BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 7 2005, pages 975–980
doi:10.1093/bioinformatics/bti109

*Phylogenetics*

# An alternative model of amino acid replacement

Gavin E. Crooks* and Steven E. Brenner

Department of Plant and Microbial Biology, 111 Koshland Hall #3102, University of California, Berkeley, CA 94720-3102, USA

Received on October 2, 2004; revised and accepted on October 25, 2004
Advance Access publication November 5, 2004

**ABSTRACT**

**Motivation:** The observed correlations between pairs of homologous protein sequences are typically explained in terms of a Markovian dynamic of amino acid substitution. This model assumes that every location on the protein sequence has the same background distribution of amino acids, an assumption that is incompatible with the observed heterogeneity of protein amino acid profiles and with the success of profile multiple sequence alignment.

**Results:** We propose an alternative model of amino acid replacement during protein evolution based upon the assumption that the variation of the amino acid background distribution from one residue to the next is sufficient to explain the observed sequence correlations of homologs. The resulting dynamical model of independent replacements drawn from heterogeneous backgrounds is simple and consistent, and provides a unified homology match score for sequence–sequence, sequence–profile and profile–profile alignment.

**Contact:** gec@compbio.berkeley.edu

## INTRODUCTION

During evolution, a protein's amino acid sequence is altered by the insertion and deletion of residues and by the replacement of one residue by another. In principle, the alignment of protein sequences and the subsequent detection of protein homologs and the inference of protein phylogenies requires a dynamical model of this sequence evolution. The most common and widely used residue replacement dynamics is the standard Dayhoff model, which assumes that the substitution probability during some time interval depends only on the identities of the initial and replacement residues and that the dynamics is otherwise homogeneous along the protein chain, between protein families and across evolutionary epochs. In other words, under this model the dynamics of amino acid substitution resembles a continuous time, first order Markov chain (Dayhoff *et al*., 1972, 1978; Gonnet *et al*., 1992; Jones *et al*., 1992; Müller and Vingron, 2000).

However, it has long been known that this widely used Markovian substitution model is fundamentally unsatisfactory. One major problem is that the short and long time substitution dynamics are incompatible (Gonnet *et al*., 1992; Benner *et al*., 1994; Müller and Vingron, 2000). Benner *et al*. (1994) suggest that this is because at short evolutionary times the patterns of substitution are influenced by single base mutations between neighboring codons, whereas for more diverged sequences the genetic code is irrelevant and the patterns of replacement are dominated by the selection of chemically and structurally compatible residues.

A more serious problem with the Dayhoff Markovian model is that it assumes that every residue in every protein has the same background distribution of amino acids and that protein sequences rapidly evolve to this uninteresting equilibrium. In actuality, the amino acid background distribution varies markedly from one residue position to the next, as can be seen, for example, in Figure 1. These large site-to-site variations are stable across relatively long evolutionary time-scales, and they account for the success of protein hidden Markov models and other profile based multiple sequence alignment methods. (See, for example, Sjölander *et al*., 1996; Durbin *et al*., 1998.) Profile methods can detect substantially more remote homologies than pairwise alignment (Park *et al*., 1998; Green and Brenner, unpublished data). In short, the dynamics of amino acid substitution are not Markovian, stationary, nor homogeneous, and the prediction of rapidly decaying sequence correlations is at odds with the success of profile based remote homology detection.

A natural solution to the limitations of the Markov model is to assume that residue replacement is governed by different Markov processes for each position, each process potentially possessing its own background distribution and substitution probabilities. The appropriate Markov matrix for a particular protein position is chosen based upon predictions of the protein structure, or directly from the sequence data (Goldman *et al*., 1996, 1998; Thorne *et al*., 1996; Topham *et al*., 1997; Koshi and Goldstein, 1998; Dimmic *et al*., 2000; Lartillot and Philippe, 2004). However, this approach is both computationally and conceptually complex.

Here, we propose that the observed sequence correlation between diverged homologs is principally due to the heterogeneous, stable, background distribution of each protein site and, therefore, that a Markovian amino acid replacement dynamics is overly complicated and possibly unnecessary for the accurate construction of protein sequence alignments and phylogenies. As an alternative, we construct a dynamical model of amino acid replacement that explicitly assumes that each protein site has a different equilibrium distribution of the 20 canonical amino acids (which we refer to as that site's amino acid background, $\theta$) and that the residue distribution of each site rapidly (relative to evolutionary time-scales) relaxes to this local, site-specific equilibrium. These distributions themselves conform to a probability distribution of backgrounds, $P(\theta)$, which we may discover by studying many families of homologous proteins (Fig. 2). We do not need to model the short time dynamics with any great accuracy, since the alignment of highly conserved homologs is relatively straightforward. Therefore, we will assume that replacement residues

*To whom correspondence should be addressed.


**Fig. 1.** The amino acid background distribution of a site within a protein is often stable across large evolutionary time-scales, but varies markedly from one site to another. This figure illustrates the helix-turn-helix motif from the CAP family of homodimeric DNA binding proteins. The height of each letter corresponds to the amino acid frequency in a multiple alignment of 100 diverse, homologous sequences. (For details, see Crooks *et al*., 2004a,b; Schneider and Stephens, 1990.) These background distributions are determined by structural, functional and evolutionary constraints. For example, positions 180, 181 and 185 are critical to the sequence-specific binding of the protein to DNA, the conserved glycine at position 177 is located on the inside of the turn between the helices, and the buried sites 172, 176, 178, 183, 187 and 190 contain mostly hydrophobic residues. It should be noted that the correlations inherent in these background distributions are far stronger than can be explained by local structural features (such as burial and secondary structure) alone (Crooks & Brenner, 2004a,b).

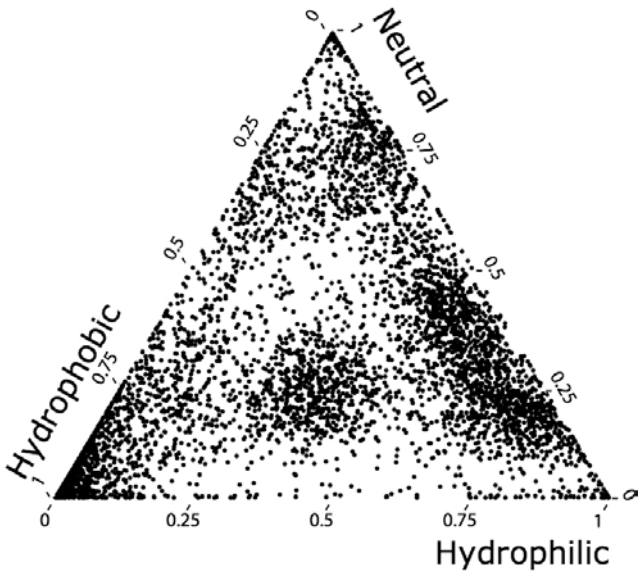


**Fig. 2.** The distribution of amino acid backgrounds is heterogeneous and multimodal. This ternary scatter plot represents 5000 randomly sampled distributions drawn from the `dist. 20comp` Dirichlet mixture model (Karplus, 1995, http://www.cse.ucsc.edu/research/compbio/dirichlets/) of amino acid backgrounds. Each has been projected onto the three-dimensional subspace of hydrophobic (C, F, I, L, M, V, W, Y), neutral (A, G, P, S, T) and hydrophilic (D, E, H, K, N, Q, R) residues. Major peaks in the probability density are located around (0.5, 0.25, 0.25) hydrophobic/neutral/hydrophilic, and towards the vertices of this ternary plot. Thus, if we observe several homologous hydrophilic residues (e.g. position 179 or 184 in Fig. 1) we can be reasonably confident that additional homologous residues will also be hydrophilic.

are randomly sampled from the background distribution of that site. Consequentially (and in direct contrast to the Markov model) the replacement residue is conditionally independent of the initial amino acid at all times. Note that multiple substitutions are statistically equivalent to a single substitution (since a single mutation is sufficient to relax a site to local equilibrium) and that a residue can be replaced by the same amino acid type. The resulting dynamic is a site-specific, continuous time, zeroth order Markov chain, similar in spirit to Felsenstein's (1981) model of nucleic acid substitution. The crucial difference is that the initial and replacement residues are non-trivially correlated because both have been sampled from the same background distribution, whereas non-homologous residues are drawn from different backgrounds. Bruno (1996) has used essentially the same dynamical model discussed here, albeit without incorporating the background prior distribution, to find maximum likelihood estimates of site-specific amino acid frequencies.

Under our residue replacement model, the principal origin of sequence correlation between diverged homologs is the background distribution of each protein site. This is also the central idea underlying profile based multiple sequence alignment algorithms. Therefore, we are not proposing a radically different method for homolog detection or sequence alignment; rather we are proposing a concrete and consistent dynamics for the underlying evolutionary process. The implications of this dynamics can be readily extended to cover not only profile–sequence based alignment, but also profile–profile and pairwise sequence–sequence alignment. Moreover, when we consider pairwise, sequence alignment below, we find that our model is essentially equivalent to the standard pairwise alignment methods, as they are used in practice. This alternative dynamical model of amino acid replacement is biologically reasonable, conceptually straightforward and can adequately explain many of the observed patterns of homolog sequence correlation without invoking a Markovian dynamic.

The correlations between pairs of homologous residues can be summarized by an amino acid substitution matrix, $S$, whose entries represent the log probability of observing the homologous pair of amino acids $q_{ij}$ in a properly aligned pair of homologous proteins, against the probability $p_i\,p_j$ of independently observing the residues in unrelated sequences (Altschul, 1991):

$$S_{ij} = \frac{1}{\lambda} \log \frac{q_{ij}}{p_i\,p_j}. \qquad (1)$$

Units of one-third bits are traditional for substitution matrices (base 2 logarithm, $\lambda = 1/3$, $\approx \frac{1}{10}$ digits), although the scaling is arbitrary. Assuming conditionally independent replacements, we can directly construct the large time limit substitution matrix from the background probability distribution, $P(\theta)$. Fortunately, this large distribution has previously been investigated and parameterized by fitting many columns from multiple alignments of homologous protein sequences to a mixture of Dirichlet distributions (Karplus, 1995; Durbin *et al*., 1998). Figure 2 displays a projection of the `dist.20comp` parameterization (Karplus, 1995) and Figure 3 displays the corresponding substitution matrix. The mathematical details of matrix construction are given below.

At shorter evolutionary times there is a significant chance that no mutation has occurred at all, resulting in an enhanced probability of amino acid conservation. Let $c$ be the probability of zero mutation events; then the substitution matrix, adjusted for the possibility of zero mutations, is

$$S_{ij}(c) = \log \frac{c p_i \delta_{ij} + (1-c) q_{ij}}{p_i\,p_j}, \qquad (2)$$

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -2 | -2 | -2 | 0 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -1 | -2 | -1 | 0 | -1 | -3 | -2 | -1 | -2 | -1 | -1 |
| R | -2 | 6 | -1 | -1 | -4 | 1 | 0 | -3 | 0 | -3 | -3 | 2 | -2 | -3 | -2 | -1 | -1 | -2 | -2 | -3 | 3 | -1 | -1 |
| N | -2 | -1 | 7 | 1 | -3 | 0 | 0 | -1 | 0 | -5 | -4 | 0 | -3 | -4 | -2 | 0 | -1 | -3 | -2 | -4 | 3 | -1 | -1 |
| D | -2 | -1 | 1 | 7 | -4 | 0 | 1 | -1 | -1 | -6 | -5 | 0 | -4 | -5 | -1 | 0 | -1 | -4 | -3 | -5 | 0 | -2 | -2 |
| C | 0 | -4 | -3 | -4 | 12 | -3 | -4 | -3 | -3 | -1 | -2 | -4 | -1 | -2 | -3 | -2 | -1 | -2 | -2 | 0 | -3 | 5 | -2 |
| Q | -1 | 1 | 0 | 0 | -3 | 6 | 1 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -2 | -2 | -3 | 0 | 1 | -1 |
| E | -1 | 0 | 0 | 1 | -4 | 1 | 5 | -2 | -1 | -4 | -4 | 1 | -3 | -4 | -1 | -1 | -1 | -3 | -3 | -4 | 0 | -1 | -1 |
| G | -1 | -3 | -1 | -1 | -3 | -2 | -2 | 7 | -2 | -6 | -5 | -2 | -4 | -5 | -2 | -1 | -2 | -4 | -4 | -5 | -2 | -2 | -2 |
| H | -2 | 0 | 0 | -1 | -3 | 0 | -1 | -2 | 9 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -1 | 0 | 0 | -3 | 0 | -1 | -1 |
| I | -2 | -3 | -5 | -6 | -1 | -3 | -4 | -6 | -3 | 5 | 2 | -4 | 1 | 0 | -4 | -4 | -2 | -1 | -1 | 3 | -4 | -2 | -2 |
| L | -2 | -3 | -4 | -5 | -2 | -3 | -4 | -5 | -3 | 2 | 5 | -3 | 2 | 1 | -3 | -3 | -2 | -1 | -1 | 1 | -4 | -2 | -2 |
| K | -2 | 2 | 0 | 0 | -4 | 1 | 1 | -2 | -1 | -4 | -3 | 5 | -2 | -4 | -1 | -1 | -1 | -3 | -3 | -3 | 1 | -1 | -1 |
| M | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -2 | 1 | 2 | -2 | 7 | 1 | -3 | -2 | -1 | 0 | 0 | 1 | -3 | -2 | -1 |
| F | -2 | -3 | -4 | -5 | -2 | -3 | -4 | -5 | -1 | 0 | 1 | -4 | 1 | 7 | -3 | -3 | -2 | 3 | 3 | 0 | -3 | -2 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -4 | -3 | -1 | -3 | -3 | 8 | -1 | -2 | -3 | -3 | -3 | -2 | -2 | -2 |
| S | 0 | -1 | 0 | 0 | -2 | 0 | -1 | -1 | -1 | -4 | -3 | -1 | -2 | -3 | -1 | 4 | 1 | -3 | -2 | -3 | 0 | -1 | -1 |
| T | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | -1 | -2 | -2 | -1 | -1 | -2 | -2 | 1 | 5 | -2 | -2 | -1 | -1 | -1 | -1 |
| W | -3 | -2 | -3 | -4 | -2 | -2 | -3 | -4 | 0 | -1 | -1 | -3 | 0 | 3 | -3 | -3 | -2 | 12 | 3 | -2 | -3 | -2 | -1 |
| Y | -2 | -2 | -2 | -3 | -2 | -2 | -3 | -4 | 0 | -1 | -1 | -3 | 0 | 3 | -3 | -2 | -2 | 3 | 8 | -2 | -2 | -2 | -1 |
| V | -1 | -3 | -4 | -5 | 0 | -3 | -4 | -5 | -3 | 3 | 1 | -3 | 1 | 0 | -3 | -3 | -1 | -2 | -2 | 5 | -4 | -2 | -2 |
| B | -2 | 3 | 3 | 0 | -3 | 0 | 0 | -2 | 0 | -4 | -4 | 1 | -3 | -3 | -2 | 0 | -1 | -3 | -2 | -4 | 3 | -1 | -1 |
| Z | -1 | -1 | -1 | -2 | 5 | 1 | -1 | -2 | -1 | -2 | -2 | -1 | -2 | -2 | -2 | -1 | -1 | -2 | -2 | -2 | -1 | 3 | -1 |
| X | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -2 | -1 | -2 | -2 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | -1 | -1 |

**Fig. 3.** This substitution matrix [Equation (1), a conventional description of amino acid replacement propensities] has been directly constructed from the `dist.20comp` Dirichlet mixture model of amino acid background probabilities (Fig. 2; Karplus, 1995), using the conditionally independent substitution model of amino acid replacement [Equations (3)–(10)]. As a consequence of the heterogeneity and stability of amino acid background probabilities—illustrated in Figures 1 and 2—the amino acid identity of a pair of alignable, homologous residues is non-trivially correlated over long evolutionary time-scales, simply because both residues are drawn from the same background. Scores are in units of $\frac{1}{3}$ bits, rounded to the nearest integer.

where $\delta_{ij}$ is the Kronecker delta function. A reasonable default model for the conservation probability $c$ would be to assume that substitutions are Poissonian. Then $c = \exp(-t/\tau)$, where $\tau$ is the mean time between replacements. Note that although replacement is Markovian (albeit zeroth order), the dynamic decay of residue correlations at a position is not, due to the heterogeneity of the amino acid background at that position (an unobserved hidden variable).

Equation (2) gives the log odds of aligned residues, given the inter-sequence divergence. Conversely, given a prior on the parameter $c$ and a fixed alignment we can invert Equation (2) and estimate the divergence between sequences. Conserved residues indicate small divergences and unconserved pairs argue for large divergences, although different pairs are weighted differently.

As evolution proceeds, the background distribution of a site may itself change, due, for example, to a change in structure of that part of the protein. This will result in a loss of homology signal under our model, and it may no longer be possible to align the diverged residues, nor to recognize them as homologs. This is schematically illustrated in Figure 4.

Various families of substitution matrices have been developed, including PAM, BLOSUM and VTML. Different members of the same family represent different degrees of sequence divergence. In principle, we should match the divergence inherent in the substitution matrix to the divergence of the pair of sequences we wish to align (Altschul, 1993). However, this is computationally expensive, and, in practice, a single matrix is chosen based on its ability to align remote homologs, on the grounds that matching close homologs is relatively easy (Brenner *et al*., 1998). Under the Markov model, the chosen matrix has no particular significance. On the other hand, under our model there is a natural, non-trivial, long time limit matrix [Fig. 3; Equation (2), $c = 0$]. This matrix represents the sequence

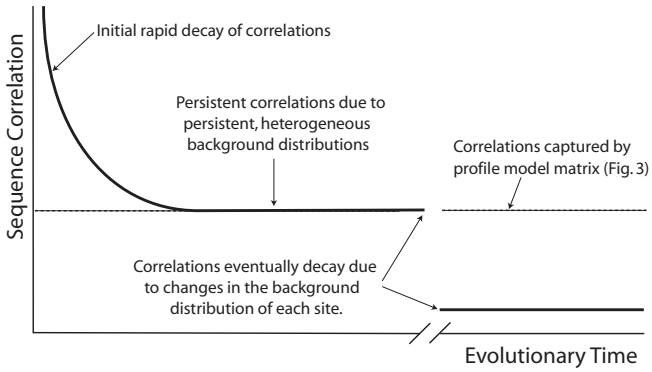


**Fig. 4.** Schematic representation of the decay of sequence correlation with evolutionary time. There is an initial rapid reduction in correlation on the time-scale of single residue substitutions. Under the standard Markov model, this exponential decay would continue. However, under the profile model the correlations instead limit towards a plateau value, due to the heterogeneity of the background amino acid distribution. These are the correlations captured by the substitution matrix of Figure 3. Finally, over a second, much longer time-scale, the sequence correlations decay towards insignificance due to changes in the site-specific background distribution.

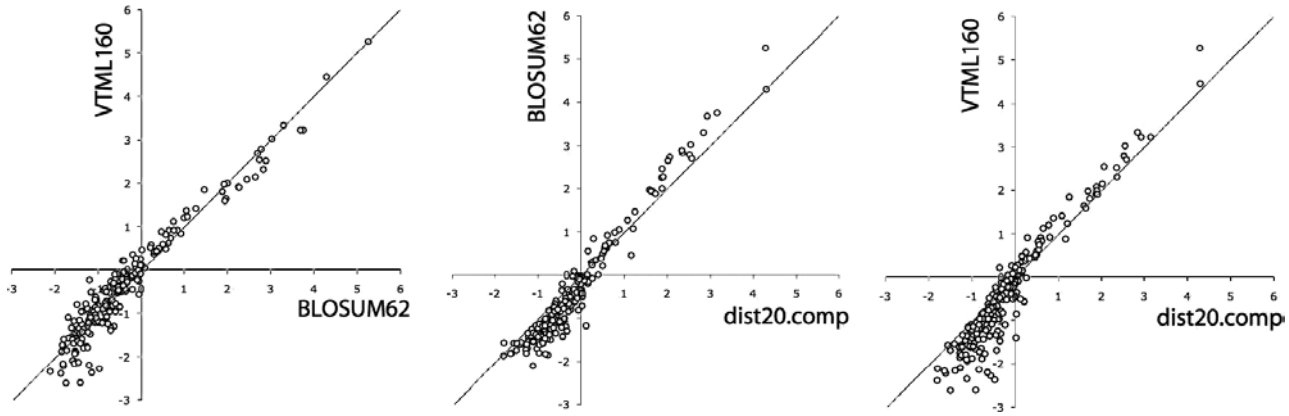


**Fig. 5.** Comparison between the bit scores of BLOSUM62 (Henikoff and Henikoff, 1992), VTML160 (Müller *et al*., 2002) and `dist.20comp` (Fig. 3) substitution matrices. All three matrices reflect similar levels and patterns of sequence divergence, but have been derived using very different approaches. The BLOSUM matrices are empirical, the VTML family are based upon the Markov model of amino acid substitution and the `dist.20comp` matrix is based upon the conditionally independent substitution model.

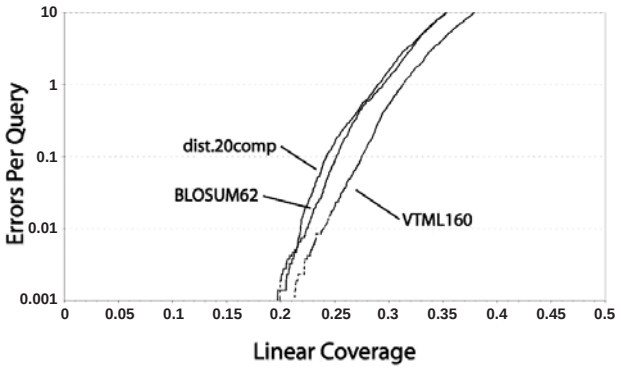


**Fig. 6.** The substitution matrices, BLOSUM62, VTML160 and `dist.20comp`, are of comparable effectiveness under Green and Brenner's (2002) evaluation of pairwise remote homology detection. A set of about 2800 sequences (none of which share more than 40% sequence identity) are collated from the SCOP (Structural Classification of Proteins) database (version 1.57) (Murzin *et al*., 1995; Brenner *et al*., 2000). SCOP reliably clusters these sequences into groups of homologs using structural information. Each sequence is matched against the dataset using Smith–Waterman alignment (Smith and Waterman, 1981), a particular substitution matrix and appropriate gap penalties (Green and Brenner, 2002). The results are shown as a plot of errors per query against (linearly normalized) coverage, the average fraction of true homologs that are found for each sequence. There is a trade-off between accuracy and coverage; the bottom right of the above graph is ideal; high coverage with few errors. We used Bayesian bootstrap resampling to estimate that the standard deviation of the coverage is about 0.02 at 0.01 errors per query (Green and Brenner, 2002; Zachariah *et al*., 2005, Price, Crooks, Green and Brenner, unpublished). Thus, there is a statistically significant, but relatively small and (in practice) unimportant variation in homology coverage between the three matrices. Note that both the BLOSUM and the VTML matrices have been directly trained upon pairwise alignment data, and may therefore be favored in this pairwise alignment test.

correlations at any time after the first few mutations, and before the underlying amino acid background itself diverges (Fig. 4).

Figures 5 and 6 demonstrate that the `dist.20comp` matrix represents a similar level of evolutionary divergence, and similar patterns of substitution as BLOSUM62 and VTML160, two substitution matrices commonly used for pairwise sequence alignment and remote homology detection. These three matrices have been created using very different evolutionary models; the `dist.20comp` matrix is based upon our heterogeneous, background/independent substitution model, and the `dist.20comp` background distribution is, in turn, derived from the columns of many multiple alignments of homologous protein sequences; the popular BLOSUM matrices are empirically derived from the BLOCKS database of reliable protein sequence alignments (Henikoff and Henikoff, 1992; Henikoff *et al*., 2000); and the classic PAM (Dayhoff *et al*., 1978) and modern VTML (Müller *et al*., 2002) matrix families are explicitly based upon the Markovian model of amino acid replacement. In a recent evaluation of pairwise remote homology detection, the VTML160 matrix was found to be more effective than any other VTML, PAM or BLOSUM matrix (Green and Brenner, 2002). However, as can be seen in Figure 6, the difference in remote homology detection ability of the three matrices is relatively small.

In summary, the important sequence correlations can be adequately explained by assuming conditionally independent replacements drawn from background distributions that vary from site to site, but are stable over evolutionary time-scales. The standard, Markovian model of amino acid replacement is unnecessary, overly complicated and inconsistent with observed substitution patterns.

This alternative, heterogeneous background, independent substitution model may be particularly useful for simultaneous sequence alignment and phylogenetic tree reconstruction, since it is necessary to align pairs of close homologs at the leaves, and multiply align many remote homologs at the interior nodes of the tree. Therefore, a simple (yet realistic) evolutionary dynamic that is consistent across a wide range of divergence times, and that leads naturally to sequence–sequence, sequence–profile and profile–profile alignment algorithms, may be advantageous.

## MATHEMATICAL DETAILS

A collection of homologous residues can be represented by a 20-component canonical amino acid count vector, $n = \{n_1, \ldots, n_{20}\}$.

The total number of counts can be 1, if the observation is taken from a single sequence, or many if the collection represents an entire column of a multiple sequence alignment or some other related set of residues.

In general, we wish to estimate whether two collections of homologous residues are related, given that detectably homologous residues are drawn from the same background amino acid distribution. The appropriate test statistic is the log odds of sampling the two amino acid count vectors from the same, but unobserved, background distribution, against the probability of independently sampling the two count vectors from different distributions:

$$S(n^1, n^2) = \log \frac{P(n^1, n^2)}{P(n^1)P(n^2)}. \tag{3}$$

The probability of independently sampling a particular collection of homologous residues $n$ from the background amino acid profile from which those residues are drawn, $\theta = \{\theta_1, \ldots, \theta_{20}\}$, follows the multinomial distribution;

$$\begin{aligned} P(n|\theta) &= \mathcal{M}(n|\theta) \\ &= \frac{1}{M(n)} \prod_{i=1}^{20} \theta_i^{n_i}, \quad M(n) = \frac{\prod_i n_i!}{(\sum_i n_i)!}. \end{aligned} \tag{4}$$

This is the multivariant generalization of the common binomial distribution.

The probability distribution of background distributions $P(\theta)$ has been studied and measured by collating columns from many multiple protein sequence alignments. Since this is a very large, multi-modal probability it is necessary to parameterize the distribution into a convenient representation. Typically, a mixture of Dirichlet distributions is used (Karplus, 1995; Sjölander *et al*., 1996; Durbin *et al*., 1998):

$$P(\theta) = \sum_{k=1,m} \rho_k \mathcal{D}(\theta|\alpha^k) \tag{5}$$

The $m$ mixture coefficients $\rho_k$ sum to one. The $k$th Dirichlet distribution is itself parameterized by the 20-component (canonical amino acids) non-negative vector $\alpha^k$,

$$\mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{20} \theta_i^{(\alpha_i - 1)}, \quad Z(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(A)}, \tag{6}$$

where $A = \sum_i \alpha_i$.

Dirichlet mixtures are used to model the background probability partially because Dirichlet distributions are naturally conjugate to the multinomial distribution (and therefore mathematically convenient) and partially because a Dirichlet mixture can approximate the true distribution with a reasonably small number of parameters. The underlying assumption is that the probability distribution is smooth, but lumpy (Fig. 2).

If we do not know the particular background from which the observations have been drawn, then we must average over all backgrounds to find the probability of observing a particular count vector:

$$\begin{aligned} P(n) &= \int \mathrm{d}\theta\, P(n|\theta) P(\theta) \\ &= \int \mathrm{d}\theta \mathcal{M}(n|\theta) \sum_k \rho_k \mathcal{D}(\theta|\alpha^k) \\ &= \int \mathrm{d}\theta \sum_k \rho_k \frac{\prod_{i=1}^{20} \theta_i^{n_i}}{M(n)} \frac{\prod_{i=1}^{20} \theta_i^{(\alpha_i^k - 1)}}{Z(\alpha^k)} \\ &= \sum_k \rho_k \frac{1}{Z(\alpha^k)} \frac{1}{M(n)} \int \mathrm{d}\theta \prod_{i=1}^{20} \theta_i^{(n_i + \alpha_i^k - 1)} \\ &= \sum_k \rho_k \frac{Z(n + \alpha^k)}{Z(\alpha^k) M(n)}. \end{aligned} \tag{7}$$

The last line follows because the product in the previous line is an unnormalized Dirichlet with parameters $(n + \alpha^k)$. Therefore, the integral over $\theta$ must be equal to the corresponding Dirichlet normalization constant, $Z(n+\alpha^k)$. The final result is a mixture of multivariate negative hypergeometric distributions (Johnson and Kotz, 1969). The negative hypergeometric is an under-appreciated distribution [e.g. Equation (11.23) of Durbin *et al*., 1998] which bears the same relation to the hypergeometric as the negative binomial does to the binomial distribution. The multivariant generalization appears in this case as the combination of a Dirichlet and a multinomial. Confusingly, the negative hypergeometric distribution is sometimes called the inverse hypergeometric, an entirely different distribution, and vice versa.

The probability of independently sampling two count vectors, $n^1$ and $n^2$, from the same undetermined background is

$$\begin{aligned} P(n^1, n^2) &= \int \mathrm{d}\theta\, P(n^1|\theta) P(n^2|\theta) P(\theta) \\ &= \int \mathrm{d}\theta \mathcal{M}(n^1|\theta) \mathcal{M}(n^2|\theta) \sum_k \rho_k \mathcal{D}(\theta|\alpha^k) \\ &= \sum_k \rho_k \frac{1}{Z(\alpha^k)} \frac{1}{M(n^1)} \frac{1}{M(n^2)} \\ &\quad \times \int \mathrm{d}\theta \prod_{i=1}^{20} \theta_i^{(n_i^1 + n_i^2 + \alpha_i^k - 1)} \\ &= \sum_k \rho_k \frac{Z(n^1 + n^2 + \alpha^k)}{Z(\alpha^k) M(n^1) M(n^2)}. \end{aligned} \tag{8}$$

Combining Equations (7) and (8) with the log likelihood ratio, Equation (3), generates a generic profile–profile sequence alignment score that is valid whether the number of counts is small or large:

$$S(n^1, n^2) = \log \frac{\sum_k \rho_k \dfrac{Z(n^1 + n^2 + \alpha^k)}{Z(\alpha^k) M(n^1) M(n^2)}}{\sum_k \rho_k \dfrac{Z(n^1 + \alpha^k)}{Z(\alpha^k) M(n^1)} \sum_k \rho_k \dfrac{Z(n^2 + \alpha^k)}{Z(\alpha^k) M(n^2)}} \tag{9}$$

For the particular case that one of the count vectors contains only a single observation, this score reduces to the standard sequence–profile score frequently used by hidden Markov model protein

sequence alignment (Sjölander *et al*., 1996). This is inevitable, since the underlying mathematics is the same.

If both count vectors contain only a single observation, then this profile–profile score reduces to a pairwise substitution matrix. Note, given that $n_x^1 = \delta_{xi}$ and $n_x^2 = \delta_{xj}$ (where $\delta_{xj}$ is a Kronecker delta function), then all but the $j$th element of the product $Z(\delta_{xj} + \alpha^k)/Z(\alpha^k)$ cancels. Thus,

$$
\begin{aligned}
S_{ij} &= \log \tfrac{q_{ij}}{p_i p_j}, \\
p_i &= \textstyle\sum_k \rho_k \frac{\alpha_i^k}{A^k}, \\
q_{ij} &= \begin{cases} \displaystyle\sum_k \rho_k \frac{\alpha_i^k \alpha_j^k}{A^k(A^k+1)}, & i \neq j, \\ \displaystyle\sum_k \rho_k \frac{\alpha_i^k(\alpha_i^k+1)}{A^k(A^k+1)}, & i = j. \end{cases}
\end{aligned}
\tag{10}
$$

Applying Equation (10) to the 20-component Dirichlet mixture, `dist.20comp` generates the pairwise substitution matrix illustrated in Figure 3.

An interesting feature of this model is that it provides a unified homology match score for sequence–sequence, sequence–profile and profile–profile alignment [Equation (9)]. As far as we are aware, this profile–profile score has not been evaluated in a profile–profile alignment algorithm, although it is a natural generalization of the established hidden Markov model profile–sequence score. However, in the large sample limit, Equation (9) reduces to the Jensen–Shannon divergence between the two empirical amino acid distributions, a measure that has shown some promise in profile–profile alignment (Yona and Levitt, 2002; Edgar and Sjölander, 2004; Marti-Renom *et al*., 2004).

## ACKNOWLEDGEMENTS

We would like to thank Richard E. Green, Marcus Zachariah, Robert C. Edgar and Emma Hill for helpful discussions and suggestions. This work was supported by the National Institutes of Health (1-K22-HG00056). GEC received funding from the Sloan/DOE postdoctoral fellowship in computational molecular biology. SEB is a Searle Scholar (1-L-110).

## REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**, 290–300.

Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.*, **7**, 1323–1332.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Bruno,W.J. (1996) Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, **13**, 1368–1374.

Crooks,G.E. and Brenner,S.E. (2004*a*) Protein secondary structure: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Crooks,G.E. and Brenner,S.E. (2004*b*) Measurements of protein sequence-structure correlations. *Proteins*, **57**, 804–810.

Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Dayhoff,M.O., Eck,R.V. and Park,C.M. (1972) A model of evolutionary change in proteins. *Atlas Protein Sequences Structure*, **5**, 89–99.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. *Atlas Protein Sequences Structure*, **5** (Suppl 3), 345–352.

Dimmic,M.W., Mindell,D.P. and Goldstein,R.A. (2000) Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac. Symp. Biocomput.*, 18–29.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Edgar,R.C. and Sjölander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, **20**, 1301–1308.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Goldman,N., Thorne,J.L. and Jones,D.T. (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, **263**, 196–208.

Goldman,N., Thorne,J.L. and Jones,D.T. (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.

Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.

Green,R.E. and Brenner,S.E. (2002) Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proc. IEEE*, **90**, 1834–1847.

Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.

Henikoff,S. and Henikoff,J.G. (1992) Amino-acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Johnson,N.L. and Kotz,S. (1969) *Discrete Distributions*. John Wiley, New York.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Karplus,K. (1995). Regularizers for estimating distributions of amino acids from small samples. Technical report, University of California, Santa Cruz.

Koshi,J.M. and Goldstein,R. (1998) Models of natural mutations including site heterogeneity. *Proteins*, **32**, 289–295.

Lartillot,N. and Philippe,H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.

Marti-Renom,M.A., Madhusudhan,M.S. and Sali,A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.*, **13**, 1071–1087.

Müller,T., Spang,R. and Vingron,M. (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, **19**, 8–13.

Müller,T. and Vingron,M. (2000) Modeling amino acid replacement. *J. Comput. Biol.*, **7**, 761–776.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I.S. and Haussler,D. (1996) Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Thorne,J.L., Goldman,N. and Jones,D.T. (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, **13**, 666–673.

Topham,C.M., Srinivasan,N. and Blundell,T.L. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.

Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.

Zachariah,M.A., Crooks,G.E., Holbrook,S.R. and Brenner,S.E. (2005) A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins*, **58**, 329–338.