Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 318 (2003) 169–175
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

# Human disease genes: patterns and predictions

Nick G.C. Smith[a,*], Adam Eyre-Walker[b]

[a] *Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden*
[b] *Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton, UK*

Received 15 December 2002; received in revised form 29 April 2003; accepted 20 June 2003

Received by W. Makalowski

**Abstract**

We compared genes at which mutations are known to cause human disease (disease genes) with other human genes (nondisease genes) using a large set of human–rodent alignments to infer evolutionary patterns. Such comparisons may be of use both in predicting disease genes and in understanding the general evolution of human genes. Four features were found to differ significantly between disease and nondisease genes, with disease genes (i) evolving with higher nonsynonymous/synonymous substitution rate ratios (Ka/Ks), (ii) evolving at higher synonymous substitution rates, (iii) with longer protein-coding sequences, and (iv) expressed in a narrower range of tissues. Discriminant analysis showed that these differences may help to predict human disease genes. We also investigated other factors affecting the mode of evolution in the disease genes: Ka/Ks is significantly affected by protein function, mode of inheritance, and the reduction of life expectancy caused by disease.


*Keywords:* Deleterious mutations; Nearly neutral theory; Substitution rate variation; Protein function; Nonsynonymous/synonymous rate ratio; Human disease gene prediction

## 1. Introduction

We have compared the evolution of those genes at which mutations are known to cause genetic disease in humans (termed disease genes) relative to those genes which are not yet known to cause disease (provisionally nondisease genes) using a set of human–rodent alignments. There are two principal motivations behind this study. The first reason is concerned with the practical benefits to human health. We already know many genes which cause genetic disease, but we certainly have not identified all such genes. The sequencing of the human genome offers the prospect of identifying potential disease genes by sequence analysis. By performing sequence comparisons between humans and other mammalian species, we can characterize the pattern of evolution at disease genes relative to nondisease genes. If there are strong differences in evolutionary patterns between disease and nondisease genes, then the comparative approach might prove a useful method for identifying potential disease genes.

Why might we expect disease-causing genes to show unusual patterns of evolution? It is already well established that those nonsynonymous DNA mutations which cause disease are atypical both in their rate and pattern of evolution. For example, nonsynonymous mutations causing disease are enriched at highly conserved amino acid positions (Notaro et al., 2000; Miller and Kumar, 2001), and the amino acid changes caused by such mutations tend to involve large changes in amino acid physicochemical properties and seem likely to have severe effects on protein stability (Miller and Kumar, 2001; Ferrer-Costa et al., 2002). We do not necessarily expect the pattern of mutations causing disease to match the pattern of substitution in those genes (e.g., because of negative selection; see Sunyaev et al., 2001), but the mutation data suggest that disease genes may show unusual patterns of substitution.

Furthermore, we can expect that disease genes may evolve at unusual rates due to selection given that mutations in disease genes have discernible phenotypic (and hence fitness) effects. If one considers the full spectrum of possible

*Abbreviations:* Ka, nonsynonymous substitution rate; Ks, synonymous substitution rate; Ka/Ks, nonsynonymous/synonymous substitution rate ratio.

* Corresponding author. Tel.: +46-18-4716466; fax: +46-18-4716310.
*E-mail address:* nick.smith@ebc.uu.se (N.G.C. Smith).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00772-8

phenotypic effects of mutations, it seems reasonable that disease genes represent an intermediate class. We can observe disease symptoms only if disease mutations are not lethal at an early stage of development, so disease mutations must have less severe consequences than lethal mutations. On the other hand, the fact that genetic disease generates discernible phenotypic symptoms means that disease mutations do have some deleterious effects. Unfortunately, the prediction of whether we expect disease genes to evolve faster or slower than nondisease genes depends on the unknown spectrum of phenotypic effects of mutations in nondisease genes: does the effect of lethal mutations in nondisease genes outweigh the effect of mutations which generate no change in phenotype?

We have used three substitution rate statistics to compare the evolution of disease and nondisease genes: Ka, Ks, and Ka/Ks. Ka is the substitution rate per nonsynonymous site which measures protein evolution, while Ks is the substitution rate per synonymous site for DNA mutations which do not affect protein sequence. Ka/Ks is the ratio of the nonsynonymous and synonymous substitution rates. If synonymous sites are neutral, as seems reasonable in mammals, the Ka/Ks rate ratio measures the rate of protein evolution relative to the mutation rate and is a useful indicator of selection pressures (Keightley and Eyre-Walker, 2000; Yang and Bielawski, 2000; Eyre-Walker et al., 2002). Ka/Ks is expected to increase as the level of negative selection decreases and as the level of positive selection increases, with the level of selection determined by the distribution of selection coefficients across sites.

The second reason for studying the evolution of human disease genes is to improve our understanding of why different genes evolve at different rates, a fundamental problem in molecular evolution (Li, 1997). Human disease genes may tell us how human genes evolve generally. We have taken advantage of a recent study which compiled phenotypic and genetic features for a large number of disease genes (Jimenez-Sanchez et al., 2001), including several features which might be expected to affect evolutionary processes and hence rates of evolution.

## 2. Materials and methods

The database of human disease genes comes from the study of Jimenez-Sanchez et al. (2001), which also provided data on many phenotypic and genetic features of disease genes (see Section 3.3). Jimenez-Sanchez et al. (2001) compiled nearly 1000 human disease genes, 97% of which are genes causing monogenic disease. We developed an automated strategy for developing sets of confirmed disease and provisionally nondisease genes. Online Mendelian Inheritance in Man (OMIM) accession numbers from the database were used to extract human protein sequences from NCBI www.ncbi.nlm.nih.gov), and these protein sequences were then compared to human protein sequences derived from a large set of human–rodent DNA alignments of confirmed orthology (Duret and Mouchiroud, 2000) using stand-alone BLAST (Altschul et al., 1997). The criterion for positive identification as a disease gene was over 99% protein identity, excluding gaps. The BLAST searches identified 392 genes in the disease set and 2038 genes in the nondisease set.

For all human–rodent DNA alignments, synonymous and nonsynonymous substitution rates, Ks and Ka, were estimated using codeml in PAML (Yang, 1997). Genes with extreme values of $Ks > 2$ were removed from the analyses, leaving 387 disease and 2024 nondisease genes. A measure of expression pattern based on EST data (Duret and Mouchiroud, 2000) was recorded for all genes: $T$ is the number of tissues for which expression was recorded in humans, out of a total of 19 tissues. Base composition was calculated as the G + C content at the third codon position, taking the average of the human and rodent sequences (GC3). Protein domains found in human proteins were taken from the InterPro database (Apweiler et al., 2001).

## 3. Results and discussion

### 3.1. Comparison of disease and nondisease genes: predicting disease genes

Table 1 summarizes various features of disease and nondisease genes. The substitution rate data clearly show that disease genes evolve faster than nondisease genes at both synonymous and nonsynonymous sites (percentage differences refer to means, $p$ values for Mann–Whitney $U$ test; Ks 8% higher, $p = 0.022$; Ka 25% higher, $p < 0.001$). The difference in substitution rates is greater at nonsynonymous sites than at synonymous sites, as shown by Ka/Ks significantly higher for disease genes (Ka/Ks 24% higher, $p < 0.001$).

We also investigated whether disease and nondisease genes differ in characters other than evolutionary rates: gene length ($L$), tissue range of expression ($T$), base composition (GC3), and protein function. Disease genes have a

Table 1
Comparison of means and standard errors for various features of 387 disease genes and 2024 nondisease genes: tissue range of expression ($T$), coding sequence length in bp ($L$), G + C content at the third codon position (GC3), synonymous and nonsynonymous substitution rates Ks and Ka, and the substitution rate ratio Ka/Ks

| | Nondisease | | Disease | | $p$ |
|---|---|---|---|---|---|
| | Mean | S.E. | Mean | S.E. | Two-tail Mann–Whitney $U$ |
| Ks | 0.7425 | 0.0127 | 0.8001 | 0.0406 | 0.022 |
| Ka | 0.0652 | 0.0016 | 0.0816 | 0.0037 | <0.001 |
| Ka/Ks | 0.0913 | 0.0022 | 0.1135 | 0.0052 | <0.001 |
| $T$ | 5.1186 | 0.1077 | 4.3359 | 0.2120 | 0.010 |
| $L$ | 1564.7 | 25.6 | 2010.3 | 84.6 | <0.001 |
| GC3 | 0.6218 | 0.0029 | 0.6287 | 0.0065 | 0.259 |

significantly narrower tissue range of expression ($T$ 15% lower, $p = 0.016$) and significantly longer coding sequences ($L$ 28% higher, $p < 0.001$) than nondisease genes, while base composition does not differ significantly between disease and nondisease genes.

We checked whether protein structure might affect the chances of a gene being involved in disease. The justification for this analysis in terms of predicting disease genes comes from studies which have noted that the occurrence of disease-causing mutations varies with protein structure (Sunyaev et al., 2001; Ferrer-Costa et al., 2002). We defined protein structure using the InterPro database of protein domains (Apweiler et al., 2001). Only the 22 domain types found in 25 or more genes were considered. We tested whether there was significant heterogeneity in the proportions of disease and nondisease genes across protein domains; in other words, were some protein domains more likely than others to be found in disease genes? A Chi-squared test revealed no significant heterogeneity ($\chi^2 = 40.2$, $df = 43$, $p = 0.59$), and thus, we have no evidence for a relationship between protein structure and disease genes, at least for the protein domain classification of protein structure.

Our analyses suggest four potential predictors of disease genes: the substitution rate ratio Ka/Ks, the synonymous substitution rate Ks, the tissue range of expression $T$ and the gene length $L$. Although the nonsynonymous substitution rate Ka also varies significantly between disease and nondisease genes, the use of Ka/Ks is a better measure of protein evolution since it controls for mutation rate (strictly Ks) variation. We performed a discriminant analysis using SPSS version 11 to build a predictive model of group membership (i.e., whether genes are classified as disease or nondisease). The purpose of the discriminant analysis is to examine whether disease gene prediction is possible; we anticipate that alternative methods such as neural nets and logistic regression may provide more powerful approaches to disease gene prediction.

As one would expect, given that all factors differ significantly between disease and nondisease genes, a highly significant discriminant function was generated (disease genes mean discriminant score 0.38 and S.E. 0.06, nondisease genes mean discriminant score $-0.07$ and S.E. 0.02; Wilk's lambda = 0.973, $\chi^2 = 64.9$, $df = 4$, $p < 10^{-12}$). Despite the highly significant discriminant function, note that the division of disease and nondisease genes explains less than 3% of the variance in the predictive factors since Wilk's lambda gives the amount of variability among factors that is not explained by group membership (see p. 319 in Zar, 1999). The value of the discriminant function, a linear combination of the four predictor factors, was used to predict group membership contingent on the known sizes of the disease and nondisease groups (note that the true sizes of the disease and nondisease groups are unknown, with the known size of the disease group an underestimate of the true value, so it may be worthwhile exploring other combinations of group sizes in future studies). With the discriminant analysis, 83.9% of the genes were correctly classified, with 374 incorrectly classified as nondisease genes, and 14 incorrectly classified as disease genes.

The 14 genes supposed to have been incorrectly classified as disease genes are the nondisease genes with the highest values of the discriminant function, and their accession numbers are as follows: AF027807, X62515, X69086, X05615, U29344, U29874, U38291, U79716, U86136, X51435, M34677, D25542, M73548, L32832. These 14 genes can be considered as the most likely of the nondisease genes to be actual disease genes. Three types of misclassification of the provisionally nondisease genes can be distinguished. First, there are those disease genes present in the database of Jimenez-Sanchez et al. (2001) but not picked up by our automated sequence analysis approach, for example, due to the presence of alternative splice products in the protein databases. Second, there are those disease genes which were not present in the database of Jimenez-Sanchez et al. (2001), but for which clinical symptoms and/or allelic variants with phenotypes are now recorded in OMIM www.ncbi.nlm.nih.gov). Third, there are those genes which are not recorded as human disease genes in OMIM. In the set of 14 genes, there are three cases of the first type (accession numbers and OMIM references: X05615 and 188450, M34677 and 306700, M73548 and 175100) and two cases of the second type X62515 and 142461, U79716 and 600514), which leaves the remaining nine genes of the third type as those most strongly predicted by our analysis to be implicated in human disease.

When the additional disease genes are taken into account (i.e., if we now count 392 disease genes and 2019 nondisease genes), we can see that although the discriminant method may not be very good at predicting all disease genes (high level of false negatives), it does seem to be fairly reliable when it does predict a disease gene (low level of false positives). If we consider the nondisease genes, only 0.45% (9 out of 2019) are misclassified as disease genes by the discriminant analysis. In contrast, 95.2% (374 out of 392) of the disease genes are misclassified, and hence, the level of false negatives is very high. However, the change from nondisease genes to disease genes does represent a greater than 10-fold enrichment of predicted disease genes from 0.45% to 4.8%. Such an enrichment of disease genes indicates that the methods developed do have some power and hence utility in disease gene prediction. Furthermore, the false positive rate is low, with an upper limit of 33% (9 out of 27), since some of the nine provisionally nondisease genes may be implicated in human disease in the future.

### *3.2. Comparison of disease and nondisease genes: evolutionary explanations*

Our investigation into human disease gene prediction has revealed significant differences between disease and nondisease genes. Although not strictly necessary for the

purposes of disease gene prediction, it is clearly desirable to try to explain such differences. Furthermore, it is useful to distinguish between those differences due to fundamental evolutionary processes and those which may be instead due to various ascertainment biases which may have affected the discovery of human disease genes and which may have generated some of the differences observed in this study. The former type of difference will presumably be more useful than the latter for predicting disease genes in the future.

The findings that disease genes have significantly higher Ka and Ka/Ks than nondisease genes, as well as the observation that disease genes are expressed in a significantly narrower range of tissues (lower *T*), are all consistent with the idea that disease genes are under weaker negative selection than nondisease genes. If it is assumed that disease mutations have phenotypic effects intermediate between lethal mutations and those mutations with no discernible phenotype, then our results suggest that the mode of nondisease gene evolution is more affected by lethal mutations than by mutations with no phenotypic effect (see Section 1). Thus disease genes, which are less affected by lethal mutations, are expected to evolve faster than nondisease genes.

It has been previously shown that both Ka and Ka/Ks are negatively correlated with the tissue range of expression (Duret and Mouchiroud, 2000), as expected if negative selection is stronger on genes with broader expression patterns. For the combined disease and nondisease data set, we find that both Ka and Ka/Ks negatively covary with *T* (Spearman's rank correlation $r = -0.199$ for Ka/Ks and $r = -0.228$ for Ka, $p < 0.001$ in both cases). Thus, the narrow range of expression of disease genes is consistent with the higher Ka and Ka/Ks of disease genes. We can ask whether the faster rate of protein evolution (Ka) and different mode of evolution (Ka/Ks) in disease genes can be fully explained by tissue expression patterns by using the residuals of Ka and Ka/Ks after linear regression against *T*. Although the magnitude of the difference decreased slightly (from 0.022 to 0.019 for Ka/Ks and from 0.016 to 0.014), the significant differences between disease and nondisease genes remained after correction for *T* (Mann–Whitney *U* tests, $p < 0.001$ for both Ka/Ks and Ka).

Thus, our results are consistent with the hypothesis that disease genes are subject to weaker negative selection on protein function than nondisease genes, partly due to a narrower tissue range of expression but mostly independent of expression range effects. However, how much confidence should we place in such conclusions? Although we have attempted to control for confounding factors, it is hard to be sure that the differences we see are solely due to a difference between disease and nondisease genes. As we learn more about substitution rate variation, the list of factors which need to be checked as potential confounding factors inexorably rises. At present, such a list might include the following: amino acid composition and protein structure (Xia and Li, 1998; Tourasse and Li, 2000; Xia and Xie, 2002), presence of a duplicate/paralog (Nembaware et al., 2002), protein function (Hurst and Smith, 1999; Nembaware et al., 2002), expression levels and tissue of expression (Duret and Mouchiroud, 2000), and many factors which vary regionally across the genome, such as gene density and tissue of expression (Lercher et al., 2002). All of these complicating factors could invalidate a simple interpretation of the rate differences between disease and nondisease genes. Unless we can control for both strong negative selection (Wilson et al., 1977; Brookfield, 2000) and strong positive selection (Hurst and Smith, 1999), it is difficult to properly test hypotheses concerning substitution rate variation.

Several other assumptions implicit in our interpretation of our results may also weaken our conclusions. In order to obtain a large set of gene alignments, we have used human–rodent pairwise comparisons. However, such distant comparisons may tell us little about current selection pressures in the human genome. Similarly, it is unknown if there is a tendency for human disease genes to also cause disease in rodents, although the fact that the orthologs of human disease genes in *Caenorhabditis elegans* tend to have viable RNAi loss-of-function phenotypes suggests that the phenotypic effects of mutations are well conserved in animal evolution (Kamath et al., 2003). A final problem is the need to extrapolate from the fitness consequences of those mutations causing disease to the evolution of all the sites in a gene.

Given that synonymous mutations in mammals are probably neutral, differences in Ks can be interpreted as mutation rate differences. How then can we explain why Ks is significantly higher for disease genes than nondisease genes? The direction of the difference is consistent with adaptive mutation rates: if disease genes are under weaker negative selection than nondisease genes, then cost–benefit considerations would lead to higher mutation rates for disease genes. We also note that the well-known positive correlation between synonymous and nonsynonymous substitution rates in mammals (Smith and Hurst, 1999; Bielawski et al., 2000; Hurst and Williams, 2000) appears to be associated with the Ks difference; there is no significant difference between disease and nondisease genes in the residuals of Ks after linear regression against Ka (Mann–Whitney *U* test, $p = 0.37$). However, since we do not yet understand the cause of the Ka–Ks correlation, this finding does not provide a proper explanation for our Ks result. In particular, if the Ka–Ks correlation in mammals is simply due to mutation rate variation according to neutral theory (Ohta and Ina, 1995), then we are left with a circular argument: both Ks and Ka are higher in disease genes because mutation rates are higher in disease genes.

We now move to perhaps the most surprising difference between disease and nondisease genes, the fact that disease genes are significantly and considerably (28%) longer than nondisease genes. This length difference cannot be fully explained by the weak negative correlation between gene

length and range of tissue expression (combined disease and nondisease data set, Spearman's rank correlation $r = -0.081$ and $p < 0.001$), since the significant length difference remains for the residuals after linear regression against $T$ values (difference in mean residuals 420 bp, two-tail Mann–Whitney $U$ test, $p < 0.001$); nor can the gene length difference be explained by the rarity of longer genes in high GC isochores (Duret et al., 1995), since the GC3 of the disease genes is slightly greater than the GC3 of nondisease genes. We suggest that the explanation for longer disease genes may be due to a mutation screen bias: the longer a gene is, the more sites there are at which a mutation may be found, and so, the more likely the identification of the disease gene. Alternatively, features known to be associated with human disease genes such as overlapping gene groups and multiple amino acid runs (Karlin et al., 2002) would be expected to be more common in longer genes.

### 3.3. Substitution rate variation within disease genes: do phenotypic and genetic characteristics affect substitution rates?

We now turn to evolutionary rate variation within the set of disease genes in the hope that such comparisons may reveal general patterns of human gene evolution. Although we did find several significant differences between disease and nondisease genes, such differences only reflect a small proportion of the total variation between genes (the discriminant analysis indicated that less than 3% of the variation in predictive factors was due to the difference between disease and nondisease genes), and so the use of disease genes should not bias our results to any great extent (see also Makalowski et al., 1996).

We considered five features from a database of human disease genes (Jimenez-Sanchez et al., 2001) which might be expected to affect selection and hence the Ka/Ks ratio: (1) function of the protein product, (2) disease frequency in the human population, (3) mode of inheritance, (4) age of onset of clinical manifestations, and (5) reduction of life expectancy (Jimenez-Sanchez et al., 2001). Note that in order to predict the effects of the features on whole gene substitution rates, we are extrapolating from the characteristics of disease mutations to those of the whole gene (except for the protein function analysis). Some of the disease character categories with few representatives were reclassified so to improve statistical power: all modes of inheritance other than autosomal recessive and autosomal dominant were reclassified as "unknown," and the protein function categories of hormone, extracellular transporter, and cell signaling were reclassified as "other."

Table 2
Comparison of Ka/Ks for different protein function classes within the disease genes

| Protein function | $n$ | Ka/Ks mean | S.E. |
|---|---|---|---|
| Channel | 17 | 0.054 | 0.022 |
| Transcription factor | 37 | 0.066 | 0.018 |
| Intracellular protein component | 15 | 0.079 | 0.024 |
| Extracellular protein component | 13 | 0.088 | 0.025 |
| Enzyme | 143 | 0.096 | 0.012 |
| Transmembrane transporter | 18 | 0.108 | 0.021 |
| Modulator of protein function | 46 | 0.135 | 0.018 |
| Receptor | 46 | 0.154 | 0.017 |

Table 3
Comparison of Ka/Ks for different reduction of life expectancy classes within the disease genes

| Reduction of life expectancy | $n$ | Ka/Ks mean | S.E. |
|---|---|---|---|
| None | 64 | 0.088 | 0.017 |
| Mild (death after 60 years) | 40 | 0.155 | 0.021 |
| Moderate (death between puberty and 60) | 62 | 0.131 | 0.015 |
| Severe (death before puberty) | 65 | 0.153 | 0.017 |

As a preliminary analysis, we looked for the effects of all five features on Ka/Ks separately, without considering the correlations between features. One-way ANOVAs, with genes grouped according to disease character category, revealed that the function of the protein product, the mode of inheritance, and the reduction of life expectancy all had significant effects on Ka/Ks ($p < 0.05$). However, given the strong covariance of protein function and the four other disease characters considered here (see Jimenez-Sanchez et al., 2001), it is important to confirm these results using multiway analysis of variance to correct for interactions between disease features. So we performed a multiway univariate analysis of variance using SPSS version 11 with a type I model and the five disease features (protein function, disease frequency, age of onset, reduction of life expectancy, and mode in inheritance) considered as fixed factors. The multiway ANOVA confirmed significant effects on Ka/Ks of protein function ($F = 6.46$, $p < 10^{-8}$), mode of inheritance ($F = 6.01$, $p = 0.003$), and reduction of life expectancy ($F = 2.74$, $p = 0.030$).

Let us consider the effect of protein function on substitution rates in more detail. A relationship between functional classification and Ka/Ks has previously been reported in human–mouse comparisons using Gene Ontology terms (Nembaware et al., 2002), but the use of a different protein function classification means that we cannot easily compare results. Much of the Ka/Ks variation in our data set is attributable to the slow evolving channels and transcription factors and fast evolving receptors (see Table 2; all post hoc pairwise comparisons of known protein function categories which are significant after Bonferroni correction [see p. 240 in Sokal and Rohlf, 1995] involve at least one of these categories). Under strict neutral theory, the variation in Ka/

Table 4
Comparison of Ka/Ks for different mode of inheritance classes within the disease genes

| Mode of inheritance | $n$ | Ka/Ks mean | S.E. |
|---|---|---|---|
| Autosomal recessive | 150 | 0.127 | 0.008 |
| Autosomal dominant | 94 | 0.087 | 0.010 |
| Unknown and other modes | 143 | 0.118 | 0.009 |

Ks is largely due to variation in the proportion of amino acids under functional constraint: the implication is that a particularly high proportion of amino acids are strongly constrained in transcription factors and channels, and a particularly low proportion of amino acids are strongly constrained in receptors. Another explanation for the low Ka/Ks of transcription factors is due to lower levels of duplication and possible adaptive evolution: transcription factors are less likely to undergo gene duplication than other genes (Conant and Wagner, 2002), and Ka/Ks is higher in those genes for which paralogs are found (Nembaware et al., 2002). The Ka/Ks of receptors might similarly be affected by positive selection: many of the two best known types of positively selected genes, immune and reproductive genes, are receptors (Yang and Bielawski, 2000). A further possible explanation for low Ka/Ks is that transcription factors often regulate many downstream genes, hence, an amino acid change in a transcription factor is highly unlikely to be neutral or beneficial to all its downstream targets (Sutton and Wilkinson, 1997).

How can we explain the significant effect of the reduction of life expectancy on Ka/Ks? In general, selection is expected to be weaker after the cessation of reproduction, so a lesser reduction of life expectancy should be expected to be less deleterious and thus cause higher substitution rates under nearly neutral theory than a greater reduction of life expectancy. This prediction utterly fails to explain the observed patterns; indeed, the strongest pattern is that Ka/Ks is lowest when there is no reduction in life expectancy (see Table 3, the only post hoc pairwise comparison which is significant after Bonferroni correction is between no reduction in life expectancy and a severe reduction). Thus, negative selection would appear to be strongest in those genes which cause no reduction in life expectancy. This result can be taken as evidence that the numerous assumptions involved in predicting substitution rates may not be justified (see Section 3.2). In particular, these results suggest that the fitness effect of a disease may be a poor predictor of the fitness effect of all the mutations in the disease gene.

The mode of inheritance of mutations is expected to affect substitution rates according to classical population genetics (Kimura, 1983; Charlesworth et al., 1987). Disease genes which are autosomal and recessive evolve with much higher Ka/Ks than disease genes which are autosomal and dominant (mean 46% higher, see Table 4). If the mode of inheritance is assumed to be similar for all nonsynonymous mutations in a gene, then this result is consistent with nearly neutral theory in which protein evolution occurs by the fixation of slightly deleterious mutations.

## References


Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, T., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A., Zdobnov, E.M., 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29, 37–40.

Bielawski, J.P., Dunn, K.A., Yang, Z.H., 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

Brookfield, J.F.Y., 2000. Evolution: what determines the rate of sequence evolution? Curr. Biol. 10, R410–R411.

Charlesworth, B., Coyne, J.A., Barton, N.H., 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130, 113–146.

Conant, G.C., Wagner, A., 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. Nucleic Acids Res. 30, 3378–3386.

Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.

Duret, L., Mouchiroud, D., Gautier, C., 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC rich isochores. J. Mol. Evol. 40, 308–317.

Eyre-Walker, A., Keightley, P.D., Smith, N.G.C., Gaffney, D., 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol. Biol. Evol. 19, 2142–2149.

Ferrer-Costa, C., Orozco, M., de la Cruz, X., 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J. Mol. Biol. 315, 771–786.

Hurst, L.D., Smith, N.G.C., 1999. Do essential genes evolve slowly? Curr. Biol. 9, 747–750.

Hurst, L.D., Williams, E.J.B., 2000. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. Gene 261, 107–114.

Jimenez-Sanchez, G., Childs, B., Valle, D., 2001. Human disease genes. Nature 409, 853–855.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P., Ahringer, J., 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421, 231–237.

Karlin, S., Chen, C., Gentles, A.J., Cleary, M., 2002. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. Proc. Natl Acad. Sci. U. S. A. 99, 17008–17013.

Keightley, P.D., Eyre-Walker, A., 2000. Deleterious mutations and the evolution of sex. Science 290, 331–333.

Kimura, M., 1983. The Neutral Theory of Evolution. Cambridge Univ. Press, Cambridge.

Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183.

Li, W.H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Makalowski, W., Zhang, J., Boguski, M.S., 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res. 6, 846–857.

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

Nembaware, V., Crum, K., Kelso, J., Seoighe, C., 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse–human orthologs. Genome Res. 12, 1370–1376.

Notaro, R., Afolayan, A., Luzzatto, L., 2000. Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. FASEB J. 14, 485–494.

Ohta, T., Ina, Y., 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. 41, 717–720.

Smith, N.G.C., Hurst, L.D., 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics 153, 1395–1402.

Sokal, R.R., Rohlf, F.J., 1995. Biometry. W.H. Freeman, New York.

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

Sutton, K.A., Wilkinson, M.F., 1997. Rapid evolution of a homeodomain: evidence for positive selection. J. Mol. Evol. 45, 579–588.

Tourasse, N.J., Li, W.H., 2000. Selective constraints, amino acid composition, and the rate of protein evolution. Mol. Biol. Evol. 17, 656.

Wilson, A.C., Carlson, S.S., White, T.J., 1977. Biochemical evolution. Ann. Rev. Biochem. 46, 573–639.

Xia, X.H., Li, W.H., 1998. What amino acid properties affect protein evolution? J. Mol. Evol. 47, 557–564.

Xia, X., Xie, Z., 2002. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. Mol. Biol. Evol. 19, 58–67.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z.H., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15, 496–503.

Zar, J.H., 1999. Biostatistical Analysis. Prentice-Hall, Upper Saddle River, NJ.