Understanding human disease mutations through the use of interspecific genetic variation

Mark P. Miller and Sudhir Kumar*

Department of Biology, Arizona State University, Tempe, AZ 85287-1501, USA

Received June 5, 2001; Revised and Accepted July 31, 2001

Data on replacement mutations in genes of disease patients exist in a variety of online resources. In addition, genome sequencing projects and individual gene sequencing efforts have led to the identification of disease gene homologs in diverse metazoan species. The availability of these two types of information provides unique opportunities to investigate factors that are important in the development of genetically based disease by contrasting long and short-term molecular evolutionary patterns. Therefore, we conducted an analysis of disease-associated human genetic variation for seven disease genes: the cystic fibrosis transmembrane conductance regulator, glucose-6-phosphate dehydrogenase, the neural cell adhesion molecule L1, phenylalanine hydroxylase, paired box 6, the X-linked retinoschisis gene and *TSC2*/tuberin. Our analyses indicate that disease mutations show definite patterns when examined from an evolutionary perspective. Human replacement mutations resulting in disease are overabundant at amino acid positions most conserved throughout the long-term history of metazoans. In contrast, human polymorphic replacement mutations and silent mutations are randomly distributed across sites with respect to the level of conservation of amino acid sites within genes. Furthermore, disease-causing amino acid changes are of types usually not observed among species. Using Grantham's chemical difference matrix, we find that amino acid changes observed in disease patients are far more radical than the variation found among species and in non-diseased humans. Overall, our results demonstrate the usefulness of evolutionary analyses for understanding patterns of human disease mutations and underscore the biomedical significance of sequence data currently being generated from various model organism genome sequencing projects.

INTRODUCTION

One central purpose of genome sequencing projects is to effect a better understanding of the genetics of disease and provide assistance with the identification of disease-associated genes (1–3). However, many human mutation databases containing genetic variation found in disease patients already exist, and new databases and database entries are rapidly accumulating (4,5). Concomitant analysis of these two types of information provides unique opportunities to identify intrinsic attributes of disease-associated human genetic variation, leading to a better understanding of the relationship between mutations and the development of disease phenotypes.

Information contained in the alignments of homologous disease-associated genes has long been recognized as an important factor for understanding contemporary deleterious genetic variation in humans (4,6). For example, in a given set of homologous genes, a large fraction of amino acid sites will be conserved even among distantly related species that diverged hundreds of millions of years ago. Variations that arose at such positions throughout evolutionary history have evidently been under strong purifying selection and eliminated from populations, suggesting that the existing amino acid residues at invariant positions are critical for proper gene function. Thus, information from interspecific alignments can indicate amino acid residues in gene products that are likely to produce disease if mutated in humans. Likewise, some positions in protein sequences vary among species, and such variable sites may indicate positions that are under less severe selective constraints. These variable positions suggest sites where residue changes can be tolerated by natural selection and provide insights into the types of amino acids that can be freely exchanged without negatively impacting protein function.

Since the logic of these statements is often used by researchers to indicate the potential for an observed amino acid change to produce disease in humans (6–10), we conducted a study to directly evaluate the extent that interspecific sequence alignments reveal common attributes of the deleterious mutations observed in humans. We performed three types of analyses using disease mutation data and homologous gene sequences from seven disease-associated genes (Table 1 and Fig. 1): cystic fibrosis transmembrane conductance regulator (*CFTR*), glucose-6-phosphate dehydrogenase (*G6PD*), neural cell adhesion molecule L1 (*L1CAM*), phenylalanine hydroxylase (*PAH*), paired box 6 (*PAX6*), the X-linked retinoschisis gene

*To whom correspondence should be addressed. Tel: +1 480 727 6949; Fax: +1 480 965 2519; Email: s.kumar@asu.edu
Present address:
Mark P. Miller, Department of Fisheries and Wildlife, Utah State University, Logan, UT 84322-5210, USA

Table 1. Disease genes, database web sites and numbers of mutations analyzed from each database

Gene	Web-address for database (reference no.)	Number of mutations analyzed (disease/polymorphic/silent)[a]
CFTR	www.genet.sickkids.on.ca/cftr	429/32/61
G6PD	http://rialto.com/favism/mutat.htm (32)	110/–/–[b]
L1CAM	dnalab-www.uia.ac.be/dnalab/l1/ (37)	48/–/–
PAH	http://www.mcgill.ca/pahdb/ (38)	270/–/–
PAX6	www.hgu.mrc.ac.uk/Softdata/PAX6/ (39)	29/–/–
RS1	www.dmd.nl/rs/rs.html	71/–/–
TSC2	expmed.bwh.harvard.edu/ts	47/18/33

[a]Disease mutations refer to those amino acid changes that produce a disease phenotype. Polymorphic mutations are amino acid changes that are presumably not disease related. Silent mutations are DNA sequence changes that do not alter the encoded amino acid.

[b]The database analyzed contained 48 type I mutations that result in chronic non-spherocytic hemolytic anemic and 62 less severe types II, III or IV mutations.

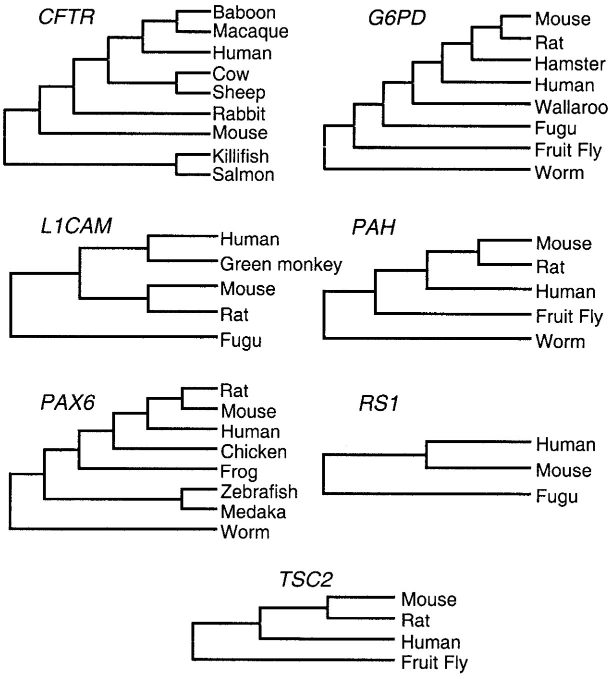

Figure 1. Evolutionary relationships used to determine the number of mutational events that have occurred at each amino acid site in each gene throughout evolutionary history. GenBank accession nos for sequences used in analyses are as follows: *CFTR*, NM_000492, AF162401, AF013753, M76128, U20418, U40227, M69298, AF000271, AF155237; *G6PD*, NM_000402, Z11911, NM_017006, AF044676, U13899, X83611, AH002543, Z73102; *L1CAM*, XM010169, AF129167, X12875, X59149, AF026198; *PAH*, K03020, X51942, NM_012619, M32802, AF119388; *PAX6*, 12736585, X63963, NM_013001, D87837, U77532, AF061252, AJ000938, U31537; *RS1*, AF014459, NM_011302, AF146687; *TSC2*: X75621, NM_011647, D50413, AF172995.

and a gene associated with tuberous sclerosis (*TSC2*). First, we determined the association between the prevalence of disease mutations and the extent to which corresponding amino acid sites in other species have been conserved throughout the evolutionary history of metazoans. Secondly, we compared the frequency of a given type of amino acid change in disease patients to frequencies obtained from interspecific comparisons. Finally, we compared the chemical property differences of amino acid changes seen among species and non-diseased humans with those observed in disease patients.

RESULTS AND DISCUSSION

The association of disease mutations and evolutionarily conserved amino acid residues

A null hypothesis describing the distribution of human genetic variation among amino acid sites in a gene can be generated assuming that point mutations occur randomly throughout that gene. If a set of mutations found in a population is representative of the random mutational process, then the number of mutations observed at a given type of site in a gene should be proportional to the frequency with which sites of that type appear in a sequence. Using information from interspecific comparisons, we tested the null hypothesis that disease-associated replacement mutations are randomly distributed among different classes of amino acid sites which were determined based on their variability among extant metazoans. This analysis permits a direct assessment of statements suggesting that disease mutations are more common at evolutionarily conserved residues. If we do not reject the null hypothesis of random association for a set of disease mutations, then mutations at conserved sites are no more important than those at variable sites for the development of the disease phenotype. In contrast, analyses will illustrate the importance of replacement mutations at conserved sites if the null hypothesis is rejected due to an overabundance of disease mutations at conserved positions and a deficiency at variable amino acid sites.

This analysis was performed for the sets of mutations obtained from each of the seven disease mutation databases examined. Type I and types II, III and IV mutations found in *G6PD* were analyzed in separate data sets. For comparison purposes, we also tested the sets of polymorphic replacement mutations and silent mutations obtained from the *CFTR* and *TSC2* databases. Our analyses revealed that the mutations observed in disease patients are not randomly distributed across sites in genes. Instead, with the exception of the analysis of less severe types II, III and IV mutations to *G6PD*, significantly more mutations were observed at invariant amino acid positions than expected by chance alone (Fig. 2). All other sites in the genes examined, even those that have experienced a single amino acid substitution over evolutionary history, showed reduced numbers of disease mutations. Polymorphic human replacement mutations and silent mutations in *CFTR* and *TSC2* were distributed at random with respect to the level of conservation of amino acid sites in the gene (Fig. 3), as were the types II, III and IV mutations to *G6PD* (Fig. 2).

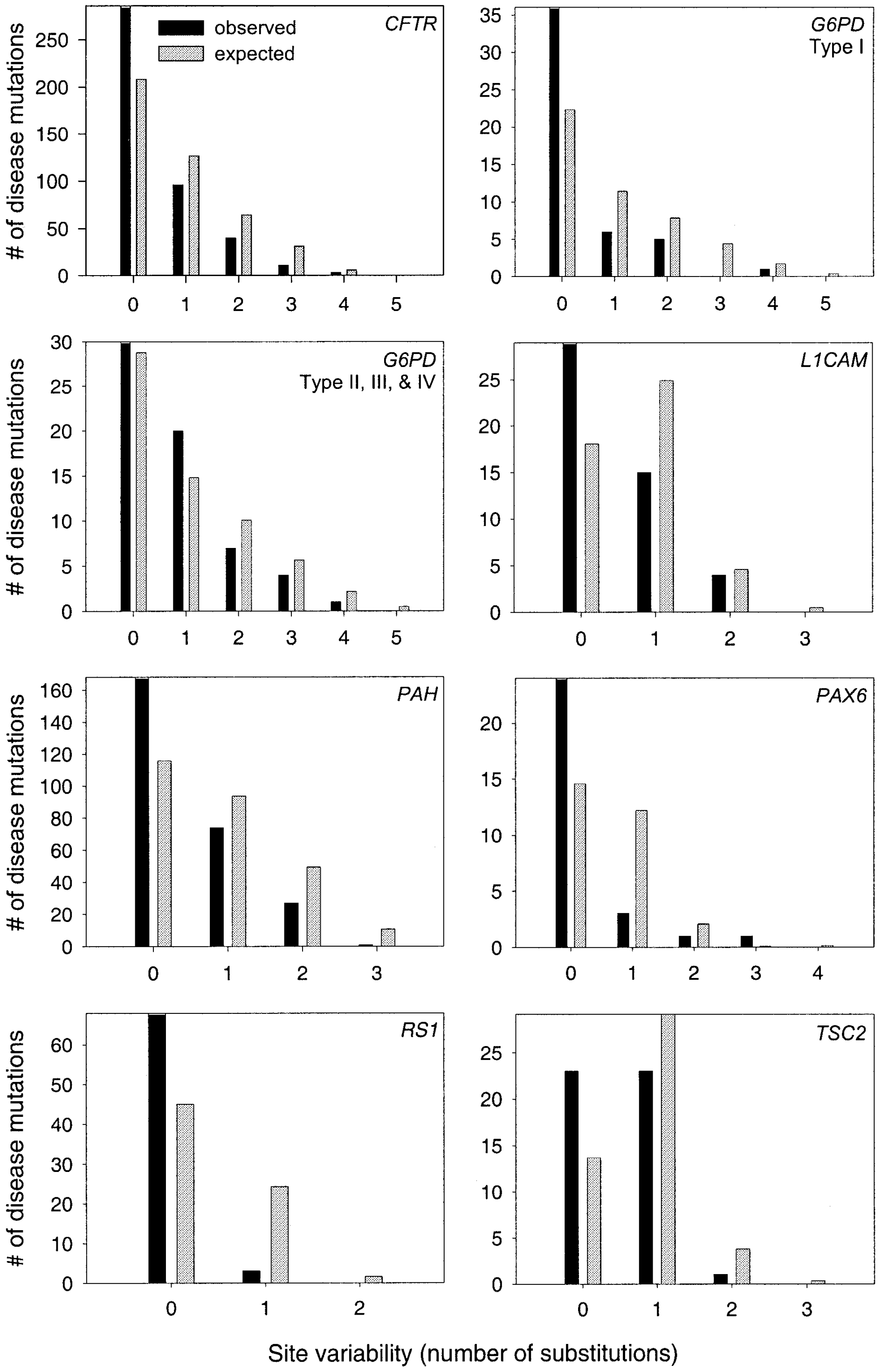

Figure 2. Observed (black bars) and expected (gray bars) numbers of disease-causing replacement mutations at amino acid sites that have experienced different numbers of substitutions among species. With the exception of types II, III and IV mutations in *G6PD*, all of the remaining genes display significant deviations from random expectations. *CFTR*, $\chi^2_{(5\,\mathrm{df})} = 60.50$, $P < 0.001$; *G6PD* type I mutations, $\chi^2_{(5\,\mathrm{df})} = 17.11$, $P < 0.005$; *G6PD* types II, III and IV mutations, $\chi^2_{(5\,\mathrm{df})} = 4.43$, $P = 0.489$; *L1CAM*, $\chi^2_{(3\,\mathrm{df})} = 11.09$, $P < 0.005$; *PAH*, $\chi^2_{(3\,\mathrm{df})} = 46.74$, $P < 0.001$; *PAX6*: $\chi^2_{(4\,\mathrm{df})} = 26.31$, $P < 0.001$; *RS1*, $\chi^2_{(2\,\mathrm{df})} = 32.10$, $P < 0.001$; *TSC2*, $\chi^2_{(3\,\mathrm{df})} = 10.13$, $P < 0.05$.

This overall pattern illustrates that disease-associated mutations are only a subset of the full complement of random mutations that may occur in a gene. These subsets mainly encompass mutations at conserved residues and reflect the evolutionary constraints placed on coding regions of genes, as the proper function of a given gene product likely requires specific amino acids at certain positions of the protein. Mutations to those residues have severe phenotypic consequences,

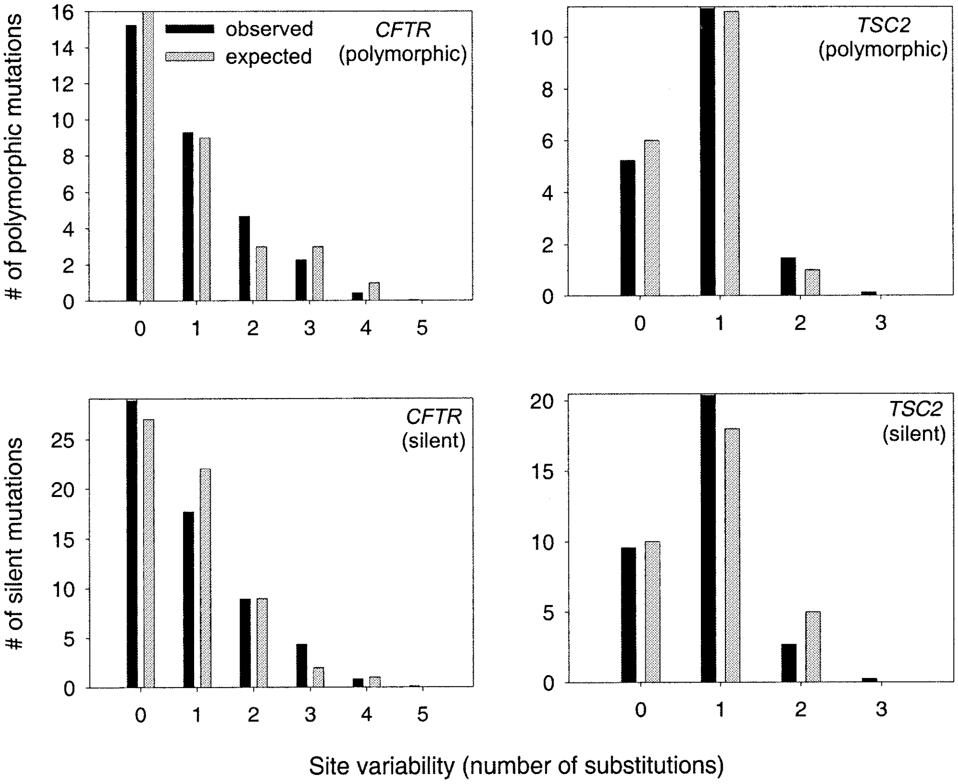

Figure 3. Observed (black bars) and expected (gray bars) numbers of polymorphic replacement and silent mutations observed in *CFTR* and *TSC2* databases at sites that have experienced different numbers of substitutions among species. Polymorphic *CFTR* mutations, $\chi^2_{(5\text{ df})} = 1.68$, $P = 0.89$; polymorphic *TSC2* mutations, $\chi^2_{(3\text{ df})} = 0.39$, $P = 0.94$; silent *CFTR* mutations, $\chi^2_{(5\text{ df})} = 2.56$, $P = 0.76$; silent *TSC2* mutations, $\chi^2_{(3\text{ df})} = 2.55$, $P = 0.46$.

thus bringing the patients in question to the attention of clinicians. Our analyses also revealed in some instances (for example, *CFTR*, *L1CAM*, *RS1*, and *TSC2*) relatively large numbers of disease mutations at sites that vary among species (Fig. 2). However, the vast majority of the disease mutations found at variable alignment positions occur at sites that have undergone single substitutions in the phylogenetic lineages least related to humans. For example, of the six type I *G6PD* disease mutations observed at sites that have changed once among the species analyzed, five occur at positions that have mutated in fish, *Drosophila* or *Caenorhabditis elegans*. Likewise, we observed 79% of the *CFTR* disease mutations, 93% of the *L1CAM* mutations, and all of the *RS1* mutations in once-changed sites at alignment positions that differ from humans only in fish lineages. Similarly, of the *TSC2*, *PAH* and *PAX6* disease mutations found at once-changed sites, 91, 99 and 100%, respectively, occur at positions that vary only in the invertebrate sequences analyzed. Therefore, the majority of the amino acids that can produce disease mutations are conserved, at least among mammals. While the inclusion of distantly related species in analyses may seem to provide counterintuitive results (and result in the observation of disease mutations at variable sites), it is important to note that from a statistical perspective, the amino acid variation introduced through the inclusion of distantly related species is critical if one is to address the influence of evolutionary history on contemporary patterns of disease mutations. Without the genetic variation observed from the inclusion of non-mammalian species, insufficient information is available to generate counts of disease mutations at non-conserved sites for analysis purposes. Thus, we recommend using as many orthologous gene sequences as possible from both closely and distantly related metazoan species to better understand the degree with which gene sites have been conserved throughout evolutionary history.

We did not reject the null hypothesis in our analyses of polymorphic replacement and silent mutations in *CFTR* and *TSC2*, indicating that these sets of mutations were randomly distributed with respect to the level of conservation of amino acid sites (Fig. 3). However, it should be noted that the biological relevance of this pattern is not equivalent for each mutation type. In the case of silent mutations, the random association of mutations with sites is a reflection of their selectively neutral nature. Silent mutations may occur anywhere in a coding region without affecting the protein product and, barring any significant codon usage bias (11,12), should show minimal effects on an organism's phenotype. However, the random distribution of polymorphic replacement mutations may indicate the presence of slightly deleterious amino acid changes in these genes. Since invariant positions in a sequence alignment suggest positions that are critical for proper protein function, only variable sites indicate positions where residue changes are likely tolerated due to relaxed selection constraints. Recent studies of *CFTR* have in fact revealed mutations that, while not responsible for the development of cystic fibrosis, are instead putatively associated with less severe disease phenotypes such as asthma, disseminated brochiectasis or chronic obstructive pulmonary disease (9,10,13,14). Several of the mutations reported in patients of these diseases were found in our list of polymorphic amino acid changes obtained from the cystic fibrosis mutation database, and many of these mutations occur at fully conserved alignment positions. Thus, the random distribution of sets of *CFTR* and *TSC2* polymorphic replacement mutations may suggest the presence of recently evolved random replacement mutations that have relatively mild, albeit deleterious, phenotypic consequences. Perhaps only those mutations found at variable positions are truly selectively neutral. The remainder, specifically those found at conserved sites, may be eliminated by natural selection over long-term evolutionary history.

Based on the calculation of amino acid sequence alignment profile scores, among other criteria, Sunyaev *et al.* (15) predicted that ~20% of the non-synonymous single nucleotide polymorphisms (SNPs) in humans are to some extent deleteriou. In their study, alignment scores quantified the agreement of amino acids found at each site within a set of homologous sequences. Any observed human amino acid that changed the site-specific score beyond a set threshold was interpreted to be damaging to the protein. Likewise, Notaro *et al* (16). compiled all available sequence data from bacteria to vertebrates and plants for *G6PD*. Based on the calculation of amino acid similarities of other extant organisms with humans, they concluded that more mutations than expected are found at moderate to highly conserved sites; however, fewer mutations were seen at fully conserved residues among the 52 species analyzed. The authors suggested that mutations at fully conserved sites may be lethal, thus accounting for their relative absence in data sets, while those at moderately conserved sites are more likely to produce the observed disease phenotype in patients. If such an interpretation is correct, then we overestimated the number of fully conserved sites among species by restricting our analyses to metazoans. However, including such a wide variety of species in analyses can be problematic, as amino acid variation observed among ancient phylogenetic lineages may be correlated with functional differences in gene products. Thus, the inclusion of such distantly related species in analyses may inappropriately indicate amino acid variation at sites that are critical to the function of the gene in humans and result in an underestimate of the number of fully conserved sites within metazoan or mammalian lineages.

While the interspecific comparison approaches of Notaro *et al.* (16) and Sunyaev *et al.* (15) address the prevalence of deleterious mutations at evolutionary conserved sites, we note that both suffer technical problems relative to our approach of estimating numbers of substitution events via phylogenetic analyses. First, the similarity based approach (16) will overestimate the variability of a given site because an identical residue may appear in multiple species of a phylogenetic lineage, indicating that it also was present in the common ancestor. This means that a single substitution at a given site may be counted multiple times depending on the species sampled for analyses. It has long been recognized that species cannot be treated as independent observations for use in statistical analyses (17), as groups of species within phylogenetic lineages share common attributes (such as amino acid sequences) due to their common ancestry. However, our approach to quantifying interspecific variation makes use of known phylogenetic relationships to naturally account for the correlated amino acid changes that appear among related organisms. In addition, the quantification of amino acid similarities tends to result in large ranges of values that need to be arbitrarily clustered to form groups for analysis purposes. Such grouping is not necessary when using tree-based statistics, as the actual count of the number of substitutions provides a natural classification (Figs 2 and 3). Secondly, when considering the alignment profile score approach (15), researchers need to establish an arbitrary criterion to eliminate closely related sequences [set at 95% similar or more in the study by Sunyaev *et al.* (15)] from analyses to avoid overly influencing the site-specific alignment score. The inclusion of excessive closely related sequences would make the site-specific scores indicate highly conserved residues when in fact the relative lack of variation is due to the inclusion of recently diverged lineages that have not had sufficient time to accumulate amino acid changes. In contrast, our phylogenetic approach permits the use of all available sequence data, even those from species closely related to humans (Materials and Methods).

Frequencies and chemical differences of amino acid changes

It is well known that the relative rates of substitution are not uniform for different pairs of amino acids. This is clearly illustrated by matrices such as PAM (18) and BLOSUM (19) that provide information on the frequency with which a given amino acid is substituted for another over the long term evolutionary history of a large number of proteins. Our analyses showed that both the type and frequency of a given amino acid change differs considerably between those observed in disease patients and those found among species. Scatterplots of the frequencies of each type of change revealed no correlation between mutations observed in disease patients and those derived from the PAM matrix or directly observed among species in our main set of seven disease genes (Fig. 4A and B). However, estimated frequencies of changes obtained from the PAM matrix and interspecific comparisons of the seven disease genes were highly correlated (r = 0.87; Fig. 4C), illustrating that both sources of data provide similar information. Figure 5 shows in more detail the differences between the amino acid substitutions observed among species and changes commonly found in disease patients. For example, in the set of amino acid changes involving glutamic acid (E), the most common substitution among species is an aspartic acid (D). The disease-associated mutation data set, in contrast, shows changes that most often involve lysine (K) (Fig. 5). Likewise, phenylalanine→tyrosine changes (F→Y) are commonly observed among species. However, this change was never seen among the 1004 disease-associated mutations analyzed. We note that these types of trends are not due to the limited sampling of disease mutations from only seven genes. To illustrate this, we computed amino acid change frequencies from an extensive collection of 10 262 disease-associated replacement mutations observed in a wide variety of disease genes. These frequencies, which were calculated from data obtained at the Human Gene Mutation Database (HGMD; 20,21), were highly correlated with disease associated amino acid change frequencies observed in our main set of seven genes (r = 0.86;

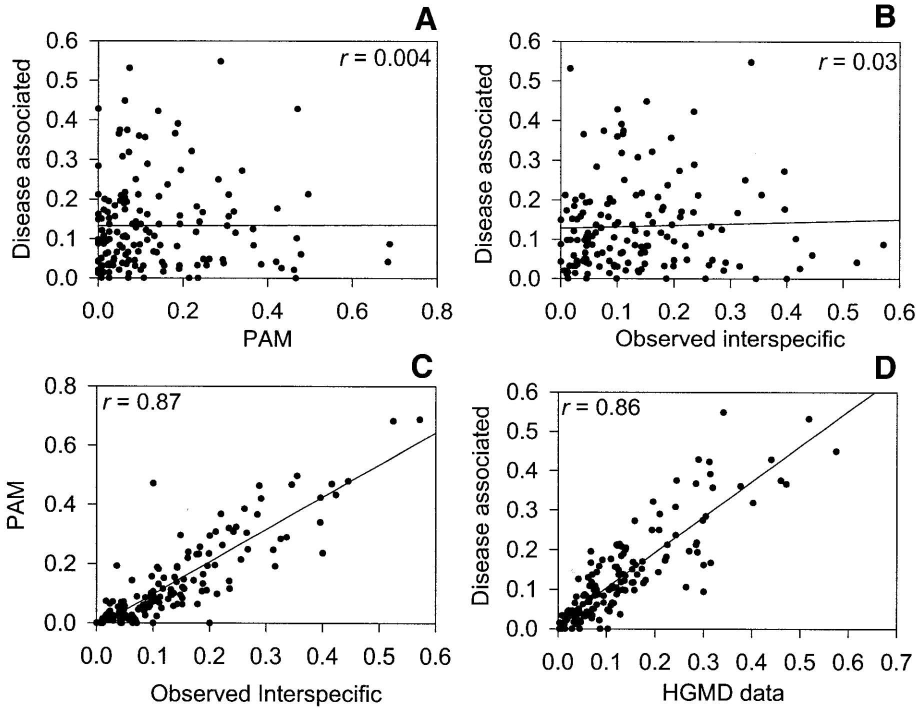

Figure 4. Scatterplots of the relative frequencies of each type of amino acid change obtained from seven disease mutation databases, interspecific comparisons of disease gene homologs, modified PAM matrix transition probabilities and data from HGMD. (**A**) Modified PAM transition matrix probabilities versus amino acid change frequencies observed in seven disease mutation databases. (**B**) Frequencies of each type of amino acid change observed among species versus amino acid change frequencies observed in seven disease mutation databases. (**C**) Frequencies of each type of amino acid change observed among species versus modified PAM matrix transition probabilities. (**D**) Frequencies of each type of amino acid change based on HGMD data versus disease associated amino acid change frequencies observed from the seven disease mutation databases examined in detail (see Discussion for more details).

Fig. 4D). Overall, this finding illustrates that there is a general pattern to the types of amino acid changes seen in disease patients and that they are not of the types accepted by natural selection over long term evolutionary history.

A logical explanation for the lack of correlation between interspecific and disease-associated amino acid substitution frequencies (Fig. 4A and B) is that differences in the chemical properties of disease-associated amino acid changes are larger than those generally tolerated by natural selection. Many researchers have noted the similar physicochemical properties of the amino acids commonly substituted among species. This has been demonstrated through the use of multiple different measures of the chemical differences of amino acids (22–24) and through analyses based on the assignment of residues to categories based on common shared characteristics (25). Studies such as these illustrate that variation in the rate of substitution of different pairs of amino acids over the long-term history of species is in part due to negative selection against amino acid changes to residues with dramatically different chemical properties. While there are many coarse classifications of amino acid residues, the single most informative residue grouping is unknown. In addition, while it is seemingly logical to label a set of possible amino acid changes as 'conservative', the alternative label of 'radical' has no quantitative component and does not accurately reflect the severity of the amino acid change. Therefore, we used the chemical difference matrix of Grantham (26), a multivariate combination of residue side chain composition, polarity and volume differences, to quantify the severity of amino acid changes.

Our analyses revealed that amino acid changes associated with the development of a disease phenotype are on average more radical (larger chemical difference) than the types of amino acid changes that are observed in interspecific comparisons of species (Table 2). In all cases, chemical differences for disease-associated mutations in humans were significantly greater than those observed for amino acid changes among species. Scores for human polymorphic amino acid changes in *CFTR* and *TSC2* were likewise significantly less than scores observed for disease-associated amino acid changes; in both cases, values for the polymorphic human variation were statistically indistinguishable from the amino acid substitutions observed among species (Table 2). In the case of *G6PD*, we found no significant differences between type I and types II, III and IV disease mutations. However, the average for the less severe mutation types was noticeably lower than that seen for the more severe type I mutations. Mean values for both disease mutation classes of *G6PD* were significantly greater than those observed for amino acid changes among species. Thus, interspecific amino acid substitutions are to some extent constrained by selection against mutated residues with dramatically different properties, and the presence of a dramatically different mutated amino acid in humans is therefore more likely to be associated with disease. Grantham's chemical difference matrix itself is negatively correlated with the relative amino acid substitution frequency matrix of

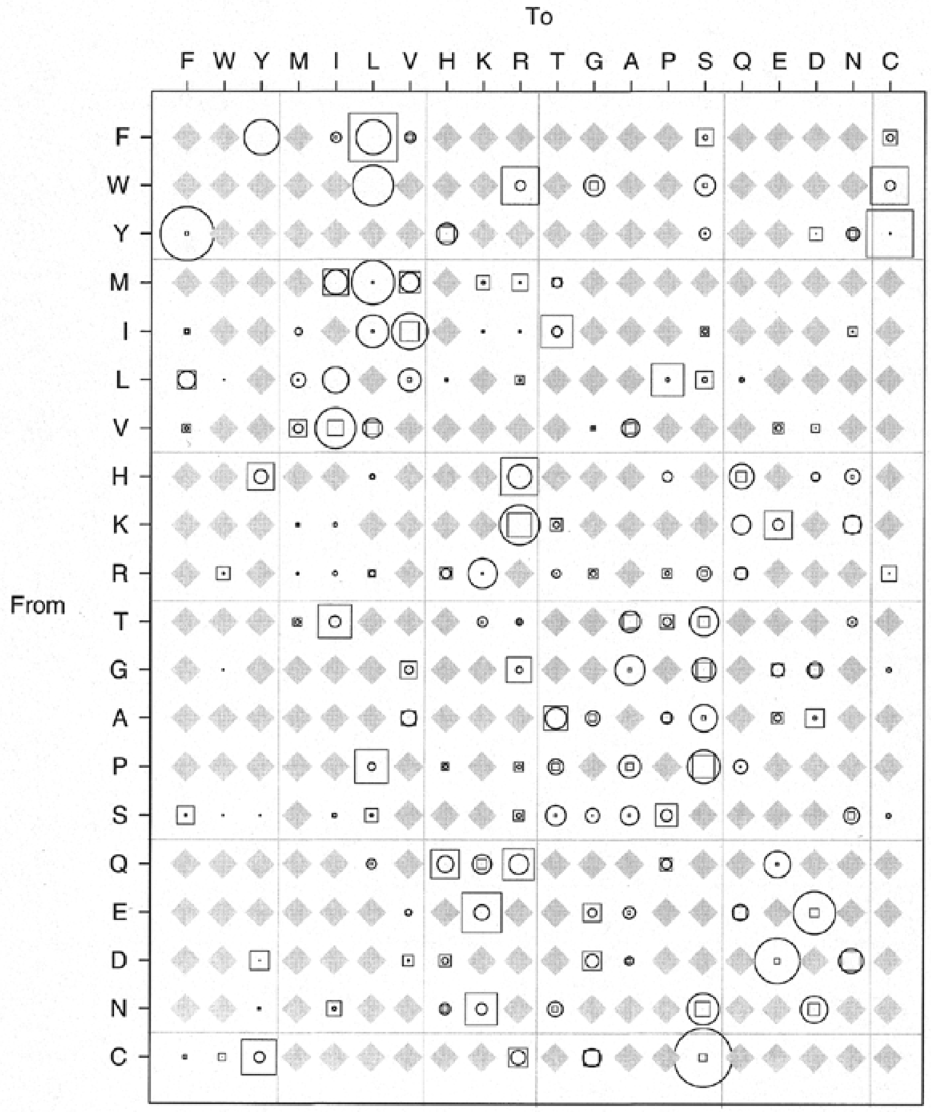

Figure 5. Plots showing the relative frequencies of amino acid changes observed in interspecific comparisons (circles) and detected in disease patients (squares). The width of the symbols is proportional to the relative frequency of a given amino acid change. Diamonds indicate amino acid changes that cannot be observed as a result of a single base mutation. Horizontal and vertical lines across the graph denote groups of amino acids with relatively similar properties, thus, residue changes to an amino acid with similar properties are found along diagonal elements of the grid.

McLachlan (27), further indicating that changes to dissimilar residues are rarely tolerated by natural selection. Likewise, Krawczak *et al.* (28) showed that Grantham's measure is correlated with the clinical observation likelihood of disease patients, indicating that individuals with mutant amino acid residues quite different from the wild-type are more likely to come to the attention of clinicians.

Although we observed no significant overabundance of less severe types II, III and IV mutations in *G6PD* at conserved sites (Fig. 2), chemical difference scores for these mutations were on average greater than those observed among species, but less severe than those associated with type I mutations. In this case, it appears that the mild phenotypic consequences of these changes are mainly due to the mutated residue's radical chemical properties relative to the wild-type rather than a tendency towards the disruption of a presumably critical amino acid residue. Miyata *et al.* (24) have also demonstrated that the chemical differences of human hemoglobin mutations resulting in severe and mild hemolytic anemia were greater than those expected by chance, and average differences for mutations producing severe phenotypes were likewise greater than those that produce milder forms.

Polymorphic amino acid changes in *CFTR* and *TSC2* were commonly observed, albeit not in excess, at invariant sites in the sequence alignments (Fig. 3). While information from the alignments suggests that such mutations are deleterious, the chemical difference scores for polymorphic mutations in these data sets are on average no different from those observed among species (Table 2), indicating that the types of amino acid changes seen in these non-diseased individuals should be functionally compatible with the structure of the protein.

Table 2. Average chemical differences of amino acid changes in disease-associated mutations, polymorphic replacement mutations and mutations observed in interspecific comparisons of humans and other metazoan species

	Average chemical difference (SE)		
Gene	Disease[a]	Polymorphic	Interspecific
CFTR[b]	88.53 (2.26)	68.84 (9.96)	55.82 (1.17)
TSC2[b]	91.32 (7.66)	56.22 (8.26)	60.53 (1.30)
G6PD[c]	95.13[d] (8.54)	–	54.58 (1.89)
	78.23[e] (6.29)		
L1CAM	111.40 (7.49)	–	61.08 (1.45)
PAH	86.33 (3.00)	–	59.01 (2.28)
RS1	109.27 (7.05)	–	51.87 (4.00)
PAX6	84.55 (9.12)	–	56.84 (2.20)

[a] For each gene, there were significant overall differences among means. Results of Kruskal–Wallace tests for each gene. *CFTR*, $\chi^2_{(2\ \mathrm{df})} = 142.3$, $P < 0.001$; *TSC2*, $\chi^2_{(2\ \mathrm{df})} = 14.01$, $P < 0.001$; *G6PD*, $\chi^2_{(2\ \mathrm{df})} = 27.20$, $P < 0.001$; *L1CAM*, $\chi^2_{(1\ \mathrm{df})} = 40.67$, $P < 0.001$; *PAH*, $\chi^2_{(1\ \mathrm{df})} = 37.68$, $P < 0.001$; *RS1*, $\chi^2_{(1\ \mathrm{df})} = 32.4$, $P < 0.001$; *PAX6*, $\chi^2_{(1\ \mathrm{df})} = 5.64$, $P = 0.018$.

[b] Mann–Whitney U tests indicate that average chemical differences for disease mutations are significantly greater than those for polymorphic amino acid changes or amino acid differences observed among species. Scores for human polymorphic and interspecific amino acid changes are not significantly different.

[c] Mann–Whitney U tests indicate that average chemical difference scores for type I and types II, III and IV mutations for *G6PD* are not significantly different, however, scores for both sets of disease mutations are significantly greater than scores for amino acid differences observed among species.

[d] Type 1 mutation.

[e] Types II, III and IV mutations.

Indeed, 14 of the polymorphic *CFTR* mutations and eight of the polymorphic *TSC2* mutations observed at invariant or once-changed sites have chemical difference scores <50, well below the average score for amino acid substitutions observed among species. Thus, in some instances, the 'conserved' sites where human polymorphism is observed may indicate less critical residues that, by chance alone, appear to be invariant based on the reference sequences used in analyses. These positions could likewise vary among unanalyzed individuals of the other species or vary among species that were not analyzed. In comparison, a total of seven polymorphic *CFTR* and two polymorphic *TSC2* mutations from our data sets have chemical difference scores >100. Perhaps, of the polymorphic variation analyzed, it is the mutations with the most radical amino acid changes that are most likely to be removed from populations over long term evolutionary history.

CONCLUSIONS

We have shown that disease-associated amino acid changes in humans differ from human polymorphic changes or variation seen among species in several ways. First, disease-associated amino acid changes are overabundant at conserved residues, thus illustrating the critical importance of such sites for proper gene function. Secondly, the types of amino acid changes producing disease are not likely to be similar to the types of amino acid substitutions commonly observed among species. Chemical differences of these uncommon disease-associated amino acid changes are more radical than the commonly encountered polymorphic amino acid variation found in humans or permitted by natural selection throughout evolutionary history. In some respects, the analyses conducted here are analogous to the many commonly employed procedures that use patterns of within and among species variation to elucidate the relative importance of neutral versus selective (purifying or positive) processes in the evolution of genes (29–31). In our case, the comparison of intraspecific disease-associated genetic variation and variation observed among species reveals several common trends and shows that comparative molecular analyses are directly applicable to the deleterious genetic variation found in humans. Thus, researchers should strive to obtain homologous gene sequences from both closely and distantly related species to obtain an evolutionary perspective towards the level of conservation of amino acid sites in disease associated genes and the types of amino acid changes that have been permitted in a given gene over time. The identification of disease gene homologues from the annotated genomes of organisms such as *Drosophila*, *C.elegans*, *Takifugu*, mouse and rat, will likely provide this evolutionary context for a wide variety of human genetic disorders.

MATERIALS AND METHODS

Data acquisition

Using available information from online databases, we obtained data for single base pair replacement mutations observed in seven disease-associated genes (Table 1). Each entry in the databases contained information on a nucleotide position, a reference nucleotide typically observed in humans at that position, and the nucleotide residue observed at that position in a disease patient. Based on nucleotide position information, we excluded all point mutations that occurred in non-coding regions, termination codons or those mutations from the database whose reference nucleotide could not be reconciled with a reference human gene sequence obtained from GenBank. Not all of the databases examined contained information on the observation frequencies of a given mutation. Therefore we included each mutation only once in our analyses to ensure that comparable data were present for each disease gene. In addition, the elimination of frequency information prevented bias in our results towards the properties of commonly observed mutations over those less frequently reported for a specific genetic disease. *CFTR* and *TSC2* databases also contained information on 'polymorphic' replacement (presumably non-disease associated) and silent mutations in sufficient quantities for analysis. In the case of G6PD, information on the severity of the disease phenotype was also available (32), which permitted us to analyze severe (type I) mutations resulting in chronic non-spherocytic hemolytic anemia and less severe (types II, III and IV) mutations resulting only in enzyme deficiencies separately. In total, we examined 1004 disease-associated replacement mutations, 50 polymorphic replacement mutations and 94 silent mutations (Table 1). We further obtained complete GenBank reference cDNA sequences for human and other metazoan species for each of the disease-associated genes (Fig. 1). cDNA sequences were aligned at the amino acid level using Clustal X (33).

Determining the association between disease mutations and evolutionarily conserved sites

We performed a test to determine whether human disease mutations are more common at evolutionarily conserved amino acids than expected by chance. Interspecific variability (i.e. degree of conservation) was quantified by calculating the minimum number of amino acid substitutions that have occurred at each amino acid site throughout the evolutionary history of the given gene. This process was performed for each position using the algorithm of Fitch (34) assuming the tree topologies given in Figure 1. Alignment gaps in the human reference sequences were excluded from analyses since they provide no information about mutations that may occur in disease patients. However, human disease mutations were in a few instances observed at alignment positions containing gaps in some of the other species analyzed. In these cases, alignment gaps were treated as 21st amino acid states. Overall, this parsimony procedure provides information on the relative extent that amino acid variation has been tolerated at different sites in each gene while simultaneously accounting for shared amino acids observed within phylogenetic lineages. Ideally, parsimony is best suited for the analysis of slow evolving genes where there is no need to account for multiple hits. In other instances, we recommend the use of maximum likelihood (ML) methods in conjunction with the known phylogenetic tree to obtain counts of numbers of substitutions at each amino acid site. In this case we estimate r_j, the relative evolutionary rate of site j, under a gamma distribution of rate heterogeneity (35) and a Poisson (or PAM) model of amino acid substitution, and rescale each r_j such that $\Sigma\, r_j = 1$. Given the total number of substitutions per site of the tree (S) and the sequence length (L), the estimated number of substitutions at site j is $S_j = r_j \times S \times L$ (35). For the genes examined here, estimated counts derived from the parsimony procedure described above and from ML-based analyses as implemented in PAML (36) were highly correlated ($r \geq 0.987$ for all genes). Thus, we rely on the parsimony-based results in the analyses that follow, as the estimates are discrete integer values that provide a convenient means of classifying individual sites in genes based on their variability.

If n_i is the observed number of amino acid sites in the alignment that have undergone i substitutions throughout evolutionary history and N is the total number of amino acid sites in the gene, then under the null hypothesis of random association we expect to observe the fraction n_i/N of the human mutations in a sample at sites that have undergone i substitutions. Thus, the expected number of human mutations at sites of type i is

$$D_i^{\mathrm{exp}} = (n_i / N) \times D, \qquad (1)$$

where D is the total number of mutations in the data set, given by $D = \Sigma_0^i D_i^{\mathrm{obs}}$, and D_i^{obs} is the observed number of human mutations at sites that have experienced i substitutions. Deviations of observed values from expectations for each gene can be evaluated through the use of χ^2 tests with i degrees of freedom where

$$\chi^2 = \Sigma_0^i (D_i^{\mathrm{obs}} - D_i^{\mathrm{exp}})^2 / D_i^{\mathrm{exp}}. \qquad (2)$$

Frequencies of different amino acid changes in disease-associated mutations and interspecific comparisons

We generated four data sets to explore the null hypothesis that frequencies of different types of amino acid changes associated with disease are comparable to those observed among species. We first tabulated the frequencies of all 1004 disease-associated amino acid changes seen in the seven databases analyzed. Based on the observed counts N_{ij} of mutations from residues i to j, we calculated the relative frequency of mutations from residues i to j as

$$M_{ij} = N_{ij} / \sum_k N_{ik}, \qquad (3)$$

where the summation involves only the set of k amino acid changes from residue i that are possible as a result of a single nucleotide mutation. Secondly, we calculated M_{ij}'s for amino acid substitutions observed among species based on the combined analysis of the interspecific alignments for each of the seven disease genes. To keep frequency data comparable to those found in the disease mutation data set, only amino acid changes that could arise via single nucleotide substitutions were considered in analyses. Furthermore, each amino acid change from residue i to j at a given site in the alignment was counted only once to account for the common ancestry of residues within phylogenetic lineages. For the third data set, we modified the PAM-1 matrix (18) to obtain information on accepted point mutation frequencies from more generalized sets of homologous proteins. The PAM matrix provides instantaneous transition probabilities for all possible pairs of amino acid substitutions. Thus, we treated the probability of change from amino acid i to amino acid j as a frequency and recalculated new relative frequencies as described above, again considering only the set of amino acid changes that can result from a single base pair substitution. Finally, we used equation 3 to compute M_{ij}'s for an extensive collection of 10 262 disease-associated replacement mutations obtained from the HGMD (20,21). This information permitted us to evaluate the extent to which disease mutations observed in a large set of disease genes are comparable to those observed in our main set of seven genes. To better understand differences in the types of amino acid changes seen among disease patients and among species, we generated simple visual depictions of the frequencies of each type of amino acid change to compare disease data to interspecific derived data. Likewise, we constructed scatterplots to evaluate the correlation of the frequencies of each type of amino acid change among data sets.

Characterizing amino acid property differences

Using Grantham's chemical difference matrix (26), we tested the null hypothesis of no difference in chemical properties among disease-associated amino acid changes, human polymorphic amino acid variation and amino acid variation observed among species for each disease gene. Because our human mutation data sets contained information on single nucleotide mutations, only a subset of the 380 possible amino acid changes could have been observed. Thus, as with our analysis of the frequencies of different amino acid changes, we only analyzed amino acid changes among species that could have been the result of a single nucleotide mutation and scored each type of amino acid change seen at a site once to account for the residue's common ancestry within a phylogenetic lineage. In the cases of *CFTR* and *TSC2*, we further analyzed the observed polymorphic amino acid changes in humans. From the chemical distance scores, we used a non-parametric Kruskal–Wallace test to determine if significant overall differences existed among the sets of human and interspecific amino acid changes. Scores for *CFTR* and *TSC2* were further subjected to Mann–Whitney U tests post hoc to determine where the significant differences lie among the disease associated, human polymorphic, and interspecific scores. This same procedure was used for comparisons of type I mutations, types II, III and IV mutations and interspecific mutation scores in *G6PD*. These analyses were conducted using SPSS 10.0 (SPSS Inc, 1999).

ACKNOWLEDGEMENTS

We thank Drs Georgia Floyd-Smith, Chris Lorson, Michael Rosenberg, Sudhindra Gadagkar, David Cutler and two anonymous reviewers for comments on this manuscript. This research was supported by grants from the National Science Foundation (DBI-9983133), National Institute of Health (HG-02096) and Burroughs Wellcome Fund (BWF-1001311) to S.K.

REFERENCES

1. Green, E.D. (2000) The human genome project and its impact on the study of human disease. In Scriver, C.R., Beaudet, A.L., Sly, W.S. and Valle, D. (eds), *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill, New York, Vol. **1**, pp. 259–298.
2. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. Horaitis, R., Scriver, C.R. and Cotton, R.G.H. (2000) Mutation databases: overview and catalogues. In Scriver, C.R., Beaudet, A.L., Sly, W.S. and Valle, D. (eds), *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill, New York, Vol. **1**, pp. 113–125.
5. Cotton, R.G. (2000) Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. *Hum. Mutat.*, **15**, 4–6.
6. Welsh, M.J., Ramsey, B.W., Accurso, F. and Cutting, G.R. (2000) Cystic Fibrosis. In Scriver, C.R., Beaudet, A.L., Sly, W.S. and Valle, D. (eds), *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill, New York, Vol. **1**, pp. 5121–5188.
7. Poyau, A., Buchet, K., Bouzidi, M.F., Zabot, M.T., Echenne, B., Yao, J., Shoubridge, E.A. and Godinot, C. (2000) Missense mutations in SURF1 associated with deficient cytochrome c oxidase assembly in Leigh syndrome patients. *Hum. Genet.*, **106**, 194–205.
8. Girodon, E., Cazeneuve, C., Lebargy, F., Chinet, T., Costes, B., Ghanem, N., Martin, J., Lemay, S., Scheid, P., Housset, B. *et al.* (1997) CFTR gene mutations in adults with disseminated bronchiectasis. *Eur. J. Hum. Genet.*, **5**, 149–155.
9. Lazaro, C., de Cid, R., Sunyer, J., Soriano, J., Gimenez, J., Alvarez, M., Casals, T., Anto, J.M. and Estivill, X. (1999) Missense mutations in the cystic fibrosis gene in adult patients with asthma. *Hum. Mutat.*, **14**, 510–519.
10. Tzetis, M., Efthymiadou, A., Strofalis, S., Psychou, P., Dimakou, A., Pouliou, E., Doudounakis, S. and Kanavakis, E. (2001) CFTR gene mutations—including three novel nucleotide substitutions – and haplotype background in patients with asthma, disseminated bronchiectasis and chronic obstructive pulmonary disease. *Hum. Genet.*, **108**, 216–221.
11. Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*, **136**, 927–935.
12. Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics*, **139**, 1067–1076.
13. Pignatti, P.F., Bombieri, C., Marigo, C., Benetazzo, M. and Luisetti, M. (1995) Increased incidence of cystic fibrosis gene mutations in adults with disseminated bronchiectasis. *Hum. Mol. Genet.*, **4**, 635–639.
14. Pignatti, P.F., Bombieri, C., Benetazzo, M., Casartelli, A., Trabetti, E., Gile, L.S., Martinati, L.C., Boner, A.L. and Luisetti, M. (1996) CFTR gene variant IVS8-5T in disseminated bronchiectasis. *Am. J. Hum. Genet.*, **58**, 889–892.
15. Sunyaev, S., Ramensky, V.V., Koch, I.I., Lathe, I.W., Kondrashov, A.S. and Bork, P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
16. Notaro, R., Afolayan, A. and Luzzatto, L. (2000) Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. *FASEB J.*, **14**, 485–494.
17. Felsenstein, J. (1985) Phylogenies and the Comparative Method. *Am. Naturalist*, **125**, 1–15.
18. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In Dayhoff, M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, pp. 345–352.
19. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
20. Krawczak, M. and Cooper, D.N. (1997) The human gene mutation database. *Trends Genet.*, **13**, 121–122.
21. Krawczak, M., Ball, E.V., Fenton, I., Stenson, P.D., Abeysinghe, S., Thomas, N. and Cooper, D.N. (2000) Human gene mutation database-a biomedical information and research resource. *Hum. Mutat.*, **15**, 45–51.
22. Epstein, C.J. (1967) Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature*, **215**, 355–359.
23. Clarke, B. (1970) Selective constraints on amino-acid substitutions during the evolution of proteins. *Nature*, **228**, 159–160.
24. Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.
25. Zhang, J. (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.*, **50**, 56–68.
26. Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
27. McLachlan, A.D. (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. *J. Mol. Biol.*, **61**, 409–424.
28. Krawczak, M., Ball, E.V. and Cooper, D.N. (1998) Neighboring-nucleotide effects on the rates of germ-line single base-pair substitutions in human genes. *Am. J. Hum. Genet.*, **63**, 474–488.
29. Hudson, R.R., Kreitman, M. and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.
30. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila. Nature*, **351**, 652–654.
31. Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.
32. Vulliamy, T., Luzzatto, L., Hirono, A. and Beutler, E. (1997) Hematologically important mutations: glucose-6-phosphate dehydrogenase. *Blood Cells Mol. Dis.*, **23**, 302–313.
33. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
34. Fitch, W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.
35. Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.
36. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
37. Van Camp, G., Fransen, E., Vits, L., Raes, G. and Willems, P.J. (1996) A locus-specific mutation database for the neural cell adhesion molecule L1CAM (Xq28). *Hum. Mutat.*, **8**, 391.
38. Scriver, C.R., Waters, P.J., Sarkissian, C., Ryan, S., Prevost, L., Cote, D., Novak, J., Teebi, S. and Nowacki, P.M. (2000) PAHdb: a locus-specific knowledge base. *Hum. Mutat.*, **15**, 99–104.
39. Brown, A., McKie, M., van Heyningen, V. and Prosser, J. (1998) The human PAX6 mutation database. *Nucleic Acids Res.*, **26**, 259–264.