J. Mol. Evol. 13, 27–34 (1979)

Journal of
Molecular Evolution

Are Evolutionary Rates Really Variable?

John H. Gillespie[1] and Charles H. Langley[2]

[1]Department of Biology, University of Pennsylvania, Philadelphia, PA. 19104
[2]National Institute of Environmental Health Sciences, Research Triangle Park, N.C. 27701, USA

Summary. Langley and Fitch (1974, 1976) have shown that the pattern of nucleotide substitutions in proteins is inconsistent with a Poisson process with constant rate. From this they conclude that the rate is temporally heterogeneous. It is pointed out in this note that a process which is temporally homogeneous but not a Poisson process is compatible with the data if the coefficient of variation of the time between substitutions is around 1.63. Furthermore, theoretical analysis of samples from neutral phylogenies shows that these samples should not appear to be samples from a Poisson process, but should deviate from a Poisson process in the same direction, though perhaps not to the same extent, as do the data.

Introduction

In two recent papers Langley and Fitch (1973, 1974) concluded that the evolutionary rates of four proteins in vertebrates are temporally variable. This conclusion has been generally accepted and is routinely cited in the secondary literature (e.g. Dobzhansky et al., 1977, p. 311). In this note we will show that the statistical analysis of Langley and Fitch (L-F) also admits an alternative interpretation: that evolutionary rates are temporally constant, but that the substitution process is more complex than the Poisson process assumed in the L-F analysis. This new interpretation motivates a reexamination of the generally accepted view that the neutral allele model of protein evolution (Kimura, 1968) will yield data which are compatible with a Poisson process. Such a reexamination shows that samples from a neutral phylogeny may not appear to be samples from a Poisson process; instead, certain aspects of the sample may deviate from a Poisson process in the same direction as do the data analyzed by Langley and Fitch. This suggests that the presently available data may not be incompatible with a constant-rate neutral allele model of evolution.

The Langley-Fitch Analysis

The L-F analysis of protein sequence data is based on a two-part null hypothesis: (1) that the rate of substitution of nucleotides is constant within a protein, and (2), that

0022-2844/79/0013/0027/$ 01.60

the substitution process is a Poisson process. Langley and Fitch rejected this null hypothesis using a likelihood-ratio technique and concluded that evolutionary rates are not constant within a protein. That is, they chose to reject only part (1) of the null hypothesis. There is nothing in their procedure to favor the rejection of part (1) over part (2), or, for that matter, over (1) and (2) jointly. To reject part (2) but not part (1) would be equivalent to claiming that the substitution process moves at a constant rate, but that the process is a stationary point process which is more complex than the Poisson process. We will show in this section that this viewpoint appears to be compatible with the data, although a final judgement must be reserved until a proper analysis of the data is achieved. We begin with a general discussion of stationary point processes.

Substitutions may be viewed as occurring at definite points in time. Rather than discussing directly the time of occurrence of these substitutions, it is more convenient to discuss the time intervals between substitutions. Thus if substitutions occurred at times $t_1^*, t_2^*, t_3^*, \ldots$, then the times between substitutions make up the sequence t_1, t_2, t_3, where $t_i = t_{i+1}^* - t_i^*$.

Let the sequence of random variables $T_1, T_2, \ldots$ represent such times between successive substitutions. If the collection $T_1, T_2, \ldots$ is made up of independent, identically distributed random variables then the substitution or point process is a renewal process. In the special case where the T_i are exponentially distributed, then the process is a Poisson process. In a L-F type of procedure it would be possible to use a renewal process, say one where the T_i are gamma-distributed, instead of a Poisson process for the null hypothesis. It is quite possible that this null hypothesis would be compatible with the data given the richer parameter space. If this were to happen, then we could conclude that the data are compatible with a constant-rate, non-Poisson process.

Rather than pursue the renewal approach, we will go one step further in generality. Let the sequence of inter-substitution times be a strictly stationary sequence. That is, let the distribution of the sequence

$$T_{n_1}, T_{n_2}, \ldots T_{n_j}$$

be the same as the distribution of the sequence

$$T_{n_{1+k}}, T_{n_{2+k}}, \ldots, T_{n_{j+k}} \quad .$$

This is tantamount to assuming that the process is translation-invariant. Obviously all of the moments of the process are temporally constant. In particular, *the rate of the process, $(ET_i)^{-1}$ is constant.* The second order moments of stationary processes are given by the autocovariance function

$$\gamma_k = \text{Covariance}\,(T_i, T_{i+k})$$

Obviously

$$\gamma_0 = \text{VAR}\,(T_i) = \sigma^2$$

Let the mean inter-substitution time be

$$\mu = ET_i$$

The theory of stationary point processes allows one to express the distribution of the number of substitutions which occur in a fixed time t in terms of the inter-substitution times. For large t the distribution of the number of substitutions is a normal distribution with mean

$$E(N_t) \sim t/\mu$$

and variance

$$VAR(N_t) \sim \sigma_a^2 t/\mu^3$$

where

$$\sigma_a^2 = \sigma^2 (1 + 2 \sum_{i=1}^{\infty} \gamma_i)$$

is a measure of the autovariance of the process. These relationships between the moments of the number of substitutions and the moments of the inter-substitution times may be found in Cox and Miller (1965, p. 361). A measure of the "clumpedness" of the process is given by

$$\kappa = \frac{VAR(N_t)}{E(N_t)} = \frac{\sigma_a^2}{\mu^2}$$

In the case of a Poisson process, where the T_i are exponentially distributed, κ is one. For renewal processes, $\gamma_i = 0$, $i \neq 0$, so

$$\kappa = \frac{\sigma^2}{\mu^2} \quad .$$

We are now in a position to roughly analyze what would happen if we used a stationary point process instead of a Poisson process as the null hypothesis in a Langley-Fitch type analysis. We will be looking to see, therefore, if a constant rate but non-Poisson process is compatible with the protein data. The Chi-Square statistic used by Langley and Fitch for a single leg of a phylogeny is

$$\chi^2 = \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} = \frac{(N_t - E(N_t))^2}{E(N_t)}$$

where Obs is the observed number of substitutions (inferred from the minimal phyletic distance procedure) and Exp is the expected number of substitutions. Using our previous observations on the moments of N_t write this, for large t, as

$$\chi^2 = \frac{(N_t - t/\mu)^2}{t/\mu}$$

This statistic may be written, using our expression for the variance, as

$$\chi^2 = \frac{\sigma_a^2}{\mu^2} \left(\frac{N_t - t/\mu}{\sqrt{\frac{\sigma_a^2 t}{\mu^3}}} \right)^2 = \frac{\sigma_a^2}{\mu^2} \chi_1^2$$

which is a constant, σ_a^2/μ^2, times a Chi-Square random variable with one degree of freedom χ_1^2. This is true because the squared quantity is a normal random variable with mean zero and variance one. Note that in the Poisson case $\sigma_a^2 = \mu^2$ and

$$\chi^2 = \chi_1^2$$

By our stationarity assumption we assume that μ and σ_a^2 are the same for all legs of a given protein. Thus the χ^2 for the protein is simply the sum of the χ^2's for each leg. This sum may be written

$$\frac{\sigma_a^2}{\mu^2} \chi_\nu^2$$

where ν is the degree of freedom and will, in general, be less than the number of legs if estimation procedures are used to determine times of splitting and for the various parameters.

We see at this juncture that the χ^2 statistic used by Langley and Fitch will not be Chi-Square distributed if the substitution process is non-Poisson. It will differ by a constant multiple, σ_a^2/μ^2. A rough idea of the magnitude of σ_a^2/μ^2 required to account for the high Chi-Square value in Langley and Fitch may be obtained by asking for the value of σ_a^2/μ^2 which will make $\sigma_a^2 \chi^2{}_\nu/\mu^2$ equal to the value of 82.4 with $\nu = 31$ degrees of freedom given in Langley and Fitch (1976). This is the among legs, over protein comparison which is the appropriate one for our purpose. Obviously if

$$\sigma_a^2/\mu^2 \approx 2.65$$

or, equivalently, if the "coefficient of variation", σ_a/μ, of the substitution process is

$$\sigma_a/\mu \approx 1.63$$

we can easily account for the large Chi-Square presented by Langley and Fitch.

To reiterate: we have shown that the Langley-Fitch data is roughly compatible with a stationary process which, because $\sigma_a/\mu > 1$, moves at a constant rate but tends to appear clumped. Thus, the occurrence of a substitution in a particular leg makes it more likely that further substitutions will occur in the leg. We must emphasize that this is an alternative view of the data. This interpretation is in no way preferable to the interpretation that the large χ^2 is due to variable rates of substitutions. Distinguishing between these two views would probably be impossible with the presently available data. It would also prove difficult because the class of Poisson processes with stationary coefficients is very large and is, from practical point of view, almost indistinguishable from the class of stationary point processes.

Sampling from Neutral Phylogenies

The analysis of the previous section naturally poses the question: what is the nature of the substitution process under the neutral allele model? The answer to the populational aspect of this question seems quite complex, but the sampling aspect may be rather simply described. In this section we will show that samples from neutral phylogenies are neither stationary, nor Poisson, and that the deviations from Poisson are in the same direction as the data. The problems with sampling neutral phylogenies all grow out of the existence of extensive polymorphism. The nature of these problems will emerge from the following considerations.

Consider an infinite-sites neutral allele model with no recombination between the sites. This model was thoroughly investigated by Watterson (1975). We imagine that a locus consists of an infinite number of nucleotides each of which will mutate no more than one time. The number of mutations per replication is assumed to be Poisson distributed with an average number of u mutations per replication. Although this model is biologically absurd, its behavior will be quite similar to that of a finite-sites model with recurrent mutation for short time intervals. As long as the number of segregating sites is considerably less than the total number of sites, we may expect the infinite-sites model to be a reasonable approximation.

Suppose we are in the enviable position of having a large number of species pairs with the two species in each pair having been genetically isolated exactly t years ago. Suppose that we sample one genome from each of the two species in each pair and count the number of substitutions which have occurred between the members of each species pair. What will be the distribution of substitutions? This sampling procedure is the one which is used in most sequence work, where one genome is characteristically presented for each species.

In the appendix we show that the distribution of the number of substitutions may be written as the sum of a Poisson and a geometric random variable. This structure results from the fact that the genomes sampled from each of the two species must have a common ancestor, but the common ancestor will have existed at some random time, T, before t. That is, knowledge of t does not imply knowledge of the exact split time of the genomes which are sampled. For large population sizes, N, and small mutation rates, u, we have the moments of the number of substitutions, N_t as:

$$E(N_t) = 2ut + \theta = \theta(1 + \alpha)$$

$$VAR(N_t) = 2ut + \theta(1 + \theta) = \theta(1 + \alpha + \theta)$$

where $\theta = 4\ N\ u$ and $\alpha = t/2\ N$. This latter quantity is the time of isolation measured in units of 2 N generations. The ratio of the variance to the mean is

$$\rho = \frac{VAR(N_t)}{E(N_t)} = 1 + \frac{\theta}{1 + \alpha} = 1 + \frac{\theta^2}{E(N_t)}$$

which is always larger than one. From the above we make two critical observations:

(1) The mean number of substitutions in samples from this simple phylogeny are not directly proportional to t. Thus samples should not appear stationary if t is used as the split time. This problem crops up because of genetic polymorphism and will be most serious for large θ.

(2) Since $\rho > 1$, samples will appear clumped. That is, we expect samples to deviate from Poisson in the direction of having too large a variance to mean ratio. This is, of course, what we observed in the previous section.

Does this indicate that the explanation for the high Chi-Square of the Langley-Fitch analysis resides in the peculiarities of sampling a neutral phylogeny? A definitive answer can only come from reexamination of the data with the proper null hypothesis. We can, however, make the following observation. From the above

$$\kappa = 1 + \frac{\theta^2}{E(N_t)}$$

In the data in Langley and Fitch for β - hemoglobins, for example, the average number of substitutions per leg is approximately 10. In order to achieve a variance to mean ratio of about 2.65 as seen in the data we would require a θ satisfying

$$1 + \frac{\theta^2}{10} = 2.65$$

or $\theta = 4.06$. This is higher than the estimates of θ based on typical electronphoretic data, but given the problems of non-identification such a θ value is not unreasonable.

These observations indicate that the rejection of the null hypothesis in the L-F analysis does not provide compelling evidence against the constant-rate neutral allele model. A proper analysis based on a neutral-allele null hypothesis would be quite complex, but the outcome should prove quite valuable.

Disscussion

There are two quite distinct aspects to this paper. The first is a purely empirical point that the analysis of Langley and Fitch does not exclude the constancy of evolutionary rates of nucleotides. The second is a theoretical point that samples from neutral phylogenies will not, in general, lead to Poisson-distributed numbers of substitutions. Overall, we would suggest that a good deal more work is required to properly assess the significance of the data on protein evolution.

Further work is also needed on models of protein evolution which incorporate natural selection in addition to mutation and genetic drift. If, for example, the random environment model of Gillespie (1977) is superimposed on the neutral model the resultant process, which will be a stationary stochastic process, may show a higher variance to mean ratio than the neutral model. This combined model may, in fact, provide the best fit to the data.

Appendix

Let t be the time of the split between the species pairs. Draw one chromosome at random from each of the two species pairs. The most recent common ancestor of these two chromosomes will have lived at some random time T *before* t. Obviously T is geometrically distributed:

$$\Pr[T = i] = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^i \quad , i = 0, 1, 2, \ldots..$$

Thus the common ancestor chromosome lived t + T generations ago. Each generation the two lineages accumulate mutations. There will be 2(t + T) replications separating the two sampled cromosomes. The total number of mutations during these 2(t + T) replications will be

$$S = X_1 + X_2 + \ldots + X_{2(t+T)}$$

where X_i is a Poisson random variable with mean u. We may write this as the sum

$$S = S_1 + S_2 = (X_1 + X_2 + \ldots + X_{2t}) + (X_1 + X_2 + \ldots + X_{2T})$$

S_1 is Poisson distributed with mean 2ut and probability generating function

$$f(s) = e^{2ut(s-1)}$$

The probability generating function of 2T is

$$\frac{\lambda}{1-s^2\,(1-\lambda)} \quad , \lambda = \frac{1}{2N}$$

so the p.g.f. of S_2 is the compound of a Poisson and a geometric:

$$g(s) = \frac{\lambda}{1-e^{2u(s-1)}\,(1-\lambda)}$$

The p.g.f. of S is the product of f and g:

$$h(s) = \frac{\lambda e^{2ut(s-1)}}{1-e^{2u(s-1)}\,(1-\lambda)}$$

If t and N → ∞ and u → 0 in such a way that $\theta = 4Nu$ and $\alpha = t/2N$ stay constant, h(s) will approach

$$h(s) \sim \frac{e^{\theta\alpha(s-1)}}{1-\theta\ (s-1)}$$

$$= \frac{F}{1-(1-F)s}\ e^{\theta\alpha(s-1)}$$

where $F = (1 + \theta)^{-1}$ is the expected homozygosity of the population.
Notice that final distribution is the sum of two independent random variables. The first is a geometric random variable with mean θ and variance $\theta\,(1 + \theta)$. This random variable represents the contribution from the time before t. The second random variable is a Poisson random variable with mean and variance $\alpha\theta$. Thus the overall mean and variance in the number of substitutions are

$$EN_t = \theta(1 + \alpha)$$

$$VAR\ N_t = \theta(1 + \alpha + \theta) \quad .$$

Note that if we set t = o we get Watterson's distribution of the number of sites which differ between a pair of randomly chosen individuals from a random mating population. This is a geometric distribution with mean θ. The probability that the two chromosomes are the same is obtained by setting s = o:

$$h(o) = F$$

as is well known.

References

Cox, D.R., Miller, H.D, (1965). The Theory of Stochastic Processes. Methuen.

Dobzhansky, T., Ayala, F.J., Stebbins, G.L., Valentine, J.W. (1977) Evolution San Francisco: Freeman and Co.

Gillespie, J.H. (1977). Sampling theory for alleles in a random environment. Nature **266**, 443–445

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature **217**, 624–626

Langley, C.H., Fitch, W.M. (1973). The constancy of evolution. A statistical analysis of the α and β hemoglobins, cytochrome c and fibrinopeptide A. In: Genetic Structure of Populations (N.E. Morton, ed.) pp. 246–262, Honolulu: Univ. Press of Hawaii

Langley, C.H., Fitch, W.M. (1974). An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3, 161–177

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**, 356–376

Received April 18/ Revised December 23, 1978