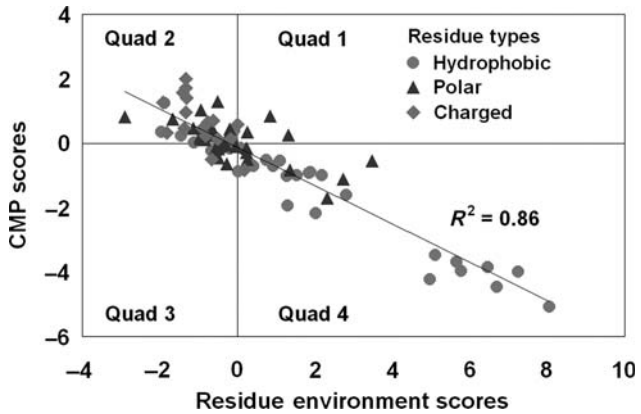

Fig. 4. CMP potential profile correlation. See Supplementary data available at *PEDS* online for color figure.

group combined (Fig. 5). It is clear from Fig. 5 that our computational characterization of these groups effectively discriminates between hydrophobic core and DNA/RNA-binding residues, as well as distinguishing between interface positions and other surface residues that are not as structurally or functionally vital.

Inferential models of mutant GVP activity

As detailed earlier, feature vectors were derived for each of the 371 GVP mutants belonging to one of three phenotypic classes based on their degree of *E.coli* growth inhibition. Training sets included three proper two-class subsets (one-against-one), as well as three versions of the complete data set of mutants relabeled to reflect only two classes (one-against-all, whereby one class is chosen as 'reference' and the other two classes are combined as 'non-reference'). Application of RF supervised classification in conjunction with the 10-fold CV testing procedure on these training sets reveals that the full and inactive pair of GVP mutant classes encode the most disparate signals in their feature vectors and are most easily distinguishable from one another, reflecting the intuitive biological notion that this pair of mutant functional classes exhibits the most significant structural differences (Fig. 6, Table II). Additionally, the higher performance measures obtained using the full/partial two-class subset over the partial/inactive subset, coupled with similarly higher measures using the full/others combined complete data set over the inactive/others combined data set, suggest that GVP mutants in the partial class are more similar to their inactive counterparts rather than to the fully active GVP mutants (Fig. 6, Table II).

The results summarized above justify the clustering of these 371 GVP mutants into two classes based on segregating fully active mutants from the others combined, which we will subsequently refer to by using the following class labels: *unaffected* (full) and *affected* (partial and inactive combined). Performance of the RF algorithm on this data set

Table I. Distribution of annotated residues

Graph Quads	Residue types				Total
	Surface[a]	Hydrophobic core[a]	DNA/RNA binding[b]	Interface[c]	
Q1	1	0	0	1	2
Q2	22	1	9	1	33
Q3	6	0	2	2	10
Q4	10	13	1	4	28
Total	39	14	12	8	73

[a]Terwilliger *et al*. (1994); [b]Skinner *et al*. (1994); [c]Stassen *et al*. (1992) and Su *et al*. (1997).

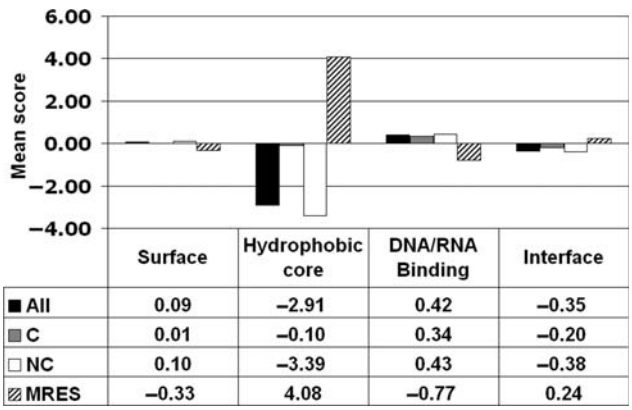

Fig. 5. Characterization of GVP residues. MRES, mean of the residue environment scores.

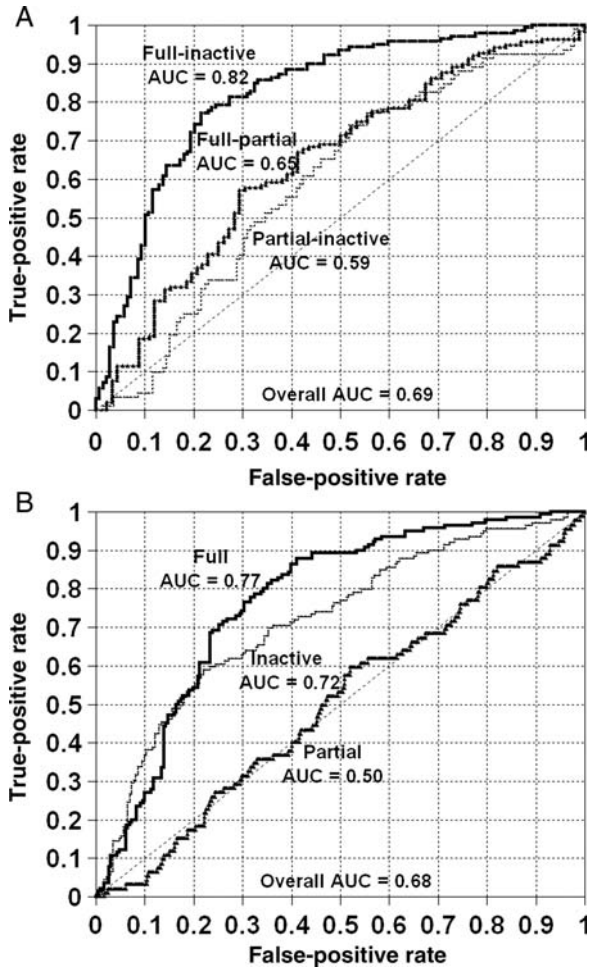

Fig. 6. (**A**) One-against-one and (**B**) one-against-all ROC curves based on 10-fold CV.

Table II. RF performance (10-fold CV)

Class	Full	Partial	Inactive	Others combined
Full		$Q = 0.63$ MCC = 0.22 BER = 0.39 AUC = 0.65	$Q = 0.77$ MCC = 0.54 BER = 0.23 AUC = 0.82	$Q = 0.73$ MCC = 0.44 BER = 0.27 AUC = 0.77
Partial			$Q = 0.59$ MCC = 0.16 BER = 0.42 AUC=0.59	$Q = 0.53$ MCC = 0.03 BER = 0.48 AUC = 0.50
Inactive				$Q = 0.66$ MCC = 0.30 BER = 0.34 AUC = 0.72

Table III. RF performance (unaffected/affected)

Method	Q	MCC	BER	AUC
10-fold CV[a]	0.71 ± 0.01	0.40 ± 0.02	0.29 ± 0.01	0.76 ± 0.01
LOOCV	0.72	0.42	0.29	0.77
66/34 split[a]	0.69 ± 0.03	0.37 ± 0.07	0.31 ± 0.04	0.74 ± 0.03

[a]Ten iterations performed for 10-fold CV and 66/34 split methods.

was evaluated based on running 10 iterations each of 10-fold CV and 66/34 stratified random split, as well as LOOCV, with relatively consistent results across all three techniques (Table III). All MCC values associated with each method are statistically different from zero ($P < 0.0001$), indicating that the RF predictions are significantly more correlated with the data compared with random guessing.

In particular, the 10-fold CV results were compared with those obtained by using a control derived from the mutant GVP data set by randomly shuffling the original unaffected/affected class labels among the mutants, for which $Q = 0.57$, MCC $= 0.10$, BER $= 0.45$ and AUC $= 0.55$. The results suggest that a model trained with this 'shuffled classes' random control cannot perform better than random guessing, highlighting the strength of signals embedded in feature vectors of the original data set.

Using the unaffected/affected two-class labeling of the 371 GVP mutants, we compared the performance of our RF model (Table II, full versus others combined) with that of other state-of-the-art approaches. Identification of other methods appropriate for making comparisons is a non-trivial issue in this regard. For example, there are a number of tools available for predicting mutant stability change (e.g. $\Delta\Delta G$); however, the property to be predicted for the GVP mutants is the effect on activity (degree of inhibition of *E.coli* growth), which in the aggregate is not directly correlated to the effect on stability. One well-known server for predicting $\Delta\Delta G$ known as PoPMuSiC (http://babylone.ulb.ac.be/popmusic/) was used to obtain predictions for the 371 GVP mutants, where increased (decreased) stability from wild type was interpreted as an unaffected (affected) prediction. PoPMuSiC predicts values of $\Delta\Delta G$ upon mutation by utilizing different combinations of database-derived torsion and amino acid pair distance potentials based on the solvent accessibility of the mutated position. As expected, the performance was extremely poor ($Q = 0.58$, MCC $= -0.02$ and BER $= 0.50$) and equivalent to random guessing; however, this is the result of an inappropriate application of the method and does not reflect the ability of the tool to generate accurate predictions. Similarly, there exist numerous servers for predicting whether a single-residue substitution has either no effect or any effect on protein function. Often the models driving the servers have been trained using only human proteins, as is the case for example with PolyPhen (http://genetics.bwh.harvard.edu/pph/), nsSNPAnalyer (http://snpanalyzer.utmem.edu/), Pmut (http://mmb2.pcb.ub.es:8080/PMut/) and PhD-SNP (http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm), which renders these tools unreliable in making accurate GVP mutant predictions. On the other hand, SIFT (http://blocks.fhcrc.org/sift/SIFT.html) and SNAP (http://cubic.bioc.columbia.edu/services/SNAP/) were trained using variant proteins from diverse organisms. In particular, SIFT is sequence-based, SNAP is structure-based, and both utilize evolutionary information available in the form of multiple sequence alignments. Our RF model moderately outperforms SNAP ($Q = 0.68$, MCC $= 0.36$ and BER $= 0.31$) and significantly outperforms SIFT ($Q = 0.60$, MCC $= 0.22$ and BER $= 0.38$).

The novelty of our structure-based approach derives from combining supervised classification (RF) with mutant attributes (EC scores) obtained using a four-body potential. These are two predictive approaches that previously have only been studied separately. As our method does not explicitly incorporate evolutionary information, it serves as an orthogonal approach that complements other methods such as SIFT and SNAP that utilize information derived from multiple sequence alignments. Our method for mutant feature vector representation has been used to develop independent models that can predict stability change (Masso and Vaisman, 2008) as well as functional change (Masso and Vaisman, 2007). Each model is trained using a data set of diverse protein mutants with either experimentally determined stability change (e.g. $\Delta\Delta G$) or functional change (i.e. effect on activity), respectively. Subsequently, each model can be used to make predictions about new, unexplored mutants with respect to the type of property change on which the model was trained. For increased (decreased) stability mutants, the mutated position and its neighbors often display favorable (unfavorable) changes in the form of positive (negative) EC scores, and hence providing an important discriminating feature that classification algorithms can exploit. Similarly with functional changes, the mutated position and its neighbors generally display EC scores that are relatively small in magnitude for unaffected mutants compared with those that are affected, which again provides the algorithms with distinguishing features for developing accurate classification models.

Next, we generated learning curves in order to assess the influence of data set size on trained RF model performance. We began by applying RF learning and 10-fold CV to each of 10 stratified random samples of 100 mutants, selected from among the 371 experimental GVP mutants, and a mean accuracy and standard deviation (SD) was calculated. Subsequent iterations involved incrementing by 50 mutants the size of the sampled data sets. The lack of plateaus in the learning curves as the data set size approaches 371 indicates that enlargement of the current mutant GVP data set may further optimize performance of the RF model (Fig. 7).

Finally as an important practical application, we employed the RF model learned from the entire training set of 371 mutants in order to predict the unaffected/affected class memberships of all remaining uncharacterized single-point GVP mutants. In particular, we had generated the feature vector input attributes for all 87 positions $\times$ 19 substitutions/position $= 1653$ mutants, leaving 1282 mutants to form a separate test set, each with an unknown functional class output attribute. On the basis of signals encoded by the input attributes of their feature vectors, the RF model generated a functional class prediction for every test set mutant. We pooled all experimental and predicted GVP mutants into the array shown in Fig. 8, which summarizes overall mutational patterns in the protein. Columns represent residue positions in GVP, and rows represent the 20 possible amino acid

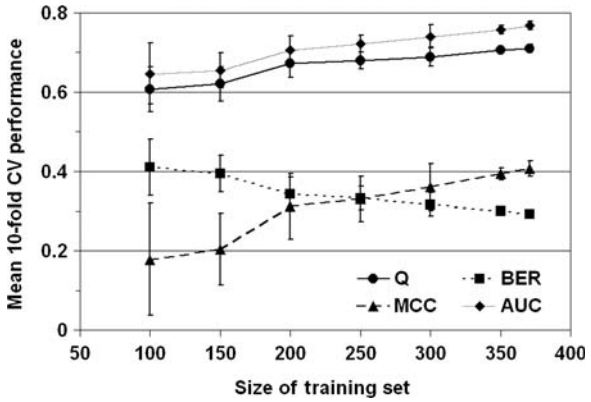

Fig. 7. Learning curves. Error bars represent ± 1 SD from the mean.

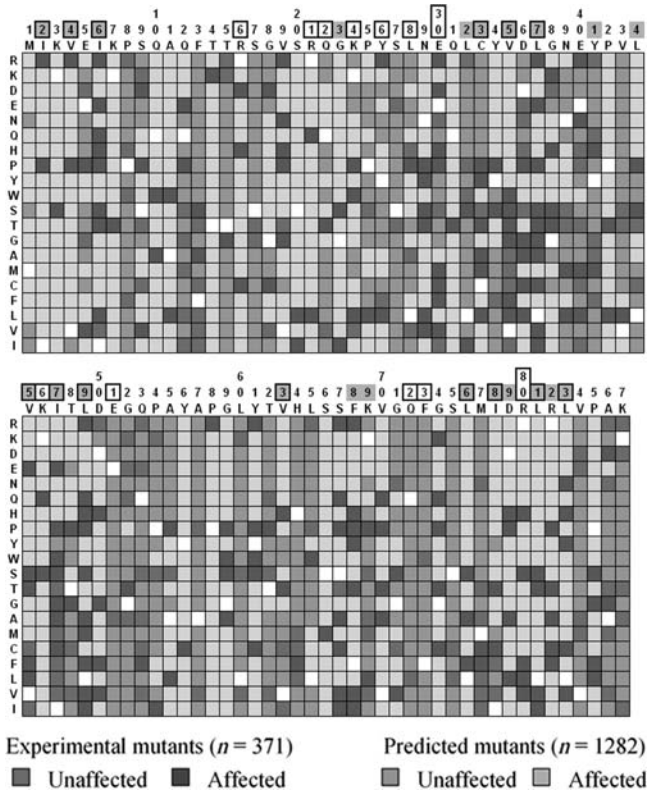

Fig. 8. GVP mutational array. Columns, native amino acids; rows, substitutions; darker shades, experimental mutants; lighter shades, predicted mutants; white squares, self-substitutions; boxed numbers, DNA/RNA binding residues; shaded numbers, interface residues; boxed and shaded numbers, hydrophobic core residues. See Supplementary data available at *PEDS* online for color figure.

replacements, arranged from top to bottom in order of increasing hydrophobicity (Kyte and Doolittle, 1982). Notably, at interface (G23, L44, F68, D79, R82), DNA/RNA binding (R16, R21, K24, Y26, E30, K46, R80) and hydrophobic core (I2, V4, I6, C33, V35, L37, V45, I47, L49, V63, L76, I78, L81, L83) positions known to be intolerant to specific types of amino acid substitutions, our predictions are well in line with the experimental GVP mutant data.

Acknowledgements

GVP ribbon diagram was produced using the UCSF Chimera package, and tessellation visualization was produced using Matlab, Version 7.0.1.24704 (R14) Service Pack 1.

References

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) *Bioinformatics*, **16**, 412–424.
Barber,C.B., Dobkin,D.P. and Huhdanpaa,H.T. (1996) *ACM Trans. Math. Software*, **22**, 469–483.
Barenboim,M., Masso,M., Vaisman,I.I. and Jamison,D.C. (2008) *Proteins*, **71**, 1930–1939.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Bordner,A.J. (2008) *Bioinformatics*, **24**, 2865–2871.
Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Breiman,L. (2001) *Mach. Learn.*, **45**, 5–32.
Bromberg,Y. and Rost,B. (2007) *Nucleic Acids Res.*, **35**, 3823–3835.
Dayhoff,M.O., Schwartz,R.M. and Orcut,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5. National Biomedical Research Foundation, Washington, DC, pp. 345–352.
Fawcett,T. (2003) *Technical Report HPL-2003–4*. Hewlett-Packard Labs, Palo Alto.
Folkers,P.J., Nilges,M., Folmer,R.H., Konings,R.N. and Hilbers,C.W. (1994) *J. Mol. Biol.*, **236**, 229–246.
Frank,E., Hall,M., Trigg,L., Holmes,G. and Witten,I.H. (2004) *Bioinformatics*, **20**, 2479–2481.
Hand,D.J. and Till,R.J. (2001) *Mach. Learn.*, **45**, 171–186.
Krishnan,V.G. and Westhead,D.R. (2003) *Bioinformatics*, **19**, 2199–2209.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Masso,M. and Vaisman,I.I. (2003) *Biochem. Biophys. Res. Commun.*, **305**, 322–326.
Masso,M. and Vaisman,I.I. (2007) *Bioinformatics*, **23**, 3155–3161.
Masso,M. and Vaisman,I.I. (2008) *Bioinformatics*, **24**, 2002–2009.
Masso,M., Lu,Z. and Vaisman,I.I. (2006) *Proteins*, **64**, 234–245.
Masso,M., Hijazi,K., Parvez,N. and Vaisman,I.I. (2008) In Mandoiu,I., Sunderraman,R. and Zelikovsky,A. (eds.), *Lecture Notes in Bioinformatics*, Vol. 4983. Springer, Heidelberg, pp. 390–401.
Mathe,E., Olivier,M., Kato,S., Ishioka,C., Vaisman,I. and Hainaut,P. (2006) *Hum. Mutat.*, **27**, 163–172.
Ng,P.C. and Henikoff,S. (2006) *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.
Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) *J. Comput. Chem.*, **25**, 1605–1612.
Provost,F. and Domingos,P. (2001) *CeDER Technical Report IS-00–04*. Stern School of Business, New York University, New York.
Qi,Y., Bar-Joseph,Z. and Klein-Seetharaman,J. (2006) *Proteins*, **63**, 490–500.
Singh,R.K., Tropsha,A. and Vaisman,I.I. (1996) *J. Comput. Biol.*, **3**, 213–221.
Skinner,M.M., Zhang,H., Leschnitzer,D.H., Guan,Y., Bellamy,H., Sweet,R.M., Gray,C.W., Konings,R.N., Wang,A.H. and Terwilliger,T.C. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 2071–2075.
Stassen,A.P., Harmsen,B.J., Schoenmakers,J.G., Hilbers,C.W. and Konings,R.N. (1992) *Eur. J. Biochem.*, **206**, 605–612.
Su,S., Gao,Y.G., Zhang,H., Terwilliger,T.C. and Wang,A.H. (1997) *Protein Sci.*, **6**, 771–780.
Terwilliger,T.C. (1995) *Adv. Protein Chem.*, **46**, 177–215.
Terwilliger,T.C., Zabin,H.B., Horvath,M.P., Sandberg,W.S. and Schlunk,P.M. (1994) *J. Mol. Biol.*, **236**, 556–571.
Vaisman,I.I., Tropsha,A. and Zheng,W. (1998) *Proceedings of the IEEE Symposia on Intelligence and Systems*, pp. 163–168.
Verzilli,C.J., Whittaker,J.C., Stallard,N. and Chasman,D. (2005) *Appl. Stat.*, **54**, 191–206.

Received April 2, 2009; revised July 19, 2009; accepted July 20, 2009

Edited by Richard Goldstein