doi:10.1016/j.jmb.2005.08.020 J. Mol. Biol. (2005) 353, 459–473

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease

Peng Yue[1,2], Zhaolong Li[1] and John Moult[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, MD 20850 USA*

[2]*Molecular and Cellular Biology Program, University of Maryland, College Park, MD 20742, USA*

**Corresponding author*

The most common cause of monogenic disease is a single base DNA variant resulting in an amino acid substitution. In a previous study, we observed that a high fraction of these substitutions appear to result in reduction of stability of the corresponding protein structure. We have now investigated this phenomenon more fully. A set of structural effects, such as reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions, is used to represent the impact of single residue mutations on protein stability. A support vector machine (SVM) was trained on a set of mutations causative of disease, and a control set of non-disease causing mutations. In jack-knifed testing, the method identifies 74% of disease mutations, with a false positive rate of 15%. Evaluation of a set of *in vitro* mutagenesis data with the SVM established that the majority of disease mutations affect protein stability by 1 to 3 kcal/mol. The method's effective distinction between disease and non-disease variants, strongly supports the hypothesis that loss of protein stability is a major factor contributing to monogenic disease. Mutant analysis is available (www.snps3d.org).

© 2005 Elsevier Ltd. All rights reserved.

Keywords: protein structure; protein stability; monogenic disease; mis-sense mutations; single nucleotide polymorphisms

Introduction

Over 1000 human genes have been identified where one or more sequence modifications are directly causative of disease.[1] These mutations may affect protein function through a number of mechanisms, such as changes in transcription, RNA processing, protein expression, folding of the polypeptide chain, stability of the folded state, post-translational modification, interactions with binding partners, and alterations to catalysis. An analysis of the Human Gene Mutation Database (HGMD)[1] has shown that the vast majority of known cases act through changes to the coding sequence, with mis-sense mutations (a single base change resulting in change of a single amino acid) by far the most common effect, accounting for greater than 60% of all monogenic disease mutations. In a previous study, we investigated the relationship between these mis-sense mutations and disease, in terms of the effect of the resulting amino acid change on protein structure and function.[2] In that work, we concluded that the most common mechanism (up to 80% of cases) by which a non-synonymous base change results in disease is destabilization of the protein structure, relative to the unfolded state. A further approximately 10% of mis-sense mutations were seen to affect some known aspect of molecular function, and the remaining 10% to operate by other, unidentified, mechanisms. That study relied primarily on visual inspection of the effect of an amino acid substitution on protein structure and function.

Here, we have developed a more objective model, focusing just on the role of destabilization of the folded structure. The primary goals are to more rigorously investigate the extent to which stability is a common factor in causing monogenic disease, and to provide a general and fully automatic stability perturbation model that can be used for analysis of the impact of non-synonymous single nucleotide polymorphisms (SNPs) found in the human population.

Two principal strategies have been developed for identifying which mis-sense base changes are most likely to be causative of disease. The most common approach makes use of the fact that the more critical

Abbreviations used: SVM, support vector machine; HGMD, Human Gene Mutation Database; PDB, Protein Data Bank.

E-mail address of the corresponding author: moult@umbi.umd.edu

0022-2836/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.

a position in a protein sequence is to viability, the more restricted are the residue types accepted there. A number of different methods for assessing the significance of amino acid conservation have been developed.[3–7] Methods that utilize sequence conservation have the advantage of including all kinds of impact on protein viability. Also, the methods can be used with any human protein for which a suitable set of sequence relatives is known, and so have wide applicability. The approach has the disadvantage that it provides no direct insight into the underlying mechanism. The second strategy is to make use of knowledge of protein structure and function. For instance, recognizing that a change occurs in a key catalytic residue, or one involved in ligand binding, or a target for post-translational modification.

Wang & Moult[2] used a structure-based model to identify amino acid substitutions likely to significantly affect protein stability as well as other contributions to function. Stability impact was assessed using a set of simple rules based on changes in hydrophobic burial, backbone strain, overpacking, and electrostatic interactions. That work forms the foundation of the present study. Other groups have combined sequence and structure strategies to varying degrees. Sunyaev[4,5] predicted the effect of mis-sense mutations using empirically derived rules which make use of a variety of data, such as functional information, hydrophobic propensity, side-chain volume change and transmembrane location,[8] together with sequence information. The rules were tested against disease causing mutations annotated by Swiss-Prot, using the variation between human and other species as a control. In Chasman's[6] method, ANOVA and principal component analysis were applied to a series of features that capture aspects of structural and sequence context. Data on the relationship between site-directed mutants and changes in phenotype for a phage protein, T4 lysozyme, and a bacterial protein, Lac repressor, were used as training and testing sets. Features showing strong discrimination between mutations affecting or not affecting the phenotype, such as the relative residue temperature factor, relative surface accessibility, relative phylogenetic entropy (sequence conservation in the protein family) and burial of charge, were selected, A probability model was then constructed based on the selected features, and used to estimate the likelihood that a given mutation will affect function. A similar probability approach has also been used to include function effects.[9] Krishnan & Westhead[7] used two machine learning methods, a decision tree and a support vector machine, to predict the impact of single amino acid changes based on a set of structural (secondary structure and surface accessibility) and sequence attributes, such as sequence conservation score calculated using ScoreCons.[10] Secondary structure and surface accessibility data were taken from the HSSP database[11] or predicted using PHD.[12]

The central hypothesis of the present work is that moderate loss of stability of the folded state of a protein molecule is frequently associated with monogenic disease. To investigate this, we must identify significant changes in the free energy difference between the folded and unfolded states of a protein molecule resulting from an amino acid substitution. A theoretically rigorous approach would be to use an appropriate integration of the energy change as one amino acid is morphed into another in the context of the protein structure. These free energy perturbation techniques[13] have been incorporated in a number of the more widely used molecular dynamics software packages. Issues of conformational sampling, appropriate representation of the unfolded state and force field accuracy have generally resulted in poor accuracy.[14] Recent results show encouraging improvement, but require care and method optimization in each case,[15] restricting large-scale application. Force field deficiencies may be reduced by parameterizing using free energy differences obtained from site-directed mutagenesis experiments.[16] The resulting model is effective at predicting this type of stability change.

We have developed a knowledge-based method that estimates whether or not an amino acid substitution reduces protein structure stability sufficiently to be potentially causative of monogenic disease. As in the earlier work,[2] we make use of the extensive literature on the effect of amino acid substitutions on protein stability, as well as knowledge of the underlying factors affecting the free energy of the folded state. We identify a set of 15 such factors that may contribute to a free energy difference, through changes in interaction energy between amino acids, effects on the entropy of the system, and the local rigidity of the structure. A machine learning technique (a support vector machine, SVM[17]) is used to partition the 15-dimensional space representing these factors into two volumes, in such a way that, as far as possible, disease causing mutations fall in one volume and non-disease causing ones in the other. Any new mutation may then be assigned a position in this space. Mutations falling in one volume are predicted to significantly decrease protein stability, and thus to be potentially disease causing. Those falling in the other volume are considered non-disease causing. Distance from the volume partitioning surface provides an approximate measure of confidence in the assignments.

The model is trained on a set of mis-sense mutations that cause monogenic disease, extracted from the HGMD.[1] A control set of residue substitutions not contributing to disease susceptibility was based on inter-species differences.[4] Stability effects are analyzed using available experimental structures of human proteins, or reliable comparative models. Jack-knifed testing shows that this model does differentiate between disease and non-disease mutations, validating the hypothesis that stability effects play a major and quite general role in monogenic disease.

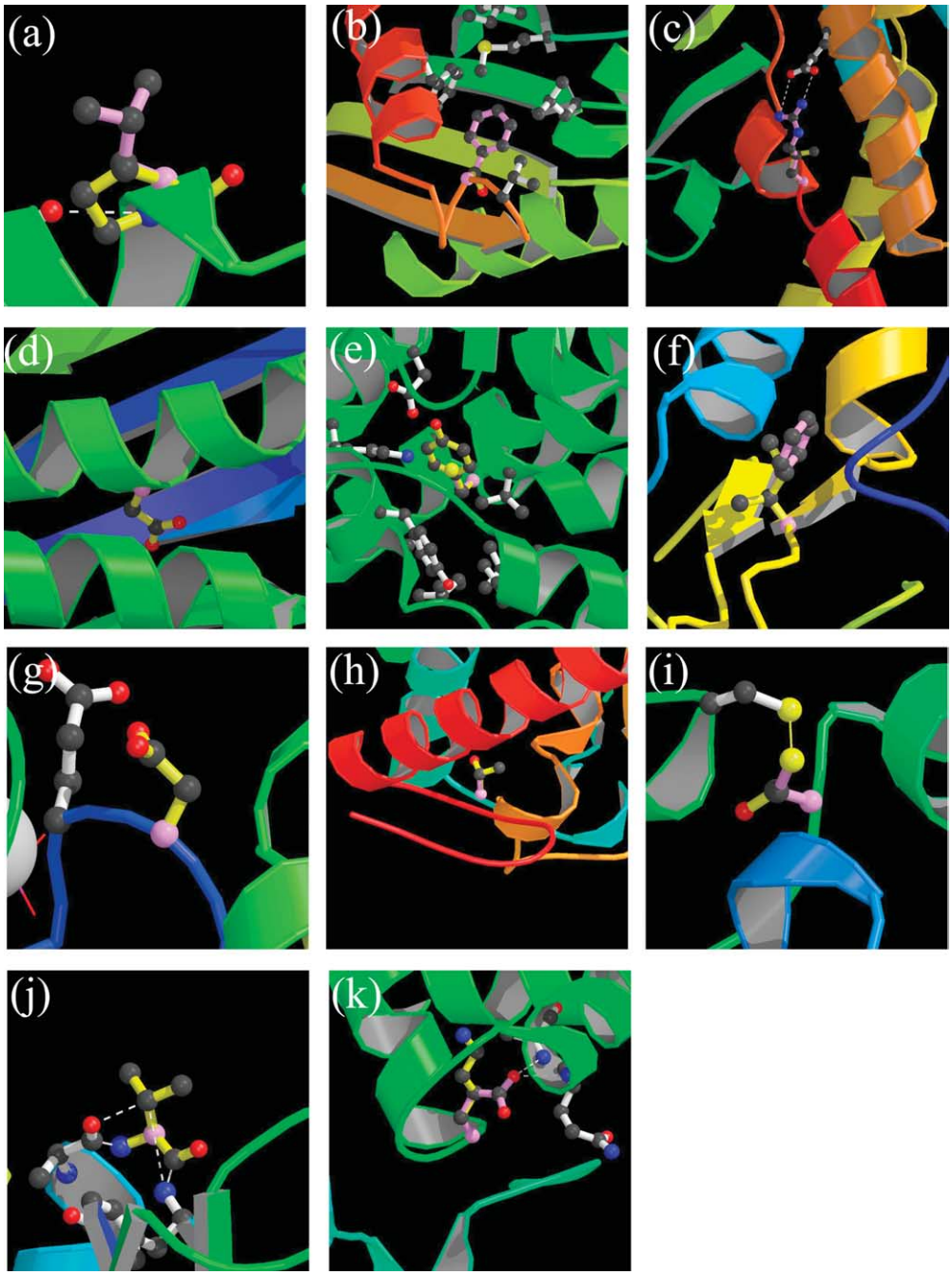

Figure 1. Examples of disease caused by structure destabilizing factors. For each case, bonds of wild-type side-chains are shown purple, and bonds of the mutant side-chains are yellow. Atoms are colored by type. In a number of cases, more than one factor is involved. The selected one is judged to be the most significant. The full model considers all factors together. Disease associations are taken from the NCBI Refseq database. (a) Loss of polar–polar interactions. L226P in galactose-1-phosphate uridylyltransferase (GALT, PDB code 1HXP_B), causing galactosemia. This mutant introduces a proline into an α-helix, resulting in the loss of a main-chain hydrogen bond, as well as loss of hydrophobic interactions of the side-chain. (b) Loss of hydrophobic interactions. F234S in GTP cyclohydrolase (GCH1, 1IR8_I), causing dopamine-responsive dystonia. A large buried non-polar side-chain is replaced by a small polar one, reducing the burial of non-polar area on folding. A cavity is also created, and there is a small gain in polar–polar energy. (c) Loss of a salt-bridge. R382L in isovaleryl coenzyme A dehydrogenase (IVD, 1IVH_C), causing isovaleric acidemia. R382 forms a salt-bridge (charge–charge interaction) in the wild-type protein, lost in this mutant. (d) Buried charge. G60D in aspartylglucosaminidase (AGA, 1APY_A), causing aspartylglycosaminuria. G60D introduces a charge group into the interior of the protein. It also causes over-packing. (e) Over-packing. C91Y acyl-coenzyme A dehydrogenase (ACADM, 1EGE_C), causing ACADM hereditary deficiency. C91Y introduces a bulky side-chain into the interior of the protein, resulting in substantial over-packing. (f) Cavity formation. F411I in glucocerebrosidase (GBA, 1OGS_A), causing Gaucher's disease. F411I replaces a large buried non-polar side-chain with a smaller one, creating an internal cavity. There is also a loss of hydrophobic interaction. (g) Electrostatic repulsion. G38D in guanine nucleotide binding protein (GNAT1, 1TAG), causing night blindness. Introduction of the aspartic acid side-chain results in an unavoidable

Results

Selection of data for analysis

As described in Materials and Methods, 10,263 disease causing mutations in 731 proteins were extracted from the HGMD.[1] Appropriate structure information was available for 37% (3768 in 243 proteins) of these mutants, forming the disease set. Three hundred and forty-six of the HGMD proteins had close orthologs in other species. The corresponding 16,682 inter-ortholog residue differences provided a set of non-disease variants. 14% (2309 in 153 proteins) of the inter-species variants had appropriate structure information, and formed the control set.

Analysis of factors likely to affect protein stability

Eleven contributions to the energy and entropy of protein stability are considered. There are four classes of electrostatic interaction: reduction of charge–charge, charge–polar or polar–polar energy, or introduction of electrostatic repulsion; three solvation effects: burying of charge or polar groups, and reduction in non-polar area buried on folding; and two terms representing steric strain: backbone strain and overpacking. The other two contributions considered are cavity formation (affecting van der Waals energy), and loss of a disulfide bridge. Figure 1 shows examples of each of these, with the corresponding disease outcome. The crystallographic temperature factor and surface accessibility of mutated residues are also considered.

Figure 2(a) shows the distribution of each of these effects in the disease and non-disease data sets (criteria used are described in Materials and Methods). The red bar shows the fraction of all disease data points classified as disease, and the green bar is the fraction all non-disease points classified as disease. An ideal factor includes a large fraction of the disease points (red bar), and no non-disease points (green bar). The 11 energy and entropy factors are ordered by the ratio of the two bar heights, with the best discriminators on the left.

Discrimination power ranges from perfect for disulfide bond breakage (the only instances are in the disease set), to none (loss of polar–polar interactions is as common in the disease set as in the control set). Coverage also varies widely, from only 3% of disease cases involving disulfide bond loss to 24% of cases involving over-packing. The last two terms capture the ability of the structure to relax to partly compensate for unfavorable energy or entropic effects. As expected, regions of lower crystallographic temperature factor contain more disease mutations than non-disease ones. Similarly, buried residues, which generally have least space to adjust to change and more other energetic restrictions, have a twofold excess of disease mutations over non-disease ones.

Greater discrimination can be achieved by taking advantage of the fact that most mutants affect more than one factor. Figure 2(b) shows some examples of discrimination using pairs of factors. For example, combining loss of a polar–polar interaction with a non-surface environment increases the ratio of disease to non-disease cases from about one to approximately three to one. Highest discrimination will be obtained with a method that considers all the factors affected by a mutation. For this purpose, each mutant is represented as a point in a 15-dimensional factor space. Eleven of the dimensions are the energy and entropy factors shown in Figure 2. One dimension is the surface accessibility of the mutated residue, relative to the unfolded state. The other three are the C^{α} temperature factor of the mutated residues, the Z value of the temperature factor, and the standard deviation of all C^{α} temperature factors. (Three dimensions rather than one are used to allow for variable scaling of the experimental values.) As described in Materials and Methods, a SVM was used to determine a surface that optimally partitions the disease and non-disease points in this space.

Accuracy of the SVM model

Figure 3 summarizes the results of the model. 74% of the 3768 mis-sense mutations in the disease dataset were assigned as disease causing, and 85% of the 2309 mis-sense mutations in the non-disease dataset were classified as non-disease. For the 82% of data points more than a distance of 0.5 from the SVM partitioning surface, the prediction accuracy increases to 79% correctly identified disease data points, and 89% correctly assigned non-disease points. The 15% false positive rate arises from defects in the model. Since only stability factors are included in the model, all mutants that act through

electrostatic repulsion with another aspartic acid. There is also limited over-packing. (h) Buried polar group. A543T in hexosaminidase B (HEXB, 1O7A_D), causing Sandhoff disease. Here a hydroxyl group is introduced in a buried non-polar environment. There is also minor over-packing. (i) Breaking of a disulfide bond. C163S in aspartylglucosaminidase (AGA, 1APY_A), causing aspartylglycosaminuria. C163S replaces one component of a disulfide bond. (j) Backbone strain. G137V in arylsulfatase B (ARSB, 1FSU), causing Maroteaux–Lamy syndrome. G137V introduces a side-chain onto a glycine residue with backbone dihedral angles unsuitable for other residue types. (k) Loss of charge–polar interaction. E167K in uroporphyrinogen decarboxylase (UROD, 1R3Q_A), causing familial porphyria cutanea tarda and hepatoerythropoetic porphyria. E167 forms charge–polar interactions with two main-chain N–H groups, providing a helix cap. The mutation removes these interactions.

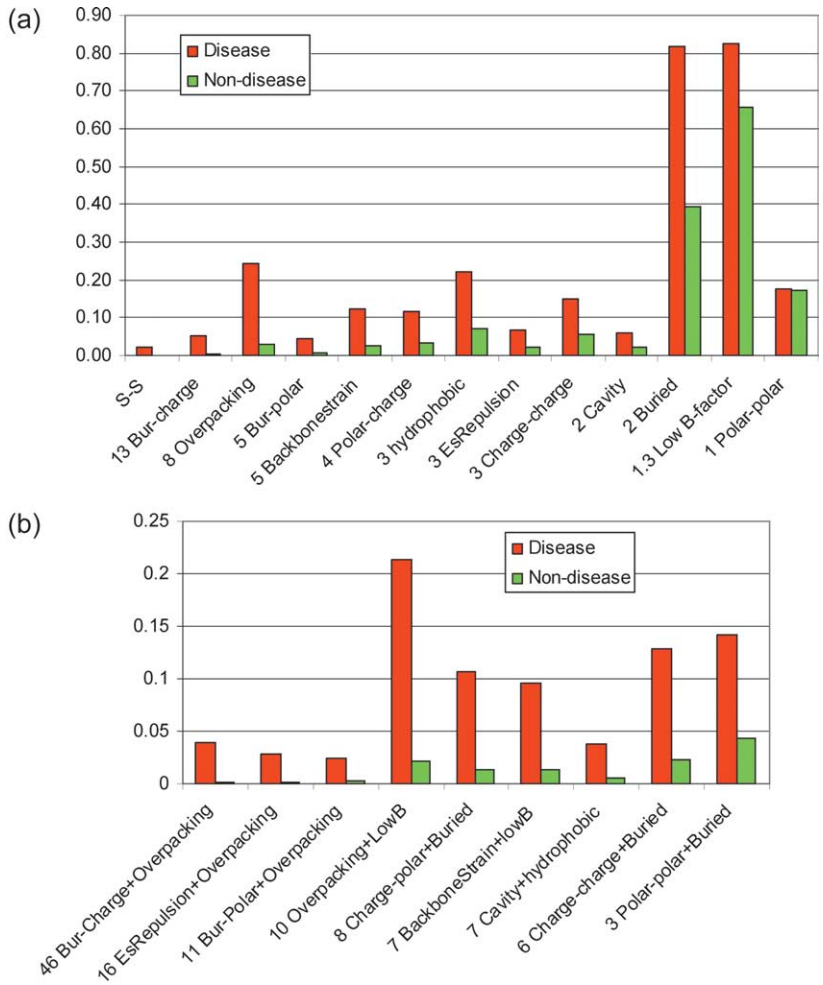

Figure 2. (a) Partitioning of each stability factor between the disease and non-disease data sets. The red bars show the fraction of disease variants covered by the corresponding factor, and the green bars show the fraction of non-disease variants covered. An ideal factor has high coverage of the disease set, and no examples in the non-disease set. Factors are ordered by the discriminatory power (ratio of disease to non-disease coverage), best discriminators to the left. The discriminatory power of each factor is included in the bar labels. The ratio ranges from infinite for breaking a disulfide bridge (no examples in the non-disease set) to 1 for polar–polar interactions (an approximately equal number of examples in the disease and non-disease sets). (b) Improvement in discrimination when two stability factors are considered together. As in (a), bars show the partitioning between disease and non-disease variants, now considering two factors at a time. Discriminatory power is considerably improved. For example, adding a non-surface requirement to loss of polar–polar interactions increase the discriminatory ratio from 1 to 3. Best discrimination is achieved when all relevant factors are considered, as in the full model.

other mechanisms, such as effects on catalysis, binding and so on, are included in the 26% false negative rate. Some fraction of false negatives is mutants included in the HGMD database that do not appear to cause disease. For example, the mutant G15D in the alpha chain of hemoglobin (HBA1) is in HGMD, but is predicted to be non-disease causing, with a confident SVM score of 0.8. The literature on this mutation[18] gives no indication of disease. Allowing for approximations in the model, a conservative conclusion is that substantially more than half of disease mutants operate at least partly through destabilization of the folded structure.

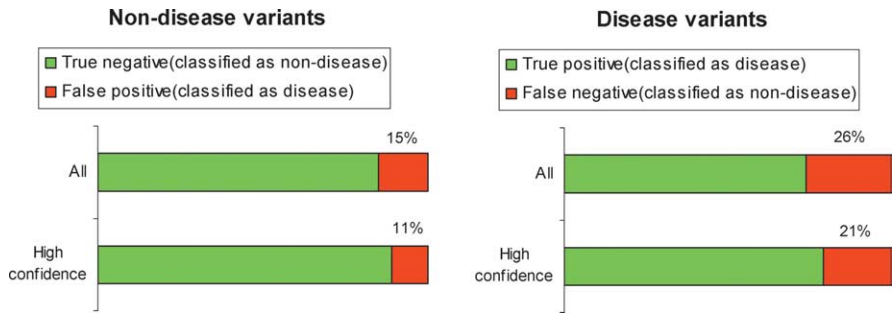

Figure 3. Evaluation of the SVM model. The right-hand panel shows the fraction of disease variants correctly identified by the model in jack-knifed testing. The model is trained only to detect variants that cause disease by destabilization of the structure, so that the false negative rate of 26% includes all other causes, as well as deficiencies in the model. The bottom bar shows the result for the more confident subset of predictions (the 80% of the data with an SVM distance greater than 0.5), with a false negative rate of 21%. The left-hand panel shows the same data for the non-disease data set. Here, the false positive rate (variants incorrectly assigned to disease) is 15% for the full set and 11% higher confidence classifications.

Model evaluation using *in vitro* mutagenesis stability data

The SVM disease model is trained entirely on disease related mutant data, containing no explicit information about stability. Evaluation of the model's performance against *in vitro* mutagenesis free energy data provides an independent test of the hypothesis that disease is strongly coupled with structure destabilization. We would expect that there should be a strong correlation between a potential disease outcome and the change in the free energy difference between the folded and unfolded states.

As described in Materials and Methods, we have run the disease trained prediction model against a set of 581 of these *in vitro* stability data, from four proteins (Table 1). Figure 4 shows the relationship between the change in free energy and the fraction of mutations that would be predicted to have a disease outcome. For mutants that stabilize or mildly destabilize the folded state (up to 1 kcal/mol) the faction of potential disease causing residues is close to the false positive rate of the model (16%). As the change in free energy increases, so does the fraction of potential disease causing mutations, reaching 90% in the 3–4 kcal/mol range, and 100% above 4 kcal/mol. These results confirm that the model is detecting destabilizing effects on structure. The observation that most potential disease classified mutations destabilize the folded state by about 2 to 3 kcal/mol suggests that real disease causing mutations will be in this range. That conclusion is supported by the fact that the distribution of SVM scores for mutants that destabilize by more than 2 kcal/mol is similar to that of the disease causing mutants (means of −0.88 and −1.00, medians of −0.68 and −0.60, respectively).

It is informative to examine the outliers in this distribution. Five (L108I, L36V, L37V and A132G in staphylococcal nuclease and S92A in barnase) of the 53 mutants that decrease stability by 3–4 kcal/mol are predicted not to be consistent with disease. The two L→V mutants differ by one methyl group, and both result in a slight loss of hydrophobic burial. There are 24 L→V mutants in the disease dataset and 37 cases in the non-disease dataset, suggesting that this class of mutant is finely balanced between disease and non-disease causing, and subtle effects tip the balance. Consistent with this, the SVM gives a low confidence score (0.14 and 0.13) for these two outliers. L108I creates no change of volume or overall hydrophobicity, so it is surprising that it is so destabilizing. There are 25 such mutations in the non-disease dataset, and only four in the disease set, suggesting that this high level of destabilization is unusual. The SVM score is also in the less confident range (0.3). The authors of the experimental study[19] suggest that loss of highly optimal van der Waals packing is primarily responsible for the large effect. The remaining two mutations, A132G and S92A, are both predicted to be non-disease causing with relatively high confidence (SVM scores 0.70 and 0.89). For A132G, there is a minor loss of hydrophobic burial. There are 36 cases of A→G mutations in the non-disease set and only 11 cases in the disease dataset. For S92A, the model identified the loss of a hydrogen bond and a slight gain of hydrophobic burial. Serrano and colleagues[20,21] note that this residue is the first residue in a beta turn between two strands. The hydroxyl group is buried, and makes two hydrogen bonds, suggesting that it is may be involved in unusually strong interactions. There are 77 cases of S→A mutants in the non-disease set and only two in the disease set, indicating that such strong polar electrostatic interactions are unusual.

Eight of the 52 mutants that increase protein stability are predicted to be consistent with disease. All but one are in staphylococcal nuclease. All increase stability by less than 1 kcal/mol. For three

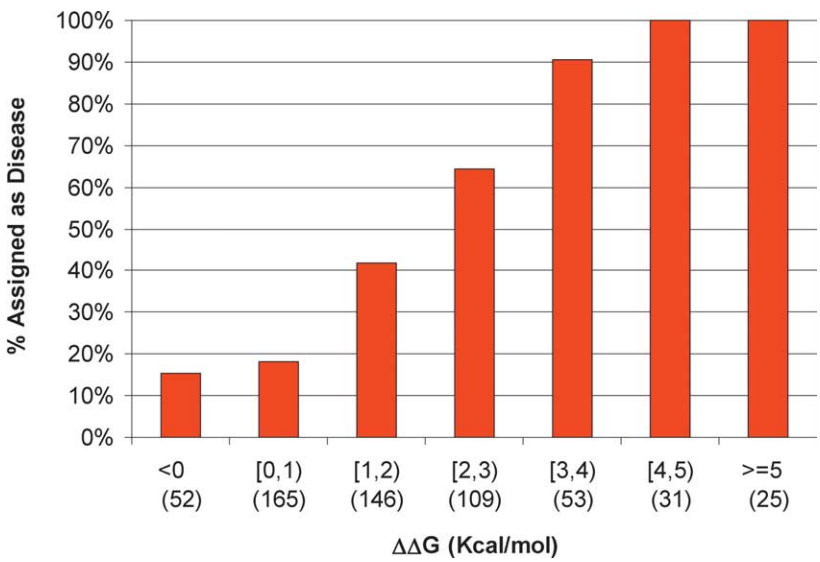

Figure 4. Application of the disease/stability model to *in vitro* site-directed mutagenesis data. The plot shows the fraction of mutants classified as consistent with disease, as a function of the free energy difference between the folded and unfolded states. For stabilizing and weakly destabilizing mutants, the disease compatible fraction is similar to the false positive rate of the model. Above 3 kcal/mol of destabilization, 90% of mutants are classified as disease compatible. The results suggest that a typical disease causing mutant destabilizes the folded state by 2–3 kcal/mol.

cases: N138G, S128A and H124F, the SVM returns a low confidence score. In none of the other cases is it clear why there is disagreement with experiment. For D21A and D21G, there is a predicted loss of charge–charge and charge–polar interactions. The distributions of these two mutations between the disease and non-disease datasets are 8/8 and 57/11, respectively. T41I is predicted to result in a large gain of hydrophobic burial, offset by the loss of a charge–polar and polar–polar interactions in a buried environment. There are 41 cases of T→I mutations in the disease dataset and 18 cases in the non-disease dataset, most with a predicted large gain of hydrophobic burial and decreased electrostatic interactions. G50A is predicted to result in backbone strain. It is probable that the structure is able to relax to accommodate the change in backbone angles. The temperature factor is moderately high, supporting this possibility. The eighth mutant, N58D, is in barnase. There is a predicted loss of polar–polar interaction and a slight gain of charge–polar interaction.

Alternative test sets

This work uses disease and non-disease related data for training and testing. Others[3,6,7] have used data on the phenotypic impact of single residue mutants in a bacterial and a phage protein. We have investigated the relationship between our assignment of disease potential and phenotypic impact in these mutagenesis sets. The data are a set of about 4000 mutants of the *Escherichia coli lac* repressor[22] and a set of about 2000 mutants of phage T4 lysozyme.[23] A total of 1987 mutations in T4 lysozyme and 3291 mutations in *lac* repressor can be modeled on to the corresponding protein structures (PDB entries 1lbh and 7lzm, respectively). Each data set was partitioned into groups based on the phenotype annotations in the literature. For *lac* repressor, these annotations are: + (wild-type phenotype, 200-fold repression of beta-galactosidase activity, but in practice some times only 8–10% of this); +s (wild-type phenotype under certain conditions, including temperature-sensitive mutations); +− (20–200-fold galactosidase repression); −+ (4–20-fold); and − (less than fourfold repression). For T4 lysozyme, the groups are: ++ (wild-type phenotype: plaque size similar to control); + (significantly smaller plaques); +/− (similar in size to +, but hazy morphology); and − (no plaques produced).

The HGMD trained SVM model was used to assign potential disease mutants in each of the phenotype categories. Figure 5 shows the results. For both proteins, a high fraction of the mutants in the most severe class of phenotype impact are assigned as disease-like (∼90% for Lac repressor and ∼100% for T4 lysozyme). However, for both proteins, about 40% of "wild-type" mutants are also assigned as consistent with disease. The probable explanation is that a rather low level of enzyme activity is needed for a wild-type classification: for T4 lyzozyme, as little of 4% residual enzyme activity may be classified as wild-type,[23] and for Lac repressor, 10% activity is some times sufficient.[22] Such low levels of monogenic disease protein activity would likely usually result in disease.

Functional analysis of single residue mutations

An advantage of the structure/stability model is that it provides mechanistic insight into why a mutant has a deleterious effect on protein function. In principle, functional roles, such as ligand binding and catalysis, may also be assigned to particular residues, and so allow more general mechanism-based analysis. As described in Materials and Methods, we have investigated this possibility

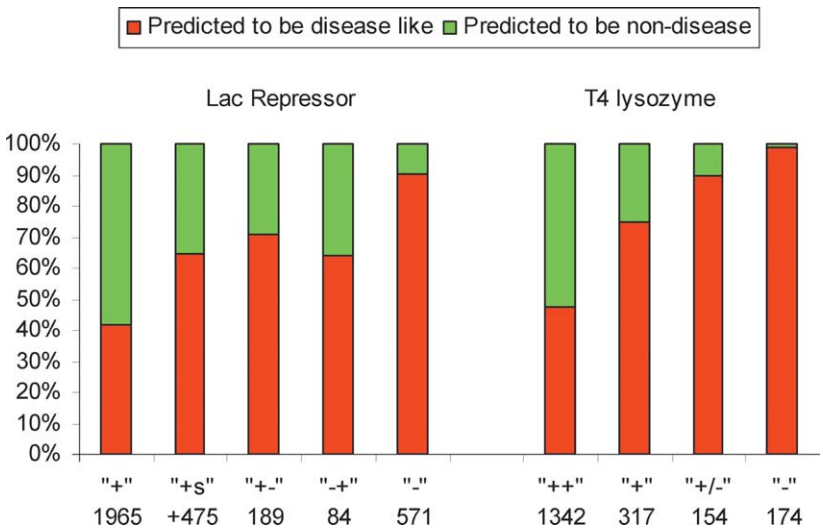

Figure 5. Application of the disease/stability model to mutants of Lac repressor and T4 lyzozyme. Symbols below the bars indicate the extent of phenotypic impact for that set of mutants, from + for the most activity to − for none. Red regions of the bars show the fraction of mutants in each category found to be compatible with disease. As expected, a high fraction of the low activity mutants are assigned as compatible with disease, but a significant fraction of the maximum activity ones are also so classified. This result is consistent with the fact that a low % of activity is sufficient for a + classification for both proteins. Numbers below each column show the number of mutants in that category.

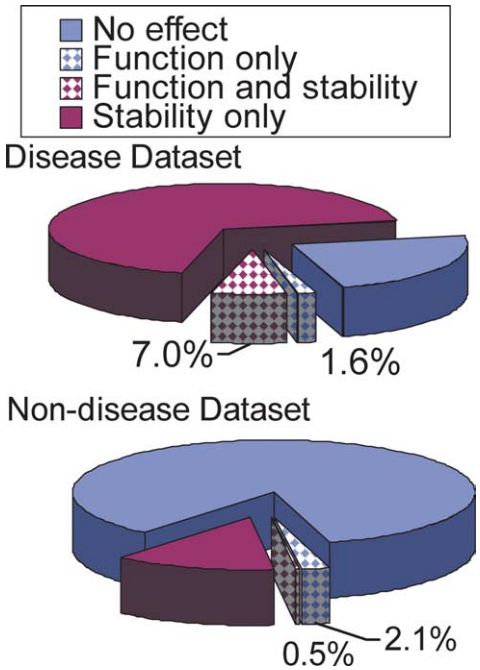

Figure 6. Distribution of direct functional effects of variants in the disease and non-disease data sets. Residue function was assigned from Swiss Prot annotation and on the basis of contacts with bound ligands. 7% of stability variants also have a known functional role, and only an additional 1.6% of false negatives are associated with function. 2.1% of correctly classified non-disease variants are assigned a functional role. Overall, few variants are assigned function, and inclusion of those in a disease classification model would slightly increase the false positive rate.

using SwissProt functional annotation and experimentally observed ligand binding. Figure 6 shows the results. For the disease set, an additional 1.6% of the mutants that were false negatives in the stability model are annotated as functionally important. Seven percent of the stability related mutants are also assigned a functional role. These low values probably reflect the incomplete assignment of function. Inclusion of these in the model would reduce the false negative rate by 1.6%. However, in the non-disease set, an additional 2.1% of mutants are assigned a functional role, leading to an increase in the fraction of false positives. Thus, we conclude that, at present, residue function annotation is too unreliable and incomplete to be useful.

Investigation of the role of protein structure accuracy

Two-thirds of the mis-sense mutations are analyzed in the context of structure models rather than experimental structures. The accuracy of these comparative models therefore plays a role in the accuracy of disease assignment. In general, accuracy of a structure model decreases with decreasing sequence identity between the structure of interest and the closest available template structure.

To investigate the significance of this factor, disease assignment accuracy was examined as a function of structure/template sequence identity, in ranges between 25% and 100% ("100%" are those cases for which an experimental structure of the human protein is available). A separate SVM model was trained and tested within each sequence ID group.

Results are shown in Table 2. Overall, disease assignment using protein models based on a structure template with more than 40% sequence identity is not significantly less accurate than that based on experimental structures. For sequence identity of 30% or lower, errors in structure models begin to have a significant effect, with increases in both the false negative and false positive rates. Multiple factors contribute to the decline in accuracy, including less reliable side-chain interactions arising from higher main-chain position errors, an increased frequency of sequence alignment errors, and higher number of insertions and deletions.[24]

Discussion

Role of protein destabilization in monogenic disease

This work tested the hypothesis that destabilization of protein structure is a major factor in human monogenic disease. A simple factor-based model of the stability impact of single residue mutants and an objective machine learning technique are used. In properly jack-knifed testing, the model is able to distinguish between mutants likely to lead to disease and those that do not, with reasonably low false negative (26%) and false positive (15%) rates. The false negative rate (those non-synonymous base changes that lead to disease not

Table 1. *In vitro* mutagenesis data from four proteins, used to test the SVM model

Protein and PDB structure	Structure class	Number of mutations (total 581)
Acyl-coenzyme A binding protein (2abd)	All alpha	30[59]
fk 506 binding protein (1fkj)	Alpha and beta	34[60,61]
Barnase (1bni)	Alpha and beta	87[20,21,62]
Staphylococcal nuclease (1stn)	All beta	430[19,42–48]

Structure class is taken from SCOP.[49]

Table 2. Disease assignment accuracy as a function of structure model quality

	Disease variants				Non-disease variants			
% Identity	# Mutants	% of Total	# Proteins	FN (%)	# Mutants	% of Total	# Proteins	FP (%)
100%	1710	35	85	25.5	672	23	50	16.7
90–99%	981	20	67	23.2	932	33	61	13.2
40–99%	1077	22	93	24.3	705	25	62	16.7
25–39%	1181	24	142	27.5	551	19	91	28.2

Each row shows data using structure models based on a given range of sequence identity to an experimental structure. Accuracy is measured by the false positive rate, FP (fraction of non-disease variants classified as disease causing), and the false negative rate, FN (fraction of disease variants classified as non-disease causing) The 100% row gives data for cases based on an experimental structure, rather than a model. Accuracy is unaffected by the use of a structure model for sequence identities down to about 40%.

so categorized) partly reflects deficiencies in the model, but also includes the fraction of mutants that act through mechanisms other than destabilization. We conclude from these results that substantially more than half of monogenic disease mutants act through a process consistent with destabilization of the folded state.

Use of the model to classify *in vitro* mutagenesis data strongly supports the role of stability in disease, and implies that a disease causing mutant typically destabilizes a protein by 2–3 kcal/mol. For most globular proteins, the free energy difference between the folded and unfolded state is between 5 kcal/mol and 15 kcal/mol,[25] corresponding to an equilibrium constant between the unfolded and folded states of between 10^{-4} and 10^{-13}. A mutant that destabilized by 2 kcal/mol would increase the concentration of the unfolded state by about two orders of magnitude, but the fraction of unfolded molecules is still so small that there would be no expected effect on function in an *in vitro* assay. *In vivo*, though, chaperone scavenging of unfolded proteins[26] may cause such a 100-fold increase in the fraction of unfolded molecules to result in a much lower steady state protein concentration.

Although loss of stability is clearly highly related to a disease outcome, it may sometimes be an effect on folding that is the actual mechanism. *In vitro* folding studies of simple proteins, such as barnase,[27] show that about 40% of mutants that affect stability also affect the folding rate. For disease mutants, folding may be slowed so much that most molecules are targeted for recycling by the quality control machinery in the endoplasmic reticulum and elsewhere.[28] Since not all mutants that affect stability also affect folding rate, if folding were the primary factor, a stability model should generate a high level of false negatives. The reasonably high accuracy of the stability model thus suggests that it is the most relevant factor. Nevertheless, without extensive experimental studies, it is not possible to know for what fraction of cases stability or folding is most relevant.

Direct experimental evidence for the role of stability is scarce, since there are very few studies of the properties of disease causing mutants in human proteins. One exception is mutants in phenylalanine hydroxylase. Excess phenylalanine is toxic, and defects in this enzyme lead to phenylketonuria (PKU). Over 100 single residue disease causing mutants are known, and a subset of these have been studied in COS cells.[29] There is a clear correlation between the set assigned as affecting stability by our model and the *in vivo* total activity and concentration, as measured by immuno-precipitation.

Why does protein stability play a prominent role in monogenic disease?

There are many mechanisms by which a single base change may affect the function of a protein *in vivo*: changes in gene regulatory regions may lead to altered transcription rates; changes in the transcribed message may lead to altered processing, particularly splicing; message changes may affect translation through, for example, altering the secondary structure properties.[30,31] Surprisingly, data for monogenic disease in HGMD suggest that all these pre-protein factors account for less than 10% of cases.[1] This number may be an underestimate of the true value, because of bias in detection methodology. Nevertheless, it is clear that protein level effects are by far the more common.

Once a polypeptide chain has been produced, non-synonymous changes (those base changes resulting in an amino acid substitution) may affect *in vivo* activity in two major ways: aspects of the protein's molecular function may be altered, particularly ligand binding, catalysis, post-translational modification, or an allosteric mechanism. The likelihood of this class of effect depends on the fraction of residues critically involved in one or more of these functions.

The second way in which non-synonymous base changes may affect *in vivo* activity is by reduction of the concentration of protein. This may come about through less successful folding, or an increase in the fraction of unfolded protein, caused by a reduction in stability. Tests with the stability model, sampling a large number of randomly chosen mutants, show that approximately half are consistent with a disease outcome. Thus, the high fraction of disease mutants associated with stability loss is likely a consequence of the higher fraction of mutants that can affect stability, compared with the other possible causes.

Distinguishing properties of monogenic disease proteins

For the 1000 or so monogenic disease proteins in HGMD, the average number of known single residue mutants leading to disease is just over 10.[1] Yet no mutants directly causative of monogenic disease are known in the remaining approximately 22,000 human proteins. What is the difference between these two sets of proteins? First, monogenic disease proteins may be abnormally unstable or have abnormally fragile folding behavior. There is very little data with which to address this possibility, but many are relatively simple metabolic enzymes, and compared with most human proteins, the least likely to exhibit this sort of fragility. A second possibility is that mutants in many of the other proteins lead to a non-viable fetus, and so are never classified as disease causing. Gene suppression in *Caenorhabditis elegans*[32] and *Saccharomyces*,[33,34] as well as limited mouse knockout data all suggest that only 10–20% of proteins are essential in this sense, and so that is unlikely explanation. Third, and most probable, monogenic disease proteins may be the subset to which the system is least robust to component failure. Analysis of non-synonymous single nucleotide polymorphisms in the human population shows a significant fraction that appear to be as deleterious to protein structure and function as those found in monogenic disease genes,[3,5,6,35] but with no disease outcome. Limited knowledge of human protein networks makes it difficult to rigorously test this possibility. Nevertheless, inspection of the pathway context of monogenic disease proteins supports this explanation. Many, such as phenylalanine hydroxylase, appear to perform unique roles, with no redundancy alternative pathways. In contrast, inspection of the pathway context of proteins containing SNPs that destabilize protein structure significantly, such as the T cell receptors,[36] usually suggests a mechanism that makes the system robust to failure of a protein component. Many different T cell receptors are involved in an antigenic response, so that reduced effectiveness of some will not have obvious disease consequences, although it may influence resistance to particular infections in subtle but significant ways.

Advantages and disadvantages of a protein structure-based approach

An advantage of the structure-based approach is that it provides a detailed atomic level model of the precise mechanism by which an amino acid change results in a change in protein properties. A disadvantage is that it is limited to stability effects, and to cases where structure is available. Use of comparative modeling allowed us to extend the number of mutants that can be analyzed. Tests showed that disease prediction accuracy is unaffected by the use of a model, down to 40% sequence identity to a known structure. This is in keeping with studies of the accuracy of structure modeling methods,[24] and also partly reflects the fact that the method does not depend on very accurate structures. Even so, only about 10% of human protein domains can currently be analyzed. The rapid advance of structural genomics[37] may quickly reduce this limitation.

Access to results

Analysis results for all missense disease and control mutations is available†.

Materials and Methods

Identification of single residue variants related to monogenic disease

Genes associated with monogenic disease were identified by checking all 16,220 human gene names in the NCBI Locuslink[38] database (as of 04/26/2002) against the Human Gene Mutation Database[1] (HGMD) (as of 02/09/2002). HGMD contains the most comprehensive collection of mutations related to monogenic disease. Most are causative of monogenic disease, although a few may be associated with disease as a result of linkage disequilibrium rather than directly causative, or contribute to complex trait disease. Later versions of HGMD include more of the latter class, and so the earlier version was preferred. A total of 731 genes containing 10,263 single residue variations were identified.

Identification of a set of single residue variants not related to disease

We also required a control set of mutants, not causative of disease. It is not known which base variants in the human population contribute to complex trait disease, and so it is not possible to use these. Following others,[4] we used non-synonymous base differences between human proteins and closely related proteins in other mammals. The justification here is that almost all variants that are fixed between species are essentially neutral and non-deleterious. To maintain compatibility between the disease and control sets, the same 731 monogenic disease proteins were used. The protein sequences of these genes were compared to all other mammalian protein sequences in SWISS-PROT,[39] using BLAST.[40] Proteins with at least 90% sequence identity over at least 80% of the full length were selected. Single residue differences in these alignments were used as a set of pseudo "mutations", providing the non-disease set. A total of 348 proteins containing 16,682 such single-residue differences to the human disease set were obtained.

Selection of sets of mutants with protein structure

Each of the 731 human proteins was checked for entries in the Protein Data Bank (as of 7/26/2004).[41] Templates for models of human proteins were taken from the PDB for cases where there was no human structure available, and there was a PDB entry for an X-ray structure at least

† www.snps3d.org

Table 3. The 15 factors included in the model of protein stability

Type	Factors
Continuous factors	Electrostatic interaction: polar–polar, polar–charge, charge–charge
	Over-packing
	Hydrophobic burial
	Surface accessibility
	Structural rigidity: crystallographic *B*-factor, *Z* score and standard deviation
Binary factors	Cavity
	Electrostatic repulsion
	Backbone strain
	Buried charge
	Buried polar
	Breakage of a disulfide bond

The effect of each single residue mutant on stability is expressed in terms of the value of one or more of these contributions to the energy and entropy. Continuous factors are represented by a continuous variable, binary factors are two state, either significantly or not significantly affecting stability.

3.0 Å resolution and with 40% or higher sequence identity to the human protein over at least 100 residues.

For the non-disease set, variants that might be partially compensated by other species differences in the same protein were eliminated as follows. All clusters of variants where there are interatomic contacts of 5 Å or less between residues were discarded. For example, A2S in the myosin light chain is a variant between human and mouse, and between human and rat. G20T, a variant between human and mouse, makes contact with the G20 position, and so both variants were excluded. The rat protein has no change at G20, so rat A2S was retained in the non-disease set.

Support vector machine

The Support Vector Machine software package SVMlight† was used to determine the partitioning surface between the disease and non-disease data in the 15-dimensional parameter space. Continuous variables were normalized in the form of a Z score (Z=(value-mean)/standard-deviation). A radial basis kernel was used, allowing for complex surface topology. For this kernel, the higher the parameter γ, the more complex the effective surface, allowing better accommodation of the data. Too higher a gamma leads to over-fitting, and less accurate prediction on new data. A γ value of 0.2 was selected, based on a series of trials. Weights were assigned to the disease and control data sets to compensate for their different sizes, such that they contributed equally to determining the partitioning surface. The distance of a data point from the partitioning surface provides an approximate measure of confidence in a prediction.

SVM model training and testing

A subset of 90% of the disease and non-deleterious variations were selected randomly to form a training set. The resulting SVM model was used to predict which of the 10% of data not included in training are disease causing. The training and testing procedure was repeated 30 times, randomly selecting the test data on each run. For each trial, the false negative rate (the fraction of disease variations mis-classified as non-disease) and false positive rate (the fraction of non-disease variants mis-classified as disease causing) in the test dataset were calculated. The average false positive and false negative rates provide the measure of the prediction accuracy.

† http://svmlight.joachims.org/

In vitro mutagenesis data

Free energy difference data from site-directed mutagenesis experiments were used to test the ability of the SVM model to identify known destabilizing mutations. Four proteins with a large number of associated site-directed mutagenesis experiments[19–21,42–48,59–62] were selected. They cover three classes of protein folds (SCOP[49] classification): all-alpha, all-beta and alpha and beta. Table 1 lists the proteins, and the number of mutants in each. More data are available in the PROTHERM database[50] but inconsistencies in format, particularly the sign convention for free energy, prevent the large scale use of these.

Comparative modeling of protein structure

Comparative models were built using the in-house APSE (Automatic Protein Structure Emulator) pipeline. Modeling protocols in APSE are based on experience with building comparative models in the CASP experiments[51] and a variety of projects. The procedure can be run in automatic or semi-automatic mode. A core backbone model is first constructed by copying regions of the chosen template structure. Alignments are obtained using CLUSTALW. Co-ordinates of side-chains conserved between the human protein and the PDB template are copied. Remaining side-chains are added using SCWRL.[52]

Where necessary, quaternary structure was taken from the PQS (protein quaternary structure) database of biological units.[53] Additional subunits are modeled in the same manner as the chain of interest. Side-chains are modeled in the multimer context.

Modeling the structure of single residue mutants

All SCWRL library backbone-dependent conformations[52] for the new side-chains were built. The conformation least damaging to stability was selected, based on the following rules. First, the conformation with the least worst over-packing was selected; i.e. if there is one conformation with an interatomic contact of 2.6 Å and another with 2.7 Å, the latter was accepted. No distinction was made between conformations with

contacts 3.0 Å or longer. If more than one conformation remained, the one with the least loss of hydrophobic area was selected. In cases where there is no loss of hydrophobic area, conformations with loss of a salt-bridge were next eliminated, then those with electrostatic repulsion, hydrogen bond loss, cavity formation, backbone strain, introduction of a buried charge, and finally, introduction of a buried polar group.

Modeling the stability impact of a single residue mutant

Table 3 lists the stability factors that provide the 15 dimensions used in assessing the impact of each mutant on protein stability. These are divided into those factors treated as continuous variables, and those treated as two state variables (significantly destabilizing or not).

Continuous factors

(1) *Electrostatic interactions*. The difference in electrostatic energy between a wild-type protein and its corresponding mutant was calculated using a simple Coulomb's law treatment, with no solvent model. The partial electrostatic interaction energy between a pair of polar or charged groups i and j is calculated in the usual manner as:

$$E_{ij} = K \sum_k \sum_l q_k q_l / r_{lk}$$

where the sums are over all atoms k of group i and atoms l of group j, the qs are the partial atomic charges in electrons, and r_{lk} is the distance between atoms l and k, in Å. K is the scaling constant (332) nominally converting energies to kcal/mol. (Absolute scale is not significant here, because of the Z score normalization.) Interactions between a pair of groups are included if the centers of charge are less than a cutoff distance d_c apart. The center of charge of a group $\mathbf{r}_c$ is defined as:

$$\mathbf{r}_c = \sum_k |q_k| \mathbf{r}_k / \sum_k |q_k|$$

where the sum is over all atoms in the group. Electrostatic group definitions and partial atomic charges are as defined by Pedersen & Moult.[54] The threshold for group–group interactions, d_c, is 5 Å. This protocol for electrostatic calculations has been shown to be effective at identity incorrect structural features in experimental structures.[55]

(2) *Overpacking*. For each mutant, the closest inter-atomic distance between the mutant residue and any neighboring residue was used.

(3) *Relative surface accessibility*. Solvent accessible surface[56] was calculated with in-house software. The relative surface accessibility of a residue is defined as the surface area of the side-chain in the folded state divided by an estimate of the average surface area in the unfolded state.[57]

(4) *Hydrophobic burial change*. The change in buried non-polar area ΔA_{NP} resulting from a single residue mutation is defined as:

$$\Delta A_{NP} = \sum_i \Delta a_i - \sum_j \Delta a_j$$

where the first sum i is the change in non-polar area on folding for all non-polar atoms in the mutant structure and the second sum j is over all non-polar atoms in the wild-type structure. The change in atomic non-polar area in folding is given by:

$$\Delta a = a_u - a_f$$

where a_u is the estimate of the average atomic surface area in the unfolded state[57] for that atom, and a_f is the calculated atomic area in the folded structure. Non-polar atoms are those assigned zero charge.

(5) *Crystallographic temperature factors*. For each experimental structure used directly or as a model, the average temperature factor $\langle B \rangle$, and standard deviation $\sigma(B)$ over all C^α atoms was calculated, and used to obtain a temperature factor Z score for each C^α: $Z_i = (B_i - \langle B \rangle)/\sigma(B)$. B_i, Z_i, and $\sigma(B)$ were used as parameters in the SVM.

Binary factors

(6) A cavity is assigned to any mutation resulting in the loss of volume of an aliphatic carbon group or greater at a zero solvent accessibility position. For example, Ala mutated to Gly, where the wild-type C^β atom has zero solvent accessibility.

(7) Electrostatic repulsion is assigned to any mutation which results in two like charged groups with an unavoidable atomic contact of less than 4.5 Å.

(8) Backbone strain is assigned to any mutation if one of the following conditions is met. (A) Replacement of a glycine residue with ϕ/ψ angles in a non-allowed region for other residue types. Allowed regions were those covering 90% of observed ϕ/ψ values, as provided in PROCHECK.[58] (B) Replacement of a *cis*-proline ($\omega = 0(\pm 60)°$) with another residue. (C) Replacement of another residue by proline, where the ϕ value is inappropriate (permitted ϕ for Pro $= -60(\pm 15)°$).

(9) Buried charge is assigned to any mutation that results in a zero solvent accessibility, electrostatically isolated, charge group.

(10) Buried polar is assigned to any mutation that results in a zero solvent accessibility polar group with no hydrogen bond. A hydrogen bond is defined as a donor to acceptor distance ≤ 2.5 Å, and an angle at the acceptor $\geq 90.0°$.

(11) Breakage of a disulfide bond is assigned to any mutation that replaces a cysteine residue in an S–S bond with a non-cysteine residue.

Evaluation of discrimination power of each stability factor

The frequencies of each stability factor in the disease and non-disease datasets were calculated. The ratio of the two frequencies defines a discrimination power. For this purpose, a threshold was chosen for each of the continuous factors. Any mutation with a value higher than the threshold was considered to destabilize protein structure. Thresholds were chosen by inspection of the distribution of values for the disease and non-disease sets, selecting levels that provide a high fraction of true positives and true negatives, while minimizing false negatives and false positives. The following values were used:

(1) Overpacking: at least one unavoidable atomic contact

of 2.5 Å or less of the mutated residue to a neighboring one.
(2) Hydrophobic burial: loss of hydrophobic burial of more than 50 $Å^2$.
(3) Electrostatic interaction: any reduction in electrostatic interaction energy, for polar–polar, charge–charge and charge–polar interactions.
(4) Buried residue: relative residue accessibility of less than 20% (i.e. the wild-type side-chain accessibility is less than 20% of the estimated average unfolded state accessibility).
(5) Moderate crystallographic temperature factor: the C^{α} temperature factor of the mutated residue has a Z score of less than $+1$ (i.e. the temperature factor is less than one standard deviation above the mean for the protein).

Identification of residues with a role in molecular function

Each mutated residue, and all residues with one or more atomic contacts of 6 Å or less to it, was checked against the SWISS-PROT feature annotation table for possible functional effects. Additionally, a check was made for atomic contacts of the mutated residue of 6 Å or less to any ligand atom in PDB entries for that protein and other X-ray structures with at least 40% sequence identity over at least 100 amino acid residues, and at 3.0 Å or better resolution.

Acknowledgements

This work was supported by grant LM07174 from the National Library of Medicine.

References

1. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. *et al*. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581.
2. Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270.
3. Ng, P. C. & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* **31**, 3812–3814.
4. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597.
5. Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucl. Acids Res.* **30**, 3894–3900.
6. Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706.
7. Krishnan, V. G. & Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.
8. Ng, P. C., Henikoff, J. G. & Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, **16**, 760–766.
9. Lau, A. Y. & Chasman, D. I. (2004). Functional classification of proteins and protein variants. *Proc. Natl Acad. Sci. USA*, **101**, 6576–6581.
10. Valdar, W. S. & Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313**, 399–416.
11. Dodge, C., Schneider, R., Sander, C. & The, H. S. S. P. (1998). database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313–315.
12. Przybylski, D. & Rost, B. (2002). Alignments grow, secondary structure prediction improves. *Proteins: Struct. Funct. Genet.* **46**, 197–205.
13. Beveridge, D. L. & DiCapua, F. M. (1989). Free energy *via* molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 431–492.
14. Mark, A. E. & van Gunsteren, W. F. (1994). Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J. Mol. Biol.* **240**, 167–176.
15. Pan, Y. & Daggett, V. (2001). Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: free energy perturbation calculations using transition and denatured states from molecular dynamics simulations of unfolding. *Biochemistry*, **40**, 2723–2731.
16. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
17. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.
18. Molchanova, T. P., Pobedimskaya, D. D. & Postnikov, Yu. V. (1994). A simplified procedure for sequencing amplified DNA containing the alpha 2- or alpha 1-globin gene. *Hemoglobin*, **18**, 251–255.
19. Holder, J. B., Bennett, A. F., Chen, J., Spencer, D. S., Byrne, M. P. & Stites, W. E. (2001). Energetics of side chain packing in staphylococcal nuclease assessed by exchange of valines, isoleucines, and leucines. *Biochemistry*, **40**, 13998–14003.
20. Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
21. Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805–818.
22. 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.
23. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88.
24. Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins: Struct. Funct. Genet.* **53**, 352–368. (Suppl. 6).

25. Privalov, P. L. (1979). Stability of proteins: small globular proteins. *Advan. Protein Chem.* **33**, 167–241.
26. Hohfeld, J., Cyr, D. M. & Patterson, C. (2001). From the cradle to the grave: molecular chaperones that may choose between folding and degradation. *EMBO Rep.* **2**, 885–890.
27. Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. VI. The folding pathway of barnase: comparison with theoretical models. *J. Mol. Biol.* **224**, 847–859.
28. Plemper, R. K. & Wolf, D. H. (1999). Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem. Sci.* **24**, 266–270.
29. Scriver, C. R., Hurtubise, M., Konecki, D., Phommarinh, M., Prevost, L., Erlandsen, H. *et al.* (2003). PAHdb 2003: what a locus-specific knowledgebase can do. *Hum. Mutat.* **21**, 333–344.
30. Shen, L. X., Basilion, J. P. & Stanton, V. P., Jr (1999). Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl Acad. Sci. USA*, **96**, 7871–7876.
31. Pelletier, J. & Sonenberg, N. (1987). The involvement of mRNA secondary structure in protein synthesis. *Biochem. Cell. Biol.* **65**, 576–581.
32. Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M. *et al.* (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.
33. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J. *et al.* (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
34. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K. *et al.* (2000). Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.
35. Yue, P. & Moult, J. (2005). Identification and analysis of deleterious human SNPs. *J. Mol. Biol.* In the press.
36. Wang, Z. & Moult, J. (2003). Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins: Struct. Funct. Genet.* **53**, 748–757.
37. Service, R. (2005). Structural biology. Structural genomics, round 2. *Science*, **307**, 1554–1558.
38. Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L. *et al.* (2004). Database resources of the National Center for Biotechnology Information: update. *Nucl. Acids Res.* **32**, D35–D40.
39. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
40. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
41. Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q. *et al.* (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucl. Acids Res.* **33**, D233–D237.
42. Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.
43. Green, S. M. & Shortle, D. (1993). Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. *Biochemistry*, **32**, 10131–10139.
44. Green, S. M., Meeker, A. K. & Shortle, D. (1992). Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry*, **31**, 5717–5728.
45. Meeker, A. K., Garcia-Moreno, B. & Shortle, D. (1996). Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **35**, 6443–6449.
46. Stites, W. E., Meeker, A. K. & Shortle, D. (1994). Evidence for strained interactions between side-chains and the polypeptide backbone. *J. Mol. Biol.* **235**, 27–32.
47. Schwehm, J. M., Kristyanne, E. S., Biggers, C. C. & Stites, W. E. (1998). Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. *Biochemistry*, **37**, 6939–6948.
48. Byrne, M. P., Manuel, R. L., Lowe, L. G. & Stites, W. E. (1995). Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry*, **34**, 13949–13960.
49. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.* **32**, D226–D229.
50. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucl. Acids Res.* **32**, D120–D121.
51. Samudrala, R. & Moult, J. (1997). Handling context-sensitivity in protein structures using graph theory: *bona fide* prediction. *Proteins: Struct. Funct. Genet.* **1**, 43–49.
52. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014.
53. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
54. Pedersen, J. T. & Moult, J. (1997). Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **269**, 240–259.
55. Oliva, M. T. & Moult, J. (1999). Local electrostatic optimization in proteins. *Protein Eng.* **12**, 727–735.
56. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
57. Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351–371.
58. Laskowski, R. A. M. M., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.
59. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999). The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nature Struct. Biol.* **6**, 594–601.
60. Main, E. R., Fulton, K. F. & Jackson, S. E. (1998). Context-dependent nature of destabilizing mutations on the stability of FKBP12. *Biochemistry*, **37**, 6145–6153.
61. Fulton, K. F., Main, E. R., Daggett, V. & Jackson, S. E.

(1999). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445–461.
62. Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992). Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**, 544–559.

Edited C. R. Matthews

(Received 7 May 2005; received in revised form 8 August 2005; accepted 10 August 2005)
Available online 31 August 2005