Topological Estimation Biases with Covarion Evolution

Huai-Chun Wang[1,2], Edward Susko[1], Andrew J. Roger[2] and Matthew Spencer[3]

[1] *Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia*

[2] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4H7, Canada*

[3] *School of Biological Sciences, University of Liverpool, Liverpool, UK*

Corresponding Author: Edward Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; Phone: (902) 494-8865; Fax: (902) 494-5130; E-mail: susko@mathstat.dal.ca

Abstract Covarion processes allow changes in evolutionary rates at sites along the branches of a phylogenetic tree. Covarion-like evolution is increasingly recognized as an important mode of protein evolution. Several recent reports suggest that maximum likelihood estimation employing covarion models may support different optimal topologies than estimation using standard rates-across-sites (RAS) models. However, it remains to be demonstrated that ignoring covarion evolution will generally result in topological misestimation. In this study we performed analytical and theoretical studies of limiting distances under the covarion model and four-taxon tree simulations to investigate the extent to which the covarion process impacts on phylogenetic estimation. In particular, we assessed the limits of a RAS model-based maximum likelihood method to recover the phylogenies when the sequence data were simulated under the covarion processes. We find that, when ignored, covarion processes can induce systematic errors in phylogeny reconstruction. Surprisingly, when sequences are evolved under a covarion process but an RAS model is used for estimation, we find that a long branch repel bias occurs.

Abbreviations: LBA, long branch attaction; LBR, long branch repel, also called long

branch repulsion; RAS, rates variation across sites; JC, the uniform rate of amino acid substitutions, equivelent to the model of Jukes and Cantor (1969) for nucleotide substitutions; the TS model, the covarion model of Tuffley and Steel (1998).

Key words: phylogeny estimation, maximum likelihood, simulation, bias, covarion, heterotachy, inconsistency, protein evolution

Introduction

Phylogenetic inference with the maximum likelihood methods relies on the correct specification of the molecular evolutionary processes. A variety of models of nucleotide and amino acid substitutions have been proposed over the years to specify the processes, including, for instance, equal rate across sites and lineages, rate variation across sites (RAS or the Γ model), proportion of invariable sites, and rate variation across lineages (such as covarions and heterotachy) (Felsenstein 2004). Model misspecification is often cited as one of the primary causes for incorrect topology estimation by the maximum likelihood methods (Gaut and Lewis 1995; Bruno and Halpern 1999).

The covarion process holds that selective pressures on an amino acid or nucleotide site are dependent on the identity of other sites that change throughout time, resulting in changes of evolutionary rates of sites along the branches of a phylogenetic tree (Fitch and Markowitz, 1970). These changes can be gradual or discrete covarion shifts (Inagaki et al., 2004). For instance, a functional change in a molecule after a lineage has split in two can result in rates of evolution for many sites in the descendents of one lineage being much faster than the other. Covarion-like evolution is increasingly recognized as an important mode of molecular evolution in structural RNA genes and protein coding genes (Miyamoto and Fitch 1995; Simon et al. 1996; Lockhart et al. 2000; Galtier 2001; Huelsenbeck 2002; Lopez et al., 2002; Misof et al. 2002; Pupko and Galtier 2002; Ané et al. 2005; Wang et al., 1997). The covarion models differ from models that have been previously shown to result in inconsistent topological estimation (Susko et al., 2004; Kolaczkowski and Thornton, 2004) in that they are stationary processes throughout the tree. Reports have been accumulated that maximum likelihood estimation with covarion models may support different optimal topology than using a model without covarions (e.g. the RAS model) (Ruiz-Trillo et al., 2004; Shalchian-Tabrizi et al., 2006; Wang et al., 2007). However, it is not clear that whether failing to account for covarion-like evolution will generally result in topological miss-estimation. In this study we performed analytical and theoretical studies of limiting distances under the Tuffley-Steel covarion

model (Tuffley and Steel, 1998) to investigate the extent to which the covarion process impacts on phylogenetic estimation. We then did four-taxon tree simulations to assess the limits of a RAS model-based maximum likelihood method to recover the phylogenies when the sequence data were simulated under the Tuffley-Steel model and a general covarion model (Wang et al., 2007), respectively. We compared the effects of different sequence lengths and amino acid substitution matrices on the simulation results.

Analytical Results

We start by considering results in an idealized four-taxon setting: amino acid Jukes-Cantor distances (Jukes and Cantos 1969) being used with neighbour-joining for tree estimation. While the setting is simplified to make analysis more tractable, we expect and, indeed, will show that similar behaviour arises with more complicated substitution processes and maximum likelihood. What we will show here is that distances that are uncorrected for rates-across-sites variation will cause a long-branch-attraction bias while distances that are corrected for rates-across-sites will show a long-branch-repels bias.

Our development is similar to Susko et al. (2004). With or without a gamma rates-across-sites adjustment, Jukes-Cantor distances between a pair of taxa, i and j, are a continuous function of the proportion of sites with different amino acids in the sequences, $\hat{p}^{(i,j)}$; we denote the distance as $d_{ij} = d(\hat{p}^{(i,j)})$. Since the proportion of sites with different amino acids converges to the probability of different amino acids for i and j at a site, $p^{(i,j)}$, we have that

$$d(\hat{p}^{(i,j)}) \to d(p^{(i,j)})$$

as the number of sites goes to infinity. It will be valuable to think of incorrectly specified distances in terms of their dependence on the true evolutionary distances. The probability of different amino acids at a site is dependent upon the pair, i and j, only through the true evolutionary distance, t, between the pair: $p^{(i,j)} = w(t)$. Thus the limiting

distance, $d(p^{(i,j)})$ is also a function $g(t) := d(w(t))$ of the true evolutionary distance between the pair.

In the case of a four taxon tree with taxa A, B, C and D there are three topologies which can be described in terms of the neighbour of A: (A, B), (A, C) and (A, D). We assume throughout that the true topology is (A, B). With $b > a$ in Figure 1, the true tree will either have long branches apart: Figure 1A, which we will call the *abaab* tree, or the true tree will have long branches together: Figure 1B, which we will call the *abaab* tree. For a four taxon tree the neighbour joining algorithm can be shown (Saitou and Nei 1987) to determine the estimated topology according to the following rules:

1. (A, B) is preferred to (A, D) if

$$d_{AD} + d_{BC} - d_{CD} - d_{AB} > 0 \tag{1}$$

2. (A, B) is preferred to (A, C) if

$$d_{AC} + d_{BD} - d_{CD} - d_{AB} > 0 \tag{2}$$

The limiting behaviour differs depending on whether the limiting distances, $g(t)$, are concave functions of the true distances, t or not. We start by considering the case that $g(t)$ is concave and the generating tree has long branches apart: the *abaab* tree. As discussed in Susko et al. (2004), with a large number of sites, (1) will for sure be satisfied so that the estimated topology will be (A, B) or (A, C). Let $b(a)$ be the solution of

$$g(2b + a) - 2g(a + b) + g(3a) = 0 \tag{3}$$

Then, as discussed in Susko et al. (2004), with a large number of sites, (2) will be satisfied, if and only if $b < b(a)$. In other words, for $b > b(a)$, the tree with long branches together, (A, C), will be estimated.

Consider now a generating Jukes-Cantor amino acid model combined with the covarion model (the TS model) described in Tuffley and Steel (1998). The TS model assumes a Markov process for rate switching along the edges of a phylogenetic tree. Rates along an edge switch from off to on and from on to off. When a site is off, no substitutions occur and when it is on, substitutions occur at a constant rate. The model has two additional parameters: s_{10} and s_{01}, the rate of transition from the off state to the on state and then corresponding rate from on to off.

Figure 2A gives the estimated distances plotted against the true distances for various choices of s_{01} and s_{10} when Jukes-Cantor distances are used that make no adjustment for the TS model or even rates-across-sites variation. The concave shapes of the plots indicate that it is a long-branch-attraction form of inconsistency that will arise. The zones of inconsistency are given in Figure 2B. Values of b and a above the lines correspond to regions where the topology with long branches together, the (A, C) topology, will be estimated. In some respects, the results are not surprising. Tuffley and Steel (1998) show that for a pair of taxa, the TS model is indistinguishable from a rates-across-sites model and it is well know that a failure to adjust for rates-across-sites variation leads to long-branch-attraction (Huelsenbeck 1995; Kuhner and Felsentstein 1995). For $s_{10} = 0.001$ and $s_{01} = 0.1$ the limiting estimated distances are almost linear functions of the true evolutionary distances giving an almost non-existent zone of inconsistency. This too is not surprising when one considers that the proportion of time that the covariance process is on, $s_{01}/(s_{01} + s_{10})$, is almost 1; at any given site it is likely there is little or no rate variation.

The case where the limiting distances are convex is different. As discussed in Susko et al. (2004), in this case with the *aaabb* generating tree that has long branches together, the trees with long branches apart, (A, C) and (A, D), will be estimated with probability approximately 1/2 each in the zone of inconsistency. This zone corresponds to values of

$b > b(a)$ where $b(a)$ is the solution of

$$2g(b + 2a) - g(2a) - g(2b) = 0 \tag{4}$$

Figure 3A plots the limiting distances against estimated distances for a number of different choices of the α parameter in gamma corrected rates-across-sites distances. The other parameters in the generating model were set to values estimated in real data (the HSP90 data considered in Wang et al. 2007): $s_{01} = 0.61$ and $s_{10} = 0.53$. With α small we see highly convex $g(t)$. In this case, with a generating *aaabb* tree that has long branches together, the trees with long branches apart will be estimated. The zones of inconsistency are plotted in Figure 3B, which shows large zones when α is small. As α gets larger, the relationship between estimated and true distance gets closer to linear with a corresponding small zone of inconsistency and, as expected given the results illustrated in Figure 2, with no rates-across-sites variation ($\alpha \to \infty$), the relationship between estimated and true distance is concave with no corresponding zone of inconsistency for the *aaabb* generating tree that has long branches together.

The situation in usual practice is more complex. Maximum likelihood estimation is frequently used in practice rather than distances methods, empirical substitution models like the JTT model (Jones et al., 1992) are used and the value of α is estimated. Still, the results here suggest that in the common setting where a rates-across-sites adjustment is made, the likely consequence is a long-branch-repels form of bias. In the following sections we investigate this through simulation.

Simulations

The two types (*abaab* and *aaabb*) of four-taxon trees were used to simulate protein sequence data. The edge lengths a and b varies from 0.05, 0.3, 0.6, 0.9, ..., 2.4, 2.7, 3.0. There are total 121 trees for the combinations of the a and b settings. 100 simulations were performed for each setting. Seq-gen-aminocov, a sequence simulator adapted from

seq-gen (Ané et al, 2005; Rambaut and Grassly, 1997), were used to simulate amino acid data under the given models (source code available from http://www.liv.ac.uk/ matts/). The simulated sequence lengths include 100, 459, 1000, 10000 and 100000 amino acids for different simulation experiments. The amino acid substitution rates include the uniform rates (Jukes-Cantos amino acid exchange matrix) and the JTT rates. Three site-rate variation models were used in this study: the RAS model (Yang, 1994), the TS model and the general covarion model (Wang et al., 2007). For the TS model we simulated the sequences with $s_{01} = 0.61$ and $s_{10} = 0.53$, which correspond to the equilibrium frequency of *on* sites (π) = 0.53 and switching speed per substitution (ν) = 0.57 in the seq-gen-aminocov parameterization. The general covarion model (Wang et al., 2007) combines the RAS models with both the TS and Galtier models, allowing evolutionary rates of sequence sites not only to switch from *on* to off and off to *on*, as in the Tuffley - Steel and Huelsenbeck models (Huelsenbeck, 2002), but also to switch among different *on* states, as in the Galtier model (Galtier, 2001). The general model has three more parameters than the TS model, including the rate of switching from *on* to *on* (s_{11}), the proportion of the covarion process (π) (1 - π is the proportion of non-covarion RAS process) and the Γ shape parameter (α) for the RAS process, in addition to s_{01} and s_{10}.

Topology estimations were conducted with PAML (Yang, 1997) under the RAS model and a uniform rate or with Puzzle (Schmidt et al., 2002) under the RAS model and JTT rates. Heatmaps, plotted with a R script, were used to show the distribution of the estimated optimal topologies for different a and b settings.

The effect of the Tuffley-Steel covarion model on phylogenetic inference

1. The simulating tree is of *abaab*-type

We simulated five data sets of different lengths under the TS model and amino acid Jukes and Cantos model (JC). The covarion parameters for the simulations were $s_{01} = 0.61$, $s_{10} = 0.53$. The estimations were conducted with PAML under the RAS model (with 4 Γ rate categories and allowing α to be optimized), the JC rate and allowing edge

length to be optimized. In this setting, the only "misspecification" of the model is the TS versus RAS process while the amino acid rate matrix (a uniform rate) - is constant.

Table 1 summarizes the results of the simulations for the *abaab*-type tree and 5 sequence length settings: 100, 459, 1000, 10000 and 100000 amino acids.

The results indicate that as the sequence lengths are very short (100 amino acids) both LBA and LBR biases are obvious. As the length increases, the numbers of both LBA and LBR decreases, until virtually become zero when the lengths are 10000 or 100000 amino acids. This suggests while ignoring the simplest covarion process has little effect on phylogenetic inference with the RAS model for a long sequence alignment, it could be problematic if the aligned sequences are very short.

Figure 4 shows two panels of heatmaps that represent respectively the proportions of the misestimated optimal topologies with regard to different settings of the edge lengths (a and b) for sequence length of 100 and 459 amino acids, respectively. The top two maps are the proportions of the AC and AD trees combined and the bottom two maps are the proportions of the AC trees only. For the sequence length of 100 amino acids, both maps indicate that the wrongly assigned AC and AD trees are distributed for all range of a (especially $a = 0.05$) and for $b > 0.05$. For length = 459 amino acids, misestimated optimal AC and AD trees are only present in $a = 0.05$ and b greater than 0.6. For length = 1000 amino acids (figure not shown), misestimated optimal AC and AD trees are only present in $a = 0.05$ and b greater than 0.9. For length = 10000 and 100000 amino acids, misestimated optimal AD and AC trees are only present in one simulation ($a = 0.05$ and $b = 3.0$) out of a total of 12100 simulations.

The estimation of the tree topologies allowed the optimization of the Γ shape parameter (α). Figure 5 shows a heatmap of the estimated α values averaged for each cell of AB trees for sequence length of 1000 amino acids. This indicate α is small when both a (< 1.0) and b (< 1.5) are small. For same a or b, α increases with b or a. Heatmaps of α for the other sequence lengths (not shown) show similar distribution of α but the average values are smaller as the sequence lengths increase.

2. The simulating tree is of *aaabb*-type

Table 2 shows the results for simulations under the TS + JC models and estimated under RAS + JC models. The simulation and estimation conditions are as shown in Table 1, with the only exception that the simulating trees were of the *aaabb*-type.

As described above, both AC and AD trees represent the LBR bias for simulations under the *aaabb*-type trees, which is supported by the comparable numbers of AC and AD trees within the same sequence length settings (Table 2). Table 2 also shows that short sequence length (100 amino acids) brings about more LBR bias. As the length increases the LBR bias slowly decreases and at the length of 100000 amino acids, the total number of the misestimated trees is still over 1000.

Figure 6 shows heatmaps for the proportions of the misestimated optimal AC + AD trees with regard to a and b for the simulations under different sequence lengths. This indicates that for length = 100 amino acids wrongly assigned AC and AD trees are distributed mainly in the region composed of a less than 1.5 and b greater than a. For length = 459 amino acids, the distribution of AC + AD trees is more restricted to the upper left corner. The shrinking distribution of the AC + AD trees continues for the simulated datasets of the even longer sequences. For instance in the case of sequence lengths = 100000 amino acids, the AC and AD trees are limited in the setting of a = 0.05 and 0.3 and b greater than 0.9.

For the simulated sequences of length = 1000 or 459 amino acids, the estimated average α for the estimated optimal AB trees are relatively small (α less than 2.0) when a less than 1.0, but b can be upto 3.0. For short sequences (sequence = 100 amino acids), α less than 2.0 only occurs when $a < 0.6$ and $b < 1.5$. But this is rather inconsistent, as some a and b settings within this range can cause α greater than 5.

The above simulations examine bias in phylogenetic inference for sequence data simulated under the TS model + the JC rate and estimated under RAS + the JC rate and demonstrated that the outcomes depend on the types of the simulating trees. If the data are simulated under the *abaab*-type trees both the long branch attraction and

long branch repel biases will be reduced with increasing sequence lengths and the biases essentially disappear when the sequence length reaches 10000 amino acids, suggesting a TS model-based covarions can be handled with a RAS model for long sequences. However, if the data are simulated under the *aaabb*-type trees both AC and AD trees represent long branch repel bias and they significantly persist even the sequence length reaches 100000 amino acids, suggesting that the covarions will likely always cause LBR bias in phylogenetic inference if the RAS method is used for estimation.

Comparing the general covarion and RAS models

Having looking at the effect of the TS model on phylogenetic inference with a RAS method, we want to see the effect of a more general covarion model. Four simulation experiments were conducted: Simulations I and II used the *abaab*-type trees for simulations. Simulations III and IV used the *aaabb*-type trees for simulations. In all simulations sequence lengths were kept at 459 amino acids.

I: data were simulated under the RAS model with $\alpha = 0.8$, JTT + 4 Γ rates. For each tree, edge lengths a and b vary from 0.05 to 3.0 and 100 data sets were simulated for each setting. II: data were simulated under the general covarion model. In addition to using the above parameters for the RAS model, covarion parameters also included: proportion of the covarion process $(\pi) = 0.71$, $s_{01} = 0.43$, $s_{10} = 0.57$, $s_{11} = 0.97$. These parameter settings were based on the optimized result for a HSP90 data set that was previously used for testing the covarion models (Wang et al., 2007). III: The sequences were generated under the RAS model. The simulation conditions are same as Simulation I except that the *aaabb* trees were used. IV: The sequences were generated under the general covarion model. Simulation conditions are same as Simulation II except that the *aaabb* trees were used.

Puzzle was then used to estimate the topologies and compute the maximum likelihoods for the datasets, withJTT + 4 Γ rates and allowing α and edge length optimization. Figure 7 shows 8 heatmaps that represent respectively the proportions of the

estimated optimal topologies with regard to the distribution of a and b for the four data sets, for each of which two heatmaps are presented: one is the proportion of the AC + AD trees and the other is the proportion of the AC trees only. Since the estimated wrong trees (i.e. the AC and AD trees) are restricted in the region made of $[b >= a, a <= 1.0]$, we computed average frequencies of AB, AC and AD trees among cells in this region and their standard errors (Table 3). For the frequency data in the defined region, the Monte Carlo standard error of the proportion of the AB trees is the square root of the sum of the proportion times (1 - the proportion) divided by the number of the AB trees over all a and b settings in the region. The standard errors of the proportions of the AC and AD trees were calculated in the same way.

This result indicates the proportions of AC and AD are not significantly different for the RAS model simulated data (data set I), but for data set II, simulated under the general covarion model, the proportion of AD is significantly greater than that of AC. This suggests that the RAS model generates approximately equal amount of trees of LBA and LBR biases, while the covarion model generates more LBR than LBA. For data sets III and IV, both the AC and AD trees represent LBR. Both the proportions of AC and AD trees are significantly increased in data set IV (simulated under the general model) than in data set III (simulated under the RAS). In summary for these simulation settings the covarion model significantly increases LBR bias compared with the RAS model.

Furthermore, we simulated sequence data under the TS + JTT models and estimated under the RAS + JTT models for both *abaab* and *aaabb* trees. The results (not shown) again demonstrate that sequence data simulated under the covarion model and estimated under the RAS model cause more LBR bias.

Discussion

Four-taxon simulated datasets have been widely used to produce controlled simulation of evolutionary processes and evaluate the success rate of different methods for

recovering phylogenetic trees (Felsenstein 2004; Huelsenbeck and Hillis 1993; Gaut and Lewis 1995; Chang 1996; Bruno and Halpern 1999; Susko et al. 2004; Kolaczkowski and Thornton 2004; Spencer et al. 2005). This has successfully identified model misspecification and paticular edge length setting that may cause LBA (Felsenstein zone) and LBR biases (Farris zone). The previous simulation studies usually focused on model misspecification in the RAS model and different substitution rates. In this study we investigated the effect of covarion process on a conventional phylogenetic method (i.e., the RAS) and we see that depending on the simulating tree, the bias could be LBA or LBR. For sequence data simulated under the abaab-type tree and the TS model, both LBA and LBR bias will appear if a RAS model is used for phylogenetic estimation. However, increasing sequence length will effectively reduces both biases. For sequence data simulated under the aaabb-type tree and estimated with a RAS method, the LBR bias will persist even very long sequences are used. The analytical results also show that LBR is the main bias for the aaabb-simulating trees, which present a limiting distance of a convex function of the true distance. Therefore, although the covarion process could cause LBA (Lockhart and Steel 2005) when using a RAS model to estimate, a LBR bias is much of a concern and it is not removed even large data sets are employed.

References

Ané, C. and Burleigh, J.G. and McMahon, M.M. and Sanderson, M.J. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. 22: 914-924

Bruno, W. J., and A. L. Halpern. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564-566

Chang J.T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci. 134:189-215

Felsenstein, J. 2004. Inferring phylogenies. Sinauer, MA.

Fitch, W.M. and Markowitz, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4: 479-593

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18: 866-873

Gaut, B.S. and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12: 152-162

Huelsenbeck, J.P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48

Inagaki, Y., E. Susko, N. M. Fast and A. J. Roger. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF1-alpha phylogenies. Mol. Biol. Evol. 21: 1340-1349

Jones DT, Taylor WR and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275-282

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21-123 in H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kolaczkowski, B. and J.W. Thornton. 2004. Performance of maximum parsomony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984.

Kuhner, M.K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468

Lockhart, P. and M. Steel. 2005. A tale of two processes. Syst. Biol. 54:948-951

Lopez, P., Casane, D. and Philippe, H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol.19: 1-7

Misof, B., Anderson, C.L., Buckley, T.R., Erpenbeck, D., Rickert, A. and Misof, K. 2002. An empirical analysis of mt 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected. J. Mol. Evol. 55: 460-469

Miyamoto, M.M. and Fitch, W. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12: 503-513

Pupko, T. and Galtier, N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. Biol. Sci. 269: 1313-1316

Rambaut, A. and Grassly, N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic tree. Comput. Appl. Biosci. 13: 235-238

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing evolutionary trees. Mol. Biol. Evol. 4:406–425

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504

Simon, C., Nigro, L., Sullivan, J., Holsinger, K., Martin, A., Grapputo, A., Franke A. and McIntosh, C. 1996. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. Mol. Biol. Evol. 13:923-932

Spencer, M. and E. Susko and A. J. Roger. 2005. Likelihood, Parsimony, and Heterogeneous Evolution. Mol. Biol. Evol. 1161-1164

Susko, E., Inagaki, Y. and Roger A.J. 2004. On inconsistency of the neighbour joining method and least squares estimation when distances are incorrectly specified. Mol. Biol. Evol. 29:1629–1642

Tuffley, C. and Steel, M. A. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63-91

Wang H-C., Spencer, M., Susko, E. and A. J. Roger 2007. Testing for Covarion-like Evolution in Protein Sequences. Mol. Biol. Evol. 24:294-305

Yang, Z. 1994. Maximum-likelihood phylogenetic estimation of from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306-311

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 15:555-556

Table 1: Simulations under TS + JC and estimation under RAS + JC for different lengths of the simulated sequences. The simulating trees are of the *abaab*-type.

Sequence length	Simulation model	Estimation model	#AB tree	#AC tree	#AD tree
100	TS + JC	RAS + JC	11125	470	405
459	TS + JC	RAS + JC	11938	32	130
1000	TS + JC	RAS + JC	12047	12	41
10000	TS + JC	RAS + JC	12099	0	1
100000	TS + JC	RAS + JC	12099	1	0

Table 2: Simulations under TS + JC and estimation under RAS + JC for different lengths of the simulated sequences. The simulating trees are of the *aaabb*-type.

Sequence length	Simulation model	Estimation model	#AB tree	#AC tree	#AD tree
100	TS + JC	RAS + JC	10106	1033	961
459	TS + JC	RAS + JC	10892	606	602
1000	TS + JC	RAS + JC	10922	596	582
10000	TS + JC	RAS + JC	11040	570	490
100000	TS + JC	RAS + JC	11094	514	492

Table 3: Proportions ± standard errors of the estimated AB, AC and AD trees in the regions of $[b >= a, a <= 1.0]$ for the 4 simulation experiments indicated in the text.

Data set	Simulation model	Estimation model	AB tree	AC tree	AD tree
I[1]	RAS + JTT	RAS + JTT	0.96 ± 0.003	0.03 ± 0.002	0.02 ± 0.002
II[1]	General + JTT	RAS + JTT	0.95 ± 0.005	0.01 ± 0.002	0.04 ± 0.004
III[2]	RAS + JTT	RAS + JTT	0.76 ± 0.008	0.12 ± 0.007	0.12 ± 0.007
IV[2]	General + JTT	RAS + JTT	0.54 ± 0.008	0.22 ± 0.008	0.24 ± 0.008

[1] The simulating trees is of the *abaab*-type.
[2] The simulating trees is of the *aaabb*-type.

Figure 1: Two types of four-taxon trees for simulation: A. of the *abaab* form; B. of the *aaabb* form.

Figure 2: The relationship between the limiting estimated distances and the true distances is given in Figure 2A. The TS model is the generating model with a Jukes-Cantor amino acid substitution process but estimated distances are Jukes-Cantor distances, uncorrected for both the TS model and rates-across-sites variation are used. Figure 2B gives the zones of inconsistency with an *abaab* generating tree that has long branches apart. All values of a and b above the boundary curves correspond to cases where the tree with long branches together will be estimated with long sequences.

Figure 3: The relationship between the limiting estimated distances and the true distances is given in Figure 3A. The TS model ($s_{01} = 0.61$ and $s_{10} = 0.53$) is the generating model with a Jukes-Cantor amino acid substitution process but estimated distances are Jukes-Cantor distances with a gamma correction for various choices of a fixed α gamma parameter. Figure 3B gives the zones of inconsistency with an *aaabb* generating tree that has long branches together. All values of a and b above the boundary curves correspond to cases where a tree with long branches apart will be estimated with long sequences.

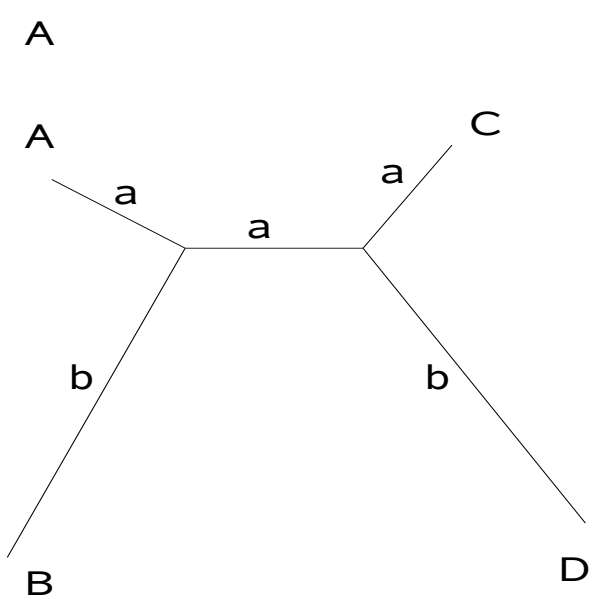
Figure 4: Heatmaps for the distribution of the proportions of estimated AC and AD trees with respect to a and b for simulations under *abaab*-type trees. Top left : proportion of AC + AD trees for length 100 amino acids; Bottom left: proportion of AC trees for length 100 amino acids; Top right: proportion of AC + AD trees for length 459 amino acids; Bottom right: Proportion of AC trees for length 459 amino acids.

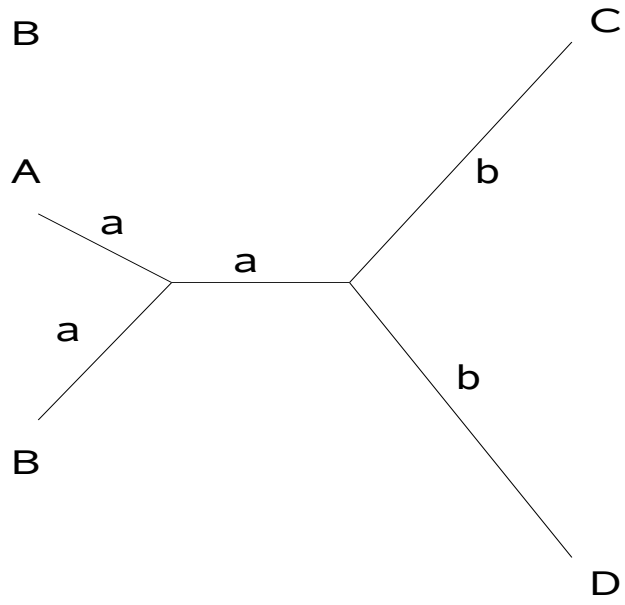
Figure 5: Heatmaps for the distribution of the estimated α with respect to a and b for simulations under *abaab*-type trees and sequence length of 100 amino acids.

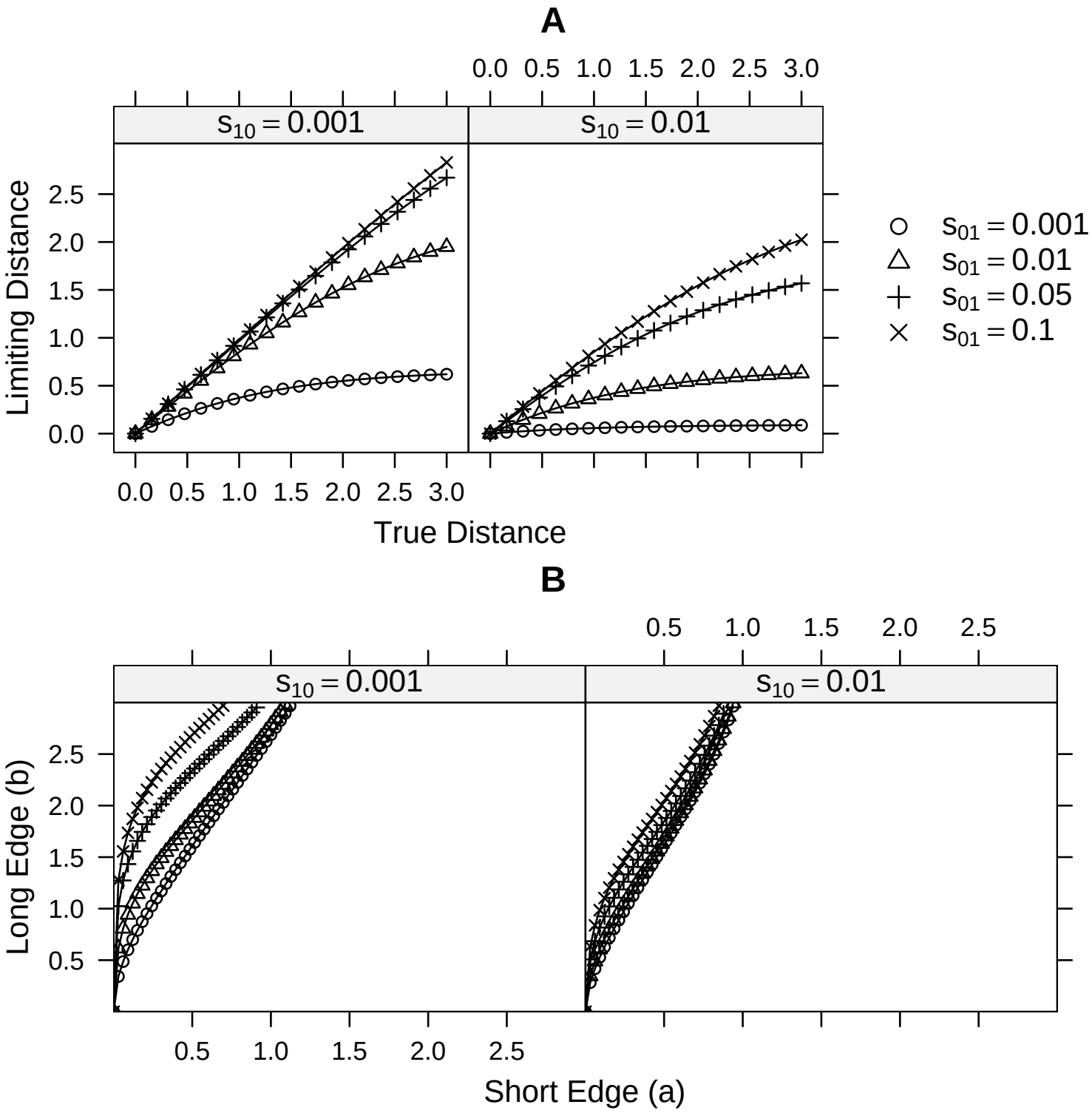
Figure 6: Heatmaps for the distribution of the proportions of estimated AC and AD

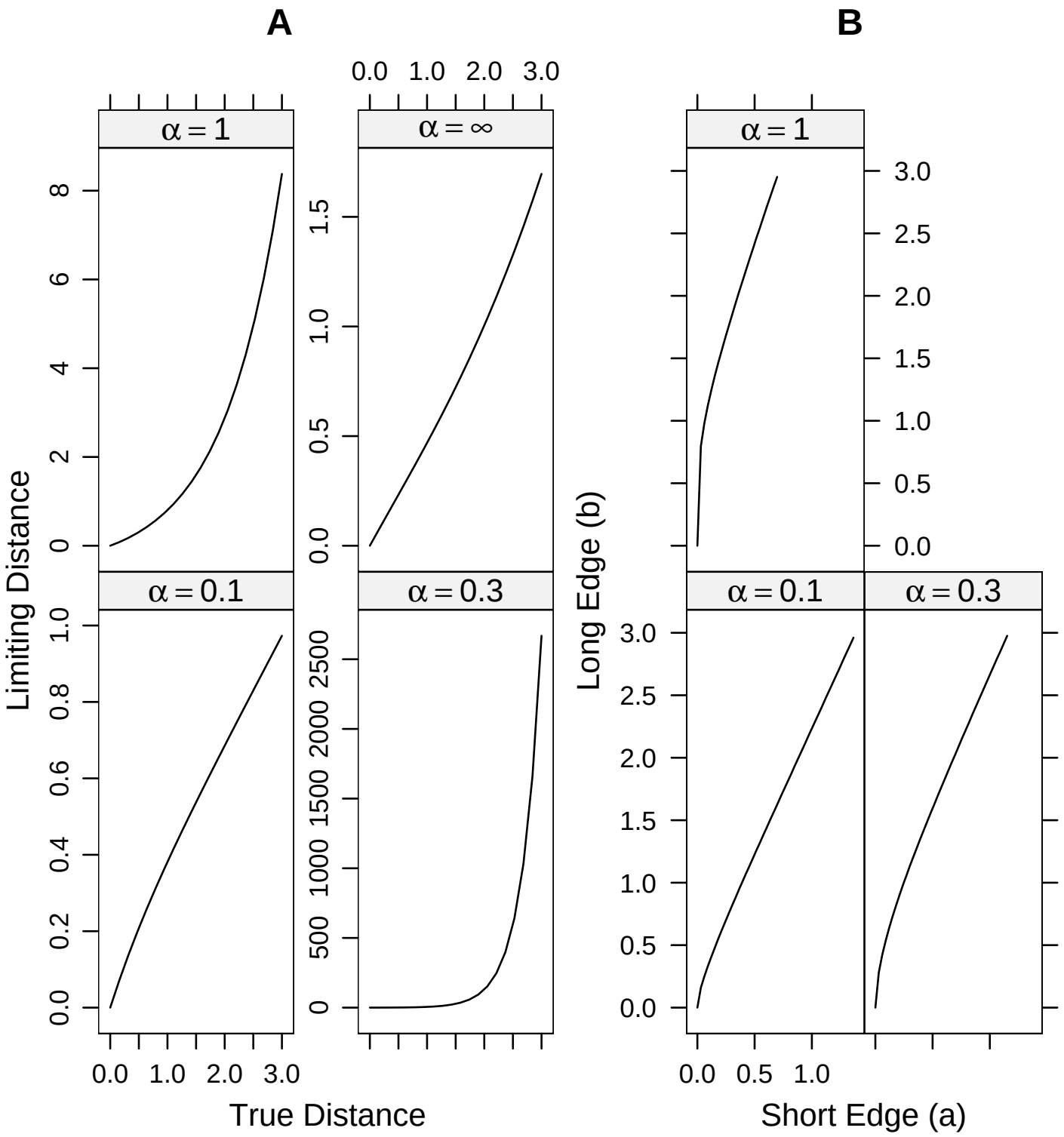
trees with respect to a and b for simulations under *aaabb*-type trees and sequence length being 100, 459, 1000, 10000 and 100000 amino acids, respectively.

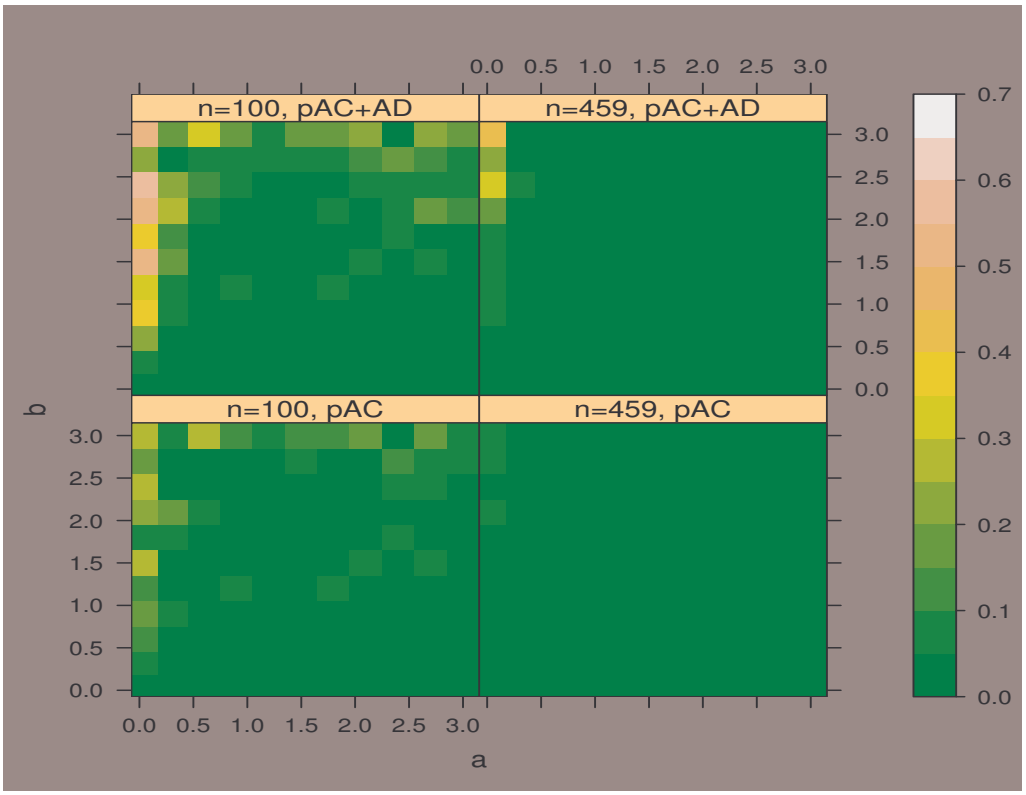
Figure 7: Heatmaps for the distribution of the proportions of estimated AC and AD trees with respect to a and b for simulations under *abaab*-type trees (the bottom two panels) *aaabb*-type trees (the top two panels). For each horizontal panel, the left map shows the data were simulated under the RAS + JTT models and the right map shows the data simulated under the general covarion + JTT models. The sequence length for all simulations was maintained at 459 amino acids and the α for the RAS and covarion models was 0.8. Four Γ rate categories were used.

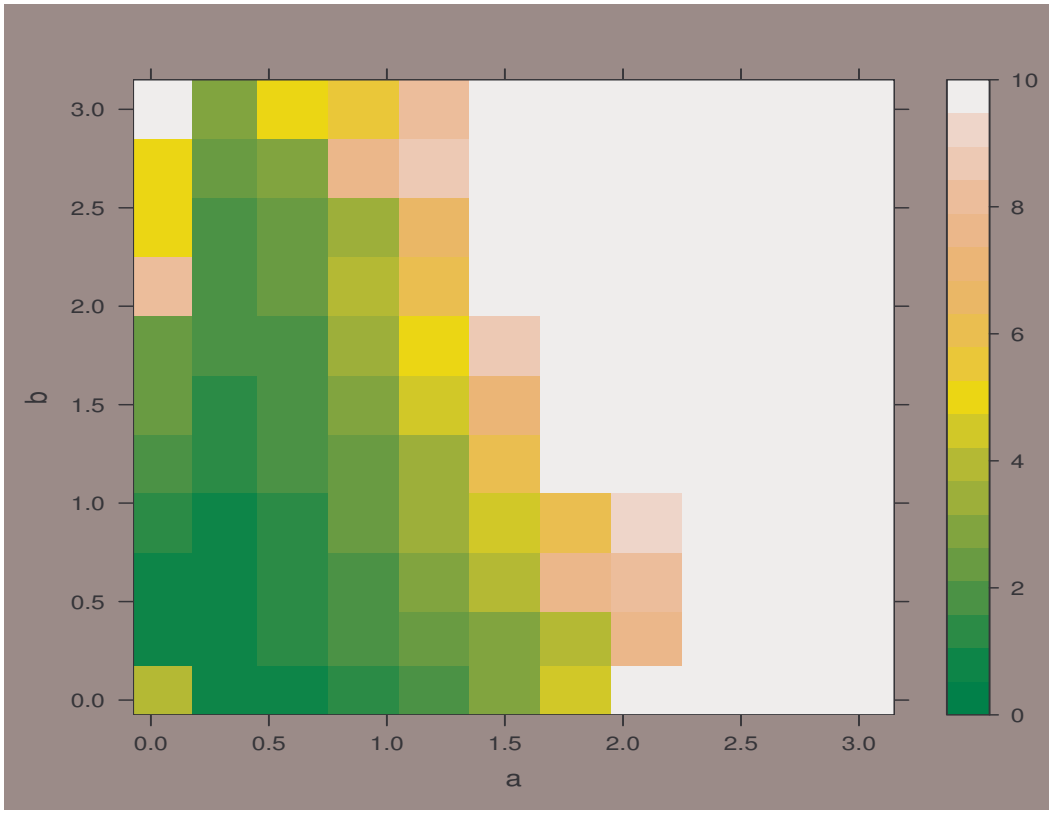

A
A
C
a
a
a
b
b
B
D

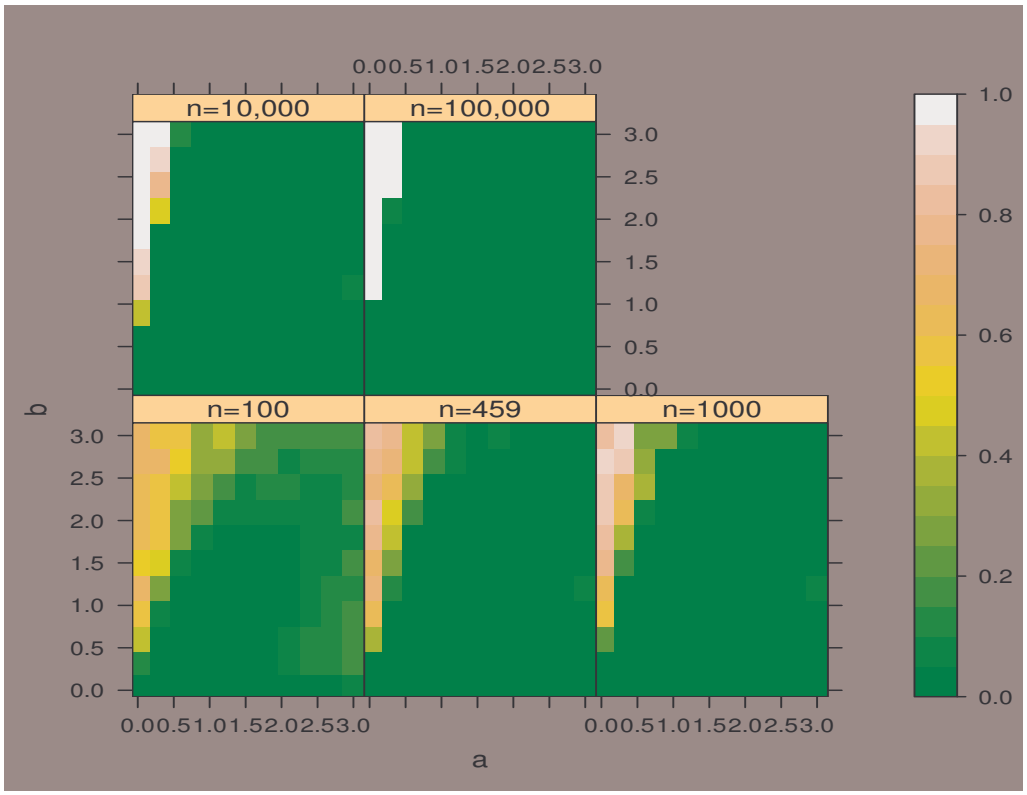

B
C
A
b
a
a
a
b
B
D

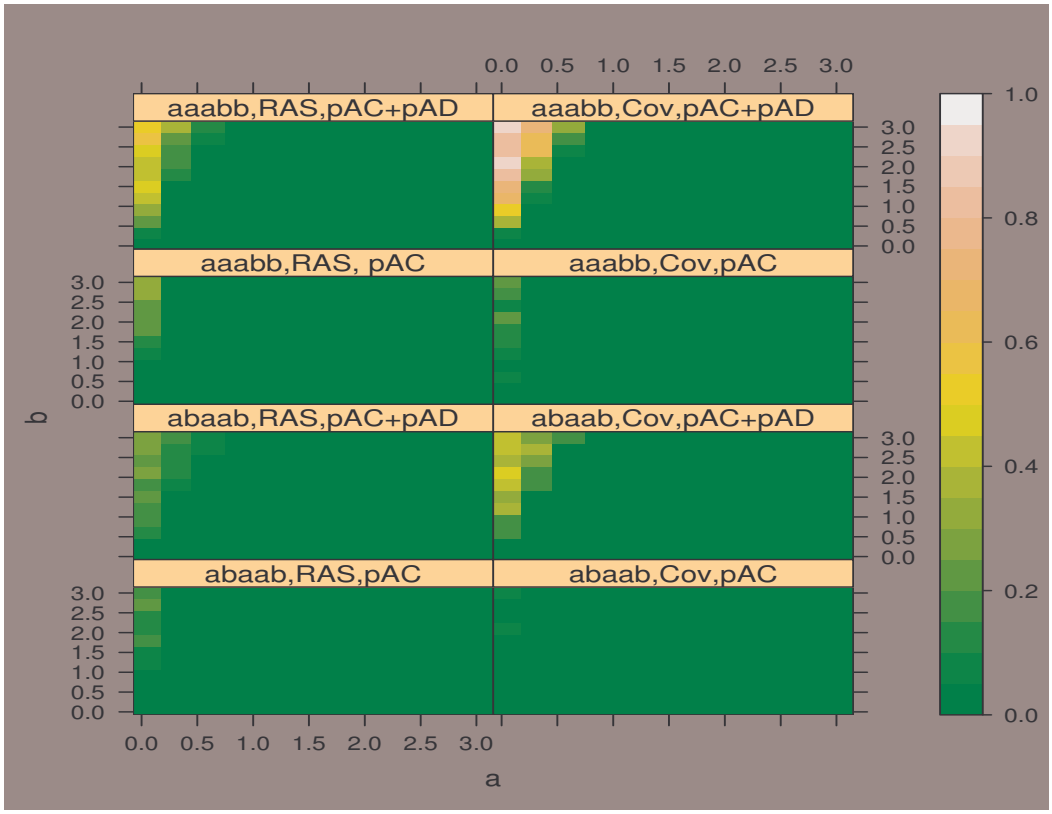

A
$s_{10} = 0.001$
$s_{10} = 0.01$
Limiting Distance
True Distance
$s_{01} = 0.001$
$s_{01} = 0.01$
$s_{01} = 0.05$
$s_{01} = 0.1$
B
$s_{10} = 0.001$
$s_{10} = 0.01$
Long Edge (b)
Short Edge (a)


A
B
α = 1
α = ∞
α = 0.1
α = 0.3
Limiting Distance
True Distance
Long Edge (b)
Short Edge (a)


n=100, pAC+AD
n=459, pAC+AD
n=100, pAC
n=459, pAC
a
b
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


b
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
a
10
8
6
4
2
0


n=10,000
n=100,000
n=100
n=459
n=1000
a
b
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8


aaabb,RAS,pAC+pAD
aaabb,Cov,pAC+pAD
aaabb,RAS, pAC
aaabb,Cov,pAC
abaab,RAS,pAC+pAD
abaab,Cov,pAC+pAD
abaab,RAS,pAC
abaab,Cov,pAC
0.0 0.5 1.0 1.5 2.0 2.5 3.0
a
b
1.0 0.8 0.6 0.4 0.2 0.0