Non random DNA evolution

CARLOS Y VALENZUELA*

Departamento de Biología Celular y Genética, Facultad de Medicina,
Universidad de Chile, Santiago, Chile

A model for testing random molecular evolution is proposed. Randomness of recurrent mutation is defined based on isotropy and zero covariance among nucleotide sites. Assuming an equal rate of mutation for the bases A, T, G, and C, in both DNA strands, a mutational matrix of transformation A, T, G, and C with 6 parameters is developed. Under this model the equilibrium proportions (F) of the bases are $F_A = F_T = (D+E)/[2(D+E+H+J)]$ and $F_G = F_C = (H+J)/[2(D+ E+H+J)]$, D, E, H, J being 4 of the 6 matrix parameters. Thus the expected $(F_A+ F_T)/(F_{G+}F_C)$ ratio can also be tested. If the average rate of mutation is 10^{-8} per nucleotide site and cell replication, the equilibrium for every site, in most species, is reached in 10^8 years. Eight DNA segments from human, bacteria, fungus and insect genomes were chosen to test these proportions and their heterogeneity among coding and non coding subsegments. While F_G was similar to F_C as expected, F_A was highly different from F_T. Huge heterogeneities were found between coding and non coding segments and among non coding segments. These results are a strong evidence for non randomness of molecular evolution.

Key terms: *DNA, molecular evolution, non-random*

INTRODUCTION

Is there sufficient time for adenine (A), thymine (T), guanine (G) and cytosine (C) nucleotides to have reached their equilibrium frequencies in DNA? Let m = 10^{-8} be the rate of recurrent mutation in a nucleotide site per cycle of cell replication. If we consider all kind of mutation, in replication, due to radiation or chemical agents, gene conversion, error in sex processes, DNA rearrangements, *etc*, this seems a reliable figure (Valenzuela & Santos, 1996).

The probability of no mutation for a particular site *in a cell cycle should be (1-m). In n cycles this probabil*ity is $(1\text{-m})^n$, which tends to 0 as n increases. Sooner or later, every site should be mutated. The question turns out to be: Is there sufficient DNA (cell) replications since the origin of a DNA segment till now for its bases to have reached the equilibrium frequencies? In 10^8 cell cycles (1/m) the probability that a site remains in the original state is $1/e^1$ = 0.3679 (assuming Poisson distribution with parameter 1, in this case). For 2/m, 3/m, 4/m, 6/m and 10/m cell cycles this probability should be $1/e^2$ = 0.1353, $1/e^3$ = 0.0498, $1/e^4$ = 0.0183, $1/e^6$ = 0.0025 and $1/e^{10}$ = 0.000045 with parameters 2, 3, 4, 6, 10, respectively.

Let's denote the number of cell cycles equal to 1/m as one replacement cycle (RC) in a site. In 10 RC the probability for a site

[*] **Correspondence to:** Dr Carlos Y Valenzuela, Departamento de Biología Celular y Genética, Facultad de Medicina, Universidad de Chile, Independencia 1027, Casilla 70061, Santiago, Chile. Phones: (56-2) 678-6456, -6302. Fax: (56-2) 737-3158.

to remain unchanged is 45 in a million. It can also be seen as, in average, 45 among 10^6 sites remain unchanged in this genome.

It is important to remark that this conclusion holds independently of the genome that carry the site or the population size where it is found. A bacterium can easily have 1000 cell cycles a year. Thus, in 100,000 years, bacteria have 10^8 cell cycles or 1 RC. In one million years (MY) the bacteria could have 10 RC, thus, all the nucleotide sites shall be changed, with the exception of an average of 45 in a million sites. A very similar situation is valid for unicellular eukaryotes.

Multicellular eukaryotes have several cell cycles between the zygote stage and the gamete emission. In general, the first gamete stem cell or germ cell appears in embryos near the 1000 cell stage, that is 10 cell cycles. Other 10 cell cycles are needed to yield 1000 (1024) gametes from one stem cell, and 20 cycles yield 10^6 gametes (Valenzuela & Santos, 1996). Thus, a minimal of 20 to 30 cell cycles occur within a generation period of multicellular organisms. Several insects, as flies, and small animals, as rats or mice, have a generation period or life span of 20 to 40 days. For example, the life span of the worm *Caenorhabditis elegans* is 15-18 days (Lakowski & Hekimi, 1996). The first germ cell appears at 7-8 cell divisions from the egg (Watson *et al*, 1987). It needs 10 cell cycles to yield 1000 sperms. Thus the number of cell cycles per generation is, at least, 17 and the time for one generational cell cycle (TGCC) is approximately 1 day. It is to be noted that most living beings have a TGCC close to one or a few days. In one year, most organisms can have 300 generational cell cycles. Only big animals or plants with long generation time have fewer cycles than that. Present human beings have a generation period of 30 years and produce around 10^9 gametes (30 cell cycles). Adding 10 cycles to yield the first germ cell, we have a TGCC near 270 days (0.75 year). This is the present human situation, but, 20 MY ago, the generation time for "human ancestors" could be more similar to small mammals than to present humans. In fact, the generation time increased in primates (Martin, 1990) and humans and mice diverged 75 MY ago (Nei, 1987). Moreover, since we consider DNA segments independent from the species at which they belong, it would be sufficient to deal with DNA segments more than 10^8 years old (any present gene), to be sure that they have reached the state of site replacement and base frequency equilibrium.

THE MATRIX MODEL

The proportion of bases at equilibrium can be obtained from the base frequency vector F = (F_A, F_T, F_G, F_C) and the mutational matrix M:

$$M = \begin{vmatrix} M_{A-A} & M_{T-A} & M_{G-A} & M_{C-A} \\ M_{A-T} & M_{T-T} & M_{G-T} & M_{C-T} \\ M_{A-G} & M_{T-G} & M_{G-G} & M_{C-G} \\ M_{A-C} & M_{T-C} & M_{G-C} & M_{C-C} \end{vmatrix}$$

Where M is the mutation rate from the base at the left into that at the right side of the subscript, in one cell cycle. $M_{A-A} = 1-M_{A-T}-M_{A-G}-M_{A-C}$; $M_{T-T} = 1-M_{T-A}-M_{T-G}-M_{T-C}$; $M_{G-G} = 1-M_{G-A}-M_{G-T}-M_{G-C}$; $M_{C-C} = 1-M_{C-A}-M_{C-T}-M_{C-G}$ (12 parameters or mutation rates are necessary to define the system). The frequency vector in the cell cycle t+1 is given by $F_{t+1} = MF_t$. Thus, the equilibrium frequency for each base is reached when $F_{t+1} = F_t$ or $F_t = MF_t$ (Nei, 1987). If all coefficients are equal the expected vector F is (1/4, 1/4, 1/4, 1/4) (Valenzuela & Santos, 1996).

We have shown that in the case transitions have different rate from transversions the expected vector is also (1/4, 1/4, 1/4, 1/4) (Valenzuela & Santos, 1996). However, if all the rates of mutation are different the elements of the vector should be consequently different. Fortunately, a realistic simplification comes from the complementary nature of DNA.

Let's assume that randomness means: i) isotropy, that is, the occurrence of any event in each base is distributed homogeneously along with DNA; ii) independence, that is, any event at a site occurs independently of the occurrence of events at any other site (the covariance is zero). Let's remember that a point mutation arises from a non repaired base change.

If we consider the base pair A-T any mutation of A implies, in the next round of replication or repair, a mutation in its complementary T. Any mutation of T implies a mutation in its complementary A. For example, if A mutates into G, this implies a mutation of the complementary T into C. This change could be produced by the previous mutation from T into C and then the implied change of the complementary A to G. Thus, under the assumption of isotropy (the rate of mutation is equal in both DNA strands), and considering both strands, the overall rate of mutation of A (in the left strand) into G is the addition of $M_{A\text{-}G} + M_{T\text{-}C}$ which is equal to $M_{T\text{-}C} + M_{A\text{-}G}$, that occurs in another nucleotide site, with T in the left strand. This is equivalent to imagine that there are two matrices M, one for the left strand and the other for the right strand.

Several attempts have been made to deal with randomness *vs* non-randomness DNA evolution by examining nucleotide substitutions (Li *et al*, 1984; Ayala *et al*, 1996). However, substitutions involve the mechanism of fixation and they are a few nucleotides in a DNA segment. Our approach deals with all the nucleotides in a DNA segment and, more general, in genomes independently of species or populations. This approach is similar to those searching for a correlation among bases in long segments of DNA (Peng *et al*, 1992; Pande *et al*, 1994).

Evidence for the equality of nucleotide substitution rates in both strands has been found with point mutations in pseudogenes (Li *et al*, 1984). It is important to remark that A and T co-mutate, as well as G and C do. Thus, the matrix M' (the fused two Ms for both strands) can be written:

$$M' = \begin{vmatrix} R & M_{T\text{-}A}+M_{A\text{-}T} & M_{G\text{-}A}+M_{C\text{-}T} & M_{C\text{-}A}+M_{G\text{-}T} \\ M_{A\text{-}T}+M_{T\text{-}A} & R & M_{G\text{-}T}+M_{C\text{-}A} & M_{C\text{-}T}+M_{G\text{-}A} \\ M_{A\text{-}G}+M_{T\text{-}C} & M_{T\text{-}G}+M_{A\text{-}C} & S & M_{C\text{-}G}+M_{G\text{-}C} \\ M_{A\text{-}C}+M_{T\text{-}G} & M_{T\text{-}C}+M_{A\text{-}G} & M_{G\text{-}C}+M_{C\text{-}G} & S \end{vmatrix} = \begin{vmatrix} R & B & D & E \\ B & R & E & D \\ H & J & S & K \\ J & H & K & S \end{vmatrix}$$

Where $R = 1\text{-}B\text{-}H\text{-}J$; $S = 1\text{-}D\text{-}E\text{-}K$.

We have only 6 different mutation rates in M'. The mutational behaviour of A is identical to that of T, and the mutation rates are equal between G and C. The conversion of (A or T) into (G or C) with the mutation rate 2H + 2J and the reverse conversion (G or C) to (A or T) with rate 2D+2E could be different. The equilibrium frequencies should be: $F_A = F_T = (D+E)/[2(D+E+H+J)]$ and $F_G = F_C = (H+J)/[2(D+E+H+J)]$.

The demonstration is straightforward either by the matrix method (Nei, 1987) or by simple algebra. Since the mutation coefficients for A are equal to those for T, at equilibrium, the matrix vector product should be equal for both bases, thus F_A should be equal to F_T and by the same reason F_G should be equal to F_C. Let $p = F_A+F_T$, $q = F_G+F_C$, $(p+q = 1)$, $u = 2(H+J)$, $v = 2(D+E)$. Then the equilibrium is reached when $up_E = vq_E$ or $up_E = v(1\text{-}p_E)$. Thus at equilibrium $p_E = v/(u+v)$, $q_E = u/(u+v)$. Since $F_A = F_T$, $F_G = F_C$, $p = F_A+F_T$ and $q = F_G+F_C$, $F_A = v/[2(u+v)]$ and $F_G = u/[2(u+v)]$. It must be remarked that this should occur in a single strand of DNA; then, it can be tested with published or bank data. For any DNA segment with more than 10 RC, independent of its function and species, $F_A = F_T$, $F_G = F_C$ and $(F_A+F_T)/(F_G+ F_C)$ should be expected to be defined proportions (random variables).

SAMPLES AND MODEL TESTING

DNA segments of human growth hormone (GH, from data bank), MOZ gene (MOZ) (Borrow *et al*, 1996) and β-globin (Nei, 1987), of *Drosophila* gene torso (*Torso*) (Sprenger *et al*, 1989), of fungus *Cryptococcus neoformans* Mtl-1 gene (Perfect *et al*, 1996), of bacteria *Pseudomonas putida* bedD gene (Fong *et al*, 1996), *Vibrio anguillarum* flaC gene (McGee *et al*, 1996) and *Thermotoga maritima* drrA and hpkA genes (Lee & Stock, 1996) were chosen to test these expected proportions. These DNA segments were taken from those that some colleagues were studying or from journals given to postgraduate students to prepare academic events. It is assumed that any DNA segment must be a random collection of nucleotides under a random model of evolution.

Table I shows the distribution of bases and their percentages among the human and *Drosophila* DNA segments according to the conventional and functional (protein synthesis) divisions. Here, coding regions are those involved in amino acid specification. Statistical significance for departure from the expected proportions or for heterogeneity was studied by the χ^2 test. Degrees of freedom are denoted as subscript.

The four DNA segments presented F_G/F_C ratios very near 1. The F_A/F_T ratio deviated very significantly from 1 in MOZ (1.25, $\chi^2_1 = 52.0$, $p << 10^{-6}$), β Globin (0.78, $\chi^2_1 = 19$, $p < 1.3x10^{-5}$) and Torso (1.12, $\chi^2_1 = 8.1$, $p < 0.0046$); it was very close to 1 in GH.

Significant heterogeneities among conventional coding-noncoding DNA segments were found for the A/T ratio in: GH ($\chi^2_3 = 18.7$, $p < 0.00033$) due to the very low frequency of T (and high frequency of A) in non coding initial segment; MOZ ($\chi^2_2 = 43.7$, $p < 10^{-6}$) mainly due to the high frequency of T found in non coding regions; Torso ($\chi^2_5 = 21.7$, $p < 0.00061$) due to the high proportion of T found in Introns. The only significant heterogeneity for the G/C ratio was found in GH ($\chi^2_3 = 8.3$, $p < 0.0412$) due to the high proportion of T in Exons.

Comparisons between coding (exons) and non coding segments were significant for the A/T ratio in MOZ ($\chi^2_1 = 43.0$, $p << 10^{-6}$) and Torso ($\chi^2_1 = 9.1$, $p < 0.0025$); and for the G/C ratio only in GH ($\chi^2_1 = 6.0$, $p < 0.0147$). Analyses of non coding segments were performed to find hetero-

Table I

A, T, G, C nucleotide bases distribution in 3 human and a *Drosophila* DNA segments (N and %)

	A	T	G	C	Tot	A	T	G	C	Tot
	HUMAN GROWTH HORMONE					HUMAN *MOZ*				
Non coding initial	66(30)	32(14)	65(29)	61(27)	224	96(24)	113(29)	101(26)	82(21)	392
Coding exons	150(23)	137(21)	160(24)	207(32)	654	1787(30)	1259(21)	1502(25)	1464(24)	6012
Non coding introns	178(22)	175(21)	246(30)	215(27)	814					
Non coding terminal	100(20)	139(27)	129(25)	140(28)	508	439(30)	484(33)	254(17)	288(20)	1465
Total	494(22)	483(22)	600(27)	623(28)	2200	2322(29)	1856(24)	1857(24)	1834(23)	7869
A/T, G/C and (A+T)/(G+C) ratios	1.023		0.963		0.799	1.251		1.103		1.132
	HUMAN β GLOBIN					*Drosophila* GENE *Torso*				
Non coding(flanking) initial	24(23)	17(17)	34(33)	28(27)	103	208(29)	201(29)	145(21)	147(21)	701
cDNA non coding(protein) initial	16(32)	12(24)	6(12)	16(32)	50	74(38)	49(25)	39(20)	33(17)	195
Coding exons	86(19)	105(24)	136(31)	114(26)	441	770(28)	615(22)	712(26)	675(24)	2772
Non coding introns	271(28)	385(39)	160(16)	163(17)	979	266(31)	322(37)	141(16)	143(16)	872
cDNA non coding(protein) terminal	35(26)	49(36)	23(17)	28(21)	135	32(33)	28(29)	16(17)	20(21)	96
Non coding(flanking) terminal	100(29)	116(34)	62(18)	65(19)	343	46(43)	35(32)	12(11)	15(14)	108
Total	532(26)	684(33)	421(21)	414(20)	2051	1396(29)	1250(26)	1065(22)	1033(22)	4744
A/T, G/C and (A+T)/(G+C) ratios	0.778		1.017		1.456	1.117		1.031		1.261

geneity among these assumed homogeneous functional regions. Significant heterogeneity for A/T ratio was found in GH (χ^2_2 = 18.2, p < 0.00012) due to low frequency of T in the non coding initial segment; and in Torso (χ^2_4 = 12.2, p < 0.0158) mainly due to the high frequency of A in initial non coding cDNA and its low proportion in Exons. For G/C ratio, there was not significant heterogeneity.

The ratio (F_A+F_T)/(F_G+F_C) was 0.7989, 1.1319, 1.4563 and 1.2612 in GH, MOZ, *β*-globin and Torso, respectively. A significant heterogeneity for this ratio, among the different DNA segment was found in MOZ (χ^2_2 = 72.0, p << 10^{-6}) mainly due to an extreme deficiency of (G+C) in the non coding terminal segment (and an excess of A+T), β-globin (χ^2_5 = 89.6, p < 10^{-6}) due to an excess of (G+C) in Exons, and in Torso (χ^2_5 = 110.0, p << 10^{-6}) due to equal proportions (A+T = 50%) and (G+C = 50%) in Exons and a very low proportion (32%) of (G+C) in Introns. Heterogeneity in this ratio was also found among non coding segments in MOZ (χ^2_1 = 12.2, p < 0.0005) due to a high (G+C) proportion in the non coding initial segment; β-globin (χ^2_4 = 31.5, p < 0.000003) and Torso (χ^2_4 = 20.0, p < 0.0005), both, also due to a high (G+C) proportion in the non coding initial segment.

Heterogeneity was searched for between Exons and non coding segments. It was significant in MOZ (χ^2_1 = 60.4, p << 10^{-6}) due to a low (G+C) proportion (a high frequency of A+T) in non coding segments; in β-globin (χ^2_1 = 59.4, p << 10^{-6}) due to a high proportion of (G+C) in Exons; and in Torso (χ^2_1 = 91.3, p << 10^{-6}) due to a low proportion of (G+C) in non coding DNA. There was heterogeneity for (A+T)/(G+C) ratio among the three human DNA segments (χ^2_2 = 96.8, p << 10^{-6}). The heterogeneity in the distribution of the four bases among all the segments was highly significant: GH (χ^2_9 = 29.2, p < 0.00061); MOZ (χ^2_6 = 123.6, p < 1.2×10^{-6}); β-globin (χ^2_{15} = 105.4, p < 2×10^{-6}); Torso (χ^2_{15} = 136.0, p < 1.2×10^{-6}). Also, there was heterogeneity of the A, T, G, C distribution among the three human DNA segments (χ^2_6 = 156.8, p << 10^{-6}).

Table II shows the bases distribution in the fungus *Cryptococcus neoformans* and bacteria *Pseudomonas putida*, *Vibrio*

Table II

A, T, G, C nucleotide bases distribution in a fungus and 3 bacteria DNA segments (N and %)

	A	T	G	C	Tot	A	T	G	C	Tot
	C neoformans FUNGUS *Mtl-1* GENE					*P putida bedD* GENE				
Non coding initial	94(28)	76(23)	75(23)	87(26)	332	45(30)	34(23)	32(21)	39(26)	150
Coding segment	303(29)	241(23)	223(22)	274(26)	1041	197(18)	176(16)	363(33)	362(33)	1098
Non coding terminal	117(30)	123(32)	79(20)	72(18)	391	121(17)	155(22)	224(31)	214(30)	714
Total	514(29)	440(25)	377(21)	433(25)	1764	363(18)	365(19)	619(32)	615(31)	1962
A/T, G/C and (A+T)/(G+C) ratios	1.168		0.871		1.178	0.995		1.007		0.590
	V anguillarum flaC GENE					*T maritima drrA* AND *hpkA* GENES				
Non coding initial	67(26)	84(33)	47(19)	57(22)	255	74(24)	108(34)	71(23)	60(19)	313
Coding segment	355(31)	270(24)	258(23)	248(22)	1131	753(32)	514(22)	645(27)	466(19)	2378
Non coding terminal	111(32)	100(28)	75(21)	67(19)	353					
Total	533(31)	454(26)	380(22)	372(21)	1739	827(31)	622(23)	716(27)	526(19)	2691
A/T, G/C and (A+T)/(G+C) ratios	1.174		1.022		1.313	1.330		1.361		1.167

anguillarum and *Thermotoga maritima*. Since the analysis is the same as in Table I the χ^2 value will be omitted.

As in human and *Drosophila* DNA segments the G/C ratio was nearer 1 than the A/T ratio in the four species, with the exception of *T maritima* where the G/C was largely over 1 (1.36, $p << 10^{-6}$), as it was the A/T ratio (1.33, $p << 10^{-6}$). *C neoformans* showed a decreased G/C ratio in the limit of significance (0.87, $p = 0.0491$) as well as an increased significant A/T ratio (1.17, $p = 0.0166$). It was remarkable the close agreement to 1 of the A/T and G/C ratios in *P putida*. Also, as in human and *Drosophila* genes there was a significant heterogeneity of bases distribution among coding and non coding segments: $p = 0.0097$ in *C neoformans* mainly due to an excess of T and a lack of C in the non coding terminal segment; $p = 0.000054$ in *P putida* mainly due to an excess of A and a deficiency of G in non coding initial segment, a lack of T in coding segment and an excess of T in the non coding terminal segment; $p = 0.054$ in *V anguillarum*, which is at the limit of signification, but with a significant excess of T in non coding initial segment; $p = 0.0000038$ in *T maritima* due to an excess of T and a deficiency of A in non coding initial segment.

The A/T ratio was heterogeneous among coding and non coding segments in *P putida* ($p = 0.0318$) because a low proportion of A in non coding terminal, *V anguillarum* ($p = 0.0205$) because of a high proportion of T in non coding initial and *T maritima* ($p = 0.0000014$) because a high proportion of T and a low one of A in non coding initial segment. No significant heterogeneities among coding and non coding DNA segments were found for the G/C ratio. The ratio $(F_A+F_T)/(F_G+F_C)$ was 1.7778, 0.5900, 1.3125 and 1.1667 in Mtl-1, bedD, flaC and, drrA and hpkA genes, respectively. This ratio was highly heterogeneous among coding and non coding segments in *P putida* ($p = 0.000028$) due to a high proportion of A+T and a low one of G+C in non coding initial segment, and *C neoformans* ($p = 0.0043$) due to a low proportion of G+C and a high one of A+T in non coding terminal segment. It was not significant in the other two bacteria. Significant heterogeneities in the distribution of the four bases between initial and terminal non coding segments were found in *C neoformans* ($p = 0.0156$) and *P putida* ($p = 0.00117$). The total heterogeneity of the four bases distribution among the DNA of the three bacteria was enormously significant ($p <<< 10^{-6}$) mainly due to the inverse and large proportion of G+C in *P putida*.

DISCUSSION

The very similar proportion of G and C agrees remarkably with the model. This agreement, the weak heterogeneity of the G/C ratio found in GH and the solely highly significant deviation in *T maritima* make the model reliable. *Thermatoga maritima* belongs to the oldest branch of eubacteria (Prescott *et al*, 1996) and has its optimal growth at 80° C (Lee & Stock 1996). Its large disagreement with the model could be due to these two features. The proportions of A and T deviated from the expected equal ratio and were extraordinarily heterogeneous in six out of the eight DNA segments. The A/T ratio was quite heterogeneous between coding and non coding segments, and the G/C ratio only showed a moderate significant heterogeneity in these segments for GH. This analysis allows to conclude that, while the behaviour of G-C could be explained by random mutation (with the exception of the heterogeneity in GH), the A-T distribution can not be explained by a simple stochastic process. However, the large deviations (several of them are beyond any computer software power) in (A+T)/(G+C) ratio or four base proportions found either in total segments, between coding and non coding segments, or among non coding segments, allow us to conclude that molecular evolution (as far as these eight segments are concerned) could not occur at random.

Our proposition (Valenzuela & Santos, 1996) –that random mutations searched for self-organizing structures, and once they were found, they were mostly preserved or maintained mainly by adaptive selection– finds support from the present study. It is important to remark that this is also valid for non coding DNA segments and points out that they have a hidden functional property that we cannot see at this moment. Accordingly, long range correlations in nucleotide sequences have been found in several DNA segments by studying nucleotide distributions as a process of random walk (Peng *et al*, 1992). These correlations show different kinds of non random nucleotide walk for intron-rich, intron-less and bacteriophage gene sequences. Also, a similar analysis allowed Pande *et al* (1994) to find non randomness in protein sequences.

The present approach is different from those searching for deviations from randomness through studies of molecular clocks, genetic distances or similarities or rates of replacements (Li *et al*, 1984; Ayala *et al*, 1996). This study test randomness directly, provided that whole DNA segments have reached their equilibrium (Valenzuela & Santos, 1996) and is mostly concerned with evolutionary "modus" or specificity, while those ones test randomness according to deviations from the expected Poisson distributions or rate of replacement heterogeneity and are concerned rather with evolutionary "tempus" or kinetics of nucleotide substitutions. However, since those studies create and test models with observed distributions, they include a large epistemic circular restriction (Valenzuela, 1994), which is not present when observed distributions are compared with theoretical random expected distributions issued from all the possible alternatives. Thus, Li *et al* (1984) found non randomness of point mutations by studying nucleotide substitutions in pseudogenes. Transitions occurred with a higher rate than expected for transversions. This study cannot be assimilated to the present one because it dealt with substitutions that occur in a few of all nucleotides. Large differences in the proportion of bases or substitution rates were not analyzed by those authors.

REFERENCES

AYALA FJ, BARRIO E, KWIATOWSKI J (1996) Molecular clock or erratic evolution? A tale of two genes. Proc Natl Acad Sci USA 93: 11729-11734

BORROW J, STANTON VP Jr, ANDRESEN JM, BECHER M, BEHM FG, CHAGANTI RSK, CIVIN CL, DISTECHE C, DUBE I, FRISCHAUF AM, HORSMAN D, MITELMAN F, VOLINIA S, WATMORE AE, HOUSMAN DE (1996) The translocation t(8;16)(p11; p13) of acute myeloid leukaemia fuses a putative acetyltransferase to the CREB-binding protein. Nature Genet 14: 33-41

FONG KPY, GOH CBH, TAN HM (1996) Characterization and expression of the plasmid-borne bedD gene form *Pseudomonas putida* ML2, which codes for a NAD$^+$-dependent *cis*-benzene dihydrodiol dehydrogenase. J Bacteriol 178: 5592-5601

LAKOWSKI B, HEKIMI S (1996) Determination of life span in *Caenorhabditis elegans* by four clock genes. Science 272: 1010-1013

LEE PJ, STOCK AM (1996) Characterization of the genes and proteins of a two-component system from the hyperthermophilic bacterium *Thermatoga maritima*. J Bacteriol 178: 5579-5585

LI WH, CHUNG-I W, CHI-CHENG L (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58-71

MARTIN RD (1990) Primate origins and evolution. London: Chapman & Hall

McGEE K, HÖRSTEDT P, MILTON DL (1996) Identification and characterization of additional flagellin genes from *Vibrio anguillarum*. J Bacteriol 178: 5188-5198

NEI M (1987) Molecular Evolutionary Genetics. New York: Columbia Univ Press

PANDE VS, GROSBERG AY, TANAKA T (1994) Nonrandomness in protein sequences: evidence for a physically driven stage of evolution. Proc Natl Acad Sci USA 91: 12972-12975

PENG CK, BULDYREV SV, GOLDBERGER AL, HAVLIN S, SCIORTINO F, SIMONS M, STANLEY HE (1992) Long-range correlations in nucleotide sequences. Nature 356: 168-170

PERFECT JR, RUDE TH, WONG B, FLYNN T, CHATURVEDI V, NIEHAUS W (1996) Identification of a *Cryptococcus neoformans* gene that directs expression of the cryptic *Saccharomyces cerevisiae* mannitol dehydrogenase gene. J Bacteriol 178: 5257-5262

PRESCOTT LM, HARLEY JP, KLEIN DA (1996) Microbiology. 3rd ed. Dubuque, IA: Wm C Brown Publ

SPRENGER F, STEVENS LM, NÜSSLEIN-VOLHARD C (1989) The *Drosophila* gene torso encodes a putative receptor tyrosine kinase. Nature 338: 478-483

VALENZUELA CY (1994) Epistemic restrictions in population biology. Biol Res 27: 85-90

VALENZUELA CY, SANTOS JL (1996) A model of complete random molecular evolution by recurrent mutation. Biol Res 29: 203-212

WATSON JD, HOPKINS NH, ROBERTS JW, ARGETSINGER-STEITZ J, WEINER AM (1987) Molecular Biology of the Gene. Vol II. Menlo Park, CA: Benjamin/Cummings Publ Co