Theor Chem Acc (1999) 101:9–15
DOI 10.1007/s002149800m86

Theoretical
Chemistry Accounts

Regular article

How possible is the detection of correlated mutations?*

P. Tufféry, M. Durand, P. Darlu

INSERM U155, Université Paris 7, case 7113, 2, place Jussieu, F-75251 Paris, France

Received: 24 April 1998 / Accepted: 8 August 1998 / Published online: 11 November 1998

Abstract. The remarkable conservation of protein structure, compared to that of sequences, suggests that, in the course of evolution, residue substitutions which tend to destabilise a particular structure must be compensated by other substitutions that confer greater stability on that structure. Given the compactness of proteins, spatially close residues are expected to undergo the compensatory process. Surprisingly, approaches designed to detect such correlated changes have led, until now, only to limited success in detecting pairs of residues adjacent in the three-dimensional structures. We have undertaken, by simulating the evolution of DNA sequences including sites mutating in a correlated manner, to analyse whether such poor results can be attributed to the detection methods or if this failure could result from a compensatory process more complex than that implicitly underlying the different approaches. Present results show that only methods taking into account the phylogenetic reconstruction can lead to correct detection.

Key words: Correlated mutations – Phylogeny – Sequence alignment

1 Introduction

It has long been suggested that, in the course of evolution, residue substitutions which tend to destabilise a particular structure must be compensated by other substitutions that confer greater stability on that structure. It is reasonable, given the compactness of protein structures, to suppose that it is residues close to each other that undergo the compensatory process. Such a hypothesis has led in recent years to increasing interest in developing approaches able to detect correlated changes in sequence evolution, with the aim to derive some spatial proximity information that could help protein structure prediction. One can refer to the pilot work of Altschuh et al. [1], who analysed the amino acid substitutions in the coat protein structure of tobacco mosaic virus and seven related viruses and showed that some pairs of positions with identical patterns of amino acid substitutions are close together.

Surprisingly, the different approaches proposed have led only to limited success in identifying pairs of residues that are close together in the structures. Among the possible explanations underlying this, three major reasons have been proposed:

1. Distant residues can also participate in the compensatory process. Such a hypothesis is compatible with the fact that the compensation can also be invoked concerning the preservation of the function of a given protein.
2. The compensation involves numerous residues. Thus a detection process based on a pair detection could be unsuited.
3. The criteria used to detect the correlated sites are inefficient or the approaches employed are not relevant. The present work focuses on this last point.

Two different paradigms underly the recent approaches that have been described:

1. Some studies are based on analysis of the observable set of sequences (taxa) [2–4]. These studies focus on the detection of the correlation that is expected to be associated with the compensatory process. Namely, the distribution of the amino acids observed at pairs of sites within a collection of aligned sequences, weighted by some residue similarity factor, should be correlated in some manner if there is some compensation between the two sites.
2. Another study [5] used a phylogenetic reconstruction to estimate the likelihood of the occurrence of correlated mutations.

The question of assessing the effectiveness of these two paradigms remains. Recent work [6] has focused on the effectiveness of approaches not requiring phylogenetic reconstruction. The authors start from a simulation of

*Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France
Correspondence to: P. Tufféry

the evolution of protein sequences, involving pairs of correlated residues. For such pairs, the residues are constrained to undergo a toggle process between two given amino acid states. They show that, for all the various methods tested, a major difficulty comes from the low ratio of the truly correlated sites detected compared to the background noise.

We present here results obtained by simulating the evolution of random DNA sequences, allowing or not the occurrence of correlated mutations, and we check the effectiveness of methods based upon the analysis of the aligned taxa sequences alone, and of an approach using the information of the phylogenetic reconstruction. The choice of simulating DNA sequences was dictated by several considerations:

1. The use of DNA sequences results in a simpler system to simulate and analyse. In particular, it allows us to avoid the problem of measuring the similarity between amino acids, which might introduce some noise in the detection of the correlated pairs. Also, the number of states falls from 20 to 4.
2. Phylogenetic reconstruction of DNA sequences is much simpler than for protein sequences.
3. It is possible to return from DNA to protein sequences.

Using such methodology, it is possible to control many factors that may interfere with the detection of the correlated pairs. The alignments are perfect and the ancestor sequences unambiguously identified if desired.

2 Methods

2.1 Simulation of DNA sequence evolution allowing correlated events along a phylogenetic tree

The simulation of the sequences was performed starting from a modified version of Seq-Gen [7]. This program allows simulation of the evolution of nucleotide sequences along a phylogeny, using common models of the substitution process such as the Hasegawa, Kishino and Yano (HKY) model [8] or the general reversible process [9]. It takes as input a phylogenetic tree, generates a random ancestor sequence and makes it evolve along the tree according to the evolution process selected. The mutation rates along each branch of the tree are a function of its length. A scaling procedure allows easy variation of the lengths of the whole tree (i.e. to simulate different mutation rates) without specifying a new input file. In our case, we have used the Kimura 2 parameters model (K2P), a simplified version of HKY that states that the base frequencies are equal, and the ratio transition/transversion of 0.5. Since our goal is to generate sets of sequences with the only interest that, at some locations, mutations are not independent, the use of more complex models is not justified.

Incorporating correlated mutational events was performed by defining clusters of correlated sites, and triggering their mutation according to the mutation of the "prototype" of the cluster. Namely, if the prototype mutates, all its correlated sites are forced to mutate (the possibility of partial correlation was not considered in the present study). Since, when we allow the cluster prototype to mutate, we trigger in fact several mutations, this might result in average branch lengths larger than desired. To avoid this, we introduce a correction on the mutation rate associated with the prototype that makes the probability of non-mutating all the sites of the cluster equal in the cases where they belong (or not) to a cluster. Finally, to make sure the variability inherent to the simulation process does not lead to large deviations between the expected and simulated branch lengths, a mechanism of verification "a posteriori" of the distance along the branch was implemented. The modifications of the program were designed so that the evolution process can incorporate (or not) correlated evolution. In the present study, mutations that occur during correlated mutational events are performed "at random" (i.e. the only condition is that a mutation occur). "Constrained" mutations (i.e. the mutation is constrained to avoid some possibilities) were not considered.

Two classical tree topologies were considered: a well balanced tree and a maximally imbalanced tree. These two topologies result in different distributions of the homologies between the taxa sequences: the maximally imbalanced tree favours a large distribution of the homology rates between the taxa sequences, while the well balanced one favours a more focused distribution. The distances between the nodes of the tree and between the nodes and the taxa were chosen close to 0.20 for the well balanced tree. For the maximally imbalanced tree, the distance between the nodes was fixed to 0.2 and the distance between the taxa and the nodes increases by 0.2 as the taxa are linked to nodes closer and closer from the root. For the well balanced tree, the distances were scaled by different values between 0.1 and 1, to obtain taxa sequence identity varying within 80–90% for 0.1, 40–60% for 0.5 and 25–50% for 1.0. For the maximally imbalanced tree the corresponding identity ranges were 99–60% for 0.03, 99–50% for 0.05, 96–35% for 0.1 and 93–30% for 0.15. For that tree, larger scalings were not considered owing to the mutation saturation.

2.2 Criteria to detect correlated mutations

Three different criteria have been used to detect correlated events.

1. Correlation measurement (G94): this criterion is that described by Gobel et al. [3]. It measures the correlation between similarity matrices describing positions i and j. Gobel et al. use a similarity index between amino acids in their similarity matrices. For DNA sequences, we use boolean similarity matrices (1 if the bases are identical, 0 elsewhere). The assessment of the signification of the score obtained is made by randomising the sequences, and counting the number of times a larger score is obtained.
2. Mutual information criterion (MIC): $\mathrm{MIC}_{ij} = H_i + H_j - H_{ij}$, where $H_i = \sum_{k=A,C,G,T} -p_k \times \ln(p_k)$, with p_k the observed frequency of the base k upon the position i of the aligned sequences, and $H_{ij} = \sum_{l=AA,AC,..,TT} -p_l \times \ln(p_l)$, with p_l the observed frequency of the base doublet l upon the position i and j of the aligned sequences. This criterion takes positive values. 0 is obtained for independent distributions for i and j. The assessment of the signification of the score obtained is made, as for G94, by randomising the sequences.
3. Phylogenetic probabilistic criterion (S94): this criterion is that of Shindyalov et al. [5]. It estimates along a tree and, given the ancestor sequences, the probability of observing C_{ij} correlated events at positions i and j. We have implemented both exact and approximate approaches and checked their concordance. For computational purposes, we use the approximate computation.

2.3 Phylogenetic reconstruction

The ancestor sequences were reconstructed by using a parsimony method, using the program PAUP [10]. The structure of the tree being constrained (the simulated tree), the ancestor sequences were reconstructed either (1) by considering all possible assignments of the internal ancestral states or (2) by choosing the accelerated transformation option (ACCTRAN) which allows us to assign an unambiguous, even though arbitrary, state of character to each internal state. In some cases, the tree reconstruction was also considered. The heuristic searches of the most parsimonious trees were performed by a simple addition of sequences and a tree-bisection-reconnection branch-swapping algorithm.

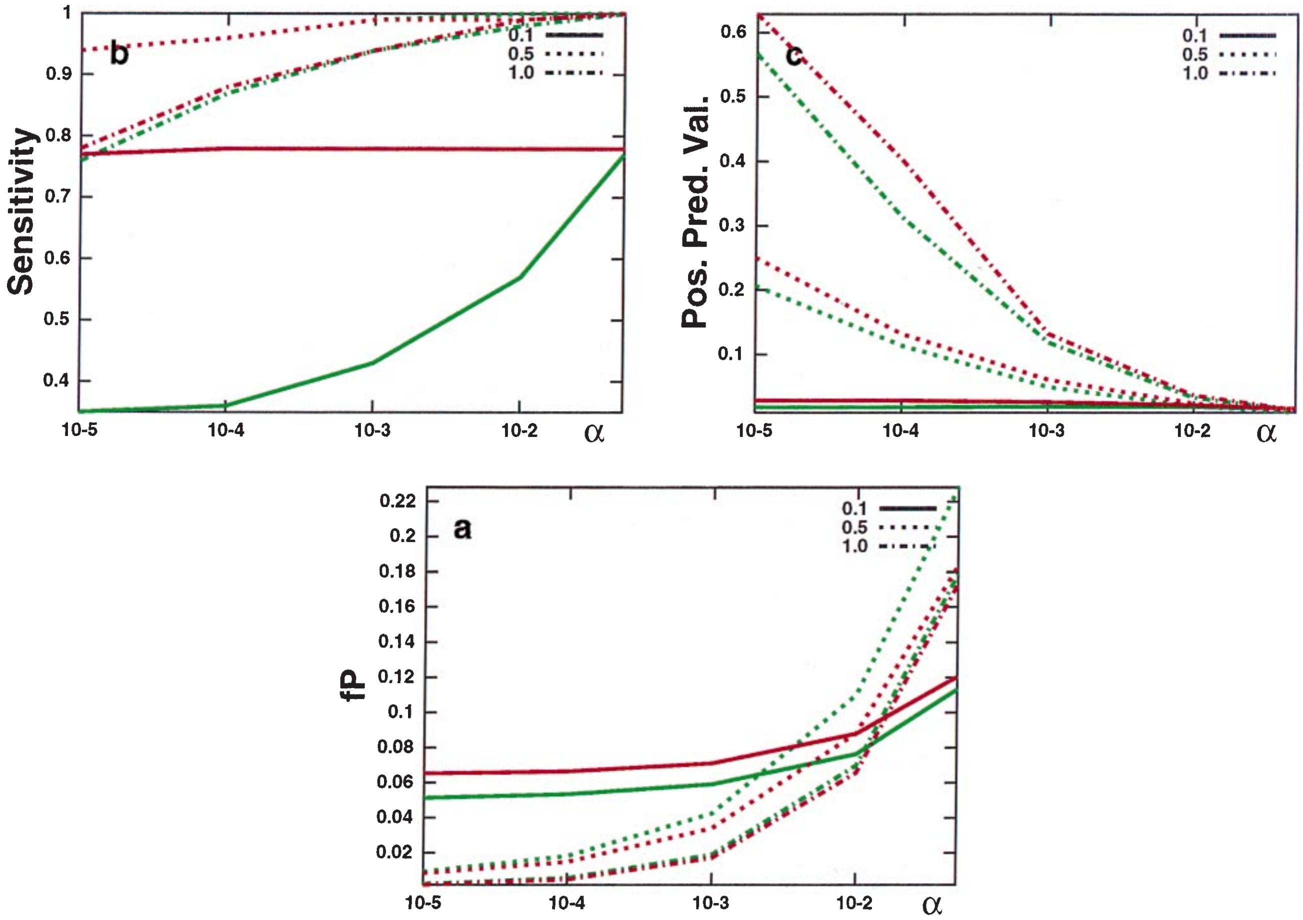

Fig. 1a–c. Results obtained with the MIC (*green*) and G94 (*red*) criteria, for 32 sequences of 300 residues. The sequences include 100 correlated pairs. The different tree scalings (0.1, 0.5, 1.0) are related to different taxa sequence identities (see methods). **a** Fraction of false positives (fP) (over the number of pairs tested) as a function of the type one error α. **b** Sensitivity of the detection of correlated pairs as a function of the type one error. **c** Positive predictive value of G94 and MIC as a function of the type one error

3 Results

3.1 Ignoring the phylogenetic information

Figure 1 shows the results obtained from the aligned taxa sequences with MIC and G94, for the well balanced tree, 32 taxa and sequence lengths of 300 residues, 100 pairs of residues undergoing correlated evolution. Figure 1a shows the fraction of false positive pairs that have been detected as correlated upon the total number of pairs tested, when no correlation was introduced. While this fraction is rather small, it corresponds in fact to a large number of false positives: for sequences of length 300, we test $300 \times 299/2 = 44850$ pairs. Thus even 10% corresponds to 4480 false positives detected. Furthermore, even for very small values of the type one error level (down to 10^{-5}), a number of zero false positives could never be reached. Figure 1b shows the sensitivity of the detection (fraction of simulated correlated pairs that have been detected) as a function of type one error. For a scaling of 0.1, both criteria obviously fail to detect correctly the correlated pairs, which can be related to the very large identity between the sequences. For other scalings, good detection is obtained for a level close to the standard 5% level. Reducing this value is systematically associated with a worse sensitivity. Also, the best sensitivity is obtained for the intermediate scaling of 0.5. This suggests that the criteria are best suited for medium range sequence identity. Figure 1c shows the positive predictive value (fraction of the truly correlated pairs detected upon the number of pairs detected) as a function of the type one error. Best ratios are obtained for smaller values of the type one error, as a consequence of the diminution of the number of false positives (Fig. 1a). In all the cases, the values remain very far from 1. When modifying the number of the correlated pairs, or the tree shape, similar results where obtained (not shown): the number of false positives remains very large. Thus, using such an approach, it seems that it is impossible to reach simultaneously a good sensitivity and a positive predictive value.

3.2 Considering the phylogenetic information

3.2.1 Applying S94 to the exact ancestor sequences

Figure 2 shows the corresponding results obtained with S94, and considering the ancestor sequences as generated. As can be seen in Fig. 2a, for a type one error (less than 10^{-3}) no false positives are detected for all

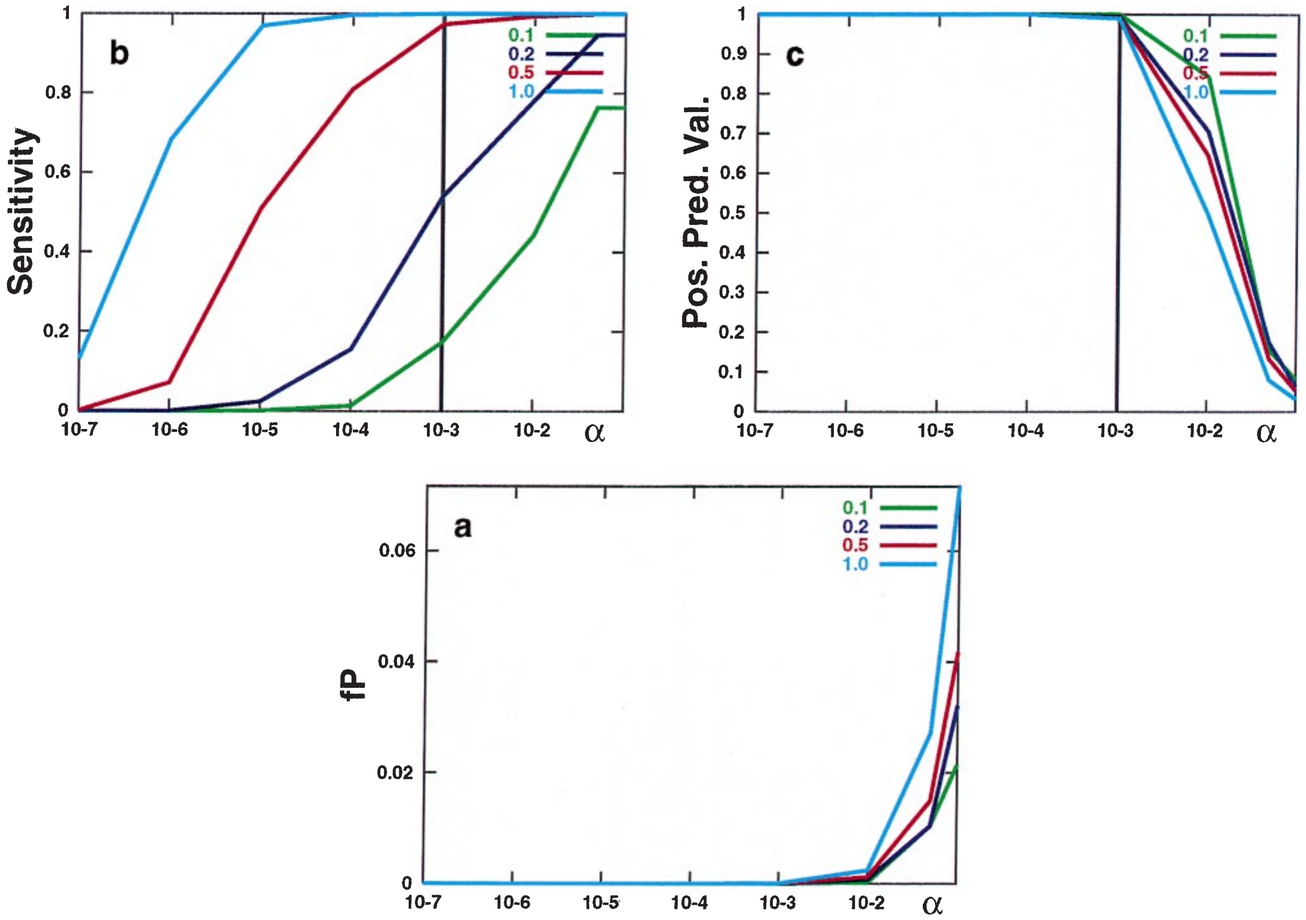

Fig. 2a–c. Results obtained with the S94 criterion, for 32 sequences of 300 residues. The sequences include 100 correlated pairs. The different tree scalings (0.1, 0.2, 0.5, 1.0) are related to different taxa sequence identities (see methods). **a** Fraction of false positives (fP) (over the number of pairs tested) as a function of the type one error α. **b** Sensitivity of the detection of correlated pairs as a function of the type one error. **c** Positive predictive value of S94 as a function of the type one error. For **b** and **c**, the *vertical black line* corresponds to the value of the type one error for which no false positives were detected (see **a**)

mutation rates. As for MIC and G94, for low scalings (0.1, 0.2), a sensitivity (Fig. 2b) of 1 is not reached. For higher mutation rates (scaling 0.5, 1.0) one sees that the sensitivity reaches 1.

This can be explained by the fact that when the taxa sequences are too identical, too few mutations occur, which prevents a satisfactory detection. In terms of sequence identity, a correct detection can only be achieved for identities less than 50% . Interestingly, for these two scalings, it is possible for a level of 10^{-3} to have both a sensitivity of 1 and no false positives, and thus to reach a perfect detection. Even for values of identity close to that obtained for maximal sequence divergence (25–30%), the approach still succeeds in detecting the correlated events. This suggests a good efficiency for S94.

These results are only slightly affected in more unfavourable conditions, and the decrease of the quality of the detection comes systematically from a worse sensitivity; the number of false positives remains always weak for a level less than 10^{-3}. Figure 3 shows the variation of the sensitivity associated with different tree (Fig. 3a), different sequence length (Fig. 3b), different number of taxa (Fig. 3c), and a different number of correlated pairs (Fig. 3d). Only the variation of the number ot taxa and of the shape of the tree introduce some modification compared to Fig. 2b. Concerning the shape of the tree, this decrease was expected for the maximally imbalanced tree since S94 does not take into account the difference between the lengths of the branches. The sensitivity remains rather good, however, since values of 1 are reached for the largest scalings. Concerning the number of taxa, our results suggest that too small a number of taxa cannot lead to a correct detection of correlated pairs. This might be due to the fact that the number of branches becomes too small to allow a sufficient power to S94; the scaling was adapted to preserve the identity between the taxa sequences comparable with that obtained with the 32 taxa tree.

3.2.2 Applying S94 to the reconstructed ancestor sequences

In a real situation, neither the phylogenetic tree nor the ancestor sequences are known. The previous results thus correspond to the best performance for the prediction. We have investigated the effect of reconstructing the ancestor sequences along the simulated tree. As shown by Fig. 4a, the fraction of sites for which one obtains

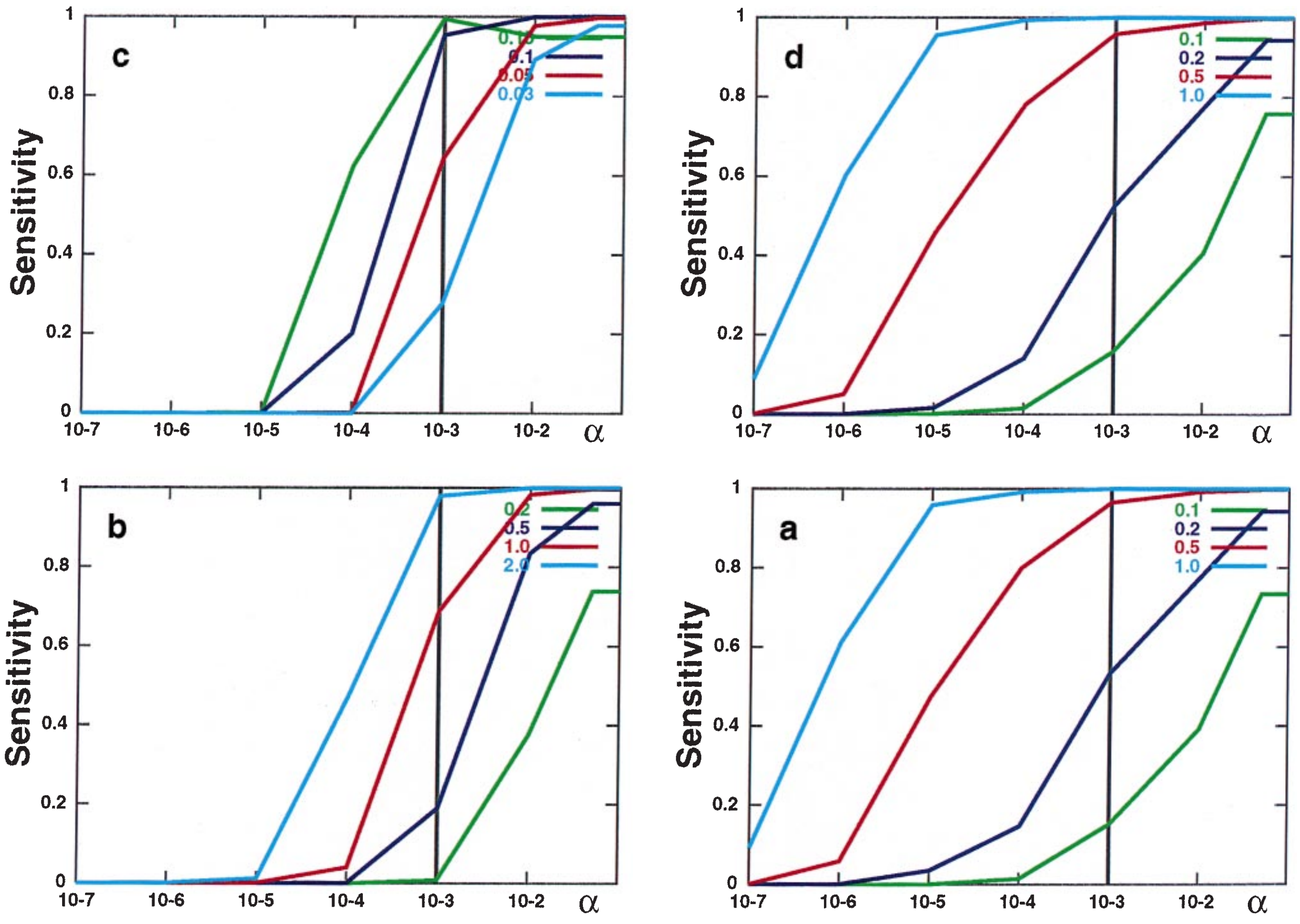

Fig. 3a–d. Results obtained with the S94 criterion. The sequences include 100 correlated pairs, except for **a**. The different different tree scalings (0.1, 0.2, 0.5, 1.0) are related to different taxa sequence identities (see methods). The *vertical black line* corresponds to the value of the type one error for which no false positives were detected (see Fig. 2a). **a** Sensitivity obtained for a well balanced tree with 32 tips, sequence length of 300 bases and 50 correlated pairs. **b** Sensitivity obtained for a well balanced tree with 16 tips, and sequence lengths of 300 bases. **c** Sensitivity obtained for a maximally imbalanced tree with 32 tips. **d** Sensitivity obtained for a well balanced tree with 32 tips, and sequence lengths of 1000 bases

non-unique ancestor reconstruction varies between 10% for a scaling of 0.1 (sequence identity close to 0.8–0.9) and 85% for low identities. Figure 4b and c shows the results obtained when applying a strict rule of not considering the sites for which there is ambiguous ancestor state reconstruction. Figure 4b shows the quality of the prediction, when considering that only the correlations occurring between sites exhibiting no ambiguity in their ancestral state reconstruction could be detected. One notes a worse sensitivity compared to that of Fig. 3b. A possible explanation is that the sites for which the ancestor reconstruction is ambiguous are also the sites that are most informative for the analysis. Figure 4c shows the sensitivity obtained if one considers that the whole set of correlated pairs, including those removed from the analysis, could be detected. Owing to the scaling effect illustrated in Fig. 4a, this overall sensitivity is very low. Finally, we have considered the possibility to force (using the ACCTRAN heuristic) the reconstruction to assign a unique ancestor sequence, thus accepting possible errors in the reconstruction of the ancestor sequences. Surprisingly, we observe (not shown) that the number of false positives does not increase much, and that a level of 10^{-3} remains the limit below which no false positives are observed. Interestingly, as shown in Fig. 4d, the partly incorrect ancestor sequence reconstruction leads to a better sensitivity compared to that of Fig. 4c.

4 Discussion and conclusions

In the present study, we have investigated how possible is the detection of the correlated pairs using different criteria. First, analyses were performed in conditions that are optimal: there is no uncertainty concerning the alignment of the taxa sequences, and the correlation between the sites is perfect (sites mutate together with no exceptions). Concerning methods based upon the analysis of the aligned taxa sequences alone, even in the best conditions we could not obtain a satisfactory detection of the correlated pairs. In most cases the detection of the truly correlated pairs was not perfect, and, above all, the number of pairs detected as correlated while they are not is very large. These results are consistent with those obtained by Pollock and Taylor [6] by simulating amino

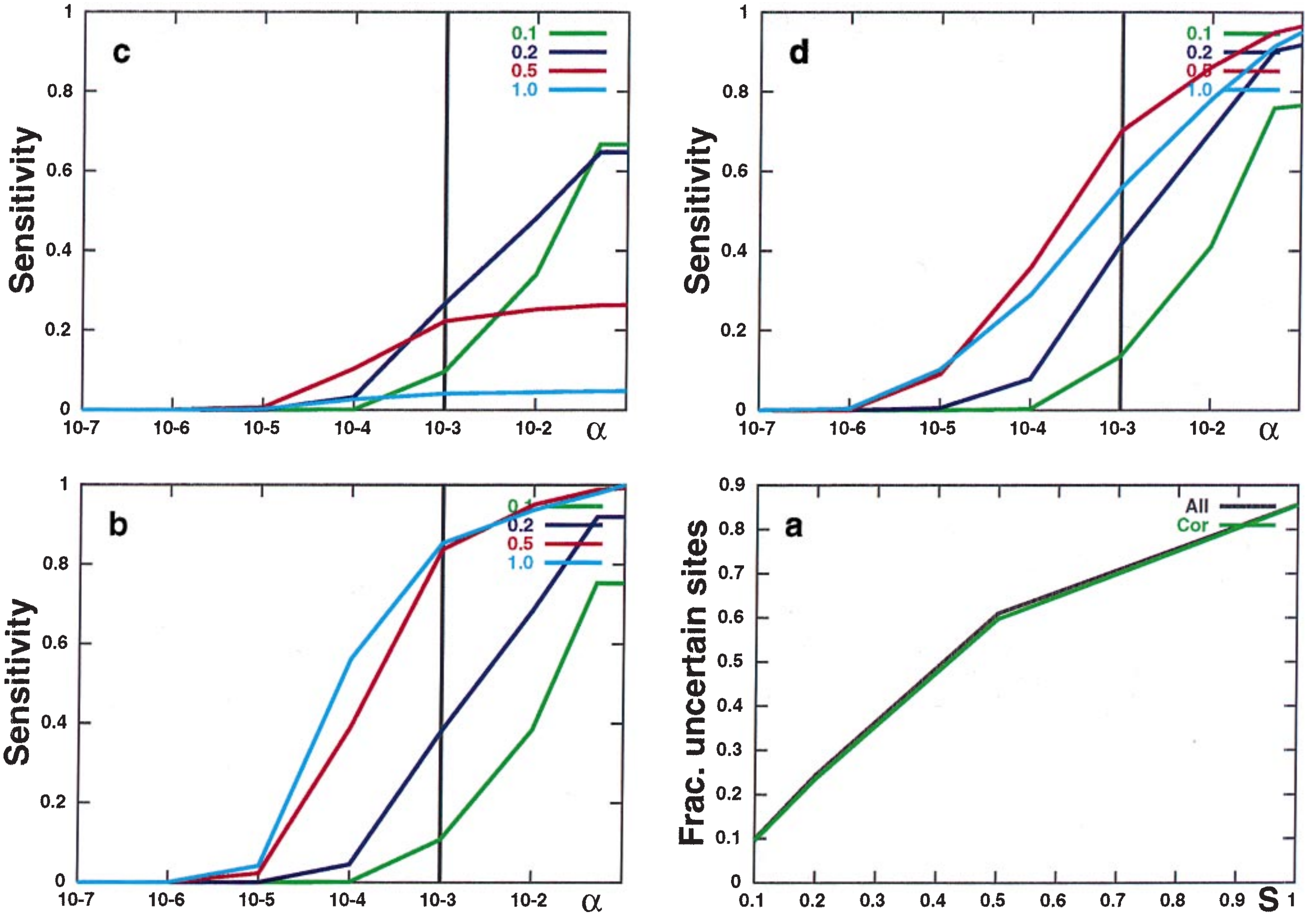

Fig. 4a–d. Effect of the reconstruction of the ancestor sequences. The reconstruction was performed for the simulated tree, using a parsimony method. **a** Fraction of the sites for which there is ambiguity of ancestral state, for all sites (*All*) or for only the correlated sites (*Cor*). **b** Sensitivity obtained when applying a strict rule of not considering the sites for which there is ambiguous ancestor state reconstruction, and when considering that only the correlations occurring between sites exhibiting no ambiguity in their ancestral states could be detected. The different tree scalings (0.1, 0.2, 0.5, 1.0) are related to different taxa sequence identities (see methods). **c** Sensitivity obtained when applying a strict rule of not considering the sites for which there is ambiguous ancestor state reconstruction, and when considering that the whole set of correlated pairs, including those removed from the analysis, could be detected. **d** Sensitivity obtained for a reconstruction of the ancestral states using the ACCTRAN heuristic

acid sequences. Thus, productive use of such approaches implies subsequent information to filter the results obtained. Olmea and Valencia [11] have proposed a method to predict residue contact that combines correlated mutations and other sources of sequence information.

As expected, the approach based upon the analysis of the phylogenetic reconstruction gives significantly better results. For some cases where we had all the exact information concerning the phylogenetic reconstruction, this approach was able to detect almost perfectly the correlated pairs. Under other conditions, the results get worse. However, the method appears particularly robust concerning the detection of false positives: as a general rule, only a few false positives were observed for type one error levels smaller than 10^{-3}. Thus, choosing this level, one can expect to detect rather well the truly correlated pairs. Shindyalov and coworkers [5] used in their study a threshold of 5% , which could be too permissive, even if one takes into account that DNA sequences were simulated here. One major drawback of the criterion comes from its sensitivity to the quality of the reconstruction. For sequence identity rates compatible with common rates observed within a family (less than 50%), the parsimony reconstruction of the ancestor sequences could not assign a unique ancestral state for more than 50% of the sites. This affects the detection dramatically: if one accepts performing the analysis from the sites where no ambiguity exists, the rate of detection of the correlated sites falls by less than 0.2 in the best cases. This bad score can be enhanced up to 0.6, however, by using a heuristic that forces the unique ancestral state. Such a value can be expected to be optimistic since, in a real study, one should accept that some uncertainty will come not only from the tree construction but also from the alignment of the taxa sequences. In this study, the parsimony reconstruction of the tree led to a unique tree for the well balanced tree, and for scaling values as large as 0.5 (corresponding to taxa sequence identity around 50%). For lower identity (scaling 1), the simulated tree was found in only 50% of the cases, and in 10% of the cases a wrong tree was unambiguously identified. Thus, the best results are obtained for a scaling of 0.5 for which a sensitivity of 60% was reached using the ACCTRAN heuristic.

Several perspectives are offered to improve this score: focusing on the reconstruction of the ancestor states, one could imagine enhancing the S94 criterion to weight differently the true non-ambiguous ancestral states. Also, the detection could be considered by starting from the probabilistic approaches of the reconstruction of the ancestor sequences. Pagel [12] has proposed such an approach for sequences of two-state characters. For nucleic acids, a four-state model is required, and it is unclear how the larger number of parameters to be estimated could be managed.

References

1. Altschuh D, Lesk AM, Bloomer AC, Klug A (1987) J Mol Biol 193:693
2. Neher E (1994) Proc Natl Acad Sci USA 91:98
3. Gobel U, Sander C, Schneider R, Valencia A (1994) Proteins 18:309
4. Taylor WR, Hatrick K (1994) Protein Eng 7:341
5. Shindyalov N, Kolchanov NA, Sander C (1994) Protein Eng 7:349
6. Pollock DD, Taylor WR (1997) Protein Eng 10:647
7. Rambaut A, Grassly N (1997) Cabios 13:235
8. Hasegawa M, Kishino H, Yano T (1985) J Mol Evol 22:160
9. Yang Z (1994) J Mol Evol 39:105
10. Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, version 3.1. Formerly distributed by Illinois Natural History Survey, Champaign, Ill
11. Olmea O, Valencia A (1997) Folding Des 2:S25
12. Pagel M (1994) Proc R Soc Lond B 255:37