Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity

Eric A. Stone[1,2] and Arend Sidow[2,3,4]

[1]Department of Statistics, [2]Department of Pathology, and [3]Department of Genetics, Stanford University, Stanford, California 94305-5324, USA

We find that the degree of impairment of protein function by missense variants is predictable by comparative sequence analysis alone. The applicable range of impairment is not confined to binary predictions that distinguish normal from deleterious variants, but extends continuously from mild to severe effects. The accuracy of predictions is strongly dependent on sequence variation and is highest when diverse orthologs are available. High predictive accuracy is achieved by quantification of the physicochemical characteristics in each position of the protein, based on observed evolutionary variation. The strong relationship between physicochemical characteristics of a missense variant and impairment of protein function extends to human disease. By using four diverse proteins for which sufficient comparative sequence data are available, we show that grades of disease, or likelihood of developing cancer, correlate strongly with physicochemical constraint violation by causative amino acid variants.

[Supplemental material is available online at www.genome.org. A Java executable of MAPP and documentation are freely available for download at http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005.]

Missense mutations that impair protein function may result in disease. For diseases caused by such deleterious mutations, a simple but plausible model presents itself: The type of disease is dependent on when and where the protein's function is required in the organism. Given the type of disease, its severity is likely determined by at least three parameters: (1) the degree to which the function of the protein is impaired by the missense mutation; (2) variants of other genes that modulate the effect of the major locus, also referred to as genetic background; and (3) the environment. We present here a predictive statistical framework for the first of these parameters, impairment of protein function by missense mutations.

Our study was motivated by several observations relevant to protein structure, function, and evolution, and by previous studies that addressed the relationship between the nature of missense variants, impairment of protein function, and resulting disease (Miller and Kumar 2001; Sunyaev et al. 2001; Mooney and Klein 2002; Ng and Henikoff 2002; Ramensky et al. 2002; Botstein and Risch 2003; Krishnan and Westhead 2003; Cai et al. 2004; Lau and Chasman 2004). In aggregate, these studies suggest that mutations in evolutionarily conserved sites tend to impair protein function and lead to disease. There is also weak evidence that protein impairment and disease severity are somewhat correlated with the physicochemical difference between the original amino acid and the missense variant. However, no transparent, quantitative relationship between evolutionary constraint, functional impairment, and disease severity has been described. We believe that there has been a methodological barrier to uncovering such a relationship.

Our analysis rests on two complementary ideas: (1) that differences in standard physicochemical properties between the "wild-type" amino acid and the missense variant are the root cause of functional impairment; and (2) that evolutionary variation among orthologs in the affected position is a sample of the physicochemical properties that are tolerated at that position. By using these two ideas as a premise, we devised MAPP (Multivariate Analysis of Protein Polymorphism), which quantifies the physicochemical variation in each column of a multiple sequence alignment and calculates the deviation of candidate amino acid replacements from this variation. The greater the deviation, the higher is the probability that a replacement impairs the function of the protein, and the greater is its predicted effect on the function of the protein. We show that the degree of constraint violation by missense changes is highly predictive of the functional impairment of the protein. Furthermore, we show with four diverse proteins for which sufficient data are available that constraint violations quantitatively translate to grades of disease, or likelihood of developing cancer.

[4]Corresponding author.
E-mail arend@stanford.edu; fax (650) 725-4905.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3804205. Article published online before print in June 2005.

Results

MAPP methodology

MAPP consists of seven steps (Fig. 1A). We first build a multiple alignment of orthologs or closely related paralogs; distant paralogs are excluded to avoid including evolutionary variation that specifies functional differences. The sequences' evolutionary relationships are inferred by standard likelihood analysis (Friedman et al. 2002), which also yields the branch lengths in substitutions per site, for the tree (Fig. 1A, step 1). Based on the topology and branch lengths of the tree, weights are calculated for each sequence that control for phylogenetic correlation among the sequences (Fig. 1A, step 2). Multiplication of the weights with the fraction of sequences carrying a particular amino acid yields the alignment summary (Fig. 1A, step 3), which we interpret by using a matrix of physicochemical property scales (Fig. 1A, step

15:978–986 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

4). The result is an estimate of the physicochemical constraints on each position in terms of the mean and variance of the property distributions observed in its alignment column (Fig. 1A, step 5). These statistics are biologically meaningful; the mean hydropathy value at a position estimates its hydrophobic character, while the variance measures the strength of that constraint. Deviations from the alignment column are obtained for each possible variant by calculating its property difference from the mean and dividing by the square root of the variance (Fig. 1A, step 6). We interpret this statistic as a signed measure of constraint violation. To compute a single score measuring the violation of constraint across all properties, we first decorrelate the properties themselves by using a principal component transformation (see Methods). This is necessary because, for example, hydropathy and polarity of the 20 amino acids are significantly correlated. The decorrelation gives rise to a new coordinate system in which each axis is a principal component; the distance from the origin to any variant is the variant's decorrelated impact score (Fig. 1A, step 7).

An impact score is thus assigned to every possible variant in the protein, here illustrated for p53 (Fig. 1B). A high impact score identifies a potentially deleterious variant by virtue of its physicochemical dissimilarity to the observed evolutionary variation, whereas low-scoring variants are less likely to compromise protein structure or function. The preponderance of high scores in p53's DNA binding domain (positions 100–300) is consistent with a corresponding high frequency of mutations that inactivate the protein in somatic tumors (Olivier et al. 2002). Should a single value be desired for each position (as a "summary constraint" or to illustrate the general mutability of the position), the median of its 20 scores is taken (Fig. 1B). Mapping the medians onto the crystal structure of p53 (Fig. 1C) indicates that the core of the protein is moderately constrained, with the individual impact scores being generally lowest for hydrophobic variants. Surface residues not involved in ligand binding vary with little consequence to protein function, whereas the residues involved in DNA or zinc binding are least tolerant to substitution. MAPP's conversion of evolutionary variation into impact scores that capture physicochemical constraint is therefore generally consistent with known principles of protein structure and function.

To compare impact scores with experimental data, we identified four mutagenesis studies (on *Escherichia coli* LacI (Markiewicz et al. 1994; Suckow et al. 1996), T4 Lysozyme (Rennell et al. 1991), HIV Protease (Loeb et al. 1989), and HIV Reverse Transcriptase (RT) (Wrobel et al. 1995) in which a large number of single-substitution protein variants were assayed for their degree of functional impairment. For each of these studies, we obtained a multiple sequence alignment of homologs and calculated the impact score for each variant that was assayed experimentally. The variants and their associated impact scores were then subdivided into three classes according to their experimentally deter-

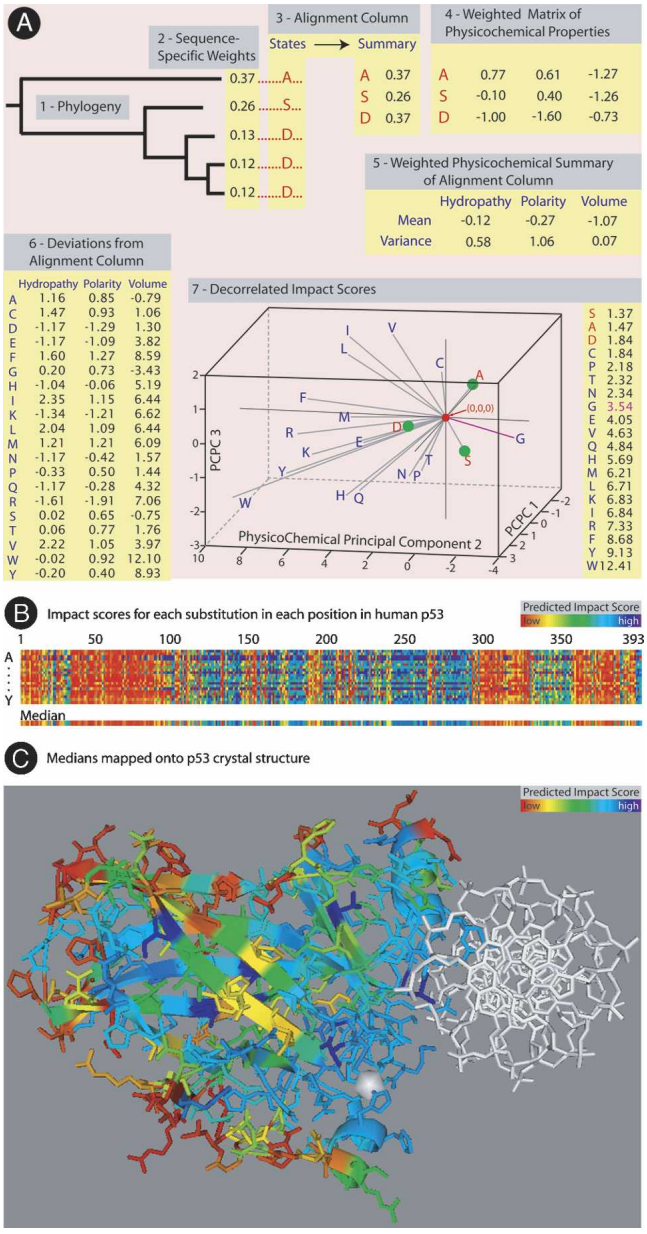

Figure 1. (*A*) MAPP's seven analysis steps. Evolutionary relationships of the protein sequences in the multiple alignment are inferred by likelihood analysis (1). Weights for each sequence are calculated to control for phylogenetic correlation (2). (The remaining steps consider each position in the protein independently and are illustrated for one such position.) Each column of the alignment is condensed into a summary in which each of the 20 amino acids is represented by the sum of the weights of those sequences carrying the amino acid at that position in the alignment (3). The summary is interpreted using a universal matrix of physicochemical property scales, only three of which are shown: hydropathy, polarity, and volume (4). The result is an estimate of the physicochemical constraints on each position in terms of the mean and variance of the property distributions observed in its alignment column (5). Deviations from the alignment column are obtained for each possible variant by calculating its property difference from the mean and dividing by the square root of the variance (6). To compute a single score measuring the violation of constraint across all properties, we first decorrelate the properties themselves by using a principal component transformation. The decorrelation gives rise to a new coordinate system in which each axis is a principal component; the distance from the origin to any variant is the variant's decorrelated impact score (7). (*B*) Each possible variant at each position in the protein is color-coded by its MAPP score, shown here for human p53. Each column corresponds to a position in human p53, in order of sequence. The spectrum of possible variants at each position reads from *top* to *bottom*, arranged alphabetically by one-letter amino-acid abbreviation. Scores for each variant are color-coded from low (red) to high (blue) as a heat map, with temperature inverse to the predicted impact of that change on the protein. The median score of possible variants at each position is shown *below* with the same color code. This median was used to color *C*. (*C*) Median MAPP scores plotted on the crystal structure of human p53 (Cho et al. 1994; DeLano 2002). Chelated Zinc and bound DNA are white.

mined activity: positive, as wild-type function; intermediate, as moderately deleterious; and negative, as strongly deleterious. Accordingly, we obtained three distributions of impact scores per mutagenesis whose medians increased substantially with the degree of a variant's functional impairment (Fig. 2A). For each mutagenesis, we also calculated a distribution comprised of the impact scores of all achievable variants and compared this "control distribution" with each of the three experimental subdivisions (see Methods). Positive variants scored lower than the control, negative variants scored higher, and the intermediate distribution resembled the control (Fig. 2A). MAPP thus captures the known phenomenon that the average substitution in a protein is mildly deleterious to structure or function (Li 1997).

For each mutagenesis, the score distribution of variants with positive activity is easily distinguished from its deleterious complement (intermediate plus negative activity). Low impact scores are clearly overrepresented (vs. control) in the positive class and underrepresented in the deleterious class, whereas for high impact scores the reverse is true (Fig. 2B). The ability to discriminate between intermediate and negative variants within the deleterious class motivated a closer comparison of MAPP scores to the HIV RT data set, the only one for which activity levels of variants were reported as continuous values (Wrobel et al. 1995). Grouping the RT variants into four classes by activity level and calculating the differences to the control distribution shows the excess of high activity variants for low scores, and low activity variants for high scores, with intermediates in between (Fig. 2C). Median scores increase significantly in successively impaired classes (Wilcoxon test: >50% vs. >5% but ≤50%, $P < 10^{-4}$; >5% but ≤50% vs. >1% but ≤5%, $P < 10^{-6}$; >1% but ≤5% vs. ≤1%, $P = 0.0008$). Most importantly for the eventual goal of distinguishing deleterious variants on a continuous scale, linear regression (appropriately using all variants as individual data points, without grouping them) revealed significant negative correlation between impact score and the logarithm of enzymatic activity ($r = -0.56$; $P < 10^{-30}$).

For a test of MAPP's predictive ability, we devised a simple probability-based rule that provides an impact-score threshold, above which variants are predicted to be deleterious and below which they are predicted to be positive (see Methods). We used this rule to assess MAPP's ability to predict the experimentally measured impact of the assayed variants in each the four mutagenesis studies (Table 1). We compared MAPP's accuracy on these data sets, ranging from 64.1%–80.4%, against that of SIFT (Ng and Henikoff 2001), the most successful sequence-based approach to classifying protein variants. SIFT's accuracy using the same alignment and mutagenesis data varied from 55.0%–78.6%, underperforming MAPP in every case (Table 1). In further contrast to SIFT, MAPP scores are widely spread across the deleterious spectrum, a feature that permits further testing of MAPP's predictive capacity. Applying the same model as before, we chose a more stringent score threshold to discriminate intermediate and negative variants. MAPP's accuracy here varies from 62.6%–76.7% (Table 1), confirming its ability to resolve strata of subfunctional variants.

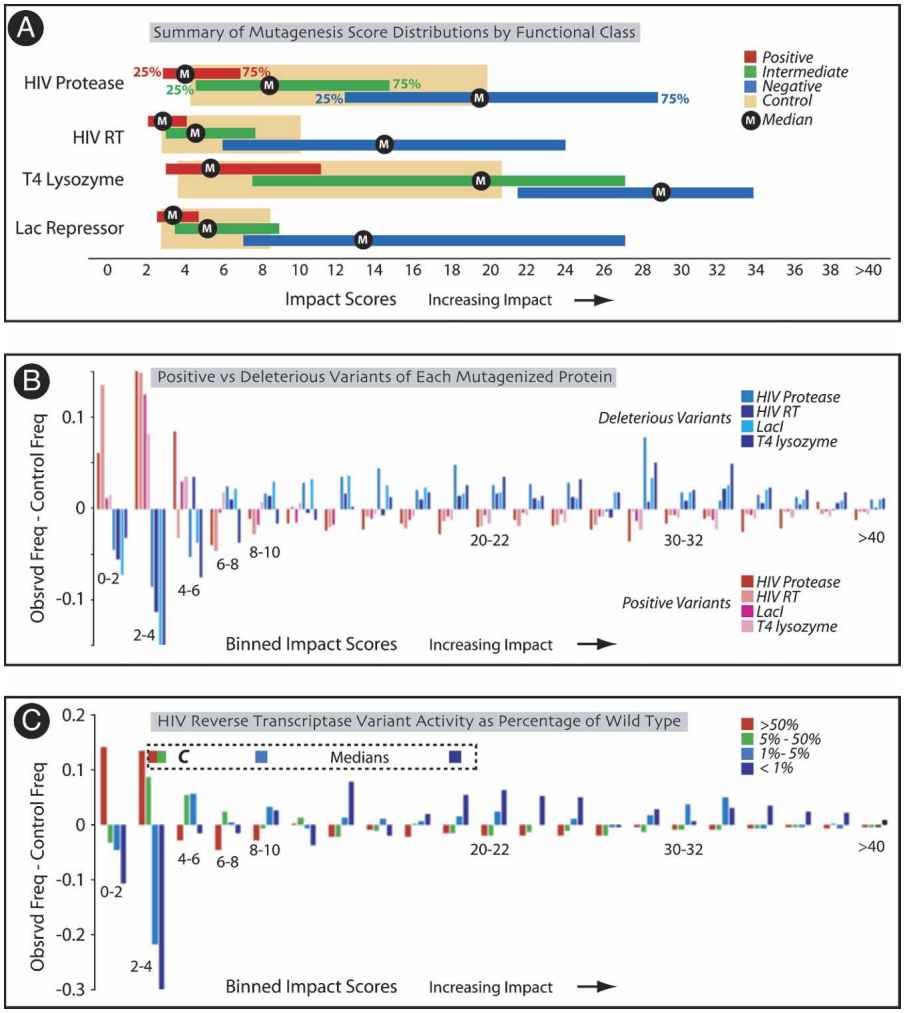

Figure 2. Comparison of MAPP scores of protein variants with mutagenesis studies. (*A*) Scores of protein variants assayed in the four mutagenesis experiments. Variants were partitioned by function as positive (red), intermediate (green), or negative (blue). The interquartile range (25%–75%) of MAPP scores for each set is shown, with the median value denoted by the M. Interquartile ranges of control distributions are in tan. (*B*) Deleterious variants (blue; intermediate plus negative from *A*) and positive variants (red; from *A*) are contrasted. MAPP scores for each set were segregated in bins of width two from zero to 40 (shown *left* to *right*); observed frequencies were calculated by dividing bin counts by the total number of variants in that set. Vertical bars show the difference between observed frequencies and control frequencies, with the latter obtained similarly from the appropriate control distribution. (*C*) Contrast between experimental distribution versus control distribution as in *B* of HIV reverse transcriptase variants. Variants were partitioned by enzymatic activity relative to wild-type (>50%, red; >5% but ≤50%, green; >1% but ≤5%, light blue; ≤1%, dark blue). Colored squares show the median MAPP score of each variant class *above* the bin to which it belongs, with C representing the median of the control distribution.

Table 1. Accuracy in classifications of variants assayed in four mutagenesis experiments

	Positive vs. deleterious			Intermediate vs. negative
	MAPP	SIFT	Increase	MAPP
HIV protease	80.4%	78.6%	1.8%	76.7%
LacI	69.2%	67.9%	1.3%	74.5%
HIV RT	64.1%	55.0%	9.1%	72.9%
T4 lysozyme	73.0%	68.3%	4.7%	62.6%

Paralogous vs. orthologous prediction accuracy

Comparative analyses such as MAPP assume that function is conserved within the aligned set of homologs under study. To the extent that this assumption is violated, constraints in positions that specify functional differences between the homologs in the alignment will be misestimated. Alignments that are enriched for orthologs are ideal for preserving specificity, but paralogs are sometimes included to capture a sufficiently diverse sample of evolutionary variation. This tradeoff is evident for LacI, whose original alignment contains a preponderance of paralogs that act on different operators and have diverse inducers: Sensitivity benefits from the deep alignment, but specificity suffers from inconsistent function (Ng and Henikoff 2002). To quantify the loss of accuracy due to inconsistency, we repeated the LacI analysis on an alignment of six orthologs. This smaller alignment outperformed the original alignment in prediction accuracy (70.7% vs. 69.2%) despite its substantially reduced sequence diversity.

The superior performance of the ortholog alignment suggests an extensive misclassification of variants in residues important to ligand binding. We explored this hypothesis by restricting the analysis to a subset of positions involved in inducer binding (Markiewicz et al. 1994; Lewis et al. 1996). MAPP's analysis of the deep, paralogous alignment was substantially less accurate for this restricted set of positions than for the rest of the protein (Fig. 3A). For these positions, use of the alignment of orthologs led to a significant improvement in accuracy (Exact test, $P = 0.004$), demonstrating better resolution of both functional and deleterious variants. To control for differences between the alignments unrelated to orthology, we analyzed 5000 alignments of six sequences, each composed of *E. coli* LacI and five sequences randomly chosen from the large alignment. None surpassed the classification accuracy of the ortholog alignment (Fig. 3B).

Together, the analyses above show that physicochemical constraint violations quantitatively govern the impairment of protein function by missense variants. We now begin to consider to what extent MAPP can illuminate missense changes in a population where direct measurements of protein activity are not available, and ask whether a quantitative relationship between physicochemical constraint violations and organismal impairment is equally detectable.

Polymorphisms in the HIV population

As a simple model, we first analyzed polymorphisms present in the HIV population within HIV-positive patients. In an untreated patient, every possible single point mutation in HIV occurs between 10^4 and 10^5 times per day (Coffin 1995). Variant proteins emerge with regularity, but the immune system of the patient imposes an extreme selective regime on the virus, whose proteins are therefore under intense purifying selection. We expect no subfunctional HIV protease variants to be present at a high frequency in infected individuals who have not been treated for their infection. MAPP soundly confirms this expectation (Fig. 4A), as no polymorphisms predicted to be deleterious appear at >5% frequency.

The frequency distribution of HIV protease variants is dramatically different within patients who have been treated with protease inhibitors (Fig. 4B). Seven subfunctional variants have now risen to a frequency >5%. All of them are known to confer resistance to protease inhibitors, and four of them are the most effective positions in conferring resistance (Rhee et al. 2003). Thus, MAPP provides evidence that these variants are indeed deleterious to the normal function of the protease, but in patients undergoing treatment, the resulting impaired fitness of the virus is outweighed by positive selection due to drug resistance.

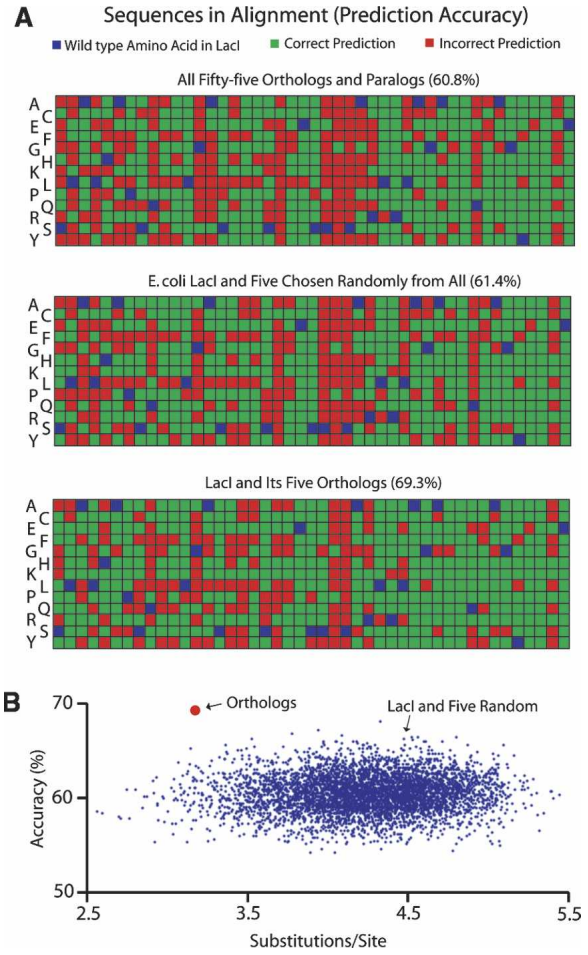

Figure 3. Effect of paralogous sequences on prediction accuracy. (*A*) Differential accuracy in MAPP classification of LacI variants implicated in ligand binding when different types of homologs are represented in the alignment. Three types of alignment are compared. Variants classified at each position are arranged from *top* to *bottom* alphabetically by one-letter abbreviation. Positions are shown *left* to *right* in increasing order (*24*). Green and red identify correct and incorrect predictions, respectively, of whether a variant is functional versus deleterious; the wild-type amino acid is blue. (*B*) Classification of 5000 alignments, each containing LacI and five sequences randomly chosen from the original alignment. Accuracy is plotted against total evolutionary divergence as measured in substitutions per site for random alignments (blue) and the single alignment of six orthologs (red).

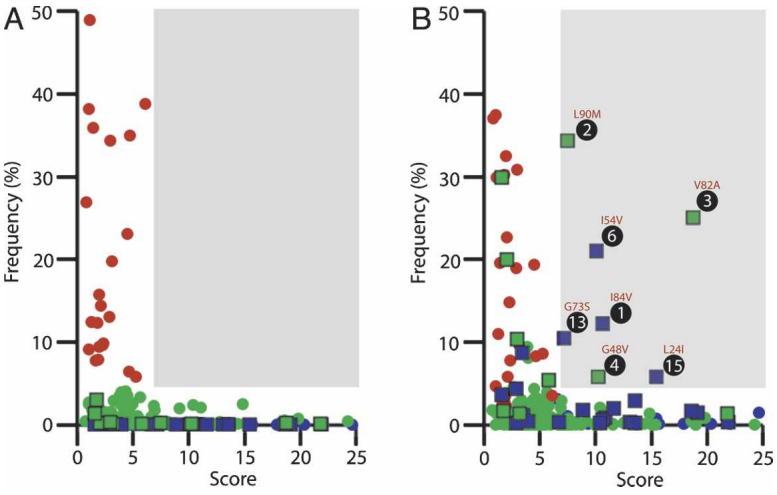

Figure 4. HIV protease variant frequencies versus MAPP score. (*A*) Frequency of each variant by position within the population of untreated individuals plotted against its MAPP score. Rare variants (frequency <0.1%) are in blue, common variants (frequency >5%) are in red, and those in between are in green. Variants known to confer resistance to protease inhibitors are the bordered squares; the remaining variants are circles. (*B*) Plot of variant frequency after treatment with protease inhibitor(s). Common variants with high MAPP scores are labeled and ranked by effectiveness in conferring resistance. Color-coding of variants is by their frequency in untreated patients, as in *A*, for comparison.

Human disease

In the HIV analysis, the quantity we compare to MAPP scores, "variant frequency," is not a direct measure of protease function but is a derivative measure that reflects both impact on the protease as well as the importance of the protease's function for the HIV organism. Now we will consider the severity of disease phenotype, which is also not a direct measure of protein function. We first turned our attention to anemias, which can be caused by the impairment of a variety of cellular processes in the blood. For pyruvate kinase (PK) (Stenson et al. 2003), glucose-6-phosphate dehydrogenase (G6PD) (Kwok et al. 2002), and β hemoglobin (HBB) (Hardison et al. 2002), well-suited mutation databases exist and sufficiently deep alignments could be generated. We first tested whether the MAPP score distributions of disease-causing variants differed significantly from the control distributions of all possible missense variants. The expectation is that the typical variant, as described by the control distribution, is deleterious, but that strong loss-of-function mutations are more deleterious than this average. This expectation is met at statistical significance for all three proteins (Wilcoxon test: G6PD, $P = 0.008$; PK, $P < 10^{-5}$; HBB, $P < 10^{-14}$) (Fig. 5A). The effect is least pronounced for the X-linked G6PD, for which the disease distribution is closer to the control than for the other proteins. This is likely due to the absence of the most severe mutations from the database because of their embryonic lethality in males (Vulliamy et al. 1993). The large differences in median MAPP scores between control and severe disease variants for PK (6.5) and HBB (15.3) mirror the results obtained from the activity of mutagenesis variants (Fig. 2). We therefore tested MAPP's ability to distinguish classes of different phenotypic severity by using HBB, the only protein of the three with sufficiently detailed clinical records.

Mutations in HBB can lead to a number of disease phenotypes, including the anemia considered above (Huisman et al. 1996). We defined four classes of HBB variants via their associated hematology: normal, erythrocytosis, anemia (all forms), and formation of Heinz bodies. Functionally normal variants scored lowest (median = 8.74), well below the control (10.72), reaffirming MAPP's ability to isolate tolerated variants (Fig. 5B). More notably, the score distributions from disease classes were grossly distinct and quantitatively consistent with clinical presentation (Fig. 5B; Huisman et al. 1996). Heinz bodies were associated with the highest scoring variants (median = 28.31), anemia variants scored much lower (18.82), and erythrocytosis variants lower still (14.46). The disease records additionally allowed further subdivision of the anemia variants by severity (mild, moderate, or severe). Again, score distributions were consistent with clinical presentation, with the median score of mild anemia variants (13.39) less than that of moderate variants (18.60) and with severe variants scoring highest (25.54) (Fig. 5C).

We next asked whether MAPP scores of p53 variants could accurately reflect the likelihood of tumor formation. As noted by several previous studies, frequency of a variant in the p53 somatic mutation database (Olivier et al. 2002) is a proxy for the likelihood that the variant will allow tumor formation, and does not primarily reflect variation in mutation rate (Walker et al. 1999; Simon et al. 2002). We stratified variants according to their count in the somatic mutation database as appearing one to four times (663 variants, 1293 mutations), five to nine times (206 variants, 1346 mutations), or greater than nine times (248 variants, 10,870 mutations). Any variants conceivable by a single missense change but not present in the database were assigned to a zero class (1201 variants). The median MAPP score of the zero class was 4.46, which is significantly lower than that of any of the other three classes and the control (Fig. 5D), confirming that it represents variants that do not disrupt the function of the protein sufficiently to cause cancer. The control distribution, once again, was centered in the mildly deleterious range, and we could not distinguish it from the distribution of low-count variants either by inspection (Fig. 5D) or by statistical test (Kolmogorov-Smirnov test, $P = 0.31$). Most importantly, MAPP was able to distinguish at statistical significance between the three classes of deleterious variants, demonstrating a significant increase in median scores (scores 1–4: 7.20; scores 5–9: 13.88; scores 10+: 20.97) for increased prevalence in the database (Fig. 5D).

Because somatic variants may be considered a special case, we also investigated the much rarer germline missense variants of p53 that have been shown to cause Li-Fraumeni syndrome (80 variants). These exhibit a skewed distribution that is statistically indistinguishable from the distribution of high-count somatic p53 variants (five or greater; Kolmogorov-Smirnov: $P = 0.25$), with a median MAPP score of 17.73, right between the median scores for variants appearing five to nine and 10 or more times in the somatic database. It is significantly different from the two distributions of mild (one to four counts) and no-effect (zero count) variants in the somatic database. Thus, physicochemical constraint violations, resulting in impaired function of the tumor suppressor protein, also capture likelihood of cancer development.

Discussion

Our results establish that physicochemical constraint violations, as inferred from evolutionary variation in orthologous se-

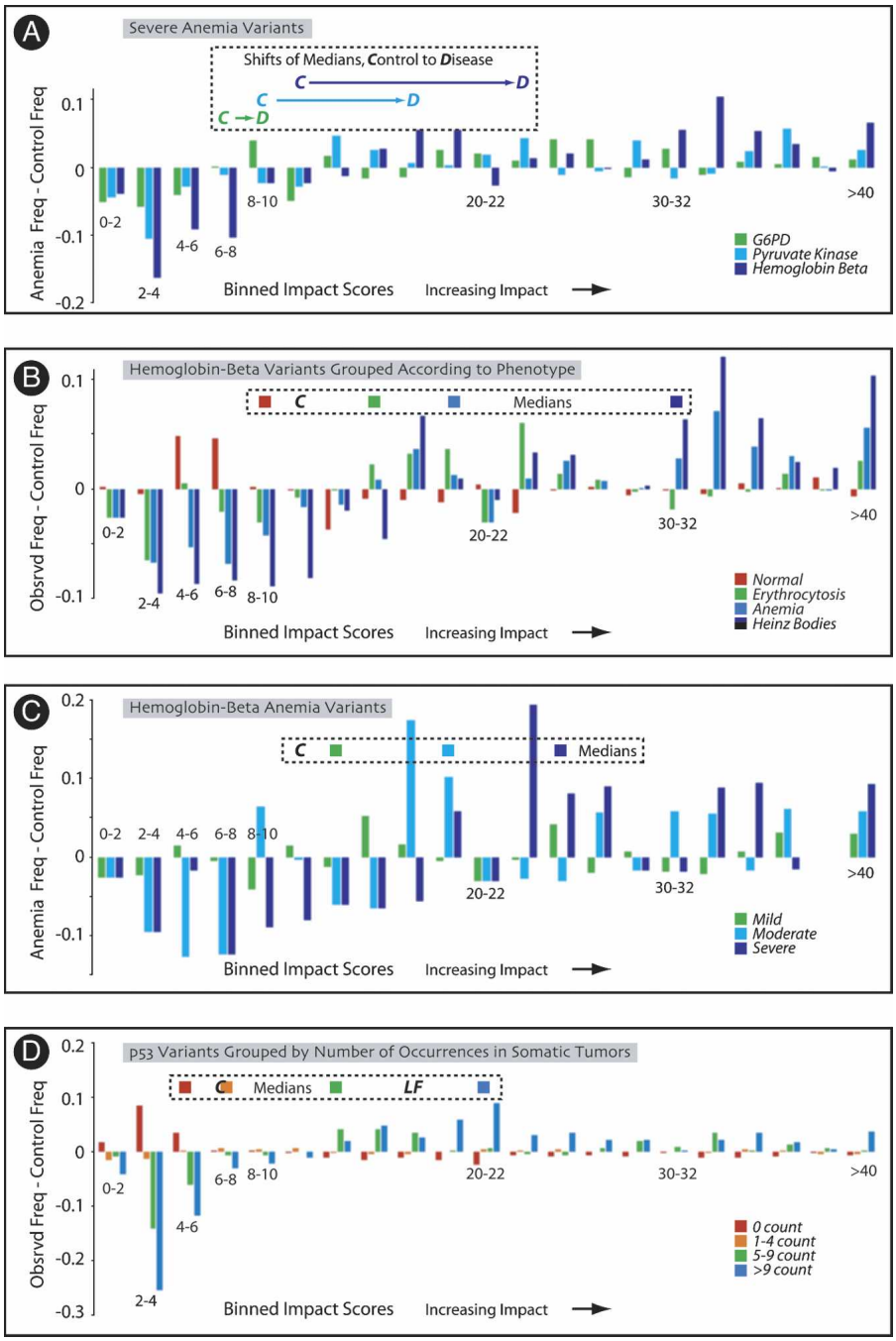

Figure 5. MAPP score distributions of disease variants grouped by severity of disease. Frequency differences between classes of variants and their respective control distributions are computed and shown as in Figure 2. Median scores of the original distributions are shown *above* the score bin into which they fall. (*A*) Frequency differences between distributions of anemia variants that cause hemolytic anemia and control distributions, for G6PD (green), pyruvate kinase (light blue), and β hemoglobin (dark blue). The shifts in medians from the control distribution (C) to the disease distribution (D) are indicated *above* their bin. (*B*) Comparisons among score distributions of β hemoglobin variants. Frequency differences and median scores are shown by phenotype (normal, red; erythocytosis, green; anemia, light blue; Heinz bodies, dark blue). (*C*) Comparisons among score distributions of β hemoglobin variants that cause anemia, graded by severity of the disease. Frequency differences and median scores are shown by severity of anemia (mild, green; moderate, light blue; severe, dark blue). (*D*) Scores of p53 variants isolated from somatic tumors. Frequency differences from control, and median scores, are shown for variants stratified by incidence (score 0, red; scores 1–4, orange; scores 5–9, green; 10+, blue). C indicates median score for control distribution; *LF*, median score for distribution of Li Fraumeni variants.

quences, are highly predictive of the impact of missense substitutions on protein function. This impairment of function is apparently a major determinant of disease severity, at least for the diseases and genes we analyzed in this study. Confounding factors such as genetic background and environment were evidently not strong enough to obscure the relationship between MAPP scores and disease severity. Previous studies that addressed missense substitutions and disease lacked explanatory power as to the underlying fundamental principles governing disease (Miller and Kumar 2001; Sunyaev et al. 2001; Mooney and Klein 2002; Ng and Henikoff 2002; Ramensky et al. 2002; Krishnan and Westhead 2003; Cai et al. 2004; Lau and Chasman 2004): The effect of physicochemical properties was marginal at best (Miller and Kumar 2001; Botstein and Risch 2003), and although comparative information seems better at classifying disease variants (Ng and Henikoff 2002; Botstein and Risch 2003), conservation in and of itself has no explanatory power. However, we have shown with MAPP that physicochemical constraint is the major determinant of protein impairment, that evolutionary variation allows quantitative estimates of constraint, and that violations of such constraints mediate disease severity via the impaired function of the protein. This result motivates a general, reductionist view of the relationship among amino acid properties, protein function, and Mendelian disease.

Previous studies also did not address the impact of missense variants on a continuous scale, focusing solely on a binary classification of deleterious versus tolerated substitutions. By contrast, MAPP can distinguish intermediate from negative variants (Table 1) and in fact allows a continuous classification, as illustrated for the RT data, because its impact scores are widely spread across the subfunctional spectrum. Should a binary classification be desired, however, then it is still advantageous to use MAPP, as shown by a comparison of its prediction accuracy to that of the most successful similar method, SIFT (Ng and Henikoff 2001). MAPP outperforms SIFT in distinguishing positive from deleterious variants (Table 1), even for the data set upon which SIFT was trained (LacI).

MAPP's predictive accuracy is complemented by the interpretability of its impact scores, which provide a trans-

parent rationalization of predictions in terms of physicochemical properties. Each variant's impact score can be dissected into individual components that measure property-specific constraint violations (Fig. 1A, step 6), effectively assigning a rationale to every prediction. Although the score of any particular variant should presently be interpreted with caution, MAPP's accuracy is likely to improve with better comparative sequence data and larger genotype collections. Eventually, predictions for major loci will be combined with those for modifiers. We speculate that aggregate constraint violations in more than one protein, perhaps in conjunction with noncoding changes, quantitatively mediate disease in a way that will be predictable in the not-too-distant future.

Methods

Multivariate analysis of protein polymorphism

MAPP uses quantitative scales measuring six physicochemical properties to evaluate missense variants: (1) hydropathy (Kyte and Doolittle 1982); (2) polarity (Stryer 1995); (3) charge (Stryer 1995); (4) side-chain volume (Zamyatin 1972); (5) free energy in α-helical conformation (Muñoz and Serrano 1994); and (6) free energy in β-sheet conformation (Muñoz and Serrano 1994). We denote these scales by s_k for $k = 1, \ldots, 6$, with k respecting the order defined above. Define the function α from the integers $1, \ldots, 20$ to the alphabet of amino acids so that represents α_1 alanine (A), α_2 represents cysteine (C), etc., and, for example, s_2 (α_1) represents the polarity of alanine. Because MAPP independently considers each position in an alignment, we describe the method (Fig. 1) for one column A, with A_i denoting the amino acid (or gap character) from sequence i.

Step 1: Alignment, treebuilding

We followed standard procedures for alignment and treebuilding, using ClustalW (Thompson et al. 1994) and SEMPHY (Friedman et al. 2002).

Step 2: Sequence weights

We derived w_i, the weight of sequence i in the alignment by using a modification of a published Brownian motion model (Altschul et al. 1989). Our method "averages" their procedure over all possible root positions of the tree, resulting in weights that capture relative phylogenetic informativeness (Stone 2004).

Step 3: Alignment summary

The alignment column is summarized by its weighted composition in a 20×1 vector $\mathbf{c}$, with

$$\mathbf{c}_m = \sum_{i=1}^{n} w_i \mathbf{1}_{\{A_i = \alpha_m\}}$$

assigned to the weighted representation of amino acid α_m in the column. The gap weight

$$g = \sum_{i=1}^{n} w_i \mathbf{1}_{\{A_i = GAP\}}$$

is distributed among the amino acids present in proportion to their weighted representation, yielding a new vector

$$\mathbf{c}' = \frac{\mathbf{c}}{1 - g}.$$

This definition is in agreement with the usual notion of alignment profile provided the sequences are ungapped and of equal weight. To avoid division by zero, we work with the modified profile

$$\mathbf{p} = (1 - .01)\mathbf{c}' + (.01)\frac{1}{20}\mathbf{1};$$

this allows for a uniform treatment (including fully conserved residues) without introducing artifacts into the analysis.

Step 4: Weighted matrix of physicochemical properties

Because the physicochemical scales we consider measure properties in incompatible units, we first standardize each to have mean zero and standard deviation one. The property mean and variance of s_k are, respectively,

$$\mu_k = \frac{1}{20}\sum_{m=1}^{20} s_k(\alpha_m)$$

and

$$\sigma_k^2 = \frac{1}{19}\sum_{m=1}^{20} (s_k(\alpha_m) - \mu_k)^2,$$

and property values are standardized by these quantities as

$$s'_k(\alpha_m) = \frac{s_k(\alpha_m) - \mu_k}{\sigma_k}.$$

We collect these standardized scales into a 20×6 matrix $\mathbf{S}$; thus $\mathbf{S}_{mk} = s'_k(\alpha_m)$, so that column k of $\mathbf{S}$ lists the ordered values of standardized scale k. The weighted matrix of physicochemical properties is $diag(\mathbf{p})\mathbf{S}$, where $diag(\mathbf{p})$ is the 20×20 matrix with entries $\mathbf{p}_m$ on the diagonal and zeros elsewhere.

Step 5: Weighted physicochemical summary

The column mean based on property k is

$$\hat{\mu}_k = \sum_{m=1}^{20} \mathbf{p}_m s'_k(\alpha_m),$$

and the column variance based on property k is

$$\hat{\sigma}_k^2 = \sum_{m=1}^{20} \mathbf{p}_m (s'_k(\alpha_m) - \hat{\mu}_k)^2.$$

Step 6: Deviations from alignment column

With respect to property k, the deviation of amino acid α_m from the column is quantified by

$$\mathbf{d}_m(k) = \frac{s'_k(\alpha_m) - \hat{\mu}_k}{\hat{\sigma}_k}.$$

The 6×1 vector of deviations for α_m will be denoted by $\mathbf{d}_m$. We interpret this statistic as a signed measure of constraint violation.

Step 7: Decorrelated impact scores

The property correlation matrix is

$$\mathbf{R} = \frac{1}{19}\mathbf{S}^T\mathbf{S};$$

$\mathbf{R}$ is a $p \times p$ matrix with $\mathbf{R}_{kl}$ equal to the correlation between scales

k and l. We use $\mathbf{R}$ as a transforming positive definite matrix to yield the impact score

$$D_m = \sqrt{\mathbf{d}_m^T \mathbf{R}^{-1} \mathbf{d}_m}.$$

Note that $\mathbf{d}_m^T \mathbf{R}^{-1} \mathbf{d}_m$ is the principal component transformation mentioned in Figure 1. Specifically, let $\mathbf{S} = \mathbf{UBV}^T$ be the singular value decomposition; then we can write

$$\mathbf{R}^{-1} = (\sqrt{19}\mathbf{B}^{-1}\mathbf{V}^T)^T(\sqrt{19}\mathbf{B}^{-1}\mathbf{V}^T).$$

Therefore

$$\mathbf{d}_m^T \mathbf{R}^{-1} \mathbf{d}_m = (\sqrt{19}\mathbf{B}^{-1}\mathbf{V}^T\mathbf{d}_m)^T(\sqrt{19}\mathbf{B}^{-1}\mathbf{V}^T\mathbf{d}_m).$$

Null distribution of impact scores

We considered that the collection of amino acids tolerated at any position could be described by a probability distribution on the range of values of the properties under consideration. To give MAPP scores a probabilistic interpretation, we made the assumption that this distribution is multivariate normal. Sequence i contributes one observation to the multivariate normal distribution, namely, the vector of property values associated with amino acid. We supposed that these observations were independent and weighted, with weights from above, so that formally

$$[s'_1(A_i) \cdots s'_6(A_i)]^T \sim N\left(\mu, \frac{1}{nw_i}\Sigma\right).$$

The vector of means $\boldsymbol{\mu} = [\mu_1 \ldots \mu_6]^T$ represents true mean of the tolerated amino acids with respect to each property, and we wish to determine whether these values are significantly different from the property vector $\mathbf{z}_m = [\, s'_1(\alpha_m) \ldots\ s'_6(\alpha_m)]^T$ of our polymorphic amino acid α_m. The impact score

$$D_m = \sqrt{\mathbf{d}_m^T \mathbf{R}^{-1} \mathbf{d}_m}$$

is an estimate of the difference between μ and $\mathbf{z}_m$ that can be used for this purpose. We assumed that Σ could be written as $\mathbf{QRQ}$ for some diagonal matrix $\mathbf{Q}$ and made the approximation that

$$\frac{n-6}{(n+1)6} D_m^2 \sim F_{6,n-6}.$$

We used the 99th percentile of this F-distribution to define classification thresholds for MAPP.

Analysis of the mutagenesis data

Variants and their associated impact scores were subdivided into three classes according to their experimentally determined activity: positive, as wild-type function; intermediate, as moderately deleterious; and negative, as strong loss of function. In the analysis of HIV Protease, we took the classes exactly as defined in the mutagenesis (Loeb et al. 1989). For HIV RT (Wrobel et al. 1995) and T4 Lysozyme (Rennell et al. 1991), four graded phenotypes were reported, and in both cases we grouped the two middle phenotypes to form our intermediate class. Detailed repression and induction phenotypes were reported for LacI variants (Markiewicz et al. 1994; Suckow et al. 1996). This complicated a tripartite categorization, and for that analysis, we proceeded as above after restricting consideration to repressor function when the inducer appeared functional. All variants were used in the classification of positive and deleterious variants.

For each mutagenesis, we also calculated a distribution comprised of the impact scores of all achievable variants for each case and compared this "control distribution" with each of the three experimental subdivisions. For the oligonucleotide-directed saturation mutageneses (HIV Protease [Loeb et al. 1989] and HIV RT [Wrobel et al. 1995]), the control was composed of all single missense variants. When amber mutations were introduced in a systematic mutagenesis (LacI [Markiewicz et al. 1994; Suckow et al. 1996] and T4 Lysozyme [Rennell et al. 1991]), the collection of variants studied formed their own control.

For the RT correlation analysis, 22 of the 366 variants were reported to have an unspecified activity level of <1% wild-type; we assumed an activity level of 1% for all variants below that threshold, which appears from the data to yield a conservative result.

Prediction accuracy, orthologs

SIFT's classification threshold was optimized for performance on a data set comprising 55 LacI-related sequences, including paralogs (Ng and Henikoff 2001). We used that same alignment for comparison of MAPP's accuracy with that of SIFT. We assessed the impact of alignment composition on prediction accuracy by restricting consideration to positions 70–80, 90–100, 190–200, 245–250, 272–277. Certain substitutions at these positions lead to variant repressor molecules that bind to the operator DNA with wild-type affinity but are incapable of induction (Lewis et al. 1996); we reasoned that paralogous sequence would be misleading for this class of variants, while orthologs would remain informative. Because the approximation we use for threshold derivations is inappropriate for small alignments, we used a score threshold of 10.49 (corresponding to $n = 12$) when analyzing six-sequence alignments. Regardless, the alignment of orthologs outperformed random alignments over a wide range of thresholds.

HIV protease variants

Data was obtained from HIVdb (Rhee et al. 2003). We considered frequency by isolate among all HIV-1 subtypes for both pre- and post-treatment analyses. The results for other configurations were similar. Consensus amino acids were omitted. We calculated the effectiveness of a position by summing the quantitative values provided for each of the seven protease inhibitors. When multiple variants at a position conferred resistance, the most effective variant was used to establish rank.

Disease variants

Control distributions for disease-associated proteins were formed from the impact scores of all achievable single missense variants. For the initial HBB analysis (Fig. 4A), we restricted attention to variants implicated in hemolytic anemia for comparison with the conditions associated with PK and G6PD; results were similar using all HBB anemia variants (Fig 4B).

Release R8 of the IARC TP53 Database (Olivier et al. 2002) contains 13,509 single missense variants of p53 isolated from somatic tumors; 1117 distinct variants occur at least once. Only 2318 p53 variants can result from a single missense mutation in human p53, suggesting that the database is nearing saturation. We therefore took those variants absent from the database to reasonably approximate a class of functional proteins.

Acknowledgments

We thank Gregory Cooper and Midori Hosobuchi for comments on the manuscript, Jonathan Binkley for supplying alignments and crystal structures, Roberta Kwok for the Java implementation of MAPP, Belinda Giardine for help with the hemoglobin data-

base, and members of the Sidow laboratory for helpful discussions. This manuscript benefited from the helpful comments of three reviewers. E.A.S. is supported by the Stanford Genome Training Program (NIH/NHGRI); A.S., by grants from NIH/NHGRI.

References

Altschul, S.F., Carroll, R.J., and Lipman, D.J. 1989. Weights for data related by a tree. *J. Mol. Biol.* **207:** 647–653.

Botstein, D. and Risch, N. 2003. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33(Suppl):** 228–237.

Cai, Z., Tsung, E.F., Marinescu, V.D., Ramoni, M.F., Riva, A., and Kohane, I.S. 2004. Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum. Mutat.* **24:** 178–184.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. 1994. Crystal structure of a p53 tumor suppressor–DNA complex: Understanding tumorigenic mutations. *Science* **265:** 346–355.

Coffin, J.M. 1995. HIV population dynamics in vivo: Implications for genetic variation, pathogenesis, and therapy. *Science* **267:** 483–488.

DeLano, W.L. 2002. *The PyMOL user's manual*. DeLano Scientific, San Carlos, CA.

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. 2002. A structural EM algorithm for phylogenetic inference. *J. Computat. Biol.* **9:** 331–353.

Hardison, R.C., Chui, D.H.K., Giardine, B., Riemer, C., Patrinos, G.P., Anagnou, N., Miller, W., and Wajcman, H. 2002. HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Hum. Mutat.* **19:** 225–233.

Huisman, T.H.J., Carver, M.F.H., and Efremov, G.D. 1996. *A syllabus of human hemoglobin variants*. The Sickle Cell Anemia Foundation, Augusta, GA.

Krishnan, V.G. and Westhead, D.R. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* **19:** 2199–2209.

Kwok, C.J., Martin, A.C., Au, S.W., and Lam, V.M. 2002. G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. *Hum. Mutat.* **19:** 217–224.

Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.

Lau, A.Y. and Chasman, D.I. 2004. Functional classification of proteins and protein variants. *Proc. Natl. Acad. Sci.* **101:** 6576–6581.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271:** 1247–1254.

Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.

Loeb, D.D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S.E., and Hutchison III, C.A. 1989. Complete mutagenesis of the HIV-1 protease. *Nature* **340:** 397–400.

Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S., and Miller, J.H. 1994. Genetic studies of the *lac* repressor, XIV: Analysis of 4000 altered *Escherichia coli lac* repressors reveals essential and non-essential residues, as well as "spacers' which do not require a specific sequence. *J. Mol. Biol.* **240:** 421–434.

Miller, M.P. and Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10:** 2319–2328.

Mooney, S.D. and Klein, T.E. 2002. The functional importance of disease-associated mutation. *BMC Bioinformatics* **3:** 24.

Muñoz, V. and Serrano, L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: Comparison with experimental scales. *Proteins* **20:** 301–311.

Ng, P.C. and Henikoff, S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* **11:** 863–874.

———. 2002. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12:** 436–446.

Olivier, M., Eeles, R., Hollstein, M., Khan, M.A., Harris, C.C., and Hainaut, P. 2002. The IARC TP53 database: New online mutation analysis and recommendations to users. *Hum. Mutat.* **19:** 607–614.

Ramensky, V., Bork, P., and Sunyaev, S. 2002. Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.* **30:** 3894–3900.

Rennell, D., Bouvier, S.E., Hardy, L.W., and Poteete, A.R. 1991. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222:** 67–88.

Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., and Shafer, R.W. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31:** 298–303.

Simon, A.L., Stone, E.A., and Sidow, A. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc. Natl. Acad. Sci.* **99:** 2912–2917.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. 2003. Human Gene Mutation Database (HGMD®): 2003 update. *Hum. Mutat.* **21:** 577–581.

Stone, E.A. 2004. "Statistical advances in interspecific data analysis." Ph.D. dissertation, Stanford University, Stanford, CA.

Stryer, L. 1995. *Biochemistry*, 4th ed. W.H. Freeman, New York.

Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. 1996. Genetic studies of the lac repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol.* **261:** 509–522.

Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A.S., and Bork, P. 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10:** 591–597.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Vulliamy, T., Beutler, E., and Luzzatto, L. 1993. Variants of glucose-6-phosphate dehydrogenase are due to missense mutations spread throughout the coding region of the gene. *Hum. Mutat.* **2:** 159–167.

Walker, D.R., Bond, J.P., Tarone, R.E., Harris, C.C., Makalowski, W., Boguski, M.S., and Greenblatt, M.S. 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: Correlation with p53 protein structural and functional features. *Oncogene* **19:** 211–218.

Wrobel, J.A., Chao, S.-F., Conrad, M.J., Merker, J.D., Swanstrom, R., Pielak, G.J., and Hutchison III, C.A. 1995. A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. *Proc. Natl. Acad. Sci.* **95:** 638–645.

Zamyatin, A.A. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* **24:** 107–123.

Received February 7, 2005; accepted in revised form April 21, 2005.