J Mol Evol (1997) 45:359–369

JOURNAL OF MOLECULAR EVOLUTION

Estimation of Evolutionary Distances from Protein Spatial Structures

Nick V. Grishin

Department of Pharmacology, The University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235-9041, USA

Received: 21 September 1995 / Accepted: 19 May 1997

Abstract. New equations are derived to estimate the number of amino acid substitutions per site between two homologous proteins from the root mean square (RMS) deviation between two spatial structures and from the fraction of identical residues between two sequences. The equations are based on evolutionary models, analyzing predominantly structural changes and not sequence changes. Evolution of spatial structure is treated as a diffusion in an elastic force field. Diffusion accounts for structural changes caused by amino acid substitutions, and elastic force reflects selection, which preserves protein fold. Obtained equations are supported by analysis of protein spatial structures.

Key words: Protein structure — RMS deviation — Molecular evolution — Evolutionary distance — Substitution rates

Introduction

The evolutionary distance d_{ij} between two homologous protein sequences i and j is equal to the number of amino acid substitutions per site occurring in these sequences since their divergence from a common ancestor. The evolutionary distance is an additive characteristic. For example, if sequences i and j originate from the sequence k[1], then $d_{ij} = d_{ki} + d_{kj}$. The distance d_{ij} is not known when two sequences i and j are given and needs to be estimated. This estimation depends on the evolutionary model of sequence change.

In contrast, the fraction of sites p_{ij} occupied by different residues in two sequences i and j is easy to calculate, but because of possible back substitutions p_{ij} is not additive: $p_{ij} \leqslant p_{ki} + p_{kj}$. If d_{ij} is small, then $d_{ij} \approx p_{ij}$.

Additivity of distances is widely used for evolution tree construction (Saitou 1988; Kishino et al. 1990; Rzhetsky and Nei 1992; Yang 1993, 1994; Tateno et al. 1994). Thus a variety of models to estimate distance d_{ij} from the fraction of different residues p_{ij} are proposed (Zuckerkandl and Pauling 1965; Dayhoff et al. 1972, 1978; Uzzel and Corbin, 1971; Holmquist et al. 1983; Wilbur 1985; Barry and Hartigan 1987; Ota and Nei 1994; Zharkikh 1994; Tajima and Takezaki 1994; Grishin 1995). Analysis of alignments (Dayhoff et al. 1978) used for estimation of relation between distance d_{ij} and fraction of identical residues $q_{ij} = 1 - p_{ij}$ misses some back substitutions for the case of lower similarity between sequences. This analysis underestimates distances. A different and independent method allowing for a relationship between d_{ij} and q_{ij} should be of a value. Such a method is presented here.

The conservation of protein spatial structure correlates with sequence conservation. The correlation between RMS (root mean square) deviation in C_α atom coordinates Δ_{ij} and fraction of identical residues q_{ij} was

Abbreviations: RMS-root mean square; definitions of symbols are as per Appendix 1

[1] Sequence k is an ancestral sequence for sequence pairs (k, i), (k, j), and (i, j).

analyzed (Chotia and Lesk 1986; Lesk and Chotia 1986; Hubbard and Blundell 1987; Flores et al. 1993; Gutin and Badretdinov 1994). This correlation was rationalized by an empirical exponential formula (Chotia and Lesk 1986) and in a more sophisticated way, based on similarities of substitution process with the diffusion in a multidimensional space (Gutin and Badretdinov 1994). In this article, the latter approach is developed: corrections are made for the upper limit of RMS deviation between two proteins with similar fold and for correlation between residues linked in a polypeptide chain. Using the derived relation between evolutionary distance d_{ij} and RMS deviation Δ_{ij}, and experimental data on correlation of identity fraction q_{ij} and RMS deviation Δ_{ij} (Chotia and Lesk 1986; Hubbard and Blundell 1987; Flores et al. 1993), d_{ij} is estimated as a function of q_{ij}. This estimate uses comparisons of three-dimensional structures and evolutionary models in terms of structural changes and thus differs methodologically from the others (Dayhoff et al. 1978—like sequence-based models). However, the resulting dependence is similar to those originating from sequence comparisons (Uzzel and Corbin 1971; Holmquist et al. 1983; Grishin 1995).

Theory

The number of sites in a protein sequence is assumed to be large. Amino acid substitutions at each site are treated as mutually independent, finite state, continuous time, homogeneous Markov processes (Takacs 1966). Each amino acid at each site is characterized by a substitution rate (infinitesimal transition probability) that is assumed to be constant.[2] The substitution process is assumed to be at equilibrium: all probability distributions are time-independent and are the same for all sequences.

Distribution of Substitution Rates Among Sites

The basic relations between the fraction u of unchanged sites (sites, at which no substitutions occurred) between two sequences, the fraction q of identical sites (sites occupied by identical amino acids in two sequences) and the distribution of rates among sites are stated here.

Under the present assumptions:

- the probability that a site with a substitution rate λ, $\lambda \geqslant 0$ remains unchanged over time t, $t \geqslant 0$ is $e^{-\lambda t}$;
- the number of substitutions per site d is proportional to time t: $d = \bar{\lambda}t$, where $\bar{\lambda}$ is the mean substitution rate, $\bar{\lambda} = \Sigma_{k=1}^{l} \lambda_k(t)/l$, $\lambda_k(t)$ is the substitution rate of the site k at time t, and l is the number of sites in the sequence, $\bar{\lambda}$ is time-independent;
- the distribution of relative substitution rates among sites $\rho(x)$ is time-independent ($\rho(x)\mathrm{d}x$ is the fraction of sites with relative substitution rate x from $x = \lambda/\bar{\lambda}$ to $x = \lambda/\bar{\lambda} + \mathrm{d}\lambda/\bar{\lambda}$, $\int_0^{+\infty}\rho(x)\mathrm{d}x = 1$, $\int_0^{+\infty}x\rho(x)\mathrm{d}x = 1$).

Therefore the fraction u of unchanged sites between two sequences separated by d substitutions per site is a Laplace transform of the density of distribution of relative substitution rates among sites:

$$u(d) = \int_0^{+\infty} \rho(x)e^{-xd}\,\mathrm{d}x. \qquad (1)$$

Due to the property of Laplace transform, v^{th} moment of the distribution $\rho(x)$ is

$$M^{(v)} = \int_0^{+\infty} x^v\rho(x)\mathrm{d}x = (-1)^v \left.\frac{\partial^v u(d)}{\partial d^v}\right|_{d=0}. \qquad (2)$$

The fraction q of identical sites is

$$q(d) = \int_0^{+\infty} q(x,d)\mathrm{d}x, \qquad (3)$$

where for the two sequences separated by d substitutions per site, $q(x,d)\mathrm{d}x$ is the fraction of identical sites with the relative subsitution rate from x to $x + \mathrm{d}x$. If $d = 0$, then $q(x,0) = \rho(x)$. Therefore, differentiating equation (3) with respect to d gives

$$\left.\frac{\mathrm{d}q}{\mathrm{d}d}\right|_{d\to 0=0} = \int_0^{+\infty} -xq(x,0)\,\mathrm{d}x = -1, \qquad (4)$$

since $\lim_{d\to 0}(\mathrm{d}q(x,d)/\mathrm{d}d) = -xq(x,0)$ and the mean value of x is 1.

If all amino acids are equally changeable, then every amino acid at a site with a relative substitution rate x is replaced by any of the rest 19 with equal probability, and $q(x) = q(x,d)$ is the solution of differential equation

$$\frac{\mathrm{d}q(x)}{\mathrm{d}d} = -xq(x) + \frac{x}{19}(\rho(x) - q(x))$$
$$\text{under initial condition } q(x,0) = \rho(x). \qquad (5)$$

Substituting the solution of (5) into Equation (3) results in the final formula

$$q(d) = \frac{19}{20}\int_0^{+\infty} \rho(x)e^{-\frac{20}{19}xd}\,\mathrm{d}x + \frac{1}{20}. \qquad (6)$$

[2] The assumptions that variations at individual sites occur independently and that the substitution rates for sites are constant, oversimplify the reality. However, these assumptions do not affect the derived later relation between the distance and the RMS deviation, as soon as the distance is directly proportional to time.

Comparison of Equations (6) and (1) yields

$$q(d) = \frac{19}{20} u\left(\frac{20}{19} d\right) + \frac{1}{20}. \qquad (7)$$

In general, when amino acids are not necessarily equally changeable, each site k is characterized by a 20×20 matrix $\mathbf{X}(k)$ of relative substitution rates with elements $x_{ij}(k) = \lambda_{ij}(k)/\bar{\lambda}$ for $i \neq j$ and $x_{ii}(k) = -\Sigma_{j=1,j\neq 1}^{20} \lambda_{ij}(k)/\bar{\lambda}$, where $\lambda_{ij}(k)$ is a substitution rate of amino acid i by amino acid j at a site k. In the case under study, matrix $\mathbf{X}(k)$ possesses 20 real nonpositive eigenvalues $\xi_i(k)$, one of which is zero; $\xi_1 = 0$, and

$$q(d) = \sum_{k=1}^{l} \sum_{i=2}^{20} c_i(k) e^{\xi_i(k)d}/l + \sum_{k=1}^{l} c_1(k)/l, \qquad (8)$$

where $c_i(k)$ are functions of $\mathbf{X}(k)$. Making use of Equations (4) and (8), we find that $\mathrm{d}q/\mathrm{d}d\big|_{d=0} = \Sigma_{k=1}^{l}\Sigma_{i=2}^{20} c_i(k)\xi_i(k)/l = -1$. If $d = 0$, then $q = \Sigma_{k=1}^{l}\Sigma_{i=1}^{20} c_i(k)/l = 1$. If $d \to \infty$ then $q \to \Sigma_{k=1}^{l} c_1(k)/l = q_\infty$. Therefore, approximating the double sum in Equation (8) by an integral, we have

$$q(d) = (1 - q_\infty) \int_0^{+\infty} g(y) e^{-\frac{1}{1-q_\infty} yd}\, \mathrm{d}y + q_\infty, \qquad (9)$$

where function $g(y)$ satisfies conditions $\int_0^{+\infty} g(y)\mathrm{d}y = 1$ and $\int_0^{+\infty} y g(y)\mathrm{d}y = 1$ and has a meaning of probability density function of relative (each non-zero eigenvalue is divided by the arithmetical mean of all non-zero eigenvalues) non-zero eigenvalues of matrices $\mathbf{X}(k)$. Comparison of Equation (9) with Equation (6) reveals that if amino acids are equally changeable, then $q_\infty = 1/20$ and $g = \rho$. Therefore under some conditions $g \approx \rho$ and from Equations (9) and (1)

$$u\left(\frac{d}{1-q_\infty}\right) \approx \frac{q(d) - q_\infty}{1 - q_\infty}, \text{ and } q_\infty \approx \frac{1}{20}. \qquad (10)$$

Thus for small distances d the fraction of unchanged sites u is close to the fraction of identical sites q: $u \approx q$. If d is larger, then $u \leqslant q$ because of back substitutions. If $d \to \infty$, then $u \to 0$ and $q \to q_\infty \approx 1/20$.

RMS Deviation and Evolutionary Distance

Consider two homologous proteins i and j of l amino acids each. The trace of a spatial structure of a protein is the set of l radius vectors $\mathbf{r}(k)$ with each vector corresponding to the C_α atom of each amino acid. Optimal superposition of two structures is given by the minimum of the RMS deviation Δ_{ij}:

$$\Delta_{ij} = \sqrt{\frac{\sum_{k=1}^{l} (\mathbf{r}_i(k) - \mathbf{r}_j(k))^2}{l}}. \qquad (11)$$

A difference vector between two optimally superimposed structures at each site k is defined as

$$\Delta\mathbf{r}(k) = \mathbf{r}_i(k) - \mathbf{r}_j(k). \qquad (12)$$

Consider the difference vectors for two structures of l residues as l realizations of a random variable $\Delta\mathbf{r}$. The minimum of Δ_{ij} implies that the mean of $\Delta\mathbf{r}$ squared is small in comparison with the mean of $\Delta\mathbf{r}^2$. Therefore, the variance of $\Delta\mathbf{r}$ is close to RMS deviation squared: $\Delta_{ij}^2 = \Sigma_{k=1}^{l} \Delta\mathbf{r}^2(k)/l$.

For the structures of two identical proteins, all difference vectors should be close to zero. Deviations from zero are caused by random reasons in the order of experimental errors (quantum mechanics limitations, refinement, resolution limits, crystal packing, etc., we do not consider substantial conformational changes) and are expected to follow normal distribution with the mean much less than Δ_{ii}, and the variance close to $\Delta_{ii}^2 = \Delta_0^2$.

Consider the evolution of a protein from an ancestral one in time t. The probability density function of the difference vectors between the ancestral protein structure and the resulting protein structure at time $t = 0$ is approximated by

$$\rho(\Delta\mathbf{r},0) = (2\pi\Delta_0^2)^{-\frac{3}{2}} e^{-\frac{\Delta\mathbf{r}^2}{2\Delta_0^2}}. \qquad (13)$$

If amino acid substitutions occur in time t, then the resulting sequence diverges from the ancestral one. This divergence causes structural deviations, and the density function changes. The normalizing condition $\int_{-\infty}^{+\infty} \rho(\Delta\mathbf{r},t)\mathrm{d}\Delta\mathbf{r} = 1$ implies that equation of continuity holds for $\rho = \rho(\Delta\mathbf{r},t)$:

$$\frac{\partial\rho}{\partial t} + \mathrm{div}\mathbf{j} = 0, \qquad (14)$$

where $\mathbf{j}$ is the current density (density passing per unit time through a unit area normal to the direction of flow). If the density is composed of l points defined by radius vectors $\Delta\mathbf{r}(k)$, then these points move in three dimensions in time as soon as the difference vectors $\Delta\mathbf{r}(k)$ change. The situation can be treated as a ''diffusion'' of points in three-dimensional space. For $l \to \infty$, Equation (14) holds.

Assumptions about $\mathbf{j}$ make it possible to solve Equation (14) under the initial condition (13) and the boundary conditions ($\rho \to 0$ for $\Delta\mathbf{r} \to \infty$ at all t), which lead to the determination of a density function $\rho(\Delta\mathbf{r},t)$ at any time $t \geqslant 0$.

Unlimited Independent Diffusion

The simplest case of diffusion originates from the assumption that each ''point'' defined by $\Delta\mathbf{r}(k)$ moves independently and in any direction with equal probability. This assumption leads to the Fick's law, obeyed for the simplest diffusion phenomena,

$$\mathbf{j} = -D\mathbf{grad}\rho, \tag{15}$$

where D is a diffusion coefficient which characterizes how fast the density changes. Substitution of (15) into (14) leads to the diffusion equation:

$$\frac{\partial\rho}{\partial t} = D\nabla^2\rho, \tag{16}$$

which under initial condition (14) has the solution

$$\rho(\Delta\mathbf{r},t) = (2\pi(2Dt + \Delta_0^2))^{-\frac{3}{2}} e^{-\frac{\Delta\mathbf{r}^2}{2(2Dt+\Delta_0^2)}}. \tag{17}$$

The variance of this distribution is equal to $2Dt + \Delta_0^2$, and the RMS deviation squared is close to the variance. Since the number of substitutions per site is proportional to time: $d = \bar{\lambda}t$, we have the following formula for RMS deviation:

$$\boxed{\Delta_{ij} = \sqrt{\alpha^2 d + \Delta_0^2}}, \tag{18}$$

where $\alpha^2 = 2D/\bar{\lambda}$. The parameter α^2 is the mean sum of the square deviations introduced by one amino acid substitution. For $d = 1/l$ (one amino acid out of l is substituted), Equation (18) can be transformed to $\alpha^2 = l\Delta_{ij}^2 - l\Delta_0^2$, where $l\Delta_{ij}^2$ is the mean sum of the square deviations between structures differing in one amino acid out of l, and $l\Delta_0^2$ is the mean sum of square deviations between structures of identical proteins of l amino acids.

The same Equation as (18) is derived with a slightly different consideration by Gutin and Badretdinov (1994). Their work shows that α^2 does not depend on the sequence length nor on the secondary structure class of the protein. One amino acid substitution alters the C_α positions of neighboring spatial structure residues, and the number of these residues as well as their distribution of deviations are approximately the same for all proteins.

Limited Independent Diffusion

Consideration in the previous section (Unlimited Independent Diffusion) implies that diffusion of density is unlimited: if $t \to \infty$, then $\Delta_{ij} \to \infty$. However, structural changes in proteins should be restricted by conservation of their fold. Assume that if $t \to \infty$, then $\Delta_{ij} \to \Delta_\infty$. A force field $\mathbf{f}$ can be introduced to preserve protein fold. The diffusion flux in a force field is

$$\mathbf{j} = -D\mathbf{grad}\rho + \rho\mathbf{f}. \tag{19}$$

Assume that the force field is elastic:

$$\mathbf{f} = -a\Delta\mathbf{r},\ a \geqslant 0. \tag{20}$$

This assumption results in the normality of the distribution $\rho(\Delta\mathbf{r},t)$ with the variance given by (A2–7) (see Appendix 2 for derivation). Since the number of substitutions per site is proportional to time: $d = \bar{\lambda}t$, the current model suggests that the RMS deviation between two structures as a function of d is

$$\boxed{\Delta_{ij} = \sqrt{\Delta_\infty^2 - (\Delta_\infty^2 - \Delta_0^2)e^{-\frac{\alpha^2}{\Delta_\infty^2}d}}}, \tag{21}$$

where $\alpha^2 = 2D/\bar{\lambda}$. If $\Delta_\infty^2 \to \infty$, then equation (21) transforms into (18). For small d

$$\Delta_{ij}^2 \approx \alpha^2\left(1 - \frac{\Delta_0^2}{\Delta_\infty^2}\right)d + \Delta_0^2. \tag{22}$$

Limited Correlated Diffusion

The distribution of the difference vectors $\Delta\mathbf{r}$ for two structures of l residues with the distance of d substitutions per site between them is normal with zero mean and variance Δ_{ij}^2 given by Equation (21). Consider the RMS deviation squared $\Delta^2 = \Sigma_{k=1}^{l} \Delta\mathbf{r}_k^2/l$ for two structures a random variable Δ^2. This random variable is proportional to the sum of squares of l normally distributed independent random variables $\Delta\mathbf{r}$ with the same mean 0 and variance Δ_{ij}^2. The mean value of Δ^2 is $\Delta_{ij}^2(d)$ given by Equation (21), and the variance is $2\Delta_{ij}^4(d)/l$. The number of residues l in a protein is usually large ($50 < l < 1000$). Therefore, the variance of the RMS deviation between any two proteins with the same d (for example, for identical proteins, where $d = 0$) should be small. Analysis of protein structures shows that it is not true, and the RMS deviations vary substantially with protein pairs (Chotia and Lesk 1986; Hubbard and Blundell 1987; Flores et al. 1993). Identical proteins (Table 1 in Flores et al. (1993), for proteins refined to a resolution 2 Å or better only) of average length of 60 residues have $\Delta_0^2 \approx 0.2 \pm 0.13$. The model of independent diffusion predicts a value of standard deviation to be $\sqrt{(2 \cdot 0.2^2)/(60)} = 0.04$. The value of 0.13 can be obtained from the model if $l = 5$. Thus changes in difference vectors are correlated.

Substitution of one residue should affect the positions of the C_α atoms of several residues and not only of the substituted one. Most of these residues are involved in secondary structure. Because the secondary structure itself is not usually altered by substitutions and random reasons (structure refinement, crystal packing, etc.), a substituted residue should cause a shift or rotation of the whole secondary structural element. This phenomenon

reduces the number of independently diffusing elements from the number of residues l to approximately the number of secondary structure elements s.

Therefore the mean value and the variance of the RMS deviation squared Δ^2 are equal to

$$< \Delta^2 > = \Delta_{ij}^2(d) = \Delta_\infty^2 - (\Delta_\infty^2 - \Delta_0^2)e^{-\frac{\alpha^2}{\Delta_\infty^2}d}, \quad Var(\Delta^2) = \frac{2\Delta_{ij}^4(d)}{s} \qquad (23)$$

respectively. This formula suggests that the error in the estimation of the mean RMS deviation corresponding to a pair of structures with d substitutions per site is $\Delta_{ij}/2\sqrt{2/s}$. Since the average number of secondary structural elements in a protein is about 10 ($s = 10$), the error in the estimation of RMS deviation is about ¼ of its value. Therefore, estimation of the distances d from the RMS deviation values Δ_{ij} is rather uncertain and is impossible to use for accurate tree construction. However, statistical analysis of data on the correlation of identity fraction q between the sequences and the RMS deviations Δ between the structures enables us to validate Equation (23) and obtain estimates of its parameters. More importantly, the proposed approach gives a relationship between the distance d and identity fraction q, independent of analysis of sequence alignments as well as an estimate of the distribution of substitution rates among sites.

Distance and Identity Fraction: Empirical Formula

Chotia and Lesk (1986) demonstrated that a simple exponential formula

$$\Delta_{ij}^2 = ae^{b(1-q)}, \qquad (24)$$

where a and b are best fit coefficients, gives a good fit to the correlation data of the RMS deviation Δ_{ij} and the identity fraction q. Combination of Equations (24), (21) and (10) leads (see Appendix 3 for the derivation) to an empirical relation between the number of substitutions per site d and the fraction of unchanged sites u:

$$u = \frac{-\ln\left(1 - \frac{\beta - 1}{\beta}e^{-\frac{\ln\beta}{\beta-1}d}\right)}{\ln\beta}. \qquad (25)$$

Therefore, under the present empirical model, only one parameter: $\beta = \Delta_\infty^2/\Delta_0^2$ is sufficient to define the relation between d and u as well as to define the distribution of substitution rates among sites. For $\beta = 1$, Equation (25) transforms to $u = e^{-d}$, which is a Poisson-correction formula (Zuckerkandl and Pauling 1965). This equation relates the fraction of unchanged residues to the distance for the case of amino acid-independent and site-independent substitution rates. This Poisson-correction formula gives the lower limit of the d estimate. For $\beta \to \infty$, Equation (25) transforms to $u \to 1$, which is the upper limit of the d estimate. Thus Formula (25) covers the range of all possible estimates and could have more general application than the exponential approximation (24). Selection of β, $\beta \geqslant 1$ gives d estimates between the lower ($\beta = 1$) and upper ($\beta \to \infty$) limits. One of these estimates [in the range of applicability of Formula (24)] is appropriate for the real case.

To find the density of relative substitution rates (1), we reverse the Laplace transform in the right part of equation (25) using the power series ($0 \leqslant (\beta - 1)/(\beta)e^{-(\ln\beta)/(\beta-1)d} < 1$ for all $\beta \geqslant 1$, $d \geqslant 0$):

$$u = \frac{-\ln\left(1 - \frac{\beta - 1}{\beta}e^{-\frac{\ln\beta}{\beta-1}d}\right)}{\ln\beta} = \frac{1}{\ln\beta}\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\beta-1}{\beta}\right)^k e^{-k\frac{\ln\beta}{\beta-1}d}. \qquad (26)$$

The inverse transform of the series in the right part of Equation (26) (the density function) is given by the series:

$$\rho(x) = \frac{1}{\ln\beta}\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\beta-1}{\beta}\right)^k \delta\left(x - k\frac{\ln\beta}{\beta-1}\right), \qquad (27)$$

where $\delta(x - k\,(\ln\beta)/(\beta - 1))$ are delta-functions ($\delta(x - k\,(\ln\beta)/(\beta - 1)) = 0$ for $x \neq k\,(\ln\beta)/(\beta - 1)$, but $\int_{-\infty}^{+\infty}\delta(x - k\,(\ln\beta)/(\beta - 1))\,dx = 1$). Thus the density function consists of discrete peaks at points $x = k\,(\ln\beta)/(\beta - 1)$ with the ''heights'' $(1)/(k\ln\beta)\,((\beta - 1)/(\beta))^k$, $k = 1,2,\ldots,\infty$. For $\beta = 1$ the distribution (27) transforms into $\rho(x) = \delta(x - 1)$, which corresponds to the case of amino acid and site independent substitution rate, that is equal to the mean substitution rate.

The substitution rate is a continuous variable ranging from 0 to $+\infty$, and the discrete character of distribution (27) is caused by an empirical exponential formula (24). The discrete distribution (27) can be approximated by a continuous one with the same as distribution (27) the first and second moments. If $u(d)$ is given by Equation (25), then the first two moments of distribution of x are found by applying Equation (2):

$$M^{(1)} = 1, \text{ and } M^{(2)} = \frac{\beta\ln\beta}{\beta - 1}. \qquad (28)$$

Two continuous distributions were used previously to approximate the distribution of relative substitution rates, namely a gamma distribution (Uzzel and Corbin 1971) and a log-normal distribution (Olsen 1987). The gamma

distribution density with the first two moments, given by (28) is

$$\rho(x) = \frac{a^a}{\Gamma(a)} x^{a-1} e^{-ax},\ a = 1 \Big/ \left(\frac{\beta}{\beta - 1} \ln \beta - 1 \right). \quad (29)$$

The log-normal distribution density with the first two moments, given by (28) is

$$\rho(x) = \frac{1}{\sqrt{2\pi a x}} e^{-\frac{1}{2a}\left(\ln x + \frac{a}{2}\right)^2},\ a = \ln \frac{\beta \ln \beta}{\beta - 1} \cdot \quad (30)$$

Distribution (27) can also be approximated by a histogram:

$$\rho(x) = \frac{\beta - 1}{k \ln^2 \beta} \left(\frac{\beta - 1}{\beta} \right)^k \text{ for } (k - 0.5) \frac{\ln \beta}{\beta - 1} <$$

$$x \leqslant (k + 0.5) \frac{\ln \beta}{\beta - 1},\ k = 1,2, \ldots$$

$$\rho(x) = 0 \text{ elsewhere.} \quad (31)$$

Parameter Estimation and Discussion

The goal is to estimate parameters Δ_0^2, Δ_∞^2, and α^2 from the data on structure comparisons, relating Δ_{ij} and q_{ij}; transform Δ_{ij} into d_{ij} using formula (21); transform q_{ij} into u_{ij} using formula (7); and estimate function $d(u)$ from the obtained pairs (u_{ij}, d_{ij}).

The lower limit of the RMS deviation could not be less than the quantum mechanics limitation and should be close to the wavelength of the carbon atom: 0.2Å. The resolution limit should increase this value. Thus the RMS deviation between independently refined structures of the same protein is expected to be close to twice the lowest limit: $\Delta_0 \approx 0.4$Å.

The mean upper limit of the RMS deviation between proteins sharing the same fold is unlikely to be more than one α-helix turn or an α-helix width: $\Delta_\infty < 5$Å.

The change in C_α positions introduced by a single amino acid substitution α is expected to be close to 1/2 of the mean difference in length between side chains of the original and the substituted amino acids. The longest side chain (Arg) is about 7Å and the shortest one (Gly) is 0Å. Assuming uniform distribution of differences over the interval from 0Å to 7Å, the estimation of the mean difference as a standard deviation of the distribution is $7/\sqrt{12} \approx 2$. Therefore $\alpha \approx \alpha^2 \approx 1$.

To obtain more precise estimates, data on the analysis of the correlation of the identity fraction q and the RMS deviation Δ, made by three independent research groups [Chotia and Lesk (1986, Table 2); Hubbard and Blundell (1987, Table 2, data on all topologically equivalent residue pairs); Flores et al. (1993, Fig. 1a, Table 1)] are used. Structurally equivalent residues in distant proteins are not always possible to match because of sequence divergence and indels in loop regions. Thus the core residues in two structures (a subset of sites present in both structures that could be superimposed) are defined, and the RMS deviation of residues in this subset, rather than in the whole structure, is calculated. Formulas derived above could be applied to any subset of the residues and not necessarily to the whole structures. Because of differences in subset definitions in the three works (Chotia and Lesk 1986; Hubbard and Blundell 1987; Flores et al. 1993), parameters Δ_0^2, Δ_∞^2, and α^2 are estimated from each dataset individually and $q_\infty = 0.05$ is assumed.

Independent of the relation between d and q, estimates of Δ_0^2, Δ_∞^2 and α^2 are obtained from the data on correlation of q and Δ (points (q, Δ)) as follows:

- Δ_0^2 is estimated as an arithmetical mean of the squares of the RMS deviations for identical proteins;
- Δ_∞^2 is estimated as the value of the RMS deviation squared for $q = 0.05$: the best linear fit to the points, corresponding to identity fraction 0.4 and lower (fit in a form $\Delta^2 = a(q - 0.05) + \Delta_\infty^2$);

Table 1. Estimates of parameters Δ_0^2, Δ_∞^2, and α^2

		Estimates		
		Data from		
Parameter	Theory	Chotia and Lesk (1986)	Hubbard and Blundell (1987)	Flores et al. (1993)
Δ_0^2	0.16	0.11 ± 0.02	0.31 ± 0.06	0.34 ± 0.06
Δ_∞^2	<25	4.51 ± 0.55	2.49 ± 0.10	20.64 ± 1.48
α^2	1	1.28 ± 0.35	0.87 ± 0.37	1.22 ± 0.44
$\frac{\ln \Delta_\infty^2 - \ln \Delta_0^2}{\frac{1}{\Delta_0^2} - \frac{1}{\Delta_\infty^2}}$		4.01 ± 0.73	0.78 ± 0.11	1.49 ± 0.33

Table 2. Estimates of the best fit parameters to model equations

	Data from		
	Chotia and Lesk (1986)	Hubbard and Blundell (1987)	Flores et al. (1993)
Expotential Equation (24)			
a	0.28 ± 0.12	0.34 ± 0.04	0.17 ± 0.08
b	3.05 ± 0.56	2.15 ± 0.14	5.18 ± 0.55
RMS residual	0.82	0.32	3.37
Diffusion approximation (33)			
α^2	1.46 ± 0.26	0.89 ± 0.07	2.67 ± 0.32
Δ_∞^2	3.86 ± 0.41	2.15 ± 0.07	20.52 ± 1.53
RMS residual	0.79	0.32	3.37

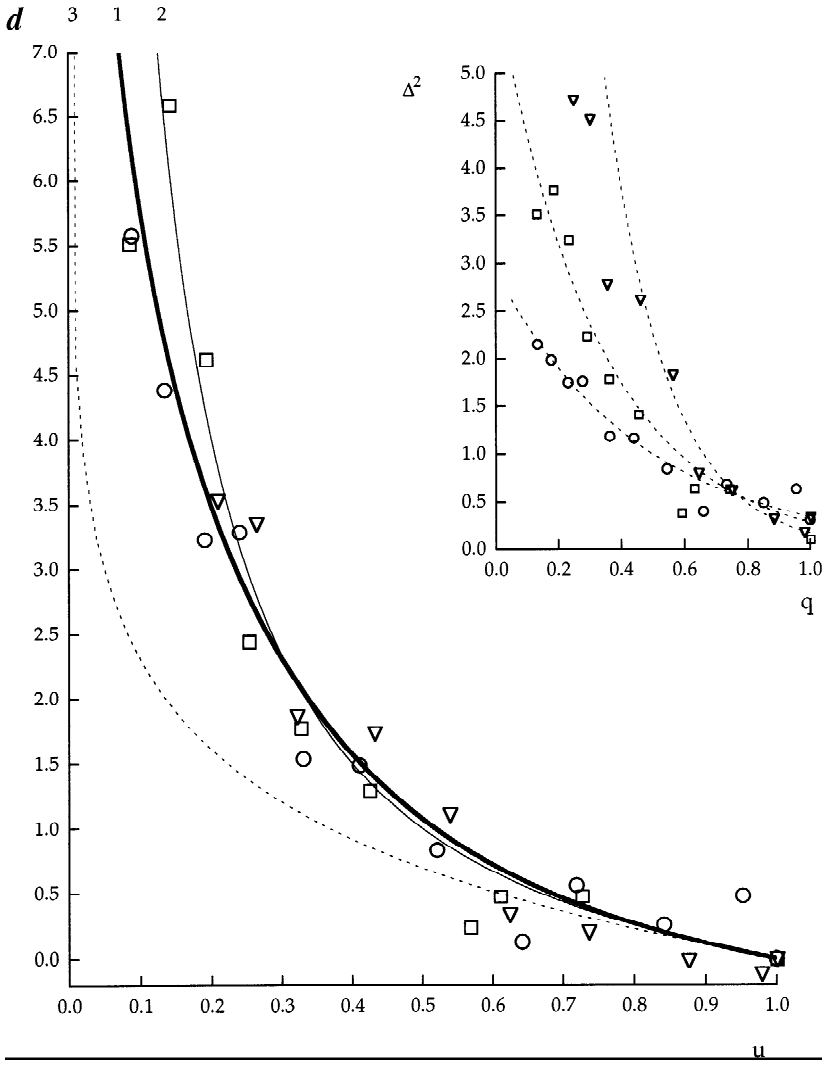

Fig. 1. Number of substitutions per site d as a function of the fraction of unchanged sites u. Transformed with Equations (21) and (7) grouped data on correlation of RMS deviation (△) and identity fraction (q) are the data of □ = Chotia and Lesk (1986); ○ = Hubbard and Blundel (1987); and ▽ = Flores et al. (1993). The thick curve (curve 1) represents Equation (34) with the most probable β value of 10 and the thin curve (curve 2) is for equation $d = 1/u - 1$ (Grishin, 1995). The dashed curve (curve 3) corresponds to the Poisson-correction formula $d = -\ln u$ [Equation (34) with $\beta = 1$]. The inset shows grouped untransformed data. The dashed curves are the best fit exponential curves (24) with parameters from Table 2.

- $\alpha^2 = -1/(1 - \Delta_\infty^2/\Delta_0^2)\, d\Delta^2/dq$, where $d\Delta^2/dq$ is estimated as the best fit slope of the linear equation $\Delta^2 - \Delta_0^2 = d\Delta^2/dq\,(q-1)$ to the points corresponding to the identity fraction 0.55 (0.65 for the data of Flores et al. 1993) and higher. Identical proteins are excluded from the fit since they were used for Δ_0^2 estimation. Only $d\Delta^2/dq$ is fitted and estimated before value of Δ_0^2 is used. These estimates are placed in Table 1. The difference in the definitions of topologically equivalent residue pairs should mostly influence the value of Δ_∞^2, and the data in Table 1 illustrate this point.

To test the applicability of exponential approximation (24) to each of the three datasets, the value of $1/(1 - q_\infty)\,(\ln\Delta_\infty^2 - \ln\Delta_0^2)/(1/\Delta_0^2 - 1/\Delta_\infty^2)$ is calculated. If the exponential formula (24) approximates the data, then from equation (A3–7, Appendix 3) $\alpha^2 = 1/(1 - q_\infty)\,(\ln\Delta_\infty^2 - \ln\Delta_0^2)/(1/\Delta_0^2 - 1/\Delta_\infty^2)$.

The Hubbard and Blundell (1987) and Flores et al. (1993) data demonstrate statistically insignificant differences between α^2 and $1/(1 - q_\infty)\,(\ln\Delta_\infty^2 - \ln\Delta_0^2)/(1/\Delta_0^2 - 1/\Delta_\infty^2)$ estimates, but the data of Chotia and Lesk (1986) deviate.

To test the applicability of Equation (21), which is derived under the assumption of limited diffusion, to the three datasets (Chotia and Lesk 1986; Hubbard and Blundell 1987; Flores et al. 1993), the equation

$$\frac{q(d) - q_\infty}{1 - q_\infty} \approx u\left(\frac{d}{1 - q_\infty}\right) = \frac{1}{\dfrac{d}{1 - q_\infty} + 1}, \qquad (32)$$

relating the distance and identity fraction (Grishin 1995) is used. Substitution of (32) into (21) leads to

$$\Delta_{ij}^2 = \Delta_\infty^2 - (\Delta_\infty^2 - \Delta_0^2)e^{-\frac{\alpha^2}{\Delta_\infty^2}\frac{(1-q_\infty)(1-q)}{q-q_\infty}}. \qquad (33)$$

The two best fit parameters Δ_∞^2 and α^2 of Equation (33) (Δ_0^2 is estimated as an arithmetical mean of squares of RMS deviations for identical proteins), and the two best fit parameters a and b of exponential approximation

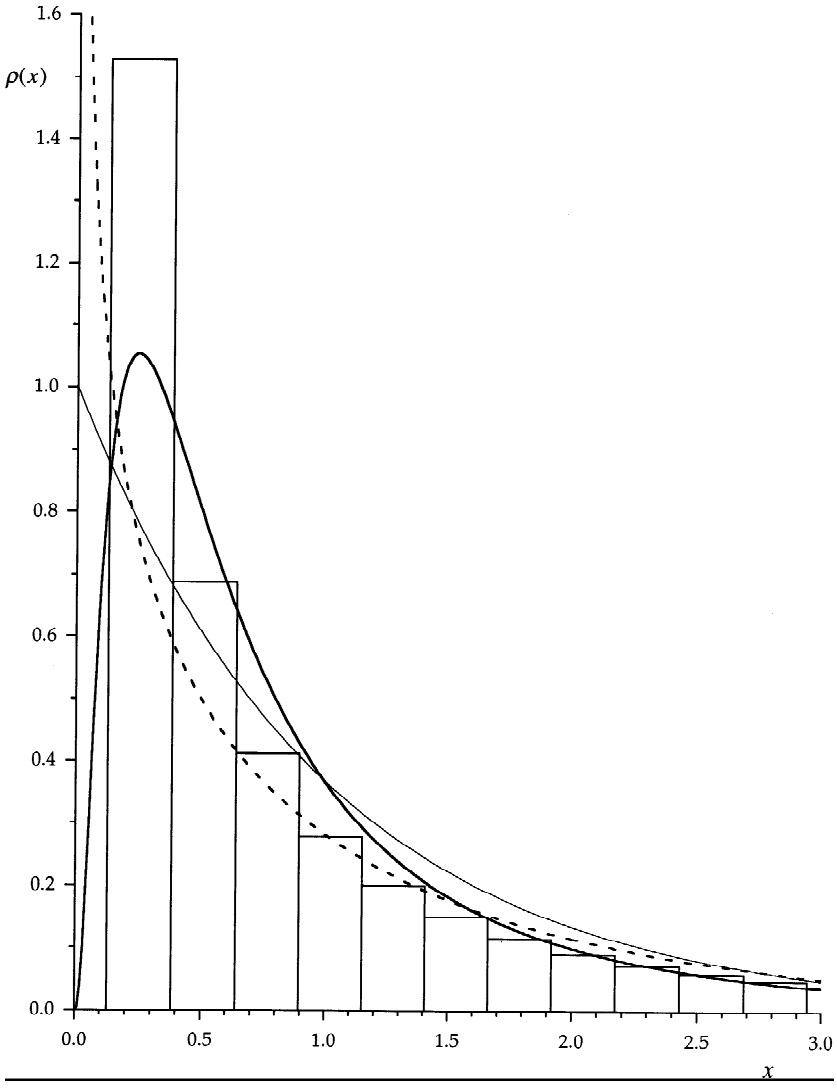

Fig. 2. Distribution of relative substitution rates $\lambda/\bar{\lambda}$ over sites. Histogram is an Equation (31) approximation of density (27) for $\beta = 10$. The thick curve corresponds to log-normal density (30) and the dashed curve is a gamma density approximation (29). For all curves $\beta = 10$. The thin curve represents exponential distribution $\rho(x) = e^{-x}$.

(24) for the data of Chotia, Hubbard, and Flores are presented in Table 2 (identity fraction between the two infinitely distant sequences ($d \to \infty$) is assumed to be $q_\infty = 0.05$). The RMS residuals for the Equation (33) fit are the same (or smaller for data of Chotia and Lesk 1986) as for the exponential fit (24), validating to a certain extent the applicability of the theoretical considerations that lead to Equation (21).

The values of Δ_0^2, Δ_∞^2, and α^2 depend on the definition of topolically equivalent residue pairs, but the relationship between d estimated from Δ with the formula (21) and q should not[3] depend on this definition. Therefore, the three data sets for the estimation of the distribution of substitution rates can be combined. Since the RMS deviation for a given identity fraction has a large variance and the three datasets have an unequal number of data points, the data in each data set were grouped into 12 intervals[4] of an identity fraction. For each interval the mean RMS deviation squared for points falling into this interval was attributed to the mean value of the identity fraction of these points (inset in Fig. 1, two data points, corresponding to identity fractions 0.175 and 0.109 from Flores et al. (1993) were excluded from the further analysis as outliers). RMS deviations Δ obtained by grouping were transformed to distances d with the formula (21) using the estimates of Δ_0^2, Δ_∞^2, and α^2 (Table 1). Identity fractions q were transformed to fractions of unchanged sites u with the formula (7). The distances d were plotted against the fractions of unchanged sites u in Fig. 1. The relation between transformed data (u, d) (Fig. 1) depends

[3] Complication arises from the fact, that definitions of the subset depend on the identity between sequences. For identical proteins all residues are included and for distant ones only 50% of residues with smallest deviations are used. That underestimates RMS deviation for pairs with lower identity fraction.

[4] The intervals were q = 0–0.15; 0.16–0.2; 0.21–0.25; 0.26–0.3; 0.1 * (n-0.9)–0.1 * n, n = 4, 5, 6, 7, 8, 9; 0.91–0.99; and 1.

on the data set much less than the relation between the untransformed data (q, Δ) (inset in Fig. 1). Transformed datapoints (u, d) were fitted to equation (25) in a form:

$$d = -\frac{\beta - 1}{\ln \beta} \ln \left(\frac{\beta}{\beta - 1} (1 - \beta^{-u}) \right). \qquad (34)$$

The best fit curve (Fig. 1, curve 1) with $\beta = 9.92 \pm 0.82$ is very similar to the relationship $d = 1/u - 1$ (Fig. 1 curve 2), which originated from the assumption of exponential distribution of substitution rates among sites. Geometric distribution $(1 - p)p^n$ adequately describing the data on the analysis of protein sequences by Holmquist et al. (1983) is generated when the Poisson parameter λ varies according to the exponential distribution. The exponential distribution is the distribution with minimal entropy when the mean value is a constant (Grishin 1995). Differences between the two curves are significant for the low identity fraction. To obtain more accurate relations and estimates of parameters than presented here, larger data sets should be analyzed. However, even the limited data sets demonstrate the applicability of the proposed method and the derived equations.

Estimates of the distribution of relative substitution rates among sites in the form of a histogram (34) and of continuous functions (29, $a = 0.6417$) and (30, $a = 0.9394$) for $\beta = 10$ are plotted in Fig. 2. The exponential distribution is shown for comparison. These results demonstrate that most of the sites in the sequence (more than expected from exponential distribution) tend to be conserved and have a substitution rate lower than the mean substitution rate.

Conclusions

New equations are derived to estimate the number of amino acid substitutions per site between two homologous proteins d from the RMS deviation between their two spatial structures Δ:

$$d = \frac{\Delta_\infty^2}{\alpha^2} \ln \frac{\Delta_\infty^2 - \Delta_0^2}{\Delta_\infty^2 - \Delta^2} \pm \sqrt{\frac{2}{s}} \frac{\Delta_\infty^2 \Delta^2}{\alpha^2 (\Delta_\infty^2 - \Delta^2)},$$

and the fraction of identical residues between their two sequences q:

$$d = -(1 - q_\infty) \frac{\beta - 1}{\ln \beta} \ln \left(\frac{\beta}{\beta - 1} \left(1 - \beta^{-\frac{q - q_\infty}{1 - q_\infty}} \right) \right)$$

$$\pm \frac{\beta - 1}{1 - \beta^{-\frac{q - q_\infty}{1 - q_\infty}}} \sqrt{\frac{q(1-q)}{l}}.$$

Values of the parameters Δ_0, Δ_∞, and α^2 depend on the definition of topologically equivalent residue pairs, and the value of $\beta = \Delta_\infty^2/\Delta_0^2$ is estimated to be equal to 10 ± 1.

Acknowledgment. The author thanks Dr. Vladislav Markin for the thorough inspection of the manuscript and a number of helpful suggestions on reorganization of the material, Dr. Vyacheslav Grishin for the help with mathematical questions, Dr. Emile Zuckerkandl for critical remarks and constant encouragement, Dr. Andrei Osterman for stimulating discussions, and Lisa Kinch for reading the manuscript critically.

References

Barry D, Hartigan JA (1987) Asynchronous distance between homologous DNA sequences. Biometrics 43:261–276

Chotia C, Lesk A (1986) The relation between the divergence of sequence and structure in proteins. The EMBO J 5:823–826

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, 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, 5, Suppl 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. Protein Science 2:1811–1826

Grishin NV (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 41:675–679

Gutin AM, Badretdinov AY (1994) Evolution of protein 3D structures as diffusion in multidimensional conformational space. J Mol Evol 39:206–209

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

Hubbard TJP, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling. Protein Engineering 1:159–171

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Lesk AM, Chotia CH. (1986) The response of protein structure to amino-acid sequence changes. Phil Trans R Soc Lond A 317:345–356

Olsen GJ (1987) Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symposia on Quantitative Biology 52:825–837

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

Rzhetsky A, Nei M. (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Saitou N (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. J Mol Evol 27:261–273

Takacs L. (1966) Stochastic process. Methuen & Co LTD, London, John Wiley & Sons Inc., NY

Tajima F, Takezaki N (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol Biol Evol 11:278–286

Tateno Y, Takezaki N, Nei M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony

methods when substitution rate varies with site. Mol Biol Evol 11:261–277
Uzzel T, Corbin KW (1971) Fitting discrete probability distribution to evolutionary events. Science 172:1089–1096
Wilbur WJ (1985) On the PAM matrix model of protein evolution. Mol Biol Evol 2:434–447
Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401
Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314
Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39:315–329
Zuckerkandl E, Pauling L. (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vodel HJ (eds) Evolving genes and proteins. Academic Press, NY, pp 97–166

Appendix 1. Symbol Definitions

t = time, $t \geqslant 0$;
d = distance between two sequences, $d \geqslant 0$;
u = fraction of unchanged sites in two sequences, $0 \leqslant u \leqslant 1$;
p = fraction of sites, occupied by different residues in two sequences, $0 \leqslant p \leqslant 1$;
q = fraction of identical sites in two sequences, $q = 1 - p$;
q_∞ = mean fraction of identical residues between two infinitely distant ($d \to \infty$) sequences with the same fold, $0 < q_\infty < 1$;
l = number of sites in a protein, $l > 0$;
s = number of secondary structure elements in a protein, $0 < s < l$;
λ = substitution rate, $\lambda \geqslant 0$;
$\overline{\lambda}$ = mean substitution rate, $\overline{\lambda} > 0$;
x = relative substitution rate, $x = \lambda/\overline{\lambda}$;
$\mathbf{X}$ = 20×20 matrix of relative substitution rates;
ξ = eigenvalue of $\mathbf{X}$, $\xi \leqslant 0$;
$\rho(x)$ = probability density function of relative substitution rates;
$M^{(\nu)}$ = ν^{th} moment of $\rho(x)$;
Δ = RMS deviation between two structures, $\Delta \geqslant 0$;
Δ_0 = mean RMS deviation between independently refined structures of the same protein, $\Delta_0 \geqslant 0$;
Δ_∞ = mean RMS deviation between structures of two infinitely distant ($d \to \infty$) proteins with the same fold, $\Delta_\infty \geqslant \Delta_0$;
α^2 = mean sum of square deviations introduced by a single amino acid substitution;
$\mathbf{r}(k)$ = radius vector of a site k in a spatial structure;
$\Delta\mathbf{r}$ = difference vector;
$\rho(\Delta\mathbf{r},t)$ = probability density function of difference vectors;
$\rho(\Delta^2, d)$ = probability density function of RMS deviations squared;
D = diffusion coefficient;
$\mathbf{j}$ = diffusion flux;
$\mathbf{f}$ = force;
β = parameter, that is close to $\Delta_\infty^2/\Delta_0^2$, $\beta \geqslant 1$;
γ = parameter, $\gamma = \alpha^2/\Delta_\infty^2$;
$\Gamma(y)$ = gamma function;

y is a real variable, $g(y)$, $v(\mathbf{x},\tau)$ are real functions, $\mathbf{x}$ is a 3-dimensional vector, a, b, c, and τ are real numbers, i, j, and k are integers;

$\cdot$ = denotes scalar product

Appendix 2

The diffusion flux in an elastic force field [substitute (20) into (19)] is

$$\mathbf{j} = -D\mathbf{grad}\rho - a\Delta\mathbf{r}\rho,\ a \geqslant 0. \tag{A2-1}$$

The diffusion equation changes (substitute (A2-1) into (14)) to

$$\frac{\partial\rho}{\partial t} = D\nabla^2\rho + a\mathbf{grad}\rho \cdot \Delta\mathbf{r} + 3a\rho. \tag{A2-2}$$

The change of variables

$$\rho(\Delta\mathbf{r},t) = e^{3at}v\left(e^{at}\Delta\mathbf{r}, \frac{e^{2at}-1}{2a}\right), \tag{A2-3}$$

where $v = v(\mathbf{x},\tau)$ is a new function, $\mathbf{x}$ is an arbitrary 3-dimensional vector, and $\tau \geqslant 0$, leads to the standard diffusion equation

$$\frac{\partial v}{\partial\tau} = D\nabla^2 v, \tag{A2-4}$$

where all derivatives are computed at the point

$$\left(e^{at}\Delta\mathbf{r}, \frac{e^{2at}-1}{2a}\right).$$

From (A2-3) if $t = 0$, then $v(\mathbf{x},0) = \rho(\Delta\mathbf{r},0)$. Solution of the diffusion equation (A2-4) under the initial condition (13) is

$$v(\mathbf{x},\tau) = (2\pi(2D\tau + \Delta_0^2))^{-\frac{3}{2}} e^{-\frac{\mathbf{x}^2}{2(2D\tau+\Delta_0^2)}}, \tag{A2-5}$$

which can be converted using (A2-3) to

$$\rho(\Delta\mathbf{r},t) = (2\pi\Delta^2(t))^{-\frac{3}{2}} e^{-\frac{\Delta\mathbf{r}^2}{2\Delta^2(t)}}, \tag{A2-6}$$

where

$$\Delta^2(t) = \Delta_\infty^2 - (\Delta_\infty^2 - \Delta_0^2)e^{-\frac{2D}{\Delta_\infty^2}t}, \quad \Delta_\infty^2 = \frac{D}{a}. \tag{A2-7}$$

Appendix 3

Combination of Equations (24) and (21) leads to

$$ae^{b(1-q)} = \Delta_\infty^2 - (\Delta_\infty^2 - \Delta_0^2)e^{-\frac{\alpha^2}{\Delta_\infty^2}d}. \tag{A3-1}$$

Coefficients a and b could be expressed in terms of Δ_∞^2, Δ_0^2, α^2. If $d = 0$, then $q = 1$ and from (A3-1)

$$a = \Delta_0^2. \tag{A3-2}$$

Differentiating Equation (A3-1) with respect to d gives

$$-abe^{b(1-q)}\frac{\mathrm{d}q}{\mathrm{d}d}=\alpha^2\left(1-\frac{\Delta_0^2}{\Delta_\infty^2}\right)e^{-\frac{\alpha^2}{\Delta_\infty^2}d}. \tag{A3-3}$$

Combining Equations (5), (A3-2) and (A3-3) for $d = 0$ ($q = 1$) yields

$$b=\frac{\alpha^2}{\Delta_\infty^2}\left(\frac{\Delta_\infty^2}{\Delta_0^2}-1\right). \tag{A3-4}$$

Substituting (A3-4) and (A3-2) into (A3-1) gives

$$e^{\frac{\alpha^2}{\Delta_\infty^2}\left(\frac{\Delta_\infty^2}{\Delta_0^2}-1\right)(1-q)}=\frac{\Delta_\infty^2}{\Delta_0^2}-\left(\frac{\Delta_\infty^2}{\Delta_0^2}-1\right)e^{-\frac{\alpha^2}{\Delta_\infty^2}d}. \tag{A3-5}$$

Designating

$$\beta=\frac{\Delta_\infty^2}{\Delta_0^2},\ \beta\geqslant 1$$

and

$$\gamma=\frac{\alpha^2}{\Delta_\infty^2}$$

transform equation (A3-5) to

$$1-q=\frac{\ln\beta+\ln\left(1-\frac{\beta-1}{\beta}e^{-\gamma d}\right)}{\gamma(\beta-1)}. \tag{A3-6}$$

If $d \to +\infty$, then $q \to q_\infty$ and equation (A3-6) can be transformed to

$$\gamma=\frac{1}{1-q_\infty}\frac{\ln\beta}{\beta-1}. \tag{A3-7}$$

Substitution of (A3-7) into (A3-6) with the help of (10) leads to

$$u=\frac{-\ln\left(1-\frac{\beta-1}{\beta}e^{-\frac{\ln\beta}{\beta-1}d}\right)}{\ln\beta}. \tag{A3-8}$$