Am. J. Hum. Genet. 63:474–488, 1998

Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes

Michael Krawczak, Edward V. Ball, and David N. Cooper

Institute of Medical Genetics, University of Wales College of Medicine, Cardiff

Summary

The spectrum of single-base-pair substitutions logged in The Human Gene Mutation Database (HGMD), comprising 7,271 different lesions in the coding regions of 547 different human genes, was analyzed for nearest-neighbor effects on relative mutation rates. Owing to its retrospective nature, HGMD allows mutation rates to be estimated only in relative terms. Therefore, a novel methodology was devised in order to obtain these estimates in iterative fashion, correcting, at the same time, for the confounding effects of differential codon usage and for the fact that different types of amino acid replacement come to clinical attention with different probabilities. Over and above the hypermutability of CpG dinucleotides, reflected in transition rates five times the base mutation rate, only a subtle and locally confined influence of the surrounding DNA sequence on relative single-base-pair substitution rates was observed, which extended no farther than 2 bp from the substitution site. A disparity between the two DNA strands was evidenced by the fact that, when substitution rates were estimated conditional on the 5′ and 3′ flanking nucleotides, a significant rate difference emerged for 10 of 96 possible pairs of complementary substitutional events. Mutational bias, favoring substitutions toward flanking bases, a phenomenon reminiscent of misalignment mutagenesis, was apparent and exhibited both directionality and reading-frame sensitivity. No specific preponderance of repeat-sequence motifs was observed in the vicinity of nucleotide substitutions, but a moderate correlation between the relative mutability and thermodynamic stability of DNA triplets emerged, suggesting either inefficient DNA replication in regions of high stability or the transient stabilization of misaligned intermediates.

Received February 9, 1998; accepted for publication May 28, 1998; electronically published July 10, 1998.

Address for correspondence and reprints: Dr. Michael Krawczak, Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, United Kingdom. E-mail: krawczak@cardiff.ac.uk

0002-9297/98/6302-0023$02.00

Introduction

The majority of germ-line mutations in human genes are thought to result from error-prone endogenous processes involving either chemical (e.g., methylation-mediated deamination of 5-methylcytosine in CpG dinucleotides), physical (e.g., DNA slippage), or enzymatic (e.g., postreplicative mismatch repair and exonucleolytic proofreading) mechanisms (Cooper and Krawczak 1993). Since the efficiency of these processes is DNA-sequence dependent, it is not surprising that both the spectrum and spatial distribution of mutations exhibit biases that reflect the influence of the local DNA-sequence environment on germ-line mutability. A large proportion of microdeletions and microinsertions in human DNA thus appear to occur as the consequence of replication slippage mediated by the presence of direct or inverted (palindromic) repeats in the immediate vicinity (Cooper and Krawczak 1993).

For single-base-pair substitutions, which constitute the majority of known lesions causing human genetic disease, the influence of the local DNA-sequence environment on mutation rates is less clear. Apart from the well-established hypermutability of CpG dinucleotides that undergo germ-line transition to TG and CA at frequencies six to seven times the base mutation rate (Cooper et al. 1995), no clear evidence has yet been presented for a strong nearest-neighbor effect on the nature of inherited nucleotide substitutions in humans in vivo. Previous studies based on phylogenetic (e.g., see Golding and Glickman 1986) or clinical data (e.g., see Todorva and Danieli 1997) have either been purely anecdotal or, instead, have been focused on single genes or gene families. A broad-based analysis of the spectrum of single-base-pair substitutions has only recently become possible, through the establishment of the The Human Gene Mutation Database (HGMD [Krawczak and Cooper 1997]), a comprehensive literature-based collection of mutations either underlying or associated with human inherited disease. Containing >7,200 different nucleotide substitutions from the coding regions of some 550 different genes, HGMD provides an unparalleled source of material for the metanalysis of germ-line mutations in human genes.

Owing to the retrospective nature of HGMD data, however, nucleotide-substitution rates can be assessed only in relative terms. Furthermore, for a mutation to be represented in HGMD, it must result in a phenotype severe enough to have come to clinical attention but, at the same time, must not be subject to preclinical selection. In order to take these considerations into account, a new analytical approach has been devised that models the likelihood that a mutation in HGMD will be observed as a function of its consequences at both the DNA and protein level (Krawczak et al. 1995; Krawczak and Cooper 1996). In an iterative procedure, estimates are obtained simultaneously for the relative occurrence rate of a given substitution type, with allowance for flanking nucleotides, and for the *relative clinical observation likelihood* of its resulting phenotype. The large number of different genes analyzed here ensures that the biases that reflect the private characteristics of individual proteins and that hamper studies of single genes or gene families are likely to be averaged out.

Our study reveals that, when neighboring mononucleotides and dinucleotides are considered as covariates, their effect on single-base-pair mutagenesis is locally confined. This finding has several important practical and theoretical implications: consideration of nearest-neighbor–dependent substitution rates should help to optimize mutation-search strategies, render phylogenetic reconstruction in molecular-evolutionary studies more realistic (Krawczak et al. 1996), and allow the private characteristics of gene- or tissue-specific mutational spectra to be identified, thereby providing the basis for a better understanding of the molecular mechanisms underlying single-base-pair substitutions in general.

Material and Methods

HGMD

HGMD is a comprehensive collection of germ-line mutations underlying or associated with human inherited disease, comprising published single-base-pair substitutions, deletions, duplications, insertions, and more-complex rearrangements in human nuclear genes (Krawczak and Cooper 1997). Although originally established for research purposes (Cooper and Krawczak 1993), the database has since acquired a much broader utility and, for this reason, was made publicly available, through the Internet, in April 1996.

By November 1997, HGMD contained 7,271 different single-base-pair substitutions (5,862 missense and 1,409 nonsense) in the coding regions of 547 different genes. HGMD entries for this category of lesion include the triplet change, with an additional flanking nucleotide logged when the mutated base lies in either the first or third position in the triplet. This information allows nearest-neighbor effects on substitution rates to be readily assessed for the 5′ and 3′ nucleotides flanking the site of mutation. For a subset of 423 of the genes, reference cDNA sequences (starting with the ATG initiation codon and ending with the stop codon) also were available, allowing the analysis of the broader DNA-sequence context for 6,885 substitutions (94.7% of the total).

It should be noted that each nucleotide substitution in HGMD has been logged only once, in order to avoid confusion between recurrent and identical-by-descent lesions. Although necessitating the systematic exclusion of multiple independent de novo mutations, this restriction is considered unlikely to bias the estimates of relative substitution rates and clinical observation likelihoods to any marked extent, because of the large sample size.

Estimation of Relative Single-Base-Pair Substitution Rates and Relative Clinical Observation Likelihoods

With the possible exception of biologically lethal dominant conditions, germ-line mutational spectra associated with genetic disease do not allow single-base-pair substitution rates to be estimated directly and in absolute terms. This shortcoming is due to the fact that, without extensive haplotyping, discrimination between recurrent mutation and identity by descent is impossible. Even if two identical-by-state lesions were recognized to be of independent origin, the actual number of meioses screened for their occurrence would remain unknown. Additionally, for a mutation to be found to be associated with disease, it must cause a phenotype sufficiently severe that it comes to clinical attention, but, at the same time, it must not be selected against antenatally or preclinically. This implies that clinically observed mutation frequencies may not reflect directly the underlying rates of occurrence.

To disentangle the effects that mutation and selection have on an observed mutational spectrum such as that found in HGMD, we have devised an iterative multivariate procedure (Cooper and Krawczak 1993; Krawczak et al. 1995; Krawczak and Cooper 1996) that takes into account the phenotypic consequences of mutation and that estimates, in relative terms, single-base-pair substitution rates and clinical observation likelihoods of phenotypic consequences. A detailed presentation of this approach is given in the Appendix.

In brief, the algorithm aims to maximize the overall likelihood of a given mutation sample, assuming that the likelihood that an individual mutation of primary type x and phenotypic consequence α will be observed equals the product of the probability of occurrence of x at the DNA level, $\mu(x)$, and the clinical observation likelihood of α, $L(\alpha)$. Maximum-likelihood estimates of relative μ and L values, also corrected for human gene-codon usage and the redundancy of the genetic code,

are obtained in iterative fashion. In the present study, mutation type x was defined either as the nucleotide substitution on its own (e.g., C→T, G→A, or A→T), as the substitution conditional on a flanking mononucleotide (e.g., CG→TG, CNT→ANT, or $GN_3A \rightarrow GN_3T$), or as the substitution conditional on a flanking dinucleotide (e.g., ATC→ATA, $TN_2TC \rightarrow AN_2TC$, or $CGN_3C \rightarrow CGN_3A$).

Phenotypic consequence α was measured in terms of the *chemical difference* between wild-type and mutant amino acid residues. This parameter, originally devised by Grantham (1974) to assess the net effect of amino acid exchanges in evolutionary comparisons, combines the three interdependent properties of composition, polarity, and molecular volume in a single, continuous quantity. For the sake of simplicity, however, the range of chemical-difference values was divided here into 11 equally sized intervals, thereby transforming α into a class variable; an additional class was introduced for nonsense mutations.

Significance Assessment (*Nearest-Neighbor Effects*)

The variability of $\mu(x)$ and $L(\alpha)$ estimates was assessed by bootstrapping, with each SD determined in 10,000 resampling simulations (Hjorth 1994). In order to test relative substitution rates for a potential strand difference, mutations were defined in terms of the immediately 5′ and 3′ flanking nucleotides (e.g., x = CTG→CAG). Estimates of $\mu(x)$ were then compared with $\mu(x^c)$, the analogous estimate for the complementary sequence (e.g., x = ATC→AGC, and x^c = GAT→GCT). Differences were deemed to be significant when $\mu(x) > \mu(x^c)$ or $\mu(x) < \mu(x^c)$ in 9,995 of 10,000 bootstrapping simulations. This threshold was adopted to allow an overall 95% significance level for the 96 comparisons involved.

The influence of neighboring mononucleotides and dinucleotides on relative single-base-pair substitution rates was assessed as follows. For a given substitution x (e.g., x = C→A), let x_i denote the mutation type defined by mononucleotide or dinucleotide η_i flanking substitution x. If it is assumed that all η_i are, a priori, equally likely to flank the wild-type nucleotide in x, then $P(\eta_i{:}x) = \mu(x_i)/\Sigma_j \mu(x_j)$ can be interpreted as the posterior probability of η_i, given that x has occurred. In the absence of any neighboring-nucleotide effects, $P(\eta_i{:}x)$ would equal $\frac{1}{4}$ or $\frac{1}{16}$ for all mononucleotides or dinucleotides η_i, respectively, so that the Euclidean distance between $P = \{P_i\} = \{P(\eta_i{:}x)\}$ and the centroid $\varepsilon = \{\varepsilon_i\}$, with $\varepsilon_i = \frac{1}{4}$ or $\varepsilon_i = \frac{1}{16}$, $d(P,\varepsilon) = [\Sigma_i (P_i - \varepsilon_i)^2]^{\frac{1}{2}}$ is a good measure of the influence of the flanking mononucleotide or dinucleotide on the relative rate of x. Whether P was significantly different from ε was again determined by bootstrapping. When the P vector emerging from the original data set was denoted by "P_o" and when the vectors from bootstrapping were denoted by "P_r," P_o was deemed significantly different from ε when

$$d(P_o,P_r) < \max[d(P_o,\varepsilon),d(P_r,\varepsilon)] \ , \qquad (1)$$

for >9,995 of 10,000 replicates P_r. Formula (1) can be interpreted as P_r being closer to P_o than ε, or at least on the same "side" of ε as is P_o. Since 10 positions surrounding each of the 12 possible substitutions were considered, this analysis involved 120 comparisons. The threshold noted above thus ensured a 95% overall significance level.

Significance Assessment (*Codon Usage*)

In a region comprising four amino acid residues both upstream and downstream of missense mutations, the relative use of ambiguous codons was tested for significant features, by means of a χ^2 statistic with 1 df. Expected frequencies were calculated from published codon-usage data for human genes (Nakamura et al. 1996). Since 55 comparisons were involved, a threshold of $0.05/55 = 9.09 \times 10^{-4}$ was adopted for the error probability. Absolute codon usage surrounding nonsense mutations was related to the total 448 HGMD reference cDNA sequences. To this end, codon usage was determined in the reference cDNA sequences around codons with the potential to mutate to a termination codon by single-base-pair substitution. The frequencies obtained were subsequently weighted, for each potentially mutable codon, in terms of its actual frequency in the HGMD sample. Significance was assessed for each of the 64 surrounding codons, by means of a χ^2 statistic with 1 df, with adoption of a threshold of $0.10/64 = 1.56 \times 10^{-3}$ for the error probability. A lower overall significance level of 90% was chosen, in order to allow for the considerably smaller number of nonsense mutations available for analysis.

Significance Assessment (*Mutational Bias toward Flanking Nucleotide*)

The importance of slippage-mediated misincorporation for the occurrence of single-base-pair substitutions was assessed by means of the relative proportion of mutations toward the immediately 5′ or 3′ flanking nucleotide. This study necessarily was limited to sites at which the wild-type nucleotide was different from the flanking base. In addition, CG→TG and CG→CA transitions explicable by the deamination of 5-methylcytosine were discarded (1,675/7,271, or 23% of the total), since the relative abundance of this type of mutation would have served to obscure any mechanistic relationship underlying the remaining substitutions. The number of mutations toward the immediately 5′ or 3′ flanking nucleotide that were expected under the null hypothesis that

substitution toward a flanking nucleotide occurs with probability $\frac{1}{3}$ was determined as described in the Appendix. Observed frequencies were tested for significant deviation from these expectations by means of a χ^2 statistic with 1 df.

Results

Throughout the following sections, a shortened form will be used to denote classes of single-base-pair substitutions and their sequence context. For any oligomers A, B, and C, "(A,B)→C" is equivalent to "A→C, B→C" whereas "A→(B,C)" means "A→B, A→C." For example, CG→(TG,CA) denotes the class of CG→TG and CG→CA transitions.

Relative Single-Base-Pair Substitution Rates

The spectrum of single-base-pair substitutions within gene-coding regions and logged in HGMD reveals a hierarchy of nucleotides that is clear-cut with respect to their propensity to undergo substitution—namely, G > C > T > A (table 1). Consistent with previous observations (Cooper and Krawczak 1990, 1993), a preponderance of transitions (62.5%) over transversions (37.5%) was observed. Most but not all of this excess can be attributed to CG→(TG,CA) mutations, which are readily explicable in terms of methylation-mediated deamination of 5-methylcytosine (5mC) on either the sense or antisense DNA strands. This type of lesion accounts, on its own, for 23.0% of all substitutions and for 36.9% of transitions. Breakdown of the data by chromosomal location revealed, however, that the proportion of CG→(TG,CA) substitutions was significantly higher for autosomal genes (1,325/5,296, or 25.0%) than for X-chromosomal genes (350/1,975, or 17.7%) ($\chi^2 = 43.21$, 1 df, $P < 10^{-5}$). In part, this disparity can be explained by a generally more pronounced CpG suppression observed in X-linked genes: the average CpG content was .0367 ± .0224 for the 401 autosomal cDNA sequences provided by HGMD and was .0286 ± .0164 for the 45 X-chromosomal cDNAs (Student's $t = 2.35$, 444 df, $P < .01$). When the CG:GC ratio was considered, in order to allow for differing G+C content in different genes and/or chromosomes, the respective average values were .4191 ± .1492 for the X chromosome and .4610 ± .1615 for autosomes ($t = 1.66$, 444 df, $P < .05$).

As outlined above (see the Material and Methods section), mutation frequencies observed in the context of human inherited disease are unlikely to reflect the true underlying rates of mutation occurrence. Since different amino acid substitutions have different effects on protein structure and function, they necessarily have come to clinical attention (and thus have entered HGMD) with different probabilities. Moreover, codon frequencies differ from one another, implying that, in a mutational event, different amino acid residues have different prior probabilities of being involved. Relative single-base-pair substitution rates corrected for these two confounding factors are presented in table 2. Although the rate estimates are broadly consistent with the frequency data in table 1, some differences are nevertheless apparent. Thus, transversions (G,A)→T are estimated to occur at only half the rates suggested by their absolute frequencies. By contrast, the corrected transition rate of T→C is some 50% higher than its frequency-based counterpart. A similar, albeit smaller bias also applies to the other three transitions (namely, C→T, G→A, and A→G).

Table 1

Spectrum of Observed Single-Base-Pair Substitutions in Gene-Coding Regions, Logged in HGMD

Original Nucleotide	No. of Substitutions by				Total
	T	C	A	G	
T	...	654	271	312	1,237
C	1,632 (940)[a]	...	371	340	2,343
A	201	163	...	538	902
G	619	453	1,717 (735)[b]	...	2,789
Total	2,452	1,270	2,359	1,190	7,271

[a] Number in parentheses is the proportion of transitions that are CG→TG.

[b] Number in parentheses is the proportion of transitions that are CG→CA.

Neighboring-Nucleotide Effects

It has been known for some time (Cooper and Youssoufian 1988) that the rate of C→T and G→A transitions in human genes is greatly increased by the presence of a 3′ guanine or 5′ cytosine residue, respectively, owing to the abundance of hypermutable methylated CpG dinucleotides (Bird 1986). The question therefore arises as to whether a neighboring-nucleotide effect might exist also for other types of nucleotide substitution. Figure 1 depicts the influence on relative single-base-pair substitution rates that is exerted by the five nucleotide positions upstream and downstream of a potential mutation site. Each data point represents 1 of 12 possible single-base-pair substitutions. Neighboring-nucleotide effects on relative mutation rates are expressed in terms of a vector-based parameter, the distance from centroid, that measures, at a given position, the deviation from independence. The larger this distance, the more skewed is the posterior probability that particular mononucleotides will be observed in the vicinity of a given substitution (see the Material and Methods section). The complete data set is available for inspection, on the HGMD Website.

Two outliers mark the strong effects that the −1 (5′)

and +1 (3′) positions have on G→A and C→T transitions, respectively. A significant influence also was noted, however, for some other substitutions: T→(C,A), A→(C,G), and G→(T,C) are biased by the nucleotide at position −1, whereas T→(C,G), C→(G,A), A→(T,G), and G→T are biased by the nucleotide at position +1. Nevertheless, it is evident that the nearest-neighbor influence decreases markedly with distance from the site of substitution (fig. 1). A significant, albeit weak effect was observed for position +2, but only for five substitutions—T→C, C→T, A→G, and G→(T,C). Interestingly, the six substitutions significantly influenced by the −1 (5′) nucleotide can be matched with their complementary substitutions being significantly influenced by the +1 (3′) neighbor, and vice versa. When nearest-neighbor effects were analyzed with allowance for neighboring dinucleotides rather than mononucleotides, more substitutions tended to exhibit a statistically significant but weaker rate dependency (data not shown). Nevertheless, the effect again was confined to a small region, unlikely to extend beyond positions −2 and +3.

In order to illustrate the impact of the computational method described in the Appendix, estimates of relative single-base-pair substitution rates, simultaneously allowing for both flanking nucleotides, also were calculated, without correction for any confounding factors. These estimates, $\mu'(x)$, simply equated the absolute number of reports of a given mutation type x multiplied by a constant factor so as to ensure that the average μ' value was unity. When allowance was made for multiple testing at an overall 95% level (192 comparisons), differences between the two estimates were deemed significant when $|\mu(x) - \mu'(x)| > 3.5 \cdot \sigma$. Here, σ denotes the SD of $\mu(x)$. In total, μ' represented a significant overestimate of μ for 51 substitutions (fig. 2), none of which involved CG→(TG,CA). On the other hand, with the exception of CCG→CTG and GCG→GTG, all rates of CG→(TG,CA) transitions would have been significantly underestimated by μ'.

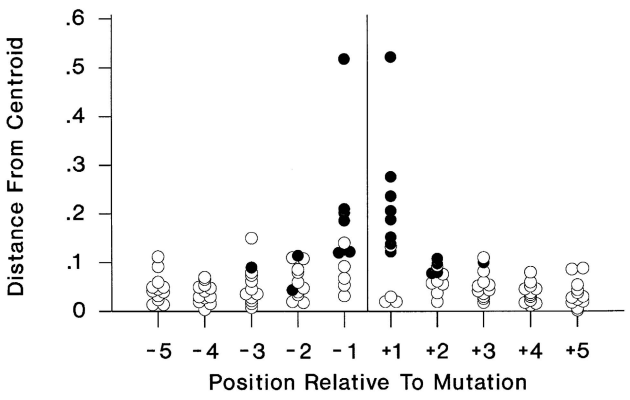

Figure 1 Influence of flanking mononucleotides on bias-corrected relative single-base-pair substitution rates. Each circle represents one of 12 possible substitutions. The vertical axis measures the importance of a flanking position for the respective substitution rate in terms of a vector-based parameter ("Distance from Centroid"; see the Material and Methods section). Substitutions with a significant neighboring-nucleotide effect are denoted by the blackened circles.

A Strand Difference in Single-Base-Pair Substitution Rates

Inspection of the relative rates of CG→(TG,CA) transitions as estimated conditional on either the upstream or downstream nucleotide, respectively (table 3), suggests that methylation-mediated deamination of CpG dinucleotides is significantly biased by the 5′ flanking nucleotide on the noncoding DNA strand (table 3, rows 1–4) but not on the coding strand (table 3, rows 5–8). By contrast, the nucleotide immediately downstream of a CpG appears to be significant for CG→TG transitions, irrespective of the DNA strand involved. Here, 3′ adenine residues are associated with the highest relative substitution rates (table 3, rows 9–12 and 13–16). Also included in table 3 (rows 17–19 and 20–22) are relative CNG→(TNG,CNA) transition rates, allowing us to address the question as to whether cytosine methylation and consequent high-frequency deamination might also

Table 2

Relative Single-Base-Pair Substitution Rates in Human Nuclear Genes Causing Inherited Disease

Original Nucleotide	Relative Substitution Rate ± SD[a]			
	T	C	A	G
T	…	1.525 ± .062	.374 ± .023	.410 ± .024
C	2.702 ± .068	…	.541 ± .028	.505 ± .028
A	.187 ± .014	.268 ± .022	…	1.127 ± .051
G	.521 ± .023	.712 ± .035	3.128 ± .078	…

[a] Based on HGMD data and corrected for confounding effects as described in the Appendix. The estimates are unitless and have been scaled so that their average, taken over all 12 substitution types, is unity.

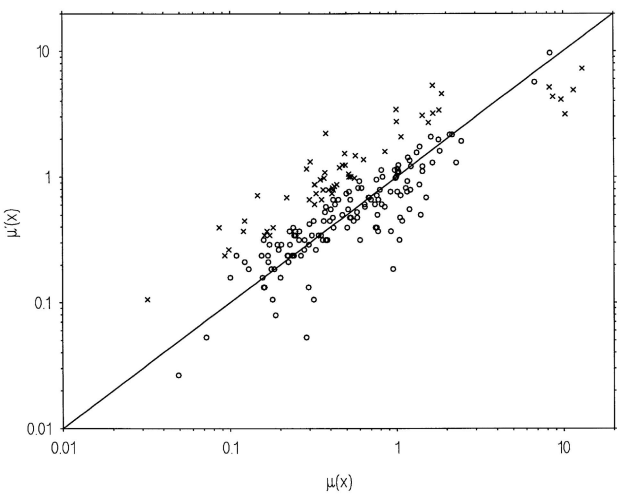

Figure 2 Bias-corrected versus uncorrected estimates of relative single-base-pair substitution rates, conditional on 5′ and 3′ flanking mononucleotides. x = mutation type; $\mu(x)$ = corrected estimate of relative substitution rate, derived as described in the Appendix; and $\mu'(x)$ = uncorrected estimate, based solely on the observed frequency of x. Pairs for which $|\mu(x) - \mu'(x)|$ is >3.5 SD of $\mu(x)$ are denoted by crosses (×).

occur at these motifs. Although the relative rates of CNG→TNG and CNG→CNA transition are not substantially higher than the average (i.e., unity) when N = G is excluded, a flanking-nucleotide effect nevertheless is noted: the data are consistent with CNG undergoing transition to TNG, on both strands, at a 50% higher rate when N = A than when N = T or N = C.

Depicted in figure 3 are relative single-base-pair substitution rates taking into account the single nucleotides immediately flanking, on its 5′ and 3′ sides, the mutated base. Each data point represents a pair of mutations, with one substitution being the complementary homologue of the other (e.g., GCA→GTA vs. TGC→TAC). Were substitution rates identical on both DNA strands, then the data points should approximate the 45° line. However, 10 significant outliers were identified (fig. 3), and these are listed in table 4. Purines and pyrimidines were involved in five cases each, with G and T consistently showing rates of substitution higher than those for C and A. Interestingly, one substitution pair involves CG→(TG,CA) transitions, and the G→A rate was found to exceed that of C→T. The same relationship holds for the other three CGN→CAN/N^{c}CG→N^{c}TG pairs (where N^c denotes the base complementary to N), although the corresponding rate differences failed to attain statistical significance once multiple testing had been allowed for. Nevertheless, individual error probabilities were .001 for N = T/N^c = A and .002 for N = G/N^c = C, indicative of a similarly strong bias for these substitution pairs. Only for N = A/N^c = T (P = .105) could a strand difference in relative substitution rate possibly be ruled out. At first sight, the higher rate for CG→CA than for CG→TG might appear to be inconsistent with the data presented in table 1 (ratio of absolute frequencies 735: 940, or .78). This discrepancy is, however, explicable in terms of CG→TG transitions being more likely to disrupt protein function than are CG→CA transitions (CGA→TGA transitions create termination codons, whereas no CG→CA transition yields a nonsense mutation).

Relative Clinical Observation Likelihoods

In the process of computing relative single-base-pair substitution rates conditional on both the 5′ and 3′ flanking nucleotides, we also estimated the relative clinical observation likelihoods of different amino acid replacements as classified by the chemical difference between the respective wild-type and mutant residues. A steady increase in clinical observation likelihood with increasing chemical difference was apparent (fig. 4). Moreover, nonsense mutations were found to be more than twice

Table 3

Neighboring-Nucleotide Effects on Relative CG→(TG,CA) and CNG→(TNG,CNA) Transition Rates

Substitution	Relative Substitution Rate ± SD[a]
CGT→CAT	10.255 ± .973
CGC→CAC	9.735 ± .787
CGA→CAA	11.527 ± .871
CGG→CAG	13.023 ± .880
ACG→ATG	8.687 ± .683
GCG→GTG	6.762 ± .491
TCG→TTG	8.276 ± .641
CCG→CTG	8.340 ± .475
TCG→TCA	15.219 ± 1.096
CCG→CCA	12.350 ± .761
ACG→ACA	10.695 ± .912
GCG→GCA	11.310 ± .844
CGA→TGA	11.888 ± .602
CGG→TGG	8.656 ± .543
CGT→TGT	7.919 ± .825
CGC→TGC	6.744 ± .498
CTG→CTA	1.815 ± .150
CAG→CAA	1.159 ± .177
CGG→CGA	1.261 ± .250
CAG→TAG	1.329 ± .084
CTG→TTG	.896 ± .168
CCG→TCG	.738 ± .196

[a] Based on HGMD data and corrected for confounding effects as described in the Appendix. The estimates are unitless and have been scaled so that they yield an average of unity for the respective dinucleotide position; relative rates therefore can be related to one another only within a subcolumn.

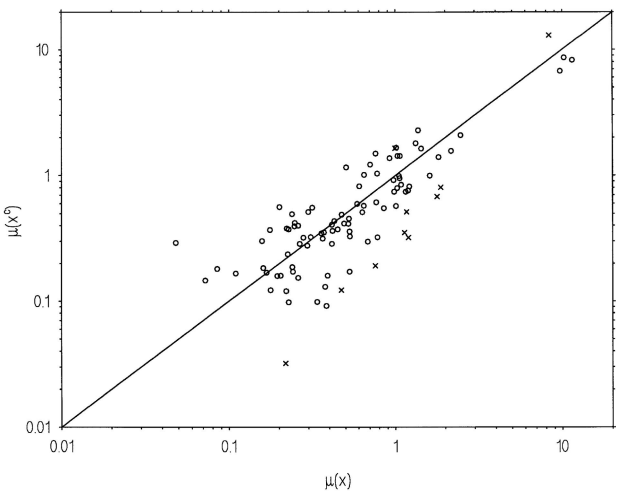

Figure 3 Bias-corrected relative single-base-pair substitution rates, conditional on 5′ and 3′ flanking mononucleotides. Each circle represents a pair of substitution rates, $\mu(x)$ and $\mu(x^c)$, with one substitution (x^c) being the complementary homologue of the other x. Pairs with significantly different nearest-neighbor–dependent rate estimates are denoted by crosses (×).

as likely to come to clinical attention as the most extreme missense mutations (chemical difference 2.0–2.2) and were three times more likely to come to clinical attention than the average amino acid change.

Codon Usage around Substitution Sites

An alternative means to study the importance of the immediate DNA-sequence environment for single-base-pair mutagenesis would be to try to identify mutation-associated sequence motifs directly. At least for missense mutations, however, this type of analysis would have been seriously hampered by the difficulty in controlling for the selective effects of the amino acid–sequence context. It therefore was decided to confine the investigation of missense mutations to surrounding synonymous codon usage—that is, by comparison of the relative frequencies only for neighboring triplets encoding one and the same amino acid (table 5). Nonsense mutations, on the other hand, are, on average, three times more likely to come to clinical attention than are missense mutations (see above), suggesting that selection acting on the immediate amino acid–sequence context of such lesions is less likely to confound their observed spectrum. General codon usage therefore was analyzed for significant features around sites of nonsense mutation (table 6).

A total of 15 codons were found to be significantly over- or underrepresented in the vicinity (i.e., four amino acid residues upstream and downstream) of missense mutations. Intriguingly, considerable overlap with the results obtained for nonsense mutations was noted: all codons but one (CTA) that were associated with nonsense mutations also were found to be (relatively) abundant around amino acid replacements. However, only one codon (GGG) was underrepresented in both data sets. Thus, no evidence was found for the existence of "mutation-protective" sequence motifs. The observed paucity of glycine residues around nonsense mutations may be suggestive of structural/functional constraints rather than of mutation repression.

A Role for Slippage-Mediated Misincorporation?

Kunkel (1985) proposed a model that sought to explain—through transient misalignment of the primer template, caused by looping out of a single template base ("misalignment mutagenesis")—nucleotide misincorporation during DNA replication. If this mechanism of slippage-mediated misincorporation were to play an important role in the generation of single-base-pair substitutions in human genes, then a substantial proportion of mutations in the HGMD data set should exhibit identity between the newly introduced base and one of the bases immediately flanking the site of mutation. The observed and expected frequencies presented in table 7 show that this is indeed the case, but only at certain codon positions. Mutation toward the 5′ flanking nucleotide has occurred significantly more often than expected at the second position of the codon but not at the first or last position; mutation toward the 3′ flanking base is favored at the first position of a codon but is disfavored at the second position. These findings suggest a mutation mechanism, at position 1 and position 2 in the codon (both of which are critical in the specification of the encoded amino acid residue), that is biased toward

Table 4

Strand Difference in Relative Single-Base-Pair Substitution Rates

Original Substitution (Relative Substitution Rate ± SD[a])	Watson Crick Homologue (Relative Substitution Rate ± SD[a])
GGT→GTT (1.166 ± .161)	ACC→AAC (.515 ± .082)
TGG→TAG (1.649 ± .128)	CCA→CTA (.994 ± .092)
CGG→CAG (13.009 ± .884)	CCG→CTG (8.351 ± .471)
CTT→CCT (1.136 ± .203)	AAG→AGG (.353 ± .098)
CTC→CCC (1.198 ± .173)	GAG→GGG (.321 ± .068)
TGC→TCC (.757 ± .128)	GCA→GGA (.192 ± .058)
CTG→CCG (1.871 ± .150)	CAG→CGG (.807 ± .126)
GGT→GAT (1.785 ± .207)	ACC→ATC (.680 ± .135)
CTG→CAG (.219 ± .044)	CAG→CTG (.032 ± .016)
CTT→CGT (.471 ± .107)	AAG→ACG (.122 ± .042)

[a] Based on HGMD data and corrected for confounding effects, as described in the Appendix. The estimates are unitless and have been scaled so that their average, taken over all 192 substitution types, is unity. Only substitutions for which one relative-rate estimate was consistently larger than its counterpart in >9,995/10,000 bootstrap simulations are included.

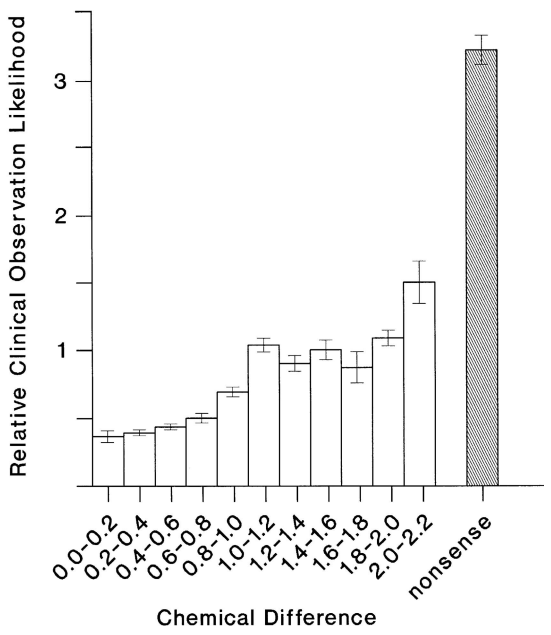

Figure 4 Relative clinical observation likelihood of amino acid substitutions, as a function of chemical difference between substituting and wild-type residues. SDs of estimates determined by bootstrapping (10,000 simulations) are demarcated by vertical bars.

the nucleotide at the other position. Inspection of the genetic code reveals that such a bias invariably serves to avoid the de novo introduction of termination codons.

Mutability and Thermodynamic Stability

The stability and melting behavior of a DNA duplex depends critically on its primary nucleotide sequence. More specifically, Breslauer et al. (1986) demonstrated that, thermodynamically, DNA duplex structures can be considered as the sum of their nearest-neighbor pairwise interactions. From the calorimetric study of 19 DNA oligomers and 9 DNA polymers, Breslauer et al. (1986) compiled tables of the transition enthalpy changes (ΔH) and the transition free-energy changes (ΔG) associated with disruption, at 1 M NaCl, 37°C, and pH 7.0, of any of the 10 possible Watson-Crick nearest-neighbor interactions. The sum of these pairwise values, taken over the primary DNA sequence, was found to correlate strongly with the experimentally observed ΔH and ΔG of 12 DNA oligomers.

In order to assess whether the nearest-neighbor influence on single-base-pair mutagenesis is associated with thermodynamic DNA stability, we attempted to relate ΔH and ΔG of a given triplet, $\omega = \omega_{-1}\omega_0\omega_{+1}$, to the average relative mutation rate at the central nucleotide of the triplet, $\omega_0, \mu_{av}(\omega) = n^{-1} \cdot \sum_{\omega'} \mu(\omega_{-1}\omega_0\omega_{+1} \rightarrow \omega_{-1}\omega'\omega_{+1})$. Here, summation is over all $\omega' \neq \omega_0$, excluding CG→(TG,CA) transitions, and n is the number of such ω' ($n = 2$ or 3). Both thermodynamic parameters of a triplet show a significant positive correlation with μ_{av} (fig. 5) when they are tested by means of a Spearman rank correlation coefficient (ΔH 0.23, $P < .05$; ΔG 0.35, $P < .005$). This implies that any increase in stability of the immediate DNA-sequence environment increases the likelihood that a given nucleotide will undergo substitution. The fact that this correlation is stronger for ΔG than for ΔH (=ΔG at 0°K) is indicative of a temperature-dependent relationship.

Discussion

Single-base-pair substitutions causing human genetic disease may arise via a number of different endogenous mechanisms. Perhaps the most important and best understood is the deamination of 5-methylcytosine, the most abundant chemically modified base in vertebrate genomes and one that is confined almost exclusively to CpG dinucleotides (Cooper and Krawczak 1993). The high rate of 5mC deamination and consequent replacement by thymine (Shen et al. 1994) renders the CpG dinucleotide a hot spot for germ-line mutation in vertebrate genomes. Another mutagenesis model often invoked to account mechanistically for germ-line nucleotide substitutions in humans involves the misincorporation of noncomplementary nucleotides as a consequence of transient template-primer misalignment ("misalignment mutagenesis"; Kunkel 1990, 1992). Misalignment is thought to be mediated by short direct or inverted sequence repeats in the immediate vicinity of the lesions. In somatic tissues, 5mC deamination also appears to be an important mechanism of single-base-pair substitution (Hollstein et al. 1991; Tornaletti and Pfeifer 1995). Indeed, the relative rate of mitotic cancer-associated CG→(TG,CA) transitions observed in the *TP53* gene, the most widely mutated gene in human tumorigenesis, is very similar to the overall germ-line rate observed in other human genes (Krawczak et al. 1995). Some somatic mutations may occur, teleologically speaking, by design—for example, the various preprogrammed hot spots for single-base-pair mutagenesis in the variable (V) regions of the mammalian immunoglobulin genes. Their hypermutability, potentiating the generation of diversity in the immune response, appears to be targeted mainly toward short RGYW (R = A/G; Y = T/C; and W = A/T) and TAA motifs (Rogozin and Kolchanov 1992). In vitro data further suggest the concomitant existence, in V regions, of mutational cold-spot motifs (e.g., TAGA; Lin et al. 1997).

What all the above mechanisms have in common is that the mutational bias involved is DNA-sequence dependent. However, since the available evidence for sequence-dependent mutational bias either has been based on in vitro data or has related only to a small number

Table 5

Codons Significantly Over-/Underrepresented among ±4 Amino Acid Residues Surrounding Missense Mutations

	CG→(TG,CA) EXCLUDED (N = 4,318)			CG→(TG,CA) INCLUDED (N = 5,556)		
CODON	No. of Codons Observed	Over-/Under-representation[a] (%)	χ^2	No. of Codons Observed	Over-/Under-representation[a] (%)	χ^2
CGA	235	35.1	24.0	282	25.9	16.9
GGT	530	27.9	38.5	653	25.3	39.8
CCT	596	23.2	36.0	748	20.7	36.7
GGT	468	21.8	21.9	567	14.9	13.2
CTT	425	18.1	13.3	532	12.4	8.3[b]
GAA	862	16.5	33.3	1,020	7.8	9.6[b]
ACT	402	16.2	11.7	497	11.0	7.1[b]
AAT	635	10.8	11.8	761	3.7	1.8[b]
AAC	715	−8.0	11.8	969	−2.7	1.8[b]
CTG	1,282	−8.6	18.6	1,764	−4.3	6.3[b]
GAG	1,023	−10.6	33.3	1,390	−5.0	9.6[b]
GTC	521	−12.9	13.4	677	−11.9	14.6
GCG	178	−21.7	12.0	244	−17.2	9.7[b]
GGG	476	−24.0	48.1	616	−21.8	49.9
CCG	147	−27.2	16.9	193	−25.4	18.9

[a] Compared with ambiguous codon usage in human genes (Nakamura et al. 1996).
[b] Value is not significant.

of motifs or genes (see above), it is unclear whether there could be specific DNA sequence motifs that are frequently or even invariably associated with single-base-pair substitutions in human genes, causing inherited disease. With the establishment of HGMD, it became possible to answer this question by utilization of a large number of known nucleotide substitutions in a wide variety of different genes.

In terms of their relative frequency of occurrence, the most important category of single-base-pair substitution in HGMD is represented by C→T and G→A transitions within CpG dinucleotides; some 23% of all single-base-pair substitutions found within the coding regions of human genes are of this type. When, through consideration of relative mutabilities, allowance is made for the confounding effects of codon usage and differential clinical observation likelihoods, this proportion translates into a mean transition rate, for either CG→TG or CG→CA, that is five times higher than the base mutation rate. This represents a considerable downward adjustment of our earlier estimate of 7.4, which was derived from a ninefold-smaller sample (Cooper and Krawczak 1993). It reflects the disappearance of an initial reporting bias in the human molecular-genetics literature, a bias that most likely was due to the limited number of mutation-detection techniques available during the 1980s and early 1990s. Before 1993, ~31% of all published single-base-pair substitutions in human genes were CG→(TG,CA). This value dropped to some 24% in 1994–95, before reaching its current level of 21%.

Intriguingly, the proportion of CG→(TG,CA) transitions is significantly higher for autosomal genes (25.0%) than for X-linked genes (17.7%), a finding that directly reflects the significantly lower frequency of CpG in the coding sequences of X-linked genes (2.9%), compared with that in autosomal genes (3.7%). The lower CpG frequency in X-chromosomal genes may itself be a consequence of a generally increased level of DNA methylation, an increase that results from the recruitment of this postsynthetic modification to play a role in X inactivation (Hornstra and Yang 1994; Jamieson et al. 1996).

For CpG dinucleotides to be hypermutable in the context of genetic disease, they must be methylated in the germ line. Since it cannot be excluded that the efficiency of both DNA methyltransferase action (Bolden et al. 1985; Smith 1994; Smith and Baker 1997) and G:T mismatch repair (Sibghat-Ullah and Day 1993) are influenced by sequence motifs flanking the CpG dinucleotide, the question arises as to whether, by virtue of their DNA sequence context, some CpG sites may be intrinsically more mutable than others. Significant differences in the relative mutation rate of CpG dinucleotides, depending on flanking nucleotides, indeed were noted in the present study (table 3). These results are consistent with those of Ollila et al. (1996), who noted a preference for 5′ pyrimidines and 3′ purines flanking mutated CpG dinucleotides, albeit in a much smaller data set derived from publicly available locus-specific mutation databases. Our findings might, however, appear to conflict with those of Clay et al. (1995), who reported a tendency toward a 5′ C and a 3′ G in CpG-poor, noncoding regions

Table 6

Codons Significantly Over-/Underrepresented among ±4 Amino Acid Residues Surrounding Nonsense Mutations

	CG→(TG,CA) EXCLUDED (N = 998)			CG→(TG,CA) INCLUDED (N = 1,328)		
CODON	No. of Codons Observed	Over-/Under-representation[a] (%)	χ^2	No. of Codons Observed	Over-/Under-representation[a] (%)	χ^2
CGA	66	42.2	8.7[b]	94	52.2	17.6
CTA	78	36.6	8.1[b]	103	35.6	10.2
ACT	134	29.5	10.0	174	26.3	10.7
CTT	109	4.7	.3[b]	174	25.6	10.2
GTT	120	31.9	10.2	151	24.7	8.1[b]
GAA	306	19.0	12.5	412	20.4	19.1
GGA	127	−23.5	11.0	189	−14.4	5.5[b]
GGG	97	−26.5	10.6	138	−21.4	9.2[b]
GGT	77	−29.1	10.3	111	−23.2	8.7[b]

[a] Compared with 448 reference (human) cDNA sequences.
[b] Value is not significant.

of vertebrate DNA, indicative of a lower mutation rate of CCGG. Such a relationship is not immediately apparent from the HGMD data. However, even if a "protective effect" such as that suggested by Clay et al. (1995) were exclusively confined to a particular CpG-containing oligomer, nothing would be known regarding its overall efficiency in coding regions. Were this efficiency low, other nearest-neighbor effects might easily have obscured its influence on the observed mutational spectrum in human genes. An alternative possibility is that a subset of methylated CpG dinucleotides might be more prone to deamination, perhaps by virtue of their occurrence in a low-melting-temperature domain (the rate of 5mC deamination in single-stranded DNA is 28-fold higher than that in double-stranded DNA [Ehrlich et al. 1986]). To test this postulate, we assessed the frequency of A and T residues flanking CGA→TGA mutations. Although A/T richness was found to be significantly increased in the vicinity of such lesions, compared with that in control DNA sequences (50.4% vs. 47.6%), this elevation was not deemed to be strong enough to suggest any particular propensity of the mutated regions to become transiently single stranded.

5mC is known to occur at low frequency in non-CpG dinucleotides within triplets of the form CpNpG (Woodcock et al. 1988; Clark et al. 1995; Kay et al. 1997), and mutations at these sites have been reported in the *NF1* gene and the *BRCA1* gene (Rodenhiser et al. 1996, 1997). This could imply that methylation-mediated C→T and G→A transitions occur at CpNpG triplets in human genes. However, since the relative rates of CpNpG→TpNpG (N ≠ G) transition and CpNpG→CpNpA (N ≠ C) transition that have been derived from the present study are not substantially higher than the average substitution rate conditional on the next-but-one nucleotide, we may conclude that methylation-mediated deamination at such triplets has not contributed significantly to the mutational spectrum observed in HGMD.

Some studies have proposed a strand bias in either CpG deamination frequency or T:G mismatch repair, with C→T transitions outnumbering G→A transitions (Skandalis et al. 1994; Leader et al. 1995). In terms of their observed frequencies, such a relationship also holds for the collection of single-base-pair substitutions that have been analyzed in the present study (table 1). However, none of the previous studies allowed either for codon usage or for the magnitude of amino acid exchange, so that the skewed transition frequencies observed probably reflected observational bias rather than intrinsic differences in mutation rate. When relative substitution rates are considered, CG→CA is estimated to occur at a probability that is 1.4-fold higher than that for CG→TG (table 3).

Although the same substitution types appear to be subject to next neighbor-effects on the coding and noncoding DNA strands, the quantitative differences in non-CpG single-base-pair substitution rates observed here confirm that the two DNA strands are not fully equivalent in terms of their rates and patterns of mutation (Wu and Maeda 1987). There are several possible (and non–mutually exclusive) reasons for this strand asymmetry. First, the four nuclear DNA polymerases, each associated with its own distinctive mutational spectrum, may be differentially involved in the synthesis of the leading and lagging strands during DNA replication (Kunkel 1992; Bambara et al. 1997). Second, since the transcriptional elongation complex is asymmetrical (Kainz and Roberts 1992), mutation rates may differ between transcribed and nontranscribed strands, on account of either unequal exposure to DNA damage or differential repair. Not only may the transiently single-

stranded nontranscribed DNA strand be particularly vulnerable to mutation (e.g., by methylation-mediated deamination; Beletskii and Bhagwat 1996), but transcription-coupled repair (Hanawalt 1994; Drapkin et al. 1994; Bhatia et al. 1996), a process that corrects lesions specifically on the transcribed DNA strand, also could account for mutation-rate differences between transcribed and nontranscribed strands. Both these mechanisms would predict a higher mutation rate for the nontranscribed as opposed to the transcribed DNA strand. This hypothesis is, however, impossible to test on the basis of mutation data logged in HGMD.

Searches for sequence motifs potentially associated with mutational hot spots were largely unsuccessful. This includes polypyrimidine runs ($\geqslant$5 bp) and the "deletion hot spot consensus sequence" (TGRRKR), two motifs previously found to be strongly associated with the occurrence of microdeletions ($\leqslant$20 bp) in human genes (Cooper and Krawczak 1993). This notwithstanding, two codons (GAA and ACT) were found to be overrepresented in the vicinity of both missense and nonsense mutations other than CG→(TG,CA) whereas GGG was underrepresented. The biological meaning of these associations, however, which point toward a mutational effect rather than toward observational bias, remains unknown.

No preponderance of direct or inverted repeats was noted in the vicinity of mutations at non-CpG dinucleotides. There is thus no evidence that misalignment, mediated by repetitive sequences surrounding a substitution site, contributes substantially to the observed mutational spectrum in human genes. However, a subtle neighboring-nucleotide effect reminiscent of misalignment-mutagenesis models nevertheless was noted; a substantial proportion of observed single-base-pair substitutions exhibited identity between the newly introduced base and one of the bases immediately flanking the site of mutation. Since this effect occurred only at a distance of 1 bp and in the absence of surrounding repeat sequences, it is potentially explicable in terms of misalignment mutagenesis involving highly localized DNA slippage, misincorporation, and realignment events, at the replication fork, between template and primer (Kunkel 1990, 1992).

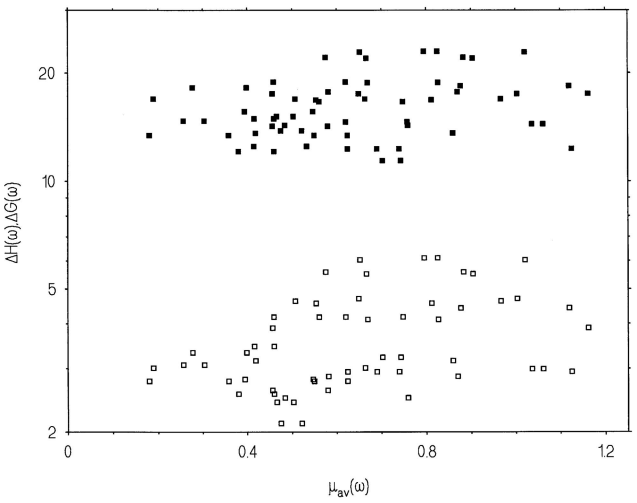

Figure 5 Thermodynamic stability and relative mutability of dsDNA triplets. $\mu_{av}(\omega)$ = bias-corrected average relative substitution rate at the central nucleotide of a triplet ω; $\Delta H(\omega)$ = change in transition enthalpy (*blackened squares*) associated with the disruption of ω; and $\Delta G(\omega)$ = change in transition-free energy at 37°C (*unblackened squares*).

Intriguingly, the mutational bias toward next neighbors occurred only at specific codon positions and exhibited directionality. Since such a phenomenon is unlikely to be explicable in terms of the primary mutational event, it is probably associated with the DNA-repair process. Implicit in this assumption, however, is that the DNA-repair machinery is able to recognize the reading

Table 7

Possible Role for Slippage-Mediated Misincorporation

	Substitutions[b]							
	5′ Flanking				3′ Flanking			
Position[a]	*N*	*F*	E(*F*)	χ^2	*N*	*F*	E(*F*)	χ^2
1	1,708	565	568.0	.01	1,645	490	390.0	33.61[c]
2	1,914	642	558.0	17.85[c]	2,025	592	659.3	10.19[c]
3	492	167	170.4	.10	485	184	164.2	3.61
Overall	4,114	1,374	1,296.4	6.78[d]	4,155	1,266	1,213.5	3.21

[a] Of mutation in codon.

[b] *N* = number of substitutions, other than CG→(TG,CA), for which the mutated nucleotide differs from the flanking nucleotide; *F* = number of substitutions toward flanking nucleotide; and E(*F*) = expectation of *F*, under a model of random mutation (for details, see text).

[c] Value is significant ($P < .05/6$, or 8.33×10^{-3}).

[d] Value is significant ($P < .05/2$, or .025).

frame and to utilize this information as a cue in effecting the repair of DNA. Consistent with such a relationship, the observed correction bias would operate in such a way as to remove newly introduced termination codons. The ability of the DNA-repair mechanism to take the reading frame into account, thereby minimizing the effects of mutation, would have had positive selective value because of the relatively deleterious nature of in-frame termination codons. Evidence for reading-frame sensitivity in the DNA-repair process has come from our previous observation that relative single-base-pair substitution rates are biased toward the avoidance of those replacements that (i) change the chemical characteristics of the encoded amino acid residue substantially and (ii) have a high likelihood of coming to clinical attention (Krawczak and Cooper 1996). We concluded from this finding that, by consideration of the genetic code, selection had optimized the DNA repair mechanism in such a way as to avoid the most hazardous of amino acid replacements, a category that certainly would include nonsense mutations.

The thermodynamic stability of DNA triplets was found to be positively correlated with the average relative rate at which the central nucleotide of a triplet undergoes substitution, a finding that implies that higher rather than lower DNA duplex stability renders a gene region more prone to single-base-pair substitution. Consistently, the absence of flanking repeat elements noted for the mutations analyzed in the present study suggests that extensive strand slippage (which would require the DNA to be single stranded) is unlikely to play an important role in the generation of single-base-pair substitutions.

A high degree of thermodynamic stability could, in principle, impair DNA replication, in various ways. First, the likelihood that DNA helicases would be incapable of unwinding the two DNA strands correctly or efficiently may be expected to be higher in regions that are more stable (Chen et al. 1992). Second, temporary reannealing of the two native DNA strands during replication might be favored and could be more enduring in such regions. In both cases, DNA polymerase activity would be seriously impeded by localized double-stranded DNA structures, which could result in either the cessation of polymerization or the skipping of one or more nucleotides, leaving a gap in the nascent DNA strand. Miscorrection during the postreplicative repair of such nicks would then introduce a single-base-pair substitution. Alternatively, the observed correlation could reflect the increased stability of at least some slippage-mediated misalignments during replication of the native and nascent DNA strands, allowing enough time for misincorporation of a noncomplementary nucleotide. In this case, however, the thermodynamic stabilities of the misaligned structures must be comparable to those of the wild-type triplets, an assumption for which there is currently no evidence. Finally, since mutagenesis depends not only on polymerase misincorporation but also on 3′ exonucleolytic proofreading, the correlation observed between triplet stability and mutagenicity can be interpreted also as an effect of surrounding–base-pair stability inhibiting proofreading (Petruska and Goodman 1985).

The clinical observation likelihood of an amino acid substitution is positively correlated with the chemical difference between the respective wild-type and mutant amino acid residue, the highest chance of coming to clinical attention being observed for nonsense mutations. A parameter closely related to chemical difference has been employed recently, by Rodin et al. (1998), to measure the selective effects of amino acid replacements in the somatic, germ-line, and evolutionary spectra of p53 mutations. On the basis of the implicit assumption that their parameter was positively correlated with clinical severity, Rodin et al. concluded that, owing to the structure of the genetic code, the likelihood of clinical observation should be much higher for CG→TG transitions than for CG→CA transitions. Indeed, the ratio of the two transition types turned out to be a good indicator of the selective pressure against loss-of-function mutations in critical regions of the p53 protein. Thus, the findings of Rodin et al. (1998) not only justify the use of chemical difference in the present context but also emphasize that clinically observed spectra of single-base-pair substitutions are likely to be biased by substitution-dependent clinical observation likelihoods.

In summary, our analysis of the largest available collection of germ-line single-base-pair substitutions in human genes has, for the first time, allowed us to disentangle the relative effects that selection and mutation have on the mutational spectrum observed in human inherited diseases. It has been shown that the relative rates at which such lesions occur at the DNA level is significantly influenced by the surrounding sequence context. However, with regard to the existence of sequence motifs that might serve to promote mutagenesis, only subtle and very localized effects were noted over and above the well-established hypermutability of CpG dinucleotides. It will be most interesting to determine whether these effects, manifesting here in the form of nearest-neighbor–dependent substitution rates and codon usage around mutational sites, also occur in noncoding DNA and in DNA from other species or can be reproduced in vitro. Similarly, our hypothesis of both a reading-frame–sensitive DNA-repair bias and a relationship between DNA duplex stability and mutability will require further independent verification by, for example, comparative evolutionary or in vitro studies. It is nevertheless hoped that our tentative conclusions may serve to guide the design of future experiments aiming to iden-

tify more specifically the mutational hot spots in individual genes, gene regions, or artificial DNA constructs.

Acknowledgments

We wish to thank Peter Stenson, Iain Fenton, and Shaun Abeysinghe, for their help in establishing and maintaining HGMD, and Peter Harper, for his continuing support and encouragement. The financial support of SmithKline Beecham and Pfizer is gratefully acknowledged. This work has been supported by the Deutsche Forschungsgemeinschaft through Heisenberg grant Kr 1093/5-1 (to M.K.).

Appendix A

Estimation of Relative Single-Base-Pair Substitution Rates and Relative Clinical Observation Likelihoods

Under the biologically meaningful assumption that the occurrence of a mutation at the DNA level and its expression at the protein level are statistically independent processes, the likelihood that an individual mutation ω of primary type $x[\omega]$ and phenotypic consequence $\alpha[\omega]$ will enter into a clinical sample equals $\mu(x) \cdot L(\alpha)$, where $\mu(x)$ is the likelihood of occurrence of x at the DNA level—for example, the single-base-pair substitution rate—and $L(\alpha)$ is the clinical observation likelihood of α. In successive iterations, relative $\mu(x)$ values are estimated via $O(x)/E(x)$, where $O(x)$ is the observed frequency of mutation type x, and $E(x)$ is the expected frequency of x when it is assumed that all mutation types are equally likely to occur at the DNA level; that is,

$$E(x) \propto \int_{\{\omega \in \Omega: x = x[\omega]\}} L(\alpha[\omega]) dP(\omega) \ . \qquad \text{(A1)}$$

In formula (A1), P can be any meaningful prior distribution on the mutation space Ω, allowing for the redundancy of the genetic code and for either codon usage or the composition of the particular DNA sequence(s) in question (see below). Similarly, relative clinical observation likelihoods are estimated in each iteration step by $O(\alpha)/E(\alpha)$, where $E(\alpha)$ is the expected frequency of α when it is assumed that all phenotypic consequences are equally likely to come to clinical attention; that is,

$$E(\alpha) \propto \int_{\{\omega \in \Omega: \alpha = \alpha[\omega]\}} \mu(x[\omega]) dP(\omega) \ .$$

This iterative algorithm represents two interwoven EM algorithms (Dempster et al. 1977) and thus converges to a (local) maximum of the overall sample likelihood. The fact that μ and L are only estimated on relative scales is irrelevant, since multiplication by any common positive constant does not change the relative location of minima and maxima on a likelihood surface. The global nature of convergence was confirmed, for each estimation process, through 1,000 independent replications using different, randomly chosen start values $0 < \mu$ and $L \leqslant 1$.

Prior Probabilities on Mutation Space

Without loss of generality, we may assume that an element of the mutation space Ω consists of a sequence of $n + m + 1$ codons plus one of nine possible replacements of the affected codon (denoted by subscript 0). Thus,

$$\Omega = \{\omega_{-n} \ldots \omega_{-1} \omega_0 \omega_1 \ldots \omega_m \wedge \omega_0 \rightarrow \omega_0'\} \ .$$

For the present analysis, the prior distribution on Ω is defined via

$$P(\omega) \propto 1/9 \cdot \Pi_{i=-n \ldots m-1} g(\omega_i \omega_{i+1}) / \Pi_{j=-n+1 \ldots m-1} f(\omega_j) \ , \qquad \text{(A2)}$$

if $\omega_0 \rightarrow \omega_0'$ results in a missense or nonsense mutation, and via $P(\omega) = 0$ if not. In formula (A2), function f denotes published codon usage in human genes (Nakamura et al. 1996), whereas g has been estimated on the basis of the 448 reference cDNA sequences provided by HGMD, which comprise a total of 309,444 codons.

It should be emphasized that, for all definitions of mutation type $x[\omega]$ and phenotypic consequence $\alpha[\omega]$ that have been used, the model given above was never overparameterized; that is both

$$\{\omega \in \Omega: x = x[\omega]\} \cap \{\omega \in \Omega: \alpha \neq \alpha[\omega]\}$$

and

$$\{\omega \in \Omega: x \neq x[\omega]\} \cap \{\omega \in \Omega: \alpha = \alpha[\omega]\}$$

always had strictly positive prior probabilities ($P > 0$) for every pair (x,α).

Expected Proportion of Substitutions toward Flanking Nucleotides

Let $\Omega(i)$ be the Ω subspace including single-base-pair substitutions other than CG→(TG,CA) that affect the ith codon position ($i = 1,\ldots,3$) and for which the wild-type nucleotide differs from the flanking base. The null hypothesis to be tested on $\Omega(i)$ is that substitution toward a flanking nucleotide occurs with probability $\frac{1}{3}$. The corresponding probability distribution, P_i, is the projection, onto $\Omega(i)$, of the distribution defined in formula (A2), with $n, m \leqslant 1$, and $m = 0$ for $i < 3$ and with $n = 0$ for $i > 1$. If we define $x[\omega] = 1$ for mutations toward the flanking nucleotide, and if $x[\omega] = 0$ otherwise, then the

expected number of mutations with $x = 1$ is proportional to

$$\int_{\{\omega \in \Omega(i):x[\omega]=1\}} L(\alpha[\omega])dP_i(\omega) \ .$$

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

HGMD, http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html (http://www.uwcm.ac.uk/uwcm/mg/msajhg1.txt [for estimates of relative single-base-pair substitution rates])

References

Bambara RA, Murante RS, Henricksen LA (1997) Enzymes and reactions at the eukaryotic DNA replication fork. J Biol Chem 272:4647–4650

Beletskii A, Bhagwat AS (1996) Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc Natl Acad Sci USA 93:13919–13924

Bhatia PK, Wang Z, Friedberg EC (1996) DNA repair and transcription. Curr Opin Genet Dev 6:146–150

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Bolden AH, Nalin CM, Ward CA, Poonian MS, McComas WW, Weissbach A (1985) DNA methylation: sequences flanking C-G pairs modulate the specificity of the human DNA methylase. Nucleic Acids Res 13:3479–3494

Breslauer KJ, Frank R, Blöcker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci USA 83:3746–3750

Chen YZ, Zhuang W, Prohofsky EW (1992) Energy flow considerations and thermal fluctuational opening of DNA base pairs at a replication fork: unwinding consistent with observed replication rates. J Biomol Struct Dyn 10:415–427

Clark SJ, Harrison J, Frommer M (1995) CpNpG methylation in mammalian cells. Nat Genet 10:20–27

Clay O, Schaffner W, Matsuo K (1995) Periodicity of eight nucleotides in purine distribution around human genomic CpG dinucleotides. Somat Cell Mol Genet 21:91–98

Cooper DN, Antonarakis SE, Krawczak M (1995) The nature and mechanisms of human gene mutation. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease, 7th ed. McGraw-Hill, New York, pp 259–291

Cooper DN, Krawczak M (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum Genet 85:55–74

——— (1993) Human gene mutation. Bios Scientific, Oxford

Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. Hum Genet 78:151–155

Dempster AP, Laird, NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B39:1-22

Drapkin R, Sancar A, Reinberg D (1994) Where transcription meets repair. Cell 77:9–12

Ehrlich M, Norris KF, Wang RYH, Kuo KC, Gehrke CW (1986) DNA cytosine methylation and heat-induced deamination. Biosci Rep 6:387–393

Golding GB, Glickman BW (1986) Evidence for local DNA influences on patterns of substitutions in the human alpha-interferon gene family. Can J Genet Cytol 28:483–496

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Hanawalt PC (1994) Transcription-coupled repair and human disease. Science 266:1957–1958

Hjorth JSU (1994) Computer intensive statistical methods. Chapman & Hall, London, pp 85–127

Hollstein M, Sidransky D, Vogelstein B, Harris CC (1991) p53 mutations in human cancer. Science 253:49–53

Hornstra IK, Yang TP (1994) High-resolution methylation analysis of the human hypoxanthine phosphoribosyltransferase gene 5′ region on the active and inactive X chromosomes: correlation with binding sites for transcription factors. Mol Cell Biol 14:1419–1430

Jamieson RV, Tam PPL, Gardiner-Garden M (1996) X-chromosome activity: impact of imprinting and chromatin structure. Int J Dev Biol 40:1065–1080

Kainz M, Roberts J (1992) Structure of transcription elongation complexes *in vivo*. Science 255:838–841

Kay PH, Harmon D, Fletcher S, Ziman M, Jacobsen PF, Papadimitriou JM (1997) Variation in the methylation profile and structure of Pax3 and Pax7 among different mouse strains and during expression. Gene 184:45–53

Krawczak M, Cooper DN (1996) Single base-pair substitutions in pathology and evolution: two sides to the same coin. Hum Mutat 8:23–31

——— (1997) The Human Gene Mutation Database. Trends Genet 13:121–122

Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E (1995) Somatic spectrum of cancer-associated single base-pair substitutions in the TP53 gene is determined mainly by endogenous mechanisms of mutation and by selection. Hum Mutat 5:48–57

Krawczak M, Wacey A, Cooper DN (1996) Molecular reconstruction and homology modelling of the catalytic domain of the common ancestor of the haemostatic vitamin K-dependent serine proteinases. Hum Genet 98:351–370

Kunkel TA (1985) The mutational specificity of DNA polymerase-α during *in vitro* DNA synthesis. J Biol Chem 260: 5787–5796

——— (1990) Misalignment-mediated DNA synthesis errors. Biochemistry 29:8003–8011

——— (1992) DNA replication fidelity. J Biol Chem 267: 18251–18254

Leader DP, Peter B, Ehmer B (1995) Analysis of CpG dinucleotide frequency in relationship to translational reading frame suggests a class of genes in which mutation of this dinucleotide is asymmetric with respect to DNA strand. FEBS Lett 376:125–129

Lin MM-Q, Zhu M, Scharff MD (1997) Sequence dependent hypermutation of the immunoglobulin heavy chain in cultured B cells. Proc Natl Acad Sci USA 94:5284–5289

Nakamura Y, Wada K, Wada Y, Doi H, Kanaya S, Gojobori T, Ikemura T (1996) Codon usage tabulated from the in-

ternational DNA sequence databases. Nucleic Acids Res 24: 214–215

Ollila J, Lappalainen I, Vihinen M (1996) Sequence specificity in CpG mutation hotspots. FEBS Lett 396:119–122

Petruska J, Goodman MF (1985) Influence of neighboring bases on DNA polymerase insertion and proofreading fidelity. J Biol Chem 260:7533–7539

Rodenhiser DI, Andrews JD, Mancini DN, Jung JH, Singh SM (1997) Homonucleotide tracts, short repeats and CpG/CpNpG motifs are frequent sites for heterogeneous mutations in the neurofibromatosis type 1 (NF1) tumor suppressor gene. Mutat Res 373:185–195

Rodenhiser DI, Chakraborty P, Andrews JD, Ainsworth P, Mancini D, Lopes E, Singh SM (1996) Heterogeneous point mutations in the BRCA1 breast cancer susceptibility gene occur in high frequency at the site of homonucleotide tracts, short repeats and methylatable CpG/CpNpG motifs. Oncogene 12:2623–2629

Rodin SN, Holmquist GP, Rodin AS (1998) CpG transition strand asymmetry and hitch-hiking mutations as measures of tumorigenic selection in shaping the p53 mutation spectrum. Int J Mol Med 1:191–199

Rogozin IB, Kolchanov NA (1992) Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim Biophys Acta 1171: 11–18

Shen JC, Rideout WM, Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res 22:972–976

Sibghat-Ullah, Day RS (1993) DNA-substrate sequence specificity of human G:T mismatch repair activity. Nucleic Acids Res 21:1281–1287

Skandalis A, Ford BN, Glickman BW (1994) Strand bias in mutation involving 5-methylcytosine deamination in the human *hprt* gene. Mutat Res 314:21–26

Smith SS (1994) Biological implications of the mechanism of action of human DNA (cytosine-5) methyltransferase. Prog Nucleic Acid Res Mol Biol 49:65–111

Smith SS, Baker DJ (1997) Stalling of human methyltransferase at single-strand conformers from the Huntington's locus. Biochem Biophys Res Commun 234:73-78

Todorova A, Danieli GA (1997) Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis. Hum Mutat 9:537–547

Tornaletti S, Pfeifer GP (1995) Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancer. Oncogene 10:1493–1499

Woodcock DM, Crowther PJ, Jefferson S, Diver WP (1988) Methylation at dinucleotides other than CpG: implications for human maintenance methylation. Gene 74:151–152

Wu C-I, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169–170