The Amino-Acid Mutational Spectrum of Human Genetic Disease

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1186/gb-2003-4-11-r72</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:4878075">http://nrs.harvard.edu/urn-3:HUL.InstRepos:4878075</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
The amino-acid mutational spectrum of human genetic disease

Dennis Vitkup*, Chris Sander†‡ and George M Church*

Addresses: *Lipper Center for Computational Genetics and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. †Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA. ‡Current address: Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA.

Correspondence: George M Church. E-mail: g1m1c1@arep.med.harvard.edu

Abstract

Background: Nonsynonymous mutations in the coding regions of human genes are responsible for phenotypic differences between humans and for susceptibility to genetic disease. Computational methods were recently used to predict deleterious effects of nonsynonymous human mutations and polymorphisms. Here we focus on understanding the amino-acid mutation spectrum of human genetic disease. We compare the disease spectrum to the spectra of mutual amino-acid mutation frequencies, non-disease polymorphisms in human genes, and substitutions fixed between species.

Results: We find that the disease spectrum correlates well with the amino-acid mutation frequencies based on the genetic code. Normalized by the mutation frequencies, the spectrum can be rationalized in terms of chemical similarities between amino acids. The disease spectrum is almost identical for membrane and non-membrane proteins. Mutations at arginine and glycine residues are together responsible for about 30% of genetic diseases, whereas random mutations at tryptophan and cysteine have the highest probability of causing disease.

Conclusions: The overall disease spectrum mainly reflects the mutability of the genetic code. We corroborate earlier results that the probability of a nonsynonymous mutation causing a genetic disease increases monotonically with an increase in the degree of evolutionary conservation of the mutation site and a decrease in the solvent-accessibility of the site; opposite trends are observed for non-disease polymorphisms. We estimate that the rate of nonsynonymous mutations with a negative impact on human health is less than one per diploid genome per generation.

Background

Several recent studies [1-6] have applied computational methods to predict potentially deleterious effects of nonsynonymous single-nucleotide polymorphisms (SNPs) in humans. SNPs represent common human alleles, usually with population frequencies greater than 1%. Both structural and evolutionary methods were used to assess potential functional effects of SNPs. It was predicted that a substantial fraction (10-30%) of human SNPs may affect protein function negatively, although the medical consequences of these SNPs remain to be established.

The main goal of the work reported here is to characterize and rationalize the overall amino-acid spectrum of disease mutations and non-disease SNPs (referred to as 'benign SNPs' below). We obtain the relative probabilities that a random
mutation (rather than an existing SNP) will cause a genetic disease while explicitly taking into account the underlying spectrum of nucleotide mutations. Such an approach will allow, in the future, the identification and characterization of highly mutable sites in the human genome which are also functionally important.

Miller and Kumar [5] performed a detailed analysis of the disease mutations and benign SNP spectra in seven human genes. While some of our results are consistent with their study, we find major differences. For example, we observe a significantly larger contribution of mutations at arginine (Arg) and glycine (Gly) to human genetic disease. We attribute the differences to the substantially larger gene set (436 genes versus 7) used in our analysis.

Results and discussion

Overall amino-acid mutational spectrum

We present the amino-acid spectra of disease mutations and polymorphisms in Figure 1. The mutations from the Mendelian Inheritance in Man (MIM) database [7] annotated in SWISS-PROT [8] were used as a source of human disease mutations. In total, 4,236 mutations from 436 genes were considered. The collection of 1,037 synonymous and nonsynonymous SNPs from the extensive analysis of haplotypes in 313 human genes [9] was used as a source of benign SNPs. There was no overlap between the disease mutations and benign sets of SNPs used in the study. The spectrum of interspecies substitutions (Figure 1d) was calculated on the basis of the PAM1 matrix [10] as described in Materials and methods.

Nearly all mutations in the current MIM database represent Mendelian disease (monogenic in etiology). It remains to be seen to what extent our results pertain to disease mutation involved in polygenic disorders. At this point, too little is known about this type of mutation, and more experimental work is required in order to understand their spectrum.

The mutation matrices in Figure 1 are sparse (that is, a large number of the matrix elements are close or equal to zero) and nonsymmetrical (in many cases the tendency of amino acid I to mutate into amino acid J is different from the tendency of amino acid J to mutate into I). The vast majority of human genetic mutations are caused by single-nucleotide changes [11,12]. Consequently, the matrices in Figure 1b,c,d represent amino-acid transitions resulting predominantly from single-nucleotide mutations in amino-acid codons. To rationalize the observed disease and benign spectra, we generated the expected mutation spectrum (Figure 1a) using the neighbor-dependent matrix of nucleotide mutation rates developed by Hess et al. [13] (see Materials and methods). The expected mutation matrix in Figure 1a represents the spectrum which would be observed if all nonsynonymous mutations were accepted (that is, there were no selection). The expected spectrum was generated for the disease genes considered and, separately, for a large collection of more than 7,000 human genes available from SWISS-PROT. These two spectra were almost identical (R = 0.98, p < 0.0001), suggesting that the expected spectrum in Figure 1a reflects general properties of all human genes (such as amino-acid codon frequencies and context-dependent nucleotide mutation frequencies). Here and throughout the paper we use the t-test statistics with n-2 degrees of freedom to estimate the significance values obtained using the t-test.

The spectrum of disease mutations was calculated separately for membrane proteins. The program TMHMM [14] was used to detect potential transmembrane regions. The disease spectrum for membrane proteins is very similar to the all-protein disease spectrum (R = 0.97, p < 0.0001 for all disease mutations in membrane proteins, 1,598 in total; R = 0.75, p < 0.0001 for disease mutations in transmembrane regions, 372 in total). Evidently, specific properties of membrane proteins and the constraints on them are not able to significantly modify the disease spectrum common to all proteins.

Correlations between the expected and the observed spectra

Close-to-diagonal mutations in Figure 1a,b,c,d represent substitutions between amino acids with similar chemical properties (conservative mutations). The interspecies substitutions (Figure 1d) contain the highest fraction of conservative mutations compared to disease mutations and benign SNPs. The frequencies of the benign SNPs, disease mutations, and interspecies substitutions are plotted versus expected frequencies in Figure 2. Benner et al. [15] showed that the genetic code affects the amino-acid substitution spectrum at early stages of divergence, whereas chemical similarities dominate at longer evolutionary distances. The correlation between the benign and expected spectra observed in our study (R = 0.78, p < 0.0001), is an expected extension of Benner’s et al. conclusion to even shorter evolutionary distances (variations within a population).

Interestingly, we also find a strong correlation of the disease mutation spectrum with the expected spectrum based on the genetic code (R = 0.71, p < 0.0001). The correlation of disease mutation frequencies with the chemical dissimilarities between original and mutant amino acids is apparent only after normalization by the expected frequencies (Figure 3a,b). Consequently, in the majority of cases the comparison of amino-acid types (wild type versus mutant) will be insufficient to distinguish neutral from disease variants.

The contribution of mutations at different amino acids to the disease spectrum is highly heterogeneous (Figure 4). Interestingly, mutations at Arg residues account for almost 15% of the disease mutations. This is a direct consequence of the well-known high mutability of Arg (due to deamination of
5′-CpG dinucleotides in Arg codons [16,17], the relatively high frequency of Arg in human proteins (<4%), and the fact that Arg mutates to residues with very different chemical properties (cysteine (Cys), glycine (Gly), histidine (His), lysine (Lys), leucine (Leu), methionine (Met), proline (Pro), glutamine (Gln), serine (Ser) and tryptophan (Trp)). The relative probability of a disease mutation at different amino acids (Figure 4b) was calculated by dividing the disease and expected frequencies. Accordingly, a random mutation at a Trp or Cys residue has the highest probability of causing a disease. This correlates well with the highest evolutionary conservation of exactly these two residues [10]. Both Trp and Cys residues play a prominent part in determining protein stability. In addition to Trp and Cys, the high probability of disease mutations at Gly may be related to important structural roles often played by this residue. For example, mutations at Gly,
which is frequently present at the turns of alpha-helices, might have a negative impact on protein structural stability. Our definition of the relative probability of disease mutations is similar to the relative clinical observation likelihood (RCOL) used by Cooper et al. in several publications (see, for example [6]). In the next section we extend the relative probabilities to interspecies comparisons.

**Probabilities of mutations or SNPs as a function of the mutation/SNP-site properties**

To complement the analysis of the amino-acid mutation matrices, we investigated how the probabilities of benign SNPs and disease mutations depend on the properties of the mutation site. Several recent studies have focused on developing evolutionary and structural approaches to predict potentially deleterious human mutations [1-4,18,19]. Here, we focus on understanding the relative mutation probabilities (see Materials and methods). Our results are in general agreement with the previous studies. The relative probabilities of disease mutations and benign SNPs are shown in Figure 5a as a function of the interspecies evolutionary conservation of the mutation site. The conservation was characterized by the relative entropy measure using homologs with more than 30% sequence identity. The probability that a random mutation will cause a genetic disease increases monotonically with an increase in the degree of site conservation, while the
The solvent accessibility of an amino-acid residue in a protein reflects the degree of the residue's exposure to the surrounding solvent in the protein structure. The relative probability of disease-causing mutations is highest in the protein interior and lowest on the protein surface (Figure 5b). The benign SNPs show the reverse trend, as their relative probability is highest on the surface and lowest in the protein interior. This is consistent with the study by Moult and co-workers [1] (see also Ferrer-Costa et al. [20] and Bustamente et al. [21]), who suggested that the dominant mechanism by which disease mutations damage protein function is a decrease in protein stability, as opposed to mutations of active-site residues (usually located on the protein surface).

Both relative entropy and solvent accessibility exclusively characterize the site of a mutation. To estimate the extent to which a given amino acid is incompatible with the residues observed at the same position in close homologs, we introduced the Grantham Ratio (GR) score based on the Grantham dissimilarity matrix [22] (see Materials and methods for a formal definition). Application of other scores, for example those based on the BLOSUM matrices, gave qualitatively similar results [3]. The GR score is the ratio of two averages - the numerator being the average dissimilarity between the mutated amino acid and the residues observed at the same site in evolution, and the denominator being the average dissimilarity within the residues observed at the site in homologous proteins. Defined in this way, a GR score smaller or close to 1 suggests that the amino acid is similar to the residues observed at the site in evolution, whereas a GR score significantly larger than 1 indicates that the amino-acid change is evolutionarily radical.

The role of purifying selection in shaping the mutation spectra is apparent from the cumulative distribution of the GR scores (Figure 6). Whereas the GR distribution for original (wild type) residues at benign sites (blue curve) is very similar to the distribution for all protein residues (black), the distribution for mutant residues at benign sites (green) clearly shows an excess of radical mutations. Importantly, the GR distribution of mutant residues at benign sites (green) is similar to the distribution for randomly generated mutations (cyan) and is quite different from the disease mutation distribution (red). Consequently, although a significant fraction of randomly arising nonsynonymous mutations are evolutionarily radical (and thus potentially deleterious) they are not, on average, as radical as the disease mutations and still have appreciable frequencies in the human population. Indeed, it was recently estimated [23] that the average reduction in evolutionary fitness due a mildly deleterious SNP with a significant frequency in the human population is in the range of 0.01-1%. The medical importance of such mildly deleterious human mutations remains to be established [24,25].

The cumulative distribution of the GR scores for disease mutations suggests that more than a half of the disease

Figure 3
Chemical dissimilarities between original and mutant amino acids versus observed mutation frequencies. (a) The amino-acid dissimilarities versus the frequencies of disease mutations. (b) The amino-acid dissimilarities versus the relative frequencies of genetic disease mutations (normalized by site mutabilities). The chemical dissimilarities between amino acids were characterized using the Grantham score [22]. Each point in the figure represents a certain type of amino-acid to amino-acid mutation. Only amino-acid transitions resulting from single-nucleotide mutations in amino-acid codons were considered. The mutation frequencies in (a) were normalized to 100. The relative frequencies in (b) were determined as the ratio between the disease and the expected mutation frequencies (see Figure 1). The relative frequencies are proportional to the relative probabilities of amino-acid mutations causing a disease. No correlation between the disease mutation frequencies and chemical dissimilarities is evident (a), but there is a significant correlation between the normalized frequencies and the chemical dissimilarities (b).
mutations are evolutionarily radical (represented by residues with GR score greater than 2). Residues with such GR scores are almost never observed in homologous sequences (blue and black curves). It is important to note that medically damaging mutations and SNPs cannot always be rationalized in terms of evolutionary radicality. Medically harmful mutations contribute to the overall mutation spectrum.
mutations may cause late-onset human diseases without strong selection in evolution. Alternatively, a particular amino-acid substitution can be damaging to a human protein but be relatively frequent in the homologous family due to compensatory mutations. Such substitutions may account for deleterious mutations with low GR scores.

**Estimation of the maximal rate of mutations with impact on human health**

From Figure 6 we can estimate the maximum rate of random mutations with significant impact on human health (that is, an impact similar to mutations currently annotated in MIM). We note that the mutation rate we estimate (a fraction of newly created deleterious mutations) is different from the fraction of existing SNPs with deleterious effects on protein function (estimated previously [1,2,23]). The comparison between the distribution of random SNP mutations (cyan) and disease mutations (red) suggests that about 10% of the randomly generated mutations have GR scores greater than 6. Such a score corresponds to approximately 40% of the disease mutations. As a result, the total rate of the disease mutations cannot be larger than one quarter of the random mutation rate. Thus, one expects, at most, 25% of random nonsynonymous mutations to be as damaging as mutations currently annotated in MIM (similar estimates are obtained using GR cutoffs larger than 6).

This estimate has a simple biochemical rationale, as mutagenesis experiments on different proteins suggest that less
than 30% of random mutations substantially damage biological function or stability of proteins [26-29].

Using the recent estimate of the human mutation rate of 175 mutations per diploid genome per generation [30] (corresponding to approximately two to three nonsynonymous mutations), we conclude that the rate of nonsynonymous mutations with serious impact on human health should be less than one per diploid genome per generation. This is probably a substantial overestimation of the rate because we assume that all human genes are as important for human health as the well-annotated disease genes currently in the MIM database. We emphasize that the rate of health-damaging nonsynonymous mutations is smaller than the total rate of deleterious human mutations, which is estimated to be larger than one [30,31].

The present analysis, together with other recent studies [1-4,23], establishes the basis for understanding the spectrum of deleterious human mutations. The amino-acid substitution matrices, such as PAM [10] and BLOSUM [32], apart from playing a fundamental role in sequence alignment, qualitatively characterize the evolutionary interchangeability of amino acids averaged over many protein families. The disease spectrum, characterized by our analysis, explores another important aspect of evolution, namely the generation of deleterious mutations. Because of all mammalian species have a broadly similarity physiology, the properties of the disease spectrum should be general, at least for mutations leading to early-onset diseases. We anticipate that understanding the disease spectrum will allow one to predict, in advance, the rates and potential medical consequences of all possible single-nucleotide mutations in the human genome.

Materials and methods
Calculation of mutation spectra
The spectrum of expected amino-acid mutation frequencies (Figure 1a) was generated using the matrix of neighbor-dependent nucleotide mutation rates obtained by Hess et al. [13] (Additional data file 1). The neighbor-dependent mutation matrix was calculated by Hess et al. on the basis of 20,200 substitutions in aligned gene/pseudogene human sequences; the relative mutation rates were calculated for the four nucleotides in all 16 possible 5' and 3' neighborhoods. To obtain the expected amino-acid mutation frequencies for a given collection of genes, we simulated all possible single-nucleotide mutations with appropriate rates, and recorded the corresponding amino-acid changes. The nucleotide mutational spectrum of individual genes may be affected by the presence of so-called mutation hot spots [33-35]. However, on average, there is only a small influence of the surrounding DNA sequence (beyond nearest 5' and 3' neighbors) on the relative nucleotide mutation rates [17].

The interspecies spectrum of amino-acid mutation frequencies (Figure 1d) was calculated on the basis of Dayhoff's PAM1 matrix. The original PAM1 matrix [10] gives the probabilities of amino-acid substitutions over small evolutionary distances. These probabilities were multiplied by the amino-acid frequencies in human genes for direct comparison with the expected, disease, and benign SNPs matrices.

Structural and evolutionary analysis of mutations
The list of disease genes obtained from SWISS-PROT was filtered using the program PSEG [36] to exclude genes with a significant fraction of low-complexity regions. As a result of the filtering, six genes for collagen proteins were excluded from the original set of 436 genes. Mutations at Gly residues constitute more than 50% of the collagen disease mutations (due to the collagen structural motif). Because of this bias, the collagen mutations were excluded from all calculations. If the collagen mutations are included, the total fraction of disease mutations at Gly (Figure 4a) increases from 12% to 15%.

Membrane proteins and transmembrane protein regions were detected using the program TMHMM [14] with standard parameters. Out of 430 disease genes, 105 (24%) were classified as membrane proteins on the basis of the presence of at least two distinct transmembrane domains. To characterize the evolutionary conservation of mutation sites we used BLASTGP to search the nrdb90 database [37] for homologs with greater than 30% sequence identity. The nrdb90 database constitutes a nonredundant merge of sequence and structural databases, which is filtered so that no pair of sequences has greater than 90% sequence identity. The homologs to each human protein were subsequently aligned using the program CLUSTALW [38] with default parameters. Only mutation sites covered by more than 10 homologous sequences (excluding gaps) were used in the evolutionary analysis. The multiple sequence alignments obtained using CLUSTALW were used to characterize the relative entropy (Kullback-Leibler distance) of the benign and disease mutation sites. The relative entropy was calculated according to the formula:

\[
\text{Relative Entropy} = \sum_n P(n) \log \frac{P(n)}{Q(n)}
\]

where the summation is over all amino-acid types n in the alignment; P(n) is the probability of the amino acid n in the column corresponding to mutation; Q(n) is the probability of the amino acid n in all columns of the multiple sequence alignment.

The multiple sequence alignments were also used to calculate the Grantham ratio (GR) score according to the formula:
To characterize the structural location of disease mutations and benign SNPs, BLASTGP [39] was used to search the Protein Data Bank (PDB) [40] for sequences homologous to known structures. Only sequences with greater than 30% identity to human sequences over the entire length of the alignment were considered. In total, the solvent accessibility were calculated for 110 benign SNPs and 840 disease mutations. The solvent accessibility of mutation sites was determined by the program NACCESS [41] using the watersphere radius of 1.4 Å. The solvent accessibility represents the relative exposure of a residue X in a protein structure compared to its exposure in the tripeptide Ala-X-Ala.

**Calculation of relative mutation probabilities**

The relative mutation probabilities shown in Figures 4b, 5a, and 5b represent conditional probabilities. Specifically, the conditional probability P(disease|descriptor), that a mutation will cause a genetic disease given a certain property (descriptor) of the mutation site, was calculated according to the formula:

\[
P(\text{disease} | \text{descriptor}) = \frac{P(\text{disease} | \text{descriptor})}{P(\text{disease})}
\]

where ‘descriptor’ represents solvent accessibility or evolutionary conservation of the mutation site, P(disease|descriptor) is the probability that a disease mutation has a given descriptor value, P(designer) is the probability that a random mutation (disease or non-disease) has a given descriptor value, and P(disease) is the probability that a random mutation will cause a genetic disease. Importantly, because P(disease) is unknown, we can only estimate P(disease|descriptor) up to a constant (assuming certain P(disease) value). Consequently, we refer to P(disease|descriptor) as relative mutation probabilities. The probability that a random mutation has a given descriptor value P(designer) was estimated by simulating random single-nucleotide mutations using the expected amino-acid mutation frequencies (Figure 1a).

**Additional data files**

The following additional data are included with the online version of this article: a list of relative mutation rates (Additional data file 1), a list of disease mutations (Additional data file 2), a list of disease mutation genes (Additional data file 3), a list of SNPs used in the analysis (Additional data file 4), and the Grantham ratio scores (Additional data file 5).

**Acknowledgements**

We thank Jay Shendure, John Aach, Patrick D’haseeleer, Daniel Segre, Peter Kharchenko, and Tzachi Pilpel for discussions. This work was supported in part by research grants from the US Department of Energy through the grant DOE DE-FG02-87-ER60565.

**References**


\[
GR = \frac{1}{n} \sum_{i=1}^{n} D(\text{Human}_\text{RES}, \text{RES}(i)) - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>1} D(\text{RES}(i), \text{RES}(j))
\]

where \(D(A,B)\) is the Grantham measure of chemical dissimilarities between amino-acid residues \(A\) and \(B\), \(\text{Human}_\text{RES}\) is the human residues at the mutation site, \(\text{RES}(i)\) is the amino acid from the \(i\)th aligned sequence homolog at the mutation site, and \(n\) is the number of aligned sequences. Qualitatively, the GR score is a measure of dissimilarity between a human amino acid and the residues seen at the same site in homologs. In total, the relative entropy and Grantham ratio were calculated for 258 benign SNPs and 2,636 disease mutations.


