Hum Genet (1990) 85:55–74

Human Genetics



# The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions

**David N. Cooper**[1] **and Michael Krawczak**[2]

[1]Molecular Genetics Section, Thrombosis Research Institute, Manresa Road, Chelsea, London SW3 6LR, UK
[2]Institut für Humangenetik der Universität, Gosslerstrasse 12d, D-3400 Göttingen, Federal Republic of Germany

Received September 22, 1989 / Revised December 5, 1989

**Summary.** Reports of single base-pair substitutions that cause human genetic disease and that have been located and characterized in an unbiased fashion were collated; 32% of point mutations were CG → TG or CG → CA transitions consistent with a chemical model of mutation via methylation-mediated deamination. This represents a 12-fold higher frequency than that predicted from random expectation, confirming that CG dinucleotides are indeed hotspots of mutation causing human genetic disease. However, since CG also appears hypermutable irrespective of methylation-mediated deamination, a second mechanism may also be involved in generating CG mutations. The spectrum of point mutations occurring outwith CG dinucleotides is also non-random, at both the mono- and dinucleotide levels. An intrinsic bias in clinical detection was excluded since frequencies of specific amino acid substitutions did not correlate with the 'chemical difference' between the amino acids exchanged. Instead, a strong correlation was observed with the mutational spectrum predicted from the experimentally measured mispairing frequencies of vertebrate DNA polymerases α and β in vitro. This correlation appears to be independent of any difference in the efficiency of enzymatic proofreading/mismatch-repair mechanisms but is consistent with a physical model of mutation through nucleotide misincorporation as a result of transient misalignment of bases at the replication fork. This model is further supported by an observed correlation between dinucleotide mutability and stability, possibly because transient misalignment must be stabilized long enough for misincorporation to occur. Since point mutations in human genes causing genetic disease neither arise by random error nor are independent of their local sequence environment, predictive models may be considered. We present a computer model (MUTPRED) based upon empirical data; it is designed to predict the location of point mutations within gene coding regions causing human genetic disease. The mutational spectrum predicted for the human factor IX gene was shown to resemble closely the observed spectrum of point mutations causing haemophilia B. Further, the model was able to predict successfully the rank order of disease prevalence and/or mutation rates associated with various human autosomal dominant and sex-linked recessive conditions. Although still imperfect, this model nevertheless represents an initial attempt to relate the variable prevalence of human genetic disease to the mutability inherent in the nucleotide sequences of the underlying genes.

*Offprint requests to:* D. N. Cooper

*What man that sees the ever-whirling wheele*
*Of change, the which all mortall things doth sway,*
*But that thereby doth find, and plainly feele*
*How mutability in them doth play*
*Her cruell cruell sports, to many men's decay?*

Edmund Spenser
(c. 1609; The Faerie Queene. Book VII,
'Two Cantos of Mutabilitie', Canto VI (i))

## Introduction

One of the most important properties of genetic material is its capacity to organize its own faithful replication within the cell. If, however, no errors were ever made during the process of DNA replication, the evolution of complex genomes would scarcely have been possible. Although in vivo studies of eukaryotic DNA replication have indicated that this replication process is extremely accurate, errors do nevertheless occur at a frequency of

between $10^{-9}$ to $10^{-11}$ misincorporated nucleotides per base (Drake 1970; Nalbantoglu et al. 1983; Thacker 1985). Whereas this is much lower than the $10^{-3}$–$10^{-6}$ error frequency observed in vitro, the disparity is thought to be a result of the efficiency of in vivo proofreading and error correction mechanisms (Roberts and Kunkel 1986; Kunkel and Bebinek 1988). Mutation rates calculated for specific human genetic diseases also vary widely (Vogel and Motulsky 1986). Intuitively, we may surmise that at least for dominant and sex-linked recessive conditions, differences in incidence/prevalence are a reflection of underlying differences in the relative mutability of certain genes in the germ-line.

The challenge now facing us is to be able to relate observed differences in the incidence/prevalence of genetic disease to the structure of the mutant genes responsible and to the function of the proteins that they encode. A major step towards this aim would be the development of the ability to make predictions regarding the probable frequency and location of lesions within any given gene by reference to the DNA sequence and genomic architecture of that gene.

With the advent of recombinant DNA and the introduction of new analytical techniques, rapid progress has been made in the characterization of gene mutations causing human inherited disease (reviewed by Cooper and Schmidtke 1989). For most conditions, a substantial proportion of these lesions is now thought to consist of point mutations together with a smaller number of deletions, the remainder being made up of insertions, inversions and complex rearrangements. Considerable variation nevertheless exists in the relative proportions of the different types of mutation found for different gene loci. This variation may reflect individual characteristics of the genes, such as length and the degree of sequence repetitivity.

Mutation of human genes may arise either as a consequence of endogenous error-prone processes, such as DNA replication and repair, or as a result of exposure to exogenous factors, e.g. chemical mutagens or ionizing/UV irradiation. Most studies to date have concentrated upon the latter category of induced mutations to the detriment of the former. This is for the simple reason that induced mutations are usually many orders of magnitude more frequent than their spontaneous in vivo counterparts. However, induced mutations probably arise by different mechanisms from 'spontaneous' mutations and may not provide a viable or realistic model for the study of in vivo mutation. Moreover, most mutation studies have been carried out in bacteria or viruses on account of their relative ease of culture and analysis. In consequence, relatively little is known of the nature of spontaneous mutation in eukaryotes and virtually nothing is known in man.

Such nucleotide substitutions are thought to arise either by chemical (e.g. deamination of 5-methylcytosine, Coulondre et al. 1978; depurination, Loeb and Preston 1986), physical (e.g. DNA slippage, Kunkel and Soni 1988) or enzymatic (e.g. postreplicative mismatch repair or exonucleolytic proofreading, Modrich 1987; Loeb and Kunkel 1982) mechanisms. Since the efficiency of all these processes is known to be sequence-dependent, it is hardly surprising that the distribution of point mutations is non-random thereby giving rise to both 'hotspots' and 'coldspots' of base substitution in eukaryotic genomes.

The study of 'naturally occurring' gene mutations is very important for a number of reasons, not the least being that the process of mutational change is fundamental to an understanding of the origins of genetic variation and, as such, plays a vital role in the evolution of living systems (reviewed by Crow and Denniston 1985). Knowledge of the relative frequency of specific mutations should also aid our understanding of the generative mechanisms that underlie gene mutation, and may provide valuable insight into the intricacies of the processes of DNA replication and repair. Finally, the drawing up of the ground-rules for assessing and predicting the probable relative frequencies and locations of specific point mutation may contribute to dramatic improvements in both the design and efficacy of mutation search strategies in molecular diagnostic medicine. For all these reasons, we should follow the advice of the early English geneticist William Bateson who, extrapolating from his own work in plant genetics, exhorted us to 'treasure our exceptions'.

Point mutations in vertebrate genes have been known for some time to be non-random. For example, the frequency of transitional changes (between T and C or between A and G), whether causing gene mutation or occurring during vertebrate gene evolution, is much higher than expected (Vogel 1972; Vogel et al. 1976; Li et al. 1984). Much of this excess of transition-type mutations has been thought to be the result of the hypermutability of the methylated dinucleotide CpG; deamination of 5-methylcytosine (5mC) to thymidine in this doublet gives rise to C → T or G → A substitutions depending upon in which strand the 5mC is mutated (Coulondre et al. 1978; Salser 1978; Duncan and Miller 1980). Deamination of 5mC to thymidine probably occurs with the same frequency as deamination of cytidine to uracil. However, whereas uracil DNA glycosylase activity (Lindahl 1982) in eukaryotic cells is able to excise uracil, thymidine being a 'normal' nucleoside is thought to be less readily detectable and removable by cellular repair mechanisms.

CpG is a 'hotspot' for mutation in vertebrate genomes as shown by (1) its under-representation or 'suppression' in genomic DNA (observed frequency in coding regions = 37% of expected from mononucleotide frequencies [Nussinov 1981; Beutler et al. 1989]), (2) the high frequency of polymorphism detected by restriction enzymes containing CpG in their recognition sequences (Barker et al. 1984; Cooper and Schmidtke 1984), (3) the high rate of change of CpG dinucleotides between DNA sequences examined in evolutionary comparison studies (Savatier et al. 1985; Cooper et al. 1987) and (4) the high frequency of CG → TG and CG → CA transitions among mutations causing human genetic disease (Cooper and Youssoufian 1988; Youssoufian et al. 1988; Cooper and Krawczak 1989). In this last category, 35% of all single base-pair substitutions causing human genet-

ic disease were found to occur within CpG dinucleotides (Cooper and Youssoufian 1988) and over 90% of these were C → T or G →A transitions consistent with the model of methylation-mediated deamination. This represents a 40-fold higher rate of mutation of 5-methylcytosine than that found for any other nucleotide. In this article, we present an updated assessment of the contribution of CpG mutation to the incidence of human genetic disease based upon a considerably larger sample of mutations.

For non-CpG mutations, marked differences in the frequency of occurrence of different single base-pair substitutions also occur and in theory may have arisen through the influence of one, or more, of a number of different factors. Thus, the non-randomness of the initial mutation event, the non-randomness of the DNA sequences under study, differences in the relative efficiency with which certain mutations are repaired, differences in phenotypic effect (and hence selection), or a bias in the clinical detection of such variants, may all in principle play a role. Whereas all of these factors could conceivably contribute to the observed non-randomness of point mutation, their individual contributions have been neither determined nor even properly estimated.

We have thus attempted here to extend our analysis of human gene mutation beyond the CpG dinucleotide to the remaining approximately 65% of point mutations that occur within other doublets. We demonstrate that these mutations are also distributed non-randomly with respect to DNA sequence, and we explore several possible mechanisms responsible for this finding. Finally, a mathematical model is presented that is empirically based upon our current knowledge of the spectrum of point mutations in human genes causing genetic disease. This model attempts (1) tentatively to predict both the location and the relative frequency of point mutations within a given gene sequence and (2) to relate a derived 'mutability quotient' for that gene to the clinically observed prevalence of the deficiency of the gene product.

## Results

### *Single base-pair substitutions causing human genetic disease*

Table 1 lists point mutations that cause human genetic disease and that have been located and characterized in as unbiased a fashion as possible. These data were collected for analysis since it was considered probable that this would provide valuable information on both the nature and frequency of spontaneous point mutations in man.

The database includes reports of disease-causing single base-pair substitutions (up until August 1989) that have been detected by DNA sequencing or oligonucleotide discrimination hybridization. A preliminary version of this list has been published previously (Cooper and Youssoufian 1988). Each entry consists of the name of the disease, the McKusick symbol for the disease gene, the base-pair change responsible for the disease phenotype, the resulting amino acid substitution at a specified codon and the literature reference under the name of the first author. Nucleotides outside the altered codon (denoted by a lower case letter) are given when the mutation occurs in the first or third position in the codon. This information is provided to allow the assessment of the potential influence of neighbouring bases on the relative likelihood of a specific mutational change. Mutations that occur within CpG dinucleotides and that are consistent with the process of methylation-mediated deamination (i.e. either C → T or G → A) are marked with an *. Several omissions have been made in the interest of obtaining an unbiased sample of point mutations causing human genetic disease.

1. In the case of single base-pair substitutions that have been reported independently more than once, the first example only has been logged. This is because of the difficulty in determining whether these mutations are identical-by-descent or whether they are truly recurrent. Adoption of this policy may result in an underestimate of the actual proportion of independent point mutations that occur in the hypermutable CpG dinucleotide.

2. Examples of point mutations in the factor VIII and factor IX genes causing haemophilia are not included because of a considerable bias in ascertainment. In these disorders, recurrent mutation is known to occur at high frequency in CpG dinucleotides, and mutation search procedures have therefore often been designed accordingly. The inclusion of these data would thus greatly increase the apparent proportion of mutations occurring within CpG dinucleotides; over 120 independent point mutations have been reported in these genes (the interested reader is referred to Cooper and Tuddenham 1991).

3. Examples of point mutations in the globin genes resulting in haemoglobin variants are not included because of (a) their disproportionately large number and (b) complexities of interpretation arising from the existence of heterozygote advantage.

4. Point mutations in the mitochondrial genome resulting in diseases such as the mitochondrial myopathies or Leber's optic atrophy are not included; mitochondrial DNA is not methylated (Groot and Kroon 1979; Castora et al. 1980) and thus mutation frequencies are unlikely to be comparable to those exhibited by the nuclear genome.

5. Point mutations in oncogenes are not included since tumorigenesis occurs in somatic cells and may be caused by the action of environmental mutagens. In addition, it is usually difficult to determine whether or not the mutation detected is itself a cause or a consequence of the transformed phenotype. This is not to say that mechanisms of mutagenesis are entirely different, e.g. CpG deamination-type mutations in the *gsp* oncogene have been shown to be a cause of pituitary tumours (Landis et al. 1989).

**Table 1.** A list of point mutations causing human genetic disease. CG→TG and CG→CA transitions are marked with an *

| Disease | Mutation | | | | Reference[a] | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gene | Base change | Amino acid change | CD | Author | Journal | Volume | Page | Year |
| ADA deficiency | ADA | CGG--CAG | Arg-Gln | 101 | Bonthron | JCI | 76 | 894 | 85* |
| ADA deficiency | ADA | AAA--AGA | Lys-Arg | 80 | Valerio | Embo J | 5 | 113 | 86 |
| ADA deficiency | ADA | CTG--CGG | Leu-Arg | 304 | Valerio | Embo J | 5 | 113 | 86 |
| ADA deficiency | ADA | gCGG--TGG | Arg-Trp | 101 | Akeson | JBC | 263 | 16291 | 88* |
| ADA deficiency | ADA | CGT--CAT | Arg-His | 211 | Akeson | JBC | 263 | 16291 | 88* |
| ADA deficiency | ADA | GCG--GTG | Ala-Val | 329 | Akeson | PNAS | 84 | 5947 | 87* |
| ADA deficiency | ADA | CCG--CAG | Pro-Gln | 297 | Hirshhorn | JCI | 83 | 497 | 89 |
| Adenylate kinase deficiency | AK | cCGG--TGG | Arg-Trp | 128 | Matsuura | JBC | 264 | 10148 | 89* |
| Adrenal hyperplasia | CA21HB | CCC--CGC | Pro-Arg | 426 | Matteson | PNAS | 84 | 5858 | 87 |
| Adrenal hyperplasia | CA21HB | AGC--ACC | Ser-Thr | 269 | Rodrigues | Embo J | 6 | 1653 | 87 |
| Adrenal hyperplasia | CA21HB | ATC--AAC | Ile-Asn | 172 | Amor | PNAS | 85 | 1600 | 88 |
| Adrenal hyperplasia | CA21HB | gCAG--TAG | Gln-Term | 318 | Globerman | JCI | 82 | 139 | 88 |
| Adrenal hyperplasia | CA21HB | cGTG--CTG | Val-Leu | 211 | Speiser | NEJM | 319 | 19 | 88 |
| Adrenal hyperplasia | CA21HB | cGTG--TTG | Val-Leu | 281 | Speiser | NEJM | 319 | 19 | 88 |
| Adrenal hyperplasia | CA21HB | AAC--AGC | Asn-Ser | 494 | Rodrigues | Embo J | 6 | 1653 | 87 |
| Aldolase A deficiency | ALDA | GAT--GGT | Asp-Gly | 128 | Kishi | PNAS | 84 | 8623 | 88 |
| Aldolase B deficiency | ALDB | tGCT--CCT | Ala-Pro | 149 | Cross | Cell | 53 | 881 | 88 |
| Amyloidotic polyneuropathy | PALB | cGTG--ATG | Val-Met | 30 | Furuya | JCI | 80 | 1706 | 87* |
| Amyloidotic polyneuropathy | PALB | TCT--TAT | Ser-Tyr | 77 | Wallace | JCI | 81 | 189 | 88 |
| Amyloidotic polyneuropathy | PALB | ATC--AGC | Ile-Ser | 84 | Wallace | AJHG | 43 | 182 | 88 |
| Angioneurotic edema | C1I | gCGC--TGC | Arg-Cys | 444 | Skriver | JBC | 264 | 3066 | 89* |
| Angioneurotic edema | C1I | CGC--CAC | Arg-His | 444 | Skriver | JBC | 264 | 3066 | 89* |
| Antithrombin III deficiency | AT3 | CGT--TGT | Arg-Cys | 47 | Duchange | NAR | 14 | 2408 | 86* |
| Antithrombin III deficiency | AT3 | CCG--CTG | Pro-Leu | 41 | Chang | JBC | 261 | 1174 | 86* |
| Antithrombin III deficiency | AT3 | CCT--CTT | Pro-Leu | 407 | Bock | Biochem | 27 | 6171 | 88 |
| Antithrombin III deficiency | AT3 | aGCA--ACA | Ala-Thr | 382 | Devraj-K. | Blood | 72 | 1518 | 88 |
| Antithrombin III deficiency | AT3 | cCGT--TGT | Arg-Cys | 393 | Thein | Blood | 72 | 1817 | 88* |
| Antithrombin III deficiency | AT3 | CGT--CAT | Arg-His | 393 | Thein | Blood | 72 | 1817 | 88* |
| Antithrombin III deficiency | AT3 | CGT--CCT | Arg-Pro | 393 | Lane | JBC | 264 | 10200 | 89 |
| α1-Antitrypsin deficiency | PI | GTG--GCG | Val-Ala | 213 | Nukiwa | JBC | 261 | 15989 | 86 |
| α1-Antitrypsin deficiency | PI | cGAG--AAG | Glu-Lys | 342 | Kidd | Nature | 304 | 230 | 83* |
| α1-Antitrypsin deficiency | PI | cCCT--TCT | Pro-Leu | 369 | Hofker | Hum Genet | 81 | 264 | 89 |
| α1-Antitrypsin deficiency | PI | gAAG--TAG | Lys-Term | 217 | Satoh | AJHG | 42 | 77 | 88 |
| α1-Antitrypsin deficiency | PI | CTG--CCG | Leu-Pro | 41 | Takahashi | JBC | 263 | 15528 | 88 |
| APRT deficiency | APRT | ATG--ACG | Met-Thr | 136 | Hidaka | JCI | 81 | 945 | 88 |
| ApoA1 deficiency | APOA1 | GAGc-GAT | Glu-Asp | 120 | Law | JBC | 260 | 12810 | 85 |
| ApoB deficiency | APOB | aCAA--TAA | Gln-Term | 2153 | Hospatta. | BBRC | 148 | 279 | 87 |
| ApoB deficiency | APOB | tCGA--TGA | Arg-Term | 1306 | Collins | NAR | 16 | 8361 | 88* |
| ApoB deficiency | APOB | CGG--CAG | Arg-Gln | 3500 | Soria | PNAS | 86 | 587 | 89* |
| ApoB deficiency | APOB | tCGA--TGA | Arg-Term | 2058 | Young | NEJM | 320 | 1604 | 89* |
| ApoE deficiency | APOE | gCGC--TGC | Arg-Cys | 158 | Funke | Clin Chem | 32 | 1285 | 86* |
| ApoE deficiency | APOE | gCGC--AGC | Arg-Ser | 136 | Emi | Genomics | 3 | 373 | 88 |
| ApoE deficiency | APOE | gCGT--TGT | Arg-Cys | 145 | Emi | Genomics | 3 | 373 | 88* |
| ApoE deficiency | APOE | gGAG--AAG | Glu-Lys | 244 | Tajima | J Biochem | 105 | 249 | 89 |
| ApoE deficiency | APOE | gCGC--TGC | Arg-Cys | 142 | Rall | JCI | 83 | 1095 | 89* |
| Diabetes (insulin resistant) | INSR | AGGt-AGT | Arg-Ser | 735 | Yoshimasa | Science | 240 | 784 | 88 |
| Diabetes (insulin resistant) | INSR | TGG--TCG | Trp-Ser | 1200 | Moller | NEJM | 319 | 1526 | 88 |
| Diabetes (MODY) | INS | TTC--TCC | Phe-Ser | 24 | Haneda | PNAS | 80 | 6366 | 83 |
| Diabetes (NIDDM) | INSR | GGC--GTC | Gly-Val | 996 | Odawara | Science | 245 | 66 | 89 |
| Ehlers-Danlos IV | COL1A3 | gGGT--AGT | Gly-Ser | 790 | Tromp | JBC | 264 | 1349 | 89 |
| Ehlers-Danlos VII | COL1A1 | ATGg-ATA | Met-Ile | 159 | Weil | Embo J | 8 | 1705 | 89 |
| Elliptocytosis | SPTA1 | AGTg-AGG | Ser-Arg | 39 | Garbarz | Blood | 72S | 41A | 88 |

**Table 1** (continued)

| Disease | Mutation | | | | Reference[a] | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gene | Base change | Amino acid change | CD | Author | Journal | Volume | Page | Year |
| Fabry disease | GLA | cCGG–TGG | Arg–Trp | 356 | Berstein | JCI | 83 | 1390 | 89* |
| Factor X deficiency | FX | cCGC––TGC | Arg–Cys | 366 | Jagadees. | JCBS | 13E | 291 | 89* |
| Factor XI deficiency | FXI | aGAA––TAA | Glu–Term | 117 | Asakai | PNAS | 86 | 7667 | 89 |
| Gangliosidosis GM2 | HEXB | CGC––CAC | Arg–His | 178 | Ohno | AJHG | 41 | A231 | 87* |
| Gangliosidosis GM2 | HEXB | aGGT––AGT | Gly–Ser | 269 | Paw | PNAS | 86 | 2413 | 89 |
| G6PD deficiency | G6PD | gAAT––GAT | Asn–Asp | 126 | Vulliamy | PNAS | 85 | 5171 | 88 |
| G6PD deficiency | G6PD | gGAT––AAT | Asp–Asn | 58 | Vulliamy | PNAS | 85 | 5171 | 88 |
| G6PD deficiency | G6PD | cGAG––AAG | Glu–Lys | 156 | Vulliamy | PNAS | 85 | 5171 | 88* |
| G6PD deficiency | G6PD | cGCC––ACC | Ala–Thr | 335 | Vulliamy | PNAS | 85 | 5171 | 88* |
| G6PD deficiency | G6PD | TCC––TTC | Ser–Phe | 188 | Vulliamy | PNAS | 85 | 5171 | 88 |
| G6PD deficiency | G6PD | cGGG––AGG | Gly–Arg | 447 | Vulliamy | PNAS | 85 | 5171 | 88* |
| G6PD deficiency | G6PD | cGTG––ATG | Val–Met | 68 | Vulliamy | PNAS | 85 | 5171 | 88* |
| G6PD deficiency | G6PD | aGAT––CAT | Asp–His | 282 | De Vita | AJHG | 44 | 233 | 89 |
| G6PD deficiency | G6PD | aGGC––AGC | Gly–Ser | 163 | Vulliamy | NAR | 17 | 5868 | 89 |
| Gaucher's disease (type 1) | GBA | CGG––CAG | Arg–Gln | 119 | Graves | DNA | 7 | 521 | 88* |
| Gaucher's disease (type 1) | GBA | AAC––AGC | Asn–Ser | 370 | Tsuji | PNAS | 85 | 2349 | 88 |
| Gaucher's disease (type 2) | GBA | CTG––CCG | Leu–Pro | 444 | Tsuji | NEJM | 316 | 570 | 87 |
| Gaucher's disease (type 2) | GBA | CCC––CGC | Pro–Arg | 415 | Widgerson | AJHG | 44 | 365 | 89 |
| Gerstmann-Straeussler syndrome | PRP | CCG––CTG | Pro–Leu | 102 | Hsiao | Nature | 338 | 342 | 89* |
| Gyrate atrophy | OAT | ATGt–ATA | Met–Ile | 1 | Mitchell | JCI | 81 | 630 | 88 |
| Gyrate atrophy | OAT | AACt–AAA | Asn–Lys | 54 | Ramesh | PNAS | 85 | 3777 | 88 |
| Gyrate atrophy | OAT | aGTG––ATG | Val–Met | 332 | Ramesh | PNAS | 85 | 3777 | 88 |
| Gyrate atrophy | OAT | AGG––ACG | Arg–Thr | 180 | Mitchell | PNAS | 86 | 197 | 89 |
| Gyrate atrophy | OAT | CTT––CCT | Leu–Pro | 402 | Mitchell | PNAS | 86 | 197 | 89 |
| Heparin cofactor 2 deficiency | HCF2 | CGC––CAC | Arg–His | 189 | Blinder | JBC | 264 | 5128 | 89* |
| HPRT deficiency | HPRT | tCGA––GGA | Arg–Gly | 50 | Wilson | JCI | 72 | 767 | 83 |
| HPRT deficiency | HPRT | ATTg–ATG | Ile–Met | 131 | Fujimori | Hum Genet | 79 | 39 | 88 |
| HPRT deficiency | HPRT | GTC––GAC | Val–Asp | 129 | Davidson | Gene | 68 | 85 | 88 |
| HPRT deficiency | HPRT | AGCt–AGA | Ser–Arg | 103 | Cariello | AJHG | 42 | 726 | 88 |
| HPRT deficiency | HPRT | TCA––TTA | Ser–Leu | 109 | Davidson | JCI | 82 | 2164 | 88 |
| HPRT deficiency | HPRT | GAT––GGT | Asp–Gly | 200 | Davidson | JBC | 264 | 520 | 89 |
| HPRT deficiency | HPRT | TTCt–TTA | Phe–Leu | 73 | Gibbs | PNAS | 86 | 1919 | 89 |
| HPRT deficiency | HPRT | cGCA––TCA | Ala–Ser | 160 | Gibbs | PNAS | 86 | 1919 | 89 |
| HPRT deficiency | HPRT | TTG–TCG | Leu–Ser | 130 | Gibbs | PNAS | 86 | 1919 | 89 |
| HPRT deficiency | HPRT | aCGA––TGA | Arg–Term | 169 | Gibbs | PNAS | 86 | 1919 | 89* |
| HPRT deficiency | HPRT | cTTC––GTC | Phe–Val | 198 | Gibbs | PNAS | 86 | 1919 | 89 |
| HPRT deficiency | HPRT | TGT––TAT | Cys–Tyr | 205 | Gibbs | PNAS | 86 | 1919 | 89 |
| HPRT deficiency | HPRT | tCAT––GAT | His–Asp | 203 | Gibbs | PNAS | 86 | 1919 | 88 |
| HPRT deficiency | HPRT | gGGC––CGC | Gly–Arg | 70 | Fujimori | JCI | 83 | 11 | 89 |
| HPRT deficiency | HPRT | CTA––CCA | Leu–Pro | 40 | Davidson | JCI | 84 | 342 | 89 |
| HPRT deficiency | HPRT | GGG––GAG | Gly–Glu | 69 | Davidson | JCI | 84 | 342 | 89 |
| HPRT deficiency | HPRT | GAT––GTT | Asp–Val | 79 | Davidson | JCI | 84 | 342 | 89 |
| HPRT deficiency | HPRT | cGCA––TCA | Ala–Ser | 160 | Davidson | JCI | 84 | 342 | 89 |
| HPRT deficiency | HPRT | cTTC––GTC | Phe–Val | 198 | Davidson | JCI | 84 | 342 | 89 |
| Hypophosphatasia | ALPL | cGCC––ACC | Ala–Thr | 162 | Weiss | PNAS | 85 | 7666 | 88* |
| Hypercholesterolaemia | LDRL | CCG–CTG | Pro–Leu | 664 | Soutar | PNAS | 86 | 4166 | 89* |
| Hyperproinsulinaemia | INS | aCAC––GAC | His–Asp | B10 | Chan | PNAS | 84 | 2194 | 87 |
| Hyperproinsulinaemia | INS | TTCt–TTG | Phe–Leu | 24 | Kwok | BBRC | 98 | 844 | 81 |
| Hyperproinsulinaemia | INS | TTCt–TTG | Phe–Leu | B25 | Kwok | Diabetes | 32 | 872 | 83 |
| Hyperproinsulinaemia | INS | CGT––CAT | Arg–His | 65 | Shibasaki | JCI | 76 | 378 | 85* |
| Hyperproinsulinaemia | INS | tGTG––TTG | Val–Leu | A3 | Awata | Diabetes | 37 | 1068 | 88 |

**Table 1** (continued)

| Disease | Mutation | | | | Reference[a] | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gene | Base change | Amino acid change | CD | Author | Journal | Volume | Page | Year |
| Immunoglobulin K deficiency | IGK | cTGC--GGC | Cys-Gly | 194 | Stavnezer | Science | 230 | 458 | 85 |
| Immunoglobulin K deficiency | IGK | gTGG--CGG | Trp-Arg | 148 | Stavnezer | Science | 230 | 450 | 85 |
| LDLR deficiency | LDLR | TGGc-TGA | Trp-Term | 792 | Lehrman | Cell | 41 | 735 | 85 |
| LDLR deficiency | LDLR | TAT--TGT | Tyr-Cys | 807 | Davis | Cell | 45 | 15 | 86 |
| LDLR deficiency | LDLR | TGCc-TGA | Cys-Term | 660 | Lehrman | JBC | 262 | 401 | 87 |
| Leprechaunism | INSR | gAAG--GAG | Lys-Glu | 460 | Kadowaki | Science | 240 | 787 | 88 |
| Leprechaunism | INSR | cCAG--TAG | Gln-Term | 672 | Kadowaki | Science | 240 | 787 | 88 |
| Maple syrup urine disease | BCKD | cTAC--AAC | Tyr-Asn | 394 | Zhang | JCI | 83 | 1425 | 89 |
| Osteogenesis imperfecta (I) | COLIA1 | tGGT--TGT | Gly-Cys | 748 | Vogel | JBC | 262 | 14737 | 87 |
| Osteogenesis imperfecta (I) | COLIA1 | tGGT--TGT | Gly-Cys | 1017 | Labhard | MBM | 5 | 197 | 88 |
| Osteogenesis imperfecta (1) | COLIA2 | GGT--GAT | Gly-Asp | 907 | Baldwin | JBC | 264 | 3002 | 89 |
| Osteogenesis imperfecta (2) | COLIA1 | cGGA--AGA | Gly-Arg | 664 | Bateman | JBC | 263 | 11627 | 88* |
| Osteogenesis imperfecta (2) | COLIA2 | GGT--GTT | Gly-Val | 256 | Patterson | JBC | 264 | 10083 | 89 |
| Osteogenesis imperfecta (4) | COLIA1 | tGGT--AGT | Gly-Ser | 832 | Marini | JBC | 264 | 11893 | 89 |
| Osteogenesis imperfecta (2) | COLIA1 | tGGC--TGC | Gly-Cys | 904 | Costanti. | JCI | 83 | 574 | 88 |
| OTC deficiency | OTC | CGA--CAA | Arg-Gln | 109 | Maddalena | JCI | 82 | 1353 | 88* |
| OTC deficiency | OTC | CGA--CAA | Arg-Gln | 26 | Grompe | JBCS | 13D | 41 | 89* |
| OTC deficiency | OTC | gCAG--GAG | Gln-Glu | 216 | Grompe | PNAS | 86 | 5888 | 89 |
| OTC deficiency | OTC | aGAA--TAA | Glu-Term | 154 | Grompe | PNAS | 86 | 5888 | 89 |
| Phenylketonuria | PAH | tCGG--TGG | Arg-Trp | 408 | Dilella | Nature | 327 | 333 | 87* |
| Phenylketonuria | PAH | CTG-CCG | Leu-Pro | 311 | Licht.-K. | Biochem | 27 | 2881 | 88 |
| Phenylketonuria | PAH | cGAA--AAA | Glu-Lys | 280 | Lynnet | AJHG | 44 | 511 | 89* |
| Porphyria | UROD | GGG-GAG | Gly-Glu | 281 | De Verne. | Science | 234 | 732 | 86 |
| Porphyria | UROD | GGG--GTG | Gly-Val | 281 | Garey | Blood | 73 | 892 | 89 |
| Protein C deficiency | PROC | cCGA--TGA | Arg-Term | 306 | Romeo | PNAS | 84 | 2829 | 87* |
| Protein C deficiency | PROC | TGGa-TGC | Trp-Cys | 402 | Romeo | PNAS | 84 | 2829 | 87 |
| Protein C deficiency | PROC | gCGG--TGG | Arg-Trp | 169 | Matsuda | NEJM | 319 | 1265 | 88* |
| PNP deficiency | NP | tGAA--AAA | Glu-Lys | 89 | Williams | JBC | 262 | 2332 | 87 |
| Rickets (vitamin D resistant) | VDR | CGA--CAA | Arg-Gln | ? | Hughes | Science | 242 | 1702 | 89* |
| Tay-Sachs disease | HEXB | CGC--CAC | Arg-His | 178 | Ohno | J Nchem | 50 | 316 | 88* |
| Tay-Sachs disease | HEXB | cGAA--AAA | Glu-Lys | 482 | Nakano | J Nchem | 51 | 984 | 88* |
| Tay-Sachs disease | HEXB | aGGT--AGT | Gly-Ser | 269 | Navon | Science | 243 | 1471 | 89 |
| TPI deficiency | TPI | GAGt-GAC | Glu-Asp | 104 | Daar | PNAS | 83 | 7903 | 86 |
| TSH deficiency | TSHB | tGGA--AGA | Gly-Arg | 29 | Hayashiz. | Embo J | 8 | 2291 | 89 |
| von Willebrand type 2a | VWF | GTC--GAC | Val-Asp | 844 | Ginsburg | PNAS | 86 | 3723 | 89 |
| von Willebrand type 2a | VWF | cCGG--TGG | Arg-Trp | 834 | Ginsburg | PNAS | 86 | 3723 | 89* |

[a] *Abbreviations*

| | |
|---|---|
| ADA | Adenosine deaminase |
| AJHG | Am J Hum Genet |
| Apo | Apolipoprotein |
| APRT | Adenosinephosphoribosyltransferase |
| BBRC | Biochem Biophys Res Commun |
| CD | Codon |
| G6PD | Glucose-6-phosphate dehydrogenase |
| HPRT | Hypoxanthinephosphoribosyltransferase |
| JBC | J Biol Chem |
| JCI | J Clin Invest |
| JCBS | J Cell Biochem [Suppl] |
| J Nchem | J Neurochem |
| LDLR | Low density lipoprotein receptor |
| MBM | Mol Biol Med |
| MODY | Maturity-onset diabetes of the young |
| NAR | Nucleic Acids Res |
| NEJM | N Engl J Med |
| NIDDM | Non-insulin dependent diabetes mellitus |
| OTC | Ornithine transcarbamylase |
| PNAS | Proc Natl Acad Sci USA |
| PNP | Purine nucleoside phosphorylase |
| TPI | Triose phosphate isomerase |
| TSH | Thyroid stimulating hormone |

6. Point mutations within intron/exon splice junctions and cryptic splice sites are not included. Such mutations are not uncommon but consideration of mutations occurring within a consensus sequence common to all genes would be expected to bias any analysis of mutation at the dinucleotide level.

*Mutation in CpG dinucleotides*

*Frequency.* The total of 139 point mutations listed causing human genetic disease are summarized in Table 2; those occurring in CpG dinucleotides account for 52 of them (37.4% of the total). However, if only CG → TG

**Table 2.** Number of point mutations ($N$) consistent with methylation-mediated deamination of 5-methyl cytosine (**b**) and of all other types (**a**). A, Adenine; C, cytosine; G, guanine; T, thymidine

**a** Point mutations causing human genetic disease other than CG → TG or CG → CA

| Initial nucleotide | Nucleotide resulting from single base-pair change | | | | Total |
|---|---|---|---|---|---|
| | A | C | G | T | |
| A | – | 0 | 8 | 2 | 10 |
| C | 7 | – | 8 | 7 | 22 |
| G | 18 | 10 | – | 14 | 42 |
| T | 4 | 10 | 7 | – | 21 |
| Total | 29 | 20 | 23 | 23 | 95 |

**b** Point mutations in CpG dinucleotides consistent with methylation-mediated deamination of 5-methyl cytosine

| Mutation | $N$ |
|---|---|
| CG → TG | 21 |
| CG → CA | 23 |
| Total | 44 |

and CG → CA mutations (i.e. consistent with methylation-mediated deamination, Table 2b) are considered, this figure falls to 44 (31.7%). This is similar to the proportion first noted by Cooper and Youssoufian (1988) with a smaller sample of human gene mutations.

Throughout the following sections, we intend to calculate expected numbers of point mutations observed within a specific dinucleotide under the assumption of equal mutability of all dinucleotides. These expectations are sequence-dependent in so far as they must account for the actual number of dinucleotides within a given sequence and for the redundancy of the genetic code. The corresponding computational methods have been outlined in detail by Vogel (1972) for calculations at the protein level, and the formulae presented there can easily be transformed for our purposes. However, the DNA sequences of many of the genes listed in Table 2 are only partially known or are difficult to access, so that a precise evaluation of the expected numbers of mutations is not possible at this stage. Therefore, we have adopted a simpler approach, calculating the expected number of mutations within a specific dinucleotide as the product of the total number of mutations observed and the known dinculeotide frequency (Setlow 1976; Nussinov 1981). The results of such calculations will, of course, differ from the precise values for a single gene sequence, but we expect that such differences will average out if a large number of genes is considered. We checked this assumption for a sample of 13 genes covering 66 of the total 139 mutations reported in Table 2 (AT3, INSR, INS, FX, FXI, G6PD, OAT, HCF2, HPRT, OTC, PAH, PROC, and VWF). For each gene, the expected number of mutations observed within a dinucleotide, $d_1 d_2$, was estimated as the product of the total number of mutations observed for that gene and the relative frequency of possible point mutations that would occur within $d_1 d_2$ and cause an amino acid change. No significant differences were found between the sum of these figures, taken over the 13 sequences examined, and the expectations based on known dinucleotide frequencies. This supports our assumption that the latter method introduces no errors worthy of note.

If we assume the 5-methylcytosine makes up approximately 1% of bases in the human genome and that all CG → TG and CG → CA mutations reported in Table 2b, (i.e. 31.7% of the total) are the result of 5mC deamination, then 5mC appears to be approximately $(0.317 \cdot 0.99)/(0.01 \cdot 0.683) \cong 46$ times more mutable than any other nucleotide. This estimate is close to that published previously (Cooper and Youssoufian 1988; Cooper and Krawczak 1989).

If all dinucleotides were equally likely to mutate, then the proportion of mutations occurring in CpG dinucleotides would be expected to be

$$2 \cdot 0.21 \cdot 0.21 \cdot 0.37 = 0.032,$$

which is twice the product of the two mononucleotide frequencies and the CpG deficiency observed in human gene coding regions (Nussinov 1981). Thus, the observed frequency of occurrence of mutation in CpG dinucleotides (0.374) is 11.7 times higher than expected.

For the average remaining dinucleotide, the observed frequency among all dinucleotides involved in a point mutation is

$$(278 - 52)/(278 \cdot 15) = 0.054,$$

whereas the expected frequency is approximately $0.25^2 = 0.0625$. Hence the observed/expected ratio is 0.86 for dinucleotides other than CpG. Although calculated on the basis of different stochastic models, the latter ratio is comparable to the one derived for CpG and we may thus infer that a CpG dinucleotide is $11.7/0.86 = 14$ times more likely to mutate than an average remaining dinucleotide. Further, 85% of mutations within CpG dinucleotides are C → T and G →A transitions consistent with methylation-mediated deamination. Interestingly, if we disregard the CG → TG and CG → CA mutations, an excess of mutations in CpG dinucleotides is still observed. A total of eight were found, a figure twice as high as expected from known mononucleotide and dinucleotide frequencies.

Among the 139 point mutations listed, we have observed transversion and transition frequencies of 37% and 63%, respectively. There is therefore a highly significant excess of transitions compared with the expected frequency (33%) [$\chi^2 = 48.9$; 1 $df$; $P < 0.001$]. These proportions are similar to those reported earlier by Vogel and Kopun (1977) and Vogel et al. (1976) for globin genes, and by Li et al. (1984) for a collection of pseudogene sequences. Most but not all of the excess of transitions over expectation can be attributed to the hypermutability of the CpG dinucleotide; when CpG mutation data are removed (44 mutations, 51% of all transitions) from the analysis, the excess of transitions is nevertheless still significant ($\chi^2 = 4.2$; 1 $df$; $P < 0.05$).

The frequency of C → T and G → A mutations in each of the four CpG-containing arginine codons (Table

3b) were compared after taking into account the differential usage of these codons in human genes. No obvious difference was noted. Thus, our own analysis does not substantiate the finding of Bains and Bains (1987) who reported differences in the mutability of CpG dinucleotides with different flanking nucleotides.

If we turn to the mutations occurring outside CpG dinucleotides (Table 2a), these appear to be unequally distributed with the order of probability of base mutation G > C > T > A. Possible reasons for this apparent non-randomness will be examined further in later sections.

*Codon choices and the CpG dinucleotide.* Selection seems to act to some extent at the level of codon usage in such a way as to reduce reliance upon the presence of the CpG dinucleotide. Table 3a illustrates the codon 'choices' made by the genes of human and eight other vertebrates for the amino acids serine, proline, threonine and alanine, one of whose codons in each case contains a CpG dinucleotide. For all codon choices, considerable avoidance of the codon containing a CpG dinucleotide is exhibited. However, despite the redundancy of the genetic code, complete avoidance of CG-containing codons within the coding sequences is not seen. It is nevertheless difficult to distinguish the possible selective pressures of mutation avoidance from those resulting from mere structural constraints at the DNA/RNA level (Cooper and Gerber-Huber 1985).

Avoidance of CG-containing codons is also seen with arginine (Table 3b) at least for codons CGA, CGG and CGT. It was anticipated that avoidance would be at its highest for CGA since deamination of 5mC in this codon could give rise to a TGA termination codon and selection against the use of CGA might be expected to be correspondingly high. However, contrary to this prediction, CGA occurs more frequently than CGT, although less frequently than CGC and CGG. As a control, we also examined 3670 bp of DNA encoding 18S rRNA, 5S rRNA, U1 snRNA and U4a snRNA genes and 8 tRNA genes. Because the expression of these genes does not result in a protein product, no especial avoidance of CGA was to be expected since a C → T transition giving rise to TGA would be no more deleterious than such a transition in any other 'codon'. Consistent with this view, a total of 63 CGA trinucleotides (1.76%) were found in all 'reading frames', exactly the proportion predicted from mononucleotide frequencies [$f(C) = 0.278$, $f(G) = 0.294$, $f(A) = 0.215$; $f(C) \cdot f(G) \cdot f(A) = 0.0172$, i.e. expected proportion is 1.72%)].

**Table 3.** Data are taken from Maruyama et al. (1986); these authors screened 40328 codons in 135 genes for humans; for 8 other vertebrates (hamster, mouse, rat, bovine, rabbit, chicken, fish, *Xenopus*), 61823 codons in 222 genes were screened

**a** Codon choices for serine, proline, threonine and alanine

| Amino acid | Codon | Humans (% occurrence) | Average 9 vertebrates (% occurrence) |
|---|---|---|---|
| Serine | TCA | 0.92 | 0.78 |
| | TCC | 1.87 | 1.89 |
| | TCG | 0.43 | 0.42 |
| | TCT | 1.43 | 1.29 |
| | AGC | 2.05 | 1.92 |
| | AGT | 0.87 | 0.91 |
| Proline | CCA | 1.18 | 1.17 |
| | CCC | 1.85 | 1.62 |
| | CCG | 0.60 | 0.47 |
| | CCT | 1.43 | 1.27 |
| Threonine | ACA | 1.41 | 1.35 |
| | ACC | 2.51 | 2.52 |
| | ACG | 0.55 | 0.59 |
| | ACT | 1.34 | 1.21 |
| Alanine | GCA | 1.27 | 1.46 |
| | GCC | 2.98 | 3.35 |
| | GCG | 0.59 | 0.65 |
| | GCT | 2.01 | 2.19 |

**b** Codon choices for arginine

| Codon | Humans (% occurrence) | Average 9 vertebrates (% occurrence) | Number of C → T or G → A[a] |
|---|---|---|---|
| CGA | 0.52 | 0.46 | 7 |
| CGC | 1.11 | 1.13 | 8 |
| CGG | 0.77 | 0.77 | 9 |
| CGT | 0.36 | 0.44 | 6 |
| AGA | 1.08 | 1.09 | – |
| AGG | 1.14 | 1.12 | – |

[a] Number of C → T or G → A mutations observed in human mutation data set (from Table 1)

*A bias in clinical detection?*

We next wished to test the hypothesis that certain base-pair changes giving rise to specific amino acid substitutions come to clinical attention more readily on account of the severity of the resulting phenotype. Such an effect could also account for the observed non-randomness of the point mutations causing human genetic disease. Several methods have been reported for assessing the relative net effect of a specific amino acid exchange. Designed as a means to make evolutionary comparisons between amino acid sequences of proteins, these methods were originally used to demonstrate that homologous proteins have evolved in such a way as to minimize the conformational effects of their amino acid replacements (Epstein 1967; Grantham 1974). Perhaps the best comparative measure of amino acid relatedness available is that devised by Grantham (1974), who combined the three interdependent properties of composition, polarity and molecular volume to assign each amino acid pair a mean chemical difference ('D value'). We used these values in order to determine whether or not, for the non-CpG mutations that we have listed, there is a correlation between the extent of the chemical difference between the amino acids exchanged and the probability of clinical detection (i.e. severity of phenotype).

Of the 576 possible point mutations (i.e. 9 different point mutations in 64 different codons), a total of 426 mutations cause an amino acid substitution without

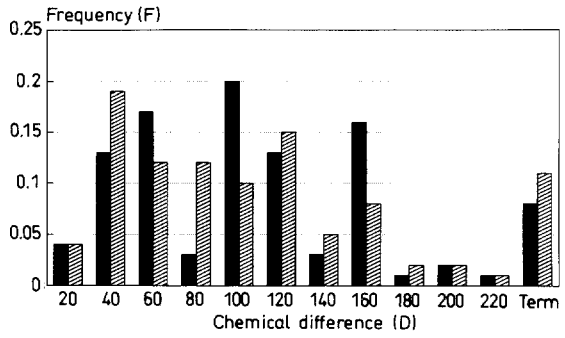


**Fig. 1.** Frequencies (*F*) of different classes of chemical differences (*D*) and termination codons (*term*) among *observed* and *possible* amino acid substitutions caused by point mutations. The number 20,..., 220 refer to intervals 0–20,..., 200–220. ■ Observed, ▨ possible

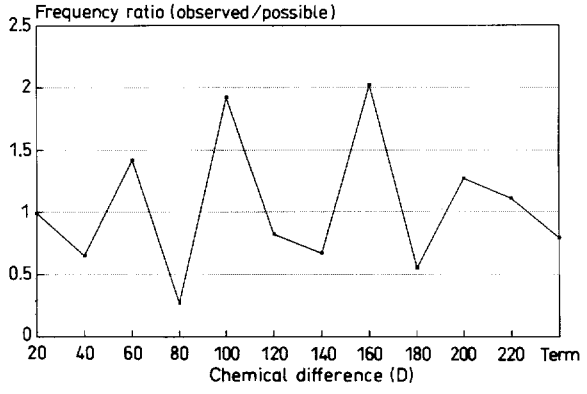


**Fig. 2.** Ratios (*observed/possible*) for the frequencies presented in Fig. 1. *Chemical difference (d)* intervals 0–20,..., 200–220

changing a CpG dinucleotide within the codon to either TG or CA. The range of D values corresponding to these amino acid substitutions was divided into 11 equal intervals (0–20, 20–40,... 200–220) and an additional class was defined for substitutions resulting in a termination codon. The frequencies of these classes among all possible point mutations are presented in Fig. 1, together with the frequencies for the 95 non-CpG mutations listed in Tables 1 and 2a. The ratios of the two frequencies (observed/possible) are presented in Fig. 2. If there were a correlation between the chemical difference corresponding to a certain amino acid exchange and the probability of clinical detection, then the curve in Fig. 2 should exhibit a systematic trend to either increase or decrease. However, Fig. 2 shows no obvious relationship between the extent of phenotypic change as assessed by the chemical difference between exchanged amino acids and the number of mutations detected.

### *Misalignment mutagenesis and the relative frequencies of single base-pair substitutions*

DNA replication occurs as a result of an accurate yet error-prone multistep process. The final accuracy is thought to be dependent upon (i) the initial fidelity of the replicative step and (ii) the efficiency of subsequent error correction mechanisms (Loeb and Kunkel 1982). Since DNA polymerases are involved in replication, recombination and repair processes, their base incorporation fidelity is probably a critical factor in determining mutation rates in the cell. We have tested the hypothesis that non-random base misincorporation during DNA replication is a major contributory factor in producing the spectrum of mutations that we have observed causing human genetic disease. To this end, we compared the observed base substitution frequencies from Table 2 with the measured in vitro base misincorporation frequencies exhibited by vertebrate DNA polymerases α and β (Kunkel and Alexander 1986). These are the major polymerases required for the replication and repair of vertebrate chromosomal (nuclear) DNA; mitochondrial DNA replication is catalyzed by DNA polymerase γ (reviewed by Roberts and Kunkel 1986). DNA polymerase δ (Kunkel et al. 1987) is much more accurate than the other polymerases, but as yet no data exist regarding the relative misincorporation frequencies. The frequency of specific base substitution errors made by purified vertebrate DNA polymerases α and β have been determined using a bacterial complementation assay to detect mutations in lacZα/M13mp2 constructs synthesized in vitro (Kunkel 1985a, b). Interestingly, the frequency of transitions is higher than that of transversions for all the purified polymerases examined in the in vitro assay (Kunkel and Alexander 1986). One model put forward to account for these errors involves the transient misalignment of the primer template during DNA replication resulting in mispair formation and incorrect base incorporation ('misalignment' or 'dislocation' mutagenesis; Kunkel and Alexander 1986). These authors calculated the frequencies of formation of specific mispairs for both polymerases α and β. Although approximate, these values may nevertheless be considered as possessing some validity relative to each other. These mispairing frequencies, together with the resulting base substitutions, are given in Table 4. The calculated sums of the polymerase mispairing frequencies were compared with the observed frequencies of single base-pair substitutions causing human genetic disease.

The Spearman rank correlation test indicated a significant correlation between the two sets of values ($S = 0.563$; $P = 0.06$). This may be interpreted as meaning that some 56% of the variance of the observed single base-pair substitution data is explicable in terms of polymerase-mediated misincorporation of bases during DNA replication. In other words, the data are consistent with a large proportion of the observed non-CpG mutations causing genetic disease having arisen through misincorporation, perhaps by a mechanism involving transient base misalignment.

### *Influence of 5′ (preceding) and 3′ (following) bases upon the frequency of point mutation*

In order to try to discern patterns of mutation at the dinucleotide level, the sequence context of each mutated base was examined. Clearly, each point mutation can be regarded as occurring within two distinct dinucleotides

**Table 4.** Mispairing frequencies of vertebrate DNA polymerase α and β and comparison with mutational spectrum causing human genetic disease. Comparison with mispairing frequencies associated with human α and rat β polymerases with the observed frequencies of single base-pair (non-CpG) substitutions causing human genetic disease. Spearman rank correlation analysis was carried out for the sum of the two corresponding frequencies

| Base-pair substitution | Frequency[a] ($\times 10^{-5}$) | | Sum ($\times 10^{-4}$) | No. of point mutations[b] |
|---|---|---|---|---|
| | Human poly α | Rat poly β | | |
| A → G | 2.56 | 47.62 | 5.02 | 8 |
| A → C | < 3.85 | < 9.09 | < 1.29 | 0 |
| A → T | 13.33 | 4.55 | 1.79 | 2 |
| T → C | 21.74 | 47.62 | 6.94 | 10 |
| T → G | 3.03 | 7.14 | 1.02 | 7 |
| T → A | 5.56 | 6.67 | 1.22 | 4 |
| G → A | 16.13 | 29.41 | 4.55 | 18 |
| G → T | 29.41 | 28.57 | 5.79 | 14 |
| G → C | 38.46 | 9.09 | 4.76 | 10 |
| C → T | 33.33 | 111.11 | 14.44 | 7 |
| C → A | 17.54 | <10.31 | < 2.79 | 7 |
| C → G | < 5.00 | <12.05 | < 1.70 | 8 |

Spearman rank correlation coefficient: 0.563; $P = 0.06$ (two-sided)

[a] Data from Kunkel and Alexander (1986); these data are not dissimilar to those mispairing frequencies associated with *Drosophila* polymerase α (Mendelman et al. 1989)
[b] Data from Table 2a

**Table 5.** Frequency of occurrence of mutations within specific dinucleotides. *rm*, Mutability relative to the dinucleotide TA; $\chi^2$, Chi-square analysis. $\chi^2 = (O - E)^2/E$ in each row. The total $\chi^2$, even if CG is excluded, is highly significant ($P < 0.001$)

| Dinucleotide | | Number of mutations | | Relative mutability (*rm*) | $\chi^2$ |
|---|---|---|---|---|---|
| Symbol | Frequency[a] (*f*) | Observed[b] (*O*) | Expected[c] (*E*) | | |
| TT | 0.097 | 6 | 26.97 | 1.38 | 16.30 |
| CT | 0.071 | 17 | 19.74 | 5.35 | 0.38 |
| AT | 0.081 | 7 | 22.52 | 1.93 | 10.69 |
| GT | 0.049 | 21 | 13.62 | 9.57 | 4.00 |
| TC | 0.057 | 16 | 15.85 | 6.27 | 0.00 |
| CC | 0.047 | 21 | 13.07 | 9.98 | 4.82 |
| AC | 0.054 | 6 | 15.01 | 2.48 | 5.41 |
| GC | 0.043 | 26 | 11.95 | 13.50 | 16.50 |
| TA | 0.067 | 3 | 18.63 | 1.00 | 13.11 |
| CA | 0.074 | 7 | 20.57 | 2.11 | 8.95 |
| AA | 0.097 | 7 | 26.97 | 1.61 | 14.78 |
| GA | 0.061 | 20 | 16.96 | 7.32 | 0.55 |
| TG | 0.074 | 23 | 20.57 | 6.94 | 0.29 |
| CG | 0.010 | 52 | 2.78 | 116.13 | 871.44 |
| AG | 0.070 | 12 | 19.46 | 3.83 | 2.86 |
| GG | 0.050 | 34 | 13.90 | 15.19 | 29.07 |
| Sum | 1.000 | 278 | 278.00 | – | 999.15 |
| | | | | CG excluded: | 121.71 |

[a] Data for human from Setlow (1976)
[b] Number of mutations per dinucleotide. Since every mutation occurs in two distinct dinucleotides, the column total represents almost twice the number of mutations given in Table 1
[c] Expected number of mutations calculated as $f \cdot 278$

depending upon whether one considers the 5′ or the 3′ neighbouring base. This is the reason that the column total of mutations causing genetic disease given in Table 5 represents twice the number of mutations listed in Table 1. An increased likelihood of mutation within specific dinucleotides would be reflected in a deviation from expected values, calculated as the product of known dinucleotide frequencies and the total number of mutations. Observed and expected frequencies of mutation are presented in Table 5.

As expected, there is a striking difference between the observed and expected frequencies of mutation for the CG dinucleotide. This dinucleotide alone contributes a $\chi^2$ of 871.44. However, even if CG is excluded, the remaining $\chi^2$ is as high as 127.71, which is significant at a level above 99.9% (14 *df*).

The mutability of each dinucleotide, $drm(d)$, relative to the least mutable one, TA, is calculated as

$$drm(d) = O(d) \cdot E(\mathrm{TA}) / [O(\mathrm{TA}) \cdot E(d)],$$

where $O$ and $E$ denote the observed and expected frequencies, respectively. This value represents an estimate of how much more likely a dinucleotide $d$ is to mutate than the dinucleotide TA, i.e. an estimate of $P(\text{mutation} \mid d)/P(\text{mutation} \mid \mathrm{TA})$. Here, $P(A \mid B)$ denotes the conditional probability of event $A$ given event $B$. For the CpG dinucleotide, the relative mutability can be split up into three parts:

$$drm(\mathrm{CG} \rightarrow \mathrm{TG}) = P(\text{mutation to TG} \mid \mathrm{CG})/P(\text{mutation} \mid \mathrm{TA})$$

$$drm(\mathrm{CG} \rightarrow \mathrm{CA}) = P(\text{mutation to CA} \mid \mathrm{CG})/P(\text{mutation} \mid \mathrm{TA}) \text{ and}$$

$$drm(\mathrm{CG}^*) = drm(\mathrm{CG}) - drm(\mathrm{CG} \rightarrow \mathrm{TG}) - drm(\mathrm{CG} \rightarrow \mathrm{CA}).$$

Of the 52 CpG mutations reported in Table 1, 21 were CG → TG and 23 were CG → CA. Thus, the total relative mutability ($drm$) of 116.13 yields

$$drm(\mathrm{CG} \rightarrow \mathrm{TG}) = 46.90$$
$$drm(\mathrm{CG} \rightarrow \mathrm{CA}) = 51.37$$
$$drm(\mathrm{CG}^*) = 17.86.$$

Ordering of the dinucleotides by virtue of their derived relative mutabilities gave the rank order: CG > GG > GC > CC > GT > GA > TG > TC > CT > AG > AC > CA > AT > AA > TT > TA. It was thought possible that the differences in mutability could be related to the relative stability of the dinucleotide pairs. The above rank order was therefore compared with that reported by McMahon and Tinoco (1978) for relative dinucleotide stabilities as measured from double-stranded RNA: GG > GC > CG > AC, AG, TC, TG > TA, AT > AA.

The similarity between two such rankings can be measured as follows. If there are $n$ elements that are present in both lists, then each pair of such elements ($E_1$, $E_2$) is assigned a value $\Delta(E_1, E_2) = 2$, if the relationships between $E_1$ and $E_2$ are strictly different in both rankings (i.e., $E_1 > E_2$ in one, $E_2 > E_1$ in the other); each pair is assigned a value $\Delta(E_1, E_2) = 1$, if the relation is strict in one ranking, but $E_1$ and $E_2$ are found to be equivalent in

the other; each pair is assigned a value $\Delta(E_1, E_2) = 0$ in the remaining cases. The sum of these values over all pairs of elements is then divided by the maximum possible sum, which is given by $n^2-n$. Subtracting this ratio from unity gives a measure of similarity.

The two rankings given above, however, include dinucleotides for which the mutabilities depend upon whether they can occur adjacent to a CpG dinucleotide. If this is so, then possible methylation of the cytosine residue in that particular CpG increases the likelihood of mutation of the original dinucleotides. Thus, for the purpose of comparison, those 44 mutations in CpG dinucleotides that were either CG → TG or CG → CA were excluded from the calculation of relative mutabilities. The corresponding new sub-order, including only dinucleotides also considered by McMahon and Tinoco (1978), is GG > TG > GC > TC > AG > AC > AT > AA > TA and the similarity value of the two rankings is 0.69 (69%).

*Relative mutability of single nucleotides*

The relative mutability of a dinucleotide, as estimated above, is proportional to its absolute mutation rate. Now we want to calculate, from a given sequence context $-b_1-b-b_2-$, a value $srm(b \rightarrow b')$ proportional to the probability that the single nucleotide $b$ mutates to $b'$ ($b'$ different from $b$). The value $srm(b \rightarrow b')$ will be defined as the sum of two values $srm_1(b \rightarrow b')$ and $srm_2(b \rightarrow b')$, where $srm_1$, and $srm_2$ take into account the two flanking nucleotides $b_1$ and $b_2$, respectively.

Let $c_1(d)$ and $c_2(d)$ be the relative frequencies of mutation at the first and second position within a dinucleotide, $d$. Since simultaneous mutations at both positions are unlikely, we may assume that $c_1 + c_2 = 1$; the two $c$-values can be estimated for each dinucleotide from the data given in Table 1 (results not shown; all $c$-values were estimated only from mutations other than CG → TG and CG → CA).

Our data on point mutations are still limited, so that the rate at which one dinucleotide mutates into another specific dinucleotide cannot be estimated accurately. We have therefore assumed that, with the exception of CpG, all mutations of a certain dinucleotide are equally likely if the position of the mutation within that dinucleotide is known. Hence, we define for a given sequence $-b_1-b-b_2-$ and a nucleotide $b'$ different from $b$

$srm_1(b \rightarrow b') = [c_1(b-b_2)\cdot drm(b-b_2)]/3$ if $b-b_2 \neq$ CG,
$srm_2(b \rightarrow b') = [c_2(b_1-b)\cdot drm(b_1-b)]/3$ if $b_1-b \neq$ CG,
$srm_1(b \rightarrow b') = [c_1(\mathrm{CG})\cdot drm(\mathrm{CG}^*)]/2$ if $b-b_2 =$ CG and $b' \neq$ T,
$srm_2(b \rightarrow b') = [c_2(\mathrm{CG})\cdot drm(\mathrm{CG}^*)]/2$ if $b_1-b =$ CG and $b' \neq$ A,
$srm_1(b \rightarrow b') = drm(\mathrm{CG} \rightarrow \mathrm{TG})$ if $b-b_2 = CG$ *and* $b' =$ T,
$srm_2(b \rightarrow b') = drm(\mathrm{CG} \rightarrow \mathrm{CA})$ if $b_1-b =$ CG and $b' =$ A, and finally
$srm_2(b \rightarrow b') = srm_1(b \rightarrow b') + srm_2(b \rightarrow b')$.

For any two sequences $-a_1-a-a_2-$ and $-b_1-b-b_2-$, the ratio $srm(a \rightarrow a')/srm(b \rightarrow b')$ gives an estimate of how much more likely a mutation of nucleotide $a$ to $a'$ is than a mutation of $b$ to $b'$.

*Relative mutability at the amino acid sequence level*

In order to correlate the nucleotide sequence of a given gene with the incidence of disease caused by point mutations within that gene, we transform our relative mutability estimates from the nucleotide level to the amino acid sequence level.

Let us consider a triplet $b_1-b_2-b_3$ coding for an amino acid $A$. For a single base change within that triplet, e.g. $b_1-b_2-b_3 \rightarrow b_1'-b_2-b_3$, the probability of this specific change can be assumed to be proportional to $srm(b_1 \rightarrow b_1')$, because multiple mutations within a triplet are unlikely. Thus, the overall probability of a change of the triplet is proportional to the sum of the *srm*-values over all single base changes within $b_1-b_2-b_3$. To calculate the relative mutability of the corresponding amino acid $A$, $arm(A)$, summation is carried out only over those single base changes that cause an amino acid substitution. Thus, the ratio $arm(A)/arm(B)$ is an estimate of how much more likely it is for $A$ than for $B$ to change as a result of a point mutation, taking into account the DNA sequence contexts of both amino acid codons.

*Computer analysis of gene coding sequences*

The calculations described in detail above were introduced into a computer program (MUTPRED) that uses the relative mutabilities of dinucleotides (*drm*-values), the relative frequencies of mutations in the first and second position of a dinucleotide (*c*-values) and the gene coding sequence of interest (GenBank, latest release used here) as input. From these data, a mutability profile of the protein can be derived by plotting the *arm*-values against the corresponding amino acid sequence. Regions prone to amino acid substitutions caused by point mutations can be recognized in such a plot by a concentration of larger bars (see Fig. 3). Additionally, the program calculates the sum of all *arm*-values, (the 'mutability quotient'), which can be taken as a measure of the overall mutability of the amino acid sequence, but which has no stochastic meaning. We shall demonstrate below, however, that the mutation quotient correlates well with the prevalence of inherited disease caused by point mutations within the corresponding gene.

*Application of MUTPRED to point mutations causing haemophilia B*

We have demonstrated here both the non-randomness of point mutation at the dinucleotide level in human genes, and a strong correlation between their mutational spectrum and that predicted from known DNA polymerase mispairing frequencies. The finding that dinucleotides differ with respect to their likelihood of mutation as a consequence of errors in a basic cellular mechanism argues strongly for the general applicability of derived dinucleotide mutability values beyond the

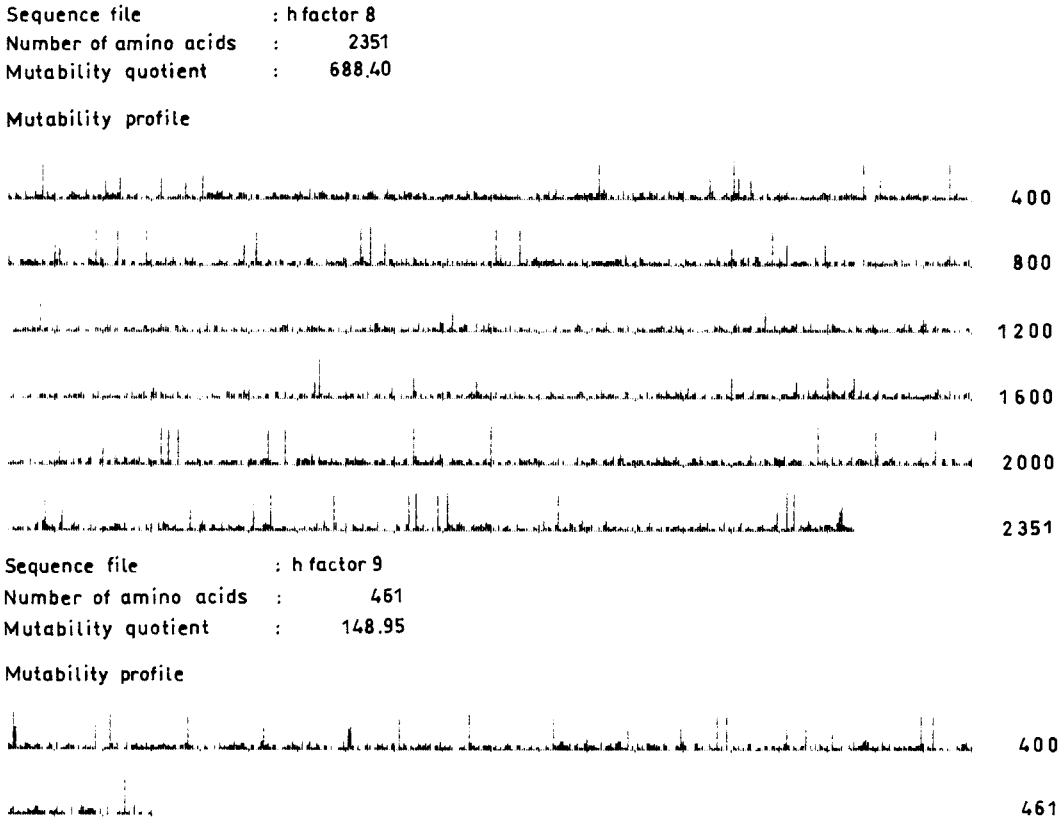


**Fig. 3.** *Mutability quotients* and *mutability profiles* of amino acid sequences of human factor VIII and factor IX proteins using MUTPRED. Each *bar* represents an amino acid, the *number* to the *right* of each *line* is the number of the last amino acid in that *line*

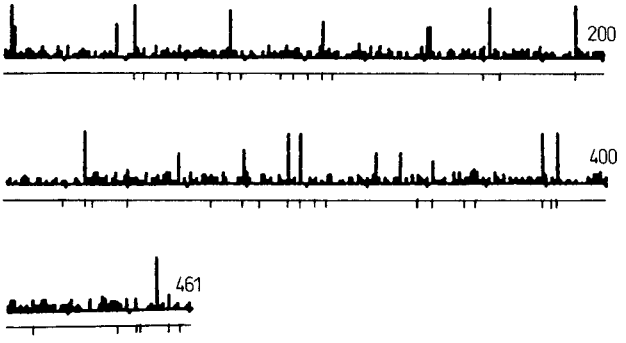


**Fig. 4.** Mutability profile of factor IX gene and observed mutational spectrum causing haemophilia B. Positions at which at least one amino acid substitution has been observed are marked by a *vertical bar*

narrow confines of our relatively small sample of point mutations.

MUTPRED may be tested empirically by directly comparing the predicted spectrum of mutations for a given gene (mutation data for which were not included in Table 1 or in subsequent calculations) with the observed mutational spectrum. The factor IX gene provides us with a unique test system; although the first characterized point mutations in this gene were biased by prior CpG site screening, the advent of direct sequencing of polymerase chain reaction (PCR)-amplified material has now led to the identification of 86 independent (51 different) point mutations that together probably represent a fair cross-section of the point mutations causing haemophilia B (reviewed by Cooper and Tuddenham 1991). A comparison of the observed and predicted mutational spectra (Fig. 4) confirms that our model has predictive value, although it performs better for CpG dinucleotides than for non-CpG dinucleotides.

As it stands, MUTPRED is however unable to predict hotspots of mutation in non-CpG dinucleotides viz. Pro 55 (CCA → GCA), Ala 390 (GCA → GTA) and Ile 397 (ATA → ACA). Whereas examples of truly recurrent mutation in non-CpG dinucleotides are unusual, the Ile 397 → Thr substitution is itself frequent (11 independently reported cases). This uniquely high level of mutation is suggestive of a high frequency sequence-directed event rather than a detection bias through phenotypic effect. We have noted a 76% homology between a 30 bp region in intron 3 of the factor IX gene (9654–9684) and the sequence immediately around codon 397 (31296–31326). Non-homologous pairing of these sequences followed by miscorrection would account for the T → C transition observed. It remains to be seen whether other cases of non-CpG recurrent mutation are explicable in this way. If DNA sequences some distance away from the actual site of mutation must be considered, mutation prediction models are likely to become very complex.

**Table 6.** Coagulation factor deficiencies: prevalence and gene mutability

| Disease | Mode of inheritance | Defective gene(s) | Coding region (bp) | Mutability quotient[a] | Prevalence[b] |
|---|---|---|---|---|---|
| Haemophilia A | Recessive, X-linked | Factor VIII | 7053 | 688 | 5195 |
| Haemophilia B | | Factor IX | 1383 | 149 | 982 |
| Von Willebrand disease (mild) | Autosomal dominant | Von Willebrand factor | 8439 | 1313 | 2215 |
| Fibrinogenaemia | | Fibrinogen α | 1923 } 4785 | 230 } 514 | |
| | | Fibrinogen β | 1473 | 157 | 10 |
| | | Fibrinogen γ | 1389 | 127 | |
| Prothrombin deficiency | | Prothrombin | 1995 | 315 | 2 |
| Factor XIII deficiency | Autosomal recessive | Factor XIIIa | 2196 } 4233 | 285 } 464 | 16 |
| | | Factor XIIIb | 2037 | 179 | |
| Factor XII deficiency | | Factor XII | 1902 | 383 | 64 |
| Factor X deficiency | | Factor X | 1584 | 224 | 37 |
| Factor VII deficiency | | Factor VII | 1578 | 261 | 35 |
| Prekallikrein deficiency | | Prekallikrein | 1971 | 200 | 3 |

[a] Explained in text

[b] Number of patients on UK Haemophilia Centre Director's National Register on 15 January 1988 (Dr. Charles Rizza, personal communication). Most conditions listed give rise to severe bleeding disorders. However, since factor XII deficiency, prekallikrein deficiency and mild von Willebrand's disease are virtually insignificant clinically, the prevalence figures given here are from chance detection and probably grossly underestimate the true prevalence of these conditions. The length of the coding region in bp includes both the native protein and leader peptide sequence

*The mutability quotient correlates with disease prevalence and relative rates of gene mutation*

It would seem reasonable to suppose that a relative measure of the mutability of a gene sequence might be obtained by simple summation of the dinucleotide mutabilities for any given DNA sequence after allowing for the redundancy of the genetic code. This mutability quotient, as calculated above, thus represents a measure of the relative likelihood that an amino acid substitution will occur within a given protein. It is calculated from the empirically-derived dinucleotide mutabilities as applied to all overlapping codon positions along the length of the gene sequence in question. We speculated that, if the primary DNA sequence were indeed an important factor in determining the location and frequency of point mutations, then the mutability quotients might correlate with clinically observed mutation rates for these genes and/or the prevalence of inherited conditions resulting from the absence or alteration of their gene products.

Clearly, the prevalence of a given condition in the general population depends not only upon the mutability of the underlying gene responsible for the disease phenotype, but also upon whether one or two mutant alleles are required for the disease phenotype to be expressed (i.e. the 'mode of inheritance'). For any comparison of this kind, it is therefore necessary to compare autosomal dominant, autosomal recessive and X-linked conditions, separately. We have attempted to compare the relative prevalence of different coagulation factor deficiencies (Dr. Charles Rizza, personal communication) with mutability quotients calculated from coagulation factor gene sequences (Table 6). The most comparable conditions listed here are the haemophilias A and B, both X-linked recessive conditions that give rise to similar bleeding phenotypes as a consequence of the defective production of factors VIII and IX, respectively. We can see from Table 6 that haemophilia A is 5.3 times more prevalent in the UK than haemophilia B. This is similar to the ratio of the mutability quotients (4.6:1). Expansion of the model to include other parameters such as gene length and number of splice junctions could in principle improve this correlation still further.

Similarly with the autosomal dominant conditions, the rank order of the disease prevalence figures and the mutability quotient values are identical (Table 6). The high mutability quotient for the von Willebrand factor gene is especially noteworthy in the light of the findings of Rodeghiero et al. (1987) who reported an extremely high incidence (0.82%) of mild autosomal dominant von Willebrand's disease (vWD) in the general population. The mildness of the symptoms of the disease in the heterozygous form almost certainly accounts for the apparent lower prevalence of vWD in the UK than haemophilia A (Table 6), since it is much less likely to come to clinical attention. No correlation was noted for the autosomal recessive disorders. This is not at all surprising because the prevalence of these rare conditions will be determined predominantly by factors such as founder effect, the degree of inbreeding, genetic drift and possibly even heterozygote advantage in a given environment.

Because the best correlation found between disease prevalence and mutability quotients was observed for the haemophilias, the comparison was extended to calculated mutation rates associated with three other X-linked conditions, Duchenne muscular dystrophy (DMD), ornithine transcarbamylase (OTC) deficiency and Lesch-Nyhan syndrome. In such X-linked recessive lethal conditions, the mutation rate may be easily calculated (reviewed by Vogel and Motulsky 1986), since roughly one third of all mutations will be new. The rank order of the

**Table 7.** Sex-linked recessive 'lethal' conditions. The length of the amino acid coding regions in bp were obtained from GenBank, latest release

| Disease | Disease gene | Coding region (bp) | Mutation rate | Reference | Mutability quotient |
|---|---|---|---|---|---|
| Duchenne muscular dystrophy | DMD | 11055 | $6 \times 10^{-5}$ | Vogel and Motulsky 1986 | 1190 |
| OTC deficiency | OTC | 1064 | $8 \times 10^{-7}$ | S. Malcolm, personal communication[a] | 118 |
| Lesch-Nyhan syndrome | HPRT | 656 | $2 \times 10^{-6}$ | Stout and Caskey 1985[b] | 68 |
| Haemophilia A | Factor VIII | 7053 | $4.4 \times 10^{-5}$ | Vogel and Motulsky 1986 | 688 |
| Haemophilia B | Factor IX | 1383 | $2.5 \times 10^{-6}$ | Vogel and Motulsky 1986 | 149 |

[a] Dr. Malcolm kindly supplied data on UK disease prevalence, from which these figures are calculated
[b] Published data used as basis for calculation of these figures

calculated gene mutation rates again agrees well with that of the mutability quotients (Table 7).

## Discussion

We have drawn up a list of 139 different single base-pair substitutions causing human genetic disease where we believe that the mutation itself has been detected and located in an unbiased fashion. For this collation, it was deemed necessary to exclude several inherited conditions (e.g. haemophilias A and B) for the reason that once hotspots of mutation had been found in CpG dinucleotides, subsequent strategies for mutation detection were designed accordingly, and of course, such mutations were preferentially found. The possibility was considered, however, that the list of point mutations was intrinsically biased from the outset since, for a mutation to have come to our attention, it must have resulted in a phenotype that was severe enough to be detectable clinically but not so severe as to have caused death in utero. This notwithstanding, the relative proportion of transitions (63%) to transversions (37%) reflects that found in a range of evolutionary sequence comparison studies. This lends weight to the view that the observed spectrum of mutations owes more to the vagaries of the replication process than to natural selection, and indicates the probable lack of any significant bias in ascertainment. A second source of bias could have been introduced by an inequality in the mutation rate between the DNA strands as observed by various authors (Vogel and Kopun 1977; Fitch 1980; Wu and Maeda 1987). Any resulting effect is however likely to be lost through an averaging-out process as the sample size increases.

Of the total number of point mutations that we have listed, some 37.4% occur within the CpG dinucleotide. This figure reduces to 31.7% if we consider only the CG → TG and CG → CA transitions compatible with the model of methylation-mediated deamination. This proportion is almost identical to that originally reported for a smaller sample of point mutations (Cooper and Youssoufian 1988) and represents a 14-fold higher probability of mutation at a CpG dinucleotide compared with that exhibited by any one of the remaining 15 dinucleotides. Clearly, whatever the cellular role(s) of DNA methylation (reviewed by Cooper 1983; Bird 1986), we pay a heavy price in terms of the increased mutability of our genes. The in vivo rate of deamination of 5mC in CpG dinucleotides has recently been measured (Cooper and Krawczak 1989); estimates derived from both clinical data and evolutionary sequence comparisons were found to be essentially identical. We have taken advantage of CpG hypermutability to optimize diagnostic screening by employing a "directed-search" strategy for detecting point mutations in the factor VIII gene causing haemophilia A (Pattinson et al. 1990).

Recently, some evidence has been put forward for the presence of cytosine methylation at dinucleotides other than CpG (Woodcock et al. 1987). These authors have further suggested (Woodcock et al. 1988) that maintenance methylases might recognise and methylate the trinucleotide $C^A_TG$ in addition to the dinucleotide CG, thus raising the possiblity of methylation-mediated deamination in $C^A_TG$ sequences. However, in our sample, only two such CAG → TAG substitutions were noted (those causing adrenal hyperplasia and leprechaunism); this is not significantly higher than random expectation in the absence of 5mC deamination. Moreover, contrary to the situation observed for CG-containing codons, CAG and CTG codons (Gln and Leu) occur more frequently in human genes than the alternative codons for these amino acids (Maruyama et al. 1986; Ohno 1988). There would therefore appear at present to be no compelling evidence for the high frequency occurrence of methylation-mediated deamination mutations at sites other than CpG. The recent observation that a eukaryotic DNA methylase methylates CA and CT dinucleotides at a 50-fold-lower rate than CG in vitro (Hubrich-Kühner et al. 1989) is not inconsistent with this finding.

In the process of this analysis, we did however come across evidence for the hypermutability of the CpG dinucleotide over and above the caused by methylation-mediated deamination; when CG → TG and CG → CA transitions are excluded from this analysis, the CpG dinucleotide still appears to mutate more frequently than one would expect on the basis of its frequency in human genes. This suggests that a second mechanism may be involved in CpG hypermutability. Whatever this turns out to be (dislocation mutagenesis?), it is probably responsible for no more than approximately 15% of point mutations (8/52 in our sample) associated with the dinucleotide.

To some extent, the redundancy of the genetic code has provided the vertebrates with the potential of ameliorating the detrimental effects of cytosine methylation through appropriate codon usage. Indeed, codons containing CpG tend to be avoided (reviewed by Cooper and Gerber-Huber 1985). However, despite this redundancy, avoidance is not complete even for the CGA (arginine) codons, which a C → T transition would transform into termination codons. Although genes that encode an RNA as an end-product rather than a protein exhibit neither CpG nor CGA avoidance, this must not immediately be taken as evidence for selection acting at the level of codon usage. It might equally well be an indirect consequence of the uniform absence of cytosine methylation in these sequences in the germline (e.g. in rRNA genes; Cooper et al. 1983).

The majority (68%) of point mutations causing human genetic disease are neither CG → TG nor CG → CA mutations, but some patterns can nevertheless be discerned among them. If substitutions of single nucleotides are considered, the ranking of mutability is guanine > cytosine > thymine > adenine in our sample of 95 'non-CpG' mutations. This non-randomness is still pronounced if dinucleotides are considered, even after allowing for differences in dinucleotide frequencies. In principle, several non-mutually exclusive explanations are possible: (1) non-randomness of the initial mutation event, (2) non-randomness of the mutation repair process and (3) differences between mutations with respect to their phenotypic consequences, and the resulting likelihood of their coming to clinical attention.

The individual contributions of the different replication and repair processes have hardly begun to be unravelled, so it is perhaps prudent first to consider the last of these explanations. Intuitively, it would seem reasonable to suppose that the phenotypic consequences of a given point mutation would be determined ultimately by the magnitude of the amino acid exchange as assessed by the resulting structural perturbation of the protein in question. We have attempted to estimate this magnitude by employing the measure of 'chemical difference' (Grantham 1974) to compare the substituted amino acid with its replacement. However, little if any relationship was found between the chemical difference value attributed to amino acid exchanges and the frequency with which those substitutions were observed. Since the validity of the measure of chemical difference is supported by its negative correlation with the frequency of favoured amino acid substitutions during evolution, we may tentatively conclude that no particular bias in ascertainment has been introduced as far as the phenotypic consequences of a specific point mutation in a particular codon are concerned. Clearly, the phenotypic consequences of a given mutation must depend not only upon the nature of the amino acid substitution, but also upon the location of that substitution within the protein. Mutational effects on protein stability have recently been reviewed (Alber 1989). In general, and with the exception of charged residues, most amino acids that make critical interactions (e.g. disulphide bonds, hydrophobic forces, hydrogen bonds etc.) are rigid or buried in the protein structure, and their mutation will be profoundly destabilizing. Our knowledge of the complex relationship between protein structure and function must however be greatly expanded before we are in a position to be able to incorporate such data into a predictive model.

In theory, base substitutions in evolutionary comparison studies might be expected to occur at different frequencies from those found to cause human genetic disease. This is because the latter are by definition always deleterious whereas the former cannot be disadvantageous either because they are present in an unselected DNA sequence or because, if present in coding sequences, they would never have become fixed in the population. However, in practice, whether one considers single base substitutions in pseudogenes (Maeda et al. 1988; Li et al. 1984; Gojobori et al. 1982; Bains and Bains 1987) where selection is no longer acting, or substitutions between closely related gene sequences (e.g. the α-interferon multigene family; Golding and Glickman 1985, 1986), the hierarchy of frequencies of single base-pair exchanges is similar to that evident from the analysis of human gene mutations causing genetic disease. This strongly suggests that the influence of selection is minimal, and points instead in the direction of non-randomness of mutation in the replication/repair process rather than towards a bias in detection.

The extremely high accuracy of eukaryotic DNA replication is thought to be achieved by means of (i) efficient initial selection of the nucleotides to be incorporated by the DNA polymerases (Roberts and Kunkel 1986), (ii) the removal of mispaired bases by 3′ → 5′ proofreading exonucleases (Kunkel 1988; Bialek et al. 1989) and (iii) post-replicative mismatch-repair processes (Modrich 1987). In comparing the specific base substitution frequencies associated with the vertebrate polymerases α and β (Kunkel and Alexander 1986) with the frequencies of the different transitions and transversions causing human genetic disease, we have effectively excluded any consideration of the efficiency of the different proofreading and post-replicative mismatch-repair mechanisms. This is because the purified polymerase preparations used in vitro lack the 3′ → 5′ exonuclease activities thought to be responsible for proofreading in vivo. The correlation observed between the two sets of data is thus consistent with the explanation that a substantial proportion of the point mutations causing human genetic disease detected to date are caused by misincorporation of bases during DNA replication. Further, it suggests that if cellular DNA proofreading and mismatch-repair mechanisms have attempted to compensate for the initial non-random distribution of mutations causing human genetic disease by correcting high frequency mutations with enhanced efficiency, they have not been so successful as to have obscured the original association. Although in prokaryotes, transition mismatches (G:T and A:C) are reportedly repaired more efficiently than transversion mismatches (G:A and C:T) (Jones et al. 1987), the analogous repair mechanisms and proofreading activities in eukaryotes are relatively uncharacterized and in some cases undoubtedly still remain to be identified. Some evidence, however, is be-

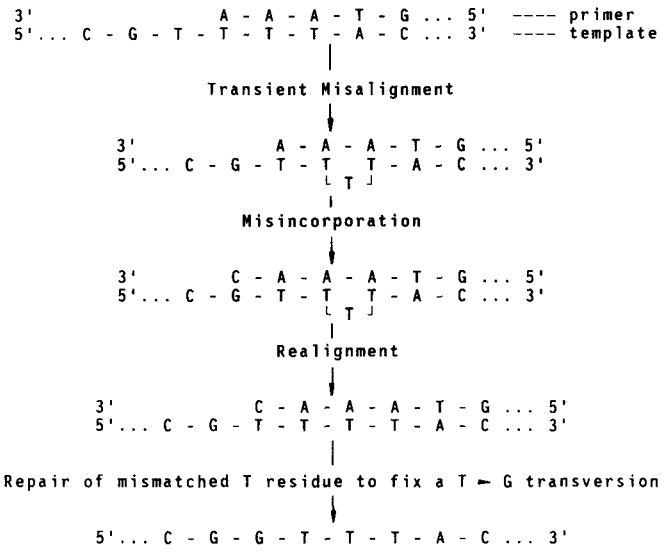


**Fig. 5.** Model of misalignment mutagenesis during DNA replication (after Kunkel and Alexander 1986; Roberts and Kunkel 1986)

ginning to emerge for differential repair of mispaired bases by the sequence-dependent proofreading activities of eukaryotic DNA polymerases (Petruska and Goodman 1985; Reyland et al. 1988). Moreover, Brown and Jiricny (1988) have reported efficient repair of G/T mismatches but with a bias in favour of T → C correction over G → A. Since repair enzymes are known to act upon unstable regions of DNA, it may be that they are less efficient in excising errors in GC-rich regions of duplex DNA (Bessman and Reha-Krantz 1977; Petruska and Goodman 1985); this would contribute still further to the non-randomness of point mutations.

Polymerase β is much more error-prone than polymerase α as far as the accuracy of base incorporation is concerned. Differences between these enzymes are probably to be expected since they differ both structurally and functionally. Polymerase β binds tightly to double-stranded DNA and fulfils a role in repair synthesis of short gaps or nicks in vivo. Polymerase α is by contrast more processive, does not bind to nicks and does not fill small gaps (reviewed by Kunkel 1985a). A proofreading 3′ → 5′ exonuclease activity appears to be associated with polymerase α but none has ever been detected with polymerase β, a finding that may explain the high frequency of errors associated with the latter enzyme.

Two different models for polymerase-associated base misincorporation may be considered. The first of these, put forward by Kunkel (1985a), is not a simple misincorporation but is the result rather of transient misalignment of the primer-template caused by looping out of a template base (Fig. 5). Immediate realignment leads to incorporation of a single mispair. 'Misalignment' or 'dislocation' mutagenesis, as it has become known, is mediated by repetitive DNA sequences in the vicinity; it provides an explanation for many point mutations not predicted by base mispairing hypotheses (e.g. Topal and Fresco 1976), which have instead invoked base analogues and tautomers of the naturally occurring nucleotides. It is consistent also with our observation that dinucleotide mutability correlates with stability since any transient misalignment must be stabilized long enough to permit misincorporation before realignment occurs. An alternative explanation involves the proposed requirement for balanced dNTP pools to maintain the accuracy of the replication process. An excess of any one dNTP can, under certain circumstances, lead to increased misincorporation of that nucleotide (Kunkel et al. 1981; Phear et al. 1987; reviewed by Meuth 1989).

In the light of the data that we have presented on the relative mutabilities of different dinucleotides, the question of the influence of neighbouring bases upon the mutability of a given nucleotide is an important one. Fersht (1979), Kunkel et al. (1981) and Mendelman et al. (1989) have shown that the in vitro substitution rate at a specific base is dependent upon the rate of incorporation of the next nucleotide. This 'next-nucleotide effect' is probably itself dependent upon the state of the available nucleotide pool and upon the local environment of the DNA sequence. The effect of the latter has been effectively demonstrated in prokaryotes, where the influence of the local DNA sequence may extend well beyond neighbouring bases (e.g. see Conkling et al. 1980). In eukaryotes, preliminary evidence has been put forward for a next-nucleotide effect in vivo, both in the evolution of α-interferon gene sequences (Golding and Glickman 1986) and in the generation of spontaneous mutations in the hamster adenine phosphoribosyl transferase (aprt) gene (Phear et al. 1987). The next-nucleotide effect is a property of DNA polymerases with proofreading activities and may be regarded as resulting from competition between the polymerization and proofreading activities of the replication complexes.

An influence of neighbouring bases may thus be expected with both models, which should not therefore be regarded as mutually exclusive. Although other factors are undoubtedly involved (Boosalis et al. 1989; reviewed by Loeb and Kunkel 1982), we surmise that misalignment mutagenesis is more important since (i) the correlation observed with polymerase mispairing frequencies is not dependent upon the presence of a proofreading activity and (ii) the most error-prone polymerase (polymerase β) apparently lacks such an activity. We further conclude that it is probably unnecessary to invoke mechanisms such as mutation caused by exogenous factors, differences in mismatch-repair systems or merely a bias in ascertainment to account for much of the observed spectrum of point mutations causing human genetic disease. We are thus left with two major contrasting mechanisms for the origin of point mutations causing human genetic disease. The majority (approximately 68%) of single base-pair substitutions appear to be replication-associated and thus replication-dependent as distinct from the time-dependent nature of 5-methylcytosine deamination (Cooper and Krawczak 1989), which accounts for approximately 32% of point mutations.

Enzymatic and chemical causes of mutation may sometimes be difficult to disentangle. For example, depurination results from the cleavage of the N-glycosylic bond that connects the purine to the deoxyribose sugar

thus removing the base but leaving the phosphodiester backbone intact (reviewed by Loeb 1985). Depurination is thought to occur at a rate of $3 \times 10^{-11}$ events/base/s under physiological conditions (Lindahl and Nyberg 1972), a figure some 30 times lower than that for deamination of 5mC to thymidine (Ehrlich et al. 1986). Depurination is in the order dG > dA > dC > dT (Loeb and Preston 1986). Apurinic sites may result from exposure to chemical agents that form bulky adducts on DNA (Loeb 1985) or from the action of DNA glycosylase in the enzymatic repair of damaged DNA (Loeb and Preston 1986). Since DNA polymerases misincorporate bases opposite abasic sites (Shearman and Loeb 1979; Kunkel et al. 1983) during replication, this could in principle be an important cause of spontaneous mutation. Deoxyadenosine has been found to be the most frequently misincorporated base with both prokaryotic and eukaryotic polymerases (Sagher and Strauss 1983; Schaaper et al. 1983; Kunkel 1984; Takeshita et al. 1987). Were depurination a major cause of spontaneous mutation, then the mutational spectrum should be heavily biased towards substitutions by deoxyadenosine. In our sample, G → A transitions are indeed the most frequently observed (Table 2a), but this excess is not dramatic, and it would appear to be unnecessary to invoke this mechanism in order to account for the observed mutational spectrum. A further chemical mutation mechanism is through damage by oxygen free radicals (Hsie et al. 1986; Richter et al. 1988); however, the spectrum of mutations in eukaryotic cells arising via this mechanism is not yet known.

In principle, a future larger sample size of human point mutations should permit more accurate and reliable conclusions to be drawn. However, as we come to understand better how point mutations arise and where they are most likely to occur, such information will increasingly be used to optimize the efficiency of searches for gene mutations and novel data will become ever harder to interpret.

Another approach to the analysis spontaneous point mutations has been to clone and sequence mutant hamster aprt genes from cell lines containing only one aprt allele. Three such studies to date (Nalbantoglu et al. 1987; DeJong et al. 1988; Phear et al. 1989) have between them reported a total of 79 different point mutations within the coding region; 26 of these are C → T transitions, which may in part be a reflection of the C-richness of the hamster aprt gene. However, only two base substitutions found were consistent with the model of methylation-mediated deamination and this is thought to be the result of a lack of methylation at CG dinucleotides within the aprt gene (Nalbantoglu et al. 1987; Phear et al. 1989). Of the remainder, a large proportion were found to occur in the dinucleotide CC. Although the rank order of dinucleotide mutabilities differs slightly from that observed with the human mutations reported here, comparison of the eight most mutable dinucleotides from both data sets nevertheless yields seven in common (GG, TG, CG, TC, CC, GC and GA), suggesting a remarkable uniformity in the process underlying mutational change in different systems. A hotspot of mutation was also detected (DeJong et al. 1988); an identical C → T transition was observed independently seven times in the third C of a CCTCCTTCC sequence. The spectrum of spontaneous point mutations in the SUP4-o gene of *Saccharomyces cerevisiae* has also been reported (Giroux et al. 1988) but interpretation of the data was lacking. Further data from other systems are now urgently required to demonstrate the generality or otherwise of these findings.

The empirically-derived values for the relative mutabilities associated with the different dinucleotides were incorporated into a mathematical model (MUTPRED) whose utility could lie in the prediction of the likely location of point mutations within human genes causing genetic disease. The spectrum of point mutations predicted by this model for the human factor IX gene was shown to resemble closely the observed spectrum of point mutations causing haemophilia B. The validity of this model is further supported by its apparent ability to predict the rank order of prevalence and/or mutation rates associated with specific autosomal dominant and X-linked recessive conditions from a knowledge only of the primary DNA sequences of the genes responsible for these disorders. The model itself does not consider point mutations that do not alter the amino acid sequence of a protein and could thus be improved by consideration of the relative likelihood of mutations in splice junctions etc. Since the model also does not consider the possibility of gene lesions other than point mutations (the proportion of other types of gene lesions may vary dramatically between different disorders, e.g. in DMD and haemophilia A, gene deletions account for approximately 60% and 4% of cases, respectively), it is perhaps a little surprising that mutability quotients correlate as well as they do with disease prevalence. Moreover, our value for the relative mutability of a CpG dinucleotide is an average figure based upon the probability ($P \cong 0.8$; Cooper 1983) that a CpG taken at random from the human genome is methylated. In practice, those CpG dinucleotides that are methylated will be more prone to deamination of 5mC than the model would suggest, whereas those that are unmethylated will be much less likely to mutate to thymidine than the 'average CpG'. Since the methylation patterns between and within different genes vary widely (see Gardiner-Garden and Frommer 1987), this all-or-nothing mutational effect of cytosine methylation illustrates the importance of establishing the methylation status of individual CpG sites within specific genes in the germline. We are currently approaching this problem by the application of genomic sequencing (Saluz and Jost 1989).

That the model presented here already appears to possess predictive value is encouraging. The enlargement of the size of the sample of human point mutations causing genetic disease, together with the possible future incorporation of new data on sequence-dependent repair processes and DNA sequences capable of predisposing a gene sequence to frameshift mutations, deletions and perhaps even insertions, promise to improve further the model and extend its applicability and reliability in a predictive context.

The genomic architecture of the locus is undoubtedly also important. More specifically, the number of exons, splice junctions and the length and nature of intronic sequences (including the number and type of repetitive elements) will probably all play a role in determining the frequency and extent of types of lesion other than single base-pair substitutions and frameshift mutations. Finally, the 'private' characteristics of individual genes must not be ignored. Of relevance will be DNA sequences encoding protein regions of structural and functional importance, e.g. those encoding proteolytic cleavage sites (which often contain CpG dinucleotides within Arg codons in coagulation factor genes), protein-binding sites, methylated CGA arginine codons (CGA → TGA creates a termination codon), etc.

The model presented above is thus likely to improve as new data emerge and are incorporated. In the meantime, we are testing the practical utility of the model in its present form by PCR/direct sequencing of gene regions predicted by the model to be especially prone to high frequency point mutation.

Our genes have evolved slowly, probably via a myriad of meandering and circuitous pathways, 'guided' through the millennia of erratic environmental influences by the moulding force of natural selection. Perhaps this hesitant evolutionary past accounts for present-day genes containing, encoded within their base sequences, the potential seeds of their own destruction. How apt in this context is the poet's description of nature; 'so careful of the type she seems, so careless of the single life' (Alfred, Lord Tennyson "in Memoriam A.H.H."; 1850)

*Acknowledgements.* We wish to thank Charles Rizza for communication of disease prevalence data from the UK Haemophilia Centre Directors National Register, David Millar, Jochen Reiss, Sue Malcolm, Michael Barfoot and Ted Tuddenham for useful discussion, Friedrich Vogel for his constructive criticism and Vijay Kakkar and Wolfgang Engel for their encouragement and support.

## References

Alber T (1989) Mutational effects on protein stability. Annu Rev Biochem 58:765–789

Bains W, Bains J (1987) Rate of base substitution in mammalian nuclear DNA is dependent on local sequence context. Mutat Res 179:65–74

Barker D, Schaefer M, White R (1984) Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. Cell 36:131–138

Bessman MJ, Reha-Krantz LJ (1977) Studies on the biochemical basis of spontaneous mutation. V. Effect of temperature on mutation frequency. J Mol Biol 116:115–123

Beutler E, Gelbart T, Han J, Koziol JA, Beutler B (1989) Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc Natl Acad Sci USA 86:192–196

Bialek G, Nasheuer H-P, Goetz H, Grosse F (1989) Exonucleotic proofreading increases the accuracy of DNA synthesis by human lymphocyte DNA polymerase αDNA primase. EMBO J 8:1833–1839

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

Boosalis MS, Mosbaugh DW, Hamatake R, Sugino A, Kunkel TA, Goodman MF (1989) Kinetic analysis of base substitution mutagenesis by transient misalignment of DNA and by miscoding. J Biol Chem 264: 11360–11366

Brown TC, Jiricny J (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54:705–711

Castora FJ, Arnheim N, Simpson NN (1980) Mitochondrial DNA polymorphism: evidence that variants detected by restriction enzymes differ in nucleotide sequence rather than in methylation. Proc Natl Acad Sci USA 77:6415–6419

Conkling MA, Koch RE, Drake JW (1980) Determination of mutation rates in bacteriophage T4 by unneighborly base pairs: genetic analysis. J Mol Biol 143:303–315

Cooper DN (1983) Eukaryotic DNA methylation. Hum Genet 64: 315–333

Cooper DN, Gerber-Huber S (1985) DNA methylation and CpG suppression. Cell Differ 17:199–205

Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83:181–189

Cooper DN, Schmidtke J (1984) DNA restriction fragment length polymorphisms and heterozygosity in the human genome. Hum Genet 66:1–16

Cooper DN, Schmidtke J (1989) Diagnosis of genetic disease using recombinant DNA. Second edition. Hum Genet 83:307–334

Cooper DN, Tuddenham EGD (1991) The molecular genetics of coagulation disorders. Oxford University Press, Oxford (in press)

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

Cooper DN, Errington LH, Clayton RM (1983) Variation in the DNA methylation pattern of expressed and non-expressed genes in chicken. DNA 2:131–140

Cooper DN, Gerber-Huber S, Nardelli D, Schubiger JL, Wahli W (1987) The distribution of the dinucleotide CpG and cytosine methylation in the vitellogenin gene family. J Mol Evol 25: 107–115

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775–780

Crow JF, Denniston C (1985) Mutation in human populations. Adv Hum Genet 14:59–123

De Jong PJ, Grosovsky AJ, Glickman BW (1988) Spectrum of spontaneous mutation at the APRT locus of Chinese hamster ovary cells: an analysis at the DNA sequence level. Proc Natl Acad Sci USA 85:3499–3503

Drake JW (1970) The molecular basis of mutation. Holden Day, San Francisco

Duncan BK, Miller JH (1980) Mutagenic deamination of cytosine residues in DNA. Nature 287:560–561

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

Epstein CJ (1967) Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 215:355–359

Fersht AR (1979) Fidelity of replication of phage Phi-X174 DNA by DNA polymerase-III holoenzyme: spontaneous mutation by misincorporation. Proc Natl Acad Sci USA 76:4946–4950

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNAs. J Mol Evol 16:153–209

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196:261–282

Giroux CN, Mis JRA, Pierce MK, Kohalmi SE, Kunz BA (1988) DNA sequence analysis of spontaneous mutations in the SUP4-o gene of *Saccharomyces cerevisiae*. Mol Cell Biol 8: 978–981

Gojobori TZ, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18: 360–369

Golding GB, Glickman BW (1985) Sequence-directed mutagenesis: evidence from a phylogenetic history of human α-interferon genes. Proc Natl Acad Sci USA 82:8577–8581

Golding GB, Glickman BW (1986) Evidence for local DNA influences on patterns of substitution in the human α-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

Groot GSP, Kroon AM (1979) Mitochondrial DNA from various organisms does not contain internally methylated cytosine in −CCGG− sequences. Biochem Biophys Acta 564:355–357

Hsie AW, Recio L, Katz DS, Lee CQ, Wagner M, Shenley RL (1986) Evidence for reactive oxygen species inducing mutations in mammalian cells. Proc Natl Acad Sci USA 83:9616–9620

Hubrich-Kühner K, Buhk H-J, Wagner H, Kröger H, Simon D (1989) Non-CG recognition sequences of DNA cytosine-5-methyltransferase from rat liver. Biochem Biophys Res Commun 160:1175–1182

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115:605–610

Kunkel TA (1984) Mutational specificity of depurination. Proc Natl Acad Sci USA 81:1494–1498

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

Kunkel TA (1985b) The mutational specificity of DNA polyemerases −α and -γ during in vitro DNA synthesis. J Biol Chem 260:12866–12874

Kunkel TA (1988) Exonucleolytic proofreading. Cell 53:837–840

Kunkel TA, Alexander PS (1986) The base substitution fidelity of eukaryotic DNA polymerases. J Biol Chem 261:160–166

Kunkel TA, Bebenek K (1988) Recent studies of the fidelity of DNA synthesis. Biochem Biophys Acta 951:1–15

Kunkel TA, Soni A (1988) Mutagenesis by transient misalignment. J Biol Chem 263:14784–14789

Kunkel TA, Schaaper RM, Beckmann RA, Loeb LA (1981) On the fidelity of DNA replication. J Biol Chem 256:9883–9889

Kunkel TA, Schaaper RM, Loeb LA (1983) Depurination-induced infidelity of DNA synthesis with purified DNA replication proteins in vitro. Biochemistry 22:2378–2384

Kunkel TA, Sabatino RD, Bambara RA (1987) Exonucleolytic proofreading by calf thymus DNA polymerase δ. Proc Natl Acad Sci USA 84:4865–4869

Landis CA, Masters SB, Spada A, Pace AM, Bourne HR, Vallar L (1989) GTPase inhibiting mutations activate the α chain of Gs and stimulate adenylyl cyclase in human pituitary tumours. Nature 340:692–696

Li W-H, Wu C-I, Luo C-C (1984) Non-randomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Lindahl T (1982) DNA repair enzymes. Annu Rev Biochem 51:61–87

Lindahl T, Nyberg B (1972) Rate of depurination of native deoxyribonucleic acid. Biochemistry 11:3610–3618

Loeb LA (1985) Apurinic sites as mutagenic intermediates. Cell 40:483–484

Loeb LA, Kunkel TA (1982) Fidelity of DNA synthesis. Annu Rev Biochem 52:429–457

Loeb LA, Preston BD (1986) Mutagenesis by apurinic/apyrimidinic sites. Annu Rev Genet 20:201–230

Maeda N, Wu C-I, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1–20

Maruyama R, Gojobori T, Aota S, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 14 [Suppl]:r151–r197

McMahon JE, Tinoco I (1978) Sequences and efficiencies of proposed mRNA terminators. Nature 271:275–277

Mendelman LV, Boosalis MS, Petruska J, Goodman MF (1989) Nearest neighbour influences on DNA polymerase insertion fidelity. J Biol Chem 264:14415–14423

Meuth M (1989) The molecular basis of mutations induced by deoxynucleoside triphosphate pool imbalances in mammalian cells. Exp Cell Res 181:305–316

Modrich P (1987) DNA mismatch correction. Annu Rev Biochem 56:435–466

Nalbantoglu J, Goncalves O, Meuth M (1983) Structure of mutant alleles of the aprt locus in Chinese hamster ovary cells. J Mol Biol 167:575–594

Nalbantoglu J, Phear G, Meuth M (1987) DNA sequence analysis of spontaneous mutations at the aprt locus of hamster cells. Mol Cell Biol 7:1445–1449

Nussinov R (1981) Nearest neighbour nucleotide patterns; structural and biological implications. J Biol Chem 256:8458–8462

Ohno S (1988) Universal rule for coding sequence construction: TA/CG deficiency − TG/CT excess. Pro Natl Acad Sci USA 85:9630–9634

Pattinson JK, Millar DS, Grundy CB, Wieland K, Mibashan RS, Martinowitz U, McVey J, Tan-Un K, Vidaud M, Goossens M, Sampietro M, Krawczak M, Reiss J, Zoll B, Whitmore D, Bradshaw A, Wensley R, Ajani A, Mitchell V, Rizza C, Maia R, Winter P, Mayne EE, Schwartz M, Green PJ, Kakkar VV, Tuddenham EGD, Cooper DN (1990) The molecular genetic analysis of haemophilia A; a directed-search strategy for the detection of point mutations in the human factor VIII gene. Blood (in press)

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

Phear G, Nalbantoglu J, Meuth M (1987) Next-nucleotide effects in mutations driven by DNA precursor pool imbalances at the aprt locus of Chinese hamster ovary cells. Proc Natl Acad Sci USA 84:4450–4454

Phear G, Armstrong W, Meuth M (1989) Molecular basis of spontaneous mutation at the *aprt* locus of hamster cells. J Mol Biol 209:577–582

Reyland ME, Lehman IR, Loeb LA (1988) Specificity of proofreading by the 3′ → 5′ exonuclease of the DNA polymerase-primase of *Drosophila melanogaster*. J Biol Chem 263:6518–6524

Richter C, Park J-W, Ames BN (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. Proc Natl Acad Sci USA 85:6465–6467

Roberts JD, Kunkel TA (1986) Mutational specificity of animal cell DNA polymerases. Environ Mol Mutagen 8:769–789

Rodeghiero F, Castaman G, Dini E (1987) Epidemiological investigation of the prevalence of von Willebrand's disease. Blood 69:454–459

Sagher D, Strauss B (1983) Insertion of nucleotides opposite apurinic/apyrimidinic sites in deoxyribonucleic acid during in vitro synthesis: uniqueness of adenine nucleotides. Biochemistry 22:4518–4526

Salser W (1978) Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harb Symp Quant Biol 42:985–1002

Saluz HP, Jost JP (1989) A simple high-resolution procedure to study DNA methylation and in vivo DNA-protein interactions on a single-copy gene level in higher eukaryotes. Proc Natl Acad Sci USA 86:2602–2606

Savatier P, Trabuchet G, Fauré C, Cheblouré Y, Gouy M, Verdier G, Nigon VM (1985) Evolution of the primate beta-globin gene region. High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J Mol Biol 182:21–29

Schaaper RM, Kunkel TA, Loeb LA (1983) Infidelity of DNA synthesis associated with bypass of apurinic sites. Proc Natl Acad Sci USA 80:487–491

Setlow P (1976) Nearest neighbour frequencies in deoxyribonucleic acids. In: Fasman GD (ed) Handbook of biochemistry

and molecular biology, vol 2: Nucleic acids, 3rd edn. CRC Press, Cleveland, pp 312–318
Shearman CW, Loeb LA (1979) The effects of depurination on the fidelity of DNA synthesis. J Mol Biol 128:197–218
Stout JT, Caskey CT (1985) HPRT: gene structure, expression and mutation. Annu Rev Genet 19:127–148
Takeshita M, Chang C-N, Johnson F, Will S, Grollman A (1987) Oligodeoxynucleotides containing synthetic abasic sites. J Biol Chem 262:10171–10179
Thacker J (1985) The molecular nature of mutations in cultured mammalian cells: a review. Mutat Res 150:431–442
Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285–289
Vogel F (1972) Non-randomness of base replacement in point mutation. J Mol Evol 1:334–367
Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180
Vogel F, Motulsky AG (1986) Human genetics – problems and approaches, 2nd edn. Springer, Berlin Heidelberg New York
Vogel F, Kopun M, Rathenburg R (1976) Mutation and molecular evolution. In: Goodman M, Tashian RE, Tashian JH (eds) Molecular anthropology. Plenum Press, New York, pp 13–33
Woodcock DM, Crowther PJ, Diver WP (1987) The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. Biochem Biophys Res Commun 145:888–894
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
Youssoufian H, Antonarakis SE, Bell W, Griffin AM, Kazazian HH (1988) Nonsense and missense mutations in hemophilia A: estimate of the relative mutation rate at CG dinucleotides. Am J Hum Genet 42:718–725