Biol Res 29: 203-212 (1996)



# A model of complete random molecular evolution by recurrent mutation

CARLOS Y VALENZUELA and JOSE L SANTOS

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

*A model for random molecular evolution based on recurrent mutation is proposed. Recurrent mutation replaces completely any original base in a nucleotidic site. This occurs if more than four times the number of reproductive cycles equal to the reciprocal of the mutation rate happen; no matter the population size, the number of nucleotides a genome has, or the taxa at which it belongs. The main results are: i) the expected distribution of DNA bases in a site is an isotetranomial distribution, where Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) occur with probability equal to 0.25; ii) the distribution of bases in a site is independent from the distribution of bases in other sites. Several expected consequences that can be contrasted with actual data are generated. Species or operational taxonomic units (OTUs) that evolved in big populations should present distances equal to zero and similarities equal to one. OTUs evolving in small populations should present distances equal to 3/4 and similarities equal to 1/4. Thus, random molecular evolution by recurrent mutation cannot yield a tree at all. The only possible tree is that produced by random fluctuations of distances according to their variances (stochastic tree). Some consequences of the model on the expected primary structure of proteins are also analyzed. There are sufficient generations for any DNA segment evolving apart during the last four hundred million years, to reach those expected base distributions.*

***Key-terms:*** *molecular evolution; randomness; recurrent mutation.*

## INTRODUCTION

Models on molecular evolution have been mostly constructed after descriptions of amino acid or nucleotide sequences from living beings (Jukes and King, 1979; Woese, 1987). From them, evolutionary theories, such as the neutral theory of evolution (Kimura, 1983), have been proposed. To found theories of evolution on data from present living beings has a severe epistemic restriction. What we see is a biased sample of what existed or was possible. A circular intellectual construction results from this procedure. Such study of evolutionary processes cannot show the most important part of evolution that led to extinction of species or individuals that did not reach our study, or to perceive how several equally probable possibilities of evolution could not be realized. The bias of models based on available information leads to overestimate random drift and overlook selection. Also, there is not an *a priori* and operational definition of randomness of evolutionary processes. The present model proposes a definition of randomness to be applied to molecular evolution and constructs a model based completely on this proposition. Some consequences of the model are contrasted with available basic information to answer the question on randomness of the evolu-

**Correspondence to:** Carlos Y Valenzuela, Departamento de Biología Celular y Genética, Facultad de Medicina, Universidad de Chile, Independencia 1027, Casilla 70061, Santiago 7, Chile. Fax: (56-2) 737-3158.

tionary process. We start with nucleic acids as they are known at present. Nucleotides are our basic units. The way from atomic elements to organic molecules is beyond the scope of this study, even though we believe that the natural process that yielded organic molecules did not occur at random. We shall not use a rigorous formal mathematical procedure in our demonstrations. A mixed biological, mathematical and current language shall be used instead, to present rigorous logical demonstrations.

## THE MODEL

### *Randomness*

To escape from the circular reasoning, we have to base our model on general properties of the hereditary material (DNA or RNA) and not on actual nucleotide sequences. Randomness, in this article, means the occurrence of equivalent possibilities or events with equal probabilities. As for example, nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) or Uracil (U) have the same probability to be found or mutate in any nucleotidic site in DNA or RNA. Any site is assumed to be independent from the others. If this is accepted, we could finish the article by concluding that, in DNA, these four bases should be found with expected probabilities equal to 0.25 at any site. Since this expectation is far from reality, we should conclude that evolution cannot be a random biotic process. We intend to show that this conclusion is not a convention, but the result of the actual properties of nucleic acids and evolutionary processes. The 0.25 probability for each base was found previously, as far as a mathematical result is concerned, by other authors (see for a revision, Weir and Basten, 1990), in a very different epistemic frame. They intended to find strategies for studying distances; but they dismissed the 0.25 value, which is the minimal limit of a similarity between two populations when the number of generations is large, because, the discrimination between populations disappears at this point. Thus, the proposition was considered unrealistic, since most known sequences do not fit this model (Arnold, 1990). Our epistemic frame is different. We need a complete random model of molecular evolution constructed independently from the known frequencies or distributions of bases in genomes. Then, we can examine exhaustively all its possible implications and test it with actual sequences. If the present known distributions do not fit the model expectations, then the conclusion should be that evolution did not occur at random, and not to search for another model built in agreement with the known present base frequencies. We will be able to propose hypotheses on the causes of such differences, only, after evaluating the differences between the random expectations and the observed distributions. This epistemic change is important because negative consequences of the incorrect use of some mathematical models applied to molecular evolution have been shown (Collen, 1994). The random occurrence of bases and their consequence on amino acid distribution in proteins have been more recently used in studies on protein folding and stability (Steipe *et al*, 1994) after applying the Boltzmann's law (Sippl, 1989). Also, a model that allows for non independence of neighbour sites has been proposed (Muse, 1995).

### *Recurrent mutation as the main evolutionary factor*

Without mutation, evolution is impossible. Selection, migration, matings and drift modify the timing of mutation diffusion, or the proportions of mutants, but they cannot counterbalance completely the process of mutants production. So, the main goal of our model is the fate of a living system under recurrent mutation. We are going to deal with mutations as a base change from a generation to another. We consider mutations that occur by any mechanism: copy errors in replication, mutation in other stages of cell cycle, chromosome changes, sexual processes, gene conversion, *etc*. In bacteria, copy errors occur with rates among $10^{-7}$ and $10^{-11}$ (Watson *et al*, 1987). Since we include all mutations, the mutation rate for a nucleotidic site can be taken as $10^{-8}$ without a large error.

### *Our biotic material*

Let us imagine a bacterium with 1,000,000 DNA base pairs. We analyze a nucleotide site whose rate of mutation (m) is $10^{-8}$ changes per reproductive cycle. To simplify, we shall refer only to one strand of DNA. At the beginning, this site is only occupied with adenine (A). This bacterium has 1,000 cycles per year. Normal populations of these bacteria include $10^{10}$ individuals. In each cycle a proportion of $10^{-8}$ sites is changed from A to guanine (G) through a transition or to C (cytosine) or T (thymine) through a transversion. First, we assume that transitions and transversions occur with equal probabilities. The proportion of the original A ($A_{or}$) should decrease until its extinction. The probability of finding A ($PA_{or}$) after the first cycle is $PA_{or} = (1 - m)$; in the second one is $(1 - m)^2$; at the *n*th cycle:

$$PA_{or} = (1 - m)^n \qquad (1)$$

This probability follows a binomial distribution with parameters m and s (s = number of sites, cycles, individuals, or original adenines) or a Poisson distribution with parameter h (expected mean of events such as muted sites, individuals, or changes of base in a site). The probability to find A in this nucleotidic site tends to zero as n increases. The original A shall be stochastically replaced by recurrent mutations. This result is independent of the population size, type of mating or drift, and it is also independent of the base we choose for the analysis. It is a well-known result in population genetics (Li, 1976).

### *The evolutionary fate of a nucleotidic site*

Is there sufficient evolutionary time for a stochastic replacement? Our bacteria are in a steady state with $10^{10}$ individuals (N). Let us calculate the fate of $A_{or}$ in 1,000,000 years. The bacteria will have $10^9$ reproductive cycles (n). According to (1):

$$PA_{or} = (1 - m)^n = 0.000045$$

We expect 45 in 1,000,000 bacteria having the original adenine. The same figure we expect to be the number of unchanged nucleotides in the 1,000,000 bases of bacteria (fixed sites). Also this result is obtained with the Poisson distribution, being $h = 10^9$ $m = 10$ (mean expected mutational events in the site). We shall refer as a cycle of replacement to the process where a number of reproductive cycles have occurred so as the expected number of mutational events equals one (h = 1); in this case a cycle of replacement includes $10^8$ reproductive cycles (1/m). In the example 10 cycles of replacement have occurred; the expected probability of occurrence of no mutation in the original A is $1/e^{10} = 0.000045$. A very important conclusion is evident. A continuous process of replacement is the rule, definitive fixation is the exception, and stochastically impossible. Since A has changed to G, C or T (and these bases have, in turn, changed to their respective three alternatives), this random molecular process should lead to the molecular polymorphism of bases in populations evolving with large number of individuals or in a set of several populations that have evolved with small population sizes. Moreover, recurrent mutation counterbalances drift deviations. Any time a base increases its frequency, more mutations are expected to occur in this base in relation to the other bases, and vice versa. While random drift is not directional, and it may increase or decrease a frequency in successive cycles, recurrent mutation shall remove any allele (base) from the population. We remark that this process is different from direct and reverse mutations. A may go to G, T or C and then back to A; but, this is not a reverse mutation. Naturally, in small populations, a transient monomorphism is the rule, but if we consider several small populations having evolved for several cycles of replacements, the polymorphism of bases at any site should be also the rule. We will analyze this type of recurrent mutation process together with the population size in examples, after answering some important features of the model. What should be the proportion of bases in the site? Is this proportion dependent on the original base?

Figure 1 shows a scheme of the process of replacement of A. In each reproductive cycle, a fraction of $10^{-8}$ As are converted to

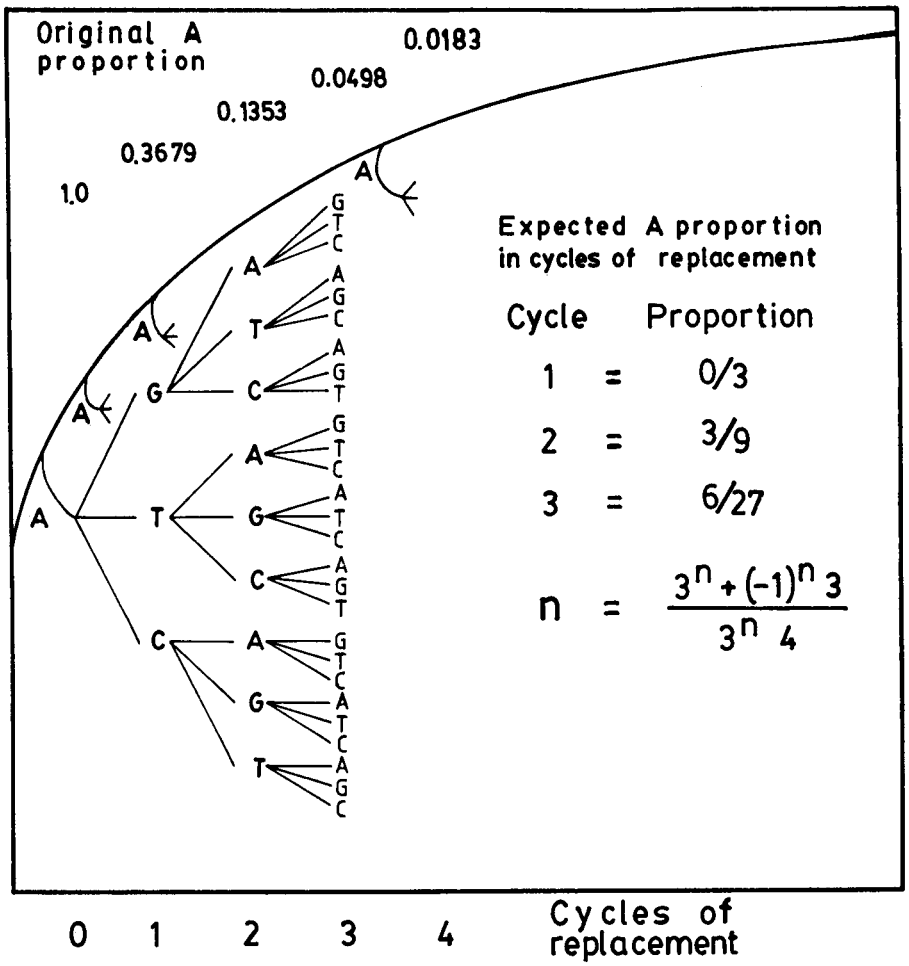


**Fig 1.** Original and expected A proportion in cycles of replacement.

G, C or T. We can calculate the proportion of A among the replaced A in successive cycles of replacement. In the origin it is 1.0, in cycle 1 is 0.0, in cycle 2 is $^3/_9 = 0.333$, in cycle 4 is 0.222 and so on. If we refer the proportion of A to four quarters, cycle 1 has 3 less As, that is $(3^1\text{-}3)/4\ (3^1) = 0.0$; cycle 2 has 3 more As, that is $(3^2 + 3)/4\ (3^2)$. The series is:

$$\text{Proportion of A among the four bases} = [3^n + 3(-1)^n]/[4(3^n)] \quad (2)$$

whose limit is $^1/_4$ when n tends to infinite. Its validity can be demonstrated by mathematical induction. The total number of possible bases in the nth generation is $3^n$; the number of As is $[3^n + 3(-1)^n]/4$; the number of As in the (n + 1)th generation is equal to the number of non-As in the nth generation, because A cannot yield A by mutation and any non A yields only one A. The number of non-A in the nth generation is $3^n - [3^n + 3\ (-1)^n]/4 = [3^n\ (4\text{-}1) - 3\ (-1)^n]/4 = [3^{n+1} + 3(-1)^{n+1}]/4$. Thus if the expression is valid for n, it is valid for n + 1 and since it is valid for n = 1 and n = 2, it should be valid for any n.

In five cycles this proportion is (243-3)/972 = 0.2469, in six, it is 0.2510. Since this conversion happens to the four bases, it is expected that, no matter the original base, we should find, in every nucleotide site each base with an equal probability (0.25). In our example, $10^6$ years yield 10 replacements; the probability to find A is 0.2500. However, this occurs with the replaced A. The total A proportion includes the unchanged original A; it is 1.0, 0.3679, 0.4686, 0.2720, 0.2776 and 0.2501 for the original, 1st, 2nd, 3rd, 4th and 10th cycles of replacement, respectively.

Now we introduce different probabilities for transitions and transversions. Transitions from A to G or from C to T or vice versa occur with probability p = 0.8. Transversions from A or G to C or T or vice versa occur with probability q/2 = 0.1 (p + q = 1). The way of a series of mutational changes of a site is then described for the binomial expansion $(p + q)^k$, **k** being the number of cycles of replacements occurred in the site.

$$(p + q)^k = p^k + kp^{k-1}q + [k(k-1)/2]p^{k-2}q^2 + \ldots kpq^{k-1} + q^k \quad (3)$$

If we consider the initial A and the final base after k cycles, a transition occurred when the final base is A or G, a transversion happened when the final base is C or T. We ask for the probability of a final transversion. In (3) a transversion occurs in any case where q has an odd exponent (an odd number of q as a factor). An even number of transversions leads A to A or G. Our problem is to obtain the limit, when k increases, of the sum of the terms in (3) with an odd number of q (Sodd-q). We subtract $(p - q)^k$ from $(p + q)^k$:

$$(p + q)^k = p^k + kp^{k-1}q + \ldots kpq^{k-1} + \ldots q^k$$
$$-(p - q)^k = p^k - kp^{k-1}q + \ldots (-1)^{k-1}kpq^{k-1} + (-1)^k q^k$$

$$= (p+q)^k - (p-q)^k = 2\,\text{Sodd-q} \quad (4)$$

Since terms with an even number of q as a factor vanish.

As p + q = 1 and 0 < p - q < 1, the limit of the difference of binomials is 1 when k increases; thus, Sodd-q, the proportion of transversions when k increases approaches $^1/_2$. C or T should be found in a proportion equal to $^1/_4$. This is another demonstration for the expected proportion of bases. We began with 100% of A and the final proportion should be $^1/_4$ for each base. Naturally this is independent of the initial base.

Any DNA (or RNA) segment that has evolved a sufficient time, will have in any nucleotidic site the four bases with equal probabilities (0.25), independently of the taxa at which it belongs. This is also true for individuals within a species (evolutionarily separated). With mutation rates equal to $10^{-8}$ this process needs four cycles of replacement to attain the 95% significant statistical level, that is 4 times $10^8$ reproductive cycles. In metazoa, the number of cell cycles needed to go from the egg to gametes should be considered to calculate the number of years per cycle of replacement. In a great deal of species, with annual life cycles, there are about 10 cell cycles to go from the egg to the first germ cell (*Caenorhabditis elegans* and fruit flies; see Watson *et al*, 1987) and 20 cell cycles to produce one million gametes from one germ cell. Thus, a species with a life cycle of 30 to 50 days will have about 300 cell reproductive cycles in a year. This is true for several invertebrates and rodents. These species, should reach the random distribution of nucleotides in two million years. This is clearly not the case. In this model, a particular base frequency may be repeated in the same site with probability equal to 1, it is only a matter of time. That is, the frequency of bases fluctuates around its expected value, but all the bases are changing in a long time. This feature of the model is completely different from models with alleles having long segments of DNA. The frequency of the alleles is expected to diverge one another in different populations with time. They are never expected to be equal after a long time.

### *The expected distribution of bases along DAN*

Since after 5 cycles of replacement the expected proportion of any base is 0.25, at any site, we can calculate the expected mean and variance of the number of A, G, T and C in a given genome, after this time (remember we are working with only one strand). Let N be the number of base pairs of the genome (1,000,000 in our case). The mean (MB), variance (VB) and standard deviation (SB) of the number of any base are:

$$MB = (^1/_4)N;\ VB = (^1/_4)(^3/_4)\,N = 3N/16;\ SB = \sqrt{3N}/4$$

In our bacteria: MB = 250,000; VB = 187,500; SB = 433.01.

Bases are expected to be distributed along with DNA according to a series of random runs with replacement (Feller, 1968). Down or upstream of a given base, any base should be found with probability 0.25. The independence of any site from the others (assumption of randomness) implies that the

covariance between two sites is zero. Or, the conditional probability for a base to be found in a site given that another base is found in any other site is equal to its unconditional probability (0.25).

### *Randomness of cold or hot spots*

Our definition of randomness allows to test whether hot spots occur at random. Since the distribution of the number of mutations per site follows a Poisson distribution (in the model), the expected number of mutations per site and reproductive cycle and their probabilities can be estimated. In $10^9$ bacteria, 10 mutated individuals per site are expected to occur. Statistically significant hot spots should be considered if 16 or more mutations are found in the site (P = 0.04875); 20 or more mutations are found with probability equal to 0.003464. Similar probabilities are obtained for cold spots with 4 and 2 mutations respectively. Thus, if the rate of mutation is randomly distributed and we expect 10 mutations in a site, then, the occurrence of twice or one fifth of this figure is sufficient to consider it as a hot or a cold spot respectively. However, we have not taken into account the possible different longitudinal properties of DNA (RNA), or we have assumed isotropy in the occurrence of mutations (equal probabilities for all the sites) along with the hereditary material. If some longitudinal properties of DNA depend on particular sequences, and the rate of mutation is one of these properties, there should be hot, normal and cold spots with different sequences and with very different rate of mutations. Evolution by recurrent mutation should be different for these DNA segments. Hot spots should auto anneal themselves faster than normal spots, and normal spots faster than cold spots. There should be a deviation towards coldspotness along with the evolutionary process, unless normal and hot spots would be maintained by selection or another evolutionary factor.

### *Nucleotide random evolution and general protein properties*

We will not examine randomness in the origin of the genetic code. The most plausible hypothesis is that the genetic code did not evolve as a random process. We are going to accept the existence of the known genetic code and study the properties of proteins under the assumption that random evolution of the hereditary material, as we have proposed, has really occurred.

A random distribution of nucleotides implies a random distribution of triplets. Any triplet should be equally present into DNA. This distribution yields 61 triplets for amino acid specification into proteins and 3 triplets of termination (Watson *et al*, 1987). The expected distribution of specific amino acid into proteins should be given for the equivalence of triplets with amino acids. Arginine, serine or leucine should be the most represented amino acids into proteins (6 triplets) and tryptophan or methionine (1 triplet each one) the fewer ones. With these expectations (number of triplets per amino acid) we can test the factual amino acid composition of proteins. As for example, bovine chymotrypsinogen (Lehninger, 1982) has 245 amino acids, only 4 arginines (expected 24.1); this occurs with probability less than 0.0001, yet significant considering 20 amino acids (see Table I). Statistical protein parameters can be estimated. The mean number of amino acids a protein has and its variance are necessary to test actual proteins. If there is a random distribution of the three termination triplets among the 61 meaningful triplets, the mean and variance of the expected number of aminoacids in a protein can be obtained. Let us walk forward from a termination triplet; let p be the probability to find an amino acid triplet and q a termination triplet (p = 61/64, q = 3/64; p + q = 1). The number of amino acids we can find (excluding the possibility to find a termination triplet in the first step) is described by the geometric series:

$$S_0 = pq + p^2q + p^3q + \ldots p^nq = q(p + p^2 + p^3 + \ldots p^n)$$

$S_0$ tends to p when n tends to infinite. The expected number of amino acids a protein has is obtained from the series:

$$S_1 = q\,(1p + 2p^2 + 3p^3 + \ldots + np^n) = p/(1-p) \text{ when n increases}$$

The expected mean number [$E(S_1)$] of amino acids is $S_1/S_0 = 1/(1-p)$. Since p is 61/64,

TABLE I

Aminoacid distribution in four eukaryotic proteins

| NUMBER OF CODONS | HUMAN CYTOCHROME-C | | BOVINE CHYMOTRYPSINOGEN | | BOVINE PROINSULINE | | BOVINE RIBONUCLEASE | |
|---|---|---|---|---|---|---|---|---|
| | Obs | Exp | Obs | Exp | Obs | Exp | Obs | Exp |
| Ala 4 | 6 | 6.8 | 22 | 16.1 | 6 | 5.3 | 12 | 8.1 |
| Arg 6 | 2 | 10.2** | 4 | 24.1*** | 4 | 8.0 | 4 | 12.2** |
| Asn 2 | 5 | 3.4 | 15 | 8.0* | 3 | 2.7 | 10 | 4.1** |
| Asp 2 | 3 | 3.4 | 8 | 8.0 | 0 | 2.7 | 5 | 4.1 |
| Cys 2 | 2 | 3.4 | 10 | 8.0 | 6 | 2.7* | 8 | 4.1 |
| Gln 2 | 2 | 3.4 | 10 | 8.0 | 5 | 2.7 | 7 | 4.1 |
| Glu 2 | 8 | 3.4** | 5 | 8.0 | 8 | 2.7** | 5 | 4.1 |
| Gly 4 | 13 | 6.8** | 23 | 16.1 | 12 | 5.3** | 3 | 8.1* |
| His 2 | 3 | 3.4 | 2 | 8.0* | 2 | 2.7 | 4 | 4.1 |
| Ile 3 | 8 | 5.1 | 10 | 12.0 | 1 | 4.0 | 3 | 6.1 |
| Leu 6 | 6 | 10.2 | 19 | 24.1 | 9 | 8.0 | 2 | 12.2** |
| Lys 2 | 18 | 3.4*** | 14 | 8.0* | 2 | 2.7 | 10 | 4.1** |
| Met 1 | 3 | 1.7 | 2 | 4.0 | 0 | 1.3 | 4 | 2.0 |
| Phe 2 | 3 | 3.4 | 6 | 8.0 | 3 | 2,7 | 3 | 4.1 |
| Pro 4 | 4 | 6.8 | 9 | 16.1 | 5 | 5.3 | 4 | 8.1 |
| Ser 6 | 2 | 10.2** | 28 | 24.1 | 3 | 8.0 | 15 | 12.2 |
| Thr 4 | 7 | 6.8 | 23 | 16.1 | 1 | 5.3 | 10 | 8.1 |
| Trp 1 | 1 | 1.7 | 8 | 4.0* | 0 | 1.3 | 0 | 2.0 |
| Tyr 2 | 5 | 3.4 | 4 | 8.0 | 4 | 2.7 | 6 | 4.1 |
| Val 4 | 3 | 6.8 | 23 | 16.1 | 7 | 5.3 | 9 | 8.1 |
| TOTAL 61 | 104 | | 245 | | 81 | | 124 | |

Amino acid with the ancient nomenclature. **Obs**, observed; **Exp**, expected.
* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. (From Lehninger, 1982).

$E(S_1) = {}^{64}/_{3} = 21.33$. The weighed sum of squares of the number of amino acids is given by the series:

$S_2 = q(1^2p + 2^2p^2 + 3^2p^3 + \ldots n^2p^n) = p(1 + p)/(1 - p)^2$ when n goes to infinite. The expected $S_2$ $[E(S_2)]$ is $S_2/S_0 = (1 + p)/(1 - p)^2$. The expected variance of the number of triplets is $E(S_2) - E(S_1)^2$.

This variance is: $VAR = p/(1 - p)^2$. In this case, it is 61 x 64/9 = 433.78. The standard deviation (SD) is 20.83. It is important to remark that in this model the smaller the number of amino acids, the higher the probability of the protein. The most frequent polypeptide should have 1 aminoacid, the second one should have two, and so on. These expected figures are so far from figures of actual proteins that no statistical tests are needed.

## DISCUSSION

The main features of our model are: 1) The expected distribution of nucleotide bases, in a site, for any set of species or individuals evolutionarily separated for a big number of generations is an isotetranomial distribution with four parameters equal to 0.25 for every base; 2) The covariance between the bases distribution in any pair of sites is expected to be zero; bases are randomly distributed along with DNA or RNA. These are conditions far from those found in living beings. Since the features of the model are well established and facts disagree with them, it is necessary to find the reasons for this disagreement. To discuss this subject we need first two extreme imaginary models with recurrent mutation and random drift or population size considered together.

I) Our population of bacteria came from one individual which in ten days yielded 33.219 generations and reached the steady state with $10^{10}$ individuals (very small genetic drift). In any reproductive cycle this figure is doubled, then, random death of half of this population reduces the figure to the steady state value. Also any reproductive cycle yields 200 mutants (A to G, C or T), from which, 100 are removed by random death. These mutants initiate their random

increase or decrease of their frequency in the population. With such a big number of individuals in the population the probability of random fluctuations of their frequency is near zero. In the next generation 100 new mutants are added which initiate their random drift way. After 1,000,000 years (10 replacement cycles) few original As are found and an equally distributed proportion of A, G, C and T is expected at any site. Random drift may only increase the variance of the frequency of the bases but not their expected proportions. Naturally, the number of first mutants decreases as the number of original As decreases, and the number of second, third and higher order mutants increases as generations occur. This picture has nothing to do with actual bacterial populations, which are, base to base, mostly identical (Woese, 1987).

II) A bacterium yielded two individuals. Random death removes one of them to produce a steady state with one bacterium (maximal genetic drift). Among the two individuals after replication, a mutant is very probable to appear (0.98) in 400,000 years, and to be selected as one of the two individuals in 2,400,000 years [$(^1/_2)^6$ (0.98) = 0.965]. In 2,400,000 years a new mutant replaces the original base with a probability higher than 0.96. However, if we examined a great deal of populations or species that have evolved similarly, we should find the four bases in every site with the same probability (0.25), because the replacement should lead to any base with the same probability as we have demonstrated. Thus, drift as in the most extreme bottle neck effect cannot explain the disagreement between this model and actual bacterial populations.

Thus, a collection of populations isolated for more than three cycles of replacement, no matter their sizes, should present an isoprobable distribution of bases at any nucleotide site. Since this is not found, an important component of evolution did not occur at random.

### *Expected molecular distances, similarities and phylogenetic trees*

The expected distances, similarities and phylogenetic trees depend on the size of populations under evolution by recurrent mutation, as the previous analysis showed. If we compare two taxa (operational taxonomic units = OTUs) evolving with big populations, as currently bacteria have done, the expected distance is zero and similarity is 1 or complete. No tree is possible for a set of OTUs. This occurs, because for every site the expected situation is a tetramorphism with the four bases occurring with probability equal to 0.25. Naturally, if we include the variance of the bases frequency, we should have a stochastic tree where similarities and distances would vary according to their variances. Random drift that is related to the historical variation in the population size, should only increase this variance. Now if populations under comparison have evolved with small sizes, and the effect of random drift is big, it is expected that at any site only one base be present. However, since there is not correlation among sites, the distance between two OTUs should be 0.75 and their similarity 0.25. Again, no tree is possible, with the exception of a stochastic tree. If populations were intermediate in size during their evolution, the expected situation is a mosaic with polymorphic and monomorphic sites, but the expected distances and similarities should be equal for every pair of OTUs separated by the same number of reproductive cycles. This situation could be that found in most actual populations, but this is not the case. Polymorphic and monomorphic sites should be randomly distributed along both genomes; this expectation is far from actual similarities and differences found in comparisons of two genomes. Thus, actual trees are expected only in OTUs separated by less than 3 cycles of replacement.

The expected distribution of amino acids into proteins is given by the random distribution of bases in triplets and the number of triplets any amino acid has. With this expectation we can test proteins and assess the distance from the actual protein to the random expectation of amino acid into this protein. Table I shows a comparison for four proteins. Statistical significance was calculated by a $\chi^2$ test with one degree of freedom, when expected values were over 5 and by a Poisson distribution elsewhere.

Unexpectedly, most amino acid frequencies agree with random expected frequencies. Four or five amino acids in each protein did not agree with random expectations. These amino acids are, mostly, the same for the four proteins. Arginine was less frequent than expected in the four proteins, while lysine was more frequent than expected in three proteins. We present this analysis only as an example of the kind of information it can provide. Also, our model predicts that the smaller the protein the higher its frequency. Most known proteins have more than 20 amino acids; however, this model indicates that the most frequent expected proteins should have less than 20 amino acids. The analysis of this disagreement could show interesting evolutionary decisions. A third kind of analyses our model allows is the distribution of nucleotides in nucleic acids or amino acid in proteins along the respective primary structure (runs). As for example, in a polypeptide with 10 amino acids, 4 of them being glutamic acid and 6 other amino acids (with 2 codons), we expect that the 4 Glus be in tandem in 7 of 210 possibilities (P = 0.033). The same analysis can be performed on DNA or RNA. The distribution of codons within those ones for the same amino acid can be studied similarly; but, it requires an analysis beyond the scope of this article.

We avoided to relate this model with those of allele distributions, for which there is an extensive literature. As in our model, the equilibrium between mutation and drift for a set of alleles is reached with frequencies that do not depend on the original ones (Kimura and Crow, 1970). However, those models do not make explicit assumptions on the randomness of the actual alleles among all the possible ones. Thus, they implicitly assume that the original and actual alleles are a random sample of all the possible alleles. The most important evolutionary problem is not whether monomorphic or polymorphic sites or loci fit random drift or selection models, but to know whether they are a biased or unbiased sample of all the monomorphic or polymorphic possibilities; and whether this occurred by drift, selection or any other set of evolutionary factors. For any known DNA segment with more than 20 bases there are more than $10^{12}$ alleles, so, it is practically impossible to decide the randomness of a set of alleles. Therefore, assuming or dealing with a set of actual alleles as a random set of all the possible alleles is an epistemic error, which leads to a circular reasoning, because models risk to be created to agree with the present data, which were assumed to be an unbiased set of possibilities. Moreover, models on allele evolution assume that the set of actual alleles is the only one which can be produced, because a new allele appears with a negligible probability; thus, these models allow mutation only among those actual alleles (Kimura, 1983; Jacquard, 1974) These restrictions are not present when dealing with abstract four bases which are the whole universe of possibilities. On the other hand, the only acceptable measurement of the mutation rate is that one made directly in each cell cycle along all the DNA (RNA) molecule. Our model presents a new viewpoint to examine actual molecular sequences to decide the randomness or selectivity of their evolution. From this model a new insight for the neutral - selective evolution debate can be proposed. Random mutation and drift explore the organizational or structural possibilities of living beings. When a stable living organization (clone, species or taxa) is reached, it is fixed and maintained mainly by selection and ecological coadaptation.

## REFERENCES

ARNOLD J (1990) Discussion on the paper by BS Weir and CJ Basten. Biometrics 46: 572-573

COLLEN K (1994) A test of the Markovian model of DNA evolution. Biometrics 50: 653-664

FELLER WJ (1968) An Introduction to Probability Theory and Its Applications. New York: John Wiley & Sons

JACQUARD A (1974) The Genetic Structure of Populations. New York, Heidelberg, Berlin: Springer-Verlag

JUKES TH, KING JL (1979) Evolutionary nucleotide replacements in DNA. Nature 281: 605-606

KIMURA M (1983) The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press

KIMURA M, CROW JF (1970) An Introduction to Population Genetics Theory. New York: Harper & Row

LEHNINGER AL (1982) Principles of Biochemistry. New York: Worth Publishers, Inc.

LI CC (1976) First Course in Population Genetics. Pacific Grove, California: Boxwood Press

MUSE SP (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics 139: 1429-1439

SIPPL MJ (1990) Calculation of conformational ensembles from potentials of mean force. J Mol Biol 213: 859-883

STEIPE B, SCHILLER B, PLÜCKTHUM, STEINBACHER S (1994) Sequence statistics reliable predict stabilizing mutations in a protein domain. J Mol Biol 240: 188-192

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

WEIR BS, BASTEN CJ (1990) Sampling strategies for distances between sequences. Biometrics 46: 551-582

WOESE CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271