doi:10.1016/S0022-2836(02)00813-6 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 322, 891–901

JMB

Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction

Christopher T. Saunders[1] and David Baker[2,3]*

[1]*Department of Genome Sciences, Box 357730 University of Washington Seattle, WA 98195, USA*

[2]*Department of Biochemistry Box 357350, University of Washington, Seattle WA 98195, USA*

[3]*Howard Hughes Medical Institute, Seattle, WA 98195 USA*

**Corresponding author*

Methods for automated prediction of deleterious protein mutations have utilized both structural and evolutionary information but the relative contribution of these two factors remains unclear. To address this, we have used a variety of structural and evolutionary features to create simple deleterious mutation models that have been tested on both experimental mutagenesis and human allele data. We find that the most accurate predictions are obtained using a solvent-accessibility term, the C^{β} density, and a score derived from homologous sequences, SIFT. A classification tree using these two features has a cross-validated prediction error of 20.5% on an experimental mutagenesis test set when the prior probability for deleterious and neutral cases is equal, whereas this prediction error is 28.8% and 22.2% using either the C^{β} density or SIFT alone. The improvement imparted by structure increases when fewer homologs are available: when restricted to three homologs the prediction error improves from 26.9% using SIFT alone to 22.4% using SIFT and the C^{β} density, or 24.8% using SIFT and a noisy C^{β} density term approximating the inaccuracy of *ab initio* structures modeled by the Rosetta method. We conclude that methods for deleterious mutation prediction should include structural information when fewer than five to ten homologs are available, and that *ab initio* predicted structures may soon be useful in such cases when high-resolution structures are unavailable.

Keywords: deleterious mutation prediction; protein mutagenesis; human disease allele; protein evolution; protein structure prediction

Introduction

The rapid discovery of single nucleotide polymorphisms (SNP) in the human genome has created an opportunity for high-throughput deleterious mutation prediction methods to discover and prioritize candidate human disease alleles from the pool of uncharacterized non-synonymous coding SNPs. Several methods have recently been developed specifically for this purpose,[1–3] which utilize information from both evolution and protein structure. In principle, a protein mutation could be deleterious either because it destabilizes structure or it disrupts a functional site involved in catalysis, ligand-binding, or interaction with another protein. For this reason, one might expect evolutionary and structural information from proteins to complement one another: structural information should help to identify destabilizing mutations, while highly conserved positions in multiple sequence alignments (msa) can help to identify functional sites.[4] While this principle is qualitatively clear, the relative importance and complementarity of these feature domains has not been well characterized with regard to computational prediction of intolerant mutations. Such characterization is important both for improving the accuracy of predictive models and for determining the relative error of predictions made when sequence or structural information is reduced or absent, as is often the case when few sequence homologs exist or only a coarse approximation to the protein structure is available.

In this study, we investigate the relative strength of a variety of structural and evolutionary features described in previous work and examine how the most effective features complement one another in simple classification models. We characterize the performance of classifiers as a function of the number of homologs available for the calculation

Abbreviations used: msa, multiple sequence alignment; PSSM, position-specific scoring matrix; SNP, single nucleotide polymorphism.

E-mail address of the corresponding author: dabaker@u.washington.edu

Table 1. Number of observations in mutation test sets

	Laboratory mutations[a]		Human alleles[b]	
Source	Deleterious	Neutral	Disease[c]	Neutral
All	1500	3706	191	87
LacI	1166	2255		
HIV-1	159	111		
T4	175	1340		

[a] Laboratory mutagenesis data were taken from two comprehensive amber suppressor assays made on Lac repressor[5] and T4 lysozyme,[7] as well as from a saturation random mutagenesis assay conducted on HIV-1 protease.[6] Only the strongest categories of deleterious and neutral mutations were compiled from each of these experiments into our test set. For the T4 lysozyme case, only observations made at 25 °C were used.

[b] To compile a test set of human disease alleles, all allele annotations on human disease associated proteins were taken from both Online Mendelian Inheritance in Man† and Swiss-Prot 40.[8] This initial disease set was culled to include only those alleles annotated in both databases and, as a final step, any alleles described as neutral polymorphic markers or somatic disease mutations were removed. The test set of human neutral alleles was compiled from naturally occurring alleles recorded in Swiss-Prot for proteins without any disease annotation. From this initial set, all alleles annotated with a speculative disease association or a known reduction of protein function were removed. Both the disease and neutral allele sets were restricted to those whose protein had a homolog of known structure with at least 3.0 Å resolution and sharing at least 40% sequence identity with the test set protein sequence.

[c] For all calculations in this study, disease alleles are treated exactly like deleterious mutations, we maintain the separate naming convention throughout only to emphasize the inherent difference between these two test sets.

of evolutionary features, which indicates how the reliability of deleterious mutation prediction is affected by the common problem of having few homologous sequences available. Finally, we explore the possibility of improving classification for cases where no experimental structure is available, by using structures generated from *ab initio* prediction methods.

Results

Mutation test sets

The predictive power of individual structure and sequence-based features was studied using two sets of mutation data compiled to test the methods developed in this study (Table 1). The first of these test sets consists of data from laboratory mutagenesis experiments. While there is a significant amount of targeted mutagenesis data available, such data often tend to be biased towards particular structural sites, residue types, or other features of interest to experimentalists. Due to this biased context, these data cannot be used to train deleterious mutation models without reducing the ability of the model to handle a wide range of mutations from many structural contexts, an important quality of any model designed to analyze coding SNPs on a genomic scale. A useful alternative to targeted mutagenesis data are data from experimental studies that probe nearly all mutations at all residues over entire proteins, where both tolerant and intolerant phenotypes can be observed. Following the suggestion made by Ng & Henikoff,[1] we have constructed a test set composed of such unbiased laboratory mutagenesis data derived from experiments on three proteins: Lac repressor,[5] human immunodeficiency virus type 1 (HIV-1) protease,[6] and bacteriophage T4 lysozyme.[7] The second set of test mutations we have constructed is composed of naturally occurring human neutral and disease alleles collected from mutations in the Swiss-Prot[8] and Online Mendelian Inheritance in Man† databases. It is apparent from Table 1 that the human allele test set is relatively undersampled, and we expect that the neutral allele set is likely to contain a number of deleterious mutations whose disease association has not yet been resolved. Therefore, the human test data are not expected to reflect the absolute accuracy of prediction on human disease alleles, but rather to confirm the relative accuracy and generality of the various classification techniques explored in this study.

Classification using individual features

The performance of individual features applied to the classification of mutants from both test sets is summarized in Table 2. For each feature, classification was performed by 20-fold cross-validation, where an optimal classification threshold was chosen for each training set and the final result reported as a balanced error term, representing the classification error when the prior probability of deleterious and neutral cases is equal. This simple feature evaluation method has been chosen over a more rigorous form of statistical hypothesis testing, such as evaluating features by the ANOVA *F*-test,[9] because we have no reason to assume that the feature distributions will conform to any particular model.

The simplest of the sequence-based features considered is a measure of the residue substitution cost, which uses the log likelihood ratio from the Blosum62 substitution table for each mutation.[10] With balanced classification errors of 40.3% and 40.1% on the laboratory mutagenesis and human allele test data, respectively, this feature is not found to be especially informative, as it uses no homolog information and simply reflects the properties of the two residues undergoing exchange. Therefore, this classification error represents an upper bound to the error expected for

† OMIM, Online Mendelian Inheritance in Man. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000, http://www.ncbi.nlm.nih.gov/omim/

sequence-based features when no homologs are available.

The entropy of a residue in a multiple sequence alignment is often used as a measure of the overall mutability of a specific site in a protein, and has appeared as a feature in previous studies of mutation prediction either directly or normalized over each protein chain.[9,11] Here, the normalized site entropy is found to be more useful for the prediction of laboratory mutants than Blosum62, with a balanced classification error on laboratory mutations of 29.0% (Table 2). This entropy score is insensitive to the actual residue substitution taking place for any given mutation and is thus orthogonal to the information in the Blosum62 score, suggesting the use of a score that combines these two factors of substitution and site-conservation.

One such score is generated by SIFT,[1] a method designed to predict deleterious mutations by creating a specialized position specific scoring matrix (PSSM) from the msa of the protein of interest and using this PSSM to derive a prediction score. A PSSM-based method similar to SIFT also appears as one component of the scoring function used by Sunyaev *et al*. for predicting intolerant mutations.[2] These methods integrate the substitution and site mutability information into a single score, allow for the calculation of consistent scores when few homologs are available, and model the tolerance of residues that belong to the same biochemical class as those already observed in the alignment. These advantages explain why the SIFT score has the lowest balanced error among sequence-based features of 22.2% for laboratory mutations and 37.4% for human alleles (Table 2). A PSSM-based score such as SIFT integrates the information contained in the two simpler sequence features previously discussed, therefore the SIFT score is the only sequence-based feature used in further prediction models for this study.

The structural features presented in Table 2 include two similar terms that reflect the structural burial of the residue at the mutation site. The first is the number of C^{β} atoms within 10 Å of the C^{β} atom of the site residue, which we refer to as the C^{β} density, and the second is the percentage of the total solvent-accessible surface area for the native residue at the mutation site, calculated using DSSP.[12] Although the solvent-accessible surface area has been used directly or indirectly as a residue burial term in previous mutation prediction methods,[2,9] we have found the C^{β} density to have similar classification performance and expect this term to be more robust to the poorly defined side-chain conformations of approximated structures, such as those modeled by *ab initio* methods. For these reasons, we elected to use the C^{β} density as the residue burial term in our classification models.

The flexibility of the side-chain at any site should be correlated with the likelihood that a residue substitution at that site is tolerated. One convenient proxy for side-chain flexibility that has been used

Table 2. Classification of test data using individual features

	% Classification error		
	Deleterious	Neutral	Balanced[a]
A. *Laboratory mutations*			
Sequence features			
Blosum62	37.8	42.8	40.3 ± 1.33
Normalized site entropy	32.8	25.1	29.0 ± 1.23
SIFT	19.9	24.4	22.2 ± 1.13
Structural features			
C^{β} density	31.1	26.6	28.8 ± 1.23
% Solvent-acc. surface area	17.4	41.9	29.6 ± 1.24
Normalized *B*-factor	23.1	48.7	35.9 ± 1.30
Normalized *B*-factor (edited)[b]	19.7	41.6	30.6 ± 1.32
Sunyaev structural rules	63.2	14.3	38.8 ± 1.32

	% Classification error		
	Disease	Neutral	Balanced
B. *Human alleles*			
Sequence features			
Blosum62	43.5	36.8	40.1 ± 5.76
Normalized site entropy	32.5	51.7	42.1 ± 5.80
SIFT	34.6	40.2	37.4 ± 5.69
Structural features			
C^{β} density	28.3	37.9	33.1 ± 5.53
% Solvent-acc. surface area	34.6	24.1	29.3 ± 5.35
Normalized *B*-factor	39.3	44.8	42.0 ± 5.80
Sunyaev structural rules	64.9	13.8	39.4 ± 5.74

Classification using single features was performed by 20-fold cross-validation where a threshold was found between deleterious and neutral cases that minimized the balanced classification error of each training set and the classification error was found from the total result of 20 rounds of test set classifications. The exception to this procedure is the Sunyeav structural rule set, in which case thresholds are specified in the rule definition, making any training procedure unnecessary.

[a] The balanced error is the classification error of cases that are weighted to normalize both the influence of deleterious and neutral cases as well as the influence of proteins that contribute to the laboratory mutagenesis test set. In this scheme, cases are first weighted to give equal total weight to each of the three test proteins for the laboratory mutagenesis test set and then, for both test sets, weights are transformed to give equal total weight to deleterious and neutral cases. We note that when using this scheme, the expectation value of the balanced prediction error for random classification is always 50%, and that the balancing of cases from the three proteins used in the laboratory mutagenesis test set is only approximate, due to the subsequent balancing of deleterious and neutral cases. The confidence interval reported for the balanced classification error is found by approximating the total of all test set classifications, either from multiple rounds of cross-validation or from a separate test set, as a binomial distribution with a probability p of producing an erroneous prediction, where p is estimated as x/n and x is the number of errors made out of a total of n predictions from the test set. We report the 95% ($\alpha = 0.05$) confidence interval, where the approximate $100(1-\alpha)\%$ confidence interval for the value of p is:

$$(x/n) \pm z_{\alpha/2}\sqrt{\frac{(x/n)(1-x/n)}{n}}$$

[b] The edited version of the normalized *B*-factor shows the results of removing the DNA-binding domain of Lac repressor from both feature normalization and classification.

in mutation prediction[9] is the crystallographic *B*-factor. We have tested the *B*-factor as a predictive feature and found classification errors of 35.9% and 42.0% on our laboratory and human test sets, respectively, which indicate that this feature is at least informative for mutation prediction. However, the contribution to the *B*-factor from static disorder and large-scale flexibility reduces its usefulness as a site-specific probe of flexibility. This problem is illustrated by the DNA-binding domain of Lac repressor, in which case the *B*-factor scores of all residues in the binding domain exceed the values of all remaining residues in the chain. When this problematic domain is removed, the classification error of laboratory mutations using the *B*-factor is reduced notably from 35.9% to 30.6% (Table 2), but such specialized feature editing is not amenable to automated prediction methods. Due to the shortcomings of the *B*-factor, we explored an alternative representation of residue flexibility, the side-chain entropy, which was used successfully by Voigt *et al*.[11] as a means to pick tolerant substitutions for directed mutagenesis experiments. Unlike the *B*-factor, the side-chain entropy represents only the conformational flexibility of an individual side-chain. We tested classification using a measure of the side-chain entropy calculated at each mutation site from the probability distribution of rotamers for the native side-chain in a force-field composed of Leonard-Jones interactions, an orientation-dependent hydrogen bond term, and an implicit solvation model (T. Kortemme & D.B., unpublished results), but this term did not yield results significantly different from those obtained with the edited *B*-factor (data not shown).

The final feature included in Table 2 is our reproduction of four structural rules that form a component of the prediction method used by Sunyaev *et al*.[2] Here, each rule is intended to identify a different class of deleterious mutation on the basis of expert knowledge, an approach similar to that taken by Wang & Moult.[3] Specifically, these rules detect hydrophobic core disruption, buried charge change, mutation to proline in an α-helix and hydrophobicity change at an exposed site (see Methods). We interpreted these structural rules as a simple classifier by predicting a mutation to be deleterious when the deleterious conditions of any of the four individual rules were met, and subsequently found the prediction error of this classifier for our test sets (Table 2), finding that its performance is comparable to that of the crystallographic *B*-factor.

To further evaluate the Sunyeav rule set, the performance of each rule is summarized individually in Table 3, showing the number of mutations that pass each rule (and thus are predicted deleterious) along with the percentage of those that actually are deleterious, and the ability of the rule to select deleterious mutations relative to random selection. The most interesting result is the particularly poor performance of the solubility change rule, which is meant to detect mutations effecting significant hydrophobicity changes on the protein surface. While anecdotal observations may suggest such a rule, it does not appear to generalize as a predictor of intolerant mutations for the laboratory mutagenesis test set, although it is possible that this rule selects mutations that lead to a reduction in fitness too subtle to be detected in the assays used to generate these test data.

Table 3. Evaluation of structural rules for deleterious mutations, due to Sunyaev *et al*.

Deleterious mutation rule	No. mutations selected	% Deleterious from selected	Log likelihood deleterious[a]
A. *Laboratory mutations*			
Hydrophobic core disruption	531	64.9	1.2
Buried charge change	14	56.0	1.0
Solubility change	21	8.0	−1.8
Proline in α-helix	94	50.5	0.8
Any rule met	621	50.2	0.8

Deleterious mutation rule	No. alleles selected	% Disease from selected	Log likelihood deleterious
A. *Laboratory mutations*			
B. *Human alleles*			
Hydrophobic core disruption	57	87.7	0.4
Buried charge change	2	100.0	0.5
Solubility change	2	50.0	−0.5
Proline in α-helix	10	83.3	0.3
Any rule met	67	84.8	0.3

[a] The log odds of a selected mutation to be deleterious *versus* random selection:

$$\log_2\left(\frac{P(\text{deleterious}|\text{selected})}{P(\text{deleterious})}\right)$$

Classification using multiple features

Simple classification models were created using the SIFT score together with selected structural terms to characterize how these features complement one another in the prediction of deleterious mutations. We have primarily relied on classification trees for this purpose, because these methods are robust to noisy data, assume no model, and are sufficiently transparent to be described as a short sequence of splitting rules. Despite this simplicity, we have found the performance of trees to be comparable to that of classification using logistic regression models. While

Table 4. Classification of test data using multiple features

	Classification tree[a]			Logistic regression classifier[b]		
	% Classification error			% Classification error		
Classifier features	Deleterious	Neutral	Balanced[c]	Deleterious	Neutral	Balanced
A. *Laboratory mutations*						
SIFT	19.9	24.4	22.2 ± 1.13	25.7	21.2	23.4 ± 1.15
SIFT + C^{β} density	18.0	22.9	20.5 ± 1.10	19.7	22.2	20.9 ± 1.10
SIFT + normalized *B*-factor	21.9	23.2	22.6 ± 1.14	25.1	21.4	23.2 ± 1.15
SIFT + Sunyaev structural rules	19.9	25.5	22.7 ± 1.14	22.5	21.9	22.2 ± 1.13
SIFT + C^{β} density + normalized *B*-factor	18.0	23.2	20.6 ± 1.10	19.5	22.1	20.8 ± 1.10
	Classification tree			Logistic regression classifier		
	% Classification error			% Classification error		
Classifier features	Disease	Neutral	Balanced	Disease	Neutral	Balanced
B. *Human alleles*						
SIFT	34.6	40.2	37.4 ± 5.69	37.7	35.2	36.4 ± 5.13
SIFT + C^{β} density	29.6	29.7	29.6 ± 4.87	28.3	35.2	31.8 ± 4.96
SIFT + normalized *B*-factor	32.4	40.7	36.5 ± 5.13	35.2	34.1	34.6 ± 5.07
SIFT + Sunyaev structural rules	39.3	41.8	40.5 ± 5.23	35.6	35.2	35.4 ± 5.10
SIFT + C^{β} density + normalized *B*-factor	27.1	34.1	30.6 ± 4.91	28.3	35.2	31.8 ± 4.96

[a] Classification trees[24] were implemented using the RPART[25] package v3.0-2 in R v1.3.0, with case weights set as described for Table 2 and splitting performed using the Gini criterion. The reported classification performance is the result of training and testing classification trees over 20 rounds of cross-validation. The result of threshold classification using only SIFT is included from Table 2 for comparison with the classification tree results.

[b] Logistic regression models were implemented using the GLM module in R v1.3.0, with case weights set as described for Table 2. All logistic regression models incorporating SIFT were trained and scored using the log transformation of this feature (SIFT values were truncated to a lower bound of 0.001 prior to taking the log). The trained logistic regression models were interpreted as classifiers by categorizing the predicted deleterious probability as greater or less than 0.5.

[c] See Table 2 for classification error calculation methods.

logistic regression models would provide the significant advantage of probability values for test cases, we consider this property to be less important than simplicity and transparency for the exploratory analysis performed in this study, and have therefore chosen to base our analysis on classification tree results.

The cross-validated errors of tree classification using a variety of features indicate that the combination of SIFT with the C^{β} density is the most accurate of the feature sets considered (Table 4), with a balanced error of 20.5% and 29.6% on laboratory and human allele test data, respectively. To further study the generality of these predictions, separate classification trees were trained on all laboratory mutagenesis data and tested for their ability to classify human alleles. These results are summarized in Table 5, in which the combination of SIFT and the C^{β} density still performs quite well, although we note this with a degree of caution due to the low significance of the human allele results. Inspection of the consensus classification tree that uses these two features reveals that the trained classifier is actually implementing a very simple heuristic in which all mutations at a buried site are considered deleterious, all mutations on the protein surface are considered tolerated, and mutations in the range of intermediate burial (having a C^{β} density value between 12 and 22) are classified according to the SIFT score (Figure 1). It is evident from the simplicity of this model that potentially informative sequence information has

Table 5. Classification of human alleles by training on laboratory mutagenesis data

	Classification tree			Logistic regression classifier		
	% Classification error			% Classification error		
Classifier features	Disease	Neutral	Balanced	Disease	Neutral	Balanced
SIFT	29.6	38.5	34.0 ± 5.05	44.1	31.9	38.0 ± 5.17
SIFT + C^{β} density	26.7	35.2	30.9 ± 4.93	32.4	27.5	29.9 ± 4.88
SIFT + normalized *B*-factor	38.9	33.0	35.9 ± 5.11	42.1	31.9	37.0 ± 5.15
SIFT + Sunyaev structural rules	38.9	33.0	35.9 ± 5.11	42.5	27.5	35.0 ± 5.08
SIFT + C^{β} density + normalized *B*-factor	25.9	35.2	30.5 ± 4.91	32.8	25.3	29.0 ± 4.84

Classifiers were trained from the complete laboratory mutagenesis test set and used to predict the mutations in our human allele test set. Classification methods are the same as those described for Table 4, except that the cross-validation procedure was not used.

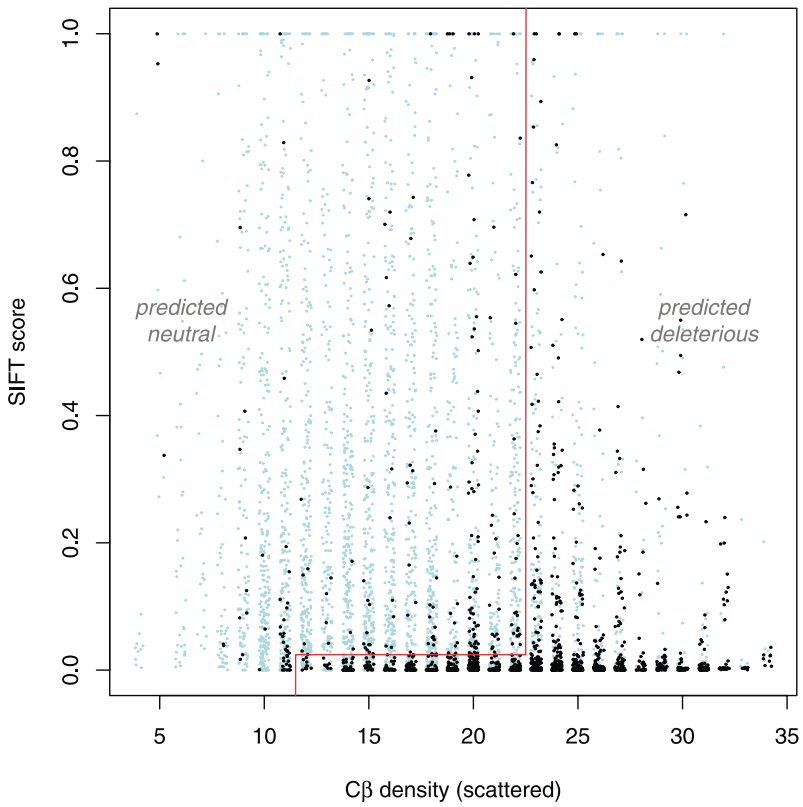

Figure 1. Classification tree trained on laboratory mutagenesis test set. Both neutral mutations (cyan) and deleterious mutations (black) from the laboratory mutagenesis test set are plotted according to the SIFT score and the C^{β} density. Superimposed in the foreground is a representation of the decision boundary (red) for the consensus classification tree resulting from cross-validated training on these mutation data. Note that the C^{β} density has been randomly scattered by ± 0.25 for visual clarity.

been ignored at sites that are sufficiently buried or exposed, suggesting the potential advantage of a linear model. However, we have tested one such model with logistic regression classification and have not found the results to be significantly different when using the same feature set (Table 4).

The dependence of the sequence and structure feature balance on msa depth was addressed by examining the relative contributions of features to the accuracy of deleterious mutation prediction as the number of available sequence homologs was varied. We tested classification using a restricted alignment of three homologs for the laboratory test set proteins (see Methods), finding that the average balanced error was 26.9% using SIFT alone and 22.4% using the SIFT/C^{β} density combination, demonstrating the increased importance of structural information when few homologous sequences are available.

To examine how homolog count influences the sequence-structure relationship in more detail, we classified the mutations in Lac repressor at all possible restricted alignment depths. Lac repressor was chosen for this purpose because it has a sufficiently deep alignment of 21 homologs and has the greatest number of mutations among the three proteins in the laboratory mutation test set. The SIFT score was calculated from randomly selected subsets of homologs for each alignment depth and used for classification of Lac repressor mutants both alone and in combination with the C^{β} density. At each alignment depth, the classification procedure was repeated 30 times using different random homolog subsets, and the balanced prediction error at each depth was calculated from the average of these repeated classifications. The results of this analysis (Figure 2) approximately quantify the increase in classification accuracy gained by adding a residue burial term when limited evolutionary information is available; the combination of the C^{β} density and SIFT calculated from only two randomly selected homologs yields prediction results that are comparable to using only SIFT calculated from the full Lac repressor alignment of 21 homologs. When only structural information is used, the results are roughly equivalent to those obtained using SIFT calculated from four homologs, yet the information from structure is complementary to that from evolution, as can be observed by the significant improvement in prediction resulting from the addition of a structural term to SIFT calculated from the full alignment of 21 sequences. This observation shows that, although structural features are clearly more useful when only a small number of homologs are available, they can still make a significant contribution

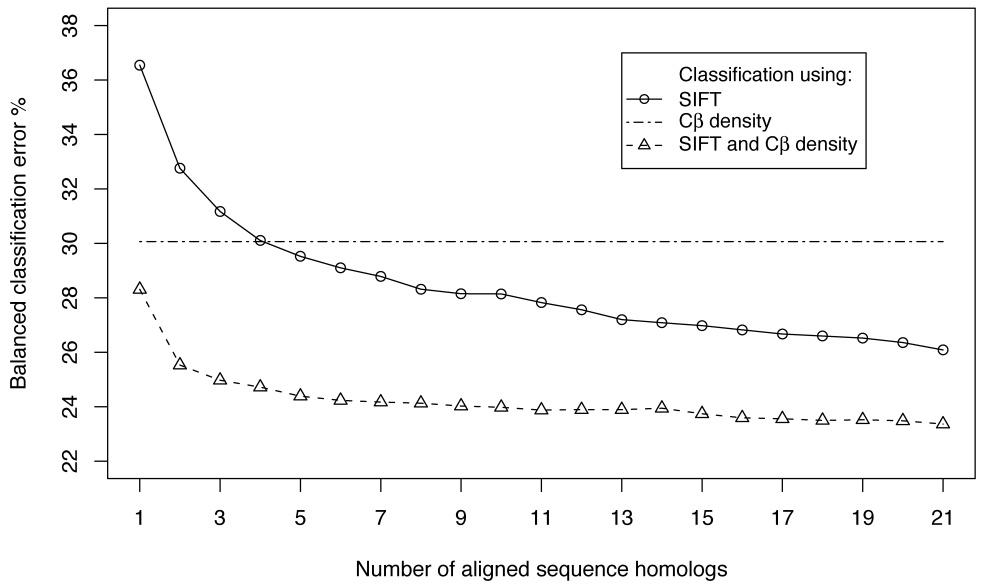

Figure 2. Classification of Lac repressor mutations with limited sequence homologs. The balanced classification error of laboratory mutagenesis data from Lac repressor is shown for classifications made using the SIFT score and the C^{β} density individually and in combination. Classification trees were trained with SIFT scores calculated from all available Lac repressor homologs, and the parameters of those classification trees were then used to categorize mutations using the SIFT score calculated from a restricted number of homologs. At each restricted homolog count, the SIFT score was calculated from a random subset of homologs and scored with the previously trained classification tree. This procedure was repeated 30 times and the results were averaged to approximate the balanced classification error for each restricted homolog count.

even when a deep msa exists due to the complementary nature of structural and evolutionary information.

Classification enhanced by *ab initio* predicted structure

The previous results suggest the possibility of using structures approximated by *ab initio* prediction methods to enhance mutation prediction for proteins lacking other sources of structural information and having few homologous sequences available. We have found that for a test set of 128 small proteins, the C^{β} density values calculated from structures predicted by the Rosetta method as part of a previous large-scale study[13] have an error distribution with a standard deviation of 4.5 (see Methods). After making the simplifying assumption that this error is normal and applying it to the actual C^{β} density for proteins in our laboratory test set, we found that the balanced classification error using the C^{β} density alone increased from 28.8% to 33.6%. Classification using this noisy C^{β} density term in combination with SIFT made very little improvement over classification using SIFT alone; however, when each protein in the laboratory test set was restricted to an alignment of three homologs (see Methods), the addition of the noisy C^{β} density to SIFT improved the classification error from 26.9% to 24.8%. Thus, with current *ab initio* prediction methods, a modest improvement in deleterious mutation prediction potentially can be obtained for proteins with few homologous sequences.

Discussion

Using a simple non-linear classification model validated on unbiased laboratory-assayed mutations, the lowest classification error found in this study was achieved using only two features: the C^{β} density and the SIFT score, yielding a balanced classification error of 20.5%. These features used alone were, respectively, the most accurate structure and sequence-based metrics considered for classification of laboratory-assayed mutations in this study, with the C^{β} density having a classification error of 28.8% and SIFT having an error of 22.2%. The feature combination of the C^{β} density and SIFT results in relatively strong cross-validated classification accuracy for the human allele test set as well as for classification of human alleles by a model trained on laboratory-assayed data, suggesting that the superior performance of this feature combination is general. Although these results indicate that a simple residue burial term can be more useful than more complex structural properties and rule sets, we anticipate that features using expert knowledge of protein structure or explicit modeling of the mutated residue in a macromolecular force-field will eventually provide more informative terms for characterizing mutations.

When fewer than five to ten homologous sequences are available, using structural information results in substantially improved prediction of intolerant mutations; for instance, when the alignments for the laboratory mutation test set are restricted to three homologs, the addition of structural information reduces the prediction error by 4.5%, compared to an error reduction of only 1.7% when full alignments are used. In Lac repressor, the addition of a structural term to the prediction when only two homologs are available adds enough information to improve the prediction accuracy to that observed when the full alignment of 21 sequences is used. We observe as well that even when a deep msa is available for Lac repressor, structure can improve the accuracy of predictions made with sequence alone, due to the complementarity of the information from structure and sequence. This is expected, because the structural information can compensate for sequence errors that persist even in deep sequence alignments, such as those arising from misalignment and unaccounted covariance among evolutionary sequence changes, and the evolutionary information encodes factors unlikely to be detectable from structure alone, such as the location of functional sites on the protein surface.

The relative importance of structural and evolutionary information in deleterious mutation prediction has an interesting parallel in the related remote homolog detection problem. The most widely used remote homolog detection methods, such as PSI-BLAST, utilize sequence information alone, but structure-based threading and fold recognition methods, many of which use sequence information as well,[14,15] can often discern more distant relationships.[16] Panchenko *et al*.[17] have examined the relative contribution of sequence and structural information to the fold recognition problem, and report that when sequence or structure is used alone the sequence information gives better fold recognition results, but that the combination of sequence and structure is superior to either used alone, a result that agrees with our observations for deleterious mutation prediction.

The demonstrated superiority of the SIFT score to a set of simpler evolutionary features is the result of SIFT's incorporation of residue substitution, conservation and biochemical class information in a single PSSM-based scoring term. One possible way of improving our methods would be to extend this PSSM-based approach by incorporating information from the local structural environment into the PSSM calculation. We have implemented such a method, which modifies the SIFT algorithm by using C^{β} density-dependent pseudocounts in the creation of the final PSSM from which the SIFT score is generated. This was accomplished by using substitution tables to generate pseudocount distributions for the PSSM, and encoding the local structure of each residue into this distribution by using a substitution table generated from residues of similar structural environment (see Methods). When this implementation of SIFT using structural pseudocounts was tested, it yielded less accurate prediction results than the previously discussed C^{β} density/SIFT classification tree method: for the laboratory and human test sets, our modification of SIFT had balanced classification errors of 21.6% and 34.9% *versus* 20.5% and 29.6% for the C^{β} density/SIFT classification tree. The relatively poor performance of this score could be explained by differences between the substitution table-based psuedocounting method[18] we used and the default SIFT method of generating pseudocounts from Dirichlet mixtures.[1] Another detrimental factor could be the quality of the structurally dependent substitution tables themselves. In order to create substitution tables for structural SIFT with sufficient counts to represent narrow ranges of the C^{β} density, we collected statistics from full PSI-BLAST alignments of a non-redundant set of known structures, resulting in alignments of lower quality than those made from smaller, more conserved sequence regions such as BLOCKS.[19] This trade-off between collecting sufficient substitution counts and the quality of the alignments used to derive these counts is likely to have adversely affected the performance of our method, and an attempt to optimize the quality of structural substitution tables for this problem in the future may prove structural variants of a PSSM-based method such as SIFT more successful.

Our results show that *ab initio* predicted structures may improve deleterious mutation prediction when experimental structures are unavailable and few homologous sequences exist. Indeed, our initial motivation for undertaking this study was to explore the use of low-resolution models produced by the Rosetta method[13] to improve deleterious mutation prediction. This approach was motivated by the observation that the C^{β} density can significantly enhance the accuracy of predictions made when using SIFT with few homologous sequences (Figure 2), suggesting that even a coarse approximation to the C^{β} density may improve prediction when the experimentally determined structure is not available. We have shown, using a noisy C^{β} density term approximating the error observed in structures modeled by *ab initio* methods, that a small improvement in prediction (reduction of error by 2.1%) can be made by combining this term with SIFT. Thus, while it is probably still premature to apply information from *ab initio* modeled structures to deleterious mutation prediction, it may be a quite useful endeavor following some improvement in protein structure prediction.

Methods

Sequence-based features

Blosum62

The Blosum62 feature for each mutation utilized the log-likelihood ratios of tolerated residue substitution

relative to random residue alignment, as calculated by the BLOSUM method[10] with sequence pairs above 62% identity clustered to reduce amino acid pair counts from recently diverged sequence homologs.

Normalized site entropy

The normalized site entropy was calculated using sequence alignments constructed by SIFT.[1] For each site in the alignment, the entropy was calculated for the probability distribution of all residues at that site, and over each protein chain these entropy values were normalized to the same mean and standard deviation.

SIFT

The SIFT algorithm for deleterious mutation prediction[1] was used in this study with alignments performed by a modification of the default "SIFT by conservation" method. In this modification, the NCBI non-redundant protein database is used for the initial PSI-BLAST[20] search and the 600 most diverged homologs from this search are taken as candidates for the final alignment. All subsequent steps in calculating the SIFT score follow the published method. This change was implemented to include a greater number of diverged homologs for proteins such as HIV-1 protease, in which case a large number of very closely related sequences mask more informative homologs when the default SIFT alignment method is used.

Test mutation structures

The structural features of proteins in the laboratory mutagenesis set were calculated from the following Protein Data Bank[21] (PDB) structures: 2lzm for T4 lysozyme, 1dif for HIV-1 protease, and 1efa for *Escherichia coli* Lac repressor. Structural features for HIV-1 protease and Lac repressor were calculated as dimers, features for all other proteins in this study were calculated from individual protein chains.

The structural features of proteins in the human allele set were calculated by searching the PDB for individual chain structures at 3.0 Å resolution or better with at least 40% sequence identity with the test set protein. Given multiple candidate structures, priority was given to the highest level of sequence identity. Alignments between the test set protein and the PDB chain sequence were made using BLAST[20] and structural features calculated on the PDB chain were transferred directly to the aligned residues of the test set protein.

Structure-based features

Residue burial

The two residue burial features used were the C^{β} density and the relative solvent accessibility. The C^{β} density feature is a count of the number of C^{β} atoms within 10 Å of the C^{β} of the mutated residue and the relative solvent-accessibility is the solvent-accessible surface area of the native residue at the mutation site as calculated by DSSP[12] and normalized by the maximum value for each residue according to Rost & Sander.[22]

Sunyaev structural rules

The Sunyaev structural rules are recreated from the structural subset of a larger set of published rules for identifying deleterious mutations.[2] We have followed the published description, except that the solvent-accessible surface area and secondary structure were calculated by DSSP, and for the human allele test set these values were calculated from an homologous structure, as described above. The structural rules included here are: (1) hydrophobic core disruption: the mutation site has less than 25% solvent accessibility and the difference in accessible surface propensity between the two residues in the mutation is greater than 0.75; (2) buried charge change: the mutation site has less than 25% solvent accessibility and the mutation entails a change in electrostatic charge; (3) solubility change: the mutation site has greater than 50% solvent accessibility and the accessible surface propensity between the two residues in the mutation is greater than 2.0; and (4) proline in an α-helix: any mutation to proline at a site predicted by DSSP to be part of an α-helix.

Classification with reduced homolog sets

The classification error of predictions made from subsets of all available homologs was found by taking an average of the classification errors from 30 rounds of prediction using a new random subset of homologs in each round. The sample classification error for each round was found by first selecting a random subset of homologs, then calculating homolog-dependent features such as SIFT from this reduced set, and finally using these features together with structural features to perform cross-validated model training and error classification (except as in Figure 2).

C^{β} density dependent substitution tables

Multiple sequence alignments were created using PSI-BLAST for each member of a non-redundant set of protein chains of known structure, with each chain having a maximum pairwise sequence identity of 50% and a minimum structural resolution of 2.5 Å.[23] Within each alignment, sequences with less than 33% sequence identity with the starting protein chain were removed. Pairwise substitution counts were then generated from these trimmed alignments using the BLOSUM method for clustering similar sequences to reduce pair counts from insufficiently diverged sequences, with the clustering threshold set at 62% sequence identity. Using the C^{β} density value of the starting protein chain at each site, substitution counts were accumulated for each C^{β} density value separately. To accumulate sufficient pair counts for every C^{β} density value in the range of 5–32, augmented count tables for each value were summed from a total of five C^{β} density values in the range -2 to $+2$ relative to the value of the augmented table. Finally, the augmented table for each C^{β} density value was converted to pair probabilities of the same type as the q_{ij} values described in the BLOSUM method.

SIFT with structural pseudocounts

The pseudocounting technique from the published SIFT method was replaced with a technique that distributes pseudocounts so as to reflect the bias of both the local structural environment and the observed residues

at an alignment site. This was accomplished by generating the pseudocount distribution using the data-dependent method of Tatusov *et al.*,[18] a technique that utilizes amino acid substitution probabilities. At each site, the table of substitution probabilities used to generate pseudocounts was selected on the basis of the local structural environment at that site.

In this study, we have implemented a version of SIFT with structural pseudocounts in which the local structural property used to generate the pseudocount distributions is the C^{β} density. C^{β} density-dependent amino acid substitution probability tables were generated for C^{β} density values from 5 to 32, and the appropriate table was used to generate the pseudocount distribution at each site in the msa, dependent on the C^{β} density of that site in the native or closest homologous structure.

Approximated Rosetta C^{β} densities

For each protein from a large-scale study of structures predicted by the Rosetta method,[13] the five highest scoring decoy structures were selected from each of the top five decoy clusters. For each residue in these decoys, the C^{β} density was calculated and the average value for each residue was taken over all five decoys. For each residue over the whole protein set, the error was found between the average decoy C^{β} density and the C^{β} density derived from the native structure. Over a test set of 128 proteins with an average size of 81 residues, we found the distribution of this decoy C^{β} density error to have a standard deviation of 4.5. To generate a rapid approximation of the decoy C^{β} density error for any protein, we made the simplifying assumption that the decoy error was normal and added a random normal deviate of standard deviation 4.5 to the true C^{β} density value wherever the approximate decoy error was used.

Acknowledgments

We thank Pauline Ng and Steven Henikoff for providing mutagenesis data, access to the SIFT source code, and for helpful discussions regarding this problem. We thank Tanja Kortemme for providing side-chain entropy calculations for several proteins. This work was supported by the HHMI.

References

1. Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874.
2. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W. C., III, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597.
3. Wang, Z. & Moult, J. (2001). SNP's, protein structure, and disease. *Hum. Mut.* **17**, 263–270.
4. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
5. Markiewicz, P., Kleina, L., Cruz, C., Ehret, S. & Miller, J. H. (1994). 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.* **240**, 421–433.
6. Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. & Hutchison, C. A., III (1989). Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.
7. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–87.
8. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
9. Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706.
10. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci.* **89**, 10915–10919.
11. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. (2001). Computational method to reduce the search space for directed protein evolution. *Proc. Natl Acad. Sci.* **98**, 3778–3783.
12. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
13. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
14. Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.
15. Panchenko, A., Marchler-Bauer, A. & Bryant, S. H. (1999). Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins: Struct. Funct. Genet.* **37**, 133–140.
16. Murzin, A. (1999). Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Struct. Funct. Genet*, **37**, 88–103.
17. Panchenko, A. R., Marchler-Bauer, A. & Bryant, S. H. (2000). Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296**, 1319–1331.
18. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci.* **91**, 12091–12095.
19. Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565–6572.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). GappedBLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
21. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
22. Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216–226.

23. Hobohm, U., Scharf, M. & Schneider, R. (1993). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
24. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*, pp. 59–129, Wadsworth, Pacific Grove, CA.
25. Therneau, T. M. & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines, *Technical Report Series No. 61*, pp. 5–13, Department of Health Science Research, Mayo Clinic, Rochester, MN, http://www.mayo.edu/hsr/techrpt.html

Edited by B. Honig

(Received 16 April 2002; received in revised form 18 July 2002; accepted 2 August 2002)