J Mol Evol (1992) 34:189–200

Journal of
Molecular Evolution

The Influence of Nearest Neighbors on the Rate and Pattern of Spontaneous Point Mutations

R.D. Blake, Samuel T. Hess, and Janice Nicholson-Tuell

Department of Biochemistry, Microbiology and Molecular Biology, University of Maine, Orono, ME 04469, USA

Summary. The numbers and local sequence environments of the two types of substitution mutation plus additions and deletions have been obtained directly in this study from differences between a large number of extant primate gene and pseudogene sequences. A total of 3786 mutations were scored in regions where similarities between pseudogene and corresponding gene sequences is ≥85%, comprising ~30% of the pseudogene database of 80,584 bp. The pattern of mutations obtained in this fashion is almost identical to that obtained by Li et al. (1984) using a slightly different, more direct approach and with a smaller database. When mutations were scored, the neighbor pairs on the 5′ and 3′ sides were also noted, leading to a large 16 × 12 matrix of transitions and transversions. Biases of varying magnitude are found in the rates of substitution of the same base pair in different local sequence environments. The overall order for the effect of the 5′ neighbor on the rates of substitution mutation of a pyrimidine is $A > C \gg T > G$, and $G > A > T > C$ for the 3′ neighbor; where these results represent the average of substitution rates for the complement purine with complement neighbors of bases ordered above. The order for the 3′ neighbor is essentially the same for the two transitions and most of the four transversions as well; however, the order for the 5′ neighbor is more variable. The overall rate for the $C{\cdot}G \rightarrow T{\cdot}A$ transition is not unusual, however the presence of a 3′ neighboring G·C pair boosts the rate substantially, presumably due to specific cytosine methylation of the CG doublet in primate DNAs. The rate of the $T{\cdot}A \rightarrow C{\cdot}G$ transition is also well above average when the 3′ neighbor is an A·T, and to a lesser extent a G·C, pair. The latter bias is typical in that it reflects the association of alternating pyrimidine–purine sequences with increasing mutation rates. The substitution of the pyrimidine in a 5′purine–pyrimidine–purine3′ sequence generally occurs much faster than in a pyrimidine tract and points to the local conformation as a major determining factor of the substitution rate. An apparent inverse relationship is found between starting and product doublet frequencies of base pairs undergoing mutations with specific 3′ neighbors, indicating that differences in intrinsic substitution rates of base pairs with specific neighbors are a key factor in producing the familiar biases of nearest-neighbor frequencies.

Key words: Mutation pattern — Substitution mutations — Nearest-neighbor effects — Transitions — Transversions — CpG — Pseudogenes

Introduction

Spontaneous mutations in selected regions of the genome are a primary basis for the evolution of species. In spacer regions, nontranscribed repetitive sequences and pseudogenes, mutations have no apparent effect on the phenotype. The highest measured substitution rates [1.3 (±1.5) × 10^{-9} substitutions per site per year in primates] have been found in pseudogenes (Li 1983; Li et al. 1985; Koop et al. 1986), indicating a relaxation of functional constraints in these sequences. These sequences are therefore uniquely suited to the study of fundamental differences in rates of the different substitution

Offprint requests to: R. D. Blake

mutations. Gojobori et al. (1982) and Li et al. (1984) examined differences between extant gene and pseudogene sequences and reconstructed ancestral sequences to study the intrinsic rates of point mutations and demonstrated that the pattern of mutation is nonrandom. Presumably the biases they observed in the pattern reflect kinetic or thermodynamic differences in one or more of the different molecular events leading to fixation of mutations: formation of particular mispaired intermediates (Topal and Fresco 1976; Shibata et al. 1991), error detection, and mispair correction (Glickman et al. 1986; Jones et al. 1987; Modrich 1987; Mendelman et al. 1989) that occur during or shortly after replication. The question we raise in this study is how these differences may be influenced by neighboring bases; with the consequence that an underlying bias exists in the formation or repair of the same mispaired intermediate in different sequence environments.

Neighbor effects may bias the mutation process at any of the different stages leading to fixation of mutations in the germ line and have been demonstrated or implicated in a variety of seemingly relevant studies. Topal et al. (1980) concluded, on the basis of in vitro competition experiments with DNA polymerase, that nearest-neighbor interactions between incoming dNTPs and the terminal base of the nascent strand influence the frequency of substitution mutations. Variable frequencies of 2-aminopurine-induced transitions found by Koch (1971) suggest a dependence on sequence context. Mendelman et al. (1989) found that the kinetics of nucleotide misincorporation by DNA polymerase α is dependent on nearest-neighbor base stacking. The local sequence environment has also been shown to differentially influence the efficiency of repair of transitions and transversions (Jones et al. 1987; Lu and Chang 1988). The classic studies of Benzer (1961), showing that sites of the rIIa and rIIb genes of T_4 mutate with different efficiencies, could be seen as reflecting different neighbor effects at the several steps in the fixation of mutations. Benzer designated sites of high efficiency as hotspots. Hotspots have been identified and characterized in the system of genes for the metabolism of lactose in *Escherichia coli* (Miller and Low 1984; Cupples et al. 1990). An influence of the neighbor environment could also be partially responsible for the biases in neighbor frequencies in sequences (Josse et al. 1961). Whether such bias is a cause or an effect of one or more of these events remains to be demonstrated and is one of the incentives of this study. Accordingly, we have extended the investigation of the pattern of point mutations to include both a larger sampling of pseudogene sequences and a consideration of neighbors on both the 5′ and 3′ sides of bases that have changed.

Methods and Results

The approach taken in this study is different and in certain respects less sensitive than that taken previously by Li et al. (1984). Ancestral sequences were not reconstructed; rather, the pattern was obtained directly from differences between extant gene and pseudogene sequences. The apparent error that is generated by assigning all mutations to the pseudogene can be shown to be $\leq$20%. Li (1983) and Li et al. (1985) have determined the average substitution rates for codons with synonymous degeneracy and have found that the average rate for fourfold degenerate codons is roughly twice that for twofold degenerate sites; whereas twofold sites are, in turn, roughly twice that for nondegenerate sites. Assuming that the ratio of the substitution rate for a particular n-fold degenerate codon to that for the same triplet in the pseudogene, R_n, is 1.0 for sixfold sites, then we obtain $R_n = 0.16, 0.33, \ldots \times, 1.0$, when $n = 1, 2, \ldots \times, 6$. These interpolated values for R_n are in surprisingly good agreement with values reported by Li et al. from a limited analysis of twofold and fourfold sites. The fraction of substitutions occurring in coding regions, f_{cds}, can then be estimated from the relationship

$$f_{cds} = \Sigma R_n f_n / 3$$

where f_n is the fraction of coding regions made up of n-fold degenerate codons. The latter can be estimated with reasonably close accuracy because the frequencies of amino acids in all globular proteins are relatively invariant (Blake et al. 1986). From these frequencies, we obtain a value for f_{cds} of ~20%, so that the estimated fraction of substitution mutations actually occurring in pseudogenes, $1 - f_{cds}$, is ~80%. With an apparent error of this magnitude, this approach has the disadvantage that the effects of mutations occurring in highly biased frequencies may be slightly modulated; however, it has the far greater advantage that it readily allows for the acquisition of larger databases, thereby improving the significance of mutations that are sparsely populated in certain neighbor environments.

The further assumption of this study is that addition and insertion mutations can be assigned exclusively to the pseudogene.

The first goal of this study was to establish the optimum alignments of genes with their corresponding pseudogene sequences. Sixty-five primate sequences totaling 80,584 bp and designated in original reports as pseudogenes, plus 70 corresponding primate gene sequences totaling 73,319 bp were obtained from the GENBANK (Release 66) and EMBL (Release 21) databases. Optimized Needleman–Wunsch (1970) type alignments of pseudogene with

```
HUMACBPA        687- CCTTCTACAACGAGCTGTGTGTGGCTG-CCAAGGAGCACCCCATGCTGCT
  (pseudogene)         * * * *** **  * * *  **  *  *** ** ***  **    *
HUMACCYBA      2003- CCGTGTTCTTTGCACT-T-TCTGCATGTCCCCCG-TCTGGCCTGGCTG-T
  (gene)             |←      (window)    →|

HUMACBPA        736- GACCA-AGG-TCCCCCTGAGCCCCAAGGCCAACCACAAGAAGATGACCCA
                     **   **  * *    * * ***  *  *** * * ** **** * * *
HUMACCYBA      2049- CCCCAGTGGCTTCCCCAGTGTGACATGGTGCATCTC-TG-C-CTTA-C-A

HUMACBPA        784- GATCATGTTGGAGATCTTCGACAGGCCAGCCATGTACGTGGCCATCCAGG
                              *    *    *   **              *  *
HUMACCYBA      2094- GATCATGTTTGAGACCTTCAACACCCCAGCCATGTACGTTGCTATCCAGG
                     |→ start of exon #4 of β-actin gene

HUMACBPA        834- CCGTGCTGTCCCTGTACACCTCTGGCC-TACCACTGACATCGTGATGGAC
                      *     *         *         *        *
HUMACCYBA      2144- CTGTGCTATCCCTGTACGCCTCTGGCCGTACCACTGGCATCGTGATGGAC

HUMACBPA        883- TACGATGACGGGGTCACCCACACTGTGCCCATCTATGAAGAGTATGCCCT
                      *  *                              *  * *
HUMACCYBA      2194- TCCGGTGACGGGGTCACCCACACTGTGCCCATCTACGAGGGGTATGCCCT

HUMACBPA        933- CCCCCATGCCATCCTGCGTGTGTGCCTGGCTGGTCAGGACCTGACTGACT
                                        *  **         * *
HUMACCYBA      2244- CCCCCATGCCATCCTGCGTCTGGACCTGGCTGGCCCGGACCTGACTGACT

 Summary: The overall fraction of identical residues in this alignment is 227/300(75.7%).
```

Fig. 1. Alignment with the algorithm FASTA (Wilbur and Lipman 1983; Lipman and Pearson 1985) of a 300-residue section of the human cytoplasmic β-actin pseudogene [GENBANK entry name: HUMACBP(AI)] with a section of the human cytoplasmic β-actin gene (HUMACCYBA). Residue 2094 of the β-actin gene begins exon 4. The section delineated by the arrows and bars indicates the window size used for the identification of regions of high sequence homology, as well as for the positions of mutations (denoted by asterisks) and for identification of the neighboring bases on either side of the mutation.

gene sequences were produced with the FASTA algorithm of Wilbur and Lipman (1983) and Lipman and Pearson (1985). An example of the alignment of a 300-bp section of the human cytoplasmic β-actin pseudogene sequence with the corresponding gene sequence is shown in Fig. 1. The overall similarity between gene and pseudogene sequences shown in Fig. 1 is 76%. Alignments that did not lead to similarities greater than 60% were eliminated from further consideration. A number of pseudogenes exhibited similarities to their designated gene sequences in the neighborhood of 50%. This is the level expected for two unrelated sequences when the inherent nearest-neighbor constraints exhibited by all natural sequences are taken into account. Random sequences of 50% (G+C) content are expected to show a 25% match at individual sites because the probability of occurrence of any one of the four bases is 0.25. However, the level of randomness in all natural sequences is limited by the constraints imposed by an underlying, uniform nearest-neighbor bias. As shown by Fitch (1983), a pair of unrelated natural sequences with the same neighbor bias is expected to exhibit a level of similarity of approximately 50%. Indeed, results of pairwise analyses with the FASTA algorithm on 50 randomly selected, unrelated primate sequences gave an average level of similarity of precisely 50%, with a standard deviation of only 3%.

Aligned pseudogene–gene sequences with similarities >60% were analyzed for the identification and isolation of regions where the level of similarity ≥85%. This was achieved by passing the aligned pair across a 20 residue-wide window and capturing those regions with three or fewer mismatches. The example in Fig. 1 shows that similarities between a β-actin pseudogene and gene are only 50% over the first two lines of 100 residues, but then jump to 90% beginning with the third line (position number 784 in the original pseudogene sequence) and continue thereafter at that high level. Mutations were scored only in the latter regions. The higher level of similarity is almost invariably associated with coding regions in gene sequences, and the example of Fig. 1 is typical. The region of 90% similarity beginning at residue 2094 corresponds precisely with the start of exon 4 of the β-actin gene sequence. It was found that an 85% or better level of similarity could be found in 24,970 bp out of 80,584 bp in the pseudogene database, when aligned with 38,063 bp out of 73,319 bp of gene sequences.

The level of gene–pseudogene similarity that was chosen for scoring mutations was selected to ensure that sequence relationships were orthologous and to

Table 1. Nucleotide substitutions in primate pseudogenes[a]

	To							Number of residues[c]	
From	A	T	C	G	Deletions	Insertions	Totals[b]	Genes	Pseudogenes
A	–	127	142	476	69	60	874	8019	5528
T	127	–	413	117	74	49	780	8117	5410
C	132	598	–	192	78	61	1061	11,368	7214
G	615	131	173	–	87	65	1071	10,558	6817
Totals	874	856	728	785	308	235	3786	38,062	24,969
Transitions = 2102 (64.8%) Transversions = 1141 (35.3%)							F_{GC} =	0.576	0.562

[a] Average standard deviations: ±28 for transitions; ±22 for transversions; ±10 for both deletions and insertions

[b] Total of changes involving the base in each row, including deletions and additions

[c] Total number of this base in gene and pseudogene sequences over the region providing acceptable homology (see text)

maximize the number of mutations in pseudogenes with a minimum ambiguity of origin. The large database of mutations needed to establish a significant next-neighbor pattern is more readily obtained in regions of fewer similarities; however, an increase in the frequency of multiple and ambiguous mutations, attributable to the time period from the gene duplication event and the origin of each pseudogene, can be expected to increase with alignments of lower similarity. So also will the potential for ambiguity in the origin of mutations increase due to gene conversion and recombinational rearrangements. The rationale for examining regions of higher similarity, therefore, is to minimize contributions from multiple changes at the same site and mutations occurring in the gene sequence. Single-site multiple mutations are nil when the similarity level is greater than 85% (Jukes and Cantor 1969; Brown et al. 1982; Gojobori et al. 1982). It was found that the pattern obtained from scoring mutations in pseudogenes with an 80% level of similarity is essentially identical to that obtained with a 90% level; although the database is much smaller in the latter case. A difference in the pattern seems to emerge when the level of similarity falls below about 80%, which is probably a consequence of some of the factors discussed above. The 85% level therefore represents an attempt to simultaneously maximize the size of the database and minimize the level of ambiguity regarding the origin of the pattern.

Table 1 shows the pattern of mutations found in 24,969 bp of pseudogene sequences exhibiting ≥85% similarity to 38,062 bp of gene sequences. The total number of transitions and transversions listed in columns 2–5 of Table 1 is about 10-fold greater than examined earlier by Li et al. (1984). Random errors due to statistical fluctuations in the numbers of each type of mutation in our finite collection of pseudogenes were determined from the population standard deviation for changes in the mean relative count of each type of mutation as the analysis proceeded with the addition of results from each new alignment. Average standard deviations determined this way were ±28 for each type of transition, ±22 for transversions, and ±11 for both additions and deletions. Transitions are along the diagonal from lower left to upper right and total 2102. There were only 1141 transversions, or slightly more than half the number of transitions, even though there were twice the number of possibilities. The ratio of transitions to transversions is 1.84, which is similar to that obtained by Li et al. (1984) [1.45 (59.2%/40.8%)]; indeed, the entire pattern is very similar to that obtained by Li et al. who assigned mutations to both gene and pseudogene sequences on the basis of reconstructed ancestral sequences. The relative fraction of deletions and insertions, given in columns 6 and 7 of Table 1, are also about the same as found by Li et al. (1984).

The total numbers of all four types of mutation of each of the four bases are indicated in column 8, followed by the numbers of each base in gene and pseudogene regions containing the mutation (columns 9 and 10). It is possible, therefore, to calculate the difference between observed frequencies of mutations and the frequencies expected if mutations were to occur randomly with respect to the presence of each of the four bases. Results are shown in Table 2. Differences in this table indicate that variations from expected are generally small. Due to random errors in the numbers of each type of mutation in Table 1, the probable error (Bevington 1969) in these values for the percent difference is ±10%. The difference for transitions is calculated from the relationship: $100[(f_{X\rightarrow S}/f_X) - 1]$, where $f_{X\rightarrow S}$ is the observed fraction of transition X → S, and f_X is the frequency of base X in gene sequences. The difference for transversions is calculated from: $100[(2f_{X\rightarrow V}/f_X) - 1]$, where $f_{X\rightarrow V}$ is the observed fraction of transversion X → V.

Li et al. (1984) made the observation that C → T and G → A transitions exceed the numbers of A → G and T → C and therefore are primarily responsible for a decrease in the fraction (G+C) com-

Table 2. Variations (percent) in the numbers of observed frequencies of mutations from expected

From	To A	To T	To C	To G	Deletions	Insertions	Totals
A	—	+5.7	+18.1	+7.5	+6.3	+21.2	+9.6
T	+4.4	—	−7.9	−3.8	+12.7	−2.2	−3.4
C	−22.5	−4.7	—	+12.7	−15.2	−13.1	−6.2
G	+5.5	−17.2	+9.3	—	+1.8	−2.9	+2.0
Overall	+7.8	+5.6	−10.2	−3.2			

position, F_{GC}, of pseudogenes. Differences from the random probability that changes will involve a particular base indicate that the net effect of these transitions on base composition is probably real, but small. The F_{GC} in scored regions of pseudogenes does indeed decrease slightly, from 0.576 to 0.562 (Table 1). It is probably more accurate to ascribe this decrease to an overall positive bias (+13.4%) in the frequency of both transitions and transversions leading to small net increases in As and Ts, and a corresponding negative bias leading to net decreases in Gs and Cs in pseudogenes. Other small biases may be found in these data, e.g., mutations involving changes to the base adenine appear to occur more frequently than expected.

When mutations were scored, the bases on either side were also noted, leading to a large 16 × 12 matrix for the numbers of transitions and transversions in specific triplet environments (Table 3). The total number of different transitions and transversions in a triplet environment is therefore 192, or 16 times the number considered when only the affected base is taken into account. Therefore, the statistical significance of differences is substantially lower than it is for the isolated pair (Table 1). From trends in the relative count of each of the 192 neighbor-dependent substitution mutations that were scored during the analysis, we estimate the average random error to be ±10 for transitions and ±7 for transversions. Despite such large statistical fluctuations, some biases in Table 3 stand out. For instance, it is found that 34.4% of all transitions and transversions involving all four bases occur when the 3′ neighboring base is a G (cf. underlined column totals at the bottom of Table 3); whereas the next highest frequency, 23.2%, occurs when the 3′ base is a C. On the 5′ side, most mutations (33.6%) occur when the neighbor is a C; whereas the smallest numbers (19.4%) occur when the neighboring base at both the 3′ and 5′ positions is a T. When the fractional presence of each base in gene sequences is taken into account (cf. Table 1), the overall order for the effect of the 5′ neighbor on the rates of substitution mutations is seen to be A > C ≫ T > G, and G > A > T > C for the 3′ neighbor.

Similarly, the percent difference between observed frequencies of mutations and those expected were calculated for each triplet environment of the mutations tabulated in Table 3. The difference was calculated from the frequency of occurrence of those triplets in gene sequences containing the base undergoing mutation. The fractions of all 64 triplets in regions of gene sequences in which mutations were scored were determined and are given in Table 4. There are substantial deviations in the observed numbers of many triplets from the numbers that would be expected from the frequencies of the four bases. For example, the triplet 5′-ACG-3′ and its complement CGT occur with less than half the frequency of the average triplet, and CTG and CAG more than twice the average. These variations are unusual even when corrected for the percent that each base occurs in gene sequences, which in this case are −60% and +54%, respectively. The low frequencies of ACG and CGT are mainly due to the ubiquitous underrepresentation of CG doublets in all metazoan organisms (Josse et al. 1961; Russell et al. 1973; Setlow 1976; Razin and Riggs 1980; Hinds and Blake 1984; Bird 1986; Ehrlich et al. 1990). Collectively, all eight CG-containing triplets differ by −51% in gene sequences (and in pseudogene sequences of recent origin as well). Although the low frequency of CG-containing triplets is unusual, other equally widespread triplet biases exist and, taken together, seem to point to a dependence of point mutations on the triplet environment as a possible contributing factor to the occurrence of doublet and triplet biases (Josse et al. 1961). Interestingly, the variances from expected of all 64 triplet frequencies show a high correlation between gene and pseudogene sequences over all regions, which is probably due to the relatively short period (~5 million years) of constraintless existence of many identifiable pseudogene sequences (Li et al. 1985) and to the high level of stringency (≥90%) in aligned regions used for scoring mutations.

Differences were calculated between the observed frequencies of mutations and frequencies expected if mutations were to occur randomly with respect to the triplet environment, and denoted in Tables 5 and 6 as the corrected Δf (in percent). The numbers of neighbor-dependent mutations scored are small so that the random error is high; therefore the assigned mutation and its complement were com-

Table 3. Nearest neighbors of bases undergoing mutation

Mutation of base denoted by *	Neighbors, 5′X–*–Y3′								
	A*A	A*T	A*C	A*G	T*A	T*T	T*C	T*G	C*A
A → G	**41**	**29**	**12**	**46**	**14**	**16**	**18**	**27**	**26**
A → C	11	2	10	10	3	12	1	5	9
A → T	7	5	16	8	0	6	4	6	11
T → C	**14**	**21**	**21**	**39**	**18**	**11**	**14**	**26**	**34**
T → A	5	5	6	10	2	9	4	4	5
T → G	4	3	3	11	6	7	3	12	4
C → T	**44**	**43**	**19**	**40**	**37**	**19**	**18**	**42**	**43**
C → A	10	0	11	3	9	4	11	4	10
C → G	12	9	24	13	23	8	15	7	18
G → A	**18**	**23**	**28**	**58**	**29**	**23**	**24**	**26**	**61**
G → C	7	12	10	9	18	23	10	6	7
G → T	2	9	9	24	5	6	6	8	6
Totals	175	161	169	271	164	144	128	173	234

Results in the first row of each of the four groups (boldface) denote transitions; underlined totals indicate the largest number of mutations within a triplet with the same 3′ neighbor

Table 4. Frequency ($\times 10^{-2}$) of occurrence of triplets in gene sequences

Base denoted by *	Neighbors, 5′X–*–Y3′								
	A*A	A*T	A*C	A*G	T*A	T*T	T*C	T*G	C*A
A	1.42	0.97	1.07	1.74	0.78	0.60	0.63	0.64	1.56
T	0.55	0.86	1.08	1.21	0.67	1.43	1.55	1.15	0.84
C	1.34	1.33	1.89	0.72	1.60	1.60	2.50	0.71	2.55
G	1.69	1.12	1.74	2.34	1.54	1.46	1.90	2.58	0.73
Sum	5.00	4.28	5.77	6.02	4.59	5.09	6.57	5.08	5.68
Δf, %	−5	−19	−10	−1	−13	−5	+3	−17	−11

bined. In any case, assignment of the strand responsible for initiating the mutation is not ordinarily possible (Wu and Maeda 1987). From the average random error in mutation counts, we estimate that deviations in the percent difference values in Tables 5 and 6 to be ±20% and ±25%, respectively. The corrected Δf_s's for transitions are summarized in Table 5, listed in rank order by the magnitude of the difference. The significance of differences within several positions of the rank order are meaningless due to statistical fluctuations in the count. Beyond this uncertainty, however, several strong biases are evident, e.g., the top four transitions, represented as 5′XCG3′(5′CGX3′) → XTG(CAX), occur as a group 184% more rapidly than expected (where X is any base, and where the bias for complement transitions in this triplet environment, C → T and G → A, have been averaged). This is an extraordinarily high rate of mutation and also reflects a relaxation of substantial constraints seen in the occurrence of the CG doublet in gene sequences (e.g., Razin and Riggs 1980; Bird 1986; Ehrlich et al. 1990).

There are other biases not involving CG. The transition XTA(TAX) → XCA(TGX) occurs as a group 47% more rapidly than expected, wheras the four XCC(GGX) → XTC(GAX) transitions occur 64% slower than expected. In summary (Table 7), the dependence of the rate of mutation on the 5′ neighboring base to the pyrimidine of the transition pair occurs in the following order: A > G > C > T; whereas the order for the 3′ neighboring base to the pyrimidine is G > A > T ≫ C. This same approximate order of neighboring base also applies to subgroups where one of the neighbors, either 3′ or 5′, is held constant. These relationships are also similar to those found above for substitution mutations overall, i.e., for both transitions and transversions. They indicate that the fastest rate for a transition is most likely to be found in the sequence context AYG, where Y is the pyrimidine, and the slowest rate is expected for TYC (Table 7). These results also indicate that the fastest transition rates are expected in an alternating purine–pyrimidine sequence, whereas the slowest are expected in polypurine-polypyrimidine stretches. The frequencies

Table 3. Extended

Neighbors, 5′X–*–Y3′								
C*T	C*C	C*G	G*A	G*T	G*C	G*G	Totals	Average
30	**73**	**60**	**26**	**15**	**17**	**26**	**476**	**30**
10	8	13	7	7	7	27	142	9
3	7	25	3	4	6	16	127	8
37	**21**	**44**	**22**	**14**	**27**	**50**	**413**	**26**
9	13	20	2	13	2	18	127	8
10	12	19	7	4	3	9	117	7
30	**29**	**73**	**36**	**20**	**43**	**62**	**598**	**37**
10	12	16	1	11	9	11	132	8
0	12	13	4	11	18	5	192	12
47	**58**	**102**	**21**	**26**	**46**	**25**	**615**	**38**
4	7	8	18	18	10	6	173	11
0	3	19	10	6	12	6	131	8
190	255	412	157	149	200	261	3243	203

Table 4. Extended

Neighbors, 5′X–*–Y3′								
C*T	C*C	C*G	G*A	G*T	G*C	G*G	Totals	Average
1.29	2.09	2.36	1.45	0.82	1.49	2.15	0.2107	1.316
0.56	2.43	3.03	0.60	0.95	1.35	2.09	0.2133	1.332
2.87	3.54	1.55	1.82	2.06	2.54	1.23	0.2987	1.868
0.70	1.47	1.30	1.94	1.71	2.58	2.94	0.2774	1.734
6.42	9.52	8.25	5.80	5.54	8.01	8.40	(38,063)	6.250
0	+27	+14	−5	−10	+11	+21		

of mutations occurring in the middle of a run of three pyrimidines (or purines), denoted by asterisks in Table 5, are substantially below expectation (−29% as a group); the lowest of any grouped set.

The results of neighbor effects on transversions are summarized in abbreviated form in Table 6. They are more tentative because of the low numbers that were scored. We note that transversions involving C with a neighboring 3′G generally occur with unusually high frequencies; +44% as a group. This bias again reflects the extraordinarily high rate of mutation of Cs that have a 5′G neighbor; and the associated relaxation of cellular constraints on the CG doublet. The next highest frequency of transversions involving C in a grouped set is only +10%, and that is when the 3′ neighbor is an A. A very strong positive bias is also found for the transversion XTA(TAX) → XGA(TCX), which occurs 36% more often than expected, whereas XTC(GAX) → XAC(GTX) occurs 24% less often. As seen above, the T·A bp in the context XTA(TAX) also exhibits an unusually high rate of transition, suggesting that, like C followed by G, a T with 5′ neighboring A undergoes more rapid substitution and represents a mutational hotspot.

Discussion

The nonrandom pattern of 3786 substitution mutations scored in 65 primate pseudogenes is essentially the same as that found by Li et al. (1984) working with a 10-fold smaller database. The principal conclusions regarding the pattern of point substitutions established in the earlier work should therefore be consulted. Explanations for many of the key features of this pattern remain speculative, e.g., for the observations that 1) the numbers of transitions are approximately twice the numbers of transversions, that is, the average rate of a transition is four times greater than the average rate of a transversion, and 2) the rates of both transitions and transversions collectively lead to a gradual increase in the numbers of As and Ts or decrease in the (G+C) content of pseudogenes over time. The net

Table 5. Effect of nearest neighbors on $C\cdot G \rightarrow T\cdot A$ and $T\cdot A \rightarrow C\cdot G$ transitions

Rank order[a]	Transition[b]	Number scored	Δf_s[c], %	Corrected[d,e] Δf_s, %
1	TCG(CGA) → TTG(CAA)	103	+57	+236 (+95)
2	CCG(CGG) → CTG(CAG)	175	+166	+194 (+74)
3	ACG(CGT) → ATG(CAT)	87	+32	+190 (+69)
4	GCG(CGC) → GTG(CAC)	120	+83	+115 (+73)
5	CTA(TAG) → CCA(TGG)	61	−7	+99 (−7)
6	GTA(TAC) → GCA(TGC)	40	−39	+58 (−8)
7	GTG(CAC) → GCG(CGC)	123	+87	+40 (+73)
8	ATG(CAT) → ACG(CGT)	69	+5	+35 (+69)
9	ATT(AAT) → ACT(AGT)	50	−24	+29 (+30)
10	ACT(AGT) → ATT(AAT)	66	0	+28 (+30)
11	ATA(TAT) → ACA(TGT)	30	−54	+25 (+14)
12	*CTT(AAG) → CCT(AGG)	83	+26	+20 (−5)
13	*CCT(AGG) → CTT(AAG)	88	+34	+18 (−5)
14	ACA(TGT) → ATA(TAT)	67	+2	+18 (+14)
15	GTT(AAC) → GCT(AGC)	26	−60	+13 (−39)
16	TTA(TAA) → TCA(TGA)	32	−51	+7 (0)
17	TCA(TGA) → TTA(TAA)	66	0	+1 (0)
18	TTG(CAA) → TCG(CGA)	52	−21	−5 (+95)
19	CTG(CAG) → CCG(CGG)	104	+58	−7 (+74)
20	ATC(GAT) → ACC(GGT)	36	−45	−10 (−34)
21	*TTT(AAA) → TCT(AGA)	52	−21	−17 (−29)
22	GCC(GGC) → GTC(GAC)	89	+36	−19 (−20)
23	GCA(TGC) → GTA(TAC)	60	−9	−22 (−8)
24	GTC(GAC) → GCC(GGC)	44	−33	−23 (−20)
25	CCA(TGG) → CTA(TAG)	69	+5	−36 (−7)
26	*TTC(GAA) → TCC(GGA)	40	−39	−37 (−50)
27	GCT(AGC) → GTT(AAC)	48	−27	−40 (−39)
28	ACC(GGT) → ATC(GAT)	45	−32	−41 (−34)
29	*TCT(AGA) → TTT(AAA)	37	−44	−47 (−29)
30	*CTC(GAG) → CCC(GGG)	47	−28	−51 (−53)
31	*TCC(GGA) → TTC(GAA)	39	−41	−58 (−50)
32	*CCC(GGG) → CTC(GAG)	54	−18	−61 (−53)
	Average	66		

[a] Transitions are arranged in rank order according to the magnitude of corrected Δf_s in column 5. The significance of the order is only good to ±20% in the corrected Δf_s of column 5

[b] An asterisk denotes a tripyrimidine tripurine sequence

[c] Δf_s denotes the difference (percent) between observed and expected frequencies of transitions: 100[(number scored/66) − 1]

[d] Corrected Δf_s denotes the difference between the observed frequencies of transitions and those expected were transitions to occur randomly with respect to the observed frequencies of the affected triplets in gene sequences

[e] The numbers in parentheses represent average values (percent) for corrected Δf_s for the forward plus reverse transition

effect of all differences in substitution rates in pseudogene sequences in primates is to increase the fraction of each base according to the order A > T > G > C and thereby to decrease the (G+C) composition in these regions. There is nothing substantive in the experimental literature that would provide a molecular explanation for these differences. The bias in transition and transversion rates apparently reflects the requirement of spontaneous transversional mutations for the occurrence of two rare tautomeric and rotameric states of the bases to form a purine–purine mispair that manages to avoid detection by repair systems, whereas transitions require only one (Topal and Fresco 1976; Kennard 1985; Carbonnaux et al. 1990).

Biases of varying magnitude are found in the rates of substitution of the same base pair in different local sequence environments. The wide range of differences between observed and expected rates of mutation in triplet environments is capable of accounting for hotspots (e.g., Benzer 1961; Miller and Low 1984; Cupples et al. 1990) and seems also to account for some of the familiar differences in nearest-neighbor frequencies (Josse et al. 1961; Russell et al. 1973; Setlow 1976; Hinds and Blake 1984). The largest bias is seen in the substitution of the C·G bp. As can be seen from Table 1, the overall rate of mutation of the C·G pair is average, i.e., almost precisely what is expected from random occurrence. However, a C·G pair with a 3′G·C neighbor, 5′CG3′(5′CG3′), undergoes transitions, and to a lesser extent transversions, far more rapidly than

Table 6. Effect of nearest neighbors on transversions

Rank order[a]	Transversion	Number scored	Δf_v[b], %	Corrected[c,d] Δf_v, %
1	CCG(CGG) → CAG(CTG)	21	+18	+172 (+129)
2	AAC(GTT) → ATC(GAT)	29	+63	+140 (+87)
3	ATA(GTT) → AGA(TCT)	16	−10	+114 (+65)
4	ACA(TGT) → AGA(TCT)	35	+96	+96 (+26)
5	ACG(CGT) → AGG(CCT)	17	−5	+86 (−21)
6	ACC(GGT) → AGC(GCT)	42	+136	+84 (+69)
7	ATA(TAT) → AAA(TTT)	11	−38	+60 (+44)
8	TCG(CGA) → TGG(CCA)	14	−21	+54 (+12)
9	TCG(CGA) → TAG(CTA)	10	−44	+51 (+30)
10	AGG(CCT) → ATG(CAT)	34	+91	+48 (+39)
11	CGC(GCG) → CTC(GAG)	14	−21	+46 (+61)
12	ATG(CAT) → AGG(CCT)	21	+18	+41 (+39)
13	CTC(GAG) → CGC(GCG)	39	+119	+39 (+61)
14	TTG(CAA) → TGG(CCA)	21	+18	+38 (−3)
15	TAG(CTA) → TTG(CAA)	11	−38	+28 (+24)
16	ACA(TGT) → AAA(TTT)	16	−10	+28 (+24)
:				
:				
49	AGA(TCT) → ACA(TGT)	15	−16	−28 (+26)
50	CGC(GCG) → CCC(GGG)	12	−33	−30 (−28)
51	CAC(GTG) → CCC(GGG)	17	−5	−32 (−31)
52	CCC(GGG) → CAC(GTG)	18	0	−40 (−31)
53	TGC(GCA) → TCC(GGA)	14	−21	−41 (+14)
54	GAC(GTC) → GCC(GGC)	10	−44	−45 (−22)
55	TTC(GAA) → TGC(GCA)	10	−44	−47 (−51)
56	ACG(CGT) → AAG(CTT)	3	−83	−53 (−9)
57	AAT(ATT) → ACT(AGT)	5	−72	−53 (−34)
58	GAC(GTC) → GTC(GAC)	8	−55	−54 (−73)
59	CCC(GGG) → CGC(GCG)	18	0	−57 (−28)
60	AGA(TCT) → ATA(TAT)	6	−66	−59 (+65)
61	TGC(GCA) → TTC(GAA)	7	−61	−61 (−51)
62	TTC(GAA) → TAC(GTA)	7	−61	−62 (−42)
63	AGG(CCT) → ACG(CGT)	9	−50	−69 (−21)
64	TAA(TTA) → TTA(TAA)	2	−89	−75 (−85)
	Average	18		

[a] Transversions are arranged in rank order according to the magnitude of corrected Δf_v in column 5. The significance of the order is only good to ±25% in the corrected Δf_s of column 5

[b] Δf_v denotes the difference (percent) between expected and observed frequencies of each transversion

[c] Corrected Δf_v denotes the difference between observed frequencies of transversions and those expected were transversions to occur randomly with respect to the observed frequencies of the affected triplets in gene sequences

[d] The numbers in parentheses represent average values (percent) for corrected Δf_v for the forward plus reverse transversion

it does when next to any of the other three neighbor pairs. The reason for this high rate of change does not result from any intrinsic lability of the C·G pair. The rate constant for spontaneous hydrolytic deamination of cytosine residues in double-stranded DNA is very low (7×10^{-13} s^{-1} at 37°C) with a half-life of ~30,000 years (Frederico et al. 1990). Moreover, deamination yields uracil, which is recognized as a lesion by the postreplication repair enzyme uracil-DNA glycosylase (Lindahl 1982). The large rate difference arises when cytosine residues 5′ to neighboring guanines are singled out for methylation, which happens mainly, but not exclusively, in regions of the genome that are inactive in expression (Ehrlich et al. 1990). 5-Methyl-cytosine is significantly more susceptible to spontaneous deamination, and, moreover, the reaction leads to the formation of thymine, a normal residue in DNA that confounds the repair process. It is undoubtedly for this reason that the CG doublet generally occurs in low frequencies in metazoan DNAs (Josse et al. 1961; Russell et al. 1973; Setlow 1976; Razin and Riggs 1980; Hinds and Blake 1984; Bird 1986; Ehrlich et al. 1990). Its global occurrence in primate DNAs is approximately 70% below that expected from the frequencies of C and G in DNA, whereas in *E. coli* DNA, which is not subject to the same pattern of methylation, the doublet GC is present in frequencies well above that expected (Hinds and Blake 1984). Given the high rate of C·G → T·A transitions in

Table 7. Summary of nearest-neighbor effects on the rates of substitution mutations[a]

	5′ neighbor	3′ neighbor
Transitions		
C·G → T·A	A > T > C > G	G > A > T > C
T·A → C·G	G > A > C > T	A > G > T > C
Overall	A > G > C > T	G > A > T ≫ C
Sequence context for the fastest rate:[b] 5′AYG3^		
Sequence context for the slowest rate: TYC		
Transversions		
C·G → G·C	A > T > G > C	A > G > C > T
C·G → A·T	C > T > G > A	G > T > C > A
T·A → A·T	G > C > A ≫ T	T > G > A ≫ C
T·A → G·C	A > C > T > G	A > G > T > C
Overall	C > T > A > G	G > A > T > C
Sequence context for the fastest rate:[b] 5′CYG3′		
Sequence context for the slowest rate: GYC		
Transitions plus transversions		
	A > C ≫ T > G	G > A > T > C
Sequence context for the fastest rate:[b] 5′AYG3′		
Sequence context for the slowest rate: GYC		

[a] Neighbors are in reference to the pyrimidine undergoing substitution

[b] Y denotes the pyrimidine of the pair undergoing mutation

the CG(CG) doublet sequence of metazoan DNAs, the expectation is that the products of the transition will be present in higher frequencies than expected, and, indeed, they are. The frequencies of the doublet TG and its complement CA in primate DNAs are the highest among all doublets (+24% greater than expected).

Transversions of the C·G pair adjacent to a 3′G·C pair also occur at high rates. However, the explanation above for transitions is less appropriate for transversions of the C·G pair. Rather, enhanced transversion may be the consequence of spontaneous alkylation of guanine at the 0–6 position, which leads, in turn, to enhancement of the rate of deamination of 5-methyl-cytosine (Fix et al. 1990). A DNA glycosylase activity appears to exist in primate cells with specificity for removing mispaired thymine residues (Wiebauer and Jiricny 1990); however, concomitant removal of the altered G may lead to error-prone repair, resulting in a high frequency of transversions.

A similar, but somewhat more subdued hotspot was found for the transition T·A → C·G and to a lesser extent for transversions of the T·A bp in the context XTA(TAX). The molecular basis for this bias is distinct from that discussed above in connection with the high rate of C·G transition, in that this transition seems to reflect a genuine lability of the T·A pair. The transition of T·A is enhanced significantly when the 3′ neighbor is an A·T, with the consequence that the presence of the TA doublet in human sequences is 30% below that expected. It is 25% below the expected in *E. coli* DNA as well; although the doublet products of the transition, CA(TG), and the same doublet products of the transition of C·G in CG(CG) discussed above are present in well above the expected frequency in both human and *E. coli* DNAs. A number of plausible explanations can be invoked in which either thymine and/or adenine instigates the change. T_{enol}·G, T·G wobble, and C_{imino}·A pairs have been proposed (Topal and Fresco 1976; Hunter et al. 1986), and there are indications that such pairs may be influenced by the local sequence context (Topal et al. 1980; Fersht et al. 1982). Also, spontaneous deamination of the adenine residue may be a factor, as the hydrogen donor and acceptor groups of adenine are switched in hypoxanthine to those of guanine.

The rate of transition of T·A is also enhanced by the presence of a 3′G·C pair, whereas neighboring 3′T·A and particularly C·G pairs retard the rate. The association of alternating pyrimidine–purine runs or tracts with enhanced mutation rates seems to be a general relationship (cf. Table 7). The effect of the 5′ neighbor is ambiguous in the two cases discussed above; however, from the overall effects of 3′ and 5′ neighbors on all substitution mutations, it would appear that the presence of a purine, A or G, as both 5′ and 3′ neighbor to the central pyrimidine undergoing the mutation is associated with an enhanced rate of mutation (Table 7). Neighboring pyrimidine residues generally have the opposite effect. This seems to point to the local conformation as the primary basis for variations in mutation frequencies. Structural studies of selected pyrimidine–purine tracts indicate that they may assume slight variants of the canonical B-structure that describes DNA with quasi-random sequences of the bases. Tracts of dA·dT have an unusually high propeller twist (25°) that makes possible a system of bifurcated hydrogen bonds, and additional water bridges (Coll et al. 1987; Nelson et al. 1987). Tracts are more apt to adopt altered global conformations because the steric conflicts between neighbors are uniform (Calladine and Drew 1986; Shakked and Rabinovich 1986; Tung and Harvey 1986). Thus, it is possible to imagine these structural features as reducing the accessibility and reactivity of the base pairs from interactions that initiate mutations and in such fashion as might extend to next neighbors and even beyond.

Acknowledgments. This work was conducted with funds granted by NIH (GM22827) and by MAES (project no. 08405), paper number 1610.

References

Benzer S (1961) On the topography of the genetic fine structure. Proc Natl Acad Sci USA 47:403–415

Bevington PR (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill, New York, chapters 1, 5, pp 1–10, 66–91

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

Blake RD, Hinds PW, Earley S, Hillyard AL, Day GR (1986) Evolution and functional significance of the bias in codon usage. In: Sarma RH, Sarma MH (eds) Biomolecular stereodynamics IV. Proceedings of the Fourth Conversation in the Discipline Biomolecular Stereodynamics. Adenine Press, Albany NY, pp 271–286

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Calladine CR, Drew HR (1986) Principles of sequence-dependent flexure of DNA. J Mol Biol 192:907–918

Carbonnaux, C, Fazakerley GV, Sowers LC (1990) An NMR structural study of deaminated base pairs in DNA. Nucleic Acids Res 18:4075–4081

Coll M, Frederick CA, Wang AH-J, Rich A (1987) A bifurcated hydrogen-bonded conformation in the d(A·T) base pairs of the DNA dodecamer d(CGCAAATTTGCG) and its complex with distamycin. Proc Natl Acad Sci USA 84:8385–8389

Cupples CG, Cabrera M, Cruz C, Miller JH (1990) A set of *lacZ* mutations in *Escherichia coli* that allow rapid detection of specific frameshift mutations. Genetics 125:275–280

Ehrlich M, Zhang X-Y, Inamdar NM (1990) Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. Mut Res 238:277–286

Fersht AR, Knill-Jones JW, Tsui WC (1982) Kinetic basis of spontaneous mutation. Misinsertion frequencies, proofreading specificities and cost of proofreading by DNA polymerases of *E. coli*. J Mol Biol 156:37–51

Fitch WM (1983) Random sequences. J Mol Biol 163:171–176

Fix DF, Koehler DR, Glickman BW (1990) Uracil-DNA glycosylase activity affects the mutagenicity of ethylmethanesulfonate: evidence for an alternative pathway of alkylation mutagenesis. Mut Res 244:1115–1121

Frederico LA, Kunkel TA, Shaw BR (1990) A sensitive genetic assay for the detection of rate constants and the activation energy. Biochemistry 29:2532–2537

Glickman BW, Fix DF, Yatagai F, Burns PA, Schaaper RM (1986) Mechanisms of spontaneous mutagenesis: clues from mutational specificity. In: Simic MG, Grossman L, Upton AC (eds) Mechanisms of DNA damage and repair. Plenum, New York, pp 425–437

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

Hinds PW, Blake RD (1984) Degrees of divergence in the *E. coli* genome from the correlation between dinucleotide, trinucleotide and codon frequencies. J Biomol Struct Dyn 2: 101–118

Hunter WN, Brown T, Anand NN, Kennard O (1986) Structure of an adenine–cytosine base pair in DNA and its implications for mismatch repair. Nature 320:552–555

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

Josse J, Kaiser AD, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem 236: 864–871

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol 3. Academic Press, New York, pp 21–132

Kennard O (1985) Structural studies of DNA fragments: the G·T wobble base pair in A, B and Z DNA; the G·A base pair in B-DNA. J Biomol Struct Dyn 3:205–226

Koch RE (1971) The influence of neighboring base pairs upon base-pair substitution mutation rates. Proc Natl Acad Sci USA 68:773–776

Koop BF, Goodman M, Xu P, Chan P, Slightom JL (1986) Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234–238

Li W-H (1983) Evolution of duplicate genes and pseudogenes. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland MA, pp 14–37

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

Li W-H, Luo C-C, Wu C-I (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

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

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Lu A-L, Chang D-Y (1988) Repair of single base-pair transversion mismatches of *Escherichia coli* in vitro: correction of certain A/G mismatches is dependent of *dam* methylation and host *mutHLS* gene functions. Genetics 118:593–600

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

Miller JH, Low KB (1984) Specificity of mutagenesis resulting from the induction of the SOS system in the absence of mutagenic treatment. Cell 37:675–682

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

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Nelson HCM, Finch JT, Luisi BF, Klug A (1987) The structure of an oligo(dA)·oligo(dT) tract and its biological implications. Nature 330:221–226

Radman M, Wagner R (1986) Mismatch repair in *E. coli*. Annu Rev Genet 20:523–538

Razin A, Riggs AD (1980) DNA methylation and gene function. Science 210:604–610

Russell GJ, McGeoch DJ, Elton RA, Subak-Sharp JH (1973) Doublet frequency analysis of bacterial DNAs. J Mol Evol 2: 277–292

Setlow P (1976) In: Fasman GD (ed) Handbook of biochemistry and molecular biology, 3rd ed, vol II. CRC Press, Cleveland OH

Shakked Z, Rabinovich D (1986) The effect of the base sequence on the fine structure of the double helix. Prog Biophys Molec Biol 47:159–195

Shibata M, Zielinski TJ, Rein R (1991) Molecular mechanism of base substitution mutations: from hydrogen bonding to molecular dynamics. In: Beveridge DL, Lavery R (eds) Theoretical biochemistry and molecular biophysics, vol 1: DNA. Adenine Press, Schenectady, NY, pp 309–319

Singer B, Grunberger D (1983) Molecular biology of mutagens and carcinogens. Plenum, New York, chapter 3, pp 15–44

Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285–289

Topal MD, DiGuiseppi SR, Sinha N (1980) Molecular basis for substitution mutations. J Biol Chem 255:11717–11724

Tung C-S, Harvey SC (1986) Base sequence, local helix structure, and macroscopic curvature of A-DNA and B-DNA. J Biol Chem 261:3700–3709

Wiebauer K, Jiricny J (1990) Mismatch-specific thymine DNA-glycosylase and DNA polymerase β mediate the correction of G·T mispairs in nuclear extracts from human cells. Proc Natl Acad Sci USA 87:5842–5845

Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80:726–730

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

Received May 29, 1991/Revised August 19, 1991