Published in final edited form as:

Proteins. 2010 July ; 78(9): 2058–2074. doi:10.1002/prot.22722.

Testing computational prediction of missense mutation phenotypes: Functional characterization of 204 mutations of human cystathionine beta synthase

Qiong Wei¹, Liqun Wang², Qiang Wang¹, Warren D. Kruger², and Roland L. Dunbrack Jr.^{1,*}

¹Program in Molecular and Translational Medicine, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111

²Program in Cancer Genetics and Signaling, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111

Abstract

Predicting the phenotypes of missense mutations uncovered by large-scale sequencing projects is an important goal in computational biology. High-confidence predictions can be an aid in focusing experimental and association studies on those mutations most likely to be associated with causative relationships between mutation and disease. As an aid in developing these methods further, we have derived a set of random mutations of the enzymatic domains of human cystathionine beta synthase. This enzyme is a dimeric protein that catalyzes the condensation of serine and homocysteine to produce cystathionine. Yeast missing this enzyme cannot grow on medium lacking a source of cysteine, while transfection of functional human CBS into yeast strains missing endogenous enzyme can successfully complement for the missing gene. We used PCR mutagenesis with error-prone Taq polymerase to produce 948 colonies, and compared cell growth in the presence or absence of a cysteine source as a measure of CBS function. We were able to infer the phenotypes of 204 single-site mutants, 79 of them deleterious and 125 neutral. This set was used to test the accuracy of six publicly available prediction methods for phenotype prediction of missense mutations: SIFT, PolyPhen, PMut, SNPs3D, PhD-SNP, and nsSNPAnalyzer. The top methods are PolyPhen, SIFT, and nsSNPAnalyzer, which have similar performance. Using kernel discriminant functions, we found that the difference in position-specific scoring matrix values is more predictive than the wild-type PSSM score alone, and that the relative surface area in the biologically relevant complex is more predictive than that of the monomeric proteins.

Keywords

mutations; phenotype prediction; cystathionine beta synthase

INTRODUCTION

A major goal of the Human Genome and Hapmap projects was to identify DNA sequence variations among individuals and populations^{1,2}. Although some genome sequence variations in noncoding regions may alter the expression levels of genes, many variations in

the coding region will alter the amino acid sequence of expressed proteins. Many mutations in coding regions are missense mutations that alter a single amino acid in an expressed protein. Some missense mutations have been associated with disease states, including cancer^{3–8}, heart disease,⁹ diabetes,^{10–12} cystic fibrosis,¹³ and Alzheimer's disease^{14–16}. The Human Gene Mutation Database presently includes 38,066 missense/nonsense mutations in coding regions that are associated with disease.^{17,18} There are also many disease and gene-specific mutation databases that contain information linking specific mutations with disease phenotypes. These databases contain many variants not present in the Ensembl¹⁹ or dbSNP databases,²⁰ because the disease-associated variants are often rare and not likely to be encountered in general sequencing projects.

Establishing inherited risk of a disease based on a known sequence variation requires two steps: (1) evidence of epidemiological association of the genetic difference with increased susceptibility of the disease and (2) a plausible biological basis that the proposed genetic difference plays a role in the disease process. Without a plausible biological basis, it is always possible that the genetic association is caused by linkage to nearby genetic differences and not the variation under study. Data from the Human Genome Project and other sources have provided an enormous number of genetic differences and increasing amounts of haplotype data.^{21,22} Although a given haplotype may be associated with disease in population studies, it is not always evident which polymorphism in the haplotype is responsible for the association. It is, therefore, useful to develop methods for identifying those missense mutations in human genes that are likely to have adverse consequences in terms of protein function and, therefore, to be potentially linked to disease.

There are several ways a missense mutation may affect protein function and lead to disease. The first is that a mutation may alter stability of a folded domain. A mutation of a residue in the hydrophobic core to a charged one may lead to unfolding and proteolytic degradation. The second is that a mutation may be in or near a site on the protein surface involved in interactions with other molecules. Such a mutation may affect only the mutated side chain or may affect the structure of the backbone, in either case significantly changing the surface presented to other molecules. Some changes may affect catalytically active residues, modifiable residues such as phosphorylation sites, and others may change the electrostatic or hydrophobic properties of the binding surface, thus affecting binding affinities for partner molecules. Mutations in interfaces between domains of a single protein chain may also disrupt interaction of the domains, altering noncontiguous binding sites or leading to instability. Unusual mutations may have long range or allosteric effects on protein folding, structure, or dynamics, even though they are not in a binding site or buried.

There are a number of proteins that have been studied exhaustively to understand the effects of mutations. Most of these studies are limited to measurements on the stability of a folded protein, and not on functional interactions of that protein with other molecules. Such systems include T4 lysozyme,^{23,24} staphylococcal nuclease,^{25–28} HIV-1 protease,²⁹ and barnase.³⁰ Studies on the *Escherichia coli* Lac repressor³¹ are particularly relevant to the study of missense mutations and disease. Early studies on this protein described the effects of buried mutations and of those that disrupted dimerization or DNA binding.³² In later studies, Miller and coworkers mutated 328 positions within Lac repressor to one of 12 or 13 amino acids to produce a set of 4039 mutations.^{31,33–35} The assays used were binding of a Lac repressor tetramer to DNA and consequently repression of β -galactosidase expression, and derepression of transcription of the β -galactosidase gene upon induction by isopropyl- β -D-1-thiogalactoside (IPTG). With the crystal structures of the Lac repressor dimer bound to DNA and to IPTG,^{36,37} mutations which had a negative effect on function were identified as confined mostly to buried sites in the monomer or inside or near the interface between monomers or the interface with DNA and IPTG.³⁵

To predict the association among SNPs, protein function, and disease, a number of computational methods have been developed. There are some popular computational methods that are publicly available and may be used to predict whether a missense mutation is likely to be neutral or deleterious, such as SIFT,^{38–40} PolyPhen,⁴¹ SNPs3D,^{42,43} PMut,^{44–46} PhD-SNP,⁴⁷ and nsSNPAnalyzer.^{48,49} These methods are trained on existing sets of mutation/phenotype association data and use sequence information from homologues, structural information, such as accessible surface area, and changes in amino acid properties to provide feature information as input to machine learning methods for phenotype prediction.

In this article, we present a new data set of mutations that have been functionally characterized *in vivo* in yeast cells. The system used here is a dimeric form of the enzyme, human cystathionine- β -synthase (CBS) and expression of this protein in a yeast strain lacking the endogenous CBS gene, *CYS4*. The protein encoded by this gene is involved in the transsulfuration pathway, catalyzing the transformation of homocysteine to cystathionine in a pyridoxal phosphate (vitamin B₆) dependent reaction.⁵⁰ CBS deficiency or classic homocystinuria is caused by mutations in the human *CBS* gene and is the most common inherited disorder of sulfur metabolism.⁵¹ Untreated CBS-deficient patients have extreme elevations in total plasma homocysteine (tHcy) and plasma methionine levels and suffer from a variety of clinical problems including thrombosis, osteoporosis, mental retardation, psychiatric disorders, and dislocated lenses.⁵²

Second, we use this new data set to perform a blind test of the six publicly available phenotype prediction methods mentioned above. It is difficult to compare the accuracy of these publicly available methods based on their published results, since the testing sets differ substantially and the distinction between training and testing sets is not always clear. The methods vary significantly on the features used as input and the computational method used for predictions. Therefore, we believe that this test is useful in assessing these methods. The test is blind because the data have not been available for training of any of these published methods. Finally, we also compare some of the commonly used sequence and structural features of the deleterious and neutral mutations from the Lac repressor data set often used to train and test computational phenotype prediction methods. We use kernel density estimation and kernel discriminant functions to produce informative data on the usefulness of various features used to predict phenotypes.

METHODS

Mutagenesis and sequencing

Random mutagenesis was performed on pHBS1-409 expressing truncated (1-409aa) human CBS, which was created using gap repair in *Saccharomyces cerevisiae* as previously described.⁵³ Before mutagenesis, a single clone was isolated in *E. coli* and the CBS ORF was completely sequenced to assure there were no starting mutations. For PCR mutagenesis, pHBS1-409 was diluted to 0.1 ng/ μ l and the ORF was amplified by PCR using primers WK89 (5'-ATGCCTTCTGAGACCCCCAGGCAG-3') and WK90 (5'-CAGGAAAGCGA-AGGAGAAGTGGGCA-3') and Taq polymerase, which is known to produce a high error rate.⁵⁴ The resulting product was then mixed with gapped pHBS Δ expression plasmid and transformed into yeast strain WY35 (*α leu2 ura3 ade2 trp1 cys4::LEU2*) using lithium acetate as previously described.⁵⁴ Colonies were selected on synthetic complete media lacking tryptophan and supplemented with 30 mg/L glutathione (SC-trp+gl). A total of 948 colonies were then picked and grown in 96-well microtiter plates containing 200 μ L SC-trp+gl media. All subsequent steps were performed in microtiter wells using multichannel pipettors to minimize the chance of “bookkeeping” errors.

For mutation determination, 40 μ L of cells were first combined with 20 μ L of 20 mM NaOH and were heated at 100°C for 15 min, followed by 30 min at 4°C. Two μ L of the resulting solution were then used as templates in two different PCR reactions. The first reaction used primers WK89 and P4 (5'CGTTGCTCTTGAACCACTTG), which amplifies the 5'-end of the CBS ORF (nt 1-982), whereas the second reaction uses primers WK90 and P5 to amplify the 3'-end (nt 762-1719). These PCR products were then sequenced using primers P5, P4, and 5'R (5'CGTTGAAGAACTCACACTTG3'). Sequencing was performed by Agencourt Biosciences (Beverly, MA).

Mutations were detected by aligning the sequencing trace files (.abi) with wild type CBS sequence using Sequencher 4.5 software (GeneCodes, Ann Arbor, MI).

Growth experiments

Three experiments were performed to assess colony growth. In the first experiment, pin transfer was used to inoculate two different plates, one containing SC-trp+gl (Cys+) media and the other containing SC-trp-gl (Cys-) media from each of the 10 96-well master plates. The total amount of media in each well was 200 μ L. After 48 h of growth at 30°, OD₆₀₀ was determined for each clone using a plate reader. In the second experiment, a multichannel pipetter was used to inoculate 5 μ L of sample from the master plates, to Cys+ and Cys- plates. Growth was again measured after 48 h. The second experiment was performed about 6 months after the first and in between, the master plates were kept at -70°C. In the third experiment, we tested 83 colonies with mutations that grew at substantially different rates in the first and second experiments. In this case, both pin and micropipette experiments were performed. With the pin transfers, the OD₆₀₀ was measured at 48 and 72 h, whereas the micropipette samples were measured after 72 h.

Mutation data set

The sequence data included clones with missense mutations, silent mutations, nonsense mutations, and multiple mutations. We processed the data to obtain the largest possible set of functional data on mutations. The first and second experiments produced substantially different growth rates for about 20% of the clones, generally much lower growth in the first experiment, which used pin transfer than in the second experiment, which used micropipette transfer. For those discrepant clones, we retested them a third time and found that the repeated analysis was identical to the second experiment. These findings suggest that the small inoculum volumes in the first experiment were likely a source of error. On the basis of these findings, we decided to use the maximal growth ratio of the first 2 experiments to determine a phenotype.

For clones with single mutations, we determined their phenotype according to maximal ratio of growth on Cys- to Cys+ from the first 2 experiments. However, a large number of clones with two, three, and four mutations were also obtained, and so we defined a number of rules for defining the assumed phenotype of single mutations within these data. We used the following procedure:

1. If a single mutation was observed by itself in a number of clones and the phenotype was sometimes neutral and sometimes deleterious, then we defined the mutation's phenotype as neutral. The rationale is that some clones may have an error in noncoding regions that affect transcription rather than the protein sequence. There were only three such mutations: Q222R, N225S, and Q242R.
2. For double mutations, if the phenotype of one of the mutations was neutral when found in a single-mutation clone, then the phenotype of the second mutation was equated to the phenotype of the double-mutation clone. Otherwise, if neither

mutation existed in a single-mutation clone, then the phenotypes of both were considered undetermined.

3. For triple mutations, if the triple phenotype was deleterious and two of them were not deleterious from single and double mutations (Steps 1 and 2), then the phenotype of the third was considered deleterious. If one of them was already defined as a deleterious mutation, then we could not determine the phenotypes of the remaining mutations.
4. For quadruple mutations, if the quadruple phenotype is deleterious and we know three of them were not deleterious mutations (Steps 1–3), then the phenotype of the fourth one was defined as deleterious. If we know one of them was deleterious, then we could not determine the phenotypes of the other mutations.

Publicly available servers

We used these mutations as a new data set to test several publicly available methods for predicting mutation phenotypes, using such measures as true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), overall predictive accuracy (ACC), Matthew's correlation coefficient (MCC),^{55,56} and the balanced error rate (BER) (defined in the Results section). For each method, certain parameters needed to be set which are described here.

SIFT—The SIFT server is available at http://sift.jcvi.org/www/SIFT_seq_submit2.html and required an input FASTA sequence (the full CBS sequence, residues 1–551) and a list of mutations. SIFT searches one of three databases to find homologous sequences, and we used the NCBI non-redundant set (Dec. 2008). It requires a median conservation score, and recommends a value of between 2.75 and 3.25, and we therefore used 3.00. It also requires a sequence cutoff for removing redundant sequences among the homologues, and we used a value of 90%. SIFT outputs a normalized probability for each amino acid type, and a value of less than 0.05 is considered deleterious; otherwise the mutation is considered neutral.

PolyPhen—The PolyPhen server is available at <http://genetics.bwh.harvard.edu/pph/>. For PolyPhen, we input the full CBS sequence and each CBS mutation manually. We used the default settings for each parameter or switch required by PolyPhen: structure database = PQS; measure used to sort hits = “identity”; whether to map to mismatch = “NO”; calculate structure parameters = “for first hit only”; calculate contacts = “for all hits”; minimal alignment length = 100; minimal identity in alignment = 0.5; maximal gap length in alignment = 20; threshold for contact = 6 Å. PolyPhen provide four predictions: benign, possibly damaging, probably damaging, and unknown. If the prediction was benign, we assigned a neutral phenotype to the prediction. If the prediction was probably damaging, we assigned a deleterious phenotype to the prediction. We calculated results separately for assigning “possibly damaging” mutations to either deleterious or neutral phenotypes.

PMut—The PMut server is available at <http://mmb2.pcb.ub.es:8080/PMut/>. We provided PMut with the full CBS sequence and the list of 204 mutations. PMut has two neural networks for phenotype prediction: a “small NN” with no hidden layer and a second neural network with a hidden layer. PMut suggests that the second neural network performs better for human proteins, so this network was used for CBS. We used both networks for Lac repressor and found the small NN had higher performance for Lac, whereas the larger network had better performance for CBS. We used a prediction value > 0.5 to assign deleterious predictions to mutations.

SNPs3D—The SNPs3D server is available at <http://www.snps3d.org/>. SNPs3D required the NCBI gene id and manual input of the mutations. SNPs3D has both a “profile model” support vector machine (SVM) and a “stability model” SVM that uses structural information. Only the profile model can be run online for input mutations, and the results here are for that model. The output from the SNPs3D profile SVM with predictions less than 0 were associated with deleterious mutations and predictions greater than or equal to 0 were associated with neutral mutations.

PhD-SNP—The PhD-SNP server is available at <http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi>. PhD-SNP required the input CBS sequence and each mutation was entered manually. We used the “sequence and profile based mode” with 20-fold cross validation, because the authors state that this is their best prediction method. The output “Disease mutation” were marked deleterious and “Neutral mutation” as neutral.

nsSNPAnalyzer—The nsSNPAnalyzer server is available at <http://snpanalyzer.utm.edu/>. nsSNPAnalyzer required the input CBS sequence and the list of 204 mutations. The program is based on a machine learning method referred to as a random forest⁵⁷ and was trained on a curated SNP dataset prepared from the SwissProt database. nsSNPAnalyzer calculates three types of information: (1) the structural environment of the SNP, including the solvent accessibility, environmental polarity and secondary structure; (2) the normalized probability of the substitution in a multiple sequence alignment; (3) the similarity and dissimilarity between the original amino acid and mutated amino acid. “Disease mutation” predictions were marked deleterious and “Neutral mutation” as neutral.

Feature calculation

We calculated several sequence and structural features for both the CBS data set and also the Lac repressor data set in order to study their predictive properties. Position-specific scoring matrices (PSSMs) were calculated from PSI-BLAST⁵⁸ searches of the nonredundant protein sequence database from NCBI,⁵⁹ and the PSSM values for both wild-type and mutant were taken from the third round of search.

Surface areas for residues in the monomer and complexes were calculated with the program *naccess*.⁶⁰

Lac mutations were obtained from Mitchell Lewis at the University of Pennsylvania (personal communication). These mutants had phenotype levels of 0, 1, 3, and 5. Mutants scored 0 or 1 were considered deleterious, and those score 3 or 5 were considered neutral.

Kernel density estimation

To explore the performance of various sequence and structural features used to label mutations, we used kernel density estimation to calculate a probability density function of each feature. The nonparametric kernel density estimate using a Gaussian kernel for any continuous variable may be written⁶¹:

$$p(x) = \frac{1}{nh \sqrt{2\pi}} \sum_{i=1,n} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right)$$

where h is a bandwidth parameter used to control the smoothing of the density estimate for n data points. After calculating kernel density estimates for both the neutral (N) and deleterious (D) mutations, Bayes' rule is used to calculate kernel discriminant functions:

$$p(\text{ph}|x) = \frac{p(x|\text{ph})p(\text{ph})}{p(x|N)p(N) + p(x|D)p(D)}$$

for phenotype ph equal to either N or D .

To optimize the bandwidth h , we used a maximum likelihood estimation procedure to maximize the probabilities of the phenotypes of the large Lac repressor data set given the feature values using a cross validation procedure.⁶² As a function of h , we maximized the value of the log likelihood:

$$L(h) = \sum_{i \in \{N\}} \log p(\text{ph}_i|x_i) + \sum_{j \in \{D\}} \log p(\text{ph}_j|x_j)$$

where the sums are taken over the set of neutral mutations, $\{N\}$ and the set of deleterious mutations, $\{D\}$. The same value of h was used for both the N and D density estimations.

The Lac repressor data are unbalanced because the set contains 1043 deleterious mutation and 2996 neutral mutations. To obtain an unbiased result, we chose a random selection of an equal number of neutral and deleterious mutations from the Lac set (1043 each). For the balanced data, we used 10-fold cross validation to select the optimal bandwidth for each feature. The bandwidths calculated from the Lac repressor data set were also used on the much smaller CBS data set.

RESULTS

Distribution of mutations

For mutagenesis, we used a truncated version of human CBS that lacks the 142 amino acid C-terminal regulatory domain which is responsible for regulation of CBS by S-adenosylmethionine. This truncated form of CBS is highly active, forms a dimer, and no longer binds SAM.^{63,64} The crystal structures of truncated forms of CBS have been determined independently by two different groups, residues 44–406 (47–403 visible in the density)⁶⁵ and a longer construct, residues 1–413 (43–397 visible in the density), that includes bound heme.⁶⁶

To identify the functional consequences of missense mutations in human CBS, we used a functional assay system in *S. cerevisiae*.⁵⁴ The basis for this assay is that wild-type human CBS can functionally correct the cysteine auxotrophy associated with deletion of the endogenous yeast CBS gene, *CYS4*. This method has been used to assess the functional effects of over 20 patient-derived alleles,^{54,67–69} where lack of growth complementation was shown to be associated with significantly impaired enzymatic activity.

We created a novel missense mutation data set in a truncated version of the human CBS enzyme using a combination of error-prone PCR and gap repair in the CBS tester strain WY35. A total of 948 colonies were sequenced and analyzed for function. Sequence analysis revealed that 591 (62.3%) of the colonies had at least one mutation that altered the amino acid sequence of the CBS protein. The vast majority of the mutations were of the

missense variety (99%), with only 10 of the colonies (1%) expressing either frame-shifted or prematurely terminated proteins. The distribution of mutations is shown in Table I.

We calculated the 12 possible nucleotide mutation rates over the 948 clones that were sequenced. The results are listed in Table II. These rates vary from 5×10^{-6} to 840×10^{-6} mutations per base pair, with A $\rightarrow$ G and T $\rightarrow$ C mutations much more frequent than others. Therefore, the collection is not evenly distributed among all possible mutations. The *E. coli* Klenow fragment and some mutants also show large differences in fidelity amongst the possible nucleotide substitutions⁷⁰ as does the Taq polymerase.⁷¹ One unexpected feature of the sequence analysis was the high frequency of the 55T $\Rightarrow$ C (S19P) mutations. We found that this mutation was present in 326 (34.4%) of the colonies, either alone (16.9%) or in combination with additional mutations (17.5%). No other mutation was observed more than four times. Because we had sequenced the starting PCR template material prior to the experiment and found it to be entirely wild type, these findings indicate that this particular nucleotide must be an extreme hot-spot for Taq polymerase.

We examined the growth for all 948 colonies in cysteine supplemented (Cys+, in the form of glutathione) and cysteine-free media (Cys-) and determined the Cys-/Cys+ growth ratios. All clones growth were assessed twice for relative growth in Cys-/Cys+ media. In the initial experiment, a very small amount of inoculum was applied to each well using pin transfer, whereas in the second experiment a larger inoculum was applied using a multichannel pipetter (see methods). About 80% of the colonies showed similar behavior in both assays, whereas about 20% of the colonies had discordant growth behavior. In nearly all cases, the Cys-/Cys+ growth ratio was substantially lower in the first experiment than in the second experiment. We suspected that this could possibly be due to the very small inoculum that occurs with pin transfer. We, therefore, performed a third experiment on all 83 discrepant colonies containing missense mutations, using both pin transfer and micropipetter transfer. In every case, at 48 h, the discrepant colonies exhibited significant growth in the pin transfer experiment and for all colonies the growth was similar to wild-type at 72 h in both pin and micropipette transfer experiments (Supporting Information). On the basis of this finding, we decided to use the maximum growth ratio of the first two experiments as the measured phenotype.

As shown in Figure 1(a), the distribution of the growth ratios for the 948 colonies is clearly bimodal with colonies with functionally impaired CBS clustering on the left (<0.4) and those with active CBS clustering on the right (>0.9) according to the maximal Cys-/Cys+ ratio of the two experiments. We then subdivided the colonies into three groups; colonies that contained no detectable mutation [Fig. 1(b)], colonies that contained only S19P [Fig. 1(c)], and 330 colonies that contained non-S19P missense mutations [Fig. 1(d)]. In the colonies with no detectable mutations and in the S19P colonies, we observed that only 4–5% of the colonies fell in the no growth range (growth ratio <0.2), and that 91–94% of the colonies fell in the high growth range (growth ratio >0.7). These results make two important points. First, the S19P mutation appears not to affect CBS function. Second, there appears to be a false no-growth rate of the assay between 4% and 5%. The basis for this rate is unclear, but it may involve plasmid rearrangements outside the sequenced region due to the gap-repair/transformation process.

As shown in Figure 1(d), the distribution of the ratios for colonies with missense mutations is clearly bimodal with colonies with functionally impaired CBS clustering on the left (<0.4) and those with active CBS clustering on the right (>0.7). A minimum in the histogram occurs between 0.3 and 0.4, so we chose 0.4 as the cutoff to assign phenotypes to the clones; that is, if the maximal value of the two Cys-/Cys+ ratios is 0.4 or less, we considered the phenotype deleterious, and those above 0.4 were considered neutral.

To produce a testing set for computational phenotype prediction, we evaluated all clones in the data set for their functional significance using the heuristic rules described in the Methods section. Using these rules, we were able to determine the functional effect of 220 unique CBS missense mutations, of which 83 were deleterious and 137 were neutral. To guarantee that each mutation had available information for structure-based properties, we deleted 16 mutations since the wild-type residues at those positions have no structural information in either of the available crystal structures of CBS. The resulting data set consists of 204 mutations, of which 79 are deleterious and 125 are neutral.

Accuracy of publicly available methods as published

The publicly available methods train on experimental data sets and use methods from simple statistics to machine learning to derive methods for predicting the phenotypes of mutations. The authors of these methods have used a number of ways to assess their accuracy, and so we first decided to calculate a consistent set of accuracy measures for all of them based on their published results on the various data sets they used. From the original publications, we were able to obtain (or calculate) the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for the testing sets used in their papers. From these we calculated the true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV). For the experimental data, we define the total number of experimentally determined positives (deleterious mutations) as $P_{\text{exp}} = \text{TP} + \text{FN}$ and the total number of experimental negatives (neutral mutations) as $N_{\text{exp}} = \text{TN} + \text{FP}$. Similarly for the predictions, we define the total number of predicted positives as $P_{\text{pred}} = \text{TP} + \text{FP}$ and the total number of predicted negatives as $N_{\text{pred}} = \text{TN} + \text{FN}$. With these values, we calculate:

$$\text{TPR} = \frac{\text{TP}}{P_{\text{exp}}}; \text{TNR} = \frac{\text{TN}}{N_{\text{exp}}}; \text{PPV} = \frac{\text{TP}}{P_{\text{pred}}}; \text{NPV} = \frac{\text{TN}}{N_{\text{pred}}}$$

In words, the TPR is the proportion of experimental positives that are correctly predicted; the TNR is the proportion of experimental negatives that are predicted correctly; PPV is the proportion of predicted positives that are predicted correctly; and NPV is the proportion of predicted negatives that are predicted correctly. Note that some publications provide the false positive rate (FPR) and false negative rate, which are defined:

$$\text{FPR} = \frac{\text{FP}}{N_{\text{exp}}} = 1 - \text{TNR}; \text{FNR} = \frac{\text{FN}}{P_{\text{exp}}} = 1 - \text{TPR}$$

We also calculate three measures that assess overall accuracy: total accuracy (ACC), Matthews correlation coefficient (MCC),⁵⁵ and balanced error rate (BER):

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{P_{\text{exp}} + N_{\text{exp}}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{P_{\text{exp}} N_{\text{exp}} P_{\text{pred}} N_{\text{pred}}}}$$

$$\text{BER} = \frac{1}{2}(\text{FNR} + \text{FPR})$$

The results on the published data are shown in Table III and discussed further below.

SIFT^{38,40} uses only sequence information by performing PSI-BLAST⁵⁸ searches to produce multiple alignments of close homologues of the query sequence and calculating position-specific probabilities of the 20 amino acids using pseudocounts from a 13-component Dirichlet mixture derived by Sjölander *et al.*⁷² Table III lists published results on SIFT's ability to predict phenotypes of some of the large mutation data sets available (Lac repressor, HIV protease, lysozyme) published by Ng and Henikoff.⁴⁰

The PolyPhen web server developed by Bork and coworkers^{41,73,74} uses both sequence and structural information to predict whether mutations are deleterious. BLAST is used to make a multiple sequence alignment from homologous proteins (30–94% sequence identity with the query), and PSIC⁷⁵ is used to construct a position-specific scoring matrix (PSSM) from this alignment. PolyPhen uses BLAST to identify proteins of known structure homologous to the query, but restricts its results to sequences in the PDB that are 50% identical to the query, and cases for which the amino acid in the structure is the same as the wild-type amino acid under study. PolyPhen uses the sequence alignment and the known structure to determine residue accessibility and proximity to ligands and interfaces with other subunits in the structure. Location of mutations in protein interfaces is determined from both the PDB structure (asymmetric units) and the hypothetical biological units given in PQS.⁷⁶ The training and testing data were obtained from HGVB⁷⁷ and the results on this set provided in Sunyaev *et al.*⁷³ are listed in Table III. Subsequently, the PolyPhen server was published⁴¹ with a posited TPR of 82% on disease mutations in SwissProt database 14 and a TNR of 92% on inter-species variations, although insufficient details are given to calculate the parameters in Table III.

Yue and Moul^{42,43} developed two methods in SNPs3D for identifying which nonsynonymous single mutations have a deleterious effect on protein function *in vivo* using support vector machines (SVM). The first method, the Stability model,⁴² uses experimental or predicted protein structures to estimate deleterious effects of mutations on the stability of folded proteins. They extracted 10,263 disease-causing mutations in 731 proteins from HGMD¹⁸ and the same 731 monogenic disease proteins were used to obtain 16,682 mutations not related to disease (control set). Appropriate structure information was available for only 37% mutations in the disease set and 14% mutations in the control set, which were used as the training and testing data for stability model. A set of factors were used to approximate the stability effect of a mutation on a protein structure, including electrostatic interactions, over-packing and cavities, hydrophobic burial, surface accessibility, structural rigidity from crystallographic B-factors, backbone strain, buried charged or polar residues, and breakage of disulfide bonds. If the query proteins have no 3D structure, SNPs3D uses homology modeling to model the query proteins' 3D structures. Homology modeling was performed with the APSE (automatic protein structure emulator) pipeline.⁴² Alignments were obtained using CLUSTALW.⁷⁸ Coordinates of side-chains conserved between the human protein and the PDB template are copied. Remaining side-chains are added using SCWRL.⁷⁹ Ten-fold cross validation was used to train the SVM. The final TPR and TNR rate for the stability model were equal to 74% and 85%, respectively. More details are provided in Table III.

The second model (the profile model)⁴³ was based on analysis of sequence families related to human proteins. Five features were used to train the SVM: (1) a PSSM score for the

mutant residue; (2) the entropy at each position in the alignment, (3) the mean entropy over the sequence, (4) the standard deviation of the entropy over all positions, and (5) the entropy at each position of the alignment expressed as a Z score. The level of conservation at a mutation position and the probability of observing such a variation within the family successfully distinguished between disease mutations and nondeleterious mutations, with a TPR and TNR of 80% and 91%, as listed in Table III.

PMut uses sequence and structural information to train a neural network model to predict phenotypes. A set of disease-associated mutations was obtained from SwissProt by searching the keywords DISEASE, VARIANT, and HUMAN80 (9334 disease-related mutations in 811 human proteins; 1319 disease-related mutations in 90 proteins with structural information) of training and testing data set. For a neutral training and testing data set consisting of 11,372 mutations (888 with structural information), the neutral mutations from the Lac repressor were used as well as a second set from variants occurring in members of the protein family from other species of the 811 proteins of the disease-associated set. PMut uses different kinds of sequence information to label mutations, such as scoring mutation matrices (PAM4081 and BLOSUM6282), volume parameters, secondary structure properties, hydrophobicity descriptors, sequence potential, and surface accessibility. Table III shows the predictive results of PMut on different testing data sets.

PhD-SNP is based on sequence and profile information. The authors used three datasets obtained from SwissProt80: “HumVar” (21,185 different single point mutations, of which 12,944 are disease-related and 8241 are neutral polymorphisms), “HumVarProf” (8718 different single point mutations, of which 3852 are disease-related and 4866 are neutral polymorphisms), and “NewHumVar” (935 single point mutations, of which 149 are disease-related and 786 are neutral polymorphisms). The authors trained three models: One is based on sequence information and a support vector machine (SVM-Sequence), one is based on profile information and a support vector machine (SVM-Profile), and the other is a decision tree that uses both SVMs (PhD-SNP). Table III gives the results of PhD-SNP on the different testing datasets.

nsSNPAnalyzer use structural and evolutionary information and the SIFT score as the features of the mutations. It uses human nsSNPs which were extracted via analysis of the VARIANT field in the corresponding Swiss-Prot entries. The authors searched against the ASTRAL database⁸³ to find the homologous structures, so the final data set contains 532 neutral and 3686 deleterious mutations. This data set was split into training data (502 neutral and 3511 deleterious mutations) and testing data (30 neutral and 175 deleterious mutations). These are highly unbalanced data sets in favor of deleterious mutations. Table III gives the results of nsSNPAnalyzer on the training data set and testing data set, showing that the method may be undertrained in predicting neutral mutations (NPV < 30).

Testing phenotype predictors on the CBS data

We used the 204 mutations of the CBS data set to test the public online programs SIFT, PolyPhen, SNPs3D, PMut, PhD-SNP and nsSNPAnalyzer. Five of these publicly available methods provide dichotomous predictions on whether mutations are “deleterious” (positive) or “neutral” (negative). PolyPhen uses empirically derived rules to predict that a mutation is “probably damaging,” “possibly damaging,” “benign,” or “unknown.” We calculated accuracy results in two ways: assuming Polyphen’s “possibly damaging” mutations are deleterious (positive) and assuming “possibly damaging” are neutral (negative). We also calculated a consensus prediction of these six methods, by excluding one at a time. If three or more of the remaining five methods predict the phenotype of a mutation is neutral or deleterious, then we predict the mutation’s phenotype is neutral or deleterious, respectively.

Table IV gives the results of these six public methods tested with CBS mutations along with the consensus predictions (lower half of Table IV).

The results of the blind test using the CBS mutations can be evaluated with different accuracy measures. The simplest is total accuracy over the 204 mutations. With this measure, PolyPhen is the most accurate at 71%, if we assume that Polyphen's "possibly damaging" mutations are neutral mutations. nsSNPAnalyzer, SIFT, and PMut come in at 66–67%, whereas SNPs3d and PhD-SNP have ACC scores of ~60%. However, the CBS mutation data set is somewhat unbalanced with 39% of the mutations deleterious and 61% neutral. Therefore, we evaluated the Matthews correlation coefficient (MCC) and the balanced error rate (BER) that try to account for the unbalance. With the MCC, SIFT, and PolyPhen come out on top, followed by nsSNPAnalyzer, and significantly ahead of the remaining methods. Lower BER is better than higher BER, and again SIFT, PolyPhen, and nsSNPAnalyzer show higher accuracy than PMut, SNPs3D, and PhD-SNP. PMut has the lowest TPR and SNPs3D has the lowest TNR, and therefore both methods have BERs significantly worse than SIFT, PolyPhen, and nsSNPAnalyzer. All the methods, except PMut and PolyPhen (poss. damaging = neutral), have false positive rates between 40% and 60%.

It is also of interest to compare the MCC and BER rates in Table IV with those in Table III, which contains results from the publications of the various methods tested. Because those data sets were also sometimes unbalanced, the MCC and BER comparisons are more reliable indicators of performance than total accuracy. Finally, the consensus predictions have TPR of 80–82% and TNR of 60–68%, a higher TNR than most of the individual methods. The consensus of five methods (excluding SNPs3D) results in the highest accuracy and Matthews' correlation coefficient at 0.478.

We examined nine mutations that were predicted incorrectly by all six methods. One of these was a deleterious mutation, K97R, predicted to be neutral by all six methods. There does not seem to be any obvious reason why K97R is deleterious. Eight mutations were neutral but predicted to be deleterious by all six methods: R121C, I127T, R164C, Q222R, L268Q, C275R, V319E, and Y381C. It is of interest that three of these mutations are to cysteine, and one is from cysteine. Because cysteine is a rare amino acid, it is likely that sequence-based methods are heavily weighted to consider mutations to or from cysteine as deleterious. Also, Cys often plays specific structural roles in disulfides or in ion binding, and thus are often heavily conserved. The methods may overemphasize the role of cysteines in a protein that does not have disulfides and only one cysteine (C52) involved in binding the heme iron.

In the CBS structure, R121 makes a salt bridge with E239 and is completely conserved, and thus one might expect a mutation to be deleterious. However, a cysteine in its place is probably solvent accessible. L268 is buried and surrounded by hydrophobic residues, and so one would expect that L268Q would be a deleterious mutation. Similarly, V319 makes hydrophobic contacts, although is close enough to the surface that glutamic acid at this position may be solvent accessible. It is a glutamic acid in a small number of CBS orthologues. A threonine in place of I127 seems feasible in the structure, even though the residue is always I or V in CBS orthologues.

It is important to test whether the results are significantly affected by the choice of threshold in separating deleterious and neutral mutations. In Figure 2, we plot the overall accuracy and the Matthews correlation coefficient for all six methods as a function of the deleterious/neutral threshold. We observe that the overall accuracy [Fig. 2(a)] and the Matthew's correlation coefficient [Fig. 2(b)] of the methods are flat when the threshold falls in the

range [0.3, 0.7] excluding PMut and SNPs3D, which change at a higher threshold. Figure 2(b) shows that the top three methods are all fairly similar and indeed, swap ranks depending on what the deleterious/neutral threshold is set to.

Properties of features used to predict phenotypes

To verify that the CBS data set is not unusual in some way, we compared some of the common properties used in prediction methods with those of the Lac repressor data set. This was accomplished through the use of non-parametric density estimation for such features as PSSM scores and accessible surface area. We first calculated the density of a feature for the deleterious mutations and the neutral mutations separately, that is, $p(x|ph = D)$ and $p(x|ph = N)$. With Bayes' rule, we can invert these probabilities to obtain the kernel discriminant functions $p(ph = D|x)$ and $p(ph = N|x)$. As an example, we show in Figure 3 the kernel density estimations and the kernel discriminant functions for both the neutral and deleterious mutations for the balanced Lac repressor data set using the PSSM scores of the wild-type residue types. The optimal kernel width for both functions was chosen to maximize the log likelihood of the discriminant functions using 10-fold cross validation (see Methods). Although the probability densities are only slightly shifted from each other, at least at the lower and higher end of the PSSM score range, the value of the feature is predictive of phenotype. The probabilities of neutral and deleterious phenotypes are equal at the PSSM scores of 4.5.

To compare the CBS mutation data set with the Lac repressor data set, in Figure 4, we plot several commonly used features in phenotype prediction: wild-type PSSM score, the difference between wild-type and mutant PSSM scores, surface area of the wild-type residue in the monomeric protein, and surface area of the wild-type residue in the biological unit (CBS from PDB entry 1M5465:84 and Lac from PDB entry 1EFA36:37). The data indicate that the CBS mutation set is fairly similar to the large Lac repressor data set, although a bit noisier due to its smaller size.

Both Lac and CBS are multimeric proteins. Lac acts as a tetramer, although only the dimer structure is available for a DNA bound form from an X-ray crystal structure (a recent structure contains a non-DNA-bound tetramer that might also be used to predict phenotypes⁸⁵). Full-length CBS is a tetramer, but the truncated form used here is a dimer, and only the structure of truncated dimeric CBS is available (i.e. missing the C-terminal regulatory "Bateman domain"⁸⁶). Because protein-protein interactions in the multimeric structures (and protein-DNA in the case of Lac) are important for function, we postulate that surface area measured in the biologically relevant assemblies is likely to be a better predictor of phenotype than surface area calculated from the monomeric proteins. This is shown in Figure 5, where we show the deleterious discriminant function for surface areas calculated from monomers and the relevant biological assemblies (dimeric CBS plus heme and PLP, dimeric Lac with bound IPTG and DNA). The results show that surface areas calculated on the biological assemblies improves the discriminant function for both systems, although the effect is somewhat larger for Lac than for CBS. The exact role of dimerization for CBS is unknown, whereas the Lac dimer binds to a symmetric binding site on DNA.

The kernel discriminant functions calculated here can be used predictively where for each mutation, the feature value x is used to predict that the mutation is either deleterious or neutral, depending on which value is larger: $p(ph = D|x)$ or $p(ph = N|x)$. For the features shown in Figure 4, several accuracy measures are given in Table V. For both systems, PSSM difference between wild-type and mutant residue is most predictive according to the overall accuracy, the MCC and the BER. It is also evident that the accessible surface areas in the biologically relevant complexes are more predictive than the surface areas of the monomeric proteins according to the same three measures, although this is more evident for

MCC and BER for CBS, because the CBS data set is somewhat unbalanced. This result emphasizes the need for accurate determination of biological assemblies from X-ray crystallographic structures. It is interesting to note that for Lac, accessible surface area in the biologically relevant complexes has the highest true positive rate and the highest negative predictive values. Although important for understanding the consequences of mutations, biological complexes are not always well determined and are not completely consistent from publicly available annotation sources, which should be used with caution.^{87,88}

Structural analysis of the mutations

We analyzed some structural features of the wild-type residues in the 204 mutations for which we have phenotypes. First, we classified the mutations into three groups using relative surface area (RSA) in the dimeric structure of CBS with bound heme and PLP: buried ($0\% \leq \text{RSA} < 15\%$), intermediate ($15\% \leq \text{RSA} < 30\%$) or exposed ($\text{RSA} \geq 30\%$). The percentages of these classes that are deleterious or neutral are given in Table VI. We also calculated hydrogen bonds of the wild-type side chains to both the backbone and other side chains, as well as the hydrophobic contacts of the wild-type side chains to determine if there were significant differences between deleterious and neutral mutations in each class (exposed, intermediate or neutral). These are given in the Supporting Information.

It is not surprising that mutations that are exposed in the dimer/ligand-bound structure are predominantly neutral at 82%. There does not seem to be any structural tendency among the deleterious exposed mutations compared to the neutral exposed mutations in terms of hydrogen bonds or favorable Van der Waals interactions. For instance, 4 of 14 exposed deleterious mutations make hydrogen bonds, whereas 20 of 62 exposed neutral mutations do so.

What is surprising is that buried mutations are about equally likely to be deleterious as neutral (52–48%). This does not change if a different cutoff is used to separate buried from intermediate mutations (at 15% RSA), because 52% of these mutations are neutral. There are not significantly more hydrophobic interactions of the deleterious mutations (35 of 50 with contacts) than by the neutral mutations (33 of 47 with contacts). Buried deleterious mutation side chains are not much more likely to be hydrogen bonding (19 of 50) than those of neutral mutations (14 of 47).

One can view the mutations in a different way by sorting the mutations the wild-type and/or the mutant residue types. For instance, 7 of 9 mutations from Leu to Pro are deleterious. Mutations to proline will eliminate any hydrogen bond the Leu backbone *N* may have. This occurs for all seven of the deleterious mutations and neither of the neutral mutations. Also, proline has a very limited range of the backbone dihedral angle ϕ from -80° to -40° . It also severely restricts the backbone conformation of the previous residue in the structure, strongly suppressing conformations in the α_P and α_L regions of the Ramachandran map.⁸⁹ One of the deleterious Leu to Pro mutations (L251P) has a bad ϕ dihedral angle of -140° . The other six deleterious Leu to Pro mutations all have left neighbors in either α_P (L81P, L124P, L136P, L230P, L241P) or α_L conformations (L101P). Neither of the two neutral Leu to Pro mutations has either of these problems.

DISCUSSION

A number of methods for prediction of neutral or deleterious phenotypes for missense mutations have been developed based on machine learning and other computational techniques. These methods have been trained on existing mutation data sets, such as those for the Lac repressor, HIV protease, and lysozyme, as well as large data sets of mutations associated with disease, such as HGMD, and presumed neutral mutations found in closely

related orthologous proteins. These methods have been assessed by their developers using various measures and on quite different test sets. An independent assessment in the spirit of the CASP90 and CAPRI91 conferences therefore seems in order.

In this paper, we have presented a set of 204 mutations of the human enzyme cystathionine beta synthase. The assay used is the complementation of the same enzyme deleted from the yeast strain WY35, and the comparison of growth on Cys+ and Cys- media. Therefore, the assay is a functional test of the protein. It not only has to fold properly and dimerize but it must also bind its substrates and cofactors in order to function. Any mutation that disrupts folding kinetics, protein stability, or intermolecular associations necessary for function would be expected to result in low growth or no growth on medium lacking a source of cysteine (in this case, glutathione).

Conversely, a mutation of a residue that does not have significant interactions with other molecules would be expected to be neutral, although some such mutations may affect folding or long-range electrostatics that may be important for function. However, we found that for CBS, while mutations on the surface of the biologically relevant complex (dimer plus cofactors) were largely neutral in phenotype, buried mutations were about equally split between deleterious and neutral. This is likely to be due to the tendency of the genetic code to preserve the physical characteristics of amino acids that arise from single nucleotide substitutions. If we define conservative mutations as those that differ in volume by at most one heavy atom and also maintain the polar or hydrophobic nature of the residue type (with the exception Ile-Thr mutations, which are both β -branched and commonly substitute for one another), then of the 47 buried neutral mutations, 20 (43%) are conservative mutations (Val-Ile, Lys-Arg, Asn-Asp, Ser-Thr, etc.), whereas only 10 of the 50 (20%) buried deleterious mutations are conservative. Also, of the 14 exposed deleterious mutations, only one is conservative (7%) using the same definitions.

We have compared six publicly available methods for their accuracy, and found that the SIFT and PolyPhen servers are the top predictors, depending on which measure is used, with nsSNPAnalyzer not far behind. Except for PMut, the true positive rates (deleterious mutations predicted as such) are in the range of 70–90%, whereas the true negative rate (neutral mutations predicted as such) are much lower at ~ 50–70%. Perhaps what is more relevant to users of such methods are the positive predictive value and negative predictive value. These are the proportions of positive or negative predictions that are correct respectively. For PolyPhen, for instance, these values are 60% and 80%, whereas the random PPV and NPV are about 40% and 60% (the proportions of positives and negatives in the CBS data set). Thus, for this data set, the prediction methods perform better than expected randomly, although not as high as the published results, shown in Table III. The results are specific for CBS, and it is possible that the methods behave better or worse relative to each other on other kinds of proteins.

It is interesting that SIFT, which does not use structural information, predicts at nearly the same rates as Polyphen and better than nsSNPAnalyzer, both of which do use structural information. It seems likely that there is room for improvement in the use of structural information in phenotype prediction, particular in using the structures of biologically relevant complexes, including protein–protein and protein–nucleic acid interactions. Modeling of mutations using side-chain prediction methods,^{79–92} and examining structural features of the mutated side chains (hydrogen bonds, steric overlaps, etc.) may also be quite helpful, as suggested by Yue and Moulton.⁴² None of the methods we were able to test performed this kind of analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant sponsor: NIH; Grant numbers: R01 GM73784, R01 HL057299

REFERENCES

1. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New goals for the U.S. human genome project: 1998–2003. *Science*. 1998; 282:682–689. [PubMed: 9784121]
2. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med*. 2009; 60:443–456. [PubMed: 19630580]
3. Carlson KM, Dou S, Chi D, Scavarda N, Toshima K, Jackson CE, Wells SA Jr, Goodfellow PJ, Donis-Keller H. Single missense mutation in the tyrosine kinase catalytic domain of the RET protooncogene is associated with multiple endocrine neoplasia type 2B. *Proc Natl Acad Sci USA*. 1994; 91:1579–1583. [PubMed: 7906417]
4. Garrigue-Antar L, Munoz-Antonia T, Antonia SJ, Gesmonde J, Vellucci VF, Reiss M. Missense mutations of the transforming growth factor beta type II receptor in human head and neck squamous carcinoma cells. *Cancer Res*. 1995; 55:3982–3987. [PubMed: 7664267]
5. Marchese CA, Bertolino F, Ceccopieri B, Vanzetti M, Scaglione D, Locatelli L, Montera M, Romio L, Resta N, Stella A, Guanti G, Mareni C. Clinical findings in a family with familial adenomatous polyposis and a missense mutation of the adenomatous polyposis coli gene. *Scand J Gastroenterol*. 1996; 31:917–920. [PubMed: 8888441]
6. Wang Y, Friedl W, Lamberti C, Ruelfs C, Kruse R, Propping P. Hereditary nonpolyposis colorectal cancer: causative role of a germline missense mutation in the hMLH1 gene confirmed by the independent occurrence of the same somatic mutation in tumour tissue. *Hum Genet*. 1997; 100:362–364. [PubMed: 9272156]
7. Shen D, Wu Y, Chillar R, Vadgama JV. Missense alterations of BRCA1 gene detected in diverse cancer patients. *Anticancer Res*. 2000; 20:1129–1132. [PubMed: 10810408]
8. Puxeddu E, Moretti S, Elisei R, Romei C, Pascucci R, Martinelli M, Marino C, Avenia N, Rossi ED, Fadda G, Cavaliere A, Ribacchi R, Falorni A, Pontecorvi A, Pacini F, Pinchera A, Santeusano F. BRAF(V599E) mutation is the leading genetic event in adult sporadic papillary thyroid carcinomas. *J Clin Endocrinol Metab*. 2004; 89:2414–2420. [PubMed: 15126572]
9. Kraus JP, Janosik M, Kozich V, Mandell R, Shih V, Sperandio MP, Sebastio G, de Franchis R, Andria G, Kluijtmans LA, Blom H, Boers GH, Gordon RB, Kamoun P, Tsai MY, Kruger WD, Koch HG, Ohura T, Gaustadnes M. Cystathionine beta-synthase mutations in homocystinuria. *Hum Mutat*. 1999; 13:362–375. [PubMed: 10338090]
10. Huang CN, Lee KC, Wu HP, Tai TY, Lin BJ, Chuang LM. Screening for the Gly40Ser mutation in the glucagon receptor gene among patients with type 2 diabetes or essential hypertension in Taiwan. *Pancreas*. 1999; 18:151–155. [PubMed: 10090412]
11. Miller SP, Anand GR, Karschnia EJ, Bell GI, LaPorte DC, Lange AJ. Characterization of glucokinase mutations associated with maturity-onset diabetes of the young type 2 (MODY-2): different glucokinase defects lead to a common phenotype. *Diabetes*. 1999; 48:1645–1651. [PubMed: 10426385]
12. Macfarlane WM, Frayling TM, Ellard S, Evans JC, Allen LI, Bulman MP, Ayres S, Shepherd M, Clark P, Millward A, Demaine A, Wilkin T, Docherty K, Hattersley AT. Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J Clin Invest*. 1999; 104:R33–R39. [PubMed: 10545530]
13. Hughes DJ, Hill AJ, Macek M Jr, Redmond AO, Nevin NC, Graham CA. Mutation characterization of CFTR gene in 206 Northern Irish CF families: thirty mutations, including two novel, account for approximately 94% of CF chromosomes. *Hum Mutat*. 1996; 8:340–347. [PubMed: 8956039]

14. Besancon R, Lorenzi A, Cruts M, Radawiec S, Sturtz F, Broussolle E, Chazot G, van Broeckhoven C, Chamba G, Vandenberghe A. Missense mutation in exon 11 (Codon 378) of the presenilin-1 gene in a French family with early-onset Alzheimer's disease and transmission study by mismatch enhanced allele specific amplification. *Mutations in brief* no. 141. Online. besancon@rockefeller1.univ.lyon1.fr. *Human Mutat.* 1998; 11:481.
15. Murayama O, Murayama M, Honda T, Sun X, Nihonmatsu N, Takashima A. Twenty-nine missense mutations linked with familial Alzheimer's disease alter the processing of presenilin 1. *Prog Neuropsychopharmacol Biol Psychiatry.* 1999; 23:905–913. [PubMed: 10509383]
16. Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F, Raux G, Camuzat A, Penet C, Mesnage V, Martinez M, Clerget-Darpoux F, Brice A, Frebourg T. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am J Hum Genet.* 1999; 65:664–670. [PubMed: 10441572]
17. Cooper DN, Ball EV, Krawczak M. The human gene mutation database. *Nucleic Acids Res.* 1998; 26:285–287. [PubMed: 9399854]
18. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD): 2003 update. *Human Mutat.* 2003; 21:577–581.
19. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T. Ensembl 2004. *Nucleic Acids Res.* 2004; 32:D468–D470. [PubMed: 14681459]
20. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311. [PubMed: 11125122]
21. Zhao H, Pfeiffer R, Gail MH. Haplotype analysis in population genetics and association studies. *Pharmacogenomics.* 2003; 4:171–178. [PubMed: 12605551]
22. Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC. SNP and haplotype variation in the human genome. *Mutat Res.* 2003; 526:53–61. [PubMed: 12714183]
23. Matthews BW. Studies on protein stability with T4 lysozyme. *Adv Protein Chem.* 1995; 46:249–278. [PubMed: 7771320]
24. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol.* 1991; 222:67–88. [PubMed: 1942069]
25. Shortle D, Meeker AK. Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation. *Proteins: Struct, Funct Genet.* 1986; 1:81–89. [PubMed: 3449854]
26. Green SM, Meeker AK, Shortle D. 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.* 1992; 31:5717–5728. [PubMed: 1610820]
27. Stites WE, Meeker AK, Shortle D. Evidence for strained interactions between side-chains and the polypeptide backbone. *J Mol Biol.* 1994; 235:27–32. [PubMed: 8289248]
28. Meeker AK, Garcia-Moreno B, Shortle D. Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry.* 1996; 35:6443–6449. [PubMed: 8639591]
29. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA III. Complete mutagenesis of the HIV-1 protease. *Nature.* 1989; 340:397–400. [PubMed: 2666861]
30. Axe DD, Foster NW, Fersht AR. Active barnase variants with completely random hydrophobic cores. *Proc Natl Acad Sci USA.* 1996; 93:5590–5594. [PubMed: 8643620]
31. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. 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.* 1994; 240:421–433. [PubMed: 8046748]

32. Schmitz A, Schmeissner U, Miller JH. Mutations affecting the quaternary structure of the lac repressor. *J Biol Chem.* 1976; 251:3359–3366. [PubMed: 931989]
33. Kleina LG, Miller JH. Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J Mol Biol.* 1990; 212:295–318. [PubMed: 2157024]
34. Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Muller-Hill B. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol Biol.* 1996; 261:509–523. [PubMed: 8794873]
35. Pace HC, Kercher MA, Lu P, Markiewicz P, Miller JH, Chang G, Lewis M. Lac repressor genetic map in real space. *Trends Biochem Sci.* 1997; 22:334–339. [PubMed: 9301333]
36. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science.* 1996; 271:1247–1254. [PubMed: 8638105]
37. Bell CE, Lewis M. A closer view of the conformation of the Lac repressor bound to operator. *Nat Struct Biol.* 2000; 7:209–214. [PubMed: 10700279]
38. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003; 31:3812–3814. [PubMed: 12824425]
39. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 2002; 12:436–446. [PubMed: 11875032]
40. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* 2001; 11:863–874. [PubMed: 11337480]
41. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 2002; 30:3894–3900. [PubMed: 12202775]
42. Yue P, Li Z, Moulton J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol.* 2005; 353:459–473. [PubMed: 16169011]
43. Yue P, Moulton J. Identification and analysis of deleterious human SNPs. *J Mol Biol.* 2006; 356:1263–1274. [PubMed: 16412461]
44. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol.* 2002; 315:771–786. [PubMed: 11812146]
45. Ferrer-Costa C, Orozco M, de la Cruz X. Sequence-based prediction of pathological mutations. *Proteins.* 2004; 57:811–819. [PubMed: 15390262]
46. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics.* 2005; 21:3176–3178. [PubMed: 15879453]
47. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics.* 2006; 22:2729–2734. [PubMed: 16895930]
48. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics.* 2005; 21:2185–2190. [PubMed: 15746281]
49. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 2005; 33:W480–W482. (Web Server issue). [PubMed: 15980516]
50. Jhee KH, Kruger WD. The role of cystathionine beta-synthase in homocysteine metabolism. *Antioxid Redox Signal.* 2005; 7:813–822. [PubMed: 15890029]
51. Mudd, SH.; Levy, HL.; Kraus, JP. Disorders in transsulfuration. In: Scriver, CR.; Beaudet, A.; Sly, W.; Valle, D., editors. *The metabolic basis of inherited disease.* New York: McGraw-Hill; 2001. p. 2007-2056.
52. De Franchis R, Sperandio MP, Sebastio G, Andria G. Clinical aspects of cystathionine beta-synthase deficiency: how wide is the spectrum? The Italian collaborative study group on homocystinuria. *Eur J Pediatrics.* 1998; 157 Suppl 2:S67–S70.

53. Shan X, Kruger WD. Correction of disease-causing CBS mutations in yeast. *Nat Genet.* 1998; 19:91–93. [PubMed: 9590298]
54. Kruger WD, Cox DR. A yeast assay for functional detection of mutations in the human cystathionine beta-synthase gene. *Hum Mol Genet.* 1995; 4:1155–1161. [PubMed: 8528202]
55. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.* 1975; 405:442–451. [PubMed: 1180967]
56. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics.* 2000; 16:412–424. [PubMed: 10871264]
57. Breiman L. Random forests. *Mach Learn.* 2001; 35:5–32.
58. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of database programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
59. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2008; 36:D13–D21. [PubMed: 18045790]
60. Hubbard, SJ.; Thornton, JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College London; 1993.
61. Silverman, BW. Density estimation for statistics and data analysis. New York: Chapman & Hall; 1986. 175 p.
62. Hall P, Kang KH. Bandwidth choice for nonparametric classification. *Ann Stat.* 2005; 33:284–306.
63. Kery V, Poneleit L, Kraus JP. Trypsin cleavage of human cystathionine beta-synthase into an evolutionarily conserved active core: structural and functional consequences. *Arch Biochem Biophys.* 1998; 355:222–232. [PubMed: 9675031]
64. Shan X, Dunbrack RL Jr, Christopher SA, Kruger WD. Mutations in the regulatory domain of cystathionine beta-synthase can functionally suppress patient-derived mutations in cis. *Hum Mol Genet.* 2001; 10:635–643. [PubMed: 11230183]
65. Meier M, Janosik M, Kery V, Kraus JP, Burkhard P. Structure of human cystathionine beta-synthase: a unique pyridoxal 5'-phosphate-dependent heme protein. *EMBO J.* 2001; 20:3910–3916. [PubMed: 11483494]
66. Taoka S, Lepore BW, Kabil O, Ojha S, Ringe D, Banerjee R. Human cystathionine beta-synthase is a heme sensor protein. Evidence that the redox sensor is heme and not the vicinal cysteines in the CXXC motif seen in the crystal structure of the truncated enzyme. *Biochemistry.* 2002; 41:10454–10461. [PubMed: 12173932]
67. Kim CE, Gallagher PM, Guttormsen AB, Refsum H, Ueland PM, Ose L, Folling I, Whitehead AS, Tsai MY, Kruger WD. Functional modeling of vitamin responsiveness in yeast: a common pyridoxine-responsive cystathionine beta-synthase mutation in homocystinuria. *Hum Mol Genet.* 1997; 6:2213–2221. [PubMed: 9361025]
68. Shan X, Dunbrack RL Jr, Christopher SA, Kruger WD. Mutations in the regulatory domain of cystathionine beta synthase can functionally suppress patient-derived mutations in cis. *Hum Mol Genet.* 2001; 10:635–643. [PubMed: 11230183]
69. Kruger WD, Wang L, Jhee KH, Singh RH, Elsas LJ II. Cystathionine beta-synthase deficiency in Georgia (USA): correlation of clinical and biochemical phenotype with genotype. *Hum Mutat.* 2003; 22:434–441. [PubMed: 14635102]
70. Minnick DT, Liu L, Grindley ND, Kunkel TA, Joyce CM. Discrimination against purine-pyrimidine mispairs in the polymerase active site of DNA polymerase I: a structural explanation. *Proc Natl Acad Sci USA.* 2002; 99:1194–1199. [PubMed: 11830658]
71. Suzuki M, Yoshida S, Adman ET, Blank A, Loeb LA. *Thermus aquaticus* DNA polymerase I mutants with altered fidelity. Interacting mutations in the O-helix. *J Biol Chem.* 2000; 275:32728–32735. [PubMed: 10906120]

72. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci.* 1996; 12:327–345. [PubMed: 8902360]
73. Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet.* 2001; 10:591–597. [PubMed: 11230178]
74. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* 2000; 16:198–200. [PubMed: 10782110]
75. Sunyaev S, Hanke J, Aydin A, Wirkner U, Zastrow I, Reich J, Bork P. Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J Mol Med.* 1999; 77:754–760. [PubMed: 10619435]
76. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci.* 1998; 23:358–361. [PubMed: 9787643]
77. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.* 2002; 30:387–391. [PubMed: 11752345]
78. 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. [PubMed: 7984417]
79. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* 2003; 12:2001–2014. [PubMed: 12930999]
80. Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr Issues Mol Biol.* 2001; 3:47–55. [PubMed: 11488411]
81. Baldi P, Chauvin Y, Hunkapiller T, McClure MA. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA.* 1994; 91:1059–1063. [PubMed: 8302831]
82. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992; 89:10915–10919. [PubMed: 1438297]
83. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* 2004; 32:D189–D192. [PubMed: 14681391]
84. Meier M, Oliveriusova J, Kraus JP, Burkhard P. Structural insights into mutations of cystathionine beta-synthase. *Biochim Biophys Acta.* 2003; 1647:206–213. [PubMed: 12686134]
85. Stenberg KA, Vihinen M. Crystal structure of a 1.6-hexanediol bound tetrameric form of *Escherichia coli* lac-repressor refined to 2.1 Å resolution. *Proteins.* 2009; 75:748–759. [PubMed: 19004002]
86. Bateman A. The structure of a domain common to archaeobacteria and the homocystinuria disease protein. *Trends Biochem Sci.* 1997; 22:12–13. [PubMed: 9020585]
87. Xu Q, Canutescu A, Obradovic Z, Dunbrack RL Jr. ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics.* 2006; 22:2876–2882. [PubMed: 17018535]
88. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL Jr. Statistical analysis of interface similarity in crystals of homologous proteins. *J Mol Biol.* 2008; 381:487–507. [PubMed: 18599072]
89. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. *J Mol Biol.* 1991; 218:397–412. [PubMed: 2010917]
90. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. *Proteins: Struct Funct Genet.* 2005; 61 Suppl 7:225–236. [PubMed: 16187365]
91. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of PRedicted interactions. *Proteins: Struct Funct Genet.* 2003; 52:2–9. [PubMed: 12784359]
92. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 2009; 77:778–795. [PubMed: 19603484]

Figure 1.

Histogram of colony growth according to the maximum value of the first two experiments. **a:** Histogram of Cys-/Cys+ ratio of 948 colonies. **b:** Histogram of Cys-/Cys+ ratio of 357 colonies with no mutations. **c:** Histogram of Cys-/Cys+ ratio of 160 colonies with only S19P. **d:** Histogram of Cys-/Cys+ ratio of 330 colonies with one or more missense mutations.

Figure 2. Two accuracy parameters as a function of the threshold for phenotype determination. **a:** The overall accuracy. **b:** The Matthew's correlation coefficient.

Figure 3.
The kernel density estimations and the kernel discriminant functions for neutral and deleterious mutations of Lac repressor using the PSSM scores of the wildtype residue type.

Figure 4.

Comparison of structure-based and evolutionary information in determining phenotype using kernel discriminant functions. **a:** PSSM values of CBS and Lac mutations. **b:** The difference between wild-type and mutant PSSM scores of CBS and Lac mutations. **c:** The surface area of the wild-type residue in the monomeric protein. **d:** The surface area of the wild-type residue in the biological unit.

Figure 5. The deleterious discriminant function for surface areas calculated from monomers and the relevant biological assemblies. **a:** CBS data. **b:** Lac data.

Table I

Distribution of Mutations

No. of mutations	No. of colonies	% Total	Mean Cys ⁻ /Cys ⁺	
			First Exp.	Second Exp.
0	357	37.7	0.82	0.94
1	356	37.6	0.65	0.78
2	169	17.8	0.48	0.60
3	47	5.0	0.38	0.50
4	13	1.4	0.33	0.28
5	6	0.6	0.18	0.31
Total	948	100.0	0.66	0.78

Table II

Mutation Rates in 948 Sequenced Clones

Base change	Count	Rate ($\times 10^{-6}$)
T $\rightarrow$ A	8	43
T $\rightarrow$ C	141	770
T $\rightarrow$ G	5	27
A $\rightarrow$ T	23	88
A $\rightarrow$ C	12	46
A $\rightarrow$ G	219	840
C $\rightarrow$ A	7	110
C $\rightarrow$ T	36	22
C $\rightarrow$ G	3	9
G $\rightarrow$ A	49	120
G $\rightarrow$ C	2	5
G $\rightarrow$ T	8	20

Published Results for Phenotype Predictions

Table III

Program	Testing data	No.	TPR	TNR	PPV	NPV	ACC	MCC	BER	P	N
SIFT	LacI	4004	57	78	66	71	68	0.359	0.325	44	56
	HIV-1 protease	336	82	70	85	66	78	0.512	0.240	67	33
	T4 Lysozyme	2015	72	59	45	82	63	0.289	0.345	32	68
PolyPhen	SwissProt disease-assoc.(1551)+ species polymorphisms(800)	2351	69	91	94	60	77	0.569	0.200	66	34
SNPS3D-profile	HGMD monogenic disease+	21286	80	90	86	85	86	0.707	0.150	44	56
SNPS3D-stability	Homologous proteins	6077	74	85	89	67	78	0.573	0.205	62	38
PMUT	SWP-Lac ^d	11588	95	90	97	82	94	0.817	0.075	81	19
PHD-SNP	SWP-Evol ^b	20706	79	84	80	83	81	0.631	0.185	45	55
	PDBst ^c	2207	88	90	93	84	89	0.772	0.110	60	40
	HumVar	21185	76	70	80	65	74	0.455	0.270	61	39
nsSNPAnalyzer	NewHumVar	935	74	73	34	94	73	0.363	0.265	16	84
	OutPhD-SNP08 ^d	34314	72	80	76	76	76	0.522	0.240	48	52
	≥ 10 homologous sequence	4013 ^e	79	58	93	28	76	0.274	0.318	87	13
< 10 homologous sequence		4013 ^f	79	62	94	30	77	0.315	0.292	87	13
		205 ^e	69	70	93	28	69	0.282	0.307	85	15
		205 ^f	76	70	94	33	75	0.352	0.270	85	15

TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value; ACC, total accuracy; P and N are the percentage of positives (deleterious) and negatives (neutrals) in the data sets.

^aThe neutral mutations are extracted from LacI.

^bThe neutral mutations are extracted from the evolutionary model.

^cStructure-based case.

^dAvailable at <http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/OutPhD-SNP08.txt>.

^eUsing support vector machine method.

^fUsing random forest method.

Table IV

The Results of Public Methods Tested with CBS Data

Methods	TPR	TNR	PPV	NPV	ACC	MCC	BER
SIFT	89	52	54	88	66	0.411	0.297
PolyPhen	87 ^a	53	54	87	66	0.404	0.299
	72 ^b	70	60	80	71	0.408	0.291
SNPs3D	82	47	50	81	61	0.300	0.353
PMut	44	79	57	69	66	0.250	0.382
PhD-SNP	70	53	48	73	59	0.220	0.388
nsSNPAnalyzer	80	59	55	82	67	0.382	0.305
Excluding SIFT	81 ^b	67	61	85	73	0.470	0.259
Excluding PolyPhen	82^b	63	59	85	71	0.445	0.273
Excluding SNPs3D	81 ^b	68	62	85	73	0.478	0.255
Excluding PMut	80 ^b	60	56	82	68	0.390	0.301
Excluding PhD-SNP	81 ^b	65	59	84	71	0.447	0.271
Excluding nsSNPAnalyzer	80 ^b	63	58	83	70	0.419	0.285

The lower half of the table shows consensus predictions, excluding one method at a time. TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value; ACC, total accuracy. The best values for each measure among the individual methods and among the consensus methods are shown in bold type. This is the largest value for all except BER, for which smallest value is best.

^aPolyPhen “Possibly damaging” mutations considered deleterious (positive).

^bPolyPhen “Possibly damaging” mutations considered neutral (negative).

Accuracy Measures of Four Features when Used Alone on the CBS and Lac Repressor Data Sets

Table V

Features	CBS (unbalanced)						
	TPR	TNR	PPV	NPV	ACC	MCC	BER
PSSM	37	85	60	68	66	0.247	0.392
PSSM difference	58	74	59	74	68	0.327	0.337
Relative ASA of complex	61	62	51	72	62	0.226	0.384
Relative ASA of monomer	38	75	49	66	61	0.140	0.434

Lac							
PSSM	63	77	73	67	70	0.364	0.302
PSSM difference	74	74	74	74	74	0.476	0.262
Relative ASA of complex	85	55	65	79	70	0.420	0.300
Relative ASA of monomer	57	71	66	62	64	0.283	0.360

The largest value for each accuracy measure for each protein system is in bold type. TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value; ACC, total accuracy; MCC, Matthews correlation coefficient; BER, balanced error rate.

Table VI

Structural Positions of Mutations

	Buried	Intermediate	Exposed
Deleterious	50 (10) 52%	15 (5) 48	14 (1) 18
Neutral	47 (20) 48%	16 (6) 52	62 (21) 82
Total	97 (30) 100%	31 (11) 100%	76 (23) 100%

The number of conservative mutations is given in parentheses for each subset. These consist of CS, IT, IV, KR, KQ, LM, LV, MV, ND, NS, ST, VA