PROTEINS: Structure, Function, and Genetics 53:806–816 (2003)

Prediction of Deleterious Functional Effects of Amino Acid Mutations Using a Library of Structure-Based Function Descriptors

Sanna Herrgard, Stephen A. Cammer, Brian T. Hoffman, Stacy Knutson, Marijo Gallina, Jeffrey A. Speir, Jacquelyn S. Fetrow, and Susan M. Baxter*

Cengent Therapeutics, Inc., San Diego, California

ABSTRACT **An automated, active site-focused, computational method is described for use in predicting the effects of engineered amino acid mutations on enzyme catalytic activity. The method uses structure-based function descriptors (Fuzzy Functional Forms™ or FFFs™) to automatically identify enzyme functional sites in proteins. Three-dimensional sequence profiles are created from the surrounding active site structure. The computationally derived active site profile is used to analyze the effect of each amino acid change by defining three key features: proximity of the change to the active site, degree of amino acid conservation at the position in related proteins, and compatibility of the change with residues observed at that position in similar proteins. The features were analyzed using a data set of individual amino acid mutations occurring at 128 residue positions in 14 different enzymes. The results show that changes at key active site residues and at highly conserved positions are likely to have deleterious effects on the catalytic activity, and that non-conservative mutations at highly conserved residues are even more likely to be deleterious. Interestingly, the study revealed that amino acid substitutions at residues in close contact with the key active site residues are not more likely to have deleterious effects than mutations more distant from the active site. Utilization of the FFF-derived structural information yields a prediction method that is accurate in 79–83% of the test cases. The success of this method across all six EC classes suggests that it can be used generally to predict the effects of mutations and nsSNPs for enzymes. Future applications of the approach include automated, large-scale identification of deleterious nsSNPs in clinical populations and in large sets of disease-associated nsSNPs, and identification of deleterious nsSNPs in drug targets and drug metabolizing enzymes. Proteins 2003;53:806–816.**

Key words: single nucleotide polymorphism; SNP; human genome; Fuzzy Functional Form™; protein structure

BACKGROUND

Single nucleotide polymorphisms (SNPs) are sequence differences in single nucleotide positions that exist at least in 1% of the population.[1] SNPs occur on average every 1,000–2,000 bases,[2,3] and account for about 90% of human genetic variation.[4] There are roughly 500,000 SNPs in protein coding regions,[4–6] of which more than half do not lead to amino acid substitutions (synonymous SNPs).[2,7] Those SNPs that exist in protein coding regions and do lead to amino acid substitutions are termed non-synonymous SNPs, or nsSNPs. NsSNPs that alter the structure or function of the encoded proteins cause most of the known monogenic disorders (disorders caused by variation in one gene).[8] Experimental efforts to identify nsSNPs that impact biological processes can be greatly facilitated by using computational methods to predict nsSNPs that are likely to have deleterious effects on protein structure or function. We have developed a fully automated computational method that uses protein sequence and structure information to predict mutations that are likely to exhibit effects specifically on protein catalytic activity. Focusing on enzyme functional sites is key for identifying nsSNPs that might affect biological function, and, thus, such methods have application in the pharmaceutical industry.

In order to predict nsSNPs and other amino acid mutations with significant effects on protein biochemical function, sequence and structure features most likely to affect protein structure and function need to be

---

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Sanna Herrgard and Stephen A. Cammer contributed equally to this work.

Sanna Herrgard's present address is Ambit Biosciences, 9875 Towne Centre Drive, San Diego, CA 92121.

Stephen A. Cammer's present address is Graduate Program in Bioinformatics and UCSD Libraries, 9500 Gilman Drive 0699, La Jolla, CA 92093-0699.

Jacquelyn S. Fetrow's present address is Departments of Physics and Computer Science, Wake Forest University, Winston–Salem, NC 27106.

*Correspondence to: Susan M. Baxter, Cengent Therapeutics, Inc., 5830 Oberlin Drive, Suite 200, San Diego, CA 92121. E-mail: susanbaxter@geneformatics.com

Received 4 November 2002; Accepted 20 February 2003

identified. Recently, Sunyaev et al.,[9] Ng and Henikoff,[10] Chasman and Adams,[11] and Saunders and Baker[12] have investigated protein characteristics to determine their effects on protein structure and function in order to predict the effects of mutations or nsSNPs. Among these features are molecular rigidity as measured by the crystallographic *B*-factor,[11,12] residue conservation,[10–12] compatibility of the substitution with amino acid residues observed at the site in related proteins,[9–12] proximity to the active site or to a ligand,[9,11,12] and electrostatic charge or hydrophobicity changes caused by substitutions of buried residues.[9,11,12] Other protein features investigated include the impact of a substitution at the hydrophobic core on the solubility of the protein, on destruction of a disulfide bond or on insertion of a proline residue into an α-helix.[9,11,12] The large number of protein sequence and structure characteristics investigated raises the question whether a minimal set of features can be identified to successfully predict effects of nsSNPs and other amino acid mutations.

The work presented here describes an automated, active site-focused method utilizing structural information for predicting effects of amino acid changes on enzyme catalytic activity. The active site focus of the method allows us to identify specifically those mutations that are more likely to affect the catalytic activity of the enzyme rather than its folding or stability. This focus is attained using Fuzzy Functional Forms™ (FFFs™),[13–15] which are structural motifs that represent functional sites and are used to identify enzyme active sites in protein structures. Once the location of the site is identified, a three-dimensional (3D) active site sequence profile (Cammer et al., submitted) is created based on the residues located in the structural vicinity of the key functional residues. The profile includes information on allowed residue variability between the sequence under consideration and related proteins. Amino acid changes are then analyzed based on the conservation of the residue position in the active site profile and the compatibility of the substitution with other residues known to occur at that position. Recent work demonstrated the need for inclusion of structural information in deleterious mutation prediction.[12] Utilization of structural information inherent in FFFs and 3D active site profiles produces a tool of reasonable predictive value, even when a high-resolution experimental structure is not available.

In the work described here, the active site-focused approach was first used to study the utility of certain protein features for predicting changes that impact an enzyme's active site and hence, its biochemical activity. A minimal set of features was then selected and used to develop three methods to identify the effect of amino acid mutations on catalytic activity. The results demonstrate that the effect of the mutation can be accurately identified in 79–83% of the test cases across 14 different enzymes. An in-depth study is presented for 21 *Escherichia coli* aspartate aminotransferase (1arg) mutations. Finally, applicability of the approach to large-scale analysis of nsSNPs in sequences where structures have not been experimentally determined and implications for pharmacogenomics and the drug discovery process are discussed.

RESULTS

Overview

The results of this study are presented in three parts. In the first part, the correlation of the following three features to the biochemical effect of the mutation are investigated: proximity of the substituted residue position to the active site; degree of amino acid conservation at the substitution site; and compatibility of the substitution with residue types observed at the site among related proteins. Based on the results obtained in the first part of the study, three methods are developed to predict deleterious mutations for residues within 17 Å of the enzyme active site. The methods and their accuracy for predicting deleterious mutations across 14 different enzymes are presented in the second part of the study. Cross-validation demonstrates that the methods are capable of predicting the effects of mutations not used as part of the training. In the third part, predictions for 21 mutations in *E. coli* aspartate aminotransferase are analyzed in detail to illustrate the strengths and weaknesses of the methods.

Protein Data Set

In order to assess the predictive value of various parameters as accurately as possible, three conditions were established for collection of the mutational data set from the literature: (1) the mutation must exist in an enzyme with a known 3D structure; (2) the effect of the mutation on the k_{cat} and K_M values of the enzyme must be experimentally determined; and (3) enzymes across different enzymatic functions and different Enzyme Commission (EC) classes must be included in the study. In summary, the data set contains 128 mutated residue positions, of which 117 are located within 17 Å of the enzyme active site. Data for 239 individual amino acid changes at these 117 residue positions were available for this study. For example, the information in Table I indicates that amino acid changes are found at 16 residue positions in two of the aspartate aminotransferase structures (1arg and 1oxo). Fourteen of these positions occur within 17 Å of the protein active sites. Twenty-five different residue changes or mutations are identified at these 14 positions. Focusing on position 43 in 1arg, we identify two changes or mutations: His to Ala, and His to Asn (see Supplementary Table, mutational_data_herrgard.xls).

The complete data set covers amino acid changes across 14 different enzymatic functions and across all six major EC classes. A summary of the data set is presented in Table I, and the full data set can be found in the supplementary file mutational_data_herrgard.xls. Because the method focuses on the effect of nsSNPs on enzyme function, each mutant protein in the data set has experimentally determined enzymatic k_{cat}/K_M data. The k_{cat}/K_M value is an expression of the catalytic activity of the enzyme; therefore, by comparing the k_{cat}/K_M for the mutant to the wild-type protein, a quantitative measure of the effect of the mutation on enzyme activity can be

TABLE I. Summary of the Data Set Used in the Study

Enzyme family	E.C. number	No. of positions where mutations occur	No. of positions within 17 Å of active site	No. of total mutations at positions within 17 Å of the active site
Alcohol dehydrogenase (1axe, 1tch)	1.1.1.1	7	7	7
Manganese peroxidase (1mnp)	1.11.1.13	3	3	3
Thymidylate synthase (1nje, 1tis, 1axw)	2.1.1.45	21	17	85
Aspartate aminotransferase (1arg, 1oxo)	2.6.1.1	16	14	25
Nucleoside diphosphate kinase (1ndl)	2.7.4.6	1	1	1
Acetylcholinesterase (2ace)	3.1.1.7	3	3	4
Cysteine protease: papain-like family (1bp4, 1ctc, 1ppo)	3.4.22.x	14	13	17
Serine carboxypeptidase (1ysc)	3.4.16.5	4	4	13
Subtilisin (1cl3, 1sup, 1mpt)	3.4.21.62	14	14	23
Carbonic anhydrase II (2cba)	4.2.1.1	6	6	6
Fructose-1,6-bisphosphate aldolase (1ado, 1zen)	4.1.2.13	11	10	10
Xylose isomerase (1a0c, 1bhw)	5.3.1.5	21	18	38
Muconate cycloisomerase (1muc)	5.5.1.1	6	6	6
Class II tRNA synthetase (1htt)	6.1.1.21	1	1	1
Total	14	128	117	239

obtained. Mutations with k_{cat}/K_M values equal to or less than 20% of the wild type are defined here to be deleterious. We chose to use a quantitative assessment of enzyme efficiency as the basis for the method, and a k_{cat}/K_M ratio of 0.2 seemed to be a reasonable value from a biochemical standpoint. This relatively large reduction of enzymatic activity would certainly hamper biological function of a protein both in vitro and in vivo. Because of the requirement for experimentally determined k_{cat} and K_M values, the data set is not large. We do not expect such a data set to yield exact determination of the accuracy of the method, but rather to demonstrate the utility of the approach.

Summary of Scoring Functions

The impact of changing an amino acid at each position is investigated by calculating conservation and substitution scores from the 3D active site profiles. As designed for this study, the 3D active site profiles define the allowed residue types for each position within 17 Å of a given active site. The conservation score takes into consideration the types of residues that are observed at analogous positions in similar proteins, whereas the substitution score indicates whether each individual substitution is compatible, from a chemical standpoint, with the amino acid types that are identified by the conservation score. Both conservation and substitution scores range from 1.0 (high impact) to 0.0 (low impact).

The conservation score indicates the likely impact of a mutation at a position based on the chemical variability of residues found at that position. It is determined from the types of amino acids found at each residue position in the 3D profiles (see Methods). A score of 1.0 (high impact) is given to fully conserved positions, 0.75 to positions in which variations are found within one chemical group (e.g., {LIV} and {KR} form two different chemical groups, as defined in the Methods), 0.5 to positions with residues in two chemical groups, 0.25 to positions with three chemical groups, and 0 to positions with more than three chemical groups. For instance, position 39 in *E. coli* aspartate aminotransferase receives a low-impact conservation score of 0.25. The 3D conservation profile indicates that residues Val, Ala, Ile, Thr, and Leu are allowed at position 39 and these residues are found in three chemical groups ({LIV}, A, and {ST}). Thus, amino acid changes at this position are not likely to adversely affect enzyme activity.

The substitution score indicates the likely impact of an amino acid change at a particular position. A high impact substitution score of 1.0 is given to substitutions that are not represented at the site in the 3D active site profile and are not chemically conserved (as defined by the modified Dayhoff groups,[16] see Methods), 0.5 is given to residues that are not present in the profile but are chemically conserved, and 0 to residues that are present in the profile. For instance, mutation Val39Leu in *E. coli* aspartate aminotransferase receives a low-impact substitution score. The 3D conservation profile indicates that residues Val, Ala, Ile, Thr, and Leu are allowed at position 39; thus, Val39Leu receives a low-impact substitution score of 0 because the substitution (Leu) is represented in the profile.

Correlation of Protein Structure and Function Features With Enzyme Activity

To investigate whether proximity of a mutation to the enzyme active site correlates with deleterious effects on enzyme catalytic activity, the distance of the residue position (where mutation occurs) from the key active site residues was the first feature analyzed. Thus, the minimum distance between the C^{α} atom position of the substituted residue and the C^{α} atom position of the nearest key active site residue, as defined by the FFF, was calculated for a data set of 128 positions of mutated residues. These data demonstrate that 94% (17 out of 18) of the mutations of key active site residues have deleterious effects on enzyme catalytic activity [Fig. 1(A)]. This is an expected result. FFFs are designed to identify key residues important for the chemistry and function of an active site; thus,

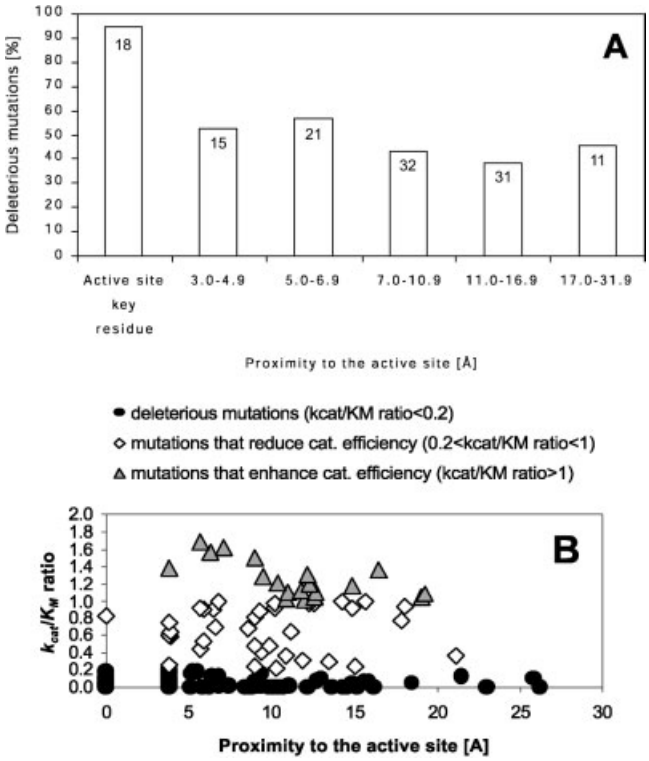

Fig. 1. Correlation of deleterious mutations with distance to functional site. **A:** A histogram showing the percentage of deleterious mutations found within six distance ranges from the active site. The total number of mutated positions in each category is shown in the bar. Deleterious mutations are defined as having less than 20% of the experimentally measured catalytic activity (k_{cat}/K_M) of the wild-type. **B:** The k_{cat}/K_M ratio is plotted against the distance from the FFF identified active site for deleterious (closed circles) and non-deleterious mutations (open diamonds, $0.2 < k_{cat}/K_{Mv} \leq 1$; gray triangles, $k_{cat}/K_{Mv} > 1$). Only mutations with k_{cat}/K_M mutant to wild-type ratios ≤ 2.0 are plotted.

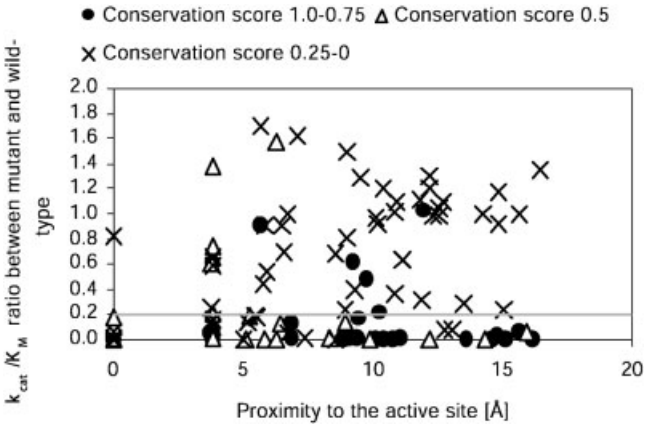

Fig. 2. The effect of residue conservation on catalytic activity. Conservation score for a position and the effect on catalytic activity plotted mutation at that position is plotted vs. the minimum distance between the position and one of the active site residues. Residues having high (1.0–0.75, filled circles) or medium (0.5, open triangles) conservation scores are likely to have low k_{cat}/K_M mutant to wild-type ratios, and residues with low conservation scores (0.25–0.0) are more likely to have k_{cat}/K_M mutant to wild-type ratios 0.2 or higher. Deleterious mutations (deleterious mutation defined as experimentally measured k_{cat}/K_M ratio between mutant and wild-type ≤ 0.2) fall below the horizontal line on the graph. Only k_{cat}/K_M mutant to wild-type ratios ≤ 2.0 are shown.

mutation of these residues would be expected to eliminate or severely limit catalytic function. The other data in Figure 1 indicate that if the amino acid change is not at an FFF-identified catalytic residue, the distance to these residues is not generally correlated with the effect of the mutation on catalytic activity. Figure 1(B) shows that 93% (65/70) of the deleterious mutations fall within 17 Å of one of the C^{α} positions of an active site residue recognized by the FFF. The remaining deleterious mutations (5/70) are more than 19 Å from C^{α} positions of active site residues recognized by the FFF. Our methodology is geared toward analyses of active sites and so we limited further analyses to the 117 residue positions within 17 Å of the active sites in our data set since we had a reasonable set of data for the testing of our methods. Applications of the method, however, are not confined by this limit.

The relationship between residue conservation at a particular position and the effect of a mutation at that position was investigated by calculating conservation scores for the 117 residue positions where mutations were found. Figure 2 illustrates the position conservation score and the effect of the mutation on catalytic activity vs. the minimum distance between the position and one of the active site residues identified by the FFF. The five conservation score values (1.0 , 0.75, 0.5, 0.25, and 0) allow for the assessment of the effect of conservation, particularly as it relates to distance from the active site, upon the catalytic efficiency of an enzyme. Upon evaluation of the conservation score data (as shown in Fig. 2), it becomes apparent that a score threshold of 0.5 is sufficient for analysis of the impact of the residue change. Thus, although five score values are used to assess the conservation at each residue position, only two are required for an adequate prediction of the effect of such a change in residue.

The effect of a mutation is correlated with the degree of conservation. Eighty-two percent of substitutions at highly conserved positions (conservation score 1.0–0.75), and 83% of substitutions at positions with medium conservation score (0.5) exhibit a k_{cat}/K_M ratio that is less than 20% of the wild-type k_{cat}/K_M. In contrast, only 59% of the positions with conservation score 0.25 and 14% with score zero exhibit this effect on catalytic activity when mutated. These results demonstrate that, in general, mutations at residue positions with a conservation score ≥ 0.5 have a higher probability of exhibiting a deleterious effect on catalytic activity than those with lower conservation scores. Analysis of the conservation score variation against the experimental k_{cat}/K_M ratio values, using a χ^2 test, yields a statistically significant P value of $1e^{-8}$ when calculated for all five conservation score values, and $1e^{-9}$ when calculated using only the 0.5 threshold for conservation value, supporting our use of this threshold to recognize trends in the data.

A substitution score was calculated for the full data set of 239 single amino acid changes that occur at the 117 positions or substitution sites located within 17 Å of the enzyme active sites. Eighty-seven percent of amino acid changes with a substitution score of 1.0 exhibit deleterious

TABLE II. Prediction Accuracies and Matthews Correlation Coefficients for the Three Prediction Methods[†]

	Method A	Method B	Method C
Total positions at which mutations occur	117	117	117
Mutations at those positions	NA	NA	239
Deleterious positions/mutations	62	62	164
Non-deleterious positions/mutations	55	55	75
Deleterious mutations predicted correctly	46	50	149
Non-deleterious mutations predicted correctly	44	43	49
Deleterious mutations falsely predicted	16	12	15
Non-deleterious mutations falsely predicted	11	12	26
% Deleterious mutations predicted correctly	74	81	91
% Non-deleterious mutations predicted correctly	80	78	65
% All mutations predicted correctly	77	79	83
Matthews correlation coefficient	0.54	0.59	0.59

Method A: Residues at which mutations are predicted to be deleterious have a conservation score of 0.5 or higher. Method B: Residues at which mutations are predicted to be deleterious have a conservation score of 0.5 or higher, or are key active site residues. Method C: Residues at which mutations are predicted to be deleterious have a conservation score of 0.5 or higher, or are key active site residues, or the mutation has a substitution score of 1.0.

[†]NA: The number of substitutions found at the 117 positions is not relevant to the calculations forMethods A and B.

effects, whereas 76% of the changes with a substitution score of 0.5, and 39% with a substitution score of 0.0, exhibit deleterious effects. This result suggests that mutations with high substitution scores (1.0) can have deleterious effects. Looking at the combination of conservation and substitution scores, a high percentage (92%) of positions with a conservation score of 0.5–1.0 and a substitution score of 1.0 are deleterious. Thus, mutations with a combination of at least a medium conservation score (0.5–1.0) and a high impact substitution score are very likely to be deleterious. These observations were used to develop simple methods using a minimal set of protein features to predict amino acid changes that exhibit deleterious effects on catalytic activity.

Methods for Prediction of Deleterious Mutations

The analysis presented above suggests that two features can be used to predict mutations and nsSNPs with deleterious effects on enzyme catalytic activity: degree of amino acid conservation at the substitution site and compatibility of the substitution with residue types observed at that site among related proteins. We hypothesized that the following conditions could be used in some combination to predict whether or not a mutation within 17 Å of the active site would be deleterious: (1) the substitution site has a conservation score of 0.5 or higher; (2) the changed residue is a key active site residue; or (3) the mutation receives a substitution score of 1.0. To examine the combination of conditions that would optimize prediction accuracy, three different sets of predictions were performed: for Method A condition (1) is true; for Method B either condition (1) or (2) is true; and for Method C one or more of conditions (1), (2), or (3), is true. The first two predictions were performed using the data set of 117 residue positions or substitution sites at which mutations within 17 Å of the active site are found. The third prediction was performed using the full data set of 239 mutations at those 117 positions in the set of 14 enzymes.

The prediction accuracy of the three methods was evaluated using two different measures. First, the ability of the method to predict deleterious and nondeleterious mutations as well as the overall prediction accuracy was assessed. Second, the Matthews correlation coefficient[17] was calculated to investigate whether the method tends to under- or overpredict deleterious effects. The results of applying the three prediction methods are presented in Table II. As expected, the prediction accuracy for deleterious mutations and the overall prediction accuracy increase when more conditions are included in the prediction method. The highest prediction accuracy for deleterious mutations (91%) as well as the highest overall prediction accuracy (83%) are achieved when all three prediction conditions are applied (Method C). The data also show that while the prediction accuracy for the deleterious mutations increases from 74 to 91%, the prediction accuracy for the non-deleterious mutations decreases from 80 to 65%. Thus, Method C shows overprediction of nondeleterious effects, as well as better prediction of deleterious mutations (Table II). The aim of the method is to identify all possible deleterious mutations or nsSNPs in large data sets so that they can be singled out for further studies. From this perspective, slight overprediction of nondeleterious effects might be tolerated if high coverage in predicting deleterious effects is achieved and the cost of testing more false positive predictions is acceptable. The Matthews correlation coefficient is 0.54 for Method A and 0.59 for Methods B and C. This analysis indicates that the prediction accuracy is statistically significant for Methods B and C, and that Method A performs worse than the two other methods. Method C was chosen for subsequent evaluation, recognizing that it does result in overprediction of nondel-

eterious mutations and its correlation coefficient indicates overall performance similar to Method B (Table II).

In the cross-validation tests, the prediction accuracy for the deleterious mutations varies between 67–100%, the average being 85%, whereas the prediction accuracy for the nondeleterious mutations is 71–90%, with an average of 81%. The higher than expected prediction accuracy for the nondeleterious mutations compared to the full data set (Table II) can be explained by the small number of nondeleterious mutations included in the randomly chosen test data set. The Matthews correlation coefficient varies between 0.49–0.91 with an average of 0.65 $\pm$ 0.08, which is comparable to the coefficient (0.59) obtained on the test set as a whole. The comparable Matthews coefficients obtained in the cross-validation studies demonstrate that the method is robust. Although the external data sets used for testing the method here are small (23 mutations), the method is not overly sensitive to training set composition and is capable of predicting mutational effects that are not included in the training set.

To investigate whether the prediction accuracy of deleterious mutations is comparable across different enzymes, the prediction accuracy of Method C was calculated for each of the 14 enzymes individually. Figure 3 shows that the false predictions are distributed fairly evenly among the different enzyme functions and that the prediction accuracy is $\geq$ 80% for every enzyme with more than 20 mutations (92% for thymidylate synthase, 80% for aspartate aminotransferase, 83% for subtilisin, and 84% for xylose isomerase). The prediction accuracy varies more for enzymes with fewer than 20 mutations (Fig. 3), probably due to the limited sample size. The prediction accuracy for deleterious mutations across different enzyme functions suggests that the method can be used to predict automatically effects of mutations and nsSNPs on enzyme catalytic activity in large-scale pharmacogenomic studies.

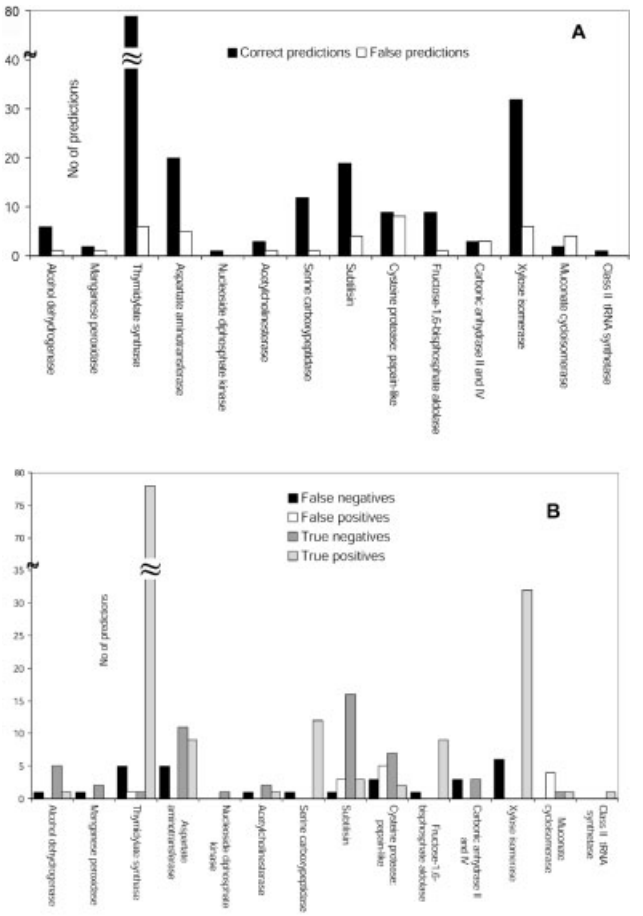

Fig. 3. Accuracy of the predictions of deleterious and nondeleterious mutations. **A:** Number of correct (black bars) and false (white bars) predictions of deleterious mutations for each of the 14 enzymes included in the test data set. **B:** Number of false negatives (black bars), false positives (white bars), true negatives (dark gray bars), and true positives (light gray bars) for each of the 14 enzymes included in the test data set.

Analysis of the Correct and False Predictions in *E. coli* Aspartate Aminotransferase

To demonstrate some of the strengths and weaknesses of Method C, its performance on predicting effects of amino acid changes in *E. coli* aspartate aminotransferase (1arg) was analyzed. *E. coli* aspartate aminotransferase contains 11 residue positions at which substitutions occur and 21 different amino acid changes at these positions. Eight of these mutations are known to exhibit deleterious and 13 are known to exhibit non-deleterious effects on enzyme activity (Table I; for details, see structure 1arg in the supplementary file mutational_data_herrgard.xls). The FFF-identified key active site residues (Tyr 225, Lys 258, and Arg386) and the mutated positions in the *E. coli* aspartate aminotransferase structure are displayed in Figure 4.

All eight deleterious mutations reported for residues His143, Pro195, Tyr225, and Arg386 were correctly assessed by prediction Method C. The FFF designed to identify the aspartate aminotransferase active site indicates that Tyr225 and Arg386 are key active site residues (involved in cofactor and substrate binding, respectively). Thus, Method C designates substitutions at these positions as deleterious. His143 is one of the three histidine residues located beneath the coenzyme pyridine ring in the active site,[18] is fully conserved, and, therefore, has a high impact conservation score of 1.0. Pro195 and has a medium conservation score of 0.5. Pro195 is actually fully conserved, but has only a medium conservation score due to the inclusion of the double mutant P138A/P195A structure (1bdq) in the 3D active site profile. This result demonstrates the need to eliminate structures of engineered or mutant proteins, and their concomitant confounding effects, from the sequence set used to determine 3D profiles.

The remaining data set contains 13 nondeleterious amino acid changes at seven positions: Val39, Cys82, Pro138, Cys191, Cys192, Cys270, and Cys401. Five substitutions at positions 39, 82, 138, 270, and 401 were correctly assessed by our method. None of the amino acids at these positions is a key active site residue and all have low-impact conservation and substitution scores consistent with the parameters defined by Method C. Seven nondeleterious mutations have been reported for position 191,

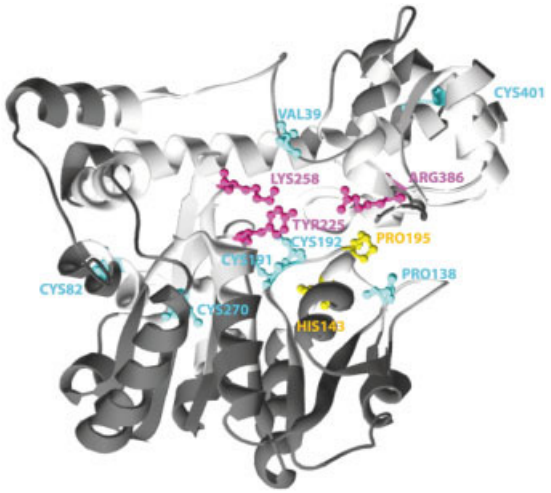

Fig. 4. Location of the active site and mutated residues in the three-dimensional structure of aspartate aminotransferase from *E. coli* (1arg). The protein backbone is shown as a ribbon, with the lighter ribbon indicating the residues within 17 Å of the FFF-identified active site residues. Side chains of the key catalytic residues identified by the FFF, Lys 258, Tyr 225, and Arg 386, are shown as pink ball and stick models. Mutation of two of these, Arg 386 and Tyr 225, as well as mutations at Pro 195 and His 143 (shown in yellow), cause deleterious effects on catalytic activity. All deleterious mutations were correctly predicted in *E. coli* aspartate aminotransferase. Positions at which nondeleterious mutations occur are shown as cyan ball and stick models. Five of these, mutations at positions 39, 82, 138, 270, and 401, were correctly assessed by Method C.

three of which were correctly assessed and four of which were incorrectly assessed. The four substitutions (Phe, Trp, Tyr, and Arg) identified by Method C as deleterious mutations are overpredicted as a result of high substitution scores. The 3D active site profile for *E. coli* aspartate aminotransferase contains only residues Cys, Leu, Thr, and Ser at this position, thus Phe, Trp, Tyr, and Arg are scored as high-impact substitutions because they do not fall in the same chemical groups as Cys, Leu, Thr, and Ser. The 3D active site profile contains 108 sequences with cysteine at this position, while only three sequences contain Leu, Thr, or Ser, indicating that Cys has been very well conserved at this position. It has been proposed that Cys191 is actually a nonessential residue that has been well conserved because there is no pathway of neutral mutations to the conservative mutation Ala191, suggesting that Cys191 is a "frozen accident."[19] The nondeleterious Cys192Ser mutation was also falsely assessed as deleterious due to a medium conservation score (0.5) for the residue. The overall correct prediction accuracy for the 21 mutations in *E. coli* aspartate aminotransferase is 76%.

DISCUSSION

The automated method presented here integrates a minimal set of features describing protein sequence, structure, and function to predict whether engineered mutations are likely to have a deleterious effect on enzyme catalytic activity. Deleterious is defined here as a k_{cat}/K_M ratio of the mutant enzyme that is ≤20% of the wild-type k_{cat}/K_M. FFFs are used to identify active sites and protein structure coordinates are used to calculate proximity to the active site. Sequence information is used, together with structure information, to create 3D active site profiles (Cammer et al., submitted) needed to calculate conservation and substitution scores. The study was carried out on engineered mutant proteins but should be generally applicable to nsSNPs.

The prediction accuracy of the method was assessed using a data set of 239 single amino acid changes with known effects on enzyme k_{cat}/K_M ratio. Methods B and C achieved an overall prediction accuracy of 79–83% on this small sample of proteins (Table II). The 83% accuracy of Method C is obtained at the expense of increased false positive prediction of non-deleterious mutations. Because the data set is necessarily small, future application of this tool to larger sequence sets will demonstrate the actual accuracy of these methods.

Comparison to Other Methods

The goal of this study was to design a simple, automated method using a minimal set of protein features that is applicable to large-scale analysis of the effect of amino acid changes on enzyme function. The first step was the identification of enzyme active sites using FFFs. Mutations at specific FFF-identified residues are almost always deleterious (Fig. 1), as expected given the focus of FFF construction on identifying the key catalytic and chemically important residues.[13] In the next step, an active site profile was computed based on the residues and their structures found in proximity to the FFF. Analysis of the location of deleterious mutations shows that the majority of such mutations in this small data set are situated within 17 Å of the FFF residues (Fig. 1). Thus, we limited our analysis to mutations located in this range. Applications of the method, however, are not confined by this limit.

Once residues within 17 Å of the active site are identified, only two characteristics, residue conservation among related proteins and identity of the substituted residue, are used to distinguish between deleterious and nondeleterious mutations. These features are similar but more simplistic than the rules published by Ng and Henikoff.[10] The main difference between our method and Ng and Henikoff's method is the application a minimal set of rules specifically to functional sites and associated 3D profiles identified by the FFFs. The second difference is the ability to automate and apply the current method on a large scale, utilizing key structural information inherent in the FFFs. The third difference is the focus on enzyme activity, as distinct from structure and stability, allowing assessment of impact on biological function. Such an application is valuable for target validation in the pharmaceutical industry.

A detailed and direct comparison of the prediction accuracy of Method C to the prediction accuracies of the methods presented earlier by other groups[9–12,20] is not easily accomplished due to the differences in the data sets used for testing. Ng and Henikoff,[10] and Chasman and Adams[11] used selected protein representatives, including lacI and lysozyme, which have rich information available about effects of mutation on protein structure, folding, and

function. On the other hand, Sunyaev et al.[9] used SWISS-PROT[21] sequences with mutations implicated in affects on protein structure, function or stability, or are associated with a disease phenotype. Saunders and Baker used both sets.[12] The methods described here are directed at identifying mutations with an effect on enzyme activity, thus the data set in this study necessarily focused on amino acid changes where effects on k_{cat}/K_M had been experimentally determined. In addition, to demonstrate generality it was necessary to identify such data for a variety of catalytic functions. As a result, the test data set does not include nonenzymatic functions, such as protein-DNA or protein-protein interactions. Other groups[9,11,22] have used their results and analyses to estimate the percentage of deleterious nsSNPs, whereas we have focused mainly on developing an automated, active site-focused method to predict the effects of amino acid changes on catalytic activity of target proteins. Other methods that require structure information for their analysis have false positive errors of ~30%.[9,11] The data presented here show a false positive error rate of 22–35% on nondeleterious mutations (Methods B and C, Table II). These results indicate, with this limited test data set, that the prediction accuracy of this automated method, focused on a minimal set of features to identify effects on enzyme activity, exhibits accuracy similar to that of other published methods.

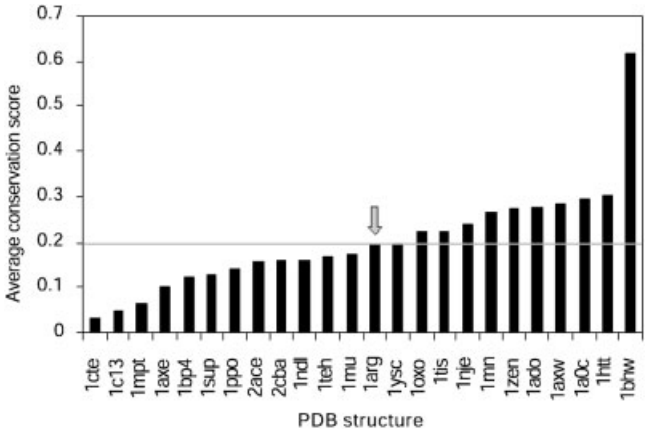

Fig. 5. Average conservation scores calculated from the 3D profile for each enzyme included in the study. Key active site residues are identified by the FFF. The 3D active site profile is then calculated for residue segments that fall within 17 Å of the active site residues (see Methods). An average conservation score was calculated by taking the mean of the individual conservation scores of all residues in the 3D profile. The average conservation score for residues in the active site profile ranges from 0.03 to 0.3, except 1bhw. These data demonstrate that a significant number of residues in the vicinity of the active site are not at all conserved. The average value (0.20) is indicated with a horizontal line. *E. coli* aspartate aminotransferase (1arg) is indicated with a gray arrow.

Limitations and Potential Improvements of Current Method

Further analysis was performed to investigate mutations that were not correctly predicted. Five factors can be associated with false predictions: inclusion of mutant sequences in the initial calculation of the 3D active site profile; prediction of effects of mutations for residues that are highly conserved but have no structural or functional role ("frozen accidents"); an unusually high or low average conservation score for the enzyme; proximity to the active site; and conformational changes at the active site. The analysis of mutations in *E. coli* aspartate aminotransferase reveals the confounding effects of including sequences exhibiting double mutations in the 3D active site profile. It also demonstrates the inability of the methods to distinguish between "frozen accidents" and residues that are conserved for fundamental reasons. Applying a more sophisticated probability-based scoring scheme, as employed by Ng and Henikoff[10,22] to calculate the conservation score could potentially reduce the error caused by inclusion of mutant sequences in the 3D active site profile. This way, one or two mutant sequences in the profile would not affect the conservation score significantly if the number of similar sequences in the profile is large enough.

The effect of the average conservation score on the prediction accuracy was investigated by calculating an average conservation score for the active site fragments (i.e., residue sets within 17 Å of an FFF residue) from the 3D profile of each enzyme included in the study (Fig. 5). It can be seen in Figure 5 that the average conservation scores are evenly distributed between 0.03 and 0.30, with 1bhw (xylose isomerase from *Actinoplanes missouriensis*) being the exception with a score of 0.62. These data demonstrate that not all residues within 17 Å of the active site are well conserved on average. An unusually high or low average conservation score can be a result of either an unusual number of similar sequences in the NCBI nonredundant protein sequence database or by an unusual degree of conservation among the sequences. While no general correlation could be observed between prediction accuracy and the average conservation score due to the limited size of the data set, we suggest that an unusual value for the average conservation score may indicate possible under- or overprediction of deleterious effects.

The fourth factor associated with reduced prediction accuracy is proximity to the active site. The prediction accuracy of Method C for deleterious mutations is only 75% for residues that are located within 3.0–6.9 Å of enzyme active site, whereas the accuracy is over 90% for residues beyond 7.0 Å (Fig. 6). The prediction accuracy for residues close to the active site is lower because the correlation between catalytic activity and residue identity does not always hold for these proximate residues. Any type of change in hydrogen bonding, van der Waals, or polar interactions in residues in close contact with the key active site residues may have a profound effect on the active site chemistry and the catalytic activity of the enzyme, irrespective of whether the residue is conserved or not. It should be noted that the mutational data set also demonstrates that mutations in residues in close contact with the key active site residues are not more likely to have deleterious effects than mutations elsewhere in the protein (Fig. 1). Taken together, these findings show that there is more uncertainty in predicting deleterious effects for mutations that occur in residues in close contact with the key active site residues. A detailed analysis of protein

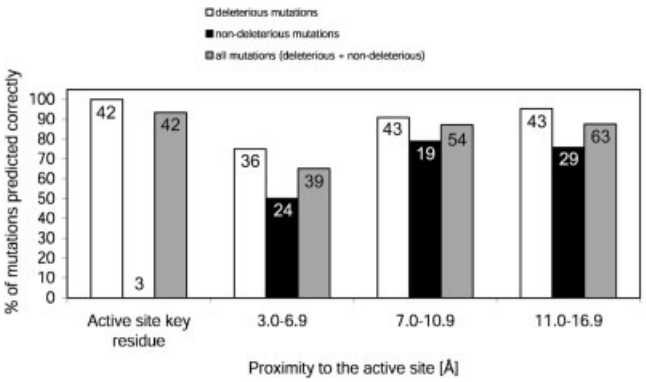

Fig. 6. Prediction accuracy of mutations. Prediction accuracy for deleterious (white bars), non-deleterious (black bars), and all (deleterious plus nondeleterious; gray bars) mutations occurring in four different distance ranges from the FFF-identified active site. The total number of mutations in each category is shown (left).

structure including electrostatic calculations could be used to gain predictive accuracy in the immediate vicinity of the key active site residues. However, these calculations are difficult to automate and are not easily done for large-scale studies.

A fifth factor that reduces prediction accuracy is structural changes that affect protein activity. Indeed, dynamic or conformational changes resulting from protein-protein interactions or cofactor binding can impact catalytic activity. If there is no experimental structure available that illustrates these conformational changes, they cannot be accounted for by the current method.

FUTURE APPLICATIONS

Although this study was carried out using enzymes with known 3D structures, the method can also be applied to proteins with no known structure. For proteins with no known structure, threading can be applied to obtain approximate structure models and these approximate models can be used in conjunction with the FFFs to identify active site residues and to create the 3D profiles. The ability of the FFFs to automatically and rapidly identify enzyme active sites using approximate structure models obtained from threading has been demonstrated for genome sequences.[14,15] Extension of these methods for application to computational models would allow structural information to be utilized in large-scale nsSNP analyses of sequences where no structure is known.

The method presented in this study has several applications in the pharmaceutical field. First, it is an effective means of utilizing structural information to identify the nsSNPs that likely cause deleterious effects on enzyme activity in disease-related sequences, and, thus, has the potential to speed up discovery of disease-related alleles. Secondly, our prediction method facilitates the identification of nsSNPs at the active sites of drug targets. Certain nsSNPs can result in the loss of drug potency and/or efficacy since they can cause changes in active site structure or chemistry that would interfere with drug binding or change protein target activity. Failure to recognize these active site changes may lead to unsuccessful or inconclusive clinical trials if occurrence of an active site nsSNP is significant in the clinical trial population. The ability to recognize deleterious nsSNPs in drug targets and to use that information to select clinical populations could be of great value to the process of bringing a drug to market. Lastly, the method could be used in efforts to engineer protein stability and activity. As the field of creating new proteins with enhanced or novel biochemical properties matures, the utility of a method for distinguishing or guiding changes in protein sequence becomes more valuable, especially for the design of protein therapeutics. This method may well aid researchers aiming to simultaneously improve upon beneficial, and avoid detrimental, changes in amino acid composition, even in the absence of an experimentally determined protein structure.

MATERIALS AND METHODS

Data Set

A data set of 14 different enzymes of known structure with mutations occurring at 128 different positions was identified from the literature. Detailed data were collected for 239 amino acid mutations that are found at the 117 positions within 17 Å of the active site. Only mutations with experimentally determined effects on enzyme k_{cat} and K_M values were included in the data set to ensure that the effects of the mutations are comparable across different enzyme classes. Table I contains a summary of the data set. The full data set for the 117 positions and 239 mutations can be found in the supplementary table, mutational_data_herrgard.xls. The data set contains mutations across all six major EC classes. Observations based on this data set were used to determine the parameters for identifying amino acid changes that affect catalytic activity. Cross-validation demonstrates that removing subsets from this data set does not change the observations on which the parameters were developed.

FFF and 3D Active Site Profile Construction

FFFs for the enzymes used in this analysis were constructed essentially as previously described.[13] Briefly, the key residues necessary for protein function and chemistry are identified. Geometric constraints, and variations of these constraints that uniquely identify the relative structure of these residues, are determined. These constraints are modified in an iterative fashion so that known functional sites are fully separated from false positives in a large data set of known PDB structures. To assure generalizability, the resulting FFF (residue identities and geometric constraints) is then cross validated against the all proteins in the training and testing sets. In this study, the FFFs are applied only to experimentally determined protein structures. However, as previously demonstrated these structural motifs can also be applied to approximate protein models.[14]

Three-dimensional active site profiles were then generated for each protein sequence in which the functional site is identified by the FFF (Cammer et al., submitted).

Briefly, active site profiles are constructed by identifying amino acid sequence fragments within 17 Å of the key active site residues. BLAST[23,24] was used to search for similar sequences in the NCBI non-redundant protein sequence database using the active site fragments as the query. BLOSUM-62 was used as a substitution matrix, and an *E*-value cut-off of 0.01 was applied. Multiple sequence alignments of the resulting fragments were generated using ClustalW.[25] This multiple sequence alignment based on the fragments surrounding the active site is defined as an active site profile.

Scoring Methods Based on Distance, Conservation, and Substitution

The methods described here integrate three aspects of protein sequence and structure and determine their effects on nsSNP analysis: proximity of the substituted residue to the active site; degree of amino acid conservation at the substitution site; and compatibility of the substitution with residue types observed at the site among related proteins. To determine the proximity of the substituted amino acid residue to the active site, the minimum distance between the Cα-atom of the substituted amino acid residue and the Cα-atoms of the key active site residues identified by the FFF was calculated.

The degree of conservation of each position at which a mutation occurs was evaluated by calculating a conservation score from the 3D active site profile. To calculate this score, amino acids were classified into groups corresponding to the chemical nature of their side chains: A, C, {DE}, {FWY}, G, H, {LIV}, M, {NQ}, P, {ST}, {KR}. The following scoring rules were applied: a score of 1.0 (high impact) was given to fully conserved positions; 0.75 to residue positions where all residue substitutions identified from the active site profile fall within one chemical group; 0.5 to those positions where residues fall within two chemical groups; 0.25 to those where residues fall in three chemical groups; and 0 to those where residues fall within more than three chemical groups.

A substitution score was calculated to evaluate whether the substituted residue is compatible, from a chemical standpoint, with the amino acid residues observed at an analogous position in similar proteins. A score of 1.0 (high impact) was given to substitutions not represented in the 3D active site profile and not chemically conserved; 0.5 to residues not present in the profile but are chemically conserved; and 0 to residues present in the profile. A chemically conserved group was defined to include all residue types in the modified Dayhoff groups[16] for each residue in the 3D active site profile. The modified Dayhoff groups were defined as follows: {VILM}, {AGSPTC}, {FWYH}, {RKH}, {DENQ}.

Evaluation of the Performance of the Prediction Methods

The performance of the prediction methods on identifying impact on catalytic activity was evaluated using two different measures. First, the percentages of correct predictions were calculated for the following sets: deleterious mutations (k_{cat}/K_M ratio between mutant and wild-type $\leq$ 0.2); nondeleterious mutations (k_{cat}/K_M ratio $>$ 0.2); and all mutations. As an example, percentage of correct predictions on all mutations was calculated as: % of all mutations predicted correctly =

$$f_d p_d + f_n p_n \tag{1}$$

where f_d is a fraction of total mutations that are deleterious, f_n is a fraction of total mutations that are nondeleterious, p_d is the % of correctly predicted deleterious mutations, and p_n is the % of correctly predicted nondeleterious mutations.

Second, the Matthews correlation coefficient[17] was used to evaluate whether the prediction method has a tendency to under- or overpredict deleterious effects. The coefficient is calculated as follows:

$$C_d = \frac{(n_d n_n) - (u_d o_d)}{\sqrt{(n_n + u_d)(n_n + o_d)(n_d + u_d)(n_d + o_d)}}, \tag{2}$$

where, n_d is the number of correctly predicted deleterious mutations, n_n is the number of correctly predicted nondeleterious mutations, u_d is the number of underpredicted deleterious mutations (mutations that were predicted to be nondeleterious, but are deleterious), and o_d is the number of overpredicted nondeleterious mutations (mutations that were predicted to be deleterious, but are nondeleterious). A correlation coefficient of $C_d = 1$ indicates a perfect agreement between the prediction and the observation, $C_d = 0$ for predictions no better than random, and $C_d = -1$ when the prediction and the observation are in total disagreement.

Cross-Validation of the Methods

A cross-validation of Method C was performed to ensure that the methods described here could predict effects of mutations that were not included in the training set. The cross-validation was performed using a data set of 117 mutated residues located within 17 Å of the active site. Ten independent tests were performed so that 80% of the mutations (94 mutations) were chosen randomly and were used for training the method, i.e., determining the appropriate cut-offs for the parameters used in the study: degree of amino acid conservation at the substitution site and compatibility of the substitution with residue types observed at the site among related proteins. Once these parameters had been determined for each cross-validation set, the resulting model was then used to predict deleterious mutations for the remaining 20% of the data withheld from the training set (23 mutations). The performance of the method in each of these 10 cases provides a measure of robustness of the process, since any inherent sensitivity to data composition should result in poor results for a portion of the cross-validation runs.

Additional Data

The mutational data set used in this study is available in the supplemental file mutational_data_herrgard.xls. The data set contains the following information for each of the

239 mutations: enzyme, structure used in this study, residue number, wild-type and mutated residue types, k_{cat}/K_M mutant to wild-type ratio, and reference.

ACKNOWLEDGMENTS

The authors thank Jeannine Di Gennaro for her help with collecting the mutational data and Ruth Feldblum for her efforts in editing and assembling the manuscript. Professors Jeffrey Skolnick and Andrzej Kolinski are warmly acknowledged for invaluable discussions.

REFERENCES

1. Brookes AJ. The essence of SNPs. Gene 1999;234:177-186.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD. The sequence of the human genome. Science 2001;291:1304–1351.
3. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, SteinGabor Marth LD, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok P–Y, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterson RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Strange-Thomann N, Zody MC, Linton L, Lander ES, Attshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–933.
4. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8:1229–1231.
5. Pirmohamed M, Park BK. Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 2001;22:298–305.
6. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 2001;294:1719–1723.
7. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:231–238.
8. Syvanen A-C. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2001;2:930–942.
9. Sunyaev S, Ramensky V, Koch I, Lathe WIII, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591–597.
10. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–874.
11. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307:683–706.
12. Saunders C, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 2002;322:891.
13. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
14. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 1998;282:703–711.
15. Fetrow J, Siew N, Di Gennaro J, Martinez-Yamout M, Dyson J, Skolnick J. Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight? Protein Sci 2001;10:1005–1014.
16. Dayhoff MO, Eck RV. Atlas of protein sequence and structure. Silver Spring: Natl Biomed Res Found; 1968.
17. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
18. Yano T, Kuramitsu S, Tanase S, Morino Y, Hiromi K, Kagamiyama H. The role of His143 in the catalytic mechanism of Escherichia coli aspartate aminotransferase. J Biol Chem 1991;266:6079–6085.
19. Gloss LM, Spencer DE, Kirsch JF. Cysteine-191 in aspartate aminotransferases appears to be conserved due to the lack of a neutral mutation pathway to the functional equivalent, alanine-191. Proteins 1996;24:195–208.
20. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–270.
21. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
22. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res 2002;12:436–446.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
25. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.