Protein Engineering vol.10 no.6 pp.647–657, 1997

Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution

D.D.Pollock[1] and W.R.Taylor

Department of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

[1]To whom correspondence should be addressed

Various methods for detecting correlation between sites were evaluated by ascertaining their ability to discriminate positively correlated sites from background correlation at randomly evolved sites. A model for generating pairwise correlations of different degrees is also described. An assortment of physicochemical vectors and similarity and difference matrices were used to discriminate correlated change. There was little difference in effectiveness between the different matrices, but there were significant differences between the matrices and the physicochemical vectors. It is shown that all methods investigated exhibit significant inability to screen out background correlation, particularly in the presence of phylogenetic relatedness between the sequences. Methods using the matrices are unable to distinguish positively correlated from negatively correlated, or compensatory, replacements.

Keywords: compensatory change/correlation/evolution/protein structure/residue contacts

Introduction

The structures of proteins are conserved to a greater degree than the sequences which determine their structures, and this leads to the hypothesis that in the course of evolution, residue substitutions which tend to destabilize a particular structure are probably compensated by other substitutions which confer greater stability on that structure. It is reasonable to suppose that sites which can compensate for a destabilizing substitution at another site are likely to be close to that site in the three-dimensional structure of the protein. For example, if a salt bond were important to structure and function, a substitution of the positively charged residue with a neutral residue would need to be compensated by a nearby residue substituting from a neutral to a positive residue. Similarly, a substitution involving a 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. Thus, if structural compensation is a general phenomenon, sites which are close together in the three-dimensional structure will tend to evolve in a correlated fashion owing to the compensation process. This line of reasoning has led to a great deal of recent interest in the development of methods for detecting correlated changes in sequence evolution, primarily as a source of distance information for use in protein structure prediction (Altschuh *et al.*, 1987a,b; Gobel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Taylor and Hatrick, 1994).

Despite the compelling logic of this line of reasoning, these studies have met with only limited success in identifying pairs of sites which are adjacent in the three-dimensional structure. Possible reasons for this failure are that distant sites are as important as adjacent sites in the compensatory process, that the number of sites which can compensate any other are so large that the pairwise correlation signal is too small to detect or that the methods used are insufficient for separating true correlation in the evolutionary process from background noise (false correlation due to random events). If the first two reasons are correct, there is nothing to be done, but if the problem lies in insufficient methods, improvement is possible. The more recent methods developed (Gobel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Taylor and Hatrick, 1994) are quite distinct, and in this paper we discuss the differences between the methods and determine the effectiveness of methods based on these studies in separating correlated signals from background evolutionary noise.

Three of the recent methods (Gobel *et al.*, 1994; Neher, 1994; Taylor and Hatrick, 1994) are similar to each other in that they detect different kinds of pairwise correlations by summing weighted contributions of the correlation evident between all taxon pairs. These methods do not rely on any kind of phylogenetic tree relating the sequences in a study, whereas the method in Shindyalov *et al.* (1994) does. Here we defer analysis of the effectiveness of the Shindyalov *et al.* (1994) method for future study, but we do analyze the effect of different kinds of phylogenetic relatedness among the sequences available on the effectiveness of the other three methods. This is important because phylogenetic relatedness will almost certainly confer some degree of positive correlation on all paired sites, but the magnitude of this spurious correlation is unknown. We develop a common framework for comparing the different methods. The methods are combinations of statistical techniques with either similarity matrices or physicochemical vectors, and so we also evaluate the effect of using different kinds of metrics. Of particular interest is the comparison of similarity matrices with the physicochemical vectors, since the similarity matrices can only detect correlation of the *magnitude* of change, whereas the physicochemical vectors can also detect correlations of the *direction* of change. A model for generating correlated sites is developed and used to infer the effectiveness of the methods in identifying signals with varying degrees of underlying causative correlation.

Methods

Sequence evolution at background and correlated sites

Sequence evolution was simulated according to either a correlated evolution model or a background model. In all cases, starting residues at sites were randomly selected according to the equilibrium residue frequencies of the model at that site. Background sites were created by evolving sequences according to the mutation data matrix (MDM) model (Jones *et al.*, 1994). An MDM is an empirically derived estimate of the average probability over all sites in all proteins of mutation from one amino acid to another over a short time period. For

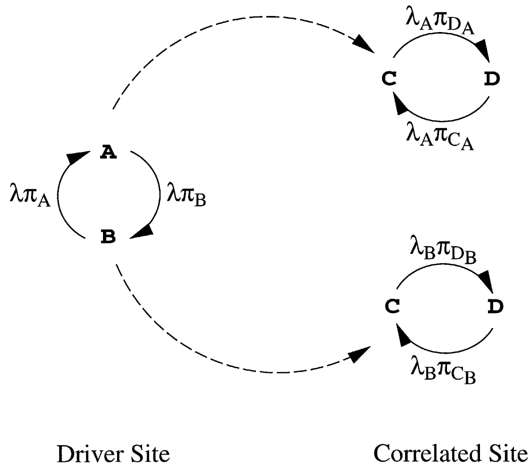

Fig. 1. Mutation parameters at correlated site pairs. A correlated pair of sites consists of a driver and a correlated site. Substitution at the driver site is controlled by two free parameters, λ and π_A. The instantaneous probability of substituting residue A for residue B at a site is $\lambda\pi_A$ dt, while the instantaneous probability of substituting B for A is λ_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 correlated 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_{DA}, and in the presence of B they are λ_B and π_{DB}.

an MDM, **M**, probabilities of change from one residue to another over time, *t*, were calculated from

$$\mathbf{P}(t) = \exp[\mathbf{M}t] = \mathbf{S}\exp[\mathbf{\Lambda t}]\mathbf{S}^{-1} \quad (1)$$

where exp[**Λ**] is the diagonal matrix of the exponentials of **M**'s eigenvalues and **S** is the matrix of corresponding eigenvectors [see Felsenstein (1981), Pagel (1994) and Goldman *et al.* (1996) for derivations]. In this study the DAY78 model was used (Dayhoff *et al.*, 1978), but any other MDM model could be easily substituted.

Correlated residue pairs were created by designating 'driver' sites, which vary randomly and independently according to one set of parameters, and associated dependent sites, which vary randomly according to parameter sets which are dependent on the state at the associated driver site (Figure 1). In this study, dependent and driver sites were limited to two residue states each. The equilibrium frequencies of residues A and B at the driver site are π_A and π_B, where $\pi_A = 1 - \pi_B$, while the rate of mutational exchange between the two residues is controlled by λ. The probability of a residue being substituted by the other residue, *i*, over time, *t*, is π_i [1 − exp(−$\pi_i\lambda t$)]. Thus, a substitution rate of 0.1 at a correlated site corresponds to a mutation rate of 10.0 PAMs at a background site.

The associated dependent sites have two similar exchange probability matrices, one for each state at the driver site. Thus there are different equilibrium frequencies and exchange rates depending on the state of the driver site. In order for the sites to be completely correlated at equilibrium, the equilibrium frequencies of the first residue should be 1.0 and 0.0 depending on the state at the driver site. The rate of exchange at the dependent site should also be high relative to the rate of exchange at the driver site.

Tree structure

Sequences were created by duplicating each sequence at the start of the simulation, and then again after a constant period of time depending on the branch length of the simulation. Prior to the second duplication, sites were changed randomly according to the branch length and the model at that site, as described above. The duplication and random change were repeated *k* times, leading to 2^k sequences at the end of the run. In order to assess the effect of tree structure, *N* sequences were generally sampled from this initial set of 2^k in one of four ways. If the first *N* sequences in the sample were picked, the resultant tree is evenly 'balanced' and all branches are of equal length (Figure 2); this was the predominant sampling method (Sokal, 1990). Alternatively, the terminal branches (those nearest the tips of the tree) were sometimes made longer than the others, and this is referred to as a balanced tree with deep terminal splits. When the first set of internal branches (those nearest the root) were made longer than the others, this is referred to as a balanced tree with shallow terminal splits. Finally, the first even but maximally 'unbalanced' tree was chosen, where the splits are occurring at regular intervals, but only along one branch (Sokal, 1990). It is impossible to separate the effect of mutation rates from those of time, and so branch lengths, *b*, and the overall tree depth, *B*, are measured in terms of λt. In the case of the MDMs, $\lambda t = 1.0$ corresponds to 1.0 PAM.

Correlation analysis

Correlation between sites was tested by one of four different methods. The first method, from Neher (1994), is a simple correlation summed over all *N* sequences in the sample:

$$r_{i,j} = \frac{1}{N} \sum_{l=1}^{N} \frac{(q_i^l - m_i)(q_j^l - m_j)}{s_i s_j} \quad (2)$$

where q_i^l and q_j^l are the values for some amino acid physicochemical vector for sequence *l* at sites *i* and *j*, and m_i and m_j and s_i and s_j are the observed means and standard deviations of that vector at sites *i* and *j*. This is referred to as the BASIC method. The other three methods are all correlations of the change between residues in sequence pairs:

$$r_{i,j} = \frac{1}{M} \sum_{l=1}^{N} \frac{(d_i^l - m_i)(d_j^l - m_j)}{s_i s_j} \quad (3)$$

where the sum is now over all sequence pairs considered, *M*, while d_i^l and d_j^l are the differences between the metric values for the residues in the sequence pair at sites *i* and *j*, respectively. In the case where each residue has a metric value, these changes can have direction (sign) in addition to magnitude, whereas if an exchange matrix (e.g. an MDM) is used, the change has only magnitude. The mean and standard deviation in these correlation methods are calculated for differences between all sequence pairs considered, rather than for metric values of each residue in each sequence.

The primary difference between these last three methods derives from which sequence pairs are considered. The second method we test is the method of Gobel *et al.* (1994), where the sum is over all $M = N^2$ sequence pairs, but site pairs where one or the other site is invariant are excluded. This will be referred to as the GOB94 method. In the case of one-dimensional physicochemical vectors, it is equivalent to Equation 10 in Neher (1994). The third method we use is also the third (and primary) method of Neher (1994), where the sum is over all *M* pairs where the residues at site *i* (and at site *j*) are non-identical (Neher's formulation is different, but the expressions are mathematically identical). This will be denoted the NEH94 method. One unusual feature of this method is

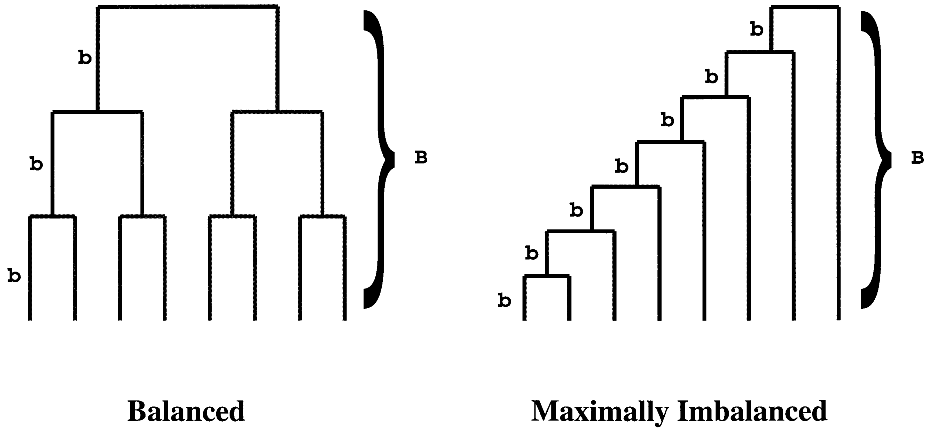

Fig. 2. Structure of trees. The structure of a balanced and a maximally imbalanced eight taxon tree. The degree of balance is dependent on the branching pattern, rather than the regularity or evenness of branching with time. Trees in this study were generally evenly branching, as depicted here, unless noted otherwise. The branch lengths, where noted, are labeled b, whereas the overall tree length (or depth) is labeled B.

that correlation values can be <–1.0 or >1.0. This can occur because the variances are calculated over all residues where a site *i* is non-identical, whereas the correlation is calculated over all sites where both *i* and *j* are non-identical. A more rigorous method would be to recalculate the variances of both sites *i* and *j* for each pairwise comparison, including in both calculations only those sequences included in the correlation calculation. We prefer to maintain consistency with the methodology of Neher (1994) for comparison purposes. This third method is also similar to the use of vectorial measures by Taylor and Hatrick (1994), except that they developed a technique for cluster analysis rather than a standard correlation measure. The cluster analysis was developed primarily to allow for the detection of negative correlation but, as we will show, the methods described above can detect negative correlation in the case of physicochemical vectors, while the GOB94 method can also detect negative correlation with the similarity matrices. Our final method is a compromise between the previous two, where the sum is over all $M = (N^2 - N)/2$ *different* sequence pairs, and will be called the NSC (no self-comparisons) method.

Physicochemical vectors considered were side-chain volume (size) (Grantham, 1974), charge (Taylor and Hatrick, 1994) and hydrophobicity (Levitt, 1976). Residue similarity matrices considered were based on either physical considerations (McLachlan, 1971), denoted MCL71, substitutional probability (Dayhoff *et al.*, 1978), denoted P120 or P250, or on substitutional probability converted to property-based similarity (Taylor and Jones, 1993), denoted TAY93S. Distance matrices were based on either the calculated distance in the three-dimensional space formed by the three physicochemical vectors, denoted GTLD, or on substitutional probability converted to Euclidean distances (Taylor and Jones, 1993; Taylor and Hatrick, 1994), denoted TAY94D. In the case of the physicochemical vectors the mean vector between all sequence pairs at a site must be zero, and both the GOB94 and the NSC methods are identical with the BASIC method.

In order to evaluate the effectiveness of a method combined with a physicochemical metric or a similarity matrix, the distributions of background correlations were compared with those of correlations at truly correlated sites. The residue pairs were clustered into bins of correlation values (width = 0.04) from −1.0 to 1.04, and the fraction of pairs falling into each bin out of all residue pairs considered for each method was recorded. In the case of the Neher (1994) method, pairwise correlations >1.0 were put into the last bin (which is empty for all other methods), and correlations <–1.0 were placed in the first bin. The expected number of background correlations above specific cut-offs is also considered, since this number will increase approximately with the square of the number of sites considered, whereas the number of physically paired sites in a protein structure (presumably proportional to the number of truly correlated sites) will increase only linearly with the number of sites.

Results

Effect of methodology

The different methods were applied to evenly balanced trees of different depths containing 8, 16 or 32 taxa and the background distribution of pairwise correlations using the MCL71 matrix was determined. A striking result is that all distributions for the GOB94 method are strongly biased towards positive values (Figure 3). This indicates that negative correlation, which should occur at random exactly as often as positive correlation, is difficult to observe with this method. It also means that a large portion of the background distribution occurs at higher positive values, which will make it difficult to discriminate these values from sites which are systematically evolving in a correlated fashion. In order to detect true correlation as efficiently as possible, the background correlation distribution should be clustered tightly around zero. Background frequencies are particularly high for larger correlation values when there are fewer taxa and higher rates along branches.

The main reason for this positive bias in the correlation appears to be the non-independence of the taxon-pair comparisons. In particular, the inclusion of self-similar comparisons tends to bias the mean correlation. This is made clear in Figure 4, where the NSC method is used and self-similar comparisons are not included. In this case, the bias has disappeared, but the distribution is extremely non-symmetrical for branch rates less than 100.0 PAMs, and larger negative correlations are still entirely absent. The reason for this inability to observe negative correlation with these methods is discussed by Taylor and Hatrick (1994) and Hatrick and Taylor (1994). One can imagine that even with perfectly negatively correlated sites, all invariant taxon pairs at the sites will make a large positive contribution

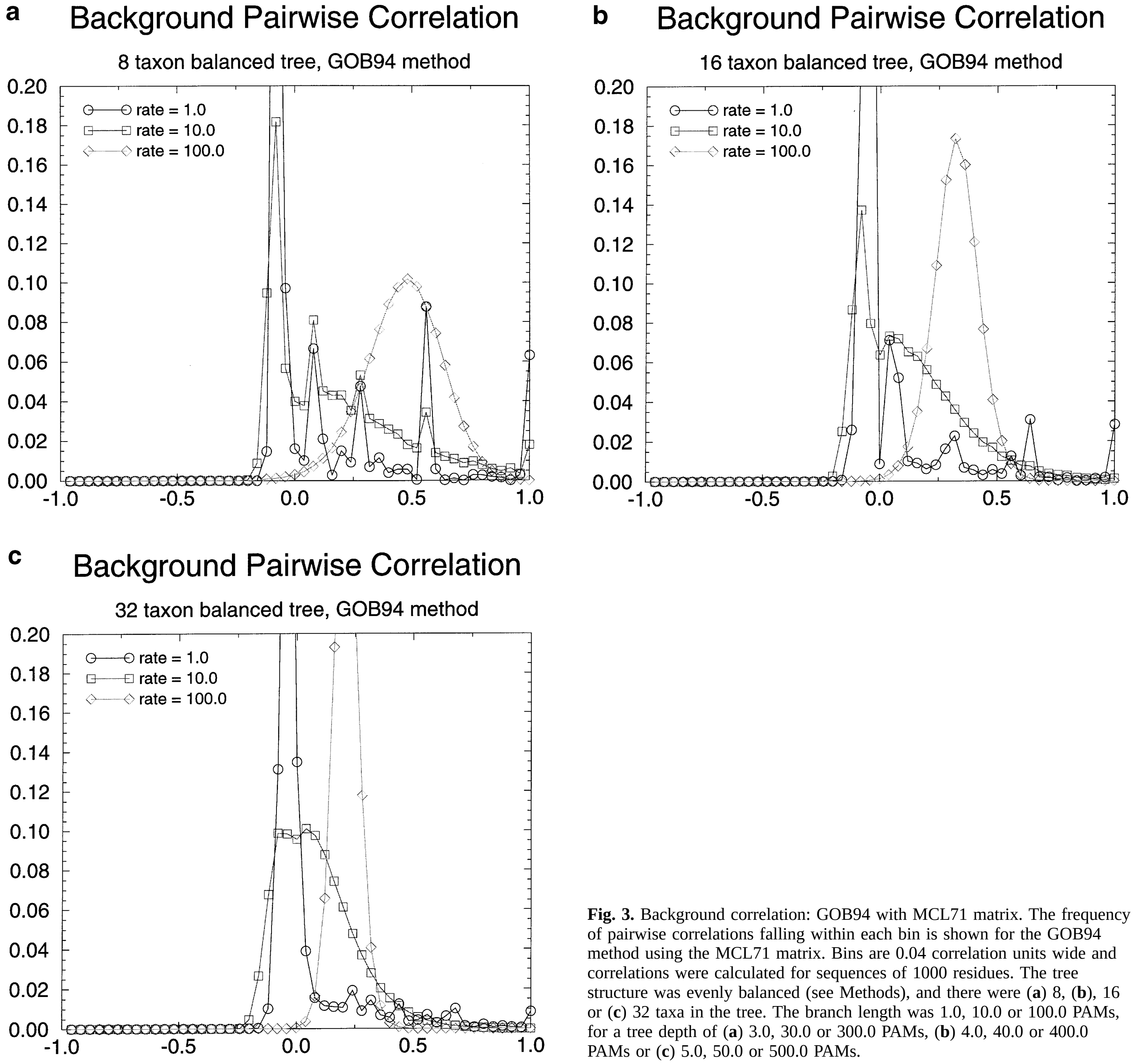

Fig. 3. Background correlation: GOB94 with MCL71 matrix. The frequency of pairwise correlations falling within each bin is shown for the GOB94 method using the MCL71 matrix. Bins are 0.04 correlation units wide and correlations were calculated for sequences of 1000 residues. The tree structure was evenly balanced (see Methods), and there were (**a**) 8, (**b**), 16 or (**c**) 32 taxa in the tree. The branch length was 1.0, 10.0 or 100.0 PAMs, for a tree depth of (**a**) 3.0, 30.0 or 300.0 PAMs, (**b**) 4.0, 40.0 or 400.0 PAMs or (**c**) 5.0, 50.0 or 500.0 PAMs.

to the pairwise correlation, which will tend to counterbalance the negative correlation in the changed (and negatively correlated) taxon pairs.

When invariant comparisons are eliminated from the correlation calculation (i.e. the NEH94 method), both the bias and the asymmetry of the background correlation frequency distribution are also eliminated (Figure 5). Although this is generally good, there is still a large effect from the tree structure causing the distribution to be decidedly non-normal. With the deepest tree, tree structure is essentially irrelevant, and the sequences are nearly independent. With the shorter trees, however, tree structure has greater effect, and the sequences are more correlated owing to that structure. Thus, the background correlation frequencies are higher for the larger positive and negative values, and the frequency distribution is extremely leptokurtic (has positive kurtosis). In addition to predicting that there will still be a large number of false-positive correlations due to background noise, this also means that statistics to calculate the significance of a correlation value, such as those used in Neher (1994), will be inaccurate since they are based on the assumption of a normal distribution.

Number of taxa and tree structure

In order to assess the effect of the number of taxa, the tails of the above distributions were compared for branch lengths of 10.0 and 100.0 PAMs (Figure 6a–c). The GOB94 and NSC methods combined with the MCL71 similarity matrix were also applied to balanced trees 30.0 PAMs deep containing 8, 16 or 32 taxa (Figure 6d). As might be hoped, increasing the number of taxa decreases the variance of the distribution and decreases the average magnitude of the background pairwise correlations. The NSC method has slightly less background than GOB94 at most correlation values and number of taxa combinations (the NEH94 method has similar values to NSC;

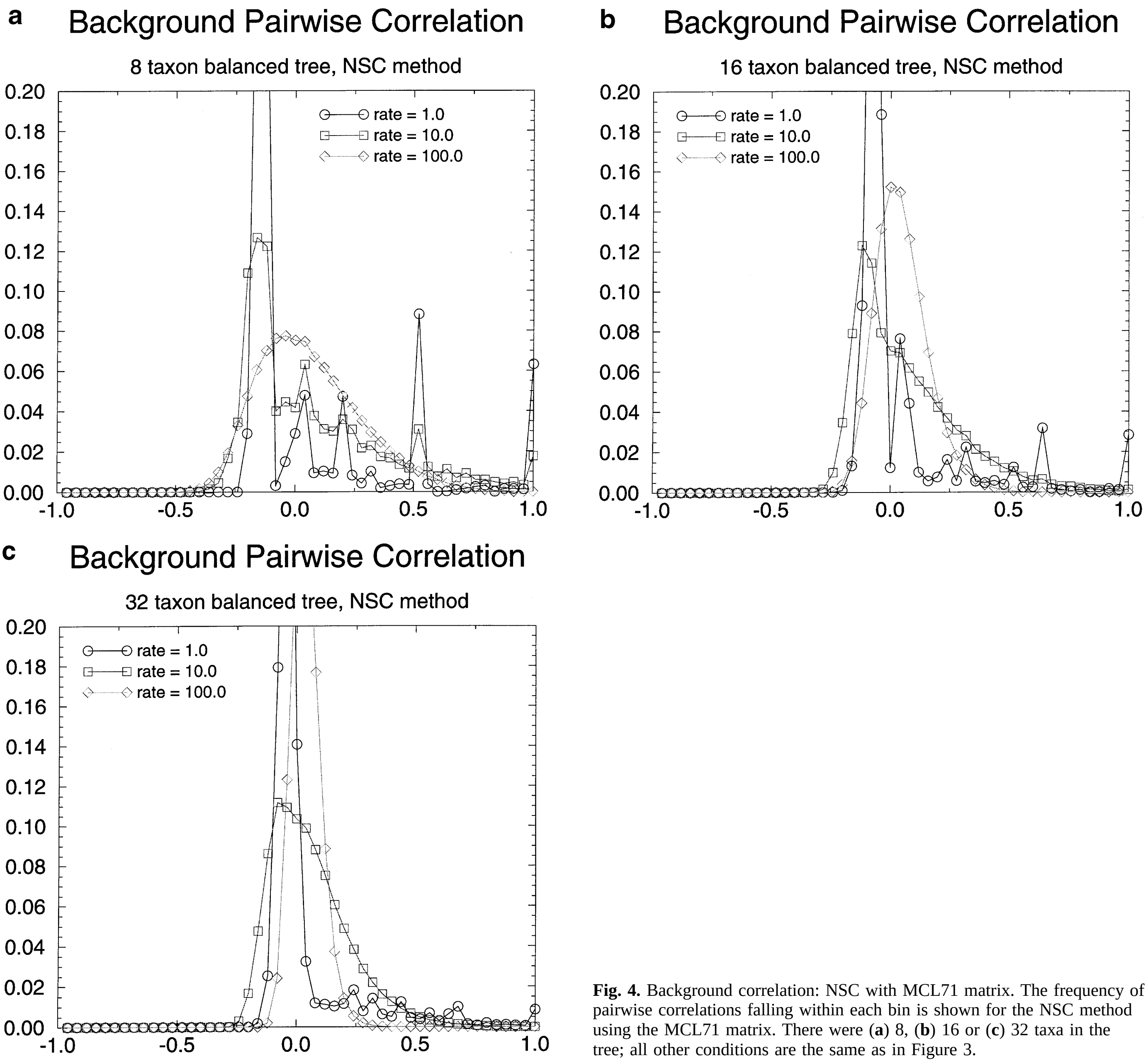

Fig. 4. Background correlation: NSC with MCL71 matrix. The frequency of pairwise correlations falling within each bin is shown for the NSC method using the MCL71 matrix. There were (**a**) 8, (**b**) 16 or (**c**) 32 taxa in the tree; all other conditions are the same as in Figure 3.

data not shown). As noted above, the GOB94 method has much higher background frequencies for the largest branch length, but when the tree depth is held constant at 30.0 PAMs, the differences in background frequencies are much greater for different numbers of taxa than for the different methods.

Even with many taxa, the number of background correlated residues above stringent cut-offs can be large (Figure 6d): for the 16 taxon tree and the NSC method, 0.54% of the distribution falls above 0.9, 1.9% falls above 0.7 and 6.2% falls above 0.5. This means that for a comparison involving 100 residues (with 4950 independent comparisons), by chance alone 27, 95 and 307 residue pairs will be expected to fall above these cut-offs, respectively. For the 32 taxon tree, the expected numbers of background pairs above these cut-offs are 6, 39 and 155.

The effect of tree structure was assessed by varying both the balance of the trees and the evenness of the branches on 16 taxon trees (Figure 7). As in Figure 6d, the trees were all 30.0 PAMs deep. The length of branches can have a considerable effect on the background correlation values of the GOB94 and NSC methods. When the terminal branches (those leading to the tips of the tree) are long relative to the internal branches, the background is slightly less than when all branches have equal length (Figure 7a). When the terminal branches are short relative to the internal branches, the background is considerably higher. The expected number of residue pairs above 0.9 for a 100 residue comparison is 511, making these methods unusable under these conditions. The NEH94 method does not suffer from this effect, and background correlation values are similar for all branch spacing (data not shown). In Figure 7b are shown the background correlation distributions for imbalanced and balanced eight taxon trees, and it is clear that the balance does not have a large effect and there is in general no strong interaction between the balance and the methods. A slight exception is that the NEH94 method background is approximately doubled for imbalanced trees at correlation values of 0.8 and higher.

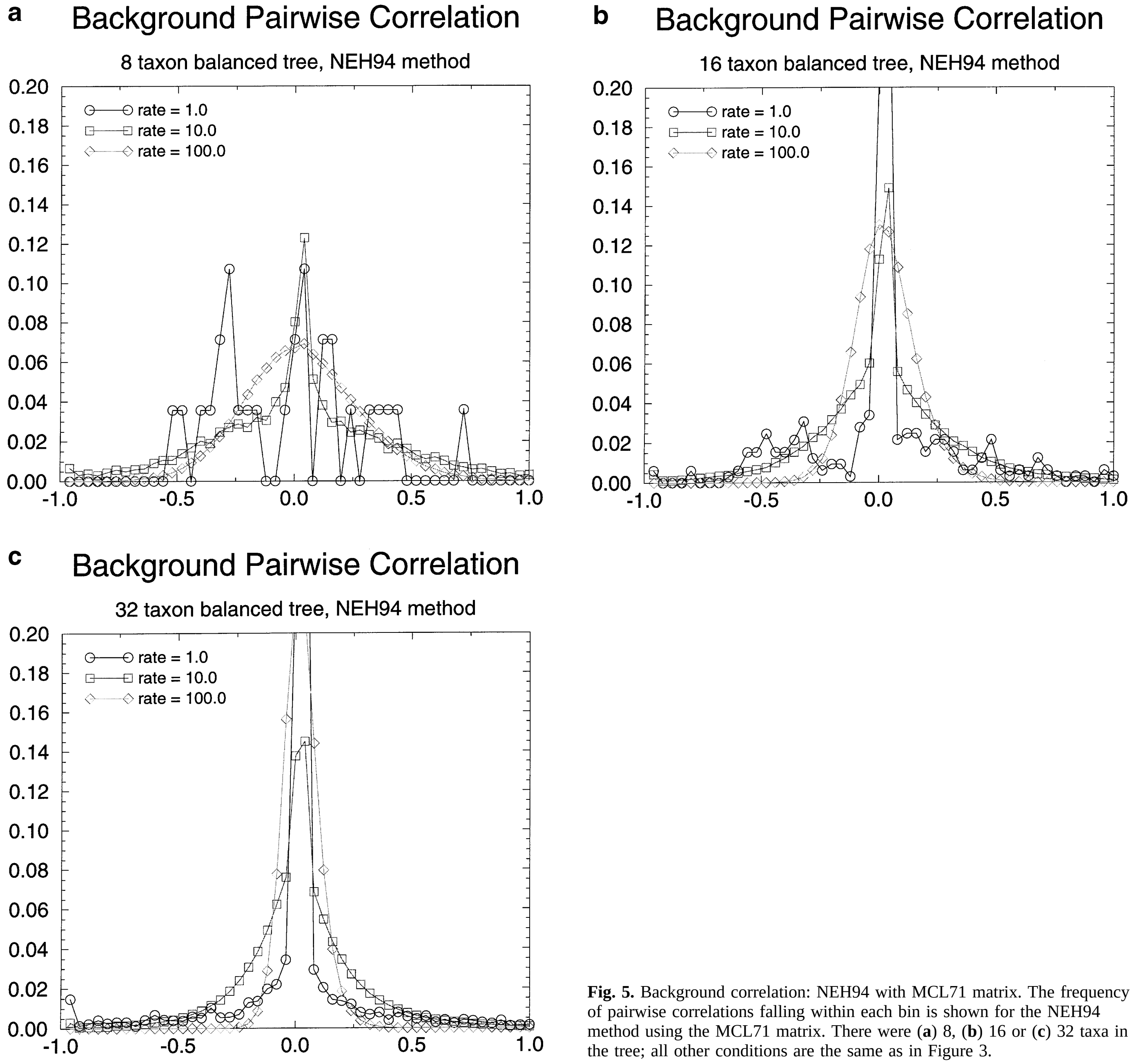

Fig. 5. Background correlation: NEH94 with MCL71 matrix. The frequency of pairwise correlations falling within each bin is shown for the NEH94 method using the MCL71 matrix. There were (**a**) 8, (**b**) 16 or (**c**) 32 taxa in the tree; all other conditions are the same as in Figure 3.

Comparison of metrics and similarity matrices

The nine different matrices and metrics were used with the NEH94 method in order to compare their background distributions. Each combination was applied to balanced trees 40.0 or 400.0 PAMs deep containing 16 taxa (Figure 8). The most distinctive difference is between the matrices and the vectors. The distributions of all the matrices have much less variance than those of the vectors at both tree depths. There is very little difference between the similarity matrices, although the TAY93S matrix has slightly greater background frequencies around zero, and slightly smaller frequencies at the tails. The distance matrices are more leptokurtic than the similarity matrices in the 40.0 PAM tree, but the differences at the tails (correlation values >0.5 and <–0.5) are minimal. The effect of tree depth is also notably different between the matrices and the vectors. At a tree depth of 400.0 PAMs, the sequences are nearly independent and tree structure is almost irrelevant, whereas at 40.0 PAMs the sequences are correlated with one another. While all the matrices behave similarly to the MCL71 matrix (investigated more thoroughly in the previous section) in that the distributions have greater variance and are leptokurtic, the physicochemical vectors (charge, size and hydrophobicity) have a greater increase in variance and do not become obviously leptokurtic. Under the simulated conditions, they are approaching a flat distribution except for a jump in frequency at 1.0 and −1.0. This means that the background distribution is high at the tails; for example, 11.9% of background correlation values fall above 0.9 for the NEH94 method using the size vector.

The distinct behavior of the physicochemical vectors as compared with the similarity matrices with the NEH94 method makes it appropriate to compare the methods again, this time using a physicochemical vector. As discussed above, the BASIC, GOB94 and NSC methods are mathematically equiva-

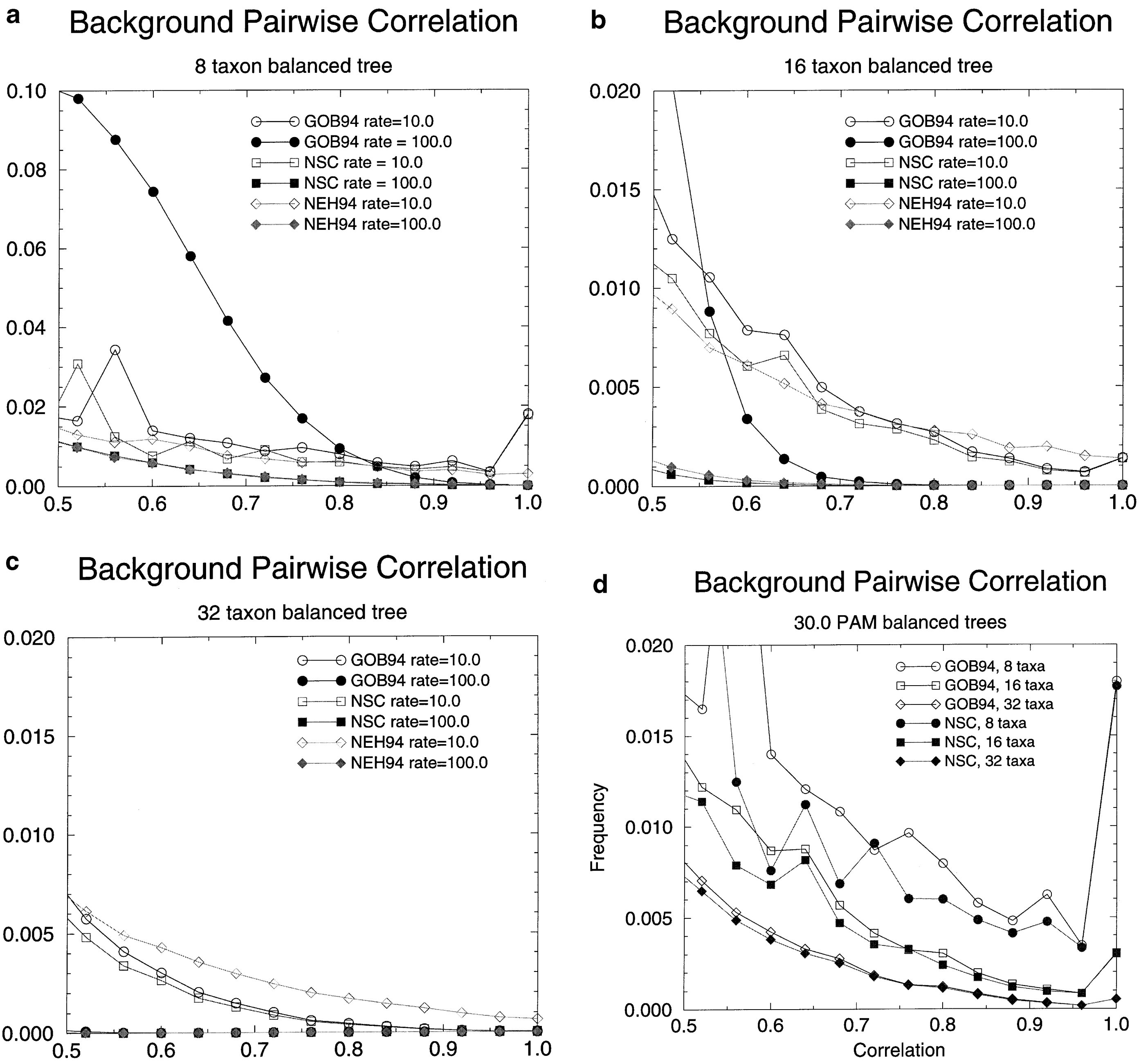

Fig. 6. Background correlation: number of taxa. The frequency of pairwise correlations falling within each bin are shown for correlation value <0.5. In (**a**), (**b**) and (**c**), the GOB94, NSC and NEH 94 methods using the MCL71 matrix are shown. Trees were balanced and contained (**a**) 8, (**b**) 16 or (**c**) 32 taxa. Results for trees with branch lengths of 10.0 and 100.0 PAMs are shown. In (**d**) the conditions were the same as in Figure 3, 4 and 5, except that the tree depth in all cases was 30.0 PAMs. The minimum separation between sequences (in the 32 taxon tree) was thus 12.0 PAMs, or ~12%. For purposes of clarity the NEH94 method is not shown, but frequencies with that method are close to the NSC method.

lent when applied to one-dimensional vectors. Therefore, we addressed this question by simulating the background distribution of correlation values for both distinct methods (BASIC and NEH94) using the size vector and the same tree structure and depths as in the previous set of simulations (Figure 9). While the distributions are similar for the trees 400.0 PAMs deep, the NEH94 method has a much broader distribution than the BASIC method with the 40.0 PAM tree. For example, the BASIC method has only 0.42% of background values above a correlation of 0.9, compared with 11.9% for the NEH94 method. This means that for a comparison involving 100 sites, by chance 589 correlation values are expected to be >0.9 for this tree structure using NEH94, whereas only 21 would be with the BASIC method. Thus, based on levels of background correlation alone, it appears that the BASIC method is likely to be preferable for physicochemical vectors.

Detection of true correlation

While background levels of correlation are extremely important in determining the efficiency of a correlation method and distance/similarity matrix or physicochemical vector combination, it is worthwhile to consider also the ability to detect signal, or true correlation. Our ability to detect true correlation in nature will, of course, depend on the number of sites which are truly correlated and the degree of correlation at those sites, both of which are at this point completely unknown. It is

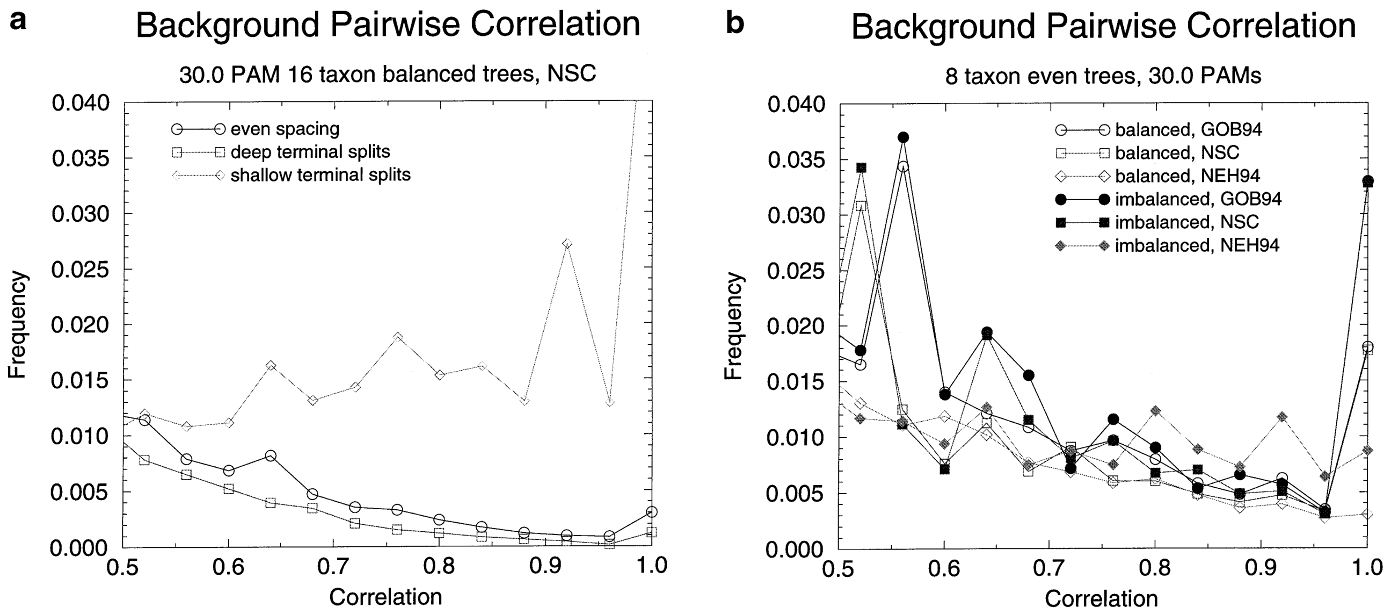

Fig. 7. Background correlation: tree structure. The percentage of pairwise correlations falling within each bin is shown for different tree structures. In (**a**), the NSC method is shown for clarity. All trees are balanced and contain 16 taxa, but the distribution of branch lengths is either even (all branches = 5.0 PAMs) or contains an excess of long branches (deep terminal splits; terminal branches are 22.5 PAMs, internal branches are 2.5 PAMs), or an excess of short branches (shallow terminal splits; first two branches are 22.5 PAMs, all others are 2.5 PAMs). The correlation frequencies for the GOB94 method are similar to those shown, while the NEH94 method is similar except in the case of the shallow terminal splits, where the frequencies closely match the even and deep terminal split spacing. In (**b**), the GOB94, NSC and NEH94 methods are shown for eight taxon trees with evenly spaced splits, but the tree structure is either balanced or imbalanced. In all cases in both (**a**) and (**b**), trees were 30.0 PAMs deep. Other conditions were the same as in Figure 3. The minimum separation of taxa (in the eight taxon imbalanced tree) was 8.58 PAMs.

possible, however, to devise ways of producing paired sites which have varying degrees of correlation expected at equilibrium. This allows us to compare how fast the correlation observed with different methods falls off as the equilibrium expectations decrease. To that end, we have devised the simple two-residue-per-site model for producing correlated sites described in the Methods section (Figure 1).

The model also aids in making clear what is meant by ‘true correlation’. When the probabilities of substitution at the dependent site change depending on the residue at the driver site, the two sites are correlated. This correlation can be effected by changing either the equilibrium residue frequencies (π_i) at the dependent site or the probable rates of substitution (λ_i). A clear consequence of the model as it is structured is that even when two sites are what we would intuitively think of as completely correlated (that is, when the equlibrium frequency of one residue at the driver site is 1.0 in the presence of a particular residue at the driver site, whereas it is 0.0 in the presence of the other residue at the driver site) they will not be completely correlated at equilibrium if the rates of exchange at the dependent site are small or those at the driver site are large. The model can also produce both positive and negative correlation by choosing residue pairs at both sites which differ in one or more physicochemical values. Thus a positive correlation will occur when larger values are associated at the two sites, whereas a negative (or compensatory) correlation will occur when larger values at one site are associated with smaller values at the other site. We note here that although compensation of physicochemical quantities is logically more likely to lead to functional compensation in the protein structure, this is not necessarily the case.

In order to observe the effect of changing the substitution rates on the correlation distributions using different methods and metrics, we ran simulations of both positively and negatively correlated sites using all available methods and both the size vector and the MCL71 matrix. The phylogenetic trees relating the sequences again contained 16 taxa and were completely balanced as before. Rates of substitution were either 0.1, 1.0 or 10.0 at both the driver and dependent sites. Equilibrium frequencies at the driver site were 0.5 and frequencies at the dependent site were either 1.0 or 0.0 in order to create completely positively or negatively correlated sites, as described above. In Figure 10 are shown the results for positive correlation when the two residues are glycine (G) and alanine (A) at both sites and the rate at the driver site is 1.0.

With the MCL71 matrix (Figure 10a), the distributions are uneven for both the GOB94 and NSC methods; this is due to the limited number of possible correlation values with only two residue types in the model. The observed correlation is always 1.0 when the rates of substitution at the dependent site are 1.0, but the mean falls off quickly as the rate decreases to 1.0, and is nearly zero for both methods when the rate is 0.1. The observed correlation decreases more quickly with the NSC method, but not by a large amount. The results shown for residues G and A are generalizable to other residue pair combinations because although the magnitudes of change will differ, the correlations are weighted by the standard deviations at each site, which normalizes the outcome. The distributions for negatively correlated, or compensatory, sites are nearly identical with those for positively correlated sites (data not shown). This means that positive and negative correlations are indistinguishable, and it occurs because the distance and similarity matrices measure only the magnitude of change, and not the direction. Although the rate of substitution at the driver site has some effect on the distributions, for the rates which were simulated (0.1, 1.0 and 10.0) this effect is small compared with the effect of the rate of substitution at the dependent site (data not shown).

The NEH94 method is always undefined with this correlated model and the MCL71 matrix; this is because with only two

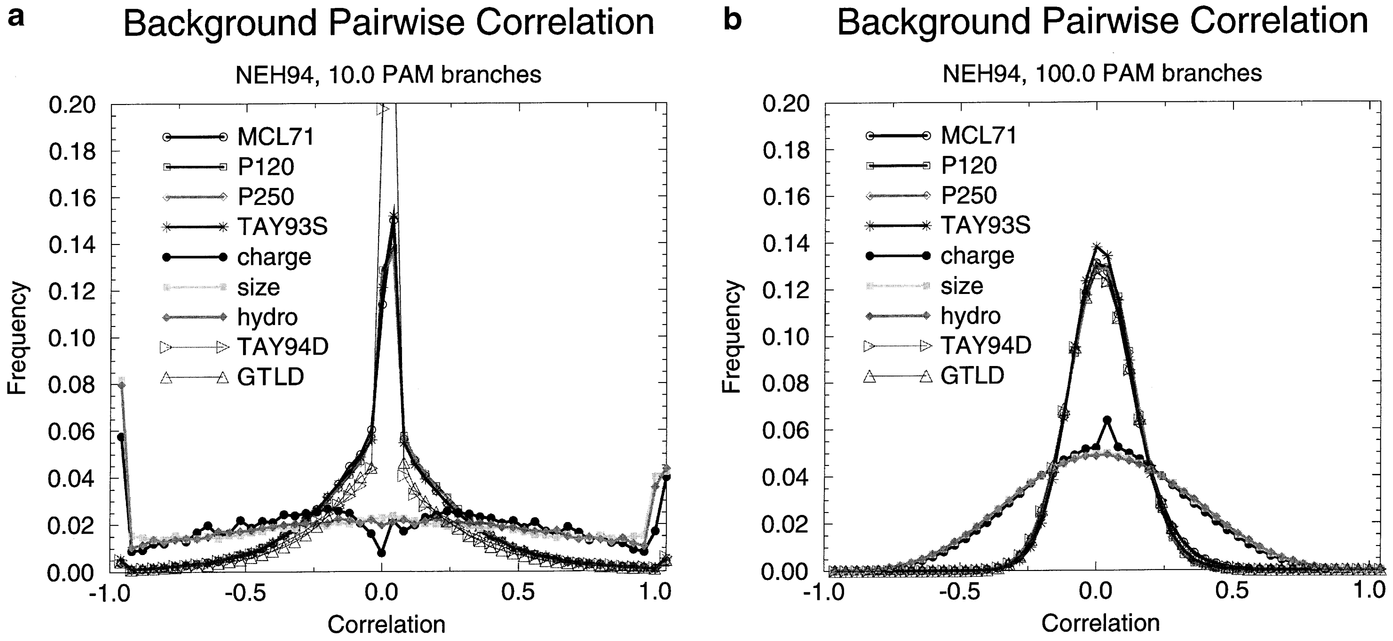

Fig. 8. Background correlation: matrices and metrics. The percentage of pairwise correlations falling within each bin is shown for different matrices and physicochemical vectors. In all cases, tree were evenly balanced and the NEH94 method of calculating correlations was used. The branch lengths were either (**a**) 10.0 or (**b**) 100.0 PAMs. There were 16 taxa in each tree, and thus tree depths were either 40.0 or 400.0 PAMs. Other conditions were the same as in Figure 3.

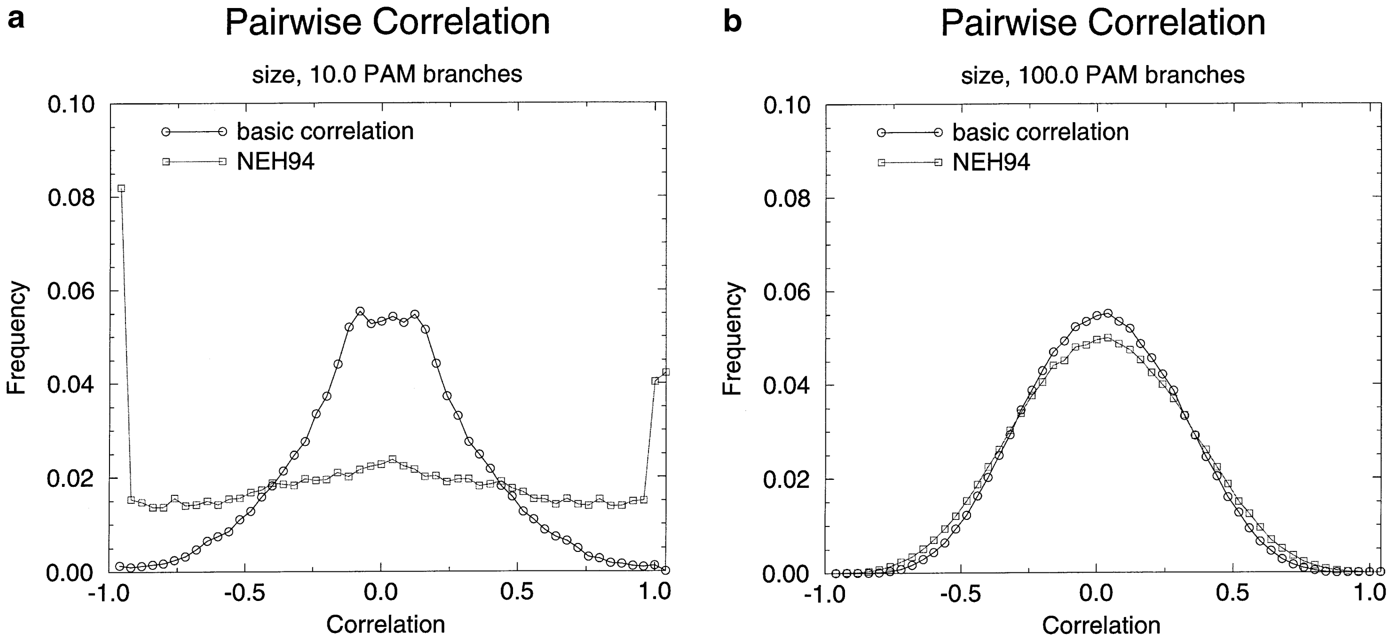

Fig. 9. Background correlation: size vector. The percentage of pairwise correlations falling within each bin is shown for the BASIC and NEH94 methods using the size vector. The branch lengths were either (**a**) 10.0 or (**b**) 100.0 PAMs; all other conditions were the same as in Figure 8.

residues at a site and a symmetric matrix, there is only one value of change when the residues differ. Since the NEH94 method is calculated only between those sequences where the residues do differ, there is no variance at a site, and the correlation is undefined. Although it might be argued that the problem is with our model for generating correlation, and not the method for assessing it, the inability of this method to detect simple correlations is a serious drawback, and even if other residues were added to generate variance at a site, it is clear that the correlation measured would be strongly dependent on the similarity of the uncorrelated residues to the correlated residues, and might have little to do with the magnitude of correlation being generated.

With the size vector, the distributions for positively and negatively correlated sites are symmetric about zero, and so only the positively correlated sites are shown in Figure 10a. As with the MCL71 data, when the substitution rates at the dependent site are 10.0, perfect correlation is observed with both useful measures. For smaller rates, mean correlation values for the NEH94 method are higher than for the BASIC method. When the substitution rates at the dependent site are 1.0, the mean correlation values are 0.936 for NEH94 and

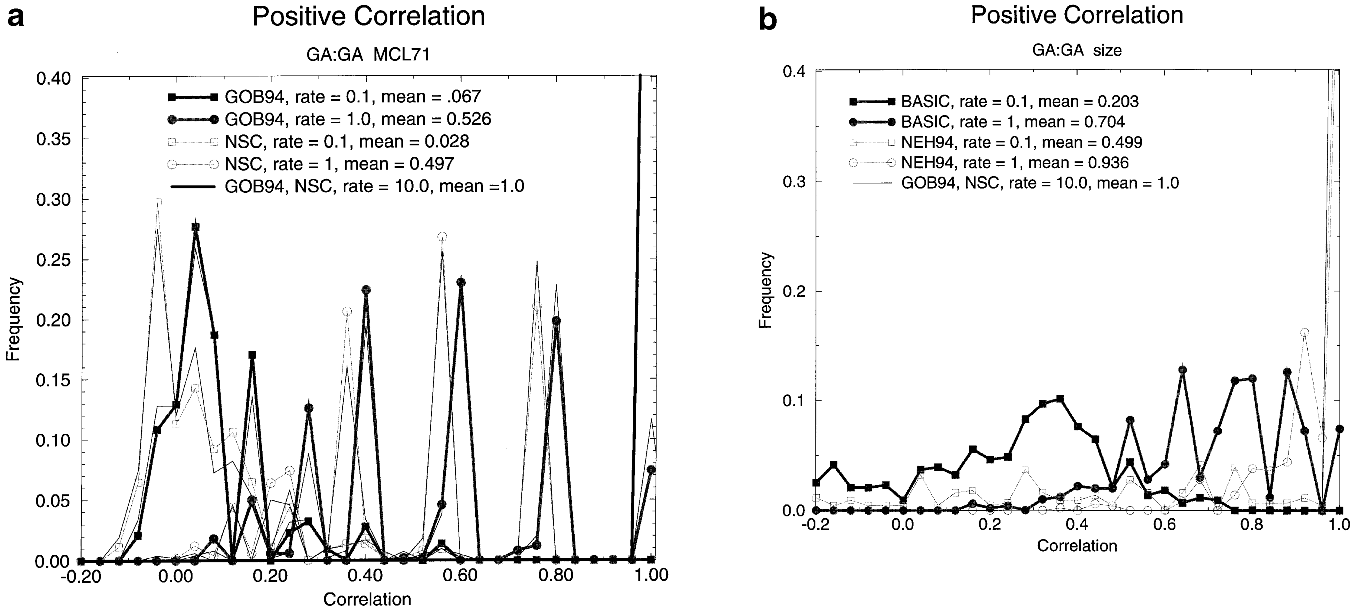

Fig. 10. Positive correlation. The percentage of pairwise correlations falling within each bin is shown for different correlation methods and rates of substitution at the dependent site (see text). Correlations were calculated for 500 correlated pairs and the tree structure was evenly balanced. There were 16 taxa in each tree, branch lengths were 1 and the rates of change at the driver sites were 1.0 (equivalent to 100.0 PAMs along a branch), while the equilibrium frequencies of both residues at the driver site were 0.5. The two residues at the driver and dependent sites were glycine (G) and alanine (A), and all site pairs were positively correlated, meaning that the equilibrium frequency of G at the dependent site was 1.0 in the presence of G at the driver site and 0.0 in the presence of A at the driver site. In (**a**), the frequency distributions of correlation values are shown for the MCL71, matrix using the GOB94 and NSC methods with dependent site rate of 0.1, 1.0 and 10.0. In (**b**), the distributions of correlation values are shown for the size vector using the BASIC and NEH94 methods and the same dependent site rates. In both cases, methods and dependent site rates are identified in the legend along with mean correlation values.

0.704 for the BASIC method. When the rates are 0.1, the mean correlation values are 0.499 for NEH94 and 0.203 for BASIC. Note that these mean correlation values are much higher than with any method in combination with the MCL71 matrix. Again, the changes in distributions for different substitution rates at the driver site are small compared with the changes for different rates at the dependent site. These results mean that neither method is *a priori* preferable over the other. This is because while the NEH94 method is better at retaining a weak signal (low dependent substitution rate), the background noise is unacceptably high. In contrast, the background for the BASIC method is for much lower correlation values close to 1.0 or −1.0, but truly correlated sites are less likely to appear in these regions. The signal-to-noise ratio of these methods when used on natural sequences will depend on the distribution of types of correlation, which is unknown.

Discussion

When similarity and distance matrices are used, out of all the methods to detect correlation considered, the pairwise comparison method which ignores invariant comparisons (NEH94), proposed by Neher (1994), appears to be the most effective at reducing random background correlation. It is unable to detect a simple positive correlation, however, and because of this the NSC method, which has only a slightly greater background at correlation values >0.5, is preferable on these criteria. The GOB94 method of Gobel *et al.* (1994), on the other hand, has an unacceptably poor ability to discriminate true correlation from background correlation under some relevant conditions, and probably should not be used for this purpose. The NEH94 method is the only method relatively insensitive to the addition of similar taxa, as long as a sufficient number of highly diverged taxa are present. Although the NEH94 method is less sensitive than the other methods to the addition of extremely similar taxa, it is likely that such additions will lead to greater inaccuracy in the correlation estimate for all methods. Tree structure generally imparts extreme non-normality in the form of kurtosis to the correlation distributions, and this invalidates significance statistics based on the assumption of a normal distribution.

When physicochemical vectors are used, the BASIC method is better than the NEH94 method at detecting strong positive or negative correlations owing to the relatively low expectation of background, or falsely positive, correlation. The NEH94 method is better at detecting weaker signals (where the substitution rate at the dependent site is slow), including some which the BASIC method is unlikely to detect at all. These will be difficult to separate from an exceedingly high level of background correlation, however. The NEH94 method was designed to detect more easily correlation with physicochemical vectors, but it appears to succeed all too well, and random correlation is enhanced at least as much as true correlation.

The results of this study put into doubt previous conclusions that the large number of correlations in residues which are not in contact with one another are due to long-range physical interactions. Rather, a significant proportion of them (if not all) are more likely due to random background noise. Increasing the number of taxa lowers the levels of background correlation, but since background pairwise comparisons increase with the square of the number of sites, a significant number of false positives are expected for correlation analyses with even a small number of residues (50–100). Most of the methodological and metric combinations succeed in improving the ratio of truly correlated sites compared with randomly evolving sites. They may thus prove useful in combination with other information for protein structure prediction in methods such as distance geometry or threading, assuming that truly correlated sites tend to be physically close. None of the methods can identify truly correlated sites without including a large number of uncorrelated sites.

The large number of residue pairs with high correlation expected simply due to background noise in the presence of phylogenetic structure makes it imperative to consider other tree-based methods to analyze pairwise correlation. The tree-free methods evaluated in this paper are likely to be optimal only in the case where the sequences themselves have no phylogenetic relationships, a situation which is highly unlikely; in any case, the alignments would be suspect under such conditions. It should be recalled that real sequences are likely to contain sites with a distribution of rates, and thus high background is likely to occur at at least some sites regardless of the average separation between taxa. The method of Shindyalov *et al.* (1994) is a reasonable start towards incorporating phylogenetic information into correlation analysis, and with such methods the information included by addition of close taxa will be beneficial. The efficiency of this method in comparison with the methods reviewed in this paper will be analyzed in a future study.

Acknowledgements

D.D.P. is a Hitchings–Elion fellow of the Burroughs Wellcome Fund.

References

Altschuh,D., Lesk,A., Bloomer,A.C. and Klug,A. (1987a) *J. Cell. Biochem.*, **S11e**, 233.

Altschuh,D., Lesk,A.M., Bloomer,A.C. and Klug,A. (1987b) *J. Mol. Biol.*, **193**, 693–707.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, pp. 345–352.

Felsenstein,J. (1981) *J. Mol. Evol.*, **17**, 368–376.

Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins: Struct. Funct. Genet.*, **18**, 309–317.

Goldman,N., Thorne,J. and Jones,D. (1996) *J. Mol. Biol.*, **263**, 196–208.

Grantham,R. (1974) *Science*, **185**, 862–864.

Hatrick,K. and Taylor,W.R. (1994) *Comput. Chem.*, **18**, 245–249.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *FEBS Lett.*, **339**, 269–275.

Levitt,M. (1976) *J. Mol. Biol.*, **104**, 59–107.

McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409–424.

Neher,E. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

Pagel,M. (1994) *Proc. R. Soc. Lond. B*, **255**, 37–45.

Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) *Protein Engng*, **7**, 349–358.

Sokal,R. (1990) *Evolution*, **44**, 1671–1684.

Taylor,W.R. and Hatrick,K. (1994) *Protein Engng*, **7**, 341–348.

Taylor,W.R. and Jones,D.T. (1993) *J. Theor. Biol.*, **164**, 65–83.

Received 6 December 1996; accepted 4 March 1997