J Mol Evol (1982) 18:360–369

Journal of Molecular Evolution

Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes

Takashi Gojobori, Wen-Hsiung Li, and Dan Graur

Center for Demographic and Population Genetics, University of Texas Health Science Center, Houston, Texas 77030, USA

Summary. The pattern of point mutations is inferred from nucleotide substitutions in pseudogenes. The pattern obtained suggests that transition mutations occur somewhat more frequently than transversion mutations and that mutations result more often in A or T than in G or C. Our results are discussed with respect to the predictions from Topal and Fresco's model for the molecular basis of point (substitution) mutations (Nature 263:285–289, 1976). The pattern of nucleotide substitution at the first and second positions of codons in functional genes is quite similar to that in pseudogenes, but the relative frequency of the transition C→T in the sense strand is drastically reduced and those of the transversions C→G and G→C are doubled. The differences between the two patterns can be explained by the observation that in the protein evolution amino acid substitutions occur mainly between amino acids with similar biochemical properties (Grantham, Science 185:862–864, 1974). Our results for the patterns of nucleotide substitutions in pseudogenes and in functional genes lead to the prediction that both the coding and non-coding regions of protein coding genes should have high frequencies of A and T. Available data show that the non-coding regions are indeed high in A and T but the coding regions are low in T, though high in A.

Key words: Neutral mutation – Transitions and transversions – Functional constraints – Base content – Substitution mutagenesis

A knowledge of the pattern of substitution mutations is important for the study of molecular evolution (Fitch 1967; Vogel and Kopun 1977; Kimura 1981) and for understanding the molecular basis of substitution mutations (Topal and Fresco 1976). In studying this pattern some authors (Fitch 1967; Vogel and Kopun 1977) have used the electrophoretic variants of hemoglobin. The pattern thus obtained may be biased because such variants are unlikely to have the same fitness and because electrophoresis can detect only mutations that change the electrophoretic mobility of the protein. Another approach to this problem is to study revertants of mutants with known base pair changes, using physiological and genetic tests (Fowler et al. 1974; Sinha and Haimes 1980). One drawback of this approach is that mutations can be studied only at particular sites and often only in certain directions. Here we propose to infer this pattern from DNA sequences for pseudogenes. As pseudogenes are apparently subject to no functional constraint, all mutations in them would be selectively neutral and would become fixed in the population with equal probability. Thus the pattern of nucleotide substitutions in pseudogenes would reflect the pattern of spontaneous substitution mutations.

Our results for the pattern of substitution mutations are useful for understanding the mechanism of substitution mutagenesis. As part of their formulation of DNA replication, Watson and Crick (1953) suggested that transition mutations might be due to the occurrence of a purine-pyrimidine (pu-pyr) pair with a base in one of its unfavored tautomeric forms. More recently, Topal and Fresco (1976) have extended the Watson-Crick concept of complementary base pairing to a wider range and proposed a more general model for the molecular basis of substitution mutations. This model has gained support from experimental studies (Topal and Fresco 1976; Sinha and Haimes 1980; Fersht and Knill-Jones 1981). It is therefore interesting to compare our results with prediction from this model.

It is also interesting to know the pattern of nucleotide substitutions in functional genes. Some authors have studied this pattern by using protein sequences (Fitch 1967; Vogel and Kopun 1977). Here we use DNA sequences. Our results for the patterns of the nucleotide substitutions in pseudogenes and functional genes may be used to predict the base contents in non-coding and coding regions in protein coding genes.

Offprint requests to: Wen-Hsiung Li

0022–2844/82/0018/0360/$02.00

Pattern of Nucleotide Substitution

Pseudogenes

A pseudogene is a DNA segment that shows high homology to a functional gene but contains nucleotide changes such as frameshifts and nonsense mutations that prevent its expression. Although some authors have argued possible functions of pseudogenes, their arguments were not substantiated (Proudfoot 1980). The pattern of nucleotide substitutions in pseudogenes can be studied as follows. Let us use the human globin pseudogene ψαl (Hψαl) as an example. This pseudogene was apparently derived from a duplication of its functional counterpart, human αl (Hαl), sometime after the divergence among man, mouse, and rabbit (Proutfoot and Maniatis 1980; Li et al. 1981). Figure 1 shows a probable evolutionary scheme for Hψαl, Hα, mouse αl (Mαl) and rabbit α (Rα); (for sequences, see Proudfoot and Maniatis 1980; Michelson and Orkin 1980; Nishioka and Leder 1979; Heindell et al. 1978). Aligning Hψαl with Hαl, we can see the nucleotide differences between the two sequences. We assume that each of these differences arose from a single substitution, neglecting the possibility of multiple substitutions. To decide the direction of substitutions we infer the ancestral sequence of Hαl, Mαl, and Rα by assuming that the ancestral nucleotide at a site is the one that requires the minimum number of substitutions to account for the nucleotide differences at that site among the three sequences. (There are some sites at which the ancestral nucleotides cannot be determined uniquely). We then attribute a nucleotide difference between Hψαl and Hαl to a substitution in Hψαl if the ancestral nucleotide at that site can be determined uniquely and is the same as the nucleotide in Hαl, but exclude it from comparison if otherwise. (We also exclude deletions, insertions, and non-coding regions). In this manner we can infer the nucleotide substitutions in Hψαl and the proportion of nucleotide changes from one type of nucleotide to another (see Table 1). Using the same ancestral sequence and the alignment between mouse pseudogene (Mψα3) and Mαl Nishioka et al. 1980), we can infer the nucleotide substitutions in Mψα3; we have excluded the 30 nucleotides that have been aligned with the nucleotides starting from position 91 to 120 of Mαl because the mismatches in this region probably occurred due to insertion rather than nucleotide substitution (Li et al. 1981). The β globin pseudogenes in rabbit (Rψβ2), goat ($G\psi\beta^X$), and mouse (Mψβh3) can be studied by using the alignments given in the original publications (Lacy and Maniatis 1980; Cleary et al. 1980; Jahn et al 1980) and the ancestral sequence of human β (Hβ), human δ (Hδ), mouse β major and β minor ($M\beta^{maj}$, $M\beta^{min}$), rabbit β1 alleles 1 and 2 (Rβ1(1), Rβ1(2)), and the partial sequence of goat β^A ($G\beta^A$) that has been published; for sequences see Lacy and Maniatis (1980), Cleary et al. (1980), Jahn et al. (1980), Lawn et al. (1980), Spritz et al. (1980), Konkel et al. (1979) and Hardison et al. (1979). Similarly, the human V_K immunoglobulin pseudogene ($H\psi V_K$) can be studied by using the alignment between $H\psi V_K$ and its functional counterpart (HV_KK101) and the ancestral sequence of HV_K and MV_KK2 (mouse V_K immunoglobulin K2); the ancestral nucleotide cannot be determined at a site at which HV_KK101 and MV_KK2 differ (for sequences and alignments, see Bentley and Rabbits (1980) and Nishioka and Leder (1980)). In the case of *Xenopus* 5S rRNA pseudogene (Xψ5S) we can attribute all of the nucleotide differences between Xψ5S (Jacq et al. 1977) and its functional counterpart in *Xenopus*, X5S, to Xψ5S. The justification for this is that the degree of divergence between the two sequences is quite small (Table 1) and that the rate of nucleotide substitution is extremely slow in 5S rRNA genes (Hori 1975) but very fast in pseudogenes (Kimura 1980; Miyata and Yasunaga 1981; Li et al. 1981). The same justification can also be made for attributing all of the nucleotide differences between human U1 RNA pseudogene (HψU1) and its functional counterpart, HU1 (Denison et al. 1981), to HψU1, because the rate of nucleotide substitution in U1 RNA is extremely slow (T. Gojobori, unpublished).

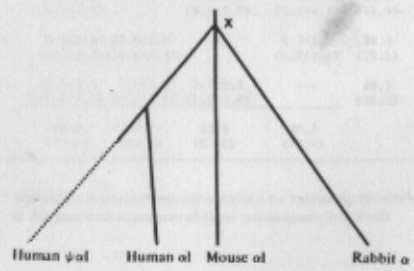

Fig. 1. Plausible phylogenetic tree for human ψαl, human αl, mouse αl, and rabbit α

The results of our analysis will be shown in terms of the sense strand of the gene, i.e., the untranscribed strand. Thus, an A→G substitution means that an A base in the sense strand is replaced by G. Because A is complementary to T and G to C, A→G actually means that an A:T pair is replaced by a G:C pair. Similarly, A→T means A:T → T:A, C→A means C:G → A:T, and so on.

Table 1. Proportions of base substitutions in pseudogenes and in functional genes

Comparison	A↔T	A↔C	A↔G	T↔A	T↔C	T↔G	C↔A	C↔T	C↔G	G↔A	G↔T	G↔C	Total	
Pseudogenes: all positions of codons														
Human ψα1	1/68	1/68	7/68	3/65	2/65	1/65	13/136	25/136	5/136	13/113	2/113	6/113	79/382	= 0.21
Mouse ψα3	1/70	2/70	5/70	3/79	1/79	5/79	1/111	9/111	1/111	7/88	3/88	1/88	39/348	= 0.11
Rabbit ψβ2	2/88	4/88	5/88	3/104	6/104	3/104	6/101	10/101	3/101	12/132	4/132	5/132	63/425	= 0.15
Goat $\psi\beta^K$	1/45	1/45	3/45	1/43	0/43	0/43	2/44	7/44	3/44	9/69	3/69	3/69	33/201	= 0.16
Mouse ψβh3	3/28	1/28	1/28	0/26	0/26	0/26	2/31	3/31	0/31	4/33	3/33	0/33	17/118	= 0.14
Human ψV_κ	2/77	3/77	3/77	0/79	7/79	2/79	3/89	9/89	3/89	7/85	6/85	3/85	48/330	= 0.15
Human ψU1	0/32	1/32	5/32	2/37	2/37	0/37	2/34	9/34	0/34	3/42	4/42	2/42	30/145	= 0.21
Frog ψ5S	3/24	1/24	0/24	0/20	0/20	2/20	0/31	5/31	1/31	3/31	0/31	0/31	15/106	= 0.14
Functional genes: first and second positions of codons														
α globin genes	2/219	0/219	5/219	1/171	2/171	0/171	4/231	2/231	6/231	8/213	1/213	8/213	39/834	= 0.05
β globin genes	4/240	6/240	4/240	3/189	2/189	2/189	3/192	2/192	6/192	7/243	3/243	2/243	44/864	= 0.05
ACTH genes	0/327	3/327	7/327	1/165	0/165	0/165	3/264	7/264	4/264	15/354	3/354	6/354	49/1110	= 0.04

From Table 1 we can see some features of nucleotide substitutions in pseudogenes. We note, for example, that there is nonrandomness in the direction of substitution; e.g., in Hψα1 A changed more often to G than to T or C. However, in order to have an easier interpretation and to be able to pool data together we need to do some mathematical manipulations. Let P_{ij} be the proportion of base changes from the ith type to the jth type (i,j=A,T,C or G) and

$$f_{ij} = \frac{P_{ij}}{\sum_i \sum_{j \neq i} P_{ij}} \times 100\%$$

e.g., in Hψα1 P_{AT} = 1/68 (Table 1). Then f_{ij} x 100 represents the expected number of base changes from the ith type to the jth among every 100 substitutions in a random sequence, i.e., in a sequence in which the four bases are equally frequent. We shall call the f_{ij}'s the relative substitution frequencies. Table 2 shows the f_{ij} values in the matrix form for six pseudogenes and the average of the eight pseudogenes used in this study; the average is weighted by the number of substitutions in a pseudogene. The four elements in the diagonal from the right upper corner to the left lower corner are the f_{ij} values for transitions while the other eight elements are for transversions; e.g., f_{ij} = 14.2 % for the transition A→G in Hψα1.

The f_{ij} values vary considerably from pseudogene to pseudogene, probably largely because of chance effects. However, the values are usually higher for transitions than for transversions, though those for the two transversions C→A and G→T tend to be higher than that for the transition T→C. In particular, the values for the two transitions C→T and G→A are in most cases the highest and the second highest, respectively. The sum of the f_{ij} values for the four transitions is given in the brackets at the right upper corner of each matrix. This sum varies from 46% to 67% but is higher than 50% in six of the eight cases; the two exceptions (not shown) are likely due to chance because the numbers of substitutions in these two cases are only 15 and 17. Thus, transitions occur on the average more frequently than transversions. We also note that the relative frequencies for transversions are on the average roughly equal to one another, except that those for C→A and G→T are somewhat higher.

To see the trend in the change of base content in pseudogenes we have computed the sum of the f_{ij} values in each column of the matrix and presented the

Table 2. Relative substitution frequencies (%) in pseudogenes and functional genes

Pseudogenes

Human ψα1

	A	T	C	G	[59.5]
A	---	2.0	2.0	14.2	18.2
T	6.4	---	4.2	2.1	12.7
C	13.2	25.1	---	5.1	43.6
G	15.8	2.4	7.3	---	25.5
	35.4	29.7	13.5	21.4	

Human ψV_κ

	A	T	C	G	[56.1]
A	---	4.5	6.8	6.8	18.1
T	0.0	---	15.4	4.4	19.8
C	5.9	17.6	---	5.9	29.4
G	14.3	12.3	6.1	---	32.7
	20.2	34.4	28.3	17.1	

Functional genes

α genes (Hα1, Hα1, Rα)

	A	T	C	G	[44.6]
A	---	5.0	0.0	12.6	17.6
T	3.2	---	6.5	0.0	9.7
C	9.6	4.8	---	14.3	28.7
G	20.7	2.6	20.7	---	44.0
	33.5	12.4	27.2	26.9	

Mouse ψα3

	A	T	C	G	[54.1]
A	---	3.2	6.3	15.8	25.3
T	8.4	---	2.8	14.0	25.2
C	2.0	17.9	---	2.0	21.9
G	17.6	7.5	2.5	---	27.6
	28.0	28.6	11.6	31.8	

Goat $\psi\beta^K$

	A	T	C	G	[57.1]
A	---	1.6	3.6	10.7	17.9
T	3.7	---	0.0	0.0	3.7
C	7.3	25.5	---	10.9	43.7
G	20.9	7.0	7.0	---	34.9
	31.9	36.1	10.6	21.6	

β genes (Hβ, $H\beta^{ma}$), Rβ1(1))

	A	T	C	G	[33.0]
A	---	8.3	12.4	8.3	29.0
T	7.9	---	5.2	5.2	18.3
C	7.7	5.2	---	15.5	28.4
G	14.3	6.1	4.1	---	24.5
	29.9	19.6	21.7	29.0	

Rabbit ψβ2

	A	T	C	G	[51.8]
A	---	3.9	7.7	9.7	21.3
T	4.9	---	9.8	4.9	19.6
C	10.1	16.8	---	5.1	32.0
G	15.5	5.2	6.4	---	27.1
	30.5	25.9	23.9	19.7	

Human ψU1

	A	T	C	G	[65.7]
A	---	0.0	3.8	18.8	22.6
T	6.5	---	6.5	0.0	13.0
C	7.1	31.8	---	0.0	38.9
G	8.6	11.6	5.7	---	25.7
	22.2	43.2	16.0	18.8	

ACTH genes (human, rat, bovine)

	A	T	C	G	[57.3]
A	---	0.0	5.8	13.6	19.4
T	3.8	---	0.0	0.0	3.8
C	7.2	16.8	---	9.6	33.6
G	26.9	5.4	10.8	---	43.1
	37.9	22.2	16.6	23.2	

Average (8 pseudogenes)*

	A	T	C	G	[55.6] (52.0)
A	---	4.7±1.9 (5.3±2.3)	5.2±0.8 (5.7±0.8)	11.4±1.6 (12.1±1.8)	21.3 (23.1)
T	4.5±1.0 (4.7±1.1)	---	6.2±1.8 (6.7±1.9)	4.6±1.8 (5.1±2.1)	15.3 (16.5)
C	8.3±1.4 (9.1±1.7)	22.0±1.8 (18.0±1.5)	---	4.7±1.0 (4.8±1.1)	35.0 (32.1)
G	16.0±1.1 (15.2±1.4)	7.0±1.5 (7.7±1.8)	5.5±0.8 (5.4±0.8)	---	28.5 (28.3)
	28.8 (29.2)	33.7 (31.0)	16.9 (17.8)	20.7 (22.0)	

Average (3 functional genes)

	A	T	C	G	[45.5]
A	---	4.2±2.0	6.3±2.8	11.5±1.4	22.0
T	5.0±1.2	---	3.7±1.6	1.7±1.4	10.4
C	8.1±0.6	9.4±3.3	---	13.0±1.5	30.5
G	20.9±3.1	4.8±0.8	11.5±3.8	---	37.2
	34.0	18.4	21.5	26.2	

* The values in parentheses are obtained by excluding the nucleotide sites were the CG dinucleotide appeared to have occurred in the ancestral sequences of these pseudogenes (see text)

results in the row below the matrix. These sums represent the numbers of substitutions that result in A, T, C, and G, respectively, among every 100 substitutions in a random sequence. We note that the sums under A and T are larger than those under G and C. We have also computed the relative frequency that one type of base is replaced by any of the others. This quantity is given by the sum of the f_{ij} values in each row and is presented in the column under the brackets. The sums are larger for C and G than for A and T so that C and G are more likely to be replaced than A and T. From these two observations, we may conclude that pseudogenes will tend to be rich in A and T bases. (This conclusion applies to both strands because A and T are complementary.) The equilibrium frequencies of A, T, C, and G can be obtained by using the formula given by Wright (1969) and Tajima and Nei (1982) and are 0.28, 0.38, 0.14, and 0.20, respectively.

It is now known that in addition to base mispairing the transition C→T can also arise from conversion of methylated C residues to T residues upon deamination (Coulondre et al. 1978; Razin and Riggs 1930). This effect will elevate the frequencies of C:G → T:A and G:C → A:T, i.e., f_{CT} and f_{GA}. As about 90% of methylated C residues in eukaryotic DNA occur at the 5'-CG-3' dinucleotides (Razin and Riggs 1980), this effect should be expressed mainly as changes of the CG dinucleotides to TG or CA. When a gene becomes a pseudogene such changes would no longer be subject to any functional constraint and can therefore contribute significantly to C→T and G→A transitions if the frequency of CG is relatively high before silencing of the gene occurs. Upon examination of the eight pseudogene sequences and their ancestral sequences we find (i) the majority of the CG dinucleotides, 34 out of 46, have changed to TG or CA, (ii) among the 77 C→T transitions 24 were due to CG→TG, and (iii) among the 58 G→A transitions 12 were due to CG → CA. (Note the asymetry that CG → TG has occurred twice as often as CG → CA). Thus, methylation of C residues appears to have contributed significantly to the frequencies of C→T and G→A and can explain partly the higher values of f_{CT} and f_{GA} over f_{TC} and f_{AG} and also partly the higher value of f_{CT} over f_{GA}. The substitution pattern obtained by excluding all nucleotide sites where the CG dinucleotide appear to have occurred in the ancestral sequences of these pseudogenes is given in the parentheses in the "average matrix" in Table 2. This pattern is somewhat different from that obtained without excluding the CG dinucleotides. In particular, the differences among the relative frequencies of the four transitions become somewhat less conspicuous and the relative frequencies for the transversions become slightly higher except for G→C. The sum of the relative frequencies of transitions is now reduced to 52%. The equilibrium frequencies of A, T, C, and G compouted from this matrix are 0.27, 0.36, 0.16, and 0.21, respectively.

The pattern of substitution mutations inferred from the electrophoretic variants of hemoglobin (Fitch 1967; Vogel and Kopun 1977) is different from that inferred from pseudogenes. In particular, the transition C→T is rare in the former but very frequent in the latter. The rarity of this transition in electrophoretic variants is probably due to natural selection (see below).

Functional Genes

The pattern of nucleotide substitution in functional genes can be studied in the same manner as above. For example, using the ancestral sequence of Hol, Mol, and Rα, we can infer the nucleotide substitutions in these three sequences. It is of course difficult to decide the direction of substitution at a nucleotide site at which the ancestral nucleotide cannot be determined uniquely and we have therefore not included such sites in our analysis. We have also not included the third position of codons because the functional constraint at this position appears to be weak in the majority of cases and we are interested in knowing the effect of functional constraint on the pattern of nucleotide substitution. In the same manner, we have inferred the nucleotide substitutions at the first and second positions of codons in the human, bovine, and rat adrenocorticotropic hormone (ACTH) genes (Chang et al. 1980; Nakanishi et al. 1979; Drouin and Goodman 1980). In inferring the nucleotide substitutions in Hβ, $M\beta^{maj}$ and Rβ(1) we used the same ancestral sequence as used in the case of β globin pseudogenes. The results of our analysis are given in Tables 1 and 2.

The pattern of nucleotide substitution in functional genes is quite similiar to that in pseudogenes, except that the relative frequency of C→T has been drastically reduced whereas those of C→G and G→C have been doubled (Table 2). As mentioned above, the transition C→T is also rare in the electrophoretic variants of hemoglobin (Fitch 1967; Vogel and Kopun 1977). In this respect, the pattern of base replacements inferred from the electrophoretic variants of hemoglobin is more similar to the pattern of nucleotide substitution in functional genes than that in pseudogenes. Because of the relatively low frequency of C→T, transition substitutions in functional genes occur somewhat less frequently than transversion substitutions. The equilibrium frequencies of A, T, C and G computed from the average matrix for the three functional genes are 0.30, 0.36, 0.17, and 0.17, respectively.

Discussion

Reliability of the Results

There are at least three factors which can affect the reliability of the results obtained above. They are the assumption of no multiple substitutions at each site, the reliability of alignment, and the sampling effect. In the case of the substitution pattern for functional genes, the first two factors should cause no serious error because in all cases the degree of sequence divergence is only 5% (Table 1) and the alignment can easily be made. The sampling effect, however, can be a serious problem for two reasons. First, the total number of substitutions observed is rather small, only 132 (Table 1), so that the pattern obtained is expected to be subject to random errors. Second, as only three different kinds of sequences are used, the pattern obtained can be very biased. Therefore, the present result should be taken as very tentative.

In the case of the substitution pattern for pseudogenes the assumption of no multiple substitutions is likely to be violated at a number of sites but should introduce no serious error because the degree of sequence divergence is still not high (Table 1). The second factor is also unlikely to cause any serious error because all the pseudogenes we used can still be easily aligned with their respective functional counterpart. (Note that in those structural pseudogenes we have used only the "exons" so that alignment was facilitated by using the reading frame in the functional genes as a reference.). The only exception is the mouse ψα3. In this case, however, we have excluded the 30 nucleotides where the alignment does not appear to be reliable (see Li et al. 1981). The total number of substitutions used to infer the substitution pattern for pseudogenes is 324 (Table 1). Although this is not a small number the sampling effect is probably not negligible. This is because the error rate of DNA replication at a site can be drastically affected by its neighboring DNA sequence (Topal and Fresco 1976; Sinha and Haimes 1980; Topal et al. 1980; Fresco et al. 1980). To get a reliable general pattern we need not only a large number of substitutions but also many different kinds of pseudogenes.

Base Contents in Coding and Non-Coding Regions

As mentioned in the Introduction the main aim of studying the pattern of nucleotide substitutions in pseudogenes is to know the pattern of spontaneous point mutations. If the pattern obtained from pseudogenes is reliable, we can predict the equilibrium frequencies of A, T, C and G in DNA sequences that are subject to no stringent functional constraint. To see how well our prediction agrees with actual data, we have compared the equilibrium base frequencies predicted by the pattern in pseudogenes with the base contents in the non-coding regions of some protein coding genes. The reason for choosing these regions is because they do not appear to require any sequence specificities other than those signals for transcriptional control, RNA processing, and translational control. Similarly, to test the prediction by the substitution pattern for functional genes, we have also studied the base contents at the first and second positions of codons in protein coding genes. The results are given in Table 3.

Table 3. Base contents and AT richness in the coding and non-coding regions of protein coding genes

Gene	Coding regions[a]						Non-coding regions					
	No. of bases	A (%)	T (%)	G (%)	C (%)	A+T (%)	No. of bases	A (%)	T (%)	G (%)	C (%)	A+T (%)
Mouse α globin (1)[b]	282	25.9	20.2	27.7	26.2	46.1	1012	23.5	22.7	28.2	25.6	46.2
Human β globin (2)	292	26.4	21.6	28.4	23.6	48.0	1608	27.7	35.9	17.7	18.7	63.6
Human ε globin (3,4)	292	30.1	22.9	24.7	22.3	53.0	2074	32.6	25.1	23.4	18.9	57.7
Human ACTH (5)	442	28.3	14.0	33.3	24.4	42.3	537	22.9	17.5	31.7	27.9	40.4
Chicken ovalbumin (6,7)	770	30.8	25.1	24.7	19.5	55.9	1885	33.1	29.2	18.4	19.3	62.3
Rat preproinsulin II (5)	218	19.3	25.2	28.0	27.5	44.5	813	22.1	26.7	25.7	25.5	48.8
Mouse γ_{2b} immunoglobulin (9)	672	31.4	21.0	22.9	26.7	52.4	862	26.8	23.9	23.5	25.8	50.7
Mouse λ_I immunoglobulin (10)	230	24.8	21.3	26.1	27.8	46.1	427	29.3	31.4	18.3	21.1	60.7
Mouse J_κ region (11)	104	30.8	22.1	28.8	18.3	52.9	1541	28.6	30.4	24.1	16.9	59.0
Total	3302	28.5	21.5	26.5	23.5	50.0	10759	28.7	27.9	22.6	20.9	56.5

a In the coding regions only the first and second positions of codons are included

b Data sources: (1) Nishioka and Leder (1979); (2) Lawn et al. (1980); (3) Baralle, Shoulders, Proudfoot (1980); (4) Baralle et al. (1980); (5) Chang et al. (1980); (6) McReynolds et al. (1978); (7) Robertson et al. (1979); (8) Lomedico et al. (1979); (9) Tucker et al. (1979); (10) Barnard et al. (1978); (11) Sakano et al. (1979)

The average frequencies of A, T, C, and G in the non-coding regions are 0.29, 0.28, 0.21, and 0.23 (Table 3) while the equilibrium frequencies predicted from the substitution pattern in pseudogenes are 0.28, 0.38, 0.14, and 0.20 when the CG dinucleotides are included in the analysis and 0.27, 0.36, 0.16, and 0.21 when the CG dinucleotides are excluded (see above). Thus the observed and predicted frequencies of A and G agree well with each other but the observed frequencies of T and C are very different from their predicted values. These two discrepancies are apparently too large to be due to chance effects alone. One possible explanation for the large discrepancies is that the substitution pattern obtained from the eight pseudogenes does not represent accurately the pattern of spontaneous point mutations. On theoretical grounds (see below), the frequencies of the four transitions are expected to be roughly equal to one another. In the pseudogenes, however, the frequency of T→C is considerably lower than those of the other transitions whereas that of C→T is very high. This is the main reason why the predicted frequency of T is very high and that of C is very low. Had the predicted frequencies of T and C been closer to each other, the above two discrepancies would have been smaller. Another possible explanation is that the observed base frequencies in the non-coding regions do not represent accurately the equilibrium frequencies expected from the pattern of spontaneous point mutations. This can arise if some parts of the non-coding regions are actually subject to significant functional constraints. Note also that nucleotide changes in non-coding regions may often arise from insertion or deletion rather than from point mutation and this can cause biases in base contents.

The average frequencies of A, T, C, and G in the coding regions are 0.28, 0.22, 0.24, and 0.27 (Table 3) while the equilibrium frequencies predicted from the pattern given in Table 2 are 0.30, 0.36, 0.17, and 0.17. Thus, the observed frequencies of T, C, and G are very different from their predicted values. As mentioned above, the substitution pattern for functional genes is based on very limited data and is probably not reliable.

It is interesting to note that the non-coding regions tend to be rich in A and T and the average proportion of A + T is about 57%. This is consistent with earlier observations (see, e.g., van den Berg 1978). On the other hand, the coding regions tend to be high in A but low in T and the average proportion of A + T is about 50%.

Molecular Basis of Substitution Mutations

So far the most plausible model for the molecular basis of substitution mutations is the one by Topal and Fresco (1976). In this model the authors make two principal assumptions. First, there is a wider set of complementary base pairs than A:T and G:C that are compatible with the steric constraints of a regular DNA helix (see their Fig. 1). The non-Watson-Crick complementary pairs constitute mispairs of two types, pu-pyr and pu-pu; pyr-pyr mispairs cannot occur. The pu-pyr mispairs are A*:C, A:C*, G*:T and G:T* while the pu-pu mispairs are $A^*:A_{syn}$, $A^*:G_{syn}$, $G^*:A_{syn}$, and $G^*:G_{syn}$, in which * denotes a minor tautomeric form and syn denotes the syn form. The frequency with which such mispairs occur is due mainly to the manifestation of the equilibrium constants for the isomerization processes required for their formation. The pathways for substitution mutations are as follows (see their Table 1): (I) Transitions arise from pu-pyr mispairing and can occur with each strand of a base pair. For example, the transition A:T → G:C can arise from A*:C, A:C*, G:T*, or G*:T. (II) Transversions arise from pu-pu mispairing but can occur only on the strand with the purine template residue. For instance, the transversion A:T → T:A can arise only from $A^*:A_{syn}$, where A* must be on the template strand. Second, there are two opportunities to express the relevant isomeric equilibria involved in particular mispairing events during the process of adding a base, first during catalytic incorporation of the new base on the growing strand, and again during a checking step. The essential concepts and the pathways proposed have gained support from experimental studies (Topal and Fresco 1976; Sinha and Haimes 1980; Fersht and Knill-Jones 1981). The quantitative aspects of the model have, however, not been well examined. Our results for the pattern of substitution mutations allow us to look into some of these aspects.

Under the pathways proposed the following properties should hold. First, f_{AT} should on the average be equal to f_{TA}, for the two transversions A→T and T→A are complementary and arise from the same pathway, i.e., from $A^*:A_{syn}$ mispairing. Similarly, we should have $f_{AC}=f_{TG}$, $f_{CA}=f_{GT}$, and $f_{CG}=f_{GC}$. These equalities are seen to hold fairly well in the pooled data, though none of them holds for all pseudogenes. Second, all transitions should be equally frquent, i.e., $f_{AG}=f_{TC}=f_{CT}=f_{GA}$, for they all arise from the same kinds of mispairing. Our results show that none of these equalities holds (Table 2). Although the two transitions A→G and T→C are complementary, f_{AG} is two times higher than f_{TC}. The two transitions C→T and G→A are also complementary, but f_{CT} is considerably higher than f_{GA}. Moreover, f_{CT} and f_{GA} are considerably higher than f_{TC} and f_{AG}. There are two possible explanations for the violation of these equalities. First, in addition to base mispairing there may be some other factors that can contribute significantly to base changes. Methylation of cytosine is such a factor. As mentioned above, this factor can partly account for the elevated frequencies of C→T and G→A. Second, the substitution pattern obtained does not represent accurately the pattern of spontaneous point mutations (see the discussion above).

Using estimated frequencies of unfavored tautomers and syn isomers, Topal and Fresco (1976) have attempted to predict the rates of substitution mutations (see their Table 2). According to their predictions, we should have the following: (I) The transversions A→C, T→G, C→G, and G→C should on the average be twice as frequent as the transversions A→T, T→A, C→A, and G→T. Our results show, instead, that all the transversions are just as frequent except C→A and G→T, which are twice as frequent as all the others. This suggest that A_{syn} and G_{syn} may occur with a more similar frequency than Topal and Fresco (1976) assumed and that the base pair mediating C→A and G→T transversions may be more stable than the other transversion mediating base pairs. Indeed, it is interesting to note that $G^*:A_{syn}$ mediates these two but no other transversions and that this is the only Topal-Fresco base pair proposed that requires a tautomeric shift in G because of a bad van der Waals interaction with A rather than for H-bonding reasons (M.D. Topal, personal communication). (II) The sum of the frequencies of all transitions should be one order higher than the sum of all transversions. (Compare the sum of such frequencies in Table 2 of Topal and Fresco (1976).) In their in vitro study of the revertant frequencies at four amber sites in φX174 DNA, Sinha and Haimes (1980) found that transitions at any site were much more frequent than transversions. Our results, however, show that the frequency of transitions is only about 10% higher than that of transversions (Table 2). The data of Fowler et al. (1974) on the revertant frequencies in the tryptophan synthetase A gene of *Escherichia coli* also suggests that the frequency of transversions is of the same order of magnitude as that of transitions. Moreover, the in vitro study of base mispairing by Fersht and Knill-Jones (1981) also does not support Topal and Fresco's prediction. Thus, the difference in frequency between transitions and transversions may not be so large as predicted by Topal and Fresco (1976). These authors have noted from Fowler et al.'s (1974) data that the overall fidelity observed for transversions is somewhat less than they expected and suggested that this could be due to a low efficiency in removal of syn than anti residues by the exonuclease. The accuracy of their predictions, of course, depends also on the reliability of their assumptions on the frequencies of minor tautomers and syn isomers.

Recently, in a comparison of mitochondrial DNA sequences from man and apes Brown et al. (1982) have found that the relative frequency of transition is ten times higher than that of transversion. This fits Topal and Fresco's (1976) prediction. It will be interesting to see whether the relative frequencies of transition and transversion in nuclear DNA are really different from those in mitochondrial DNA.

Effect of Functional Constraints

We noted earlier that, although the pattern of nucleotide substitutions at the first and second positions of codons in functional genes is quite similar to that in pseudo-

Table 4. Correlation between the relative relative frequency of base substitution and the chemical distance between bases

Sequences	N	Relative substitution frequencies						Correlation coefficient	Rank correlation
		C ↔ T	G ↔ A	C ↔ G	C ↔ A	T ↔ G	T ↔ A		
Pseudogenes									
Human φαl	49	27.6	33.5	10.1	17.2	4.6	7.0	-0.60	-0.77[a]
Mouse φαJ	25	23.8	35.1	4.1	11.4	13.0	12.7	-0.22	-0.03
Rabbit φβ2	41	21.8	26.5	10.0	19.7	12.4	9.6	-0.51	-0.49
Goat $\varphi\beta^k$	22	31.1	32.3	19.5	9.9	3.5	3.7	-0.61	-0.83[a]
Mouse φβh	13	20.1	22.0	0.0	16.0	24.1	17.8	+0.47	+0.37
Human φV_κ	28	21.4	21.6	10.5	13.8	25.8	6.9	+0.15	+0.09
Weighted average	178	24.7	29.2	9.7	15.4	12.2	8.8	-0.43	-0.49
Human φG1	30	38.2	27.3	5.7	10.8	11.4	6.5	-0.13	+0.03
Frog φ5S	15	29.0	17.4	5.8	7.5	18.0	22.4	+0.57	+0.60
Functional genes									
α globin genes	39	11.2	33.3	35.0	9.6	2.6	8.3	-0.81[a]	-0.89[a]
β globin genes	44	10.4	22.6	19.6	20.1	11.3	16.2	-0.82[a]	-0.77[b]
ACTH genes	49	16.8	40.5	20.4	13.0	5.4	3.8	-0.80[b]	-0.89[a]
Weighted average	132	13.1	32.4	24.5	14.4	6.5	9.2	-0.88[a]	-1.00[c]
Chemical distance		120.5	76.5	84.5	89.0	153.0	142.5		

N denotes the number of substitutions; a 5% significant level; b on the border line of 5% significant level; c 1% significant level

genes, some conspicuous differences do occur between them. These differences must be caused by the functional constraints in the functional genes. To understand what kinds of functional constraints were responsible for the differences we shall study the relationship between base exchangeability and amino acid exchangeability because the majority of base changes at the first and second positions of codons cause changes in amino acids. Vogel and Kopun (1977) found that a positive correlation exists between base exchangeability and amino acid exchangeability. Our approach to this problem is somewhat different from theirs.

Grantham (1974) has found that the exchangeability between amino acids in evolution is largely determined by their similarity in physicochemical properties. He defined the chemical distance between two amino acids as a function of their differences in composition (the atomic weight ratio of noncarbon elements in end groups or rings to carbons in the side chain), polarity and molecular volume. He found that the relative substitution frequency of amino acids is negatively correlated with this distance. Using Grantham's distance, we may difine the chemical distance between two nucleotide bases, say A and G, at the first position of codons as the average distance over the 16 codon pairs in each of which the two codons differ only at the first position – one has A and the other has G. (For example, ACC and GCC, ACA and GCA, and ACG and GCG are three such codon pairs). When a codon pair involves a termination codon, the chemical distance between the two codons is assumed to be two times the maximum distance (215 between tryptophan and cysteine) of all possible amino acid pairs. If the two codons are both termination codons, the codon pair is excluded from comparison because such a situation cannot arise in nature. The chemical distance between two bases at the second position of codons is defined in the same way. The average distance between bases at the first and second positions of codons is given at the bottom row in Table 4. Above the average chemical distances we have shown the relative substitution frequencies at the first and second positions of codons in functional genes and in the six structural pseudogenes. (In the cases of human ψU1 and frog ψ5S, all substitutions were included).

We note from Table 4 that if the average chemical distance between two bases is large the relative substitution frequency in the functional genes is lower than that in the pseudogenes but the situation is reversed if the average chemical distance is small. For example, the average chemical distance between C and T is large (120.5) and the relative substitution frequency between C and T in the functional genes is only half of that in the pseudogenes. On the other hand, the average chemical distance between C and G is small (84.5) and the relative substitution frequency between C and G in the functional genes is two times higher than that in the pseudogenes. Thus, the chemical distance between bases defined above appears to be a good indicator of the stringency of functional constraints.

We have also computed the moment-product correlation coefficient and the rank correlation coefficient between the average chemical distance and the relative substitution frequency (see the last two columns of Table 4). We note that a significant negative correlation exists between the two quantities in the functional genes but not in the pseudogenes.

We mentioned in the Introduction that the pattern of substitution mutations inferred from the electrophoretic variants of hemoglobin may be biased. Indeed, the frequency of C→T found was considerably lower than that in pseudogenes. This bias is apparently caused by natural selection because the average chemical distance between C and T is large (Table 4).

In a recent study of the patterns of codon usage and of nucleotide substitutions in human α- and β-globin genes, Modiano et al. (1981) observed that the relative frequency of T ⇔ non-T substitutions are much lower than those of other substitutions. They concluded that organisms have adopted a device to reduce the rates of T ⇔ non-T mutations, because such mutations often cause drastic phenotypic effects. Our results do not support their conclusion. In the pattern of substitution mutations inferred from pseudogenes the relative frequency of T ⇔ non-T mutations is 4.5% + 6.2% + 4.6% + 4.7% + 22.0% + 7.0% = 49% (see Table 2), which is by no means less frequent than other types of mutations. On the other hand, in the pattern of nucleotide substitutions in functional genes the relative frequency of T ⇔ non-T substitutions is only 28.8%, which is indeed less frequent than other types of substitutions. We may therefore conclude that the low relative frequency of T ⇔ non-T substitutions observed by Modiano et al. (1981) is due to natural selection and not due to any intrinsic mechanism of substitution mutagenesis. Actually such a mechanism is difficult to evolve because it should also reduce the frequency of A ⇔ non-A mutations, for these mutations will lead to T ⇔ non-T mutations in the complementary strand.

Acknowledgements: We thank Drs. P. Majumder, M. Nei and M.D. Topal for comments and suggestions. This work was supported by research grants from NIH and NSF.

References

Baralle FE, Shoulders CC, Goodbourn S, Jeffreys A, Proudfoot NJ (1980) The 5' flanking region of human ε-globin gene. Nucleic Acids Res 8:4393–4404

Baralle FE, Shoulders CC, Proudfoot NJ (1980) The primary structure of the human ε-globin gene. Cell 21:621–626

Bentley DL, Rabbits TH (1980) Human immunoglobulin variable region genes – DNA sequences of two V_κ genes and a pseudogene. Nature 288:730–733

Bernard O, Hozumi N, Tonegawa S (1978) Sequences of mouse immunoglobulin light chain genes before and after somatic changes. Cell 15:1133–1144

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

Chang ACY, Cochet M, Cohen SN (1980) Structural organization of human genomic DNA encoding the pro-opiomelanocortin peptide. Proc Natl Acad Sci USA 77:4890–4894

Cleary ML, Haynes JR, Schon EA, Lingrel JB (1980) Identification by nucleotide sequence analysis of a goat pseudoglobin gene. Nucleic Acids Res 8:4791–4802

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

Denison RA, van Arsdell SW, Bernstein LB, Weiner AM (1981) Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc Natl Acad Sci USA 78:810–814

Drouin J, Goodman HM (1980) Most of the coding region of rat ACTHβ-LPH precursor gene lacks intervening sequences. Nature 288:610–613

Fersht AR, Knill-Jones JW (1981) DNA polymerase accuracy and spontaneous mutation rates: Frequencies of purine-purine, purine-pyrimidine, and pyrimidine-pyrimidine mismatches during DNA replication. Proc Natl Acad Sci USA 78:4251–4255

Fitch WM (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J Mol Biol 26:499–507

Fowler RG, Degnen GE, Cox EC (1974) Mutational specificity of a conditional *Escherichia coli* mutator, *mut* D5. Mol Gen Genet 133:179–191

Fresco JR, Broitman S, Lane A-E (1980) Base mispairing and nearest-neighbor effects in transition mutations. In: Alberts B (ed) Mechanistic studies of DNA replication and genetic recombination. ICN-UCLA Symposia on Molecular and Cellular Biology, vol. 19. Academic Press, New York, p 753–768

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

Hardison RC, Butler III ET, Lacy E, Maniatis T, Rosenthal N, Efstratiadis A (1979) The structure and transcription of four linked rabbit β-like globin genes. Cell 18:1285–1297

Heindell HC, Liu A, Paddock GV, Studnicka GM, Salser WA (1978) The primary sequence of rabbit α-globin mRNA. Cell 15:43–54

Hori H (1975) Evolution of 5S RNA. J Mol Evol 7:75–86

Jacq C, Miller JR, Brownlee GG (1977) A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell 12:109–120

Jahn CL, Hutchinson III CA, Phillips SJ, Weaver S, Haigwood NL, Voliva CF, Edgell MH (1980) DNA sequence organization of the β-globin complex in the BALB/c mouse. Cell 21:159–168

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Konkel DA, Maizel Jr. JV, Leder P (1979) The evolution and sequence comparison of two recently deverged mouse chromosomal β-globin genes. Cell 18:865–873

Lacy E, Maniatis T (1980) The nucleotide sequence of a rabbit β-globin pseudogene. Cell 21:545–553

Lawn RM, Efstratiadis A, O'Connel C, Maniatis T (1980) The nucleotide sequence of the human β-globin gene. Cell 21:647–851

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Lomedico P, Rosenthal N, Efstratiadis A, Gilbert W, Kolodner R, Tizard R (1979) The structure and evolution of the two nonallelic rat preproinsulin gene. Cell 18:545–558

McReynolds L, O'Malley BW, Nisbet AD, Fothergill JE, Givol D, Fields S, Robertson M, Brownlee GG (1978) Sequence of chicken ovalbumin mRNA. Nature 273:723–728

Mickelson AM, Orkin SH (1980) The 3' untranslated regions of the duplicated human α-globin genes are unexpectedly divergent. Cell 22:371–377

Miyata T, Yasunaga T (1981) Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. Proc Natl Acad Sci USA 78:450–453

Modiano G, Battistuzzi G, Motulsky AG (1981) Nonrandom patterns of codon usage and of nucleotide substitutions in human α- and β-globin genes: An evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA 78:1110–1114

Nakanishi S, Inoue A, Kita T, Nakamura M, Chang ACY, Cohen SN, Numa S (1979) Nucleotide sequence of cloned cDNA for bovine corticotropin-β-lipotropin precursor. Nature 278:423–427

Nishioka Y, Leder P (1979) The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell 18:875–882

Nishioka Y, Leder P (1980) Organization and complete sequence of identical embryonic and plasmacytoma κ V-region genes. J Biol Chem 255:3691–3694

Nishioka Y, Leder A, Leder P (1980) Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. Proc Natl Acad Sci USA 77:2806–2809

Proudfoot NJ (1980) Pseudogenes. Nature 286:840–841

Proudfoot NJ, Maniatis T (1980) The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell 21:537–544

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

Robertson MA, Staden R, Tanaka Y, Catterall JF, O'Malley BW, Brownlee GG (1979) Sequence of three introns in the chick ovalbumin gene. Nature 278:370–372

Sakano H, Hüppi K, Heinrich G, Tonegawa S (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature 280:288–294

Sinha NK, Haimes MD (1980) Probing the mechanism of transition and transversion mutagenesis using the phage T4 DNA replication apparatus in vitro. In: Alberts B (ed) Mechanistic studies of DNA replication and genetic recombination. ICN-UCLA Symposia on Molecular and Cellular Biology, vol. 19. Academic Press, New York, p 707–723

Spritz RA, DeRiel JK, Forget BG, Weissman SM (1980) Complete nucleotide sequence of the human δ-globin gene. Cell 21:639–646

Tajima F, Nei M (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J Mol Evol 18:115–120

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

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

Tucker PW, Marcu KB, Newell N, Richards J, Blattner FR (1979) Sequence of the cloned gene for the constant region of murine γ2b immunoglobulin heavy chain. Science 206:1303–1306

Van den Berg J, van Ooyen A, Mantei N, Schamböck A, Grosveld G, Flavell RA, Weissmann C (1978) Comparison of cloned rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. Nature 276:37–44

Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180

Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967

Wright S (1969) Evolution and the genetics of populations, vol. 2. University of Chicago Press, Chicago, p 26

Received October 26, 1981 / Revised April 14, 1982