Article No. jmbi.1999.2601 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 287, 187–198

JMB

Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure

David D. Pollock[1]*, William R. Taylor[1] and Nick Goldman[2]

[1]*Division of Mathematical Biology, National Institute for Medical Research The Ridgeway, Mill Hill London NW7 1AA, UK*

[2]*Department of Genetics University of Cambridge Cambridge, CB2 3EH, UK*

**Corresponding author*

The identification of protein sites undergoing correlated evolution (coevolution) is of great interest due to the possibility that these pairs will tend to be adjacent in the three-dimensional structure. Identification of such pairs should provide useful information for understanding the evolutionary process, predicting the effects of site-directed substitution, and potentially for predicting protein structure. Here, we develop and apply a maximum likelihood method with the aim of improving detection of coevolution. Unlike previous methods which have had limited success, this method allows for correlations induced by phylogenetic relationships and for variation in rate of evolution along branches, and does not rely on accurate reconstruction of ancestral nodes. In order to reduce the complexity of coevolutionary relationships and identify the primary component of pairwise coevolution between two sites, we reduce the data to a two-state system at each site, regardless of the actual number of residues observed at that site. Simulations show that this strategy is good at identifying simple correlations and at recognizing cases in which the data are insufficient to distinguish between coevolution and spurious correlations. The new method was tested by using size and charge characteristics to group the residues at each site, and then evaluating coevolution in myoglobin sequences. Grouping based on physicochemical characteristics allows categorization of coevolving sites into positive and negative coevolution, depending on the correlation between equilibrium state frequencies. We detected a striking excess of negative coevolution (corresponding to charge) at sites brought into proximity by the periodicity of the α-helix, and there was also a tendency for sites with significant likelihood ratios to be close in the three-dimensional structure. Sites on the surface of the protein appear to coevolve both when they are close in the structure, and when they are distant, implying a role for folding and/or avoidance of quaternary structure in the coevolution process.

Keywords: coevolution; protein residues; protein structure; maximum likelihood; molecular evolution

Introduction

There has been a great deal of recent research on methods for detecting correlated changes in protein sequence evolution (Altschuh *et al*. 1987; Taylor & Hatrick, 1994; Gobel *et al*. 1994; Neher, 1994; Shindyalov *et al*., 1994; Pollock & Taylor, 1997; Pazos *et al*., 1997; Chelvanayagam *et al*., 1997). It is expected that the residues at some sites will strongly affect the evolution of certain other sites which are close in the three-dimensional structure of the protein. At such sites, a substitution which partly destabilizes the protein structure or function could be corrected by a subsequent (or simultaneous) substitution at an adjacent site. For example, a substitution involving reduction of volume in the protein core might cause a destabilizing pocket which only one or a few adjacent residues would be capable of filling without strain. Assuming such relationships exist and are stable through the evolutionary processes of divergence

Present address: D. D. Pollock, Department of Integrative Biology, Valley Life Sciences Building, University of California, Berkeley, CA 94720-3140, USA.

Abbreviations used: MLEs, maximum likelihood estimators; *LR*, likelihood ratio; *RD*, residue disequilibrium value.

E-mail address of the corresponding author: dpollock@socrates.berkeley.edu

0022-2836/99/110187–12 $30.00/0

and diversification, the record of the relationship should eventually be detectable in alignments of extant protein sequences. These coevolved changes are of interest for understanding processes of protein evolution. Detection of these changes is also useful for generating hypotheses of structural/functional relationships between residues, and if they are indeed symptomatic of three-dimensional proximity even a noisy signal might be of aid in conjunction with other methods for protein structure prediction from multiple sequences.

Studies attempting to identify correlated changes have, however, met with only limited success in identifying pairs of sites which are adjacent in the three-dimensional structure (although potentially better results are obtained when applied to docking; Pazos *et al*., 1997). There are a number of complications which could account for this failure, including the possibilities that: (1) distant sites may be as important as adjacent sites in the compensatory process; (2) the number of sites involved in compensation is so large that the pairwise correlation signal is too small to be detected; (3) the coevolutionary relationships between sites change too quickly with evolutionary time; or (4) the methods used are insufficient for separating correlation in the evolutionary process (coevolution) from background noise (false correlation due to random events). If the first three complications account for most of the problem, there is nothing to be done, but the analysis by Pollock & Taylor (1997) indicated that many current methods may inadequately discriminate coevolution from background noise. Thus, improved methodology, particularly incorporation of phylogenetic tree structure, may yet identify paired and coevolving sites considerably better than previous methods.

While the method by Shindyalov *et al*. (1994) incorporated tree structure, simplifying assumptions were made which may have strongly affected the results. In particular, they used a phylogenetic reconstruction method (UPGMA) which performs poorly in the face of branch length variation, and also assumed accurate reconstruction of ancestral sequences using parsimony techniques, which are known to reconstruct ancestral nodes poorly (Yang *et al*., 1995b). Chelvanayagam *et al*. (1997) used a novel weighting function to compensate for phylogenetic structure, but the statistical effects of this weighting are largely unknown.

The importance of both the protein structure prediction problem and the more general problem of understanding the dynamics of the evolution of protein sequences, and consideration of the limited success which early analyses have had, make it imperative that analytical methods for detecting coevolution should be improved and that these methods should define precisely what it is that they are measuring. Here, we develop a simple maximum likelihood methodology for residue coevolutionary analysis, and show that it has the ability to discriminate coevolution from apparent correlations which are in fact no more than random effects. The method assumes a constant coevolutionary relationship between sites, and may therefore be limited to use with moderately closely related protein families, but rather than being a theoretical limitation, this is more a comment on the possible true nature of coevolution in protein sequences. The method is designed to measure the underlying evolutionary relationship between two sites in a protein, and the degree to which coevolution between the sites explains the data better than a model of independent evolution would. The basis for this method was first developed with reference to correlations in morphological features and RNA sequence (Pagel, 1994; Schoniger & von Haeseler, 1994; Rzhetsky, 1995), and the relationship to these applications is also discussed. We also develop a precise quantification of the distribution of our test statistic. Since we would like to know the relationship between structural distance and significant coevolution, adequate estimation of random effects is extremely desirable. The method is tested with specific examples where site pairs are categorized by linear distance along the sequence and degree of exposure to solvent, and residues at sites are grouped according to charge and size. The methods developed are quite general, however, and these are only some examples of how it could be used. The method is tested on vertebrate myoglobin sequences, revealing significant excesses of coevolving sites for proximal residues and showing a striking signal arising from α-helical regions near the surface.

Theory

Maximum likelihood methodology

A number of authors have introduced likelihood methods which explain correlation between sites in the specific case of the evolution of sites involved in RNA structure (Schoniger & von Haeseler, 1994; Rzhetsky, 1995; Muse, 1995). The RNA methods benefit from the precise knowledge of RNA secondary structure, and (relative to protein structure) the strict and concise rules which govern its formation. In particular, a discrete contiguous segment can be hypothesized to form a ''stem'' structure in which the pairwise relationships between the sites are pre-defined, and the likelihood of correlated relationships between all pairs in this stem can be tested simultaneously. In the case of protein evolution, the prediction of structure is considerably less precise, and the (mostly unknown) rules which govern its formation are apparently extremely complicated.

Proteins are built up from a pool of twenty amino acid residues. These amino acid residues can occur in any combination among the sequences at each site, permitting a complexity of coevolutionary relationships between sites which defies analysis within the limits of current sequence availability and computational power. Within a phylogenetic context of small divergence levels between

sequences, however, most sites in a protein generally exist in a limited number of residue states. It is also likely that among these states, any coevolution between two sites will have a strong primary component which has the possibility of being detected, while other weaker and less detectable components can initially be ignored. This component might, for example, be related to residue charge or size. In this vein, we reduce the number of states at each site to two (for example, positively and negatively charged residues, or large and small residues), and test a simple model of coevolution between them.

The two-state independent model

We designate the two states A and a, where A might be for example a set of large residues, and a the complementary set of smaller residues. Creating a Markov process model following Felsenstein (1981), for a single site the instantaneous rate matrix governing substitution between the two states (A and a) has both a rate parameter, λ, and an equilibrium frequency parameter π_A, such that the instantaneous rate of substituting state j for state i ($i \neq j$) is equal to $\lambda\pi_j$ (where $\pi_A + \pi_a = 1$). The matrix of transition probabilities at time t is then:

$$\mathbf{P}_{ij}(t) = \begin{cases} \exp[-\lambda t] + \pi_i(1 - \exp[-\lambda t]) & \text{if } i = j \\ \pi_j(1 - \exp[-\lambda t]) & \text{if } i \neq j. \end{cases} \quad (1)$$

The substitution process in this model (as for the subsequent correlated process) is reversible since $\pi_i\mathbf{P}_{ij}(t) = \mathbf{P}_{ji}(t)\pi_j$, allowing considerable computational efficiency in later calculations (Felsenstein, 1981).

The pairwise dependent model

A generalized Markov process model of correlated change in pairs of two-state variables was introduced by Pagel (1994) for comparative analysis of discrete characters in a phylogenetic context. Here, we use a special case of his model which is reversible and requires only one more parameter than when the sites are independent. If the second site in the pair has two states designated B and b, with equilibrium frequencies π_B and π_b (where $\pi_B + \pi_b = 1$), then the matrix of instantaneous transition rates for the paired sites is:

$$\begin{array}{c} \\ AB \\ Ab \\ aB \\ ab \end{array} \begin{array}{c} \begin{array}{cccc} AB & Ab & aB & ab \end{array} \\ \begin{pmatrix} -\sum_{AB} & \lambda_B\pi_{Ab}/\pi_A & \lambda_A\pi_{aB}/\pi_B & 0 \\ \lambda_B\pi_{AB}/\pi_A & -\sum_{Ab} & 0 & \lambda_A\pi_{ab}/\pi_b \\ \lambda_A\pi_{AB}/\pi_B & 0 & -\sum_{aB} & \lambda_B\pi_{ab}/\pi_a \\ 0 & \lambda_A\pi_{Ab}/\pi_b & \lambda_B\pi_{aB}/\pi_a & \sum_{ab} \end{pmatrix} \end{array}$$

$$\mathbf{Q} = \quad (2)$$

where Σ_{ij} is the sum of off-diagonal elements for row (residue combination) ij, and λ_A and λ_B are the rate parameters governing substitution at the two loci, A and B. In this model, there is only one more degree of freedom than the total in the two independent models for each site. (There are five free parameters in the dependent model: two rate parameters, and three independent π_{ij}; the π_{ij} must sum to 1, and the π_i and π_j are completely constrained by the π_{ij}.) This extra degree of freedom can be represented by the residue disequilibrium value, $RD = \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}$, which is analogous to the standard linkage disequilibrium measure. Thus, if the equilibrium linked residue frequencies at the two sites (e.g. π_{AB}) are not equal to the product of the equilibrium residue frequencies at each site (e.g. $\pi_A\pi_B$), there is some degree of dependence between the two sites. The RD measure can be either negative or positive, and if the states are assigned according to some physico-chemical vector of characteristics for each amino acid, the sign corresponds to either compensation or anti-compensation of the residues. For example, if there is a greater probability of negative charge at one site when there is positive charge at the other, this will lead to compensatory evolution between the sites and negative RD values.

The substitution probabilities for the coevolving model can be calculated using $\mathbf{P}(t) = \exp[\mathbf{Q}t]$. See Pagel (1994) and Muse (1995) for complete descriptions of such calculations in evolutionary processes. Note that although the instantaneous rates of double transitions are zero, the probabilities of double substitutions over any time period greater then zero will be positive. Eigenvectors, inverse eigenvectors, and eigenvalues necessary for determining $\exp[\mathbf{Q}t]$ were calculated using standard numerical procedures.

Tree topology and model testing

Rather than use the preceding evolutionary model to construct a phylogenetic tree (which would in principle be possible), if we are given a phylogenetic tree we can use it to test the evolutionary model based on likelihood calculations. We use the methodology by Felsenstein (1981) which shows how to use matrices of transition probabilities ($\mathbf{P}(t)$ above) to calculate efficiently (*via* a ''pruning algorithm'') the likelihood of a model, its

relevant parameters, and a phylogenetic tree consisting of branching order and branch lengths, given the data (for more details, see Felsenstein, 1981; Pagel, 1994; Muse, 1995). In this study, the phylogenetic tree was constructed prior to correlation analysis, and only one tree was used. It is possible that if this tree is incorrect, there will be some effect on the analysis. The results reported by Yang *et al*. (1995a) suggest, however, that as long as the tree is approximately correct, the effects on parameter estimation can be minimal.

The likelihood ratio test statistic

The factor involved in assessing whether two sites are coevolving is to decide whether the data are significantly more likely under the dependent model than under the independent model. An appropriate way to do this is with the likelihood ratio test statistic, defined in the standard fashion as $LR = -2\ \ln(L_I/L_D)$, where L_I and L_D are the maximum likelihood values for the independent and dependent models, respectively. The larger this ratio is, the greater the statistical support is for the dependent model compared to the independent model. In our situation, the *LR* statistic is expected to have a distribution which is asymptotically chi-squared with one degree of freedom, which would be an extremely convenient assumption. Monte Carlo simulations showed, however, that the actual distribution varied with tree structure, rate of substitution, and equilibrium state frequencies (results not shown). Therefore, it is necessary to do considerable calculations to estimate the distribution of this statistic. Tests of the significance of coevolution at a site were made using the Monte Carlo parametric bootstrapping technique developed by Goldman (1993) from the method described by Cox (1962), which is applicable to both nested and non-nested models.

Briefly, the required distributions were estimated by repeated simulation of data under the null (independent) hypothesis and analysis of these data under the independent and dependent models. For estimating *LR* distributions of real proteins, frequency and rate parameters for the independent model must be calculated. In this study, parameters used were the maximum likelihood estimators (MLEs) for each site in the protein. For each replication and each site, data were simulated as follows: the state of an arbitrary node on the tree was randomly assigned from the equilibrium distribution of states (based on the MLEs for the site), and then the states of the other nodes were assigned progressively by moving outward from the initial node, making assignments based on the probabilities of change ($\mathbf{P}(t)$) calculated as described above for the independent model, until all nodes had been assigned. Branch lengths (t) were taken from the reconstructed phylogenetic tree. This was repeated until all sites in the protein had been independently simulated for each replication. For each replicate, likelihood maxima under the dependent model were found for each possible pair in order to get a parametric bootstrap estimate of the distribution of the likelihood ratio values. For real proteins, all variable sites in the protein were re-simulated, and all pairwise comparisons in the regenerated protein were considered in generating the *LR* distribution. This gave 5003 comparisons for the size metric simulation, and 2259 comparisons for the charge metric (note that different sites can be invariable with respect to different physico-chemical characteristics). The *LR* scores are necessarily positive, since the independent model is a special case of the dependent model (the models are nested and hence $L_D \geqslant L_I$), but the scores were sometimes divided into two sets based on the sign of *RD*, defined above. The errors in the cumulative distributions at all points can be estimated by using the variance for a binomial estimate (Rice, 1995), assuming independence of comparisons, and they are negligibly small for these numbers of comparisons.

In order to understand the behavior of a statistic under optimal conditions, it is desirable to create a situation where the parameters (specifically the site-specific rates and equilibrium frequencies and the tree structure) can be precisely controlled. When this was done (see below), frequency and rate parameters and the tree branch lengths and structure were determined at the outset and were identical for each site for a given set of conditions. Rate parameters tested were chosen to cover the range of values observed in a real protein (i.e. myoglobin), and 1000 independent pairs were generated for each parameter set.

Simulated evolution of coevolving sites

It is also useful to assess the power of the method by controlling the degree of coevolutionary relationships between sites. To this end, we simulated sequence evolution on phylogenetic trees according to a coevolving two-state model (Figure 1). Coevolving residue pairs were created by first designating ''driver'' sites, which vary randomly and independently according to a defined set of parameters for the Markov process model in equation (1), as described by Pollock & Taylor (1997). Each driver site was then associated with a paired ''dependent'' site, which varied according to one of two different models depending on the state at the driver site. We used this model to generate 1000 coevolving pairs for each set of conditions, which were then evaluated by the detection method described above. The degree of coevolution generated can be controlled easily by adjusting the rates and equilibrium frequencies at the dependent site. Thus, for example, in order to allow the sites to appear completely correlated at equilibrium, the equilibrium frequencies at the dependent site were 1.0 when the driver site was in one state, or 0.0 when the driver site was in the alternative state. The rates of exchange for both models at the dependent site were varied simul-

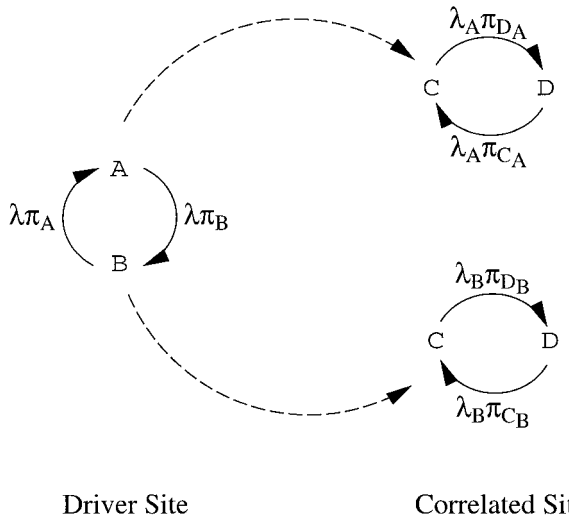

Figure 1. Mutation parameters at coevolving site pairs. A coevolving pair of sites consists of a driver and a dependent site. Substitution at the driver site is controlled by two free parameters, λ and π_A. The instantaneous probability of substituting residue A for residue B at a site is $\lambda\pi_A\,dt$, while the instantaneous probability of substituting B for A is $\lambda\pi_B\,dt$, where $\pi_B = 1 - \pi_A$. In this model, π_A and π_B are the equilibrium frequencies of residues A and B. At the dependent site, the instantaneous probability of substituting residue C for residue D is dependent on the state at the driver site. In the presence of A at the driver site, the controlling parameters are λ_A and π_{D_A}, while in the presence of B they are λ_B and π_{D_B}.

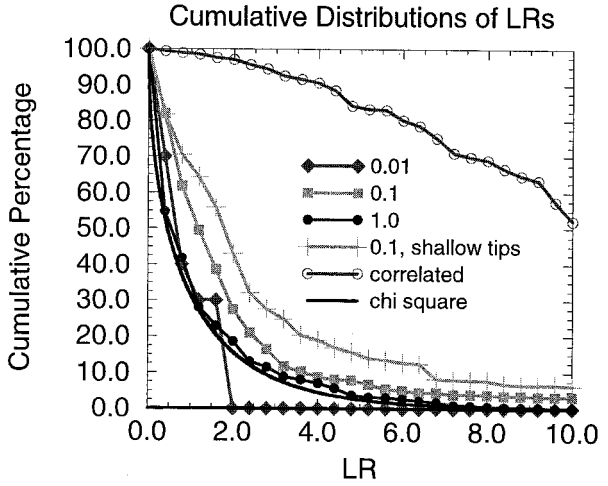

Figure 2. Testing significance and power. Cumulative distributions of likelihood ratios for simulated evolution. Independent sites were simulated along a balanced evenly branching tree (all branch lengths were equal) with probabilities of substitution of 0.01 (diamonds), 0.1 (squares), and 1.0 (filled circles) per branch. These per branch rates are equivalent to 0.3, 3.0 and 30.0 substitutions per site over the entire tree, thus covering most of the range of rates observed in the myoglobin example below. Independent sites were also simulated along a balanced tree with a substitution rate of 0.1 along internal branches and 0.011 along the terminal branches (cross hatches), thus summing to 2.18 probable substitutions per tree. Coevolving sites were simulated along a evenly balanced branching tree with probability of substitution per branch of 0.1 at the driver site and 0.1 at the dependent site, or 3.0 substitutions per tree (open circles; see the text for definitions of driver and dependent sites). Also shown for comparative purposes is a chi-squared distribution with one degree of freedom (thick line, no symbol).

taneously. Coevolution is most detectable when these rates are high (Pollock & Taylor, 1997).

Results

Background distribution and detection of simulated coevolution

When sites were randomly evolved on a 16-taxon, evenly branching, balanced tree, it was clear that the distribution of *LR*s for pairs of such sites did not always closely match the chi-squared distribution with one degree of freedom, which is included for comparative purposes (Figure 2). The distributions in this Figure are presented as (reverse) cumulative distributions, so curves represent the percentages of sites with the specified or greater *LR* value. This enables one to easily read the *LR* cutoff point where the remaining probability falls below a particular critical value (α). When the number of substitutions per branch was moderately low (0.1), the cumulative percentage in the simulations was up to 50% greater than for the theoretical asymptotic chi-squared distribution. For larger substitution rates per branch (1.0), the simulated cumulative distribution matched the chi-squared distribution much more closely, but for the smallest rate (0.01) the cumulative distribution of *LR*s was erratic relative to the chi-square. This second effect occurred not because of limited sampling, but because for smaller rates it is unlikely that more than one substitution will occur over the entire tree, and therefore the number of likely sequence patterns is limited. Large differences in the given tree structure were seen to have large effects on the cumulative distributions of likelihood ratios for independently evolving sites. For simulations on a tree with the same structure as before, and with internal branch lengths of 0.1, but with terminal tips one-ninth of the length of the deeper branches, the cumulative distribution departed even further from the chi-squared distribution. The effect with this tree was particularly noticeable for large likelihood ratios (e.g. $>\chi^2_{1,95\%} = 3.84$), where the observed cumulative percentage can be two to three times that for the evenly branching tree. Thus, use of a chi-squared distribution in significance testing as an approximate substitute for the real *LR* distribution is not generally adequate. Note that here we are addressing the effect of large and consistent differences in relative branch lengths of the true or given tree,

rather than potential errors in tree topology and branch length reconstruction, as discussed earlier.

Simulations showed, however, that the power of tests using the likelihood ratio statistic can be quite good. Figure 2 shows the cumulative distribution of sites which had coevolved according to the model described earlier, with $\lambda = \lambda_A = \lambda_B = 1.0$, and most pairs have significant *LR*s. On the curve for independently evolving sites with a rate of 1.0, a significance level (α) of 0.05 is achieved at *LR* approximately 4.6. The power (fraction of coevolving sites above this level, β) for these conditions can be read from the uppermost curve of Figure 2, and is 0.86. For $\alpha = 0.025$, β is 0.79; for $\alpha = 0.01$, β is 0.71; and even for $\alpha = 0.001$, β is still 0.52. Previous studies have shown that the observed correlation levels at coevolving sites are much more sensitive to λ_A and λ_B than λ (Pollock & Taylor, 1997), and this is also the case here: increasing λ_A and λ_B leads to perfect correlation in most cases, while decreasing λ_A and λ_B leads to reduced power (data not shown). It must be remembered, however, that these significance and power levels apply to tests of individual pairs of sites, while in real applications the number of comparisons performed can easily number in the thousands, so the effect of multiple comparisons has to be considered.

Coevolution and structure in myoglobin

Site pairs in vertebrate myoglobin were divided into those pairs which were separated by either five or fewer residues in the primary sequence, or by six or more residues, and likelihood ratios were calculated for each pair. The division was made because a question of interest is whether significant coevolution between sites can be linked to distance in the three-dimensional structure for prediction purposes, and for this it is not helpful to include sites adjacent in the primary sequence. As sites close in the primary sequence are likely to be close in the three-dimensional structure, and are therefore likely to have coevolved in a way leading to increased *LR*s, it is interesting to analyze this predictable effect separately. Residues at each site were partitioned into two states according to either a size or a charge metric, where the partition point was determined by the mean value of the metric at that site. These sites were further divided into positive and negatively coevolving sites according to the sign of the coevolution measure, *RD*. Pairs separated by six or more residues in which both residues were buried or exposed were also analyzed separately.

The MLE rate and frequency parameters were estimated for each site when segregated by both size and charge (Figure 3), and these MLE values were used to simulate independent evolution at each differentiable site (104 sites for size, 70 sites for charge). For those site pairs categorized by size and separated by six or more residues (5003 comparisons), the (reverse) cumulative distributions of

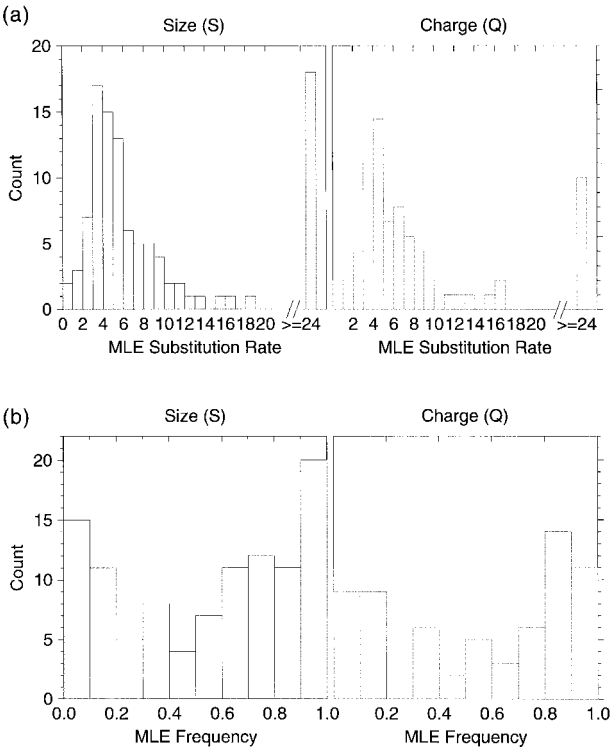

Figure 3. Maximum likelihood estimators. The distribution of the maximum likelihood estimators of the (a) rate and (b) equilibrium frequency parameters for myoglobin sites segregated by size (S) and charge (Q). There were 104 sites which could be differentiated by size and 70 sites which could be differentiated by charge. Sites invariable with respect to each physicochemical metric were not included. Each of these estimators are used for simulating independent evolution of sites in myoglobin in order to estimate the background distribution of *LR*s for the 5356 size-segregated and 2415 charge-segregated pairwise comparisons evaluated in this study. For size-segregated frequencies, the mean was 0.54(±0.03) standard errors, while for charge-segregated frequencies it was 0.54(±0.04). For size-segregated rates, the mean was 21.2(±5.0), with a skew of 4.61(±0.24). Rates are probabilities of substitution per site over the entire tree. For charge segregated rates, the mean was 14.6(±3.8), with a skew of 4.97(±0.29).

*LR*s from the simulation match the observed distributions more closely than does the chi-squared distribution with one degree of freedom (Figure 4). For sites separated by six or more residues and categorized by charge (2259 comparisons), the simulations again matched the true curves more closely than the chi-squared value for large *LR* values (Figure 5). Thus, reliance on the chi-squared distribution for significance levels would lead to large overestimates of the excess number of coevolving sites for all standard significance levels (0.05 and below). This confirms that for protein data, in a situation where a real tree structure and the effect of correlation between multiple site pair comparisons is explicitly accounted for, the chi-squared approximation is unreliable. Therefore,

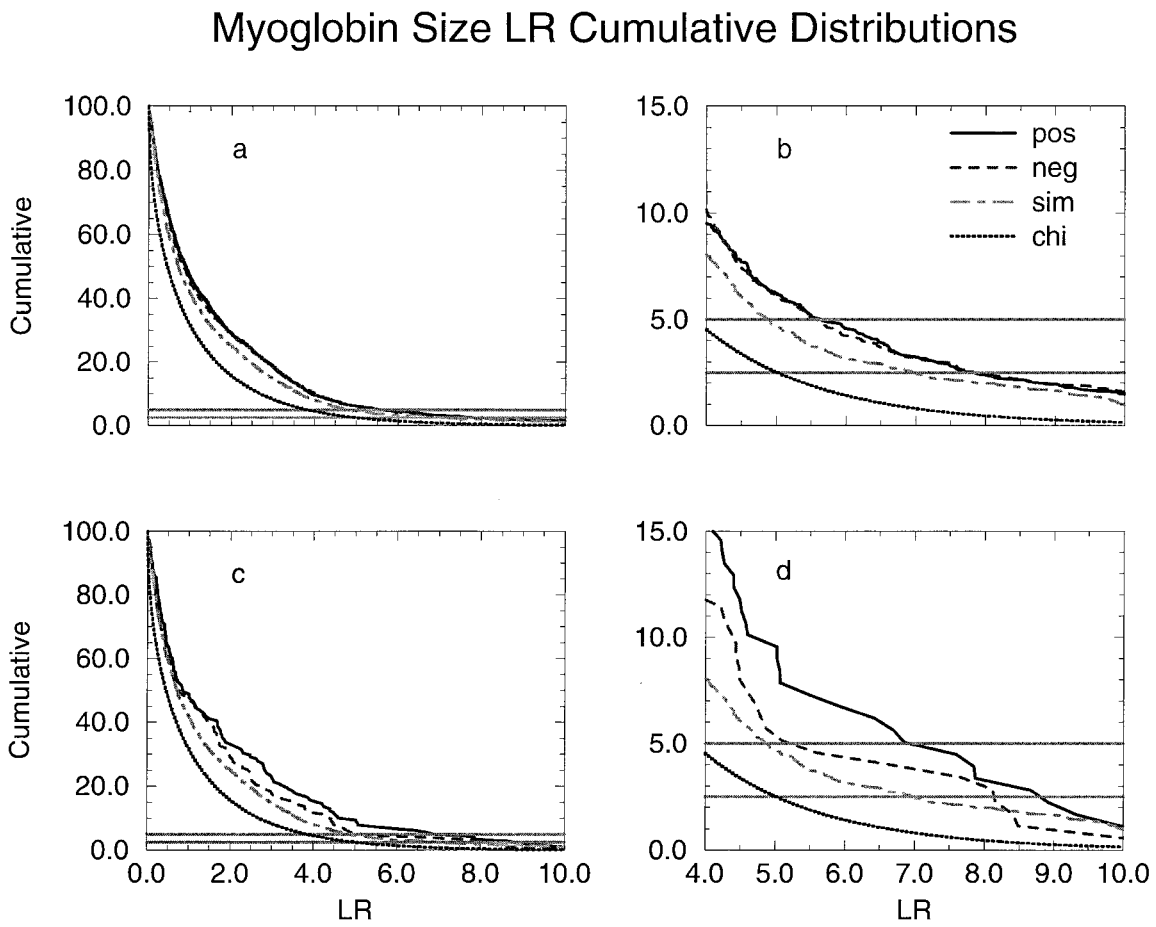

Figure 4. Cumulative distributions of positive and negative coevolution: size. The reverse cumulative probability distribution of *LR* values is shown for sites with positive (continuous line) and negative (broken line) *RD* values when segregated according to size, along with the simulated expectation (dot-dashed line) and the chi-squared distribution with one degree of freedom (dotted line). Solid horizontal lines correspond to the 5% and 10% cutoff levels and the right graphs are simply enlargements of the left graphs to better visualize the region where these lines intersect the data curves. (a), (b) The distribution for those sites separated by six or more residues; (c), (d) sites separated by five or fewer residues.

critical values for significance levels were henceforth determined from the simulations for the appropriate physico-chemical metric. If the *LR* values for the 5356 total size and 2415 total charge comparisons on which the simulation distributions are based were independent, the error in these cumulative distributions at the 0.05 cutoff would be 0.0030 for size, and 0.0044 for charge. Since there is some correlation between these comparisons, the error could be somewhat larger, but repeated simulations did not appear qualitatively different (data not shown).

Table 1 shows the observed numbers of positively and negatively coevolving site pairs, and the numbers expected beyond critical *LR* values determined as outlined above. The significance of the bias (positive to negative ratio different than 1.0) and the deviation from the numbers expected at the 0.05 and 0.01 significance levels were evaluated with the chi-squared test for goodness of fit. The bias in positive *versus* negative *RD* values for sites separated by six or more residues was found to be non-significant for the charge grouping, but extremely significant for the overall size grouping, and significant for both the buried and exposed size groupings. Assuming that the simulated cumulative distributions are an accurate reflection of the true distributions, there are extremely significant excess numbers of large *LR*s for pairs separated by six or more residues for both the size and charge metrics (Table 1). For example, with the charge metric we expect 5%, or 113, of the 2259 pairs tested to exceed the 0.05 significance cutoff of 4.813, but in fact we observe 158 such pairs. For the size metric, we expect 250 pairs over the 5% cutoff (4.880), but we observe 319. Excess numbers are also very or extremely significant for the 0.01 significance level. The pairs with LR values above the 0.05 and 0.01 cutoff were not significantly biased towards either positive or negative coevolution. These excess numbers of pairs presumably occur due to sites which have been coevolving, but such sites cannot be separated from those sites which have likelihood ratios beyond the significance threshold due to chance, since there were so many comparisons made. For example, with the

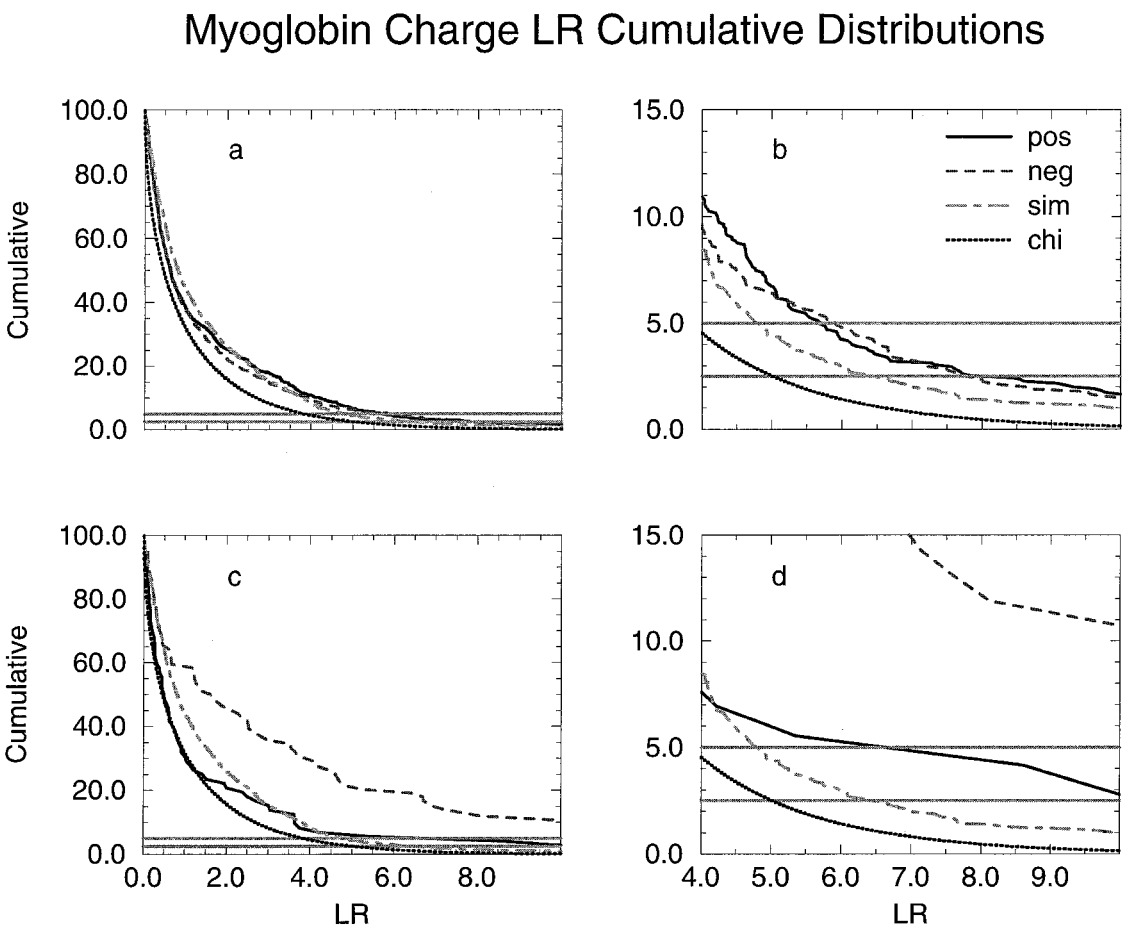

Figure 5. Cumulative distributions of positive and negative coevolution: charge. The reverse cumulative probability distribution of *LR* values is shown for sites showing positive and negative *RD* values when segregated according to charge. The lines and labels are as defined in the legend to Figure 4. (a), (b) The distribution for those sites separated by six or more residues; (c), (d) sites separated by five or fewer residues.

charge metric, there are an extra 45 pairs above the 0.05 cutoff, but the total of 158 pairs includes 113 pairs which are probably above the cutoff due to random effects.

The results from the sites separated by five or fewer residues in the primary sequence are more striking. While the excess numbers of size-segregated *LR* values are moderate for this group, the excess numbers of negatively coevolved charge-segregated sites with large *LR* values are three to ten times what is expected by chance (see Table 1 for excess numbers of sites with *LR* values above the 0.05 significance level). This indicates that negative, or compensatory, coevolution is playing an important role for these site pairs, and that this method is sensitive to detecting such correlation when it exists and is strong. For both positive and negative coevolution, the size metric produces only a slight excess of pairs with large *LR* values. This indicates that while size interactions may play some role in the evolution of proteins, they are neither necessarily strong nor consistent in a pairwise fashion on the time scale of the myoglobin tree, even for closely linked pairs.

Distance distributions of significant pairs

Since there are excess numbers of site pairs with likelihood ratio values above the 0.01 and 0.05 cutoffs for both the size and charge segregation procedures, it is of interest to see whether a close C^{α} distance between these pairs is associated with these excesses. This evaluation is possible here because the three-dimensional structure of myoglobin is known, but has implications for when the method is applied to proteins of unknown structure. For both the size and charge groupings, there is an apparent excess of negatively coevolved pairs which are close in the three-dimensional structure, although positively coevolved close pairs are in apparent excess only for the size grouping (Figure 6). The excess numbers below 20 Å are significant in these three cases (data not shown). When the buried and exposed groupings were analyzed separately (Figure 7), the number of paired sites closer than 20 Å was significant only for the exposed pairs segregated according to size. These last comparisons generally suffer from small numbers, however, and were not further divided into

Table 1. Counts and chi-squared test values for myoglobin structural categories

Category	*LR* α	Observed (±)	Bias $\chi^2_{[1]}$	Excess $\chi^2_{[1]}$
Q6+	All	2259(1149/1110)	0.673	n.a.
(charge vector, separated by six or more residues)	0.05	158(85/73)	0.911	18.91*** (14.78***/4.96*)
	0.01	36(19/17)	0.111	8.041** (5.28*/2.90)
Q6 + bur 0.2	All	253(114/139)	2.470	n.a.
(buried residues, score > 0.2)	0.05	17(6/11)	1.471	1.575
Q6 + exp 0.2	All	962(505/457)	2.395	n.a.
(exposed residues, score < 0.2)	0.05	74(42/32)	1.351	25.45***
Q1-5	All	158(72/84)	0.923	n.a.
(separated by < =five residues)	0.05	22(4/18)	8.91**	26.49***
S6+	All	5003(2632/2371)	13.62***	n.a.
(size vector, separated by six or more residues)	0.05	319(167/152)	0.705	19.95***
	0.01	76(39/37)	0.053	13.62*** (7.86*/5.77*)
S6+ bur 0.2	All	918(496/422)	5.965*	n.a.
(buried residues, score > 0.2)	0.05	58(32/26)	0.621	3.358
	0.01	18(6/12)	2.000	8.560***
S6 + exp 0.2	All	1590(841/749)	5.323*	n.a.
(exposed residues, score < 0.2)	0.05	95(48/47)	0.011	3.181
	0.01	14(9/5)	1.143	0.253
S1 − 5	All	353(178/175)	0.025	n.a.
(separated by <=five residues)	0.05	26(17/9)	2.46	4.158*

Chi-squared tests for positive/negative bias and excess above random expectation for observed and expected counts of positively and negatively correlated pairs when divided by partitioning vector, significance of likelihood ratio, and by exposure to solvent. Significance levels are indicated by: * < 0.05, ** < 0.01, *** < 0.001, n.a. not applicable. Other values are non-significant.

positively and negatively coevolved subcategories for that reason. Nevertheless, the trend in each grouping is towards more close pairs than expected. There are many more exposed charge pairs than expected beyond the 0.05 cutoff, and the unusual distribution in Figure 7(d) indicates that in this case there are excesses of both close and distant sites. The close sites are presumably due to direct charge interactions, but the distant sites are more difficult to explain: possibilities include maintenance of an isoelectric composition (or a dipole moment) across the molecule, avoidance of self-complementarity (and thus polymerization) at opposite ends of the molecule, or avoidance of compatibility of distant secondary structural segments which would interfere with the folding process.

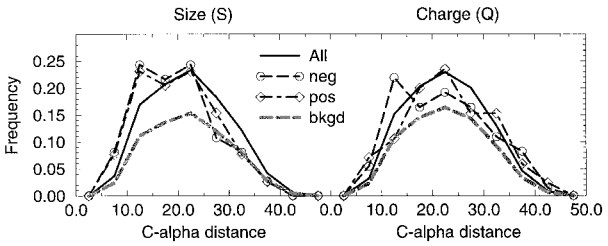

Figure 6. Distance frequencies for myoglobin site pairs. Frequency distributions for the C^{α} distances between pairs of sites in myoglobin are shown for distance bins of 5.0 Å. Distributions for all size and charge segregated pairs are shown along with the respective distributions of negatively (neg) and positively (pos) coevolved pairs with *LR* values greater than either the 0.01 cutoff (size), or the 0.05 cutoff (charge). Also shown are the predicted background frequency distribution of sites with *LR*s greater than the appropriate significance value (bkgd). See Table 2 for numbers in distributions.

Closely linked pairs and secondary structure

For residues separated by one to five residues in the primary sequence, the appropriate null hypothesis is that there will be no association between the degree of separation and the amount of coevolu-

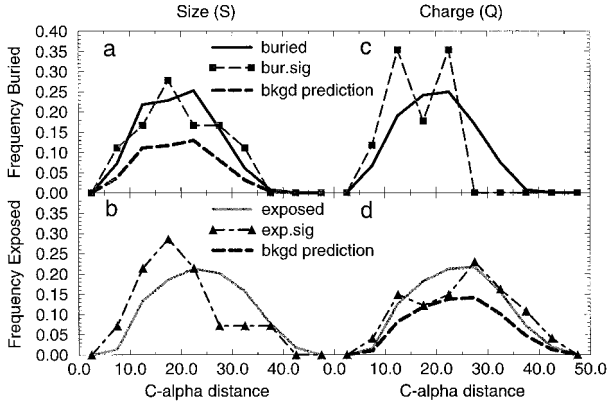

Figure 7. Distance frequencies for exposed and buried site pairs. Frequency distributions for the C^{α} distances between pairs of sites in myoglobin which are (a), (c) buried or (b), (d) exposed are shown for distance bins of 5.0 Å. Distributions for all buried and exposed (a), (b) size and (c), (d) charge segregated pairs are shown along with the respective distributions of buried and exposed pairs (bur.sig and exp.sig) with *LR* values greater than either the (a), (b) 0.01 cutoff (size), or the (c), (d) 0.05 cutoff (charge). Predicted background distributions (bkgd prediction) and numbers are as in Table 2.

Table 2. Counts of charge (Q) segregated sites by separation along sequence

	Separation				
	1	2	3	4	5
Neg Q, $LR > 4.0$	2	2	8	8	4
Total Q	31	27	35	36	27

tion between sites. Thus, if the degree of separation (and therefore secondary structure) has no effect, residue pairs with *LR*s greater than any particular probability cutoff should be evenly distributed among the different separations. Charge-segregated pairs with negative *RD* values and *LR*s greater than 4.0 (which are in extremely significant excess) have many more separation values of three and four, however, than values of one, two or five (Table 2). When testing the general hypothesis that there is any (unknown) effect of degree of separation, this observation is not significant at the 0.05 level (for a 2×5 G test of homogeneity, i.e. a generalized likelihood ratio test of the goodness of fit of a model for multinomial cell probabilities (Rice, 1995) with four degrees of freedom, $G_4 = 6.44$). The separation difference peak at three and four is strongly coincidental with the periodicity (3.6) of an α-helix, however, and in this all-helical protein the hypothesis that coevolutionary interactions should be more common between residues separated by three or four residues is a structurally reasonable hypothesis. A 2×2 G test of homogeneity between separations of one, two or five, and separations of three or four, is significant at the 0.05 level ($G_1 = 5.14$). Note that this effect is entirely different from the α-helix hydrophobicity periodicity which is the basis of many current structure prediction methods. The α-helix periodicity effect causes a slight bias towards separation distances of three or four in the totals before any application of the *LR* test. This bias is not significant in a 2×2 G test of homogeneity ($G_1 = 1.95$), and regardless is fully accounted for in the homogeneity tests. An interesting visual observation is that these coevolving pairs appear more often on the ends of α-helices, and near the surface of the protein (as would be expected for charged residues), although this was not statistically quantified.

Discussion

The likelihood methods developed here have good statistical power to detect coevolved residue substitution in pairwise comparisons in simulations, and both positive and negative coevolution are detected in myoglobin sequences. The excess number of site pairs undergoing coevolution must be estimated by simulation. In the myoglobin example the strongly coevolving sites tend to be closer in the three-dimensional structure, and there are statistical excesses of large *LR* values. Sites which are adjacent in the linear sequence have a slight bias towards positive size coevolution. There is also a large excess of negatively coevolving sites with large *LR* values, and the linear separation of the excess sites corresponds to the periodicity of an α-helix. Preliminary analysis of vertebrate lactate dehydrogenase sequences suggests that the strongest coevolutionary signal among sites separated by less than five residues is also from the α-helices (data not shown), indicating that this may be a general phenomenon. In addition to their general interest for understanding the processes of protein evolution and generating hypotheses of structural interactions between residues, these observations might prove useful in both secondary and tertiary structure prediction. For example, Pazos *et al*. (1997) have had some success predicting domain and dimer contacts using a simple methodology with low power for detecting coevolution (Pollock & Taylor, 1997), and thus the promise of enhanced signal detection of this kind is very encouraging. Including phylogenetic tree information appropriately has been helpful in other multiple sequence analysis/protein structure contexts (Goldman *et al*., 1996), and this approach appears to be beneficial here, too. Also, buried α-helices are generally hard to detect by other means (Benner, 1996), and so a strong signal which is often at the ends of such helices would be extremely helpful. In a comparison of all sites in a single protein, the signal is not strong enough to definitively overwhelm the background, so that (based on the parametric estimates of the background) even for the example of negative charge coevolution at separation distances of three or four an individual pair has only about a 75% probability of being due to a coevolutionary relationship. The problem is worse for the more distant comparisons, since the number of comparisons, and therefore the number of false positives due to background noise, goes up with the square of the number of sites included. With the methods developed here, the strength of the signal is directly estimated, and since the coevolutionary signal is independent of the signals used by most current methods, it can potentially be used to augment methodologies which can make use of a noisy signal; for example, distance geometry (Aszodi *et al*., 1997), or sequence/structure threading (Jones *et al*., 1992; Taylor, 1997).

The tests performed here were based on pre-conceived notions of the charge and size relationship of protein residues. There is no guarantee, however, that these physico-chemical properties are those which drive coevolution between residues in the protein structure. This methodology can evaluate coevolutionary relationships based on any partition of the residues into two groups, and thus there is a great deal of flexibility to search for partition types which are particularly indicative of proximity in the three-dimensional structure. Results here and in previous studies indicate that there is a real signal to be detected, and that it can prove useful in protein structure prediction. The greatest problem is separating this signal from

noise inherent in such a huge multiple comparisons analysis. The method described here uses evolutionary information to help in this separation, and quantifies the result. Future study will be directed towards determining the partitions which carry the greatest signal to noise ratio. These will also be useful in giving structural meaning to the coevolutionary process.

Unlike previous methods (Pollock & Taylor, 1997), the method developed and analyzed here will not become biased by, and indeed is able make statistical use of, similar sequences. It allows for tree structure, including variation in branch length, in a statistically robust fashion, and does not assume precise reconstruction of ancestral nodes. A major assumption of this kind of analysis (and any other coevolutionary/correlation analysis), that coevolutionary relationships are static with time, is more likely to be upheld in trees without extensive pairwise divergence levels. This is important, since such recent trees must include many sequences in order to provide the necessary information for coevolutionary analysis, and may include many closely related taxa. More deeply branching trees may still remain useful for detecting correlation in the relatively conserved protein core, and may be necessary in order to get enough evolutionary change to detect a signal at the slow-evolving core sites. Such questions of evolutionary sequence change *versus* evolutionary structural change and their effect on coevolution are important topics for future study.

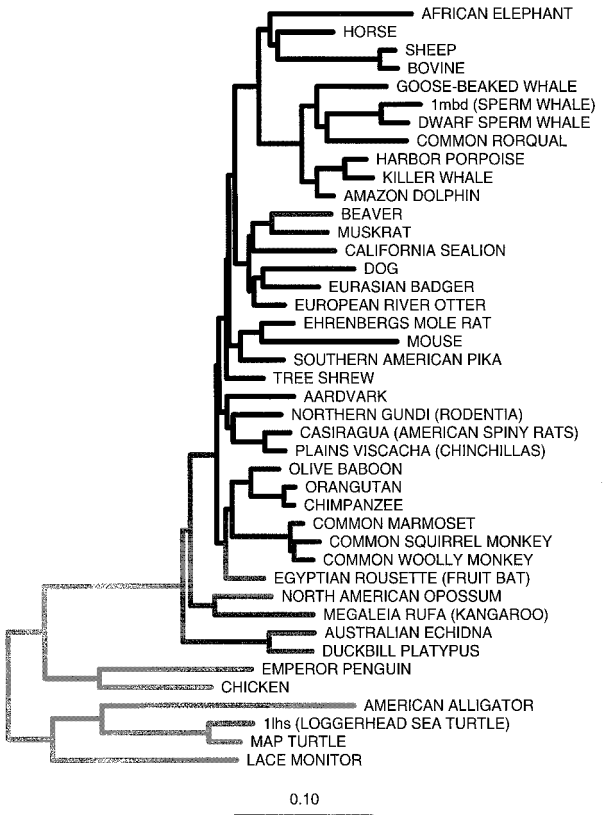

Figure 8. Phylogenetic tree of myoglobin sequences used in this study. The scale bar corresponds to molecular distance, i.e. the average probability of substitution per site for the branch length shown.

Materials and Methods

The real data analysis example of the relationship between coevolution and structural features was carried out on an alignment of tetrapod myoglobin sequences. This family began with 67 sequences collected from the TREMBL databank (Bairoch & Apweiler, 1998), which were aligned using MULTAL (Taylor, 1998) followed by adjustments by eye. Phylogenetic trees were created using the PROTDIST and NEIGHBOR programs from the PHYLIP computer package (Felsenstein, 1993), and the resulting tree was trimmed to remove branches which joined the tree in conflict with known tetrapod phylogeny, and to avoid long unbroken branches whenever possible. The remaining 42 aligned sequences were used in all further analysis, and this alignment is available at http://ib.berkley.edu/labs/slatkin/david/MYOG/. The SwissProt abbreviations for these sequences are: MYG_PAPAN, CASFI, CALJA, PONPY, LAGLA, ORYAF, PANTR, CTEGU, ROUAE, TUPGL, PROGU, SPAEH, LAGMA, ZALCA, LUTLU, MELME, ONDZI, OCHPR, PHOPH, HORSE, INIGE, ZIPCA, KOGSI, ORCOR, BALPH, ELEMA, SAISC, SHEEP, ALLMI, APTFO, CHICK, GRAGE, BOVIN, CANFA, DIDMA, TACAC, MACRU, VARVA, MOUSE, ORNAN, and 1mbd_PDB and 1lhs_PDB. The phylogenetic tree used is shown in Figure 8. Available myoglobin sequences more than 98% similar to the sequences listed were not included in the analysis. The three-dimensional structure used for all distance calculations was from sperm whale (1mbd; resolution = 1.4 Å), and was obtained from the Brookhaven Protein Data Bank (Bernstein *et al*., 1977). Reading of PDB structures and three-dimensional distance calculations was performed using subroutines kindly provided by A. Aszodi. Division of residues into buried and exposed was made using a score of 0.2 in the CONESCORE program from DRAGON (Aszodi *et al*., 1997). Expected background frequency distributions of C^{α} distances (the distances in angstroms between the C^{α} atoms of each residue pair) were calculated by multiplying numbers from the overall background distribution by the number expected beyond a significance cutoff, and dividing by the number actually observed.

Acknowledgments

D.D.P. is a Hitchings-Elion fellow of the Burroughs Wellcome Fund. N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research.

References

Altschuh, D., Lesk, A., Bloomer, A. C. & Klug, A. (1987). Correlation of coordinated amino-acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 643-707.

Aszodi, A., Munro, R. E. & Taylor, W. R. (1997). Distance geometry based comparative modelling. *Fold. Design*, **2**, S3-S6.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Benner, S. A. (1996). Predicting the conformation of proteins from sequence data. In *Protein Engineering: Principles and Practice* (Cleland, J. L. & Craik, C. S., eds), pp. 71-100, Wiley-Liss, NY.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Chelvanayagam, G., Eggenschwiler, A., Kneckt, L., Gonnet, G. H. & Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**, 307-316.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Stat. Soc. ser. B,* **24**, 406-424.

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

Felsenstein, J. (1993). *PHYLIP (phylogeny inference package) version 3.5*, University of Washington, Seattle.

Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309-317.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182-198.

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

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86-89.

Muse, S. V. (1995). Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics,* **139**, 1429-1439.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA,* **91**, 98-102.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. Roy. Soc. London ser. B,* **255**, 37-45.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interactions. *J. Mol. Biol.* **271**, 511-523.

Pollock, D. D. & Taylor, W. R. (1997). Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.* **10**, 647-657.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, 2nd edit., Duxbury Press, Belmont, CA.

Rzhetsky, A. (1995). Estimating substitution rates in ribosomal RNA genes. *Mol. Biol. Evol.* **141**, 771-783.

Schoniger, M. & von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3**, 240-247.

Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349-358.

Taylor, W. R. (1997). Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902-943.

Taylor, W. R. (1998). Dynamic sequence databank searching with templates and multiple alignment. *J. Mol. Biol.* **280**, 375-406.

Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341-348.

Yang, Z., Goldman, N. & Friday, A. (1995a). Maximum likelihood trees from DNA sequences - a peculiar statistical estimation problem. *Syst. Biol.* **44**, 384-399.

Yang, Z., Kumar, S. & Nei, M. (1995b). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics,* **141**, 1641-1650.

Edited by G. Von Heijne

(Received 19 June 1998; received in revised form 25 January 1999; accepted 29 January 1999)