ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Molecular Phylogenetics and Evolution 26 (2003) 1–7

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

An index of substitution saturation and its application

Xuhua Xia,[a,d,*] Zheng Xie,[b] Marco Salemi,[c] Lu Chen,[d] and Yong Wang[d]

[a] *Department of Biology, University of Ottawa, Ottawa, Ont., Canada*
[b] *Institute of Environmental Protection, Hunan University, Changsha, China*
[c] *Rega Institute for Medical Research, K. U. Leuven, Leuven B-3000, Belgium*
[d] *Department of Microbiology, University of Hong Kong, Hong Kong*

Received 17 October 2001; received in revised form 24 July 2002

Abstract

We introduce a new index to measure substitution saturation in a set of aligned nucleotide sequences. The index is based on the notion of entropy in information theory. We derive the critical values of the index based on computer simulation with different sequence lengths, different number of OTUs and different topologies. The critical value enables researchers to quickly judge whether a set of aligned sequences is useful in phylogenetics. We illustrate the index by applying it to an analysis of the aligned sequences of the elongation factor-1α gene originally used to resolve the deep phylogeny of major arthropod groups. The method has been implemented in DAMBE.

Keywords: Substitution saturation; Entropy; Topology; Phylogenetic signal; DAMBE

1. Introduction

The reliability of results from molecular phylogenetics of sequence data depends on how well the analysis deals with the following five problems, aside from the quality of sequencing. The first is the reliability of sequence alignment, i.e., the correct identification of homology. The second is whether substitution rates vary substantially over sites, which has been demonstrated to result in wrong trees recovered from sequence data (Kuhner and Felsenstein, 1994). The third is whether nucleotide frequencies, or the set of variable sites, change along different lineages, i.e., the problem of nonstationarity (Lake, 1994; Lockhart et al., 1994). The fourth is the well-known problem of long-branch attraction that can be caused by a variety of factors. The last, but not the least important, is whether some or all sequences in the data set have already lost phylogenetic information due to substitution saturation (Lopez et al., 1999; Philippe and Forterre, 1999). The paper deals with the last problem.

Substitution saturation decreases phylogenetic information contained in the sequences, and has plagued the phylogenetic analysis involving deep branches, such as major arthropod groups. In the extreme case when sequences have experienced full substitution saturation, the similarity between the sequences will depend entirely on the similarity in nucleotide frequencies that often does not reflect phylogenetic relationships (Xia, 2000, pp. 49–58). To avoid the problem of substitution saturation, researchers typically would use conservative genes, such as the elongation factor-1α (EF-1α), which is one of the most abundant proteins in eukaryotes (Lenstra et al., 1986) and catalyzes the GTP-dependent bindings of charged tRNAs to the ribosomal acceptor site (Graessmann et al., 1992). Because of its fundamental importance for cell metabolism in eukaryotic cells, the gene coding for the protein is evolutionarily conservative (Walldorf and Hovemann, 1990), and consequently has been used frequently in resolving deep-branching phylogenies (Baldauf et al., 1996; Cho et al., 1995; Friedlander et al., 1998; Lopez et al., 1999; Regier and Shultz, 1997).

Protein genes consist of codons, in which the third codon position is the most variable, and the second the most conservative (Xia, 1998; Xia et al., 1996). The third

* Corresponding author. Fax: +613-562-5486.
E-mail address: xxia@uottawa.ca (X. Xia).

1055-7903/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S1055-7903(02)00326-3

codon position is often not excluded from the analysis, mainly for two reasons. First, excluding the third codon position would often leave us with few substitutions to work on. Second, substitutions at the third codon position likely conform better to the neutral theory of molecular evolution than those at the other two codon positions. Consequently, the former may lead to better phylogenetic estimation than the latter, especially in estimating divergence time (Yang, 1996a). However, these two potential benefits of using substitutions at the third codon position may be entirely offset if the sites have experienced substitution saturation and consequently contain no phylogenetic information.

There are currently four main approaches for finding whether molecular sequences contain phylogenetic information. The first approach involves the randomization or permutation tests (Archie, 1989; Faith, 1991). The second employs the standard g_1 statistic for measuring the skewness of tree lengths of alternative trees (Swofford, 1993). Both approaches suffer from the problem that, as long as we have two closely related species, the tests will lead us to conclude the presence of significant phylogenetic information in the data set even if all the other sequences have experienced full substitution saturation. This problem is also shared by the third approach implemented in the RASA program (Lyons-Weiler et al., 1996). The forth approach (Steel et al., 1995; Steel et al., 1993) has just been implemented in DAMBE (Xia, 2000; Xia and Xie, 2001) with a few extensions. Its main disadvantages are that its computation is clumsy with more than four taxa, that it associates specifically with the parsimony method, that that it has not been developed further after so many years.

Here, we present a new entropy-based index of substitution saturation. Standard statistical tests can be used to test whether a set of molecular sequences has experienced substitution saturation. The index is illustrated by its application to the EF-1α sequences.

2. Materials and methods

2.1. Basic concepts

Suppose N aligned sequences with L nucleotides each. Designate the nucleotide frequencies for all sequences as P_A, P_C, P_G, and P_T. In the extreme case when there is no substitution at all, then the nucleotides at each site will all be identical, with the frequency of one nucleotide being 1 and the frequencies of the other three nucleotides all being zero. In terms of information theory, the entropy at this site i is then

$$H_i = -\left(\sum_{j=1}^{4} p_j \log_2 p_j\right) = 0, \tag{1}$$

where $j = 1, 2, 3$, and 4 corresponding to nucleotide A, C, G, and T, and p_j is the proportion of nucleotide j at site i. Substitutions will lead to polymorphic sites at which the H_i value will be larger than 0. The maximum value of H_i is 2 when nucleotide frequencies at each site are represented equally. The mean and variance of H for all L sites are then simply

$$\bar{H} = \frac{\sum_{i=1}^{L} H_i}{L}, \mathrm{Var}(H) = \frac{\sum_{i=1}^{L} (H_i - \bar{H})^2}{L-1}. \tag{2}$$

When sequences have experienced full substitution saturation, then the expected nucleotide frequencies at each nucleotide site are equal to the global frequencies P_A, P_C, P_G, and P_T, i.e., the pooled nucleotide frequencies from all sequences. The distribution of the nucleotide frequencies at each site then follows the multinomial distribution of $(P_A + P_C + P_G + P_T)^N$, with the expected entropy and its variance expressed as follows:

$$H_{FSS} = -\left(\sum_{N_A=0}^{N}\sum_{N_C=0}^{N}\sum_{N_G=0}^{N}\sum_{N_T=0}^{N}(N!/(N_A!N_C!N_G!N_T!) \times P_A^{N_A}P_C^{N_C}P_G^{N_G}P_T^{N_T}\sum_{j=1}^{4} p_j \log_2 p_j\right), \tag{3}$$

$$\mathrm{Var}(H_{FSS}) = \sum_{N_A=0}^{N}\sum_{N_C=0}^{N}\sum_{N_G=0}^{N}\sum_{N_T=0}^{N}(N!/(N_A!N_C!N_G!N_T!) \times P_A^{N_A}P_C^{N_C}P_G^{N_G}P_T^{N_T}\left(\sum_{j=1}^{4} p_j \log_2 p_j - H_{FSS}\right)^2, \tag{4}$$

where N_A, N_C, N_G, and N_T are smaller or equal to N and subject to the constraint of $N = N_A + N_C + N_G + N_T$, $j = 1, 2, 3$, and 4 corresponding to A, C, G, and T, and $p_j = N_j/N$. These equations are presented only for clarity, not for computation efficiency. The subscript FSS in H_{FSS} stands for full substitution saturation.

Theoretically, the test of substitution saturation can be done by simply testing whether the observed $\bar{H}$ value in Eq. (2) is significantly smaller than H_{FSS}. If $\bar{H}$ is not significantly smaller than H_{FSS}, then the sequences have experienced severe substitution saturation. This test is an extension of the conventional test involving two sequences using the percentage difference (p), i.e., the number of different sites divided by the sequence length, between the two sequences. At full substitution saturation, the expected p value equals $(1 - P_A^2 - P_C - P_G^2 - P_T^2)$, and a test of whether the two sequences have experienced full substitution saturation can be done by testing whether the observed p is significantly smaller than the expected p at full substitution saturation. One might note that, with two sequences, H_{FSS} is also equal to $(1 - P_A^2 - P_C - P_G^2 - P_T^2)$.

Our index of substitution saturation is defined as

$$I_{ss} = \bar{H}/H_{FSS}. \tag{5}$$

We can see intuitively that the sequences must have experienced severe substitution saturation when I_{ss} approaches 1, i.e., when $\bar{H}$ equals H_{FSS}. However, the test of $\bar{H} = H_{FSS}$ is only theoretically useful because the sequences will fail to recover the true phylogeny long before the full substitution saturation is reached, i.e., long before I_{ss} reaches 1. For this reason, we need to find the critical I_{ss} value (referred to hereafter as $I_{ss.c}$) at which the sequences will begin to fail to recover the true tree. Once $I_{ss.c}$ is known for a set of sequences, then we can simply calculate the I_{ss} value from the sequences and compare it against the $I_{ss.c}$. If I_{ss} is not smaller than $I_{ss.c}$, then we can conclude that the sequences have experienced severe substitution saturation and should not be used for phylogenetic reconstruction.

It is difficult to arrive at an analytical formulation of $I_{ss.c}$. However, intuition tells us that it may depend on the topology of the true tree, the number of OTUs (N_{OTU}), the sequence length (SeqLen), nucleotide frequencies and the transition/transversion ratio. We use computer simulation to evaluate the effect of these factors on $I_{ss.c}$.

2.2. Computer simulation

We use the program EVOLVER in the PAML package (Yang, 2000) for the evolutionary simulation. The simulated sequences evolve according to the F84 model, which is the model implemented in the DNAML program in PHYLIP (Felsenstein, 1993). The α/β ratio varied from 1 to 10. The nucleotide frequencies of the four nucleotides varied from 0.1 to 0.9, subject to the constraints that the summation equals 1. It turns out that the effect of the transition/transversion ratio and the nucleotide frequencies on $I_{ss.c}$ is negligible compared to the effect of the other three factors, i.e., topology, N_{OTU}, and SeqLen.

We have used two extreme topologies (Fig. 1), one being perfectly symmetrical and the other extremely asymmetrical. The N_{OTU} values are 4, 8, 12, 16, 20, 24, 28, and 32. When N_{OTU} values are 12, 20, 24, and 28, there is no perfectly symmetrical topology as in Fig. 1a, and multiple quasi-symmetrical topologies were used. For example, when $N_{OTU} = 12$, then we obtain multiple topologies by randomly pruning of a four-OTU symmetrical subtree from the symmetrical 16-OTU topology.

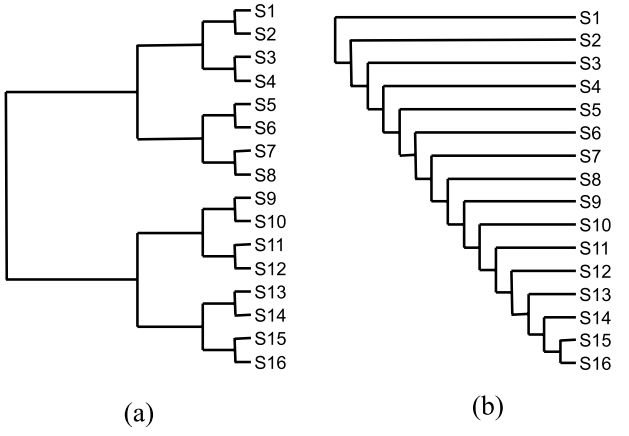

Fig. 1. Two extreme topologies used in simulation: (a) symmetrical; (b) asymmetrical.

We used six different SeqLen values, i.e., 500, 1500, 2500, 3500, 4500, and 5500, for each of the N_{OTU} values and for each of the two extreme topologies. Longer sequences apparently would alleviate the effect of substitution saturation as long as the sequences have not experienced full substitution saturation, i.e., the $I_{ss.c}$ value should be greater with a set of long sequences than with a set of short sequences, everything else being equal.

The tree length varies from 1 to 29 for the symmetrical topology and from 1 to 19 for the asymmetrical topology, with interval of 2 (i.e, tree length = 1, 3, 5,. . ., and so on). For a given topology and N_{OTU}, the longer the tree length, the greater the substitution saturation and the greater the I_{ss} value. Our purpose is to find out at which I_{ss} value the sequences will be too substitutionally saturated to recover the true tree. This particular I_{ss} value is taken as the $I_{ss.c}$ value. By doing a large number of simulations, we can determine $I_{ss.c}$ empirically for a given SeqLen, a given N_{OTU} and a given topology.

We did not use trees with tree length shorter than 1 for the following reason. When the tree length gets smaller towards zero, there will be few substitutions and the true tree will fail to be recovered not due to substitution saturation with too many substitutions, but due to too few substitutions. This study focuses on substitution saturation, but not on the opposite of substitution saturation.

Each topology was fed into EVOLVER and simulated 100 times. Phylogenetic reconstruction is done to find the proportion of trees (P_{true}) with the same topology as the input topology (i.e., the known true topology). The neighbor-joining (NJ) method with the F84 distance, and the maximum likelihood method with the same F84 model as that used in simulation, were used for phylogenetic reconstruction of the simulated sequences when N_{OTU} equals 4, 8, or 16. The two phylogenetic methods yield essentially the same P_{true} values, and only the neighbor-joining method is used for reconstructing topologies with $N_{OTU} > 16$. Phylogenetic reconstruction is all done with DAMBE (Xia and Xie, 2001).

2.3. Illustrative data set

Regier and Shultz (1997) used 21 sequences of the EF-1α gene from major arthropod groups and putative outgroups. Four sequences (U90056, U90060, U90051, and U90061) were excluded from this analysis because of the existence of multiple unresolved nucleotides, e.g., U90061 (a limpet) contains eight unresolved sites. One

sequence (U90064, a polychaete) was excluded because aligning it with others necessitates indels of multiple codons. These sequences were excluded because we do not want to complicate the problem of substitution saturation with sequencing quality and alignment problems. The remaining 16 nucleotide sequences were first translated into amino acid sequences, aligned, and the nucleotide sequences were aligned against aligned amino acid sequences by using DAMBE (Xia and Xie, 2001).

3. Results and discussion

3.1. Simulation studies

The ability of phylogenetic methods in recovering the true tree decreases with the total tree length (i.e., the degree of substitution saturation), but the effect of substitution saturation is alleviated by increasing SeqLen (Fig. 2). The relation between P_{true} and the tree length (TL) is fitted with the following purely descriptive equation

$$P_{\text{true}} = 1 - e^{-e^{B-C\cdot\hbar TL}} \quad (6)$$

for each combination of N_{OTU} and SeqLen. For the symmetrical topology, the fit is almost perfect in all cases, with r^2 values greater than 0.965. We first define the critical tree length (TL_c) as the tree length when $P_{\text{true}} = 0.95$. For example, TL_c is about 4 when $N_{\text{OTU}} = 8$ and SeqLen = 500 (Fig. 2), but equals 7.5 when $N_{\text{OTU}} = 8$ and SeqLen = 3500. I_{ss} increases with TL asymptotically and $I_{\text{ss.c}}$ is defined as the I_{ss} value corresponding to TL_c. For the asymmetrical topology, the fit is worse and there is often no tree length at which the true tree is recovered 100%. However, there is always a tree length at which P_{true} is the largest. When P_{true} decreases to 95% of the maximum P_{true} value, the tree length is taken as TL_c, and the $I_{\text{ss.c}}$ value corresponding to TL_c is then used as the $I_{\text{ss.c}}$ value for the asymmetrical topology.

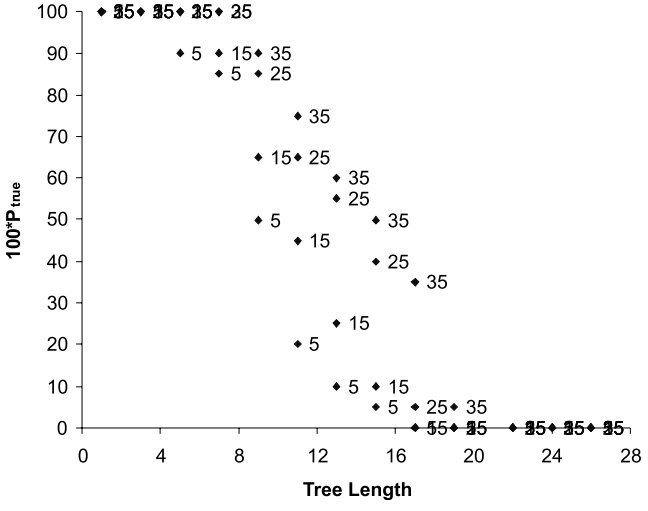

Fig. 2. The proportion of true trees found (P_{true}) depends on the tree length and the sequence length. The data shown are for $N_{\text{OTU}} = 8$. The pattern remains the same with different number of OTUs. Data labels indicate the sequence length in 100, i.e., a number of 5 means 500 bp.

Note that the P_{true} value in Fig. 2 will decrease when the tree length (TL) approaches zero, so that Eq. (6) will no longer be applicable. However, in our simulation, we do not use very small TL values because a very small TL value, implying the rarity of substitution saturation, has nothing to do with substitution saturation. Eq. (6) is applicable for the range of TL values used in our simulation.

The $I_{\text{ss.c}}$ value depends not only on SeqLen, but also on the topology and N_{OTU} in the tree (Fig. 3). These $I_{\text{ss.c}}$ values allow us to judge whether a set of sequences is useful in phylogenetic reconstruction. When an observed I_{ss} value is significantly smaller than $I_{\text{ss.c}}$, we are confident that substitution saturation is not a problem. This will be illustrated later with the elongation factor-1α sequences.

There are three points worth noting in Fig. 3. First, for any given SeqLen, $I_{\text{ss.c}}$ decreases with the increasing N_{OTU}. This point is intuitively obvious and has already been noted before (Ritland and Clegg, 1990).

Second, the decrease of $I_{\text{ss.c}}$ with N_{OTU} is much more severe for the asymmetrical topology (Fig. 3b) than the symmetrical topology (Fig. 3a), i.e., an asymmetrical topology is much more susceptible to substitution saturation than the symmetrical one. Thus, if one suspects that his/her OTUs are likely to be phylogenetically related by an asymmetrical topology, he/she should increase the sequence length. For example, for $N_{\text{OTU}} = 8$ and $I_{\text{ss}} = 0.73$, one can avoid the problem of substitution saturation as long as SeqLen > 500 bp if the true topology is symmetrical (Fig. 3a). However, if the true topology is extremely asymmetrical, one would need SeqLen > 3000 bp to overcome the problem of substitution saturation with the same N_{OTU} and I_{ss} (Fig. 3b). That an asymmetrical topology demands more sequence data to recover is often ignored, and some programs, e.g., RASA (Lyons-Weiler et al., 1996), produce measures of phylogenetic signals that are phylogeny-free. It is clear from Fig. 3 that such phylogeny-free measures of phylogenetic signals cannot be very useful.

Third, $I_{\text{ss.c}}$ values increase with SeqLen (Fig. 3), i.e., increasing SeqLen can alleviate the problem of substitution saturation. However, the increase of $I_{\text{ss.c}}$ with SeqLen soon levels off and one can gain rather little by increasing SeqLen beyond 4000 bp (Fig. 3). For recovering deep phylogenies, one is less wise to use very long but highly variable sequences than to use shorter but highly conservative sequences.

One might have noted that the $I_{\text{ss.c}}$ values in Fig. 3a are a bit too small for trees with $N_{\text{OTU}} = 12, 20, 24$, or

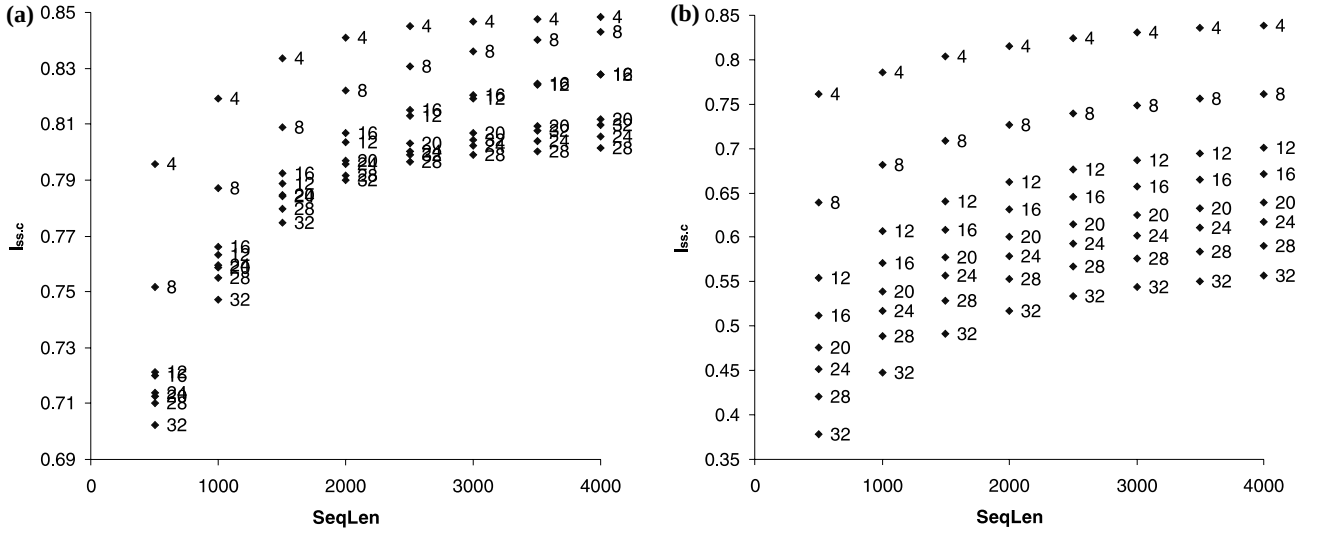

Fig. 3. The critical index of substitution saturation ($I_{ss.c}$) depends on the sequence length (SeqLen) and the number of OTUs (N_{OTU}). Data labels are N_{OTU} values: (a) with a symmetrical topology; (b) with an asymmetrical topology.

28. This is caused by the fact that a tree with one of these N_{OTU} values cannot be perfectly symmetrical, and even a slight deviation from perfect symmetry can decrease the $I_{ss.c}$ value.

3.2. Application of the method to real sequences

For the first, second, and third codon positions of the EF-1α sequences, the I_{ss} values are 0.2093, 0.1115, and 0.6636, respectively. The critical $I_{ss.c}$ value, given $N_{OTU} = 16$ and SeqLen = 350, is 0.7026 if the true tree is symmetrical, and 0.4890 if the true tree is asymmetrical, both being highly significantly greater than the observed I_{ss} values at the first and the second codon positions. Thus, there is little substitution saturation at the first and second codon positions. The resulting phylogenetic trees for the first and for the second codon positions are shown in Fig. 4a and Fig. 4b, respectively.

For the third codon position of the EF-1α sequences, the observed I_{ss} value of 0.6636 is not significantly different from the $I_{ss.c}$ value of 0.7026 for the symmetrical topology ($P = 0.130$, two-tailed t test) and is significantly greater than the $I_{ss.c}$ value of 0.4890 for the asymmetrical topology, suggesting that the third codon position have experienced so much substitution saturation that it is only marginally useful when the true tree is symmetrical and useless if the true tree is asymmetrical. The phylogenetic tree reconstructed with the third codon positions

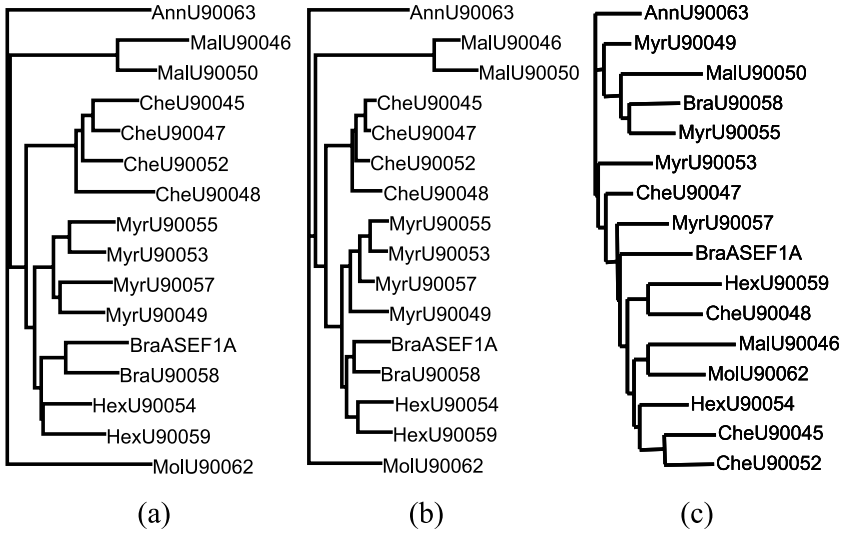

Fig. 4. Phylogenetic reconstruction based on the elongation factor-1α sequences: (a) and (b) based on the first and second codon positions, respectively; (c) based on the third codon positions only.

(Fig. 4c) is absolutely absurd. It is wise of Regier and Shultz (1997) to have used amino acid sequences rather than the nucleotide sequences of the EF-1α sequences for phylogenetic reconstruction.

We now discuss the applicability of our test with more complicated models of substitution, in particular, the rates-across-sites (RAS) model (Uzzel and Corbin, 1971) and the covarion model of substitution (Fitch, 1971; Fitch and Markowitz, 1970). The RAS model has been studied recently in relation to phylogenetic reconstruction (Chang, 1996; Yang, 1996b) because most commonly used phylogenetic methods fail to recover the true tree when sequences evolved according the RAS model (Kuhner and Felsenstein, 1994). A typical RAS model includes a proportion (P) of invariable sites (i.e., sites so strongly constrained by purifying selection that mutations at these sites are lethal), and the rest of sites evolving at the Γ-distributed rates (Yang, 1996b). Our test is applicable to the $(1-P)$ fraction of the sites. DAMBE (Xia, 2000) includes a function for estimating P that can then be inputted into the test of substitution saturation. If one assumes $P=0$ when $P>0$, then our test of substitution saturation would be conservative.

The covarion hypothesis (Fitch, 1971; Fitch and Markowitz, 1970) has been the subject of intensive studies in recent years (Galtier, 2001; Lockhart et al., 1998, 2000; Lopez et al., 1999; Miyamoto and Fitch, 1995; Penny et al., 2001; Pupko and Galtier, 2002). A simulation study (Penny et al., 2001) showed that molecular sequences evolving according to the covarion model is much less susceptible to substitution saturation than evolving according to the RAS model. The reason for this is easy to understand and can be illustrated with a simple example. Suppose, we have an amino acid sequence in which each site will evolve to substitution saturation in 10 million years. Now suppose the amino acid sites can be classified into two equal-sized groups (designated Groups 1 and 2). At time 0, those in Group 1 start evolving while those in Group 2 stay the same. After evolving for five million years, those in Group 1 are frozen in their current states and do not change any more, while those in group 2 start to evolve. At the end of 10 million years, each site would have *effectively* evolved for only five million years and are therefore far from substitution saturation. Although this is not exactly what has been done in the simulation (Penny et al., 2001), the essence is the same. Such a covarion substitution process does not seem to cause any complication in our test of substitution saturation.

Acknowledgments

The study is supported by RGC grants from Hong Kong Research Grant Council (HKU7265/00M, HKU7212/01M) and a grant from University of Ottawa to X.X. M.S. is supported by a research fellowship from the Fonds voor Wetenschappelij K Onderzoek-Vlandereen. We thank the reviewer for providing many helpful comments, suggestions, and references.

References

Archie, J.W., 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38, 219–252.

Baldauf, S.L., Palmer, J.D., Doolittle, W.F., 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA 93, 7749–7754.

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

Cho, S., Mitchell, A., Regier, J.C., Mitter, C., Poole, R.W., Friedlander, T.P., Zhao, S., 1995. A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1α recovers morphology-based tree for heliothine moths. Mol. Biol. Evol. 12, 650–656.

Faith, D.P., 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40, 366–375.

Felsenstein, J., 1993. PHYLIP 3.5 (phylogeny inference package). Department of Genetics, University of Washington.

Fitch, W.M., 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1, 84–96.

Fitch, W.M., 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, 579–593.

Friedlander, T.P., Horst, K.R., Regier, J.C., Mitter, C., Peigler, R.S., Fang, Q.Q., 1998. Two nuclear genes yield concordant relationships within Attacini (Lepidoptera: Saturniidae). Mol. Phylogenet. Evol. 9, 131–140.

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

Graessmann, M., Graessmann, A., Cadavid, E.O., Yokosawa, J., Stocker, A.J., Lara, F.J.S., 1992. Characterization of the elongation factor 1-α gene of *Rhynchosciara americana*. Nucleic Acids Res. 20, 3780.

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

Lake, J.A., 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA 91, 1455–1459.

Lenstra, J.A., Van Vliet, A., Carnberg, A.C., Van Hemert, F.J., Möller, W., 1986. Genes coding for the elongation factor EF-1α in *Artemia*. Eur. J. Biochem. 155, 475–483.

Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., Howe, C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15, 1183–1188.

Lockhart, P.J., Huson, D., Maier, U., Fraunholz, M.J., Van De Peer, Y., Barbrook, A.C., et al., 2000. How molecules evolve in eubacteria. Mol. Biol. Evol. 17, 835–838.

Lopez, P., Forterre, P., Philippe, H., 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49, 496–508.

Lyons-Weiler, J., Hoelzer, G.A., Tausch, R.J., 1996. Relative Apparent Synapomorphy Analysis (RASA) I: the statistical measurement of phylogenetic signal. Mol. Biol. Evol. 13, 749–757.

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

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53, 711–723.

Philippe, H., Forterre, P., 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49, 509–523.

Pupko, T., Galtier, N., 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. R. Soc. Lond. B. Bio 269, 1313–1316.

Regier, J.C., Shultz, J.W., 1997. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. Mol. Biol. Evol. 14, 902–913.

Ritland, K., Clegg, M., 1990. Optimal DNA sequence divergence for testing phylogenetic hypotheses. In: Molecular Evolution. Alan R. Liss, New York, pp. 289–296.

Steel, M., Lockhart, P.J., Penny, D., 1995. A frequency-dependent significance test for parsimony. Mol. Phylogenet. Evol. 4, 64–71.

Steel, M.A., Lockhart, P.J., Penny, D., 1993. Confidence in evolutionary trees from biological sequence data. Nature 364, 440–442.

Swofford, D., 1993. Phylogenetic Analysis Using Parsimony. Illinois Natural History Survey, Champaign, IL.

Uzzel, T., Corbin, K.W., 1971. Fitting discrete probability distributions to evolutionary events. Science 172, 1089–1096.

Walldorf, U., Hovemann, B.T., 1990. *Apis mellifera* cytoplasmic elongation factor 1α (EF-1α) is closely related to *Drosophila melanogaster* EF-1α. FEBS 267, 245–249.

Xia, X., 1998. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol. Biol. Evol. 15, 336–344.

Xia, X., 2000. Data analysis in molecular biology and evolution. Kluwer Academic Publishers, Boston.

Xia, X., Xie, Z., 2001. DAMBE: data analysis in molecular biology and evolution. J. Hered. 92, 371–373.

Xia, X., Hafner, M.S., Sudman, P.D., 1996. On transition bias in mitochondrial genes of pocket gophers. J. Mol. Evol. 43, 32–40.

Yang, Z., 1996a. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42, 587–596.

Yang, Z., 1996b. Among-site rate variation and its impact on phylogenetic analysis. TREE 11, 367–372.

Yang, Z., 2000. Phylogenetic analysis by maximum likelihood (PAML). University College, London.