JOURNAL OF MOLECULAR EVOLUTION

Estimation of the Number of Amino Acid Substitutions Per Site When the Substitution Rate Varies Among Sites

Nick V. Grishin

Department of Pharmacology, University of Texas Southwestern Medical School, Dallas, TX, 75235, USA

Received: 24 March 1995 / Accepted: 15 May 1995

Abstract. A general model for estimating the number of amino acid substitutions per site (d) from the fraction of identical residues between two sequences (q) is proposed. The well-known Poisson-correction formula $q = e^{-d}$ corresponds to a site-independent and amino-acid-independent substitution rate. Equation $q = (1 - e^{-2d})/2d$, derived for the case of substitution rates that are site-independent, but vary among amino acids, approximates closely the empirical method, suggested by Dayhoff et al. (1978). Equation $q = 1/(1 + d)$ describes the case of substitution rates that are amino acid-independent but vary among sites. Lastly, equation $q = [\ln(1 + 2d)]/2d$ accounts for the general case where substitution rates can differ for both amino acids and sites.

Key words: Amino acid substitutions — Evolutionary distance — PAM scale — Dayhoff et al.'s distance — Gamma distance

Introduction

For evolutionary tree construction it is necessary to transform the number of identical residues (nonadditive characteristics) between two aligned protein sequences into the number of substitutions (distance, additive characteristics) which were likely to occur between these sequences. The simplest estimation is based on the Poisson-correction method, first used by Zuckerkandl and Pauling (1965): $d = -\ln q$, where q is the fraction of identical residues and d is a mean number of amino acid substitutions per site between two protein sequences. This estimation, however, assumes all amino acids and all sites to be equally changeable.

To avoid this limitation, Dayhoff et al. (1972, 1978) introduced an empirical method of estimating distances which multiplies a "mutation probability matrix" for various evolutionary distances. Dayhoff's method allows different amino acids to have different substitution rates, while all sites are still equally changeable. Thus the substitution rate of a site is equal to the substitution rate of the amino acid that currently occupies the site. The results were presented in a form of a conversion table (Dayhoff et al. 1978: Table 36, p. 375).

For the case of substitution rates different among sites, Ota and Nei (1994) proposed an equation based on the assumption that the rate of amino acid substitution varies approximately according to the gamma distribution. This assumption was suggested by Uzzell and Corbin (1971), but did not have a strong theoretical basis. The formulas obtained included an additional parameter that was estimated by fitting the equation to the points from the table given by Dayhoff et al. (1978: Table 36, p. 375). This estimation does not seem to be appropriate because Dayhoff's model uses different substitution rates for different amino acids, but those rates are site-independent. On the other hand, the assumption of Uzzell and Corbin (1971) takes into account rate differences for sites, and not amino acids. Nevertheless, a reasonable agreement with the Dayhoff data points was obtained for the equation

$$d = a\left(q^{-\frac{1}{a}} - 1\right) \tag{1}$$

using $a = 2.0290$ for q between 0.15 and 1.00. (The square deviation sum is 0.143.)

In this paper an equation that gives a better approximation of Dayhoff's points than equation (1) is derived (the square deviation sum is 0.006):

$$q = \frac{1 - e^{-2d}}{2d} \tag{2}$$

The derivation does not use Dayhoff's points to find the best fit. Unlike the method of Dayhoff, this method allows a simple calculation of variances and covariances.

Additionally, under some assumptions the equation is derived, which allows for the substitution rate to vary among sites, but not amino acids:

$$q = \frac{1}{1 + d} \tag{3}$$

and for the general case, when the substitution rate varies for both amino acids and sites:

$$q = \frac{\ln(1 + 2d)}{2d} \tag{4}$$

All derivations are based on the same general approach and each of them is a special case for a special selection of distributions.

General Assumptions

Traditionally, amino acid substitutions at each site are treated as mutually independent, finite state, continuous time, homogeneous Markov processes (Takacs 1966). The following assumptions are made:

1. For each site, the probability that in a very short time more than one substitution occurs is very small compared to the probability that exactly one substitution occurs.
2. Each amino acid at each site is characterized by a substitution rate (infinitesimal transition probability) that is a constant. Thus the substitution rate does not depend on evolutionary lineages, on amino acids at the other sites, on amino acids that were and will be at this site, and is time-independent.
3. The substitution process is at equilibrium: that means amino acid frequencies and all other probability distributions are time-independent and are the same for all sequences.

General Consideration

Consider an amino acid sequence of n sites, undergoing single amino acid substitutions. Each amino acid A_i ($i = 1, \ldots, 20$) at each site j ($j = 1, \ldots, n$) is characterized by its substitution rate λ_{ij}. Assumptions 1 and 2 imply that the probability of the site j with amino acid A_i being unchanged over time t is $e^{-\lambda_{ij}t}$. We can consider the distributions of substitution rates among sites for each amino acid A_i. Assumption 3 implies that these distributions are time-independent. Let $\rho_i(\lambda)d\lambda$ be the probability that a randomly chosen site among those occupied by amino acid A_i has a substitution rate in the range from λ to $\lambda + d\lambda$. Then the probability of a site with amino acid A_i being unchanged over time t is $\int_0^{+\infty} \rho_i(\lambda)e^{-\lambda t}d\lambda$. Let f_i be the probability that a randomly chosen site in the sequence is occupied by amino acid A_i. For a large n, the probabilities f_i are close to the amino acid frequencies. According to assumption 3, the probabilities f_i are time-independent. The probability of a site being unchanged over time t is close to the fraction of unchanged sites (u). Thus

$$u \approx \sum_{i=1}^{20} f_i \int_0^{+\infty} \rho_i(\lambda)e^{-\lambda t}d\lambda \tag{5}$$

According to the definition of substitution rate and assumption 1, the mean number of substitutions Δr that occurred in the sequence over a small time interval Δt is

$$\Delta r = n \sum_{i=1}^{20} f_i \int_0^{+\infty} \rho_i(\lambda)\lambda\Delta t d\lambda = n \sum_{i=1}^{20} f_i \bar{\lambda}_i \Delta t \tag{6}$$

where $\bar{\lambda}_i$ is the mean substitution rate of a site, occupied by amino acid A_i. Thus under the assumptions 2 and 3 the number of substitutions per position $d = r/n$, that occurred over time t, is

$$d = \sum_{i=1}^{20} f_i \bar{\lambda}_i t = \bar{\lambda} t \tag{7}$$

where $\bar{\lambda}$ is the mean substitution rate. Let us call identity fraction (q) the fraction of sites occupied by identical amino acids in ancestral sequence and a sequence after d substitutions per position. The fraction of unchanged sites (u) is smaller than the fraction of identical sites (q), because of last-step back substitutions. If in the chain of substitutions that occurred at particular site, the last substitution restores the amino acid that was at this site in ancestral sequence, then this site is indistinguishable from the unchanged one. The number of these last-step back substitutions should be proportional to the fraction of changed sites and for large q will be small. Assume that the fraction of unchanged sites (u) is close to the identity fraction ($q \approx u$). (This assumption, though, does not exclude from consideration other than last step back substitutions.) Formulas based on this assumption give the lower estimate of d. Under this assumption if $d \rightarrow \infty$, then q approaches 0 and not a positive value close to 0.05 as in the real case. The assumption is made for simplicity of consideration and because the other assumptions about distributions could affect the results in a more drastic way. Substituting $t = d/\bar{\lambda}$ [from equation (7)] in equation (5), we finally get

$$q \approx \int_0^{+\infty} \sum_{i=1}^{20} f_i \rho_i(\lambda) e^{-\frac{\lambda}{\bar{\lambda}} d} \, d\lambda \tag{8}$$

Because of the assumed time-independence and equilibrium (assumptions 2, 3) the equation holds for the comparison of two present-day sequences.

Specific Models

Formula (8) gives the general solution to the relation between the number of substitutions per site and the identity fraction. For its practical use we made additional assumptions about density functions f_i and $\rho_i(\lambda)$. Consider four cases.

1. The Poisson correction formula can be obtained from equation (8) under the assumption that substitution rates are independent of sites and amino acids. Thus functions $\rho_i(\lambda)$ are Dirac's δ-functions with the same mean $\bar{\lambda}$: for all $i = 1, \ldots, 20$, $\rho_i(\lambda)$ is zero for $\lambda \neq \bar{\lambda}$. In this case, from (8) we have

$$q = \sum_{i=1}^{20} f_i e^{-d} = e^{-d} \tag{9}$$

2. Dayhoff's treatment (Dayhoff et al. 1978) assumes that substitution rates are site-independent but differ for different amino acids. Thus functions $\rho_i(\lambda)$ are Dirac's δ-functions with different mean $\bar{\lambda}_i$. From equation (8) we have

$$q = \sum_{i=1}^{20} f_i e^{-\frac{\bar{\lambda}_i}{\bar{\lambda}} d} \tag{10}$$

Further, to replace the sum in equation (10) by an integral, we approximate a discrete distribution $f_i = f(\bar{\lambda}_i)$, defined on 20 points $\bar{\lambda}_i$ covering the interval $\lambda_{\min} \leq l \leq \lambda_{\max}$ by a continuous distribution with density function $g(l)$. In these terms, equation (10) is

$$q \approx \int_0^{+\infty} g(l) e^{-\frac{l}{\bar{\lambda}} d} \, dl \tag{11}$$

If the values of f_i and $\bar{\lambda}_i$ are independent, $\bar{\lambda}_i$ are more or less evenly distributed over the interval $0 \leq \lambda_{\min} \leq l \leq \lambda_{\max}$ and f_i are close to each other for different i, uniform distribution on the interval $\lambda_{\min} \leq l \leq \lambda_{\max}$ can give a good approximation of $g(l)$. The three conditions, specified above, are fulfilled for Dayhoff's values of amino acid frequencies and mutabilities. (The correlation coefficient between frequencies and mutabilities is 0.122; see mutabilities in table 21 on p. 347, Dayhoff et al. 1978; and the amino acid frequencies vary from 0.01 to 0.09.) We choose the uniform distribution with the same mean as the real distribution of $\bar{\lambda}_i$: $\bar{\lambda} = \Sigma_{i=1}^{20} f_i \bar{\lambda}_i$. For simplicity we take 0 as the left border for the interval because the smallest $\bar{\lambda}_i$ is close to 0. Then the right border is $2\bar{\lambda}$ and the density function is:

$$g(l) = \begin{cases} \frac{1}{2\bar{\lambda}} & \text{for} \quad 0 \leq l \leq 2\bar{\lambda} \\ 0 & \quad\quad\; \text{elsewhere} \end{cases} \tag{12}$$

Substituting (12) into (11), we get the final formula

$$q = \int_0^{2\bar{\lambda}} \frac{1}{2\bar{\lambda}} e^{-\frac{l}{\bar{\lambda}} d} \, dl = \frac{1 - e^{-2d}}{2d} \tag{13}$$

The formula

$$q = \frac{1 - e^{-2d}}{2d} \tag{14}$$

gives a remarkable approximation of Dayhoff's points (Fig. 1) with a square deviation sum of only 0.006. This is better than the formula obtained by fitting an equation to Dayhoff's points (Ota and Nei 1994, square deviation sum 0.143).

The approximate variance of d obtained by the delta method using equation (14) is

$$V(d) = \frac{d^2 q(1 - q)}{(1 - q(2d + 1))^2 n} \tag{15}$$

If we follow Bulmer (1991), the approximate covariance of d between sequences i and j (d_{ij}) and d between sequences k and l (d_{kl}) is given by

$$Cov(d_{ij}, d_{kl}) = \frac{4 d_{ij}^2 d_{kl}^2 (p_{ijkl} - p_{ij} p_{kl})}{(e^{-2d_{ij}}(2d_{ij} + 1) - 1)(e^{-2d_{kl}}(2d_{kl} + 1) - 1)n} \tag{16}$$

where p_{ijkl} is the proportion of those sites at which both i differs from j and k differs from l, p_{ij} is a proportion of those sites at which i differs from j, and p_{kl} is a proportion of those sites at which k differs from l. Equation (14) may be solved quantitatively: d is a limit of a sequence of numbers ($k \rightarrow \infty$) defined as $d(k + 1) = [1 - e^{-2d(k)}]/2q$. Few iterations are enough with an initial value of $d = -\ln q$, even $d = (1 - q^2)/2q$ is a good approximation.

3. Assume that substitution rates depend on sites, but not on amino acids. Then equation (8) has the form

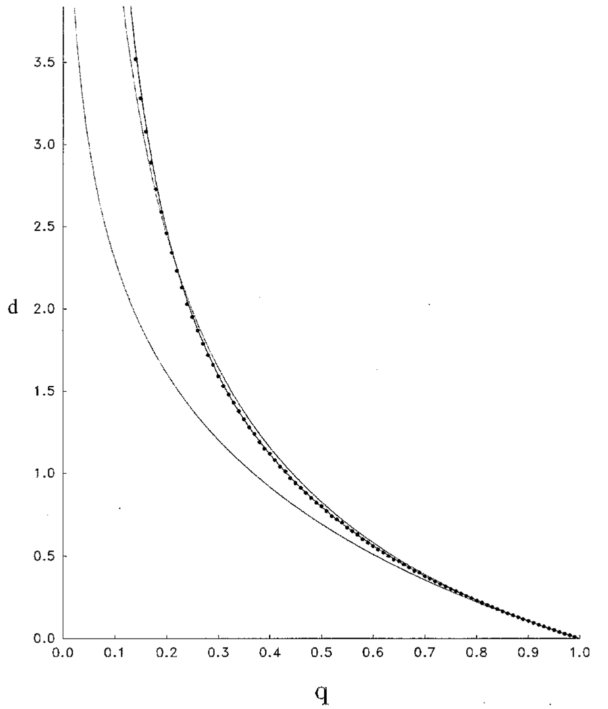

Fig. 1. Relationship between the proportion of identical amino acids (q) and the estimates of mean number of amino acid substitutions per site (d) for a pair of sequences:

1. the Poisson-correction method	$d=-\ln q$
2. the gamma distance	$d=2.03(q^{-0.49}-1)$
3. proposed method	$q=\frac{1-e^{-2d}}{2d}$

Dots indicate the distances obtained by Dayhoff et al.'s (1978) method.

$$q=\int_0^{+\infty}\rho(\lambda)e^{-\frac{\lambda}{\bar{\lambda}}d}\,d\lambda \tag{17}$$

Function $\rho(\lambda)$ should be estimated from the analysis of homologous sequences. We provide a theoretical reason for our choice of the function, making use of assumption 3 about equilibrium state of the system under consideration. For a physical system, the density function at equilibrium is usually the one that maximizes the entropy $S=-\int_0^{+\infty}\rho\ln\rho d\lambda$. ($\rho$ is a density function.) The postulate of maximum entropy at equilibrium can be assumed by default. The mean value of $\rho(\lambda)$ according to (6) and (7) is $\bar{\lambda}$. Assume also that this mean substitution rate is characteristic of a sequence and does not change: the number of substitutions over equal time intervals is constant regardless of the distribution of substitution rates among sites. In the class of continuous distributions with a given mean, the exponential distribution is the one with the maximal entropy. That is, if the differentiable function $\rho(\lambda) \geqslant 0$, defined for $\lambda \geqslant 0$ satisfies the following conditions:

$$\int_0^{+\infty}\rho(\lambda)d\lambda=1$$

$$\int_0^{+\infty}\lambda\rho(\lambda)d\lambda=\bar{\lambda}$$

$$-\int_0^{+\infty}\rho(\lambda)\ln\rho(\lambda)d\lambda \text{ is maximal}$$

then

$$\rho(\lambda)=\frac{e^{-\frac{\lambda}{\bar{\lambda}}}}{\bar{\lambda}}$$

(See footnote 1.)

The exponential distribution is the special case of the gamma distribution $\rho(\lambda)=\beta^{\alpha}/\Gamma(\alpha)\,\lambda^{\alpha-1}e^{-\beta\lambda}$ for $\alpha=1$. The choice of an exponential distribution $\rho(\lambda)=\exp(-\lambda/\bar{\lambda})/\bar{\lambda}$ in equation (17) provides the formula

$$q=\int_0^{+\infty}\frac{e^{-\frac{\lambda}{\bar{\lambda}}-\frac{\lambda}{\bar{\lambda}}d}}{\bar{\lambda}}d\lambda=\frac{1}{1+d} \tag{18}$$

The usage of an exponential distribution has some support also from analysis of protein sequences (Holmquist et al. 1983).

4. Assume that substitution rates are both site-dependent and amino-acid-dependent. The use of the exponential distribution $\rho(\lambda)=e^{-\lambda/l}/l$ and of the uniform distribution (12) will result in the formula

$$q=\int_0^{2\bar{\lambda}}\frac{1}{2\bar{\lambda}}dl\int_0^{+\infty}\frac{e^{-\frac{\lambda}{l}-\frac{\lambda}{\bar{\lambda}}d}}{l}d\lambda=\frac{\ln(1+2d)}{2d} \tag{19}$$

In summary, a general model, which is suitable for the estimation of the number of amino acid substitutions per site from the fraction of identical sites between two sequences, has been presented. One of the formulas gives a very close approximation of the points calculated by the empirical method of Dayhoff et al. (1978).

Acknowledgments. I thank Dr. Emile Zuckerkandl for constructive comments and helpful suggestions, Dr. Vladislav Markin, Dr. Margaret Phillips, and Dr. Andrei Osterman for critical reading of the manuscript.

References

Bulmer M (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol Biol Evol 8:868–883

[1] The postulate of maximum entropy works not only for statistical physics. For example, we desire to find a distribution of random errors in measurements. If the measurements are made by the same instrument, the "precision" of all measurements is constant. That makes a variance of a distribution constant. It is easy to show that the distribution with maximal entropy among all possible continuous distributions with the given variance is the normal distribution.

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In; Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC, pp 89–99

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Holmquist R, Goodman M, Conroy T, Czelusniak J (1983) The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19:437–448.

Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 38:642–643

Takacs L (1966) Stochastic process. Methuen & Co, London: John Wiley & Sons, New York; pp 28–46.

Uzzell T, Corbin KW (1971) Fitting discrete probability distribution to evolutionary events. Science 172:1089–1096

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins Academic Press, New York, pp 97–166.