PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers

Maxim Barenboim,[†] Majid Masso, Iosif I. Vaisman, and D. Curtis Jamison*[‡]

Department of Bioinformatics and Computational Biology, George Mason University, Manassas, Virginia 20110, USA

ABSTRACT

There is substantial interest in methods designed to predict the effect of nonsynonymous single nucleotide polymorphisms (nsSNPs) on protein function, given their potential relationship to heritable diseases. Current state-of-the-art supervised machine learning algorithms, such as random forest (RF), train models that classify single amino acid mutations in proteins as either neutral or deleterious to function. However, it is frequently the case that the functional effect of a polymorphism on a protein resides between these two extremes. The utilization of classifiers that incorporate fuzzy logic provides a natural extension in order to account for the spectrum of possible functional consequences. We generated a dataset of single amino acid substitutions in human proteins having known three-dimensional structures. Each variant was uniquely represented as a feature vector that included computational geometry and knowledge-based statistical potential predictors obtained though application of Delaunay tessellation of protein structures. Additional attributes consisted of physicochemical properties of the native and replacement amino acids as well as topological location of the mutated residue position in the solved structure. Classification performance of the RF algorithm was evaluated on a training set consisting of the disease-associated and neutral nsSNPs taken from our dataset, and attributes were ranked according to their relative importance. Similarly, we evaluated the performance of adaptive neuro-fuzzy inference system (ANFIS). The utility of statistical geometry predictors was compared with that of traditional structural and evolutionary attributes employed by other researchers, revealing an equally effective yet complementary methodology. Among all attributes in our feature set, the statistical geometry predictors were found to be the most highly ranked. On the basis of the AUC (area under the ROC curve) measure of performance, the ANFIS and RF models were equally effective when only statistical geometry features were utilized. Tenfold cross-validation studies evaluating AUC, balanced error rate (BER), and Matthew's correlation coefficient (MCC) showed that our RF model was at least comparable with the well-established methods of SIFT and PolyPhen. The trained RF and ANFIS models were each subsequently used to predict the disease potential of human nsSNPs in our dataset that are currently unclassified (http://rna.gmu.edu/FuzzySnps/).

Proteins 2008; 71:1930–1939.

Key words: single nucleotide polymorphism; computational geometry; artificial intelligence; algorithms; protein conformation; statistical models; complex disease; tertiary protein structure; computational biology; amino acid substitution.

INTRODUCTION

The ability to predict the effects of nonsynonymous SNPs (nsSNPs) on protein function is important for the success of genetic disease association studies. According to the common disease-common variants hypothesis, common diseases involve common variants that may affect a significantly larger percentage of the population than Mendelian diseases and result from both genetic and environmental factors.[1,2] Direct association analysis examines those nsSNPs which have a high probability of not only being associated with a disease, but which also may directly contribute to disease.

There are three general approaches described in the literature for identifying deleterious nsSNPs: empirical, probabilistic, and machine learning methods.[3–7] Attributes employed with these methods can also be divided into two main subgroups: either structure-based or evolution-based. The core theory behind these methods is that deleterious mutations can either destabilize structure or disrupt a functional site, such as a ligand-binding, cat-

[†] Current address: Clinical Brain Disorders Branch, Genes, Cognition, and Psychosis Program, National Institute of Mental Health, NIH, Bethesda, MD 20892, USA. E-mail: barenboimm@mail.nih.gov

[‡] Current address: College of Informatics, Northern Kentucky University, Highland Heights, KY 41099, USA. E-mail: jamisond1@nku.edu

*Correspondence to: D. Curtis Jamison.

Received 7 June 2007; Revised 11 September 2007; Accepted 20 September 2007
Published online 10 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21838

alytic, or protein–protein interaction site.[8] Destabilizing mutations should preferably be determined by a structure-based approach; however, functional sites are much easier to identify by scoring conserved positions through a phylogenetic analysis, as performed by the SIFT software, in which prediction is based on conservation built purely on orthologous protein alignments.[9] However, such a prediction could be unreliable if there are few homologs available. The combination of structural and evolutionary attributes has been demonstrated to improve prediction by the PolyPhen software, which predicts possible impact of an amino acid substitution on the structure and function of a human protein by using straightforward physical and comparative considerations.[10] However, when there are not enough structural parameters, its classification is based predominantly on comparative analysis. Thus, structural attributes are complementary to evolutionary ones, rather than overlapping.

Previous studies reflect widely differing notions of the meaning of a deleterious mutation. For example, in several studies an amino acid substitution is referred to as deleterious if the experimentally measured activity of the mutant protein is diminished to some degree relative to that of the native protein.[3,4] In other studies, a mutation is considered deleterious if the difference in the free energy of unfolding ($\Delta\Delta G = \Delta G_{\mathrm{mutant}} - \Delta G_{\mathrm{native}}$), a measure of mutant stability, is experimentally determined to be negative.[5] Additionally, the datasets utilized in these reports include single point mutants of proteins belonging to a variety of organisms, including viruses and bacteria. A comparative study utilizing a variety of datasets revealed that the classifier trained on a collection of human nsSNP, derived from the Swiss-Prot variant pages,[11] provided the best approach for predicting deleterious mutations.[12] The specific focus of this manuscript is on human nsSNPs resulting in mutant proteins that are either associated with a disease (daSNPs) or are simply neutral polymorphisms (ntSNPs).

Machine learning approaches have been applied previously for identifying disease-associated human nsSNPs, including implementations of the neural network,[13] the decision tree,[14] and the support vector machine[15] algorithms. Recently, random forest (RF)[16] classification of human nsSNPs was also investigated, with the use of predictors representing both the structural environment of the mutated protein positions as well as evolutionary information about those sites.[17] In this report, we evaluate the performance of RF learning based on a training set of human nsSNPs in proteins with known three-dimensional (3D) structures. However, the feature vectors for our variant proteins consist of computational geometry and statistical potential attributes obtained by means of Delaunay tessellation of protein structure. The alpha shape method (a computational geometry application of Delaunay tessellation) was previously used successfully to classify nsSNPs based on the geometric location of the mutated amino acid positions in protein structures, without the explicit utility of machine learning tools.[18] Here, we define an alternative set of computational geometry (V, sT) and topological location (S – surface, U – undersurface, B – buried) predictors and utilize a Delaunay tessellation-derived statistical potential in order to define an empirical measure (ΔQ) of the overall structural impact that the amino acid replacement has on the protein.[19] The remaining components in our 17-dimensional feature vector for each nsSNP characterize the physicochemical characteristics of the native and replacement amino acids at the mutated position and their degree of similarity.

Performance of the RF algorithm on our dataset of nsSNPs, based on 10-fold and leave-one-out (jackknife) cross-validation (CV), as well as random splits of the dataset into separate training and test sets, is measured by calculating area under the ROC curve (AUC), balanced error rate (BER), and Matthew's correlation coefficient (MCC). Our RF results are compared with those of Bao and Cui[17] in order to compare the effectiveness of our alternative predictors. Additional comparisons are made with the well-established methods of SIFT[9] and PolyPhen.[10] Similar to other supervised learning tools, the trained RF model classifies each nsSNP as either disease-associated or neutral, and a confidence measure is provided for each prediction. However, the functional effects of nsSNPs on their respective proteins are frequently best described as lying somewhere in between these two extremes, belonging to either class up to some degree. To explicitly account for the spectrum of possible impact, we also apply an adaptive neuro-fuzzy inference system (ANFIS) and evaluate its performance.[20] Finally, we obtain two separate yet complementary predictions for each of the nsSNPs that are currently unclassified by utilizing both the trained RF and ANFIS models.

MATERIALS AND METHODS

Selection of nsSNPs

The data collection process leading to the set of nsSNPs used for this study has been previously detailed elsewhere.[19] Briefly, data describing 3905 human nsSNPs with three-dimensional (3D) models were extracted from the Swiss-Prot variant pages,[11] including amino acid position, status, primary sequence location, and 3D template structure. A variant was considered only if its template was an X-ray structure with better than 2.5 Å resolution. A template here refers to either the structure of the wild type (wt) protein from which the nsSNP is derived, or a homolog that shares at least 70% sequence identity. On the basis of the variant status, each nsSNP was assigned to one of three subsets: disease-associated

(2017 daSNPs), neutral (1004 ntSNPs), or unclassified (884 unSNPs).[21]

In subsequent filtering steps, nsSNPs were deleted from their respective subsets and not considered in the analysis if they satisfied the following prohibitive conditions. First, after mapping the variant target sequence onto the template structure retrieved from the Brookhaven Protein Data Bank (PDB), an nsSNP was excluded if the native Swiss-Prot amino acid at the variant position was not identical to the aligned residue in the PDB file.[22] Second, an nsSNP was removed if its corresponding template PDB structure file contained either nonconsecutive residue numbering or gaps in the sequence. Lastly, if proline or glycine appeared as the native or replacement amino acids at a mutated residue position defining an nsSNP, then the nsSNP was no longer considered. The last two conditions were necessary in order to accurately characterize each nsSNP using the Delaunay tessellation methodology as described in the following section. After completing these filtering steps, the subsets were reduced to 919 daSNPs mapped onto 91 distinct PDB structures, 432 ntSNPs mapped onto 129 PDB structures, and 568 unSNPs mapped onto 41 PDB structures.[19] We refer to this collection of 1919 nsSNPs as the *complete* set.

Characterization of nsSNPs based on Delaunay tessellation of PDB structures

To perform the Delaunay tessellation of a PDB structure, each constituent amino acid is initially abstracted to a point by considering the 3D coordinates of its C_α atom. The Delaunay tessellation of the protein structure, represented as such a discrete collection of points in 3D space, yields an aggregate of space-filling, nonoverlapping, irregular tetrahedral simplices.[23,24] Each simplex objectively defines a quadruplet of nearest-neighbor amino acids at its vertices. A reference set of 1208 proteins of high crystallographic resolution and low sequence and structure similarity were individually tessellated, and based on the complete collection of simplices that were produced, a log-likelihood score was calculated for each of the 8855 possible unordered amino acid quadruplets.[23,24] The score of a quadruplet measures the propensity for observing the four amino acids as forming nearest-neighbor simplices in proteins relative to the random chance of occurrence, and the complete set of quadruplets along with their respective scores constitutes the four-body, knowledge-based, statistical contact potential. The *qhull* algorithm was used to perform the Delaunay tessellations.[25] An in-house suite of Java programs was used for preprocessing of the PDB structure files, which includes checking for consecutive residue numbering and the absence of gaps in the protein sequence, and postprocessing of the output data from *qhull*.

Given a tessellatable PDB coordinate file of an nsSNP template, the total potential or topological score Q_{wt} of the protein is calculated as the sum of the scores of the quadruplets defined by all of the simplices in the protein tessellation.[19,26] The topological score of the protein with the amino acid replacement at the variant position Q_{mut} is also obtained from the same template structure tessellation, by simply replacing the amino acid identity at the C_α point corresponding to the variant position and recalculating the sum of the quadruplet scores associated with all of the simplices.[19,26] Note that quadruplet scores are altered only for simplices that use the variant C_α position as a vertex. The use of the template structure tessellation for computing Q_{mut} is justified for the following reasons: an amino acid substitution generally causes minimal local structural shifts in the backbone, every amino acid is represented by the backbone C_α coordinate, and the tessellation is robust to small perturbations in the C_α points. Exceptions to this approach occur with glycine and proline residues given the strain imposed on the protein backbone when they are involved as either native or replacement amino acids.[27–29]

Specific attributes of an nsSNP can be gleaned from the Delaunay tessellation of the template structure. First, $\Delta Q = Q_{\mathrm{mut}} - Q_{\mathrm{wt}}$ provides an empirical measure of the structural impact because of the residue replacement.[26] Next, considering only the tetrahedral simplices for which the variant residue position participates as a vertex, we define V = mean simplex volume and sT = mean simplex tetrahedrality, where tetrahedrality quantifies the degree of distortion of a simplex from the ideal tetrahedron.[24,30] Finally, a topological location (surface, undersurface, or buried) is assigned to an nsSNP based on the variant C_α position. Surface (S) is used when the point lies on the convex hull of the tessellation, undersurface (U) indicates that a tetrahedral edge connects the C_α position to a surface point, and buried (B) refers to all others.

Predictors

By considering S, U, and B as Boolean binary variables, only one of which takes on the value 1 while the other two are set to 0 for an nsSNP, there are a total of six predictors based on Delaunay tessellation with the inclusion of the real-valued variables ΔQ, V, and sT. Next, consider the following clustering of the amino acids according to their side chain polarity: F = hydrophobic = (A, V, L, I, M, P, F), L = charged = (D, E, R, K), and P = polar = (N, Q, W, S, T, G, C, H, Y). This three-letter alphabet generates nine Boolean binary variable predictors (FF, FL, FP, LF, LL, LP, PF, PL, PP), which describe the polarity of the native (first letter) and replacement (second letter) amino acids defining the nsSNP. Finally, the 20 amino acids may be segregated into the six subgroups {(A, S, T, G, P), (V, L, I, M), (R, K, H), (D, E, N, Q), (F, Y, W),

Table I
Example of a Feature Vector With all 17 Predictors

Categorical			Continuous			Categorical		
Group		Substitution[a]	Statistical potential	Computational geometry		Topology		
C	*NC*	FF FP . . . PP	ΔQ	*V*	*sT*	*B*	*U*	*S*
1	0	1 0 . . . 0	−1.3	20.8	0.1	1	0	0

[a] 3-letter alphabet (see text).

(C)} based on similarities in physicochemical characteristics. Replacement of an amino acid with another from within the same subgroup is considered to be a conservative (*C*) substitution, otherwise the substitution is labeled nonconservative (*NC*).[31] The Boolean binary variables *C* and *NC* for an nsSNP yield two additional predictors. Hence, altogether there were 17 predictors that were considered for this study (Table I). An extensive analysis of the relationships between some of the predictors characterizing the SNPs in the *complete* set was previously undertaken and published elsewhere.[19]

Random forest machine learning and evaluation of classifier performance

The random forest (RF) algorithm relies on bagging (bootstrap aggregating) to generate multiple bootstrapped training sets of nsSNPs using the original set of da- and nt-SNPs, from which an ensemble of classification trees is trained and predictions are made via majority vote.[16] Additionally, rather than using all of the available predictors, RF selects a fixed-size random subset of these attributes to split at every node encountered in each of the growing trees. It has been shown that RF generally achieves better performance than other currently available learning schemes.[32–34] For this work, we utilized the RF implementation available as part of the Weka (Waikato Environment for Knowledge Analysis) suite of machine learning tools.[35] The algorithm parameters chosen included the generation of 100 trees for each experiment, and in cases where all 17 of the predictors were available, only five were randomly selected to split at every tree node.

We investigated classifier performance by applying 10-fold CV, randomized percentage split, and leave-one-out (jackknife) testing procedures. Using 10-fold CV, the dataset instances are each randomly assigned to one of 10 equally sized subsets. One subset is then held-out as the other nine subsets (90% of the original dataset) are combined to train an RF model, which is subsequently used to predict a class membership (disease-associated or neutral) for each of the nsSNPs in the held-out set. The process is repeated 10 times so that each subset serves once as a held-out set. In the end, 10 models are generated which together provide a single class prediction for every instance in the original dataset. Next, in a randomized percentage split, say 60/40 for example, 60% of the nsSNPs in the original dataset are used to train an RF model, which is subsequently used to generate a class prediction for the remaining 40% of nsSNPs in the original dataset that were not selected for training. In both testing procedures stratification is applied as the subsets are generated, ensuring that the proportions of da- and nt-SNPs in each subset are similar to those in the original dataset. Unlike these first two approaches, the leave-one-out method is deterministic because each nsSNP in the original dataset forms a subset consisting of a single instance. Leave-one-out can therefore be described as *N*-fold CV, where $N =$ size of the original dataset.

On the basis of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the test sets (where positive and negative refer to generic classes), we evaluated the procedures described earlier by using the Matthew's correlation coefficient (MCC),[36] 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 the balanced error rate (BER), defined as

$$\mathrm{BER} = \frac{1}{2} \times \left(\frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} + \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \right)$$

A χ^2 test can be applied to assess statistical significance of the MCC, where the test statistic is $\chi^2 = N \times \mathrm{MCC}^2$ with one degree of freedom.[37] An additional measure of classifier performance that is insensitive to class skew is the area (AUC) under the receiver operating characteristic (ROC) curve.[38] The ROC curve is a plot of the true positive rate = TP/(TP + FN) versus the false positive rate = FP/(FP + TN) in the unit square. By ranking the instances according their predicted probabilities for membership in the positive class, an ROC curve is obtained by counting the number of instances actually belonging to that class that rank above varied probability thresholds.

Table II
Frequency and Distribution of the Complete Set Based on SIFT

	daSNPs		ntSNPs		unSNPs	
	OK (%)	Low	OK (%)	Low	OK (%)	Low
Tolerant	173 (20)	NA	135 (75)	NA	144 (27)	NA
Intolerant	689 (80)	NA	45 (25)	NA	390 (73)	NA
Total	862 (100)	57	180 (100)	252	534 (100)	34

OK—confidence of prediction is acceptable by SIFT; Low, confidence of prediction is unacceptable.

Adaptive neuro-fuzzy inference system (ANFIS)

The ANFIS schema was constructed by using the Fuzzy Logic toolbox in Matlab.[20] ANFIS consists of a six layer feed-forward neural network: an input layer of three neurons, a fuzzification layer with six membership functions, a layer with nine fuzzy rules, a normalization layer with nine neurons, a defuzzification layer with nine neurons, and a final layer with a single summation neuron.[20] ANFIS complexity increases exponentially with the number of input layer neurons, and those considered here correspond to the ΔQ, V, and sT predictors obtained by applications of Delaunay tessellation.[19] The second fuzzification layer consists of generalized Gaussian activation functions of the form

$$y_i = \frac{1}{1 + \left[\left(\frac{x_i - c_i}{a_i}\right)\right]^{2b_i}}$$

for each neuron i, and the third layer corresponds to Sugeno-type fuzzy rules of the form

$$\text{IF} \quad x_1 \text{ is } a_1$$

$$\text{AND} \quad x_2 \text{ is } b_1$$

$$\text{THEN} \quad y = f_i = k_{i0} + k_{i1}x_1 + k_{i2}x_2$$

for each neuron i, where a_1, a_2, a_3 are fuzzy sets (membership functions) for input x_1, and b_1, b_2, b_3 are fuzzy sets (membership functions) for input x_2.[20,39,40] ANFIS uses a hybrid-learning algorithm that combines the least squares estimator and the gradient descent method and makes it possible to optimize antecedent and consequent parameters.[20] Unlike the RF statistical machine learning approach, fuzzy logic allows us to consider each SNP as an element of a fuzzy set belonging simultaneously to the disease and neutral classes, each with a degree of membership in the interval [0, 1]. Such an approach is especially useful when considering SNPs that participate in polygenic diseases, where there is a possibility that the SNPs can additively contribute to the onset of disease.

RESULTS

nsSNP dataset and predictor details

As described in the Methods section, our *complete* set of human nsSNPs accumulated for this study consists of 919 daSNPs, 432 ntSNPs, and 568 currently unclassified SNPs (unSNPs). We trained classifiers using the collection of da- and nt-SNPs, which were subsequently used to predict class memberships for each of the unSNPs. Each nsSNPs in the *complete* set was run through SIFT, which provided a prediction as to whether the associated protein is tolerant or intolerant to the particular residue substitution. The SIFT program generated a *low confidence* warning if the sequences used in making a SNP prediction were not diverse enough. Among the nsSNPs in the *complete* set, 82% (862 daSNPs, 180 ntSNPs, and 534 unSNPs) were acceptably predicted by SIFT without the *low confidence* warning. Additionally, 80% (689) of the daSNPs were intolerant, while 75% (135) of the ntSNPs were tolerant (Table II).

There is a statistically significant difference (determined by t-test) between the subsets of da- and nt-SNPs acceptably predicted by SIFT, based on values of the ΔQ, V, and sT predictors (Table III). This result was previously observed for the full subsets of da- and nt-SNPs in the *complete* set, regardless of SIFT confidence, and satisfies a necessary requirement for these parameters to be useful as predictors.[19] On the basis of a similar analysis using the tolerant and intolerant subsets of the da- and nt-SNPs acceptably predicted by SIFT, we again obtained a statistically significant difference between the intolerant daSNPs and the tolerant ntSNPs, while no such difference was observed between the tolerant daSNPs and the intolerant ntSNPs (data not shown). We define the *strict* set as the collection of 689 intolerant daSNPs, 135 tolerant ntSNPs, and 534 unSNPs acceptably predicted by SIFT, and classifiers trained with the *strict* set were compared with those of the *complete* set.

Performance of the random forest algorithm

Employing our full feature set of 17 predictors to characterize each of the da- and nt-SNPs in the *complete* and *strict* training sets, RF performance was evaluated based on 10-fold CV, 60/40 randomized split, and leave-one-

Table III
Mean Values of Statistical Potential and Computational Geometry Predictors for da- and nt-SNPs of the Complete Set Acceptably Predicted by SIFT

Subset	Count	ΔQ[a]	V[b]	sT[a]
daSNP	862	-0.76 ± 2.35	21.8 ± 11.2	0.17 ± 0.07
ntSNP	180	-0.01 ± 1.49	25.5 ± 20.1	0.25 ± 0.12

[a] $P < 0.001$.
[b] $P < 0.018$.

Table IV
Evaluation of RF Performance With the Full Feature Set (17 Predictors)

Set	Method	MCC	BER	AUC
C	Tenfold CV	0.436 ± 0.011	0.294 ± 0.007	0.797 ± 0.002
	60/40 Split	0.427 ± 0.032	0.301 ± 0.015	0.791 ± 0.019
	Jackknife	0.460	0.285	0.799
S	Tenfold CV	0.328 ± 0.017	0.374 ± 0.010	0.810 ± 0.009
	60/40 Split	0.328 ± 0.089	0.374 ± 0.036	0.811 ± 0.024
	Jackknife	0.347	0.367	0.802

Ten independent iterations were performed for the tenfold CV and 60/40 split procedures.
Sets: C = *complete;* S = *strict.*

Table V
Tenfold CV RF Performance With Subsets of the Statistical Potential and Computational Geometry Predictors

Set	Predictors	MCC	BER	AUC
C	ΔQ, sT	0.250	0.376	0.698
	ΔQ, V	0.337	0.340	0.726
	ΔQ, sT, V	0.369	0.323	0.754
S	ΔQ, sT	0.142	0.431	0.690
	ΔQ, V	0.166	0.430	0.686
	ΔQ, sT, V	0.307	0.365	0.780

Sets: C = *complete;* S = *strict.*

out CV (jackknife) using each set (Table IV). In all cases, MCC values were statistically significant ($P < 0.001$). Next, given the ability of the ΔQ, V, and sT predictors to effectively discriminate between daSNPs and ntSNPs in both the *complete*[19] and *strict* (Table III) sets, we investigated RF performance in the case where the SNPs in the associated training sets were characterized by 3D feature vectors consisting of only these attributes. For this analysis, splitting at every tree node during RF learning is based on the utilization of all three predictors rather than a random selection of a subset of features. On the basis of the performance results of 10-fold CV experiments (Table V), it is clear that RF classification using ΔQ, V, and sT is slightly less effective than using all 17 predictors (Table IV). On the other hand, RF performance is diminished substantially if fewer than all three predictors are used (Table V).

A comparison of results obtained for the *strict* and *complete* sets (Tables IV and V, Fig. 1) suggests that implicitly including evolutionary information through SIFT does not provide an additive, much less synergistic, effect on predictive performance. Factors contributing to a lack of improvement in *strict* set performance measures over those of the *complete* set include the significantly smaller size of the *strict* set, and the likelihood that evolutionary information is simply complementary to our attributes and does not provide an advantage for prediction. Hence, for this study, we concentrated on predicting a class membership for each of the 568 unSNPs in the *complete* set using two RF models trained with the da- and nt-SNPs of the *complete* set, based on employing all 17 predictors as well as using only three (ΔQ, V, and sT) attributes.

Ranking relative contributions of the individual predictors

Our focus on ΔQ, V, and sT as a fundamental subset of the 17 predictors was not only influenced by the statistically significant difference in mean values of these attributes among daSNPs versus ntSNPs (Table III), but also supported by results obtained from a search of the most influential attributes characterizing the *complete* and *strict* training sets. In particular, using the *CfsSubsetEval* attribute evaluator program in Weka,[35] subsets of the 17 predictors were evaluated for how highly correlated they were with the class (neutral or disease-associated) while also displaying low intercorrelation with one another. Subsets were selected for evaluation based on a greedy hill-climbing approach using the *BestFirst* search method program in Weka, starting with a random selection of attributes and following a bidirectional search in which all possible additions or deletions of single attributes are examined at each step.[35] The procedure was augmented with backtracking, whereby a maximum of five consecutive, nonimproving attributes were allowed. The results confirmed that ΔQ, V, and sT were consistently the most highly ranked attributes (Table VI).

Comparison of RF with ANFIS, SIFT, and PolyPhen

Next, ANFIS was applied to the *complete* set of da- and nt-SNPs in conjunction with a 10-fold CV testing

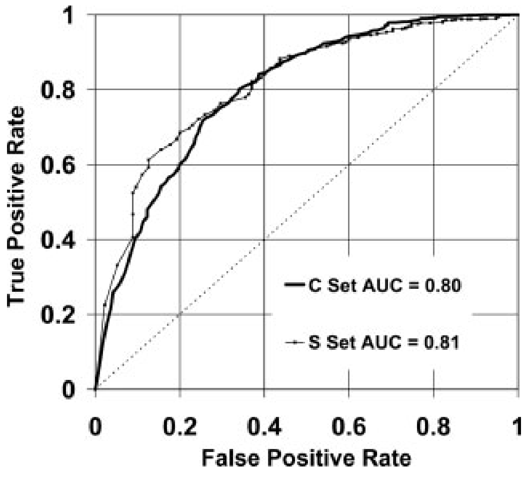

Figure 1
Tenfold cross-validation ROC curves based on random forest (RF) learning with all 17 attributes. Sets: C = complete; S = strict.

Table VI
The Most Influential Predictors and Their Rankings

	Ranked attributes			
Set	1	2	3	4
C	ΔQ	V	sT	
S	ΔQ	V	sT	S – surface

Sets: C = *complete;* S = *strict.*

procedure. Performance of ANFIS was measured by computing the area (AUC) under the corresponding ROC curve. The ANFIS system, which utilizes three predictors (ΔQ, V, and sT) as input layer neurons, yielded an AUC that was equivalent to that obtained by RF with three predictors, but did not perform as well as RF based on all 17 attributes (Fig. 2).

An overall summary comparing our performance results with those of other well-established techniques is provided in Table VII. As with our study, Bao and Cui[17] also applied the RF algorithm in conjunction with 10-fold CV; however, they had an added advantage given their use of 1000 trees in conjunction with RF learning (compared with our use of only 100 trees). Since their SNP dataset was also significantly larger than ours, it would be expected that if both feature sets were also equally informative, then their 10-fold CV performance measures would exceed ours. However, using our respective training sets, MCC, BER, and AUC values based on their structural and evolutionary predictors were lower than those based on our statistical potential and computational geometry attributes, suggesting that our predictors are more effective. To provide a more direct comparison, we identified 819 SNPs in our *complete* set that were also used in the training set of 4013 SNPs utilized by Bao and Cui.[17] Application of our RF algorithm and a 10-fold CV procedure on this *overlap* set, which makes up only 20% of the full Bao and Cui training set, resulted in AUC performance measures of 0.77 (based on all 17 predictors) and 0.75 (based only on the ΔQ, V, and sT predictors). Although Bao and Cui[17] did not provide an explicit AUC value in their manuscript regarding their RF and 10-fold CV procedure applied to their full training set, an estimate of 0.75 was calculated by carefully examining the associated ROC curve provided in Figure 1 of the article. Next, SIFT was used to generate predictions for the da- and nt-SNPs in the *complete* set in a direct comparison of methods[9]; however, SIFT performance measures were lower than those based on RF, regardless of whether RF employed all 17 predictors or simply the three most highly ranked attributes ΔQ, V, and sT (Fig. 2 and Table VII).

Finally in the case of PolyPhen, we began with the set of 1276 experimentally validated nsSNPs in their training set.[10] Next, we only retained those SNPs that were predicted by PolyPhen to be either benign or probably damaging, while those that were predicted as unknown or possibly damaging was discarded. From this set, we eliminated all SNPs that were not associated with a protein structure. This subset was further reduced by removing unSNPs and considering only those already assigned as disease-associated or neutral. Only 57 of these remaining da- and nt-SNPs had associated 3D structures that were tessellatable, and a direct comparison revealed that our RF classifier is comparable to PolyPhen (Table VII).

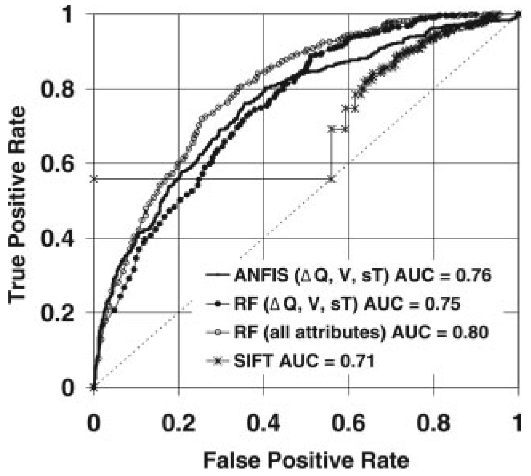

Figure 2
Comparison of tenfold cross-validation ROC curves obtained by using the disease-associated and neutral nsSNPs in the complete dataset, based on random forest (RF) learning with 17 attributes, RF learning with 3 attributes, ANFIS with 3 attributes, and SIFT.

DISCUSSION

Structure-based rules are well adapted to determine the change in stability of a protein upon mutation, and they are also able to account for a change in functionally important sites.[28] However, evolutionary rules are better at revealing the functional sites. It has been demonstrated that the application of a four-body, knowledge-based statistical potential obtained by means of Delaunay tessellation correlates well with the stability changes in

Table VII
Comparison of RF With Previously Published Methods

Method (n = daSNPs + ntSNPs)	MCC	BER	AUC
RF – *Complete* set (n = 1351)	0.436	0.294	0.797
RF – Bao and Cui (n = 4013)	0.315	0.292	—
SIFT (n = 1351)	0.190	0.413	0.708
RF – *Complete* set (n = 57)	0.340	0.263	0.771
PolyPhen (n = 57)	0.302	0.283	—

mutants.[27] Additionally, the alpha shape methodology has established the correlation of residue position location, defined by computational geometry, with SNP phenotype.[18] Hence, these studies support our inclusion of statistical potential (ΔQ) and topological location (S, U, B) attributes in the feature set. Loss of protein structure stability has been suggested to be a major cause of monogenic disease.[41] We hypothesize that our predictors are able to distinguish between disease-associated and neutral SNPs by empirically capturing changes in protein stability because of mutation. The effectiveness of additional computational geometry parameters, such as volume (V) and tetrahedrality (sT), may also result from a correlation with stability change, rather than by identification of functional sites.

Implicit relationships between predictors may also contribute to classification performance. For example, the mean ΔQ score of buried substitutions with hydrophobic to hydrophilic changes in the "complete" set is -1.97428 ± 2.7508, while that of all other buried substitutions is -0.1789 ± 1.9941. Application of a t-test revealed a statistically significant difference between these means ($P = 5.75 \times 10^{-16}$). Additionally, the mean ΔQ score of all 725 deleterious buried substitutions is -0.7805 ± 2.4576, while that of the remaining 301 nondeleterious buried substitutions is 0.1130 ± 1.5592. Again, application of a t-test revealed a statistically significant difference between these means ($P = 3.03 \times 10^{-12}$).

Our data shows that the SIFT program was not able to determine the impact of an nsSNP on the corresponding protein with high enough confidence for ~20% of nsSNPs in our *complete* set of tessellated proteins. Identification of these nsSNPs as either disease-associated or neutral could be predicted with the help of tools that utilize statistical potential and computational geometry attributes. In fact, these attributes could provide an additional level of prediction confidence in general by compensating for multiple sequence alignment errors that arise due to the use of databases tainted with paralogous and polymorphic sequences.

Statistical geometry-based attributes derived from Delaunay tessellation are orthogonal to the evolutionary attributes utilized by approaches such as SIFT and PolyPhen. By combining our statistical geometry attributes together with the types of substitutions and their spatial locations, in combination with the Random Forest or ANFIS algorithms, we have developed a unique method that is complementary to these currently available techniques. In the situation where the current methods provide results with high confidence, our approach can be used to further support and confirm their results. On the other hand, in cases where substitutions are not predicted with high confidence for SNP classification by the current methods due to e.g. the lack of a sufficient number of orthologous proteins, our approach provides a necessary and reliable alternative.

Every method has its limitations. Possible lack of orthologous proteins is a major limitation for evolutionary-based methods. Application of our method is limited by the fact that structures are not available for all wild-type proteins. However, as the number of crystal structures continues to increase, the precision of this approach will be enhanced and its coverage will be extended. Another approach to overcome this limitation would be to use 3D modeled structures of the proteins of interest.

Bao and Cui have similarly utilized an implementation of the RF algorithm for classification of nsSNPs, working with a larger combined dataset of da- and nt-SNPs consisting of 4013 total samples.[17] In contrast to our statistical potential and computational geometry attributes, their predictors included structural environment parameters,[42] secondary structure assignments based on the STRIDE program,[43] and tolerance to substitutions based on SIFT scores.[9] Tenfold CV applied to their training set generated performance measures of MCC = 0.315 and BER = 0.292 (Table VII). A direct comparison with our results is difficult since the combined da- and nt-SNPs in our *complete* dataset consist of only 1351 samples, and the significantly smaller size of our dataset has the potential to hinder performance. However, based on 10 iterations of 10-fold CV using RF machine learning, we achieved a significantly higher mean MCC = 0.436 and an essentially identical mean BER = 0.294 (Tables IV and VII) by including all 17 attributes. The clear importance of the three most highly ranked attributes ΔQ, V, and sT alone was elucidated by corresponding performance measures of MCC = 0.369 and BER = 0.323 (Table V).

For validation, Bao and Cui used an RF model trained with the set of 4013 samples to classify 205 samples contained in an independent test set, achieving MCC = 0.352 and BER = 0.270.[17] Analogously, we performed 10 iterations of a randomized 60/40 split of the da- and nt-SNPs in the *complete* set, using all 17 attributes, and yielding mean MCC = 0.427 and mean BER = 0.301 (Table IV). Based upon these detailed 10-fold CV and independent test set validation comparisons, RF models built upon rules derived from our statistical potential and computational geometry attributes clearly yield important tools for the prediction of disease alleles in the case of nsSNPs.

Neutral SNPs might be as yet unidentified contributors to diseases.[15] Moreover, we are currently unable to define precise rules for each possible situation, since a particular SNP could belong to both the disease and neutral classes to some degree, depending on the presence of other SNPs in the genome and external environmental factors. Thus, the fuzzy logic approach of ANFIS seems quite appropriate, allowing us to assess the disease potential of nsSNPs and to select the most promising nsSNPs for further investigation.

ANFIS is able to train membership functions by itself as well as create its own fuzzy rules. A major limitation of ANFIS is that its implementation with all available attributes is not computationally feasible. Having 17 attributes in the input vector would require having 3^{17} membership functions. In our experiments with a truncated 3D input vector consisting of only the ΔQ, V, and sT predictors, performance of the ANFIS system was essentially equivalent to that of RF with the same attributes (Fig. 2). In any case, if the data set is large enough, it is quite appropriate to use ANFIS for fine-tuning of membership functions and fuzzy rules.[20]

CONCLUSIONS

Currently, experimental and structural data on nsSNPs contributing to complex diseases is very scarce. However, *ab initio* predicted structures may compensate for the lack of experimental and homologous structures and make it possible to explore nsSNPs known to contribute to complex diseases.

Predictors based on computational geometry and statistical potential parameters are orthogonal to those used by other researchers. It has been demonstrated that trained RF models, using relatively few attributes, are at least as effective as other well-established methods for predicting nsSNP phenotype. The RF and ANFIS models are competitive in their predictions; however, ANFIS is based on a rather different conceptual approach. While RF classification is binary in nature and provides a more robust, computationally effective, and versatile method that is not dependent on the number of attributes, ANFIS is capable of explicitly accounting for the degree of class membership.

REFERENCES

1. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177–182.
2. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–1517.
3. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307:683–706.
4. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19:2199–2209.
5. Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 2004;20 (Suppl 1):I63–I68.
6. Needham CJ, Bradford JR, Bulpitt AJ, Care MA, Westhead DR. Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. BMC Bioinformatics 2006;7:405.
7. Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, Stinson J, Forrest WF, Bazan JF, Seshagiri S, Zhang Z. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 2007;67:465–473.
8. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 2002; 322:891–901.
9. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–3814.
10. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–3900.
11. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
12. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: be mindful of your training data! Bioinformatics 2007; 23:664–672.
13. Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics 2004;5:120.
14. Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics 2006;7:217.
15. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol 2006;356:1263–1274.
16. Breiman L. Random forests. Mach Learn 2001;45:5–32.
17. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 2005;21:2185–2190.
18. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. Structural location of disease-associated single-nucleotide polymorphisms. J Mol Biol 2003;327:1021–1030.
19. Barenboim M, Jamison DC, Vaisman II. Statistical geometry approach to the study of functional effects of human nonsynonymous SNPs. Hum Mutat 2005;26:471–476.
20. Jang JSR. Anfis: adaptive network-based fuzzy inference system. IEEE Trans Syst, Man Cybernetics 1993;23:665–685.
21. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 2004;23:464–470.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
23. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–221.
24. Vaisman II, Tropsha A, Zheng W. Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis. Proc IEEE Symp Intelligence Syst 1998: 163–168.
25. Barber CB, Dobkin DP, Huhdanpaa HT. The quickhull algorithm for convex hulls. ACM Trans Math Software 1996;22:469–483.
26. Masso M, Lu Z, Vaisman II. Computational mutagenesis studies of protein structure-function correlations. Proteins 2006;64:234–245.
27. Carter CW, Jr, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001; 311:625–638.
28. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–270.
29. Masso M, Vaisman II. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. Biochem Biophys Res Commun 2003;305:322–326.
30. Medvedev NN, Voloshin VP, Naberukhin Y. Structure of simple liquids as a percolation problem on the Voronoi network. J Phys A: Math Gen 1988;21:L247–L252.
31. Dayhoff MO, Schwartz RM, Orcut BC, editors. A model for evolutionary change in proteins, Vol. 5. Washington D.C.: National Biomedical Research Foundation; 1978. pp 345–352.
32. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003;19:1636–1643.

33. Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. Proc Natl Acad Sci USA 2003;100:9608–9613.
34. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003; 43:1947–1958.
35. Witten IH, Frank E. Data mining. San Francisco: Morgan Kaufmann, Elsevier; 2005.
36. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405:442–451.
37. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–424.
38. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.
39. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst, Man, Cybernetics 1985;15:116–132.
40. Negnevitsky M. Artificial intelligence: a guide to intelligent systems. Harlow, England: Addison Wesley; 2005.
41. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 2005;353:459–473.
42. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
43. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.