doi:10.1006/jmbi.2001.4510 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 307, 683–706

JMB

Predicting the Functional Consequences of Non-synonymous Single Nucleotide Polymorphisms: Structure-based Assessment of Amino Acid Variation

Daniel Chasman* and R. Mark Adams

Variagenics, 60 Hampshire Street, Cambridge MA 02144, USA

**Corresponding author*

We have developed a formalism and a computational method for analyzing the potential functional consequences of non-synonymous single nucleotide polymorphisms. Our approach uses a structural model and phylogenetic information to derive a selection of structure and sequence-based features serving as indicators of an amino acid polymorphim's effect on function. The feature values can be integrated into a probabilistic assessment of whether an amino acid polymorphism will affect the function or stability of a target protein. The method has been validated with data sets of unbiased mutations in the lac repressor and lysoyzyme. Applying our methodology to recent surveys of genetic variation in the coding regions of clinically important genes, we estimate that approximately 26-32 % of the natural non-synonymous single nucleotide polymorphisms have effects on function. This estimate suggests that a typical person will have about 6240-12,800 heterozygous loci that encode proteins with functional variation due to natural amino acid polymorphism.

Keywords: single nucleotide polymorphism; SNP; human genome; structural model; protein structure

Introduction

Recent surveys of human genetic diversity have estimated that there are about 250,000-400,000 common single nucleotide polymorphisms (SNPs) in protein coding sequences of the genome.[1,2] Analysis of their functional effects is a crucial aspect of current genomic science. Coding SNPs (cSNPs) are interesting, in part, because some of them, termed non-synonymous SNPs (nsSNPs), introduce amino acid polymorphisms into their encoded proteins. nsSNPs are proportionally less prevalent than synonymous SNPs that do not affect protein sequence, presumably as a consequence of selection against the functional disruptions of amino acid variation.[1,2] However, it might be expected that a significant fraction of molecular functional diversity in the human population remains attributable to effects on protein function caused by nsSNPs. For example, the kinetic parameters of enzymes, the DNA-binding properties of proteins that regulate transcription, the signal transduction activities of transmembrane receptors, and the architectural roles of structural proteins are all susceptible to perturbation by nsSNPs and their associated amino acid polymorphisms. Amino acid polymorphisms can also influence the efficacy and toxicity of drugs, as has been shown for cytochrome P450 2D6, TPMT, and the β_2-adrenergic receptor among others.[3–8]

Structural analysis of amino acid polymorphisms provides a powerful mechanistic explanation of their effects on function. Very early in the molecular analysis of genetic variation, the strengths of structural analysis were demonstrated for the case of amino acid mutations in hemoglobin. Here, the molecular basis of the clinical effects caused by mutations could be inferred as soon as the structural information became available.[9,10] These pioneering studies recognized crucial links between the structural disposition of residues and potential effects of mutations on function, including the destabilizing effects of introducing charged

Abbreviations used: SNP, single nucleotide polymorphism; cSNP, coding SNP; nsSNP, non-synonymous SNP; CWRU, Case Western Reserve University; WI, Whitehead Institute; DBH, dopamine beta-hydroxylase; ALDR1, aldose reductase; PTGS2, prostoglandin synthase.

E-mail address of the corresponding author: dchasman@variagenics.com

0022-2836/01/020683–24 $35.00/0

residues into the hydrophobic core of a protein, and the functional disruptions of mutations in protein residues that contact the iron or the heme ligands.

As the structure databases have grown and been analyzed by computational methods, the understanding of the relationship between structure and the effects of amino acid substitution on function has continued to deepen. Many studies have shown that a model residue's solvent accessibility is important for anticipating whether its mutation will affect function (for example see Bowie *et al*.[11] and Dao-pin *et al*.[12]). In a similar way, Alber *et al*. demonstrated a strong relationship between molecular rigidity measured by a crystallographic *B*-factor and the tolerance to mutation for the case of lysozyme.[13] Several groups have noted a systematic intolerance to mutation in residues that are either extremely conserved in phylogeny or confined in their identity to particular classes of amino acid residues.[14–16] Others point to relationships between functional effects and hydrophobicity or residue volume.[11] More recently, Sunyaev *et al*. have started to examine the relationship between structural features and either human disease causing mutations or common human nsSNPs.[17] They identified structural features that are significantly associated with the disease causing polymorphisms, and found that a surprisingly large fraction, about 45 %, of the prevalent nsSNPs share these structural features. For about half of these structurally important residues, the polymorphism represents an amino acid substitution that apparently is not found in interspecies variation.

The current number of known protein structures is still far less than the number of known human protein sequences, but this discrepancy does not diminish the importance of structural analysis for understanding the effects of nsSNPs and their amino acid polymorphisms on function. It is accepted generally that proteins with similar amino acid sequence will exhibit a high degree of structural homology, even when they are only distantly related. For example, in hemoglobin and myoglobin (sharing only 25 % amino acid identity in sequence) the structural dispositions of many corresponding residues are extremely well conserved. The same principle serves as the underlying basis of classifications of proteins according to fold families (e.g. DALI, SCOP)[18,19] and structure prediction methods like threading[20,21] and homology modeling.[22,23] Shared structural properties can be extremely precise, for example in the conservation of the residues that coordinate the heme in hemoglobin and myoglobin. They can also be more general, as when the corresponding residues in two proteins are both hydrophobic and buried in the hydrophobic core. For the majority of proteins, structural information is not available; but for those proteins with sequence homology to a protein of known structure, much structural information can be inferred. For the human genome, about 30 % of the protein sequences are likely to be homologous to known crystal or NMR structures.[22] The current high-throughput structure initiatives and theoretical modeling techniques will increase this proportion dramatically in the next few years.[24–26]

We have identified a set of generic structure and sequence-based features that serve as predictors of whether amino acid polymorphisms in a target protein will affect its function. The values for the features are useful whether they are derived from a structure of a target protein or from a structure of a protein homologous to the target protein. We have developed a statistical model that uses the values for the predictive features and training data to assign a probability that a polymorphism will affect function. More importantly, we have outlined a formalism for evaluating and implementing structure and sequence-based features of amino acid polymorphisms as predictors of effects on function in general. Applying our methods to the recent surveys of nsSNPs from the Case Western Reserve University (CWRU) and Whitehead Institute (WI) genome centers, we make predictions about potential effects on function for 23 % (50/216) and 39 % (65/168) of the nsSNPs, respectively. Extrapolating our analysis of these surveys to the entire genome, we estimate that 26-32 % of the natural non-synonymous polymorphisms in the human population are likely to affect protein function.

Results

Overview

Our approach is to use structural models and a standardized analysis of the structural and phylogenetic disposition of a modeled polymorphic residue to estimate whether it can be expected to affect protein function. The approach requires a formalism that includes the identification of a suitable protein structure for three dimensional modeling of the target polymorphism, the development of standard features for analysis of the modeled polymorphism, and integration the values of the features into a prediction about a polymorphism's potential effects on function. Once we have validated our methods with data sets of exhaustive unbiased mutations in the lac repressor and lysozyme, we apply them to the analysis of potential functional effects of amino acid polymorphisms derived from recently published nsSNPs in clinically relevant proteins.[1,2]

The predictive features

In order to evaluate the ability of the structure-based and phylogeny-based features (Table 1, Materials and Methods and Figure 1 for structural neighborhoods) to discriminate between mutations that either affect or do not affect protein function, we analyzed the feature values for mutations in

Table 1. Features used to described modeled polymorphisms

Feature name	Description
A. *Continuously-valued environment features*	
Residue accessibility	Solvent accessible area of model residue
Residue relative accessibility	Accessibility relative to maximum accessibility for model residue
Relative residue phylogenetic entropy[a]	Phylogenetic entropy of model residue normalized to average and SD in phylogenetic entropy for other residues in the same PDB chain
Neighborhood relative phylogenetic entropy[b]	Phylogenetic entropy of model residue's structural neighborhood relative to average phylogenetic entropy of other collections of the same number of residues from the same PDB chain
Relative residue *B*-factor	*B*-factor of model residue normalized to average and SD in *B*-factor for other residues in the same chain
Neighborhood relative *B*-factor[2]	*B*-factor of model residue's structural neighborhood relative to average *B*-factor of other collections of the same number of residues from the same PDB chain
B. *Categorically-valued features*	
Unusual AA	One of the amino acid residues in the polymorphism is not in the phylogenetic profile
Unusual AA by class	One of the amino acid residues in the polymorphism is not in the smallest amino acid class that includes the phylogenetic profile.[31]
Rare AA	The polymorphism includes an amino acid that occurs less than 10% of the time in the phylogenetic profile
Buried charge	The model residue is buried and the polymorphism includes a charged amino acid. Often a special case of Unusual AA
Turn breaking	The polymorphism occurs at a glycine or proline in a turn. Often a special case of Unusual AA
Helix breaking	The polymorphism occurs in a helical region of the model and includes a glycine or proline. Often a special case of Unusual AA
Conserved position	The polymorphism occurs at an absolutely conserved position in phylogeny. Always a special case of Unusual AA
Near conserved position	The polymorphism occurs in a structural neighborhood that includes a conserved position.
Near heterogen atom[c]	The model for the polymorphism occurs near a ligand in the model
Near interface	The model for the polymorphism occurs near a subunit interface in the model

[a] Phylogenetic entropy (or Information content) $= -\sum_{i=1}^{\text{N. Different Amino Acidresidues}} f_i \ln f_i$
Where:
f_i = Fraction of the sequences having amino acid i at that residue position in the HSSP multiple sequence alignment.

[b] Neighborhood relative feature $= \dfrac{\sqrt{N}(\langle V_n\rangle - \langle V_c\rangle)}{\sigma_{V_c}}$
Where:
N = Number of residues in the structural neighborhood.
$\langle V_n\rangle$ = Average in the value of *B*-factor or phylogentic entropy for the structural neighborhood (Materials and Methods, Figure 1).
$\langle V_c\rangle$ = Average in the value of *B*-factor or phylogenetic entropy for the PDB chain.
σ_{V_c} = Variance in the *B*-factors or phylogenetic entropy for the PDB chain.
[c] Heterogen atom is an atom in a non-standard group in a PDB file. In this analysis, heterogen atoms typically belong to ligand molecules.

the lac repressor and lysozyme mutation data sets (see Materials and Methods). These mutations represent changes to 12 or 13 alternative amino acid residues at every position in the two proteins and were constructed without *a priori* knowledge of their effects on function. As a consequence, the relationship between their features' values and effects on function can be considered to be unbiased; and standard statistical analysis can assess each feature's relative importance as a potential indicator of a mutation's effect on function. For continuously-valued features, e.g. the ''relative crystallographic *B*-factor'', we judged discrimination between two groups, mutations with or without effects on function, using Analysis of Variance (ANOVA)[27,28] with F-statistics. This procedure examines the probability of the null hypothesis that the average values and variances in the segregated groups of data not statistically different from the average value and variance in the data as a whole. For the categorically valued features, e.g. those indicating ''helix or turn breaking'' propensity and an ''unusual amino acid'' among others, we judged discrimination using the χ^2 test (Table 2).

In general, the continuously-valued environment features discriminate more strongly than the categorical features between mutations with and without effects on function (Table 2). For the lac repressor mutations modeled on the lac repressor structure and the lysozyme mutations modeled on the lysozyme structure, the statistical ranking of the environment features is essentially the same, and the purely structure-based accessibility and *B*-factor features are more important than the features that draw on phylogenetic information. For the lac repressor mutations modeled on the homologous purine repressor structure, the phylogenetic features are better predictors than the purely structure-based features, possibly due to their less

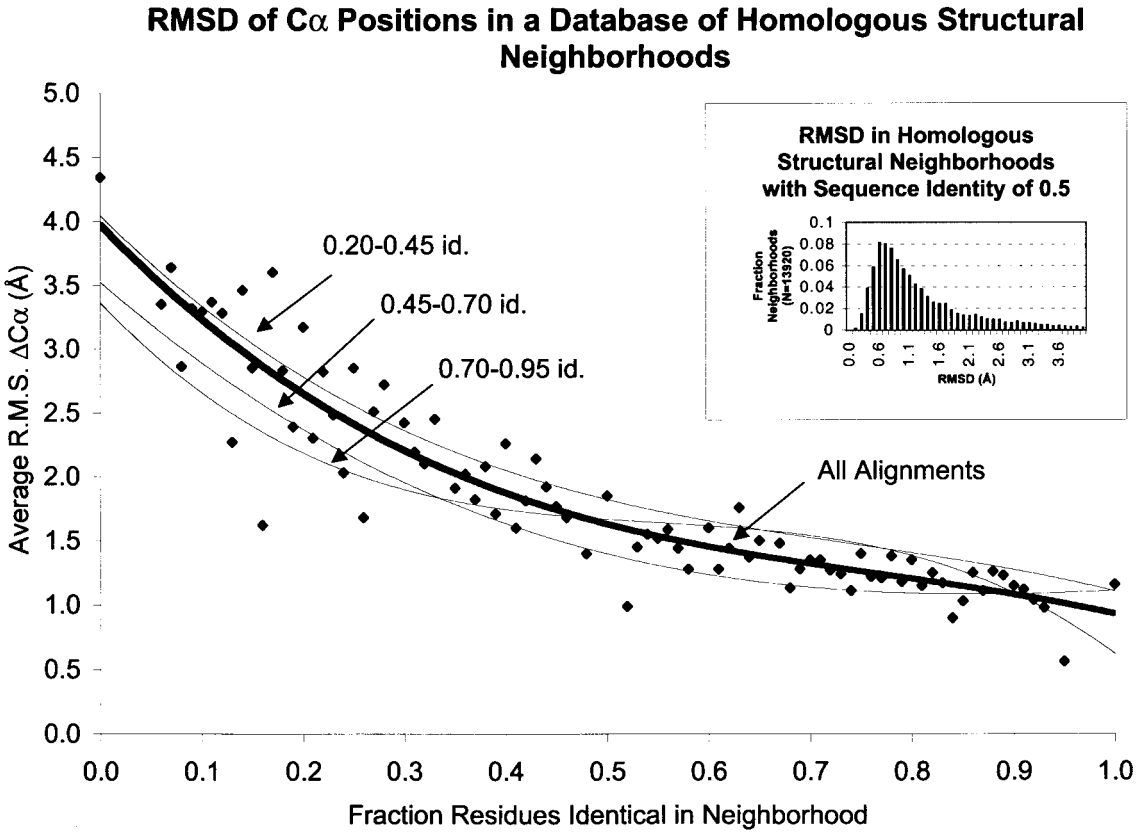

Figure 1. r.m.s.d. in C^{α} positions in a database of homologous structural neighborhoods. Each data point represents the average of the r.m.s. in C^{α} distances in homologous structural neighborhoods with the level of sequence similarity indicated on the abscissa. The thick line reflects a second order fit to these data points. The thinner lines represent the second order fits to the average r.m.s. distances of C^{α} positions in corresponding neighborhoods taken from pairs of homologous proteins with overall levels of sequence similarity as shown (0.20-0.45, 0.45-0.70, or 0.70-0.95 identical amino acid residues in the overall alignment). The distribution of the r.m.s.d. in C^{α} positions for structural neighborhoods sharing 0.5 identical residues is shown in the inset (see also Guex *et al.*[22] and Holm & Sander[37]).

critical dependence on precise structural information. The absolute statistical significance of the features in the lac repressor and lysozyme data sets is also different, and may reflect intrinsic functional differences between the two proteins or the experimental criteria for determining which mutations affected function. For the phylogenetic entropy and *B*-factor features, the neighborhood measures are roughly equivalent in statistical significance to the residue measures, emphasizing the validity of the structural neighborhood construction and the concept of local structure-based indicators of effects on function.

Among the most significant of the categorically-valued features, one is structure-based (''buried charge''), and the other is sequence-based (''unusual amino acid''). Other categorical structure-based features, for example ''near heterogen atom'' and ''near interface'', are also significant but might not be expected to be sufficiently generic for wide applicability. Alone, the ''turn breaking'' and ''helix breaking'' structural features are not reliable; they may be predictive in combination with other features.

The unusual amino acid features convey the idea that evolution has sampled the set of allowable substitutions, and that this set can be found in a multiple alignment of a complete set of sequences related to the target protein by phylogeny at the position corresponding to the polymorphic residue. When the phylogeny is incomplete, amino acid classes can be used to enumerate the family of tolerated amino acid residues in a particular environment. Indeed, the ''unusual amino acid by class'' feature is a good predictor of effects on function. Often, the unusual amino acid construction will imply the hydrophobicity measures considered by others for evaluating the functional consequences of mutations. The efficacy of the unusual amino acid features is reinforced by the observation that the ''rare amino acid'' feature predicts there will not be an effect on function (Table 2B). We conclude that for the lac repressor and lysozyme, the occurrence of a mutant amino acid in phylogeny,

Table 2. Feature discrimination between effects on function in lac repressor and lysozyme mutations

Feature name	Lac repressor mutations LacI model	Lac repressor mutations PurR model	Lysozyme mutations Lysozyme model
A. *ANOVA F-test for continuously-valued environment features*[a]			
Accessibility	9.12E-117 (2)	1.43E-45 (5)	1.39E-28 (2)
Relative accessibility	**1.14E-147** (1)	8.22E-44 (6)	**2.15E-29** (1)
Relative entropy	4.32E-73 (5)	**1.00E-112** (1)	2.45E-17 (5)
Nbhd. relative entropy	7.07E-68 (6)	4.85E-94 (2)	2.80E-09 (6)
Relative *B*-factor	3.03E-95 (4)	1.04E-75 (3)	1.72E-24 (3)
Nbhd. relative *B*-factor	2.86E-106 (3)	1.53E-58 (4)	2.77E-24 (4)
B. *Chi-squared test for categorical features*[b,c]			
Unusual AA	1.49E-16+ (5)	1.29E-29+ (4)	1.32E-05+ (5)
Unusual AA by class	9.13E-69+ (2)	**1.46E-50**+ (1)	3.30E-09+ (4)
Rare AA	ND NA	2.53E-25− NA	ND NA
Buried charge	**1.48E-83**+ (1)	6.31E-50+ (2)	**4.40E-18**+ (1)
Turn breaking	3.63E-08− NA	1.02E-06− NA	5.52E-01+ (6)
Helix breaking	2.14E-10+ (8)	4.70E-11+ (6)	5.01E-01− NA
Conserved position	4.00E-24+ (4)	1.37E-10+ (7)	3.90E-14+ (2)
Near conserved position	2.82E-11+ (6)	2.65E-32+ (3)	7.00E-10+ (3)
Near het atom	1.19E-50+ (3)	1.94E-11+ (5)	ND NA
Interface	1.19E-12+ (7)	2.80E-21+ (4)	ND NA

[a] ANOVA (Analysis of Variance) test compares the feature variance for mutations with or without effects on function to the feature variance for all mutations (see Materials and Methods). Values indicate probability of the null hypothesis occuring by chance. Numbers in parentheses indicate the ranking of the discrimination of each feature. The features with the highest discrimination are indicated by bold face font.

[b] Test compares proportion of mutations with or without the feature to the proportion of mutations with or without effects on function.

[c] +, the feature is predictive of an effect on function; −, the feature is predictive of no effect on function.

even infrequently, suggests its tolerance by a particular structural environment.

Correlation between environment features

While the environment features pertaining to solvent accessibility, crystallographic *B*-factor, and phylogenetic entropy are good predictors of a polymorphism's effect on function, they are expected also to exhibit some degree of correlation. To study these relationships, we computed correlation coefficients for each pair of these features for residues in the lac repressor, purine repressor, and lysozyme crystal structures (Table 3A). The maximum correlation value is found for ''accessibility'' and ''relative accessibility'' (0.93-0.96), and the correlation values between residue and neighborhood versions of the relative *B*-factor or the relative phylogenetic entropy are also very high (0.51-0.81). The smallest correlation is found between the phylogenetic entropy features and either the accessibility features or the *B*-factor features. The minimum correlation, found for the comparison of ''relative phylogenetic entropy'' and ''neighborhood relative *B*-factor'' in the lysozyme structure, was very low (0.09). Qualitatively, some of these findings are expected from the literature. For example, solvent inaccessible residues are known to be more rigid and therefore have a lower *B*-factor (13, see also Materials and Methods). The current analysis provides quantitative measures of the relationship between pairs of the environment features.

An exploratory study using principal component analysis suggested that the majority of the variance (0.51-0.56) in the group of continuously-valued environment features could be attributed to a single eigenvector with roughly equal proportions of the accessibility and *B*-factor features and slightly less of the phylogenetic entropy features (Table 3B). An approximate additional 0.35-0.38 of the variance can be attributed to the second and third eigenvectors, and these typically treated the two *B*-factor features as equivalent, the two phylogenetic entropy features as equivalent, and the two accessibility features as equivalent (see coefficients, Table 3B). In the third eigenvector, there is some unequal splitting of the variance between residue and neighborhood versions of the phylogenetic entropy.

Together with the ANOVA of effects on function in the lac repressor and lysozyme data sets, this result suggests that the combination of one of the two *B*-factor features, one of the two phylogenetic entropy features, and one of the two accessibility features will provide a set of parameters that describe the environment of the polymorphic residue with minimal redundancy and that are useful for predicting whether it will affect function. This conclusion was supported below by additional techniques (see maximum likelihood approach in Materials and Methods). Both the correlation analysis and the principal component analysis were performed on the entire database of structural neighborhoods (Figure 1) with essentially the same results (see Materials and Methods).

Table 3. Correlation and principal component analysis of environment features

A. *Correlation analysis*[a]

Lac repressor (LacI model PDB: 1lbh top right, PurR model PDB: 2pua bottom left)

	Access	RelAccess	RelB	NbhdRelB	RelEnt	NbhdRelEnt
Access		0.93	0.58	0.44	0.33	0.24
RelAccess	0.96		0.59	0.49	0.31	0.26
RelB	0.58	0.55		0.81	0.30	0.38
NbhdRelB	0.36	0.39	0.79		0.27	0.47
RelEnt	0.37	0.37	0.34	0.26		0.51
NbhdRelEnt	0.16	0.16	0.28	0.30	0.68	

Lysozyme (lysozyme model PDB 7lzm)

	Access	RelAccess	RelB	NbhdRelB	RelEnt	NbhdRelEnt
Access		0.92	0.66	0.36	0.31	0.26
RelAccess			0.60	0.43	0.29	0.22
RelB				0.71	0.15	0.16
NbhdRelB					0.13	0.09
RelEnt						0.51
NbhdRelEnt						

B. Principal component analysis[b]

Principal component	Variance	Total variance (fr.)	Cumulative variance (fr.)	Coefficients of principal component: Access	RelAccess	RelB	NbhdRelB	RelEnt	NbhdRelEnt
Lac repressor model (PDB ID: 1lbh) for lac repressor									
1	3.37	0.56	0.56	−0.45	−0.46	−0.46	−0.43	−0.30	−0.32
2	1.13	0.19	0.75	0.41	0.40	0.07	−0.09	−0.52	−0.63
3	0.85	0.14	0.89	−0.35	−0.29	0.43	0.57	−0.54	0.00
4	0.43	0.07	0.96	0.15	0.19	−0.31	−0.14	−0.59	0.70
5	0.17	0.03	0.99	−0.08	0.21	−0.70	0.66	0.08	−0.15
6	0.06	0.01	1.00	−0.70	0.69	0.13	−0.16	0.02	0.01
Purine repressor model (PDB ID: 2pua) for lac repressor									
1	3.24	0.54	0.54	0.46	0.46	0.46	0.40	0.36	0.29
2	1.31	0.22	0.76	−0.34	−0.35	−0.12	−0.03	0.53	0.68
3	0.96	0.16	0.92	0.38	0.37	−0.44	−0.66	0.30	0.04
4	0.28	0.05	0.96	0.13	0.18	−0.21	0.04	−0.69	0.66
5	0.18	0.03	0.99	−0.04	0.20	−0.72	0.63	0.16	−0.14
6	0.04	0.01	1.00	0.71	−0.68	−0.13	0.11	0.00	−0.01
Lysozyme model (PDB ID: 7lzm) for lysozyme									
1	3.07	0.51	0.51	−0.51	−0.50	−0.47	−0.38	−0.26	−0.24
2	1.33	0.22	0.73	0.04	0.07	0.29	0.32	−0.63	−0.64
3	0.79	0.13	0.86	0.47	0.45	−0.26	−0.66	−0.14	−0.24
4	0.50	0.08	0.95	0.05	−0.02	0.11	−0.12	−0.71	0.68
5	0.25	0.04	0.99	0.06	−0.39	0.74	−0.52	0.13	−0.09
6	0.06	0.01	1.00	−0.72	0.62	0.25	−0.20	0.03	0.02

[a] For residues in the indicated crystal structure, the correlation coeffecient (cc) was computed as:

$$cc = \frac{1}{N}\sum_{i=1}^{N}\frac{(f_{1_i} - \langle f_1\rangle)(f_{2_i} - \langle f_2\rangle)}{\sigma_{f_1}\sigma_{f_2}}$$

Where:

f_{ji} = value of feature j for ith residue, $j = 1, 2$

$\langle f_j\rangle$ = average of feature; j

σ_{f_1} = variance in feature j

N = Number of residues in the structure

[b] Principal component analysis performed as described by Lebaut *et al.*[50]

Argument for a probabilistic model for anticipating the effects of polymorphisms

We considered two alternative approaches for using the predictive features to anticipate whether polymorphisms will affect function. In the first approach, whether a polymorphism will or will not affect function would be determined by whether the values of its predictive features meet certain, predetermined criteria. For example, a polymorphic residue's relative phylogenetic entropy might be less than some threshold, or it might be within some small distance of an enzyme active site. Our first predictive method (data not shown), and independently the method of Sunyaev *et al.*,[17] used this specialized form of annotation. It is very informative; but it may not be sufficiently general. The criteria may be hard to define in a generic way and some polymorphisms that fall short of the criteria may still affect function. The alternative approach would try to assign a probability of an effect on function to any modeled polymorphism according to an integrated assessment of the values of its predictive features. The probability value could be interpreted to reflect the indeterminacy in the prediction or, more practically, to rank candidate polymorphisms according to a likelihood or confidence level for an effect on function.

Our simple probabilistic model would combine structural and sequence information, in the form of the predictive features, with the lac repressor and lysozyme mutation data for anticipating which polymorphisms will affect function. We reasoned that if the features are sufficiently generic and reflect the most fundamental aspects of protein structure, the relationship between their values and effects on function may have been sampled in an unbiased way by the lac repressor and lysozyme mutation studies. For example, the lower the relative *B*-factor for a structural neighborhood and thus the more rigid, the more likely it is intolerant to mutation. The likelihood of a mutation causing an effect on function in this neighborhood could be estimated quantitatively by comparison to mutations in neighborhoods with similar *B*-factors in the lac repressor and lysozyme data sets. To a first approximation, a generic relationship between parameter values and effects on function may be valid for many different proteins, especially since we have used relative rather than absolute measures in the design of the *B*-factor and phylogenetic entropy features. As shown below, it is a fairly accurate description of both lac repressor and lysozyme in spite of their distinct functions.

The probabilistic model

Therefore, in the simple probabilistic model, the probability that a test polymorphism (or mutation) will affect protein function is estimated as the proportion of mutations in a training data set that affect protein function from among those with predictive feature values similar to the feature values of the test mutation. Here, the training data are the lac repressor and/or lysozyme data sets. For example, the probabilistic model might use ''solvent accessibility'', ''relative residue *B*-factor'', and the unusual amino acid as predictive features. A polymorphism of glutamate to valine might occur in a residue with moderate solvent accessibility, a relatively low *B*-factor, and for which a glutamate residue but not valine appear in its phylogeny, so that the polymorphism meets the requirements of containing an unusual amino acid. The probability that this polymorphism will affect protein function would be estimated as the proportion of mutations in the lac repressor and/or lysozyme data sets that affect protein function from among those that also have moderate solvent accessibility, relatively low *B*-factors, and meet the criteria for being an unusual amino acid.

An example of a two-feature probabilistic model

To illustrate a two predictive features version of the probabilistic model, we computed probability values from the lac repressor and lysozyme mutations using only one categorically-valued sequence-based feature, unusual amino acid, and one continuously valued structure-based feature, relative *B*-factor, with a 1 SD stringency for averaging (Figure 2, Materials and Methods). For the mutations in both data sets, the same trend in probability with respect to the relative *B*-factor is observed, whether meeting the criteria for unusual amino acid or not. The local slopes of the two curves are similar ($R^2 = 0.92$, unusual amino acid, $R^2 = 0.94$, no unusual amino acid). The slight difference in calibration (about 0.15 probability units in regression analysis, not shown) for the curves from the two data sets is likely derived either from the weaker statistical significance of relative *B*-factor in the lysozyme data set or from the relatively poor phylogeny for evaluating the unusual amino acid feature for the lysozyme data set (five sequences in multiple alignment) compared to the lac repressor data set (20 sequences in the multiple alignment) (Table 2A). The lower statistical power of the unusual amino acid feature for predicting effects in the lysozyme data compared to the lac repressor data was found also in the χ^2 tests (Table 2B).

Accuracy of the probabilistic model

In order to evaluate the accuracy of the probabilistic model, we needed to convert the probability values for whether a polymorphism would affect function into a prediction. To do this, we predicted that a polymorphism with probability of 0.5 or greater would affect function at a confidence level equal to the probability value. Conversely, we predicted that a polymorphism with a probability value of less than 0.5, would not affect on function

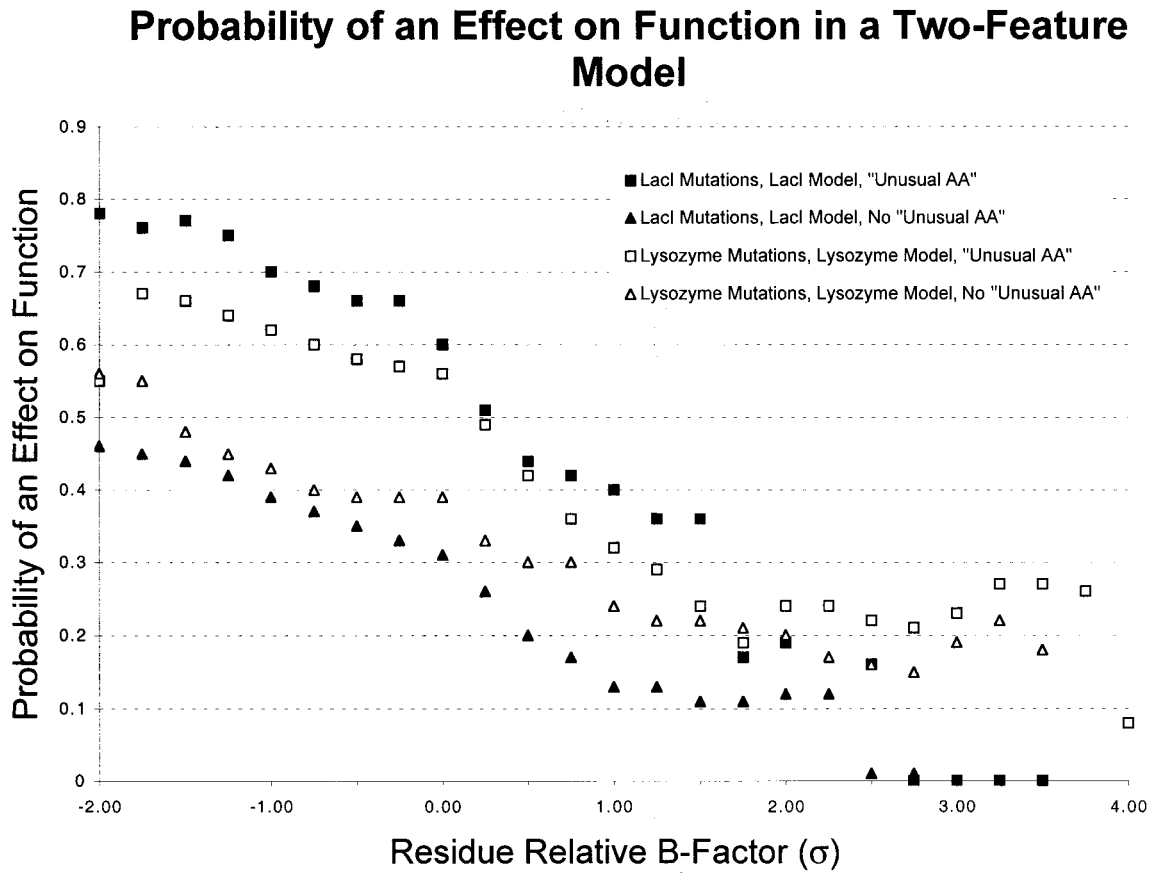

Figure 2. A two-feature example of the probabilistic model. The probability of an effect on protein function was computed dependent on the relative crystallographic *B*-factor for the modeled polymorphic residue, whether or not the polymorphism represents and unusual amino acid in phylogeny, and whether the training data are the lac repressor mutations modeled on the lac repressor structure or the lysozyme mutations modeled on the lysozyme structure. For relative *B*-factors values along the abscissa, probability values were estimated as the proportion of training mutations that affect protein function from among those within 1 SD in the relative *B*-factor and having an appropriate value of the unusual amino acid feature.

with a confidence level equal to one minus the probability value. Predictions could be judged according to their confidence level, and we report here standard measures to assess their accuracy: the number of correct and incorrect predictions of an effect and of no effect, the overall misclassification rate, the misclassification rate in predicting an effect, the misclassification rate in predicting no effect (see legend to Tables 4, 5 and 6 and Materials and Methods).

We devised two types of validation tests of the probabilistic prediction models. In the first, termed homogeneous cross-validation, each of the data sets was divided into two unequal parts, a training subset and a test subset. The training subset, comprising 90% of the mutation data, was used for selecting the subset of features by a maximum likelihood approach (Materials and Methods) and for computing probability values, and the test subset, comprising the remaining 10% of the mutation data, was used exclusively to test the predictions. Performed either with the lysozyme mutations modeled on the lysozyme structure or the lac repressor mutations modeled on the lac repressor structure, the homogeneous cross-validation test showed very significant predictive accuracy (Figure 3(a), Table 4). The overall misclassification rate ranges from a high value (0.25 lac repressor, 0.27 lysozyme) for predictions made with at least a 0.50 confidence level to a low value (0.05 lac repressor, 0.07 lysozyme) for predictions made at the 0.95 confidence level. For both, the interpretation of the probability value in terms of a confidence level for predictions is validated as is the maximum likelihood selection of categorical features. High predictive accuracy was also found for the lac repressor data modeled on the homologous purine repressor. The best structural neighborhoods in this model are about 60% identical in sequence to the lac repressor, and we note that there was no increase in the predictive accuracy for structural neighborhoods with higher sequence similarity between the target protein and the model (data not shown).

The second validation method, termed heterogeneous cross-validation, was more stringent and

Table 4. Cross-validation of probabilistic model for prediction effects of polymorphisms: homogeneous cross-validation

		Lysozyme mutations, lysozyme model (PDB: 7lzm)[a] Minimum prediction probability					Lac repressor mutations, lac repressor model (PDB: 1lbh)[b] Minimum prediction probability					Lac repressor mutations, purine repressor model (PDB: 2pua)[c] Minimum prediction probability				
Prediction	Actual	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5
Effect	Effect	3	8	24	37	43	10	34	52	66	78	36	58	83	100	115
	No effect	0	2	9	14	24	0	8	17	24	27	1	9	17	29	48
No Effect	No Effect	2	15	37	46	63	90	122	142	156	169	41	76	107	123	142
	Effect	1	7	10	18	28	3	5	22	32	43	0	4	19	30	51
Overall misclassified fraction		0.17	0.28	0.24	0.28	0.33	0.03	0.08	0.17	0.20	0.22	0.01	0.09	0.16	0.21	0.28
Predictions of an effect misclassified fraction		0.00	0.20	0.27	0.27	0.36	0.00	0.19	0.25	0.27	0.26	0.03	0.13	0.17	0.22	0.29
Predictions of no effect misclassified fraction		0.33	0.32	0.21	0.28	0.31	0.03	0.04	0.13	0.17	0.20	0.00	0.05	0.15	0.20	0.26

The environment features for all tests were ''relative accessibility'', ''residue relative phylogenetic entropy'', and ''neighborhood relative *B*-factor.'' The categorical features for the lysozyme mutations were selected by the maximum likelihood approach from among ''buried charge'', ''unusual amino acid'', ''unusual amino acid by class'', ''turn breaking'', ''helix breaking'', ''conserved position'' and ''near conserved position.'' The categorical features for the Lac Repressor mutations were selected from the group used for lysozyme combined with the ''near heterogen atom'' and ''near interface'' features.

[a] The categorical features were ''buried charge'', ''unusual amino acid by class'', ''helix breaking'', ''near conserved''. The fraction of test mutations with too little training data for prediction with these categorical features was 0.04.

[b] The categorical features were from maximum likelihood from ''unusual amino acid by class'', ''near heterogen atom'', ''near interface''. The fraction of test mutations with too little training data for prediction with these categorical features was 0.11.

[c] The categorical features were ''buried charge'', ''helix breaking'', ''near conserved position'', and ''near interface''. The fraction of test mutations with too little training data for prediction with these categorical features was 0.03.

Table 5. Cross-validation of probabilistic model for prediction effects of polymorphisms: heterogeneous cross-validation: categorical features from maximum likelihood

Training data:		Lysozyme mutations, lysozyme model (PDB: 7lzm)[a]					Lac repressor mutations, lac repressor model (PDB: 1lbh)[b]					Lysozyme mutations, lysozyme model (PDB: 7lzm)[c]				
Test data:		Lac repressor mutations, lac repressor model (PDB: 1lbh)					Lysozyme mutations, lysozyme model (PDB: 7lzm)					Lac repressor mutations pur repressor model (PDB: 2pua)				
		Minimum prediction probability					Minimum prediction probability					Minimum prediction probability				
Prediction	Actual	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5
Effect	Effect	68	227	358	483	551	30	70	156	271	341	5	9	20	45	69
	No Effect	10	33	101	233	345	8	13	49	108	166	0	1	1	76	79
No Effect	No Effect	101	259	515	666	786	232	368	436	534	644	201	328	631	654	667
	Effect	9	27	109	132	182	90	135	183	233	328	68	107	262	283	299
Overall misclassified fraction		0.10	0.11	0.19	0.24	0.28	0.27	0.25	0.28	0.30	0.33	0.25	0.24	0.29	0.34	0.34
Predictions of an effect misclassified fraction		0.13	0.13	0.22	0.33	0.39	0.21	0.16	0.24	0.28	0.33	0.00	0.10	0.05	0.63	0.53
Predictions of no effect misclassified fraction		0.08	0.09	0.17	0.17	0.19	0.28	0.27	0.30	0.30	0.34	0.25	0.25	0.29	0.30	0.31

The environment features for all tests were ''relative accessibility'', ''residue relative phylogenetic entropy'', and ''neighborhood relative B-factor.'' The categorical features in all tests were selected by the maximum likelihood approach from among ''buried charge'', ''unusual amino acid'', ''unusual amino acid by class'', ''turn breaking'', ''helix breaking'', ''conserved position'' and ''near conserved position.

[a] The categorical features selected were ''buried charge'', ''unusual amino acid by class'', ''helix breaking'', and ''near conserved position.'' The fraction of test mutations with too little training data for prediction with these categorical features was 0.42.

[b] The categorical features selected were ''buried charge'', ''unusual amino acid by class'', and ''turn breaking''. The fraction of test mutations with too little training data for prediction with these categorical features was 0.09.

[c] The categorical features selected were ''buried charge'', ''unusual amino acid by class'', ''helix breaking'', and ''near conserved position.'' The fraction of test mutations with too little training data for prediction with these categorical features was 0.70.

Table 6. Cross-validation of probabilistic model for prediction effects of polymorphisms: heterogeneous cross-validation: unusual amino acid categorical feature only

Training data:		Lysozyme mutations, lysozyme model (PDB: 7lzm)[a]					Lac repressor mutations, lac repressor model (PDB: 1lbh)[b]					Lysozyme mutations, lysozyme model (PDB: 7lzm)[c] Lac repressor mutations pur repressor model (PDB: 2pua)				
Test data:		Lac repressor mutations, lac repressor model (PDB: 1lbh)					Lysozyme mutations, lysozyme model (PDB: 7lzm)									
		Minimum prediction probability					Minimum prediction probability					Minimum prediction probability				
Prediction	Actual	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.6	0.5
Effect	Effect	0	90	321	794	997	0	17	47	306	405	0	109	369	613	748
	No Effect	0	17	129	394	594	0	7	12	124	216	0	34	124	250	372
No effect	No effect	60	388	648	976	1226	245	435	587	656	676	92	378	738	973	1177
	Effect	13	47	96	174	229	90	159	228	292	307	26	167	370	518	599
Overall misclassified fraction		0.18	0.12	0.19	0.24	0.27	0.27	0.27	0.27	0.3	0.33	0.22	0.29	0.31	0.33	0.34
Predictions of an effect misclassified fraction		NA	0.16	0.29	0.33	0.37	NA	0.29	0.2	0.29	0.35	NA	0.24	0.25	0.29	0.33
Predictions of no effect misclassified fraction		0.18	0.11	0.13	0.15	0.16	0.27	0.27	0.28	0.31	0.31	0.22	0.31	0.33	0.35	0.34

The environment features for all tests were ''relative accessibility'', ''residue relative phylogenetic entropy'' , and ''neighborhood relative *B*-factor'', and the single categorical feature was ''unusual amino acid''.

[a] The fractions of the test mutations with too little training data for these features was 0.06.

[b] The fractions of the test mutations with too little training data for these features was 0.02.

[c] The fractions of the test mutations with too little training data for these features was 0.21.

tested the accuracy of predictions made using lac repressor mutation training data and lysozyme mutation test data, or *vice versa*. The accuracy here was slightly worse than the accuracy in the homogeneous cross-validation tests but still very significant. For all predictions (i.e. those made with a confidence level of 0.50 or better), the misclassification rates are 0.28, 0.35, 0.33. for lac repressor mutations modeled on lac repressor structure, the lac repressor mutations modeled on the purine repressor structure, and the lysozyme mutations modeled on the lysozyme structure, respectively (Figure 3(b), Table 5). As with the homogeneous cross-validation, the accuracy improves with increasing confidence of predictions to about 0.10-0.27 for the more confident predictions (Table 5, minimum prediction probability 0.9, Figure 3(b)). Again, predictions on the lac repressor data modeled with the purine repressor structure and lysozyme training data were not more accurate for structural neighborhoods exhibiting higher levels of sequence similarity. For comparison, the level of accuracy for our two-state predictions (effect or no effect) in the heterogeneous cross-validation is comparable to the accuracy of residue solvent accessibility predictions scored for two states (buried or exposed) or secondary structure predictions scored for three states (helix, sheet, and other).[29]

A shortcoming of the use of as many as seven predictive features (Table 5) in the heterogeneous cross-validation was a comparative dearth of lysozyme training mutations with appropriate feature values for making predictions on the lac repressor mutations (0.42 and 0.72 lac repressor test mutations modeled on the lac repressor and purine repressor structures, respectively, with too few lysozyme training mutations to make a prediction). This limitation could be overcome by considering a probablistic model with three environment features and only the most general polymorphism-specific categorical feature, namely unusual amino acid. The results were almost as good as those from the three categorical feature model, and predictions could be made for nearly all mutations (0.06 and 0.21 failure rate for lac repressor mutations modeled on the lac repressor and purine repressor structures, respectively) (Figure 3(b) and (c), Tables 5 and 6). For the probabilistic model we describe, more complete characterization of each test mutation through the use of a greater number of features, will always demand more training data.

Application of the probabilistic model to the SNP surveys from Case Western Reserve University and the Whitehead Institute

The significant accuracy of the predictions made on the lac repressor and lysozyme mutations, even with the heterogeneous cross-validation on the purine repressor model of the lac repressor, indicated that it would be reasonable to apply the modeling and predictions to the nsSNPs in the recent CWRU and WI SNP surveys.[1,2] The polymorphic residues in these data sets were mapped onto structures of homologous proteins by completely automated procedures. Applying strict criteria for sequence similarity between the target protein and the model protein (Materials and Methods), we find that 23 % and 39 % of the nsSNPs in the CWRU and WI data sets, respectively, could be represented by residues in crystal structures from the PDB/RCSB. These values could be increased to 45 % and 50 %, respectively, by inclusion of theoretical models and NMR structures from the PDB/RCSB, but these additional models were not used for predictions. The rate of homologous matches in the structure database for the WI SNPs is higher than the estimates of 30 % for other genome-wide surveys,[22] probably due to the historical emphasis on studying the proteins chosen for the SNP surveys for clinical reasons. We used the combined lac repressor and lysozyme mutations as a training data set in a probabilistic model for anticipating which polymorphisms would affect function that included the environment features ''relative accessibility'', ''relative phylogenetic entropy'', and ''relative neighborhood *B*-factor''. Two separate groups of categorical features were selected by maximum likelihood for use with these environment features. One was selected from a group that contained the near heterogen atom and near interface features to describe polymorphisms near ligands or subunit interfaces, and the other was selected from a group with these two special cases omitted (Table 7). The probabilistic model was applied to each of the structural models for the CWRU and WI polymorphisms to yield a predicted probability that each would affect protein function (Table 7). Some modeled polymorphisms (2-22 % CWRU and 9-10 % WI, depending of whether features ''A'' or ''B'' were used) could not be assigned probability values due to an insufficient number (i.e. less than four) of similar-valued mutations in the lac repressor and lysozyme training data sets (see also heterogeneous cross-validation section and Materials and Methods). With a few exceptions, the probability values computed with features ''A'' and features ''B'' were very similar (Table 7).

CWRU and WI polymorphisms with significant likelihood of affecting function

The analysis of functional effects of the nsSNPs from the CWRU and WI surveys indicates that while most are unlikely to affect function, a few of them are likely to affect function (Table 7, top part of A and B, probability $>= 0.5$ with Features A, Figure 4, Figure 5) and we discuss the top three from each survey. One high scoring polymorphism, an alanine residue to serine in dopamine beta-hydroxylase (DBH), was detected in both SNP surveys. The DBH structure was represented by the

Table 7. A selection of non-synonymous SNPs from the CWRU and WI surveys with high and low predicted probability of affecting function

								Features A[a]			Features B[b]		
UID	Pos	aa1	aa2	hugo	snp-handle	Freq	pdb	str_id	Prob	npoints	Rank	Prob	npoints	Rank
A. *Modeled CWRU nsSNPs*														
DBH	304	A	S	DBH	CHAK00402	0.1	1phm	0.33	0.80	10	1	ND[c]		
M59783	42	H	L	ALDR1	CHAK00110	0.05	2acs	1.00	0.75	4	2	0.95	38	1
D28235	488	E	G	PTGS2	CHAK00333	0.05	1cx2	0.79	0.71	146	3	0.66	122	4
DBH	303	L	P	DBH	CHAK00401	0.05	1phm	0.19	0.69	13	4	ND3		
M32332	37	A	T	ICAM2	CHAK00582	0.05	1zxq	1.00	0.67	128	5	0.67	109	3
J00098	126	D	H	APOA1	CHAK00131	0.05	1av1	1.00	0.61	44	6	0.61	44	5
M33107	145	Q	E	KLK1	CHAK00596	0.35	1sgf	0.71	0.60	5	7	0.57	7	6
M33108	193	V	E	KLK1	CHAK00598	0.05	1sgf	0.83	0.53	218	8	0.46	199	9
M33107	77	R	H	KLK1	CHAK00595	0.05	1sgf	0.79	0.51	277	9	0.44	273	12
M61887	130	C	W	SELE	CHAK00439	0.05	1esl	1.00	0.50	88	10	0.50	88	7
X15324	244	L	R	AGT	CHAK00093	0	1atu	0.18	0.50	36	11	0.50	40	8
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J00098	184	R	P	APOA1	CHAK00132	0.05	1av1	1.00	0.08	12	42	ND[c]		
M33108	186	K	E	KLK1	CHAK00597	0.25	1sgf	0.67	0.07	72	43	0.13	39	30
DBH	197	A	T	DBH	CHAK00396	0.1	1phm	0.50	0.06	34	44	0.07	43	34
L13436	771	Q	E	NPR3	CHAK00907	0	3lck	0.57	0.05	41	45	0.08	150	33
M59783	288	T	I	ALDR1	CHAK00116	0.05	2acs	1.00	0.04	68	46	0.13	249	29
B. *Modeled WI nsSNPs*														
DBH	304	A	S	DBH	WIAF-10793	5-15 %	1phm	0.33	0.80	10	1	ND[c]		
L32771	1820	M	I	F5	WIAF-11349	5%	1kcw	0.36	0.73	157	2	0.71	146	4
L20590	291	F	S	ANX3	WIAF-11441	ND	1axn	1.00	0.71	185	3	0.66	150	5
V00520	105	S	C	GH1	WIAF-10591	5%	1hwg	1.00	0.64	39	4	0.61	31	7
M90103	168	E	K	MPL	WIAF-11243	5%	1cn4	0.43	0.63	133	5	0.56	112	10
L32769	1685	T	S	F5	WIAF-11536	ND	1kcw	0.62	0.63	194	6	0.41	140	13
M10014	191	G	R	FGG	WIAF-11492	ND	1fzf	1.00	ND[c]			0.65	20	6
L20589	251	P	L	ANX3	WIAF-11120	5%	1axn	1.00	0.51	87	7	0.39	62	14
M10014	410	M	V	FGG	WIAF-11477	ND	1fzf	1.00	0.50	4	8	0.86	14	2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J02933	295	V	D	F7	WIAF-11117	5%	1cvw	1.00	0.02	43	49	0.14	197	29
M21999	589	L	Q	F13A1	WIAF-11052	5%	1qrk	1.00	0.00	21	50	0.00	24	46
M22000	652	Q	E	F13A1	WIAF-11088	15-50 %	1qrk	0.92	0.00	22	51	0.00	29	47
M32996	331	G	S	CETP	WIAF-11016	5%	1bp1	0.57	0.00	37	52	0.00	48	48
M68516	105	K	E	PCI	WIAF-11196	5%	2pai	1.00	0.00	7	53	0.00	56	50

The Table shows all nsSNPs with greater than a probability of 0.5 of affecting function with Features A (top), and the five nsSNPs with the lowest probability of affecting function with Features A (bottom). The combined lac repressor and lysozyme mutations were used as training data for computing probability values. The column headings refer, in order, to: The unique identifier for the coding DNA sequence in Genbank, the position of the polymorphic amino acid in the translated coding sequence, the alternative amino acids encoded by the polymorphism, the HUGO locus identifier, the SNP identifier, the allele frequency, the PDB identifier, the fraction of residues in the structural neighborhood that are identical in the target protein and in the model, the probability of an effect on function using the features indicated, the number of mutations in the training data set with appropriate feature values for computing the probability of an effect on function, and the rank of the probability value using the features indicated.

[a] Features A were ''buried charge'', ''unusual amino acid'', ''unusual amino acid by class'', and ''turn breaking'' selected by maximum likelihood from among ''buried charge'', ''unusual amino acid'', ''unusual amino acid by class'', ''turn breaking'', ''conserved position'', ''helix breaking'', and ''near conserved position''.

[b] Features B were ''buried charge'', ''unusual amino acid by class'', ''near heterogen atom'', and ''near interface'', selected by maximum likelihood from among the same group used to select Features A combined with ''near heterogen atom'', and ''near interface''

[c] ND refers to predictions for modeled nsSNPs that could not be made because of insufficient training data (less than four training mutations) with similar feature values.

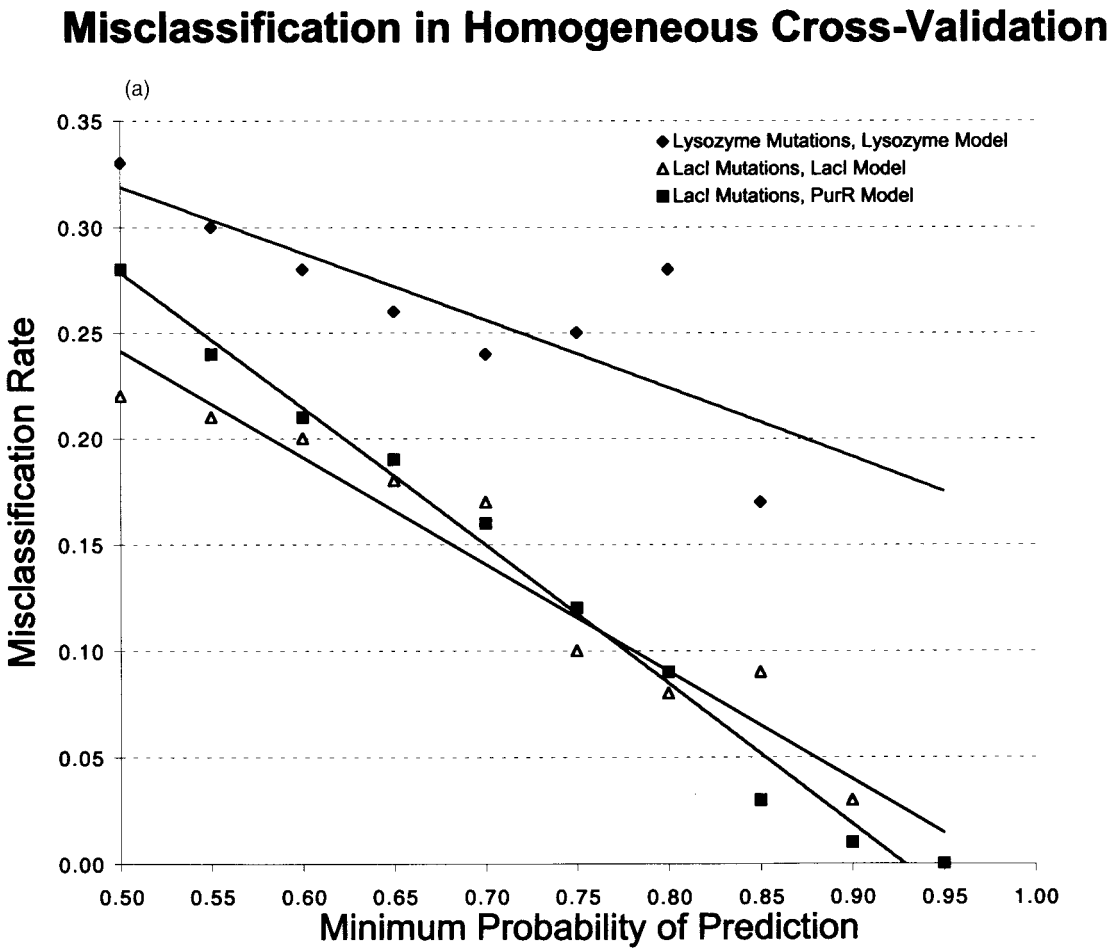

Figure 3 (*legend shown on page 698*)

crystal structure of the homologous peptidylglycine monooxygenase (PDB: 1phm), and this moderately prevalent SNP is modeled as an arginine residue that is only partially exposed to solvent (relative accessibility of 0.27). The model residue has low relative *B*-factors (−0.22 SD residue, −3.28 SD neighborhood), and intermediate to high phylogenetic entropy values (1.41 SD residue, −0.10 SD neighborhood). In spite of the great chemical difference between arginine in the structural model and either an alanine or serine residue, the serine from the polymorphism but not alanine or arginine residue represents an unusual amino acid in the phylogeny of nine sequences. Polymorphisms in aldose reductase (ALDR1) and prostoglandin synthase 2 (PTGS2) from the modeling of the CWRU nsSNPS also appear likely to have an effect on function. The histidine to leucine residue change in ALDR1 is modeled at a histidine residue of an aldose reductase structure (PDB: 2acs) and has very low solvent accessibility (2.0 Å^2 or 0.01 relative accessibility), relative *B*-factors (−0.79 residue, −3.00 neighborhood), and relative phylogenetic entropy (−0.75 residue, −3.67 neighborhood). It represents a buried charge but does not represent an unusual amino acid. It is however part of the sequence matching the PROSITE entry PS00079 (aldoketo_reductase_1) of the enzyme's active site. The glutamic acid to glycine polymorphism in PTGS2 modeled as a glutamic acid on the structure of the mouse cyclooxygenase-2 (PDB: 1cx2) represents a change at a "conserved position" in a relatively conserved neighborhood (−1.58 s.d.) that is also inaccessible (12 Å^2 or 0.06 relative accessibility) with a low neighborhood relative *B*-factor (−1.50 SD). In the WI data set, the polymorphism with the second highest score is a methionine to isoleucine residue change in Factor V (F5), which when modeled on the structure of the related ceruloplasmin (PDB: 1kcw, 37% sequence similarity overall) represents the introduction of an unusual amino acid (phylogeny of eight sequences) in a buried (0 Å^2, 0.0 relative accessibility), rigid (relative *B*-factors: −0.86 residue, −2.27 neighborhood)

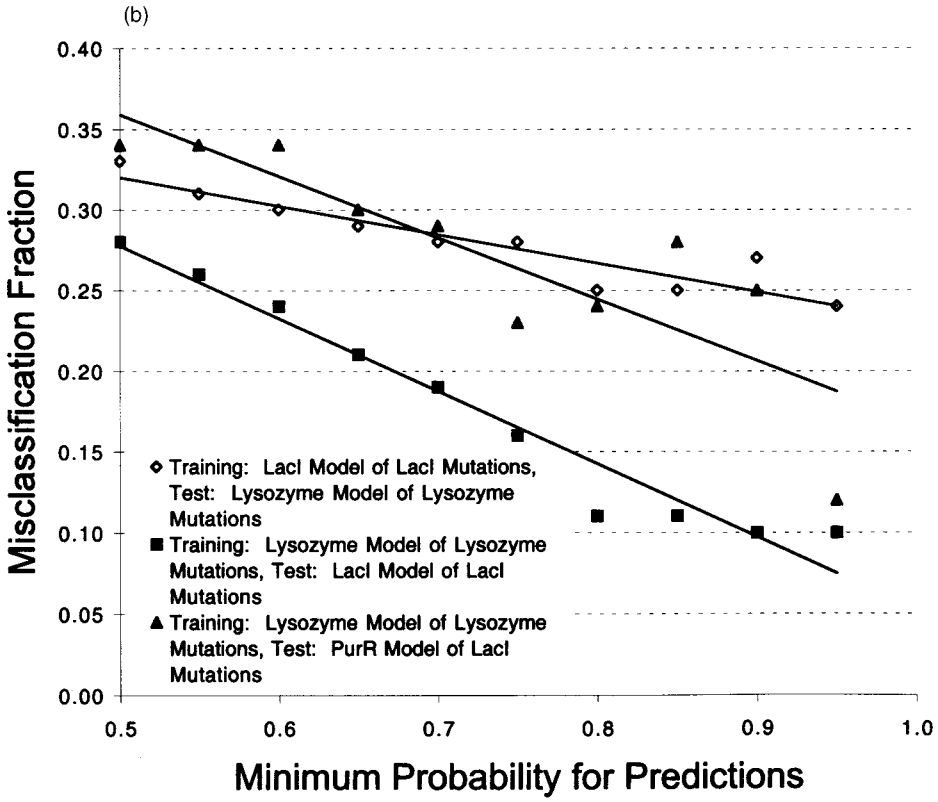

Figure 3 (*legend shown on page 698*)

and conserved (relative phylogenetic entropy: −1.48 residue, −0.73 neighborhood) residue and neighborhood. The third most likely polymorphism to affect function is a phenylalanine to serine change in annexin III also in a buried (0 Å^2, 0.0 relative accessibility), rigid (relative *B*-factors: −0.82 residue, −1.91 neighborhood), and conserved (relative phylogenetic entropy: −0.61 residue, −0.14 neighborhood) part of the structure (PDB: 1axn, 100 % sequence identity). The serine of the polymorphism represents an unusual amino acid in a phylogeny of 51 sequences.

Conversely, some of the polymorphisms have low probability values for an effect on function (Table 7A and B , bottom part, 5 models with lowest probability in each data set). As expected, these polymorphisms are modeled as solvent accessible residues in parts of structure that are relatively variable in sequence, and have high relative *B*-factors; and the values of their categorical parameters do not indicate an effect on function.

Population genetics of nsSNPs predicted to affect function

If the rate of non-synonymous cSNPs occurring in the population is proportionately lower than the rate of synonymous cSNPs due to the more immediate and potentially deleterious functional consequences of amino acid variation, it might be expected that polymorphisms with a greater probability of affecting function would have disproportionately high representation among the nsSNPs with low allele prevalence.[1,2,30] We analyzed this possibility, again using ANOVA and F-statistics, by examining the average probability values for predictions in three SNP prevalence categories (Table 8). Indeed, the predicted probability of an effect on function is lower in the high prevalence categories for both data sets but, with so few high frequency alleles, it does not reach statistical significance. Moreover, the medium frequency allele in the WI data (N = 9) in the WI data has a higher average probability of affecting function than the low frequence allele (N = 26). When the two data sets are combined, the P-value improves to 0.36.

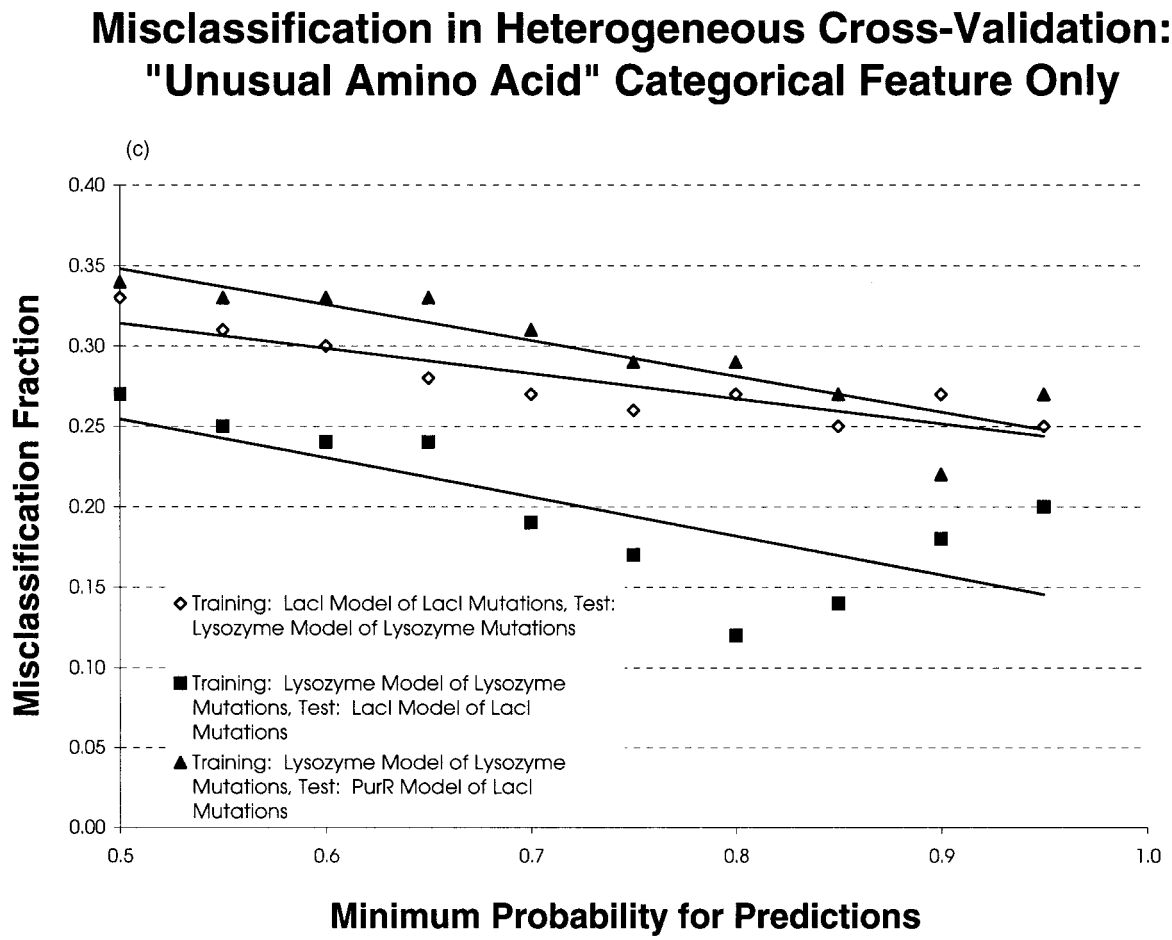

Figure 3. Cross validation tests of the probabilistic model. All figures represent plots of the misclassification fraction rate (see also Table 4) for predictions made using the probabilistic model at the minimum confidence levels indicated. Misclassification rates are plotted in increments of 0.05 in prediction confidence level units, instead of the 0.10 intervals in the table. The trend in prediction accuracy for test was also crudely fit by linear regression. The environment features for all tests were ''relative accessibility'', ''relative phylogenetic entropy'', and ''neighborhood relative *B*-factor.'' (a) Homogeneous cross-validation. For this test, 90% of the mutations from a single data set were used as the training data, and the remaining 10% of the mutations from the same data set were used as the test data (see also Table 4). The categorical features were chosen by the maximum likelihood method. They were ''unusual amino acid by class'', ''near heterogen atom'', and ''interface'', for the lac repressor mutations modeled on the lac repressor, and ''buried charge'', ''unusual amino acid by class'', ''helix breaking'', and ''near conserved'' for the lysozyme mutations modeled on the lysozyme structure, and ''buried charge'', ''helix breaking'', ''near conserved'', and ''near interface'' for the lac repressor mutations modeled on the purine repressor. (b) Heterogeneous cross-validation with maximum likelihood selection of categorical features. The outcome of the lysoyzme mutations were predicted with the lac repressor mutations as training data and *vice versa* (see also Table 5). When the lysozyme mutations modeled on the lysozyme structure were used as training data, the categorical features selected by maximum likelihood were ''buried charge'', ''unusual amino acid by class'', ''helix breaking'', and ''near conserved position''. When the lac repressor mutations modeled on the lac repressor structure were used as training data, the categorical features selected by maximum likelihood were ''buried charge'', ''unusual amino acid by class'', ''turn breaking''. (c) Heterogeneous cross-validation with only one categorical feature. The same test as performed in (b) with the single categorical feature unusual amino acid for all comparisons (see also Table 6).

Overall, the structure-based probabilistic analysis (Tables 7 and 8) suggests that proportions of modeled polymorphisms expected to affect function are 0.31-0.32 and 0.26-0.27 (depending on which categorical features are used) for the CWRU and WI data sets respectively. These values are considerably lower than the approximately 0.41-0.44 of the mutations that affect function in the combined lysozyme and lac repressor data sets. Taking together with the CWRU and WI authors' estimates of about 24,000-40,000 heterozygous loci in a typical person due to nsSNPs, these results suggest most people are expected to have about 6240-12,800 nsSNPs with effects on protein function due to amino acid variation.

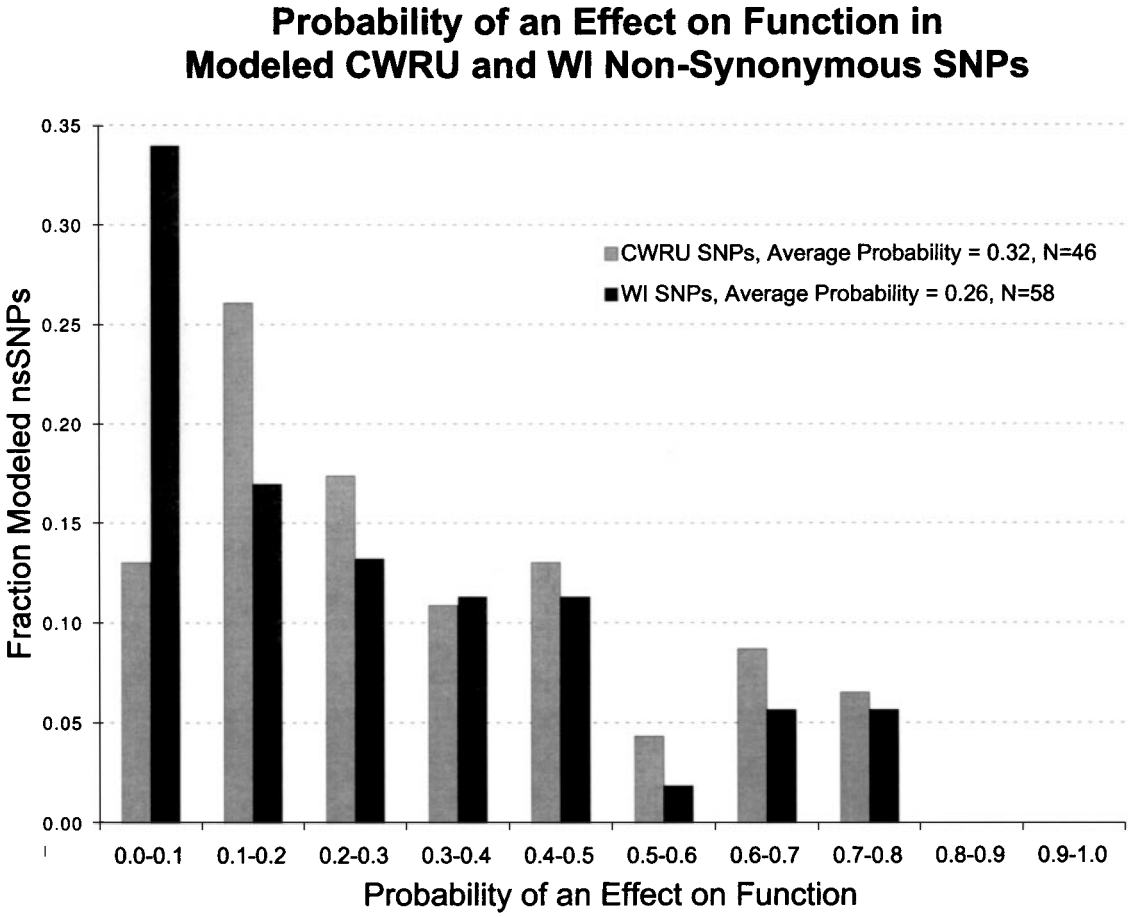

Figure 4. Histogram of probability values for an effect on function in modeled CWRU and WI nsSNPs. Probability values were determined from a probabilistic model using the combined modeled lysozyme and lac repressor mutation data sets as training data and the environment features ''relative accessibility'', ''relative phylogenetic entropy'', and ''neighborhood relative *B*-factor'' (see also Table 5). In addition, the categorical features (Features A, Table 5) ''buried charge'', ''unusual amino acid'', ''unusual amino acid by class'', and ''turn breaking'' were chosen by maximum likelihood applied the combined data set. To generate the histogram, probability values for modeled polymorphisms were segregated into bins having 0.1 probability units width, and the number of polymorphisms in each bin was graphed as the proportion of all modeled polymoprhisms in the corresponding data set (out of 46 for CWRU and 58 for WI).

Discussion

Others have noted the importance of structural, physical, and phylogenetic aspects of residues for predicting their tolerance to amino acid substitution, but as far as we know, the current study and the report from Sunyaev *et al.*[17] are the first attempts to enumerate an extensive list of predic-

Table 8. Segregation of the probability of an effect on function over nsSNP prevalence classes

Data summary	CWRU		WI		CWRU+WI	
Allele prevalence	Count	Average probability	Count	Average probability	Count	Average probability
Low (<=0.05)	39	0.32	26	0.22	65	0.28
Med (0.05-0.15)	4	0.28	9	0.30	13	0.29
High (>0.15)	3	0.27	10	0.17	13	0.19
All alleles	46	0.32	45	0.27	91	0.27
P-value (ANOVA)	0.86		0.41		0.36	

Data for predictions made with features ''A'' (Table 7). ANOVA compares estimates of the variance in the predicted probability among three allele prevalence classes by computine an F-value. P-Value is probability of the null hypothesis, i.e. no statistical difference in average predicted probability among alleles in the different frequency groups.

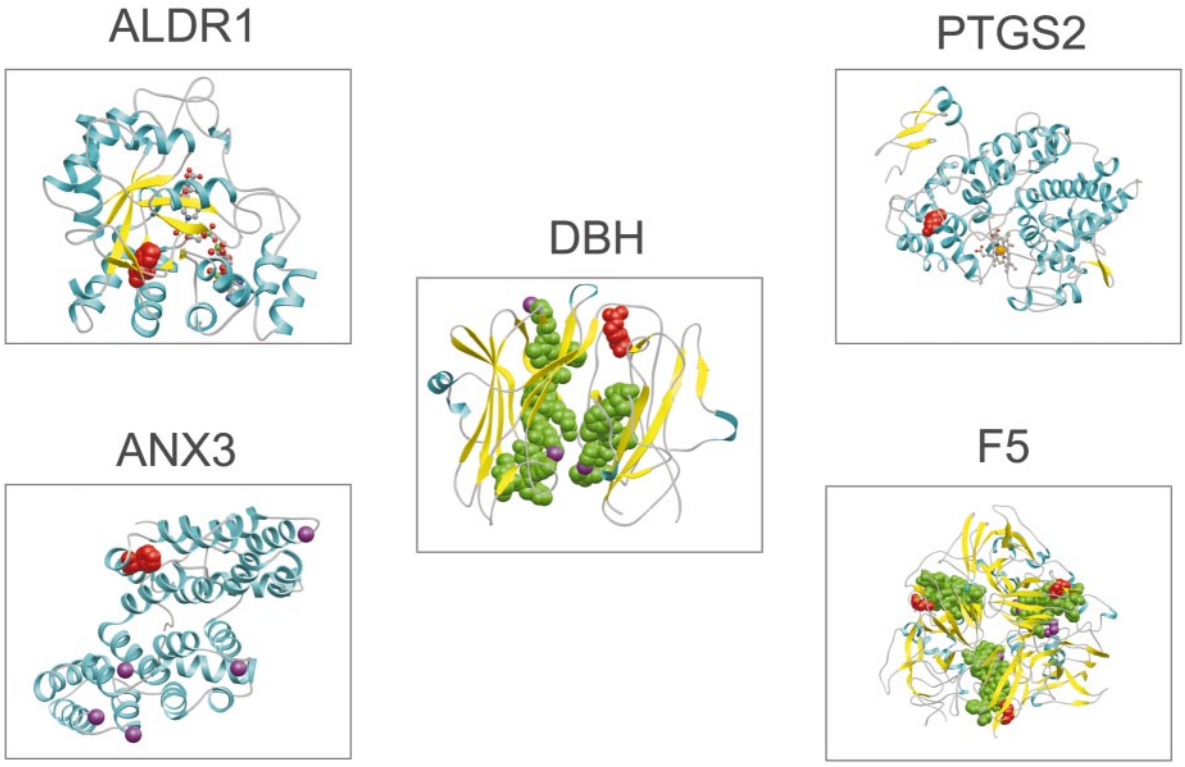

Figure 5. Images of the structural models for polymorphisms in the CWRU and WI SNP surveys with high probability values of affecting protein function. Each structural model (either of the target protein or a protein homologous to the target protein, see Table 5) is shown in a standard RIBBONS representation[51] (Target Proteins: PDB ID pairs were dopamine beta-hydroxylase (DBH):1phm, aldose reductase (ALDR1): 2acs, prostoglandin synthase 2 (PTGS2): 1cx2, Factor V (F5): 1kcw, and annexin III (ANX3): 1axn). The model residue for each nsSNP is shown in red space filling representation. Other annotations include matches to PROSITE patterns (green space filling representation), Calcium atoms (purple spheres in ANX3 model), Copper atoms (purple spheres in DBH and F5 models) and ligands (ball-and-stick representation, heme for PTGS2, NAP for ALDR1). Although not indicated in the figure, the polymorphism in ALDR1 is part of a PROSITE match (PS00079) to the enzyme's active site.

tive features and to develop a formalism for quantitatively evaluating their performance in predictions. We found a group of continuous and categorical, structure and sequence-based features that are strong predictors of an effect on function for mutations in the lac repressor and in lysosymze. For the lac repressor mutations, when the feature values were derived from the structure of the homologous purine repressor with only 30 % sequence identity, they remained very strong predictors of effects on function. In cross-validation studies, we found that feature values could be calibrated for significantly accurate predictions on both the lac repressor mutations and the lysozyme mutations. This generic calibration suggested the use of the predictive features for estimating a probability that nsSNPs in the recent CWRU and WI surveys would affect function. From the estimated probability values, we expect that the proportion of nsSNPs affecting function is substantial, about 26-32 %, and that nsSNPs may account for a significant source of molecular functional variation in human populations.

Through statistical analysis (Tables 2 and 3) and cross-validation (Tables 4, 5 and 6, Figure 3), we attempted to assess the generality of the predictive features for indicating effects on protein function. The similar statistical ranking of the features on the lac repressor and lysozyme mutations supports their general nature (Table 2), as do the cross-validation tests (Tables 5 and 6). But the absolute strength of the prediction significance in the two mutation data sets differs. These differences might be related to intrinsic differences in the tolerance of the two proteins to mutation, or to intrinsic differences in the experimental criteria for determining functional effects in the two data sets.

In the assessment of the functional consequences of natural amino acid polymorphisms, the structure-based approach we describe provides a significant enhancement over predictions based solely on the knowledge that about 40 % of random amino acid mutations affect function. To begin, the procedure is more accurate. The accuracy based on randomly assigning 40 % of the mutations to an effect on function and 60 % to no effect on function would lead to an overall misclassification rate of 0.48, with misclassification rates of predicting an effect or no effect on function of 0.60 and 0.40, respectively. These values are considerably worse

than those observed with the probabilistic model (Tables 4, 5 and 6, Figure 3), especially for the higher confidence predictions. Indeed, the improved accuracy with increasing prediction confidence validates the idea of ranking candidate polymorphisms according to their probability of an effect on function. This kind of ordering is impossible in a method that only considers the average likelihood of an effect on function among random mutations. SNP-derived amino acid polymorphisms are strongly non-random (see disproportion of nsSNPs in rare alleles in the work by Cargill *et al.*[1] and Halushka *et al.*[2]); and to the extent that the structure-based approach accurately captures the functional consequences of mutations, the current methodology is vastly more informative for this class of amino acid variation than simple extrapolation of the aggregate behavior of random mutations.

We would therefore claim not to have found a universal calibration of the features for predicting effects on function on all proteins, but rather one that is likely to be reasonably robust for well-behaved globular proteins for which there are homologous crystallographic structural models. Other structural classes of proteins (or folds) might have biological functions with different tolerance to amino acid polymorphisms. For example, the tolerance to mutation of structurally extended or dynamic but critical parts of proteins may not be revealed by low crystallographic *B*-factors. Predictions of effects on membrane proteins may require a different calibration of the feature values than are used for the small globular proteins.

Nevertheless, the features and approach we describe can be adapted for these circumstances through the our formalism, and even without modification, the features may be remarkably general (Table 1). For example, the use of relative measures for the *B*-factor and entropy features instead of absolute measures may allow respectably accurate comparisons among many different proteins and their distinct structures and folds. Conserved residues and neighborhoods will always be intolerant to amino acid substitution, and burying charges in hydrophobic cores will always be destabilizing. The methods would be strengthened by more unbiased mutation data sets in proteins with a wide range of fold classifications.[18,19] It might also be particularly helpful to have experimental data pertaining to the effects on function of exchanging each amino acid for every other amino acid in a wide variety of structural environments.

However, even as more complete sets of structure-based features, phylogeny-based features, and new mutation data become available, there may be an intrinsic limitation to the approach we describe. A long history of mutagenesis studies suggests that the effects of some mutations will not be understood in terms of the features that are analyzed in this work (for a discussion, see Ness *et al.*,[31] for examples, see Crameri *et al.*,[32] Chang *et al.*[33] and Spiller *et al.*[34]). For example, solvent accessible mutations with a high degree mobility can affect protein solubility, interactions with other molecules, or internal dynamics and folding, or even transcription and translation all in ways that may not be identified by structure or phylogeny-based features.

The current methodology finds potential application in the understanding of genetic effects in the development and improvement of pharmaceuticals.[35] At the outset of drug design, it can help anticipate functional variability in protein drug targets due to nsSNPs that might cause heterogeneous interactions of lead compounds. During the clinical phase of developing candidate drugs, it can help provide an explanation for variability in efficacy or in toxicity that may be due to genetic variability in pharmacological pathways. Finally, once an approved drug has been used by thousands of patients, the method can be used retrospectively to develop hypotheses for identifying particular polymorphisms that will serve as indicators of a drug's efficacy; and screening patients for critical polymorphisms could become a requisite aspect of drug prescription.

The genetic origins of nsSNPs are likely to be heterogeneous and may reflect neutral genetic events that become fixed, selective pressures that existed in another time, population bottlenecks, or genetic configurations that confer advantages in the current environment.[10] Unlike the lac repressor and lysozyme mutations, they are biased and their functional history can be difficult to recreate. Due to selection against the deleterious effects of amino acid substitutions, nsSNPs remaining in the population are less likely to have an effect on function than unbiased mutations
view supports this conclusion by estimating that an average of 26-32 % of the CWRU and WI nsSNPs encode amino acid variation that might not be tolerated in the modeled structural enviornment (Figure 4) (compared with 41 % and 44 % for the lac repressor and lysozyme respectively). Although the number of high frequency nsSNP alleles was limiting in our statistical tests, the greater predicted probability of effects on function for nsSNPs with lower allele frequency suggests that our predictions reflect the expected influence of genetic history on function. Confirmation of our predictions will await more nsSNP allele frequency and functional data, and of course, experimental validation of our computational methods.

Our estimates of the effects of natural polymorphisms suggests that about 6240-12,800 nsSNPs will cause heterogeneous protein function. If our predictions are approximately correct and the functional nsSNPs are distributed uniformly over the estimated 30,000[36,37] proteins encoded by the human genome, this level of functional heterogeneity would be manifest in many biochemical and signal transduction pathways. Non-coding and synonymous cSNPs affecting functions like

transcription and splicing will also contribute to heterogeneous biological function. And determining the relative functional importance of these SNPs and the non-synonymous ones is an important goal of ongoing genomic studies. In the meantime, the tools, formalism, and estimates we describe for the analysis of the nsSNPs can help understand the functional basis of natural biological variation in the human population.

Materials and Methods

Data sets

In order to evaluate structure-based features as potential indicators of effects on protein function, we considered the previously described exhaustive mutagenesis of lac repressor (about 4000 mutations) and T4 lysozyme (about 2000 mutations).[16,30] These data sets are unbiased in the sense that: (i) the mutations included 12 or 13 changes for each residue (depending on the identity of the natural residue) or roughly two-thirds of the possible mutations in the two proteins; (ii) the 12 or 13 changes represent substitutions into each of the recognized chemical classes of amino acid residues; (iii) the measurements of the effects of the mutations on function were standardized; and (iv) the mutations were not selected for their effects on function. For our analysis, outcome of the mutations was viewed in a binary fashion: either the mutation affected protein function or it did not. For example, a leucine to histidine mutation at position 133 of lysozyme caused an effect at 37 °C but not at 25 °C. This mutation was scored as having an effect. In spite of the different functions of the lac repressor and lysozyme, and the different criteria for determining effects on function, the proportions of mutations in the data sets with effects on function were very similar, 0.41 and 0.44, respectively. The mutations were mapped onto the structure of lysozyme (PDB ID: 7lzm) or the structure of lac repressor (PDB ID: 1lbh) carboxy-terminal domain as appropriate through a sequence correspondence established by a BLAST alignment. The lac repressor mutations were also mapped onto the crystal structure of the purR repressor (PDB ID: 2pua), sharing about 30 % sequence identity with the lac repressor, also through the sequence correspondence established by a BLAST alignment. The BLAST sequence correspondence differs from the structural correspondence defined by the FSSP database at seven residues out of about 333.[39]

Structural neighborhoods

We suspected that the primary effects of amino acid polymorphisms would be local, and focussed our analysis on structural models that included the polymorphic residue itself and residues that are within a small distance of the polymorphic residue (for an example, see Bordo[40] for a previous implementation of this concept). Collectively, these nearby residues are termed the structural neighborhood. Formally, we defined the structural neighborhood as the collection of residues having at least one atom within 5 Å of at least one atom in the model for the polymorphic residue.

To assess the structural similarity of homologous neighborhoods, we examined pairs of structurally aligned homologous proteins from the FSSP database to derive a comprehensive database of homologous structural neighborhoods.[39,41] Each collection of subject and representative structures in the FSSP database was cleaned by eliminating subject structures that were essentially the same in sequence. The eliminated structures included cases of crystal structures of several mutant forms of subject proteins, cases of crystal structures of varying resolution of the same subject protein, and cases of crystal structures of subject proteins solved with a selection of different ligand molecules. For each residue in each representative structure in the cleaned database, an equivalent residue was found in each of the subject structures through the sequence correspondence established by a BLAST alignment. The structural neighborhood for each residue in the representative structure was also determined, and related to a corresponding structural neighborhood in the subject structures also through the BLAST alignment. Phylogenetic data, secondary structure, and accessibility for each residue from the representative structures and its structural neighborhood were supplied by the appropriate HSSP file.[42] In all, about 350,000 representative structure residues and structural neighborhoods, their corresponding subject structures' residues and structural neighborhoods, their *B*-factor, their phylegenetic data, and their solvent accessibility were collected.

Using the FSSP coordinate transformations for the subject structures, the r.m.s.d. values for C^{α} atoms in corresponding residues of all homologous structural neighborhods were computed. These values were then compared to the fraction of identical residues in the homologous structural neighborhoods. With our procedure, the r.m.s. distance values underestimate the accuracy of representing a structural neighborhood with the corresponding structural neighborhood in a homologous proteins since the structural alignment for each structural neighborhood pair was not optimized beyond the alignment for the entire, homologous proteins. Nevertheless, it is clear that alpha carbon positions in homologous structural neighborhoods typically resemble each other even for low levels of sequence similarity, e.g. 1.5 Å and 2.0 Å average r.m.s.d. for 50 % and 30 % sequence identity respectively (Figure 1). The relationship between neighborhood sequence similarity and either the average structural similarity (Figure 1) or the distribution of the r.m.s.d. values (Figure 1, inset) can be used to provide confidence in the structural accuracy of the models of structural neighborhoods used to analyze polymorphisms. We note that a single calibration of structural neighborhood similarity as a function of sequence similarity appears to be valid for related proteins whether they are very similar in sequence or quite different (compare 0.70-0.95 sequence identity curve with 0.25-0.40 sequence identity curve). Since our structural neighborhood alignments are derived from the structural alignments of entire proteins,[39,41] it is not surprising that the r.m.s.d. values for our structural neighborhoods are in accordance with similar analyses performed for homology modeling of entire proteins (for example, see Table 1 in the work by Guex *et al.*).[22]

The predictive features

We considered an array of structure-based and sequence-based features of amino acid polymorphisms

that might serve as generalized predictors of effects on function. Some features were known from the literature, others were of our own devising. The features fell into two broad classes.

Features in the first class convey an intrinsic tolerance of a particular residue position in a protein structure to amino acid substitution. These features include the molecular rigidity represented by the crystallographic *B*-factor of the model polymorphic residue and its structural neighborhood,[13,43] the degree phylogenetic conservation represented by the phylogenetic information content or entropy for the polymorphic residue and its structural neighborhood (for example, Walker *et al.*[44] and Sander & Schneider[40]), the solvent accessibility for the polymorphic residue (see Bourie *et al.*[11]), and its proximity to ligand or cofactor molecules (Table 1). For the features pertaining to the crystallographic *B*-factor and phylogenic entropy, we considered relative measures to facilitate normalization of comparisons among different structures, e.g. for structures with very different average *B*-factors or phylogenetic entropy. Together, the six most generic of the features in this class were termed environment features. They were relative *B*-factor, neighborhood relative *B*-factor, relative phylogenetic entropy, neighborhood relative phylogenetic entropy, solvent accessibility, and relative solvent accessibility as defined by Rost & Sander.[45] All six environment features are continuously-valued. The remaining features in this class near heterogen atom (i.e. an atom in a non-standard group in a PDB file), near interface, and ''near conserved position'' are categorically-valued, assuming the value ''yes'' or 1 if certain structural criteria are met and ''no'' or 0 if not.

Features in the second class are properties of a polymorphism that indicate its potential effect on function due to the chemical differences of the alternative amino acids, sometimes in combination with the structural environment. These features are all categorically-valued. Among others, they include designations for a polymorphism that introduces a charged amino acid at a residue position that is inaccessible to solvent in the model structure (buried charge),[12] polymorphisms that substitute a glycine or a proline amino acid in a region of helical secondary structure in the model (helix breaking), polymorphisms that occur at the conserved glycine or proline in a turn (turn breaking), and polymorphisms that introduce an amino acid that is not represented in the phylogenetic profile of the polymorphic residue (unusual amino acid) or its amino acid class (unusual amino acid by class)[46] (Table 1B). Amino acid substitution matrices (e.g. BLOSUM, PAM)[47,48] could also be used to provide information about allowable mutations; but because the unusual amino acid feature is evaluated for each polymorphic residue, it will likely more accurately integrate the empirical subtlety of each particular structural environment. None of the structural aspects of any of the features relies on the detailed atomic configuration of the model, making them suitable when only structures of homologous proteins are available.

Automated construction of models for polymorphic residues

The basis of all of our analysis is the mapping of a target polymorphic residue onto the crystal structure of a target protein or its homolog, followed by the assessment of a set of structure-based and sequence-based features for the modeled polymorphic residue. The correspondence between the target polymorphic residue and a residue in a homologous structure was determined by a BLAST (v. 2.0.6[49]) alignment, selecting from among possible alignments so as to optimize the overall sequence similarity between the target protein and the model protein. Only alignments with significant P-values (i.e. $<10^{-10}$) or a sequence similarity of about 30 % overall amino acid identity were considered informative. Once the model protein and residue were identified, the corresponding PDB file was analyzed to extract *B*-factors and construct structural neighborhoods. The atomic *B*-factors for each residue were averaged to define residue *B*-factors. These residue *B*-factors were normalized by subtracting their average value in the corresponding PDB chain, and dividing by the variance to derive the ''relative residue *B*-factor'' or used in the formula in Table 1 to derive the ''neighborhood relative *B*-factor'' feature. Proximity to ligand atoms, e.g. heterogen atoms and matches to PROSITE[50] motifs was also assessed. Phylogenetic data and additional structural information needed to evaluate the predictive parameters, including solvent accessibility and residue secondary structure, were extracted from the HSSP file[42] corresponding to the model structure selected from the PDB. In the HSSP files, the phylogenetic data consists of a multiple alignment of sequences from SWISS-PROT sharing at least 30 % amino acid identity with the sequence of the model structure. Phylogenetic entropy extracted from the HSSP files was normalized as for the *B*-factors. For residues modeled on the lac repressor, the purine repressor, and lysozyme structures, the values for accessibility and relative accessibility were normalized by averages and variances (as for the *B*-factors) determined for each protein separately. Relative accessibility was determined by dividing the model residue's solvent accessibility from the HSSP file by the maximum solvent accessibility for the model residue according established values.[45] For models of the polymorphic residues from the CWRU and WI SNP surveys, the values for accessibility and relative accessibility were normalized to typical average values (42 $Å^2$ and 0.26, respectively), and variances (44 $Å^2$ and 0.26, respectively) for these parameters. Current versions of the method always normalize accessibility and relative accessibility with average values and variances derived from the PDB structure used to model the polymorphic residue. The entire procedure of mapping mutations onto exact or homologous crystal structures was automated.

Probabilistic model

We used data sets of training mutations to estimate the probability that a test polymorphism in a target protein has an effect on its function. The training mutations were the unbiased mutations in lac repressor and lysozyme modeled on their respective crystal structures as described above. Approaches using crystal structures of homologous proteins for determining feature values for the training mutations were not studied here except for the homogeneous cross-validation tests of the lac repressor mutations modeled on the purine repressor. For each test polymorphism in a target protein, environment and categorical feature values were assessed from structural models of the target

protein when possible, or from the structure of a protein homologous to the target protein. The probability that the test polymorphism affects the function of the target protein was estimated as the fraction of mutations in the training data set with effects on their protein's function from among those that had: (i) their selected continuously-valued environment feature values within a specified stringency (measured as a cartesian distance, sometimes termed the bandwidth in local regression analysis)[51] from the selected environment feature values for the modeled test polymorphism; and (ii) categorical feature values that were identical to the categorical feature values for the modeled test polymorphism. We studied the stringency parameter for comparing environment feature values and found good performance with a cartesian distance of about one standard deviation (1 SD) between the selected environment feature values of the training mutations and the selected environment feature values of the target polymorphism. As commonly found in local averaging methods, stringency values significantly less than about 1 SD in the environment features caused too little averaging, while greater values caused undifferentiated predictions of effects on function (data not shown).[51] If there were fewer than four training mutations with appropriate feature values, no prediction was made.

Principal component analysis and correlation between environment features

The computation of correlation coefficients and principal component was performed as described by Lebart *et al.*[52] using environment feature values computed from the automated modeling from the lac repressor structure (PDB ID:1lbh), the purine repressor structure (PDB ID: 2pua), and the lysozyme structure (PDB ID: 7lzm). We also examined correlation coefficients and principal components for environment feature values in the database of structural neighborhoods. We gave special attention to the *B*-factor and phylogenetic entropy features since high degree of correlation would suggest that conserved residues or neighborhoods are necessarily more rigid, and conversely that variable residues or regions of a protein are more flexible. Like the the lac repressor and lysozyme cases (Table 3), moderate correlation was found. It is maximal for the correlation of the neighborhood relative *B*-factor and neighborhood phylogenetic entropy (0.35) and minimal for the correlation of relative *B*-factor and relative phylogenetic entropy (0.26). The full correlation and principal component analysis for the database of structural neighborhoods is available upon request.

Choice of features

The most complete probabilistic model would use all of the predictive features and might provide the most accurate basis for judging a polymorphism's functional effects, but it would also contain redundant descriptions of each polymorphism caused by the correlation between pairs of the parameters. It would also require a large amount of training data for calibration. The complexity of the model can be minimized by identifying a reduced set of the most important features to be used as predictors. We noted in the ANOVA (Table 2) and in the principal component analysis (Table 3) that a robust description of the environment of a modeled polymorphic residue should include one accessibility feature, one phylogenetic entropy feature, and one *B*-factor feature. Based on the relative predictive values of the features (Table 2), we chose relative accessibility, relative phylogenetic entropy, and for the *B*-factor, either relative *B*-factor or relative neighborhood *B*-factor. The tests that are reported use the relative neighborhood *B*-factor.

The choice of features used with each training data set could also be selected by a maximum likelihood method. Here, we used the probabilistic model to compute the accumulated log likelihood of the observed effects on function in the training data set, and compared this value to the log likelihood of the observed effects estimated solely with knowledge of the fraction of mutations in the training data with effects on function:

$$\text{Log likelihood ratio} = \frac{\sum_{\text{all mutations}} \ln(l_i)}{\sum_{\text{all mutations}} \ln(\bar{l}_i)}$$

Where:

$$l_i, \bar{l}_i = \begin{cases} p_i, \bar{p} \text{ , respectively, if mutation } i \text{ has an effect on function} \\ 1-p_i, 1-\bar{p} \text{ , respectively, if mutation } i \text{ does not have an effect on function} \end{cases}$$

p_i = the probability from the probabilistic model of an effect on function given the set of environment and categorical features

$\bar{p}$ = the fraction of mutations in the training data set with effects on function

.

The best combination of features was selected as the one affording the probabilistic model the greatest improvement in predicting the likelihood of the training data. Combinations of features causing too few training mutations with appropriate feature values for computing more than 10 % of the p_i values were not considered.

To reduce the number of combinations of features tested in the maximum likelihood procedure, we typically first identified the set of three environment features that optimized the likelihood of the training data, and in a second step, up to about four categorical features that together with the selected environment features optimized the likelihood of the training data. As with the ANOVA and the correlation tests, we found that the best set of three-environment features typically included one of the two accessibility features, one of the two phylogenetic entropy features, and one of the two *B*-factor features. With these environment features, the categorical features providing the maximum likelihood of the

observed data typically included the unusual amino acid, buried charge, and conserved position features, and for the tetrameric lac repressor with its bound ligand (IPTG) or the purine repressor bound to DNA and its ligand (6MP), the near interface and near heterogen atom features (see also Tables 4, 5, 6 and 7).

Computational considerations

To speed computation of probability values, the continuously-valued environment feature values were indexed according to the formula:

$$\text{E.V. Index} = Int\left(round\left(\text{e.v.}^{*}\frac{\text{binning factor}}{\text{bandwidth}}\right)\right)$$

Where:

$$\text{e.v.} = \text{environment feature value}$$
$$\text{Bandwidth} = 1.0\sigma \text{ for all applications presented}$$
$$\text{Binning factor} = 5/2^{*} \text{ bandwidth}$$

This procedure returns five discrete environment feature indices for every two bandwidth intervals. As each new test mutation was encountered by the program, the ordered n-tuple of environment feature indices and the categorical feature values was determined. If a new feature value n-tuple corresponding to a test mutation did not exist in the cache, its probability of an effect on function was estimated from the training data. Then the n-tuple, its corresponding probability value, and the number of training mutations used in the probability determination were added to the cache. Probability values for test mutations with feature value n-tuples pre-existing in the cache were determined from cached values.

Programming and software

Scripts for all entire procedure were written in Python v.1.4 and Awk, and implemented on an SGI, Inc. O_2 computer running under IRIX6.4. All statistical analysis including the ANOVA and χ^2 tests were performed with Microsoft Excel. The molecular graphics (Figure 5) were prepared using RIBBONS.[53]

Acknowledgments

We thank Ernest Fraenkel, Scot Wolfe, Steve Orzack, Jim Freeman, Vincent Stanton, and Colin Dykes for providing comments on the manuscript and helpful discussions. We also thank Patrick Kelly for guidance in the statistical analysis of the mutation data and Greg Verdine for encouragement and critical suggestions throughout the course of this project.

References

1. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K. & Patil, N. *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231-238.
2. Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N. & Weder, A., *et al.* (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239-247.
3. Liggett, S. B., Wagoner, L. E., Craft, L. L., Hornung, R. W., Hoit, B. D. & McIntosh, T. C. *et al.* (1998). The Ile164 beta2-adrenergic receptor polymorphism adversely affects the outcome of congestive heart failure. *J. Clin. Invest.* **102**, 1534-1539.
4. Liggett, S. B. (1999). Molecular and genetic basis of beta2-adrenergic receptor function. *J. Allergy Clin. Immunol.* **104**, S42-S46.
5. Lima, J. J., Mohamed, M. H., Self, T. H., Eberle, L. V. & Johnson, J. A. (2000). Importance of beta(2)Adrenergic receptor genotype, gender and race on albuterol-evoked bronchodilation in asthmatics. *Pulm. Pharmacol. Ther.* **13**, 127-134.
6. Tai, H. L., Fessing, M. Y., Bonten, E. J., Yanishevsky, Y., d'Azzo, A. & Krynetski, E. Y., *et al.* (1999). Enhanced proteasomal degradation of mutant human thiopurine S-methyltransferase (TPMT) in mammalian cells: mechanism for TPMT protein deficiency inherited by TPMT*2, TPMT*3A, TPMT*3B or TPMT*3C. *Pharmacogenetics,* **9**, 641-650.
7. Drysdale, C. M., McGraw, D. W., Stack, C. B., Stephens, J. C., Judson, R. S. & Nandabalan, K., *et al.* (2000). Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA,* **97**, 10483-10488.
8. Sachse, C., Brockmoller, J., Bauer, S. & Roots, I. (1997). Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am. J. Hum. Genet.* **60**, 284-295.
9. Perutz, M. F. (1965). Structure and function of haemoglobin. I. A tentative atomic model of horse oxyhaemoglobin. *J. Mol. Biol.* **13**, 646-668.
10. Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution, and genic heterogeneity. In *Horizons in Biochemistry* (Kasha, M. & Pullman, B., eds), pp. 189-225, Academic Press, New York.
11. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science,* **247**, 1306-1310.
12. Dao-pin, S., Anderson, D. E., Baase, W. A., Dahlquist, F. W. & Matthews, B. W. (1991). Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry,* **30**, 11521-11529.
13. Alber, T., Sun, D. P., Nye, J. A., Muchmore, D. C. & Matthews, B. W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry,* **26**, 3754-3758.
14. Markiewicz, P., Kleina, L. G., 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.
15. Poteete, A. R., Rennell, D. & Bouvier, S. E. (1992). Functional significance of conserved amino acid residues. *Proteins: Struct. Funct. Genet.* **13**, 38-40.
16. Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B. & Muller-Hill, B. (1996). Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol.* **261**, 509-523.

17. Sunyaev, S., Ramensky, V. & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **16**, 198-200.
18. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595-603.
19. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
20. 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.
21. Marchler-Bauer, A. & Bryant, S. H. (1999). A measure of progress in fold recognition? *Proteins: Struct. Funct. Genet.* **37**, 218-225.
22. Guex, N., Diemand, A. & Peitsch, M. C. (1999). Protein modelling for all. *Trends Biochem. Sci.* **24**, 364-367.
23. Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. & Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins: Struct. Funct. Genet.* **23**, 318-326.
24. Kim, S. H. (2000). Structural genomics of microbes: an objective. *Curr. Opin. Struct. Biol.* **10**, 380-383.
25. Frishman, D., Goldstein, R. A. & Pollock, D. D. (2000). Protein evolution and structural genomics. *Pac. Symp. Biocomput.* **12**, 3-5.
26. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291-325.
27. Fisz, M. (1963). *Probability Theory and Mathematical Statistics*, Robert E. Krieger, Malabar, FL.
28. Montgomery, D. C. & Peck, E. A. (1992). *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York.
29. Rost, B. & O'Donoghue, S. (1997). Sisyphus and prediction of protein structure. *Comput. Appl. Biosci.* **13**, 345-356.
30. Sunyaev, S. R., Lathe, W. C., III, Ramensky, V. E. & Bork, P. (2000). SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16**, 335-337.
31. Ness, J. E., Del Cardayre, S. B., Minshull, J. & Stemmer, W. P. (2000). Molecular breeding: the natural approach to protein design. *Advan. Protein Chem.* **55**, 261-292.
32. Crameri, A., Whitehorn, E. A., Tate, E. & Stemmer, W. P. (1996). Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnol.* **14**, 315-319.
33. Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P. & Punnonen, J. *et al.* (1999). Evolution of a cytokine using DNA family shuffling. *Nature Biotechnol.* **17**, 793-797.
34. Spiller, B., Gershenson, A., Arnold, F. H. & Stevens, R. C. (1999). A structural view of evolutionary divergence. *Proc. Natl Acad. Sci. USA,* **96**, 12305-12310.
35. Housman, D. & Ledley, F. D. (1998). Why pharmacogenomics? Why now? *Nature Biotechnol.* **16**, 492-493.
36. Venter, J. C., Adams, M. D., Myers, E. W. *et al.* (2001). The sequence of the human genome. *Science,* **291**, 1304-1351.
37. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.
38. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67-88.
39. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316-319.
40. Bordo, D. (1993). ENVIRON: a software package to compare protein three-dimensional structures with homologous sequences using local structural motifs. *Comput. Appl. Biosci.* **9**, 639-645.
41. Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins: Struct. Funct. Genet.* **33**, 88-96.
42. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.
43. Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Advan. Protein Chem.* **46**, 249-278.
44. Walker, D. R., Bond, J. P., Tarone, R. E., Harris, C. C., Makalowski, W. & Boguski, M. S. *et al.* (1999). Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene,* **18**, 211-218.
45. Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216-226.
46. Adams, R. M., Das, S. & Smith, T. F. (1996). Multiple domain protein diagnostic patterns. *Protein Sci.* **5**, 1240-1249.
47. Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555-565.
48. Henikoff, J. G. & Henikoff, S. (1996). Blocks database and its applications. *Methods Enzymol.* **266**, 88-105.
49. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z. & Miller, W. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
50. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215-219.
51. Loader, C. (1999). *Local Regression and Likelihood*, Springer, New York.
52. Lebart, L., Morineau, A. & Warwick, K. (1984). *Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices*, Wiley, New York.
53. Carson, M. (1997). Ribbons. In *Macromolecular Crystallography, Part B* (Sweet, R. M. & Carter, C. W., Jr, eds), vol. 277 edit., pp. 493-505, Academic Press, New York.

Edited by F. Cohen

(Received 3 October 2000; received in revised form 8 January 2001; accepted 10 January 2001)