Vera-Ruiz et al. BMC Bioinformatics 2014, 15(Suppl 2):S8
http://www.biomedcentral.com/1471-2105/15/S2/S8

BMC Bioinformatics

PROCEEDINGS Open Access

Statistical tests to identify appropriate types of nucleotide sequence recoding in molecular phylogenetics

Victor A Vera-Ruiz[1], Kwok W Lau[2,3], John Robinson[1], Lars S Jermiin[2,4*]

From The Twelfth Asia Pacific Bioinformatics Conference (APBC 2014)
Shanghai, China. 17-19 January 2014

Abstract

Background: Under a Markov model of evolution, recoding, or lumping, of the four nucleotides into fewer groups may permit analysis under simpler conditions but may unfortunately yield misleading results unless the evolutionary process of the recoded groups remains Markovian. If a Markov process is lumpable, then the evolutionary process of the recoded groups is Markovian.

Results: We consider stationary, reversible, and homogeneous Markov processes on two taxa and compare three tests for lumpability: one using an *ad hoc* test statistic, which is based on an index that is evaluated using a bootstrap approximation of its distribution; one that is based on a test proposed specifically for Markov chains; and one using a likelihood-ratio test. We show that the likelihood-ratio test is more powerful than the index test, which is more powerful than that based on the Markov chain test statistic. We also show that for stationary processes on binary trees with more than two taxa, the tests can be applied to all pairs. Finally, we show that if the process is lumpable, then estimates obtained under the recoded model agree with estimates obtained under the original model, whereas, if the process is not lumpable, then these estimates can differ substantially. We apply the new likelihood-ratio test for lumpability to two primate data sets, one with a mitochondrial origin and one with a nuclear origin.

Conclusions: Recoding may result in biased phylogenetic estimates because the original evolutionary process is not lumpable. Accordingly, testing for lumpability should be done prior to phylogenetic analysis of recoded data.

Phylogeny Markov model, stationarity, homogeneity,reversibility, recoding, lumping, nucleotides, primates

Introduction

When nucleotides intentionally are recoded to a 3- or 2-state alphabet in order to focus on a subset of the possible types of substitutions (e.g., transversions [1-3]) or reduce compositional heterogeneity [4], it is no longer appropriate to use model-based phylogenetic methods that rely solely on time-reversible, 4-state Markov models. Instead, one needs to use a 3- or 2-state Markov model to approximate the evolutionary processes for the recoded sequence data. This requirement was first realised by Phillips and Penny [5], who used a time-reversible 2-state Markov model [6] to analyse *RY*-recoded nucleotide sequences, and Gibson et al. [7], who developed a time-reversible 3-state Markov model to analyse *Y*-recoded nucleotide sequences. Before these studies, other investigators had used *RY*-recoded nucleotide sequences to infer the evolutionary relationships among mammals [1-3] and among bacteria [4].

Recoding of nucleotides and/or amino acids has been used repeatedly in recent phylogenetic studies [8-31]. However, the mathematical principles underpinning the recoding of nucleotides or amino acids have not yet been adequately examined. For example, it is not yet known whether the Markovian property is maintained after recoding and how this should be tested [32]. Without

* Correspondence: Lars.Jermiin@csiro.au
[4]CSIRO Ecosystem Sciences, Canberra, ACT 2601, Australia
Full list of author information is available at the end of the article

this knowledge, we may run the risk of using a promising procedure in a manner that turns out to be inappropriate for the data.

In this paper, we take a first step by considering tests for lumpability in a Markov model of evolution for pairs of homologous nucleotide sequences (we are aware of only one paper in the phylogenetic literature where the term lumpability is used [33], but there it was used in a different context). We only consider nucleotides but believe our tests could be generalized to encompass amino acids as well. We then illustrate the performance of our tests for lumpability using simulated and real data, and show that recoding of nucleotides should be used with caution when analysing DNA phylogenetically.

Methods

The theoretical basis for recoding nucleotides

Let $\mathcal{S} = \{A, C, G, T\}$ be the set of nucleotides, and let $\mathcal{S}'$ be a partition of $\mathcal{S}$ such that $\mathcal{S}' = \{S_1, ..., S_q\}$, where $q < 4$. Then $\mathcal{S}$ is reduced by grouping, or lumping, some of the original states (i.e., *A*, *C*, *G*, and *T*) into one or two new states (i.e., *R*, *Y*, *S*, *W*, *M*, *K*, *B*, *D*, *H*, and *V*)–in molecular phylogenetics, this procedure has been called *recoding* [34]. Table 1 presents the 13 possible recoding schemes and partitions of $\mathcal{S}'$ using notation established by the NC-IUB [35]. The 13 recoding schemes fall into three major grouping categories, as shown in Table 1.

The evolutionary process for two homologous nucleotide sequences

Consider two nucleotide sequences, A and B, each with *n* independently evolving sites, which have diverged under Markovian conditions from their common ancestor on a rooted, 2-tipped tree. Let π_0 denote the initial probability vector of the nucleotide frequencies, such that $\pi_0^T = (\pi_{0_1}, \pi_{0_2}, \pi_{0_3}, \pi_{0_4})$, where, for convenience of notation, we will use the subscripts 1, 2, 3, 4 to denote *A*, *G*, *C*, *T* . Over each edge of this tree, there is a substitution process, $X(t)$ and $Y(t)$, respectively, described by the transition probabilities

$$P_{ij}^X(t) = P\left[X(t) = j | X(0) = i\right]$$

and

$$P_{ij}^Y(t) = P\left[Y(t) = j | Y(0) = i\right].$$

Let $f_{ij}(t)$ denote the theoretical joint probability of a site being in state i in A and state j in B at time t:

$$f_{ij}(t) = P[X(t) = i, Y(t) = j | X(0) = Y(0)]. \quad (1)$$

Now, let $\mathbf{F}(t) = \{f_{ij}(t)\}$ denote the joint probability matrix and let $\mathbf{P}^X(t)$ and $\mathbf{P}^Y(t)$ denote the transition probability matrices of $X(t)$ and $Y(t)$. In practice, the two matrices cannot be identified from $\mathbf{F}(t)$ without some assumptions about the evolutionary processes of the sequences A and B. We assume that the processes are globally stationary, reversible, and homogeneous (SRH) (for definitions, see [36-39], and, in more detail, [40]). Given these assumptions, take $\mathbf{P}^X(t) = \mathbf{P}^Y(t) = \mathbf{P}(t)$ and, if π_X and π_Y denote the equilibrium probability distributions of the processes $X(t)$ and $Y(t)$, take $\pi_X = \pi_Y = \pi_0 = \pi$ and write $\Pi = diag(\pi)$. Then, from (1), we get $\mathbf{F}(t) = \mathbf{P}(t)^T \Pi \mathbf{P}(t)$. The transition probability matrix can be expressed by an instantaneous rate matrix $\mathbf{R}$, such that $\mathbf{P}(t) = e^{\mathbf{R}t}$, where $R_{ij} \geq 0$ for $i \neq j$, $R_{ii} = -\sum_{i \neq j}^{4} R_{ij}$, and $\pi^T \mathbf{R} = \mathbf{0}^T$, where π^T is the equilibrium distribution of $\mathbf{R}$ [40]. Furthermore, the instantaneous rate matrix can take the form $\mathbf{R} = \mathbf{S}\Pi$, where $\mathbf{S}$ is a symmetric matrix with $s_{ij} \geq 0$ for $i \neq j$, and $s_{ii} = -\sum_{i \neq j}^{4} s_{ij}\pi_j/\pi_i$ [40]. The matrices $\mathbf{R}$ and $\mathbf{P}(t)$ can be written in terms of the

Table 1 The 13 ways of reducing a 4-letter state space ($\mathcal{S}$) to a 3- or 2-letter state space $\mathcal{S}'$.

Nucleotide-grouping Subsets	Recoding notation	Resulting $\mathcal{S}'$	Major grouping category
{{*A*, *G*}, *C*, *T*}	*R*	{*R*, *C*, *T*}	{2 : 1 : 1}
{*A*, *G*, {*C*, *T*}}	*Y*	{*A*, *G*, *Y*}	
{*A*, {*C*, *G*}, *T*}	*S*	{*A*, *S*, *T*}	
{*C*, *G*, {*A*, *T*}}	*W*	{*C*, *G*, *W*}	
{*A*, *C*, {*G*, *T*}}	*M*	{*M*, *G*, *T*}	
{*C*, {*A*, *G*, *T*}}	*K*	{*A*, *C*, *K*}	
{*A*, {*C*, *G*, *T*}}	*B*	{*A*, *B*}	{3 : 1}
{*C*, {*A*, *G*, *T*}}	*D*	{*C*, *D*}	
{*G*, {*A*, *C*, *T*}}	*H*	{*G*, *H*}	
{*T*, {*A*, *C*, *G*}}	*V*	{*T*, *V*}	
{{*A*, *G*}, {*C*, *T*}}	*RY*	{*R*, *Y*}	{2 : 2}
{{*A*, *T*}, {*C*, *G*}}	*SW*	{*S*, *W*}	
{{*A*, *C*}, {*G*, *T*}}	*KM*	{*K*, *M*}	

eigenvector decomposition of $\Pi^{1/2}\mathbf{S}\Pi^{1/2} = \mathbf{L}\Lambda\mathbf{L}^T$. In other words

$$\mathbf{R} = \mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Lambda}\mathbf{L}^T\mathbf{\Pi}^{1/2} \tag{2}$$

and

$$\mathbf{P}(t) = \mathbf{\Pi}^{-1/2}\mathbf{L}e^{\mathbf{\Lambda}t}\mathbf{L}^T\mathbf{\Pi}^{1/2}, \tag{3}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with columns containing the eigenvalues of $\mathbf{\Pi}^{1/2}\mathbf{S}\mathbf{\Pi}^{1/2}$ and $\mathbf{L}$ is a matrix with columns containing its right eigenvectors. The joint probability matrix is then symmetric and

$$\mathbf{F}(t) = \mathbf{\Pi}^{1/2}\mathbf{L}e^{2\mathbf{\Lambda}t}\mathbf{L}^T\mathbf{\Pi}^{1/2}. \tag{4}$$

Note that under these assumptions, there are only nine free parameters to be estimated: six free parameters for the off-diagonal elements of $\mathbf{S}$ (define $s^T = (s_{12}, s_{13}, s_{14}, s_{23}, s_{24}, s_{34})$) and three free parameters for π (because $\pi_4 = 1 - \sum_{i=1}^{3}\pi_i$). The time t can be fixed at 1 since modifying it is equivalent to modifying the s-parameters.

Let $\mathbf{N}$ denote the 4 × 4 divergence matrix for A and B, such that $\mathbf{N} = \{n_{ij}\}$, where n_{ij} represents the number of homologous sites that are in state i in A and state j in B. Under the model, the vector of elements of $\mathbf{N}$ has a multinomial distribution with parameters n and $\mathbf{F}(t)$; its expected value is thus $E(\mathbf{N}) = n\mathbf{F}(t)$. Because the parameters s and π are in a one-to-one relation with the elements of $\mathbf{F}$, the maximum-likelihood estimates of s and π can be obtained from the eigenvector decomposition of $\hat{\mathbf{F}}(t) = \frac{1}{2n}\left(\mathbf{N} + \mathbf{N}^T\right)$, then

$$\hat{\mathbf{\Pi}} = diag(\hat{\mathbf{F}}(1)\mathbf{1}) \tag{5}$$

and

$$\hat{\mathbf{S}} = \hat{\mathbf{\Pi}}^{-1/2}\hat{\mathbf{L}}\hat{\mathbf{\Lambda}}\hat{\mathbf{L}}^T\hat{\mathbf{\Pi}}^{-1/2}, \tag{6}$$

where $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Lambda}}$ are obtained from $\hat{\mathbf{\Pi}}^{-1/2}\hat{\mathbf{F}}(1)\hat{\mathbf{\Pi}}^{-1/2} = \hat{\mathbf{L}}e^{2\hat{\Lambda}}\hat{\mathbf{L}}^T$.

Lumpable Markov chains

The following probabilities can be defined for any given $\mathcal{S}' = \{\mathcal{S}_1, \cdot\ \cdot\ \cdot, \mathcal{S}_q\}$, where q <4, such that a *lumped process*, $X'(t)$, with a smaller number of states, is generated with transition probabilities

$$\begin{aligned} P'_{kl}(t) &= P[X'(t) = l | X'(0) = k] \\ &= P[X(t) \in \mathcal{S}_l | X(0) \in \mathcal{S}_k], \end{aligned} \tag{7}$$

and initial probabilities $\pi' = P[X'(0) = k] = P[X(0) \in \mathcal{S}_k]$.

By definition [41,42], a Markov process is lumpable if, for every starting vector π, the lumped process, defined in (7), is a Markov chain whose transition probabilities do not depend on the choice of π. A necessary and sufficient condition for $X'(t)$ to be lumpable with respect to a partition $\mathcal{S}'$ is that for every pair of subsets, $\mathcal{S}_k$ and $\mathcal{S}_l$, $\sum_{j\in\mathcal{S}_l} P_{ij}(t)$, has the same value for every state i in $\mathcal{S}_k$ [41,42]. Accordingly, if $X'(t)$ is lumpable, then the transition probabilities for $X'(t)$ for any given pair of subsets in $\mathcal{S}'$ are

$$P'_{kl}(t) = \sum_{j\in\mathcal{S}_l} P_{ij}(t), \text{ for any } i \in \mathcal{S}_k.$$

If the Markov chain is lumpable, the lumped transition matrix $\mathbf{P}'(t)$ can be expressed as a matrix function of $\mathbf{P}(t)$ as follows:

$$\mathbf{P}'(t) = \mathbf{U}\mathbf{P}(t)\mathbf{V},$$

where $\mathbf{V}$ is a 4 × q matrix, where q is the number of states in the lumped process, such that the l-th column of $\mathbf{V}$ is a vector with 1's in the components corresponding to states in S_l and 0's otherwise, and

$$\mathbf{U} = \left(\mathbf{V}^T\Pi\mathbf{V}\right)^{-1}\mathbf{V}^T\Pi,$$

is a q × 4 matrix whose k-th row is a probability vector with non-zero elements corresponding to the states in $\mathcal{S}_k$. A useful necessary and sufficient condition for lumpability [41,42] is

$$\mathbf{V}\mathbf{U}\mathbf{P}(t)\mathbf{V} = \mathbf{P}(t)\mathbf{V}. \tag{8}$$

In the case of nucleotides, the second column of Table 2 gives the conditions required for lumpability of four nucleotides.

We note that under certain conditions, such as those considered by the JC [43] and F81 [44] models, all recodings are lumpable. Conditions under which recoding of nucleotides are possible for the K2P [45] and HKY [46] models are given in Table 3.2 of [32].

Tests for lumpability

We consider three possible tests: An *ad hoc* test based on a parametric bootstrap for an index of departure from the lumpability condition [32]; a test based on a test for lumpability in Markov chains [47]; and a likelihood-ratio test.

Index test

From (8), if a Markov process is lumpable, then

$$\mathbf{M} = \mathbf{V}\mathbf{U}\mathbf{P}(t)\mathbf{V} - \mathbf{P}(t)\mathbf{V}$$

should have all elements zero. Consider the index proposed in [32]:

$$\eta = \left(\sum_{i,j} m_{ij}^2\right)^{1/2}, \text{ where } m_{ij} = \mathbf{M}. \tag{9}$$

Table 2 Conditions required for a 4-state Markovian process to be lumpable (in terms of s and π), and transformations to obtain $\tilde{\pi}$ and $\tilde{s}$ such that the lumpability holds.

$\mathcal{S}'$	Lumpability conditions	$(\tilde{\pi}, \tilde{s})$
{{A, G}, C, T}	$s_{12} = s_{23}$ $s_{14} = s_{34}$	$\tilde{s}_{13} = \tilde{s}_{23} = (\hat{s}_{13} + \hat{s}_{23})/2$ $\tilde{s}_{14} = \tilde{s}_{34} = (\hat{s}_{14} + \hat{s}_{34})/2$
{A, G, {C, T}}	$s_{12} = s_{14}$ $s_{23} = s_{34}$	$\tilde{s}_{12} = \tilde{s}_{14} = (\hat{s}_{12} + \hat{s}_{14})/2$ $\tilde{s}_{23} = \tilde{s}_{34} = (\hat{s}_{23} + \hat{s}_{34})/2$
{A, {C, G}, T}	$s_{12} = s_{13}$ $s_{24} = s_{34}$	$\tilde{s}_{12} = \tilde{s}_{13} = (\hat{s}_{12} + \hat{s}_{13})/2$ $\tilde{s}_{24} = \tilde{s}_{34} = (\hat{s}_{24} + \hat{s}_{34})/2$
{C, G, {A, T}}	$s_{12} = s_{24}$ $s_{13} = s_{34}$	$\tilde{s}_{12} = \tilde{s}_{24} = (\hat{s}_{12} + \hat{s}_{24})/2$ $\tilde{s}_{13} = \tilde{s}_{34} = (\hat{s}_{13} + \hat{s}_{34})/2$
{{A, C}, G, T}	$s_{13} = s_{23}$ $s_{14} = s_{24}$	$\tilde{s}_{13} = \tilde{s}_{23} = (\hat{s}_{13} + \hat{s}_{23})/2$ $\tilde{s}_{14} = \tilde{s}_{24} = (\hat{s}_{14} + \hat{s}_{24})/2$
{A, C, {G, T}}	$s_{13} = s_{14}$ $s_{12} = s_{13}$	$\tilde{s}_{13} = \tilde{s}_{14} = (\hat{s}_{13} + \hat{s}_{14})/2$ $\tilde{s}_{23} = \tilde{s}_{24} = (\hat{s}_{23} + \hat{s}_{24})/2$
{A, {C, G, T}}	$s_{12} = s_{13}$ $s_{13} = s_{14}$	$\tilde{s}_{12} = \tilde{s}_{13} = \tilde{s}_{14} = (\hat{s}_{12} + \hat{s}_{13} + \hat{s}_{14})/3$
{C, {A, G, T}}	$s_{12} = s_{23}$ $s_{23} = s_{24}$	$\tilde{s}_{12} = \tilde{s}_{23} = \tilde{s}_{24} = (\hat{s}_{12} + \hat{s}_{23} + \hat{s}_{24})/3$
{G, {A, C, T}}	$s_{13} = s_{23}$ $s_{23} = s_{34}$	$\tilde{s}_{13} = \tilde{s}_{23} = \tilde{s}_{34} = (\hat{s}_{13} + \hat{s}_{23} + \hat{s}_{34})/3$
{T, {A, C, G}}	$s_{14} = s_{24}$ $s_{24} = s_{34}$	$\tilde{s}_{14} = \tilde{s}_{24} = \tilde{s}_{34} = (\hat{s}_{14} + \hat{s}_{24} + \hat{s}_{34})/3$
{{A, G}, {C, T}}	$s_{12}\pi_2 + s_{14}\pi_4 = s_{23}\pi_2 + s_{34}\pi_4$ $s_{12}\pi_1 + s_{23}\pi_3 = s_{14}\pi_1 + s_{34}\pi_3$	$\tilde{s}_{23} = \frac{\tilde{s}_{12}(\hat{\pi}_2\hat{\pi}_3 - \hat{\pi}_1\hat{\pi}_4) + \tilde{s}_{14}\hat{\pi}_4(\hat{\pi}_1 + \hat{\pi}_3)}{\hat{\pi}_3(\hat{\pi}_2 + \hat{\pi}_4)}$ $\tilde{s}_{34} = \frac{\tilde{s}_{12}\hat{\pi}_1 + \tilde{s}_{23}\hat{\pi}_3 - \tilde{s}_{14}\hat{\pi}_1}{\hat{\pi}_3}$
{{A, T}, {C, G}}	$s_{12}\pi_2 + s_{13}\pi_3 = s_{24}\pi_2 + s_{34}\pi_3$ $s_{12}\pi_1 + s_{24}\pi_4 = s_{13}\pi_1 + s_{34}\pi_4$	$\tilde{s}_{13} = \frac{\tilde{s}_{12}(\hat{\pi}_1\hat{\pi}_3 - \hat{\pi}_2\hat{\pi}_4) + \tilde{s}_{24}\hat{\pi}_4(\hat{\pi}_2 + \hat{\pi}_3)}{\hat{\pi}_3(\hat{\pi}_1 + \hat{\pi}_4)}$ $\tilde{s}_{34} = \frac{\tilde{s}_{12}\hat{\pi}_2 + \tilde{s}_{13}\hat{\pi}_3 - \tilde{s}_{24}\hat{\pi}_2}{\hat{\pi}_3}$
{{A, C}, {G, T}}	$s_{13}\pi_3 + s_{14}\pi_4 = s_{23}\pi_3 + s_{24}\pi_4$ $s_{13}\pi_1 + s_{23}\pi_2 = s_{14}\pi_1 + s_{24}\pi_4$	$\tilde{s}_{23} = \frac{\tilde{s}_{13}(\hat{\pi}_2\hat{\pi}_3 - \hat{\pi}_1\hat{\pi}_4) + \tilde{s}_{14}\hat{\pi}_4(\hat{\pi}_1 + \hat{\pi}_2)}{\hat{\pi}_2(\hat{\pi}_3 + \hat{\pi}_4)}$ $\tilde{s}_{24} = \frac{\tilde{s}_{13}\hat{\pi}_1 + \tilde{s}_{23}\hat{\pi}_2 - \tilde{s}_{14}\hat{\pi}_1}{\hat{\pi}_2}$

It is clear that $\eta \geq 0$, with η being 0 only under lumpable Markovian processes. Then, the hypothesis that the Markov process is lumpable is equivalent to the hypothesis $H_0 : \eta = 0$. From the observed divergence matrix, $\mathbf{N}$ of two homologous sequences, assuming a SRH Markovian model of evolution, an estimate $\hat{\eta}$ can be used as a test statistic for H_0, where

$$\hat{\eta} = \left(\sum_{i,j} m_{ij}^2 \right)^{1/2},$$

and $\hat{\mathbf{M}} = \mathbf{V}\mathbf{U}\hat{P}(1)\mathbf{V} - \hat{\mathbf{P}}(1)\mathbf{V}$ for $\hat{\mathbf{P}}(1) = \exp(\hat{\mathbf{S}}\hat{\Pi})$.

The distribution of $\hat{\eta}$ is unknown, so we propose an approximation to it that is based on the parametric bootstrap. The estimated vectors $\hat{\pi}$ and $\hat{s}$ do not necessarily satisfy the conditions for lumpability, so we obtain $\tilde{\pi}$ and $\tilde{s}$ using the relevant equations from the third column of Table 2 as estimates that do satisfy the lumpability condition. Once the $\tilde{\pi}$ and $\tilde{s}$ vectors are calculated, a procedure similar to that shown in (2), (3) and (4) is carried out such that the matrices $\tilde{\mathbf{R}}$, $\tilde{\mathbf{P}}(1)$, and $\tilde{\mathbf{F}}(1)$ are generated under the lumpability conditions. Now B matrices can be generated by simulation under conditions of lumpability, where we take $\mathbf{N}_b^*$, with $b \in \{1, ..., B\}$, to be independent and multinomial with parameters n and $\tilde{\mathbf{F}}(1)$. From each of these simulated samples, we calculate Π_b^*, Π_b^* and $\mathbf{S}_b^*$ from $\mathbf{F}_b^*$, as in (5) and (6), and then $\mathbf{P}_b^*(1) = \exp(\mathbf{S}_b^*\Pi_b^*)$, $\mathbf{M}_b^* = \mathbf{V}\mathbf{U}\mathbf{P}_b^*(1)\mathbf{V} - \mathbf{P}_b^*(1)\mathbf{V}$, and

$$\eta_b^* = \left(\sum_{i,j} \hat{m}_{bij}^{*2} \right)^{1/2}.$$

The true P-value is then the probability that we obtain a value as large as or larger than the observed $\hat{\eta}$, so a bootstrap approximation to this P-value is the proportion of $\eta_1^*, ..., \eta_B^*$ exceeding $\hat{\eta}$.

Markov chain test

A χ^2 test to determine whether a Markov chain is lumpable with respect to a partition $\mathcal{S}'$ is available [47]. The test is based on the comparison of observed transition frequencies to their respective theoretical counterparts under the null hypothesis that the chain is lumpable. The approach does not make any assumption about reversibility or stationarity of the process. The authors used a matrix of transition counts, $\{n_{ij}\}$, to estimate the transition probabilities p_{ij}, where n_{ij} represents the number of transitions into state j from state i in one step, so the number of steps in the Markov chain is $n_{\bullet\bullet}$, where the subscript $\bullet$ indicates summation. Now, if we start from our divergence matrix **N**, where n_{ij} represents the number of sites that are in state i for sequence A and state j in sequence B, and the SRH assumptions are kept, either A or B can be assumed to be the original sequence at time 0, whereas the other one can be assumed to be the observed sequence at time 2 (since we took the edge lengths to be 1). Take A as the ancestral sequence, then the divergence matrix **N** has the same properties as a transition count matrix, and we can proceed as described in [47]. A transition probability from i to $\mathcal{S}_l$ is

$$g_{il} = \sum_{j \in \mathcal{S}_l} P_{ij}(2),$$

where $l = 1, ..., q$. From the definition of a lumpable process (7), if the Markovian process is lumpable with respect to $\mathcal{S}'$, then

$$g_{il} = \sum_{i \in \mathcal{S}_k} \sum_{j \in \mathcal{S}_l} P_{ij}(2)/\gamma_k,$$

where γ_k is the number of states that are part of the subset $\mathcal{S}_k$, and $k = 1, ..., q$. Therefore, $g_{il} = P'_{kl}(2)$ if the process is lumpable and the null hypothesis of lumpability can be expressed as $H_0 : g_{il} = P'_{kl}(2)$ for all $i \in \mathcal{S}_k$. Given the divergence matrix, **N**, estimates g_{il} and $P'_{kl}(2)$ are

$$\hat{g}_{il} = \sum_{j \in \mathcal{S}_l} \frac{n_{ij}}{n_{i\bullet}}$$

and

$$\hat{P}'_{kl}(2) = \frac{n'_{kl}}{n'_{k\bullet}} = \frac{\sum_{i \in \mathcal{S}_k} \sum_{j \in \mathcal{S}_l} n_{ij}}{\sum_{i \in \mathcal{S}_k} n_{i\bullet}},$$

where $\mathbf{N}' = \{n'_{kl}\}$ is the divergence matrix of the recoded nucleotide sequences. Jernigan and Baran [47] obtained the test statistic

$$T = \sum_{i=1}^{4} \sum_{l=1}^{q} \frac{(o_{il} - e_{il})^2}{e_{il}},$$

where

$$o_{il} = \sum_{j \in \mathcal{S}_l} n_{ij}$$

and

$$e_{il} = \frac{n_{i\bullet} n'_{kl}}{n'_{k\bullet}},$$

and showed (by pointing out that $o_{il} - e_{il}$ are a stack of q tables of size $4 \times \gamma_l$ of mean-corrected multinomials with row and column sums equal to zero) that the test statistic is distributed under H_0 as a χ^2 variable with $(q-1)\sum_{k=1}^{q}(\gamma_k - 1)$ degrees of freedom, if all cells are non-zero. In the case considered here, the degrees of freedom for any of the recoding schemes is 2.

Likelihood-ratio test

Consider estimates $(\hat{\pi}, \hat{s})$ whose values maximize a log-likelihood function

$$L(\pi, s) = \sum_{i,j} n_{ij} \ln f_{ij}(\pi, s),$$

where $\{n_{ij}\} = \mathbf{N}$, the observed divergence matrix, $\mathbf{F}_{(\pi,s)}(1) = \exp(\mathbf{S\Pi})$ and $\{f_{ij}(\pi, s)\} = \mathbf{F}_{(\pi,s)}(1)$. These matrices are obtained as shown in (5) and (6). We also want to estimate (π, s) under the constraints imposed by the null hypothesis of lumpablity, H_0. The constraints are given in the second column of Table 2. Then we can define the constrained estimates $(\tilde{\pi}, \tilde{s})$ to satisfy

$$L(\tilde{\pi}, \tilde{s}) = \max_{\pi, s \in H_0} L(\pi, s).$$

This maximization needs a new approach. We construct an orthogonal matrix, **A**, such that

$$\mathbf{A}s = \gamma,$$

where y is the response constraint vector, defined such that two values of y are zero corresponding to the two constraints. The matrix **A** will, in the case of partitions into two groups of two, contain π, so to emphasise this possible dependence, write $y = g(s|\pi)$. Also write $s = \mathbf{A}^{-1}y = g^{-1}(y|\pi)$. Then

$$L(\tilde{\pi}, \tilde{s}) = \max_{\pi, \gamma} L(\pi, g^{-1}(\gamma|\pi)).$$

The optimization process is done in two steps: the values of s, if dependent of π, are optimized given the original π set, then the π vector is optimized given the optimized values of s. This process is repeated until convergence is achieved.

From these two log-likelihood values, a log likelihood-ratio, LR, can be calculated with

$$LR = L(\hat{\pi}, \hat{s}) - L(\tilde{\pi}, \tilde{s})$$

Under the null hypothesis of lumpability, $2 \times LR$ is distributed as a χ^2 variable with 2 degrees of freedom.

Results

Assessment of accuracy

In order to check the accuracy of the tests under the null hypothesis, Monte Carlo simulations were done from a set of parameters that meets the assumption of lumpability. The parameter vectors in this case were

$$\pi^T = (0.1, 0.2, 0.3, 0.4)$$

and

$$s^T = (0.2, 0.25, 0.2, 0.2, 0.15, 0.2).$$

The joint probability distribution was calculated by the steps given in (2), (3), and (4); then, assuming a nucleotide sequence of length n = 1500, 5000 divergence matrices were calculated by Monte Carlo simulations assuming that $\mathbf{N}_i$ is multinomial with parameters $(n, \mathbf{F}(1))$ for $i = 1, \ldots, 5000$.

The accuracy of each test for lumpability was verified using a *PP* plot displaying the distribution of observed *P*-values, obtained from each test, plotted against the expected *P*-values, obtained from the uniform distribution. The linear relationship between these two sets of *P*-values (Figure 1) confirms the accuracy of the tests.

Comparisons of power

The power of each test was compared for each recoding scheme under non-lumpable conditions. To do this, we used π^T = (0.1, 0.2, 0.3, 0.4) and values of s that yield increasing values of η, as in (9), generated 3000 divergence matrices using Monte Carlo simulation, and then calculated the three test statistics and their corresponding *P*-values, using the procedures explained above, for each value of η. The power at the 5% level, is then equal to the proportion of observed *P*-values less than 0.05.

Figure 2 shows the power curves for *RY* recoding–similar power curves were obtained for the other 12 recoding schemes (results not shown). All of these results indicate that the likelihood-ratio test is the most powerful of the tests considered, followed by the Index test and, finally, by the Markov Chain test.

Cases with more than 2 homologous sequences

For general cases involving more than 2 homologous sequences, we can test for lumpability in all pairs of sequences by using the methods described above under the assumption that the evolutionary process is SRH over the whole tree (i.e., the process is globally SRH). For example, in the case of an alignment with seven sequences, there will be 21 *P*-values. A *PP*-plot with these *P*-values should yield a straight line when the data are lumpable, and deviations from this expectation when the processes are not lumpable. However, the observed *P*-values are not independent, so we need to show that this condition is not cause for concern for the dots in a *PP*-plot to be on a straight line. We give a simplified argument taken from [48]. Consider a set of observed *P*-values $P_1, \ldots, P_n$, which, if all the null hypotheses are true, are identically distributed as uniform random variables on (0, 1). Let p be any value between 0 and 1, let $I(P_j < p)$ take the value 1 if

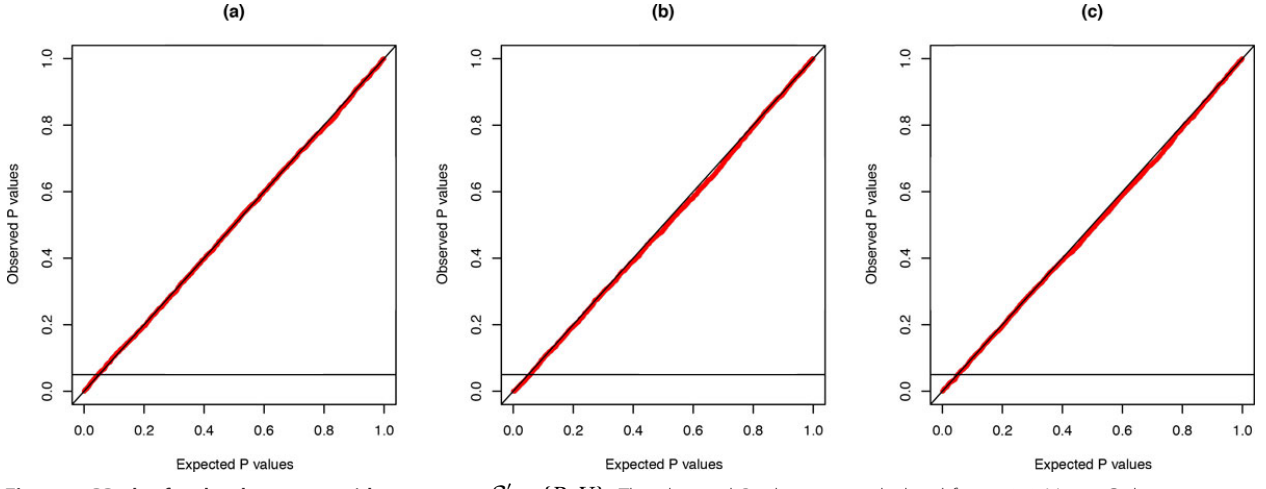

Figure 1 *PP*-plot for the three tests with respect to $\mathcal{S}' = \{R, Y\}$. The observed *P*-values were calculated from 5000 Monte Carlo experiments for each test (i.e., the Index test (a), the Markov chain test (b), and the likelihood-ratio test (c)) and charted against the expected *P*-values.

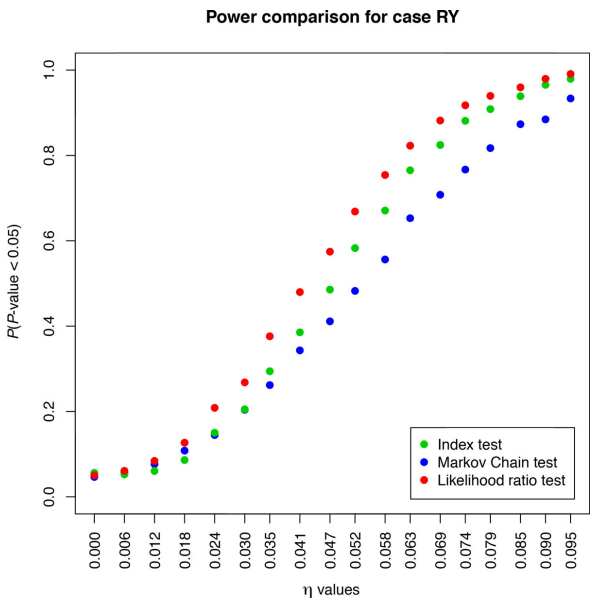

Figure 2 Power curve for partition $\mathcal{S}' = \{R, Y\}$. A total of 3000 divergence matrices were generated by Monte Carlo simulation for each triplet of points, corresponding to the three test statistics, and 500 parametric bootstraps were used during the calculation of *P*-values for the η index. The same π-vector (i.e., π^T = (0.1, 0.2, 0.3, 0.4)) was used for every triplet of points whereas the values in the *s*-vector were allowed to vary slightly for each point.

$P_j < p$ and 0 otherwise, and let $N_p = \sum_{j=1}^{n} I\left(P_j < p\right)$ be the number of observed *P*-values less than *p*. Then $E(I(P_j < p)) = P\ (P_j < p) = p$ and so the expected number of *P*-values less than *p* is $E(N_p) = np$. This implies that the plot of the observed *P*-values will lie approximately on a straight line. The dependence will cause some clustering of the observed *P*-values but the *PP*-plot will remain useful in indicating whether there is evidence against some of the hypotheses.

The *PP*-plots shown in Figure 3 were obtained from alignments of nucleotides generated under lumpable or non-lumpable conditions, with respect to *RY* recoding, on the tree shown in Figure 4 before being analysed using the likelihood-ratio test. From these two plots, it is clear that the test is able to identify cases where sequences have evolved under non-lumpable conditions.

The effect of non-lumpability on phylogenetic estimates

If a process is lumpable with respect to a given recoding scheme (e.g., *RY*), then we can obtain $\mathbf{F}_q(t) = \mathbf{V}^T\mathbf{P}(t)\mathbf{V}$ and from this, using (2, 3) and (4), we can further obtain the transition matrix of the process $X'(t)$,

$$\mathbf{P}_q(t) = \mathbf{\Pi}_q^{-1/2}\mathbf{L}_q e_q^{\mathbf{\Lambda}t}\mathbf{L}_q^T\mathbf{\Pi}_q^1/2,$$

where $\mathbf{\Pi}_q = \mathbf{V}^T\mathbf{\Pi}\mathbf{V}$. If the process is not lumpable with respect to that recoding scheme, then $X'(t)$ is not a Markov process and, although we can calculate $\mathbf{P}_q(t)$, it is *not* the transition matrix of the process $X'(t)$.

In either case, the matrix $\mathbf{P}'(t) = \mathbf{U}\mathbf{P}(t)\mathbf{V}$ can be defined and, if the process $X(t)$ is lumpable, then $P'(t) = P_q(t)$. On the other hand, if $X(t)$ is not lumpable, the elements of $\mathbf{P}'(t)$ are still given as $\mathbf{P}'_{kl}(t) = \mathbf{P}(X'(t) = l|X'(0) = k)$, but $\mathbf{P}'(t)$ is no longer a transition matrix of $X'(t)$. Conveniently, we can compare $\mathbf{P}'(1)$, the true conditional probability matrix at $t = 1$, with $\mathbf{P}_q$ (1), the false transition

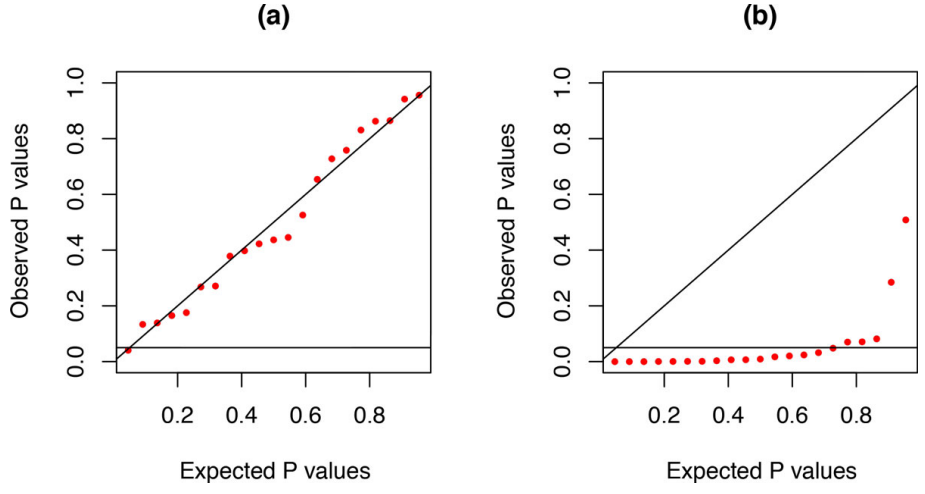

Figure 3 *PP*-plots for lumpability tests for simulated data. *PP*-plots for the likelihood-ratio test, with respect to *RY* recoding, for the randomly-generated lumpable 7-taxon data (a) and the randomly-generated non-lumpable 7-taxon data (b). Sequences comprising 2000 sites were generated on the tree shown in Figure 4 under time-reversible conditions with sites evolving under independent and identical conditions ((a): π^T = (0.1, 0.2, 0.3, 0.4), and s^T = (0.20, 0.25, 0.20, 0.20, 0.15, 0.20); (b): π^T = (0.1, 0.2, 0.3, 0.4), and s^T = (0.50, 0.25, 0.20, 0.20, 0.15, 0.20)).

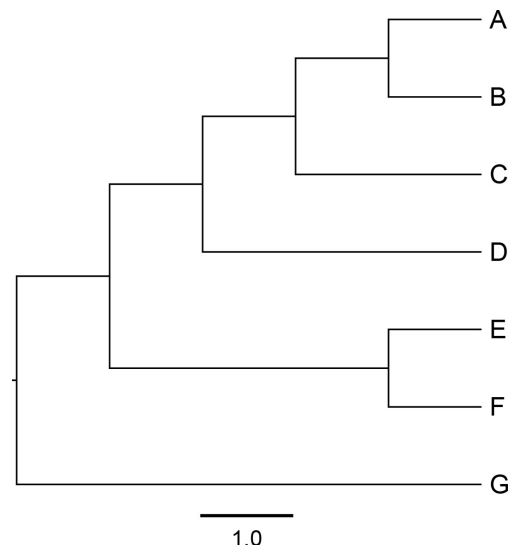

Figure 4 Tree used to generate simulated data. Tree used to generate the simulated data, which were then analysed to obtain the results shown in Figure 3. The scale bar corresponds to 1 time unit.

matrix at $t = 1$, thus allowing us to examine the effect of non-lumpability.

Figure 5 illustrates the effect on phylogenetic estimates. Figure 5a shows the tree used to simulate alignments of 3000 nucleotides under a time-reversible 4-state Markov model with π^T = (0.2, 0.3, 0.2, 0.3) and s^T = (0.2, 0.1, 0.3, 0.3, 1.0, 0.2). As can be seen from the π- and s-vectors, the lumpable condition is met for *RY*-recoding but not for *KM*-recoding. Figures 5b and 5c display the corresponding tree with the edge lengths adjusted according to, respectively, the *RY*- and *KM*-recoding schemes.

Every edge in the tree obtained from the *RY*-recoded data is shorter than the corresponding edge in the tree obtained from the original data. However, because the original process was lumpable with respect to *RY*-recoding, the relative length of each edge in the two trees is the same, the difference being equal to a scale factor

$$\rho = \frac{\sum_{i=1}^{4} \pi_i r_{ii}}{\sum_{j=1}^{2} \pi_{2j} r_{2jj}}.$$

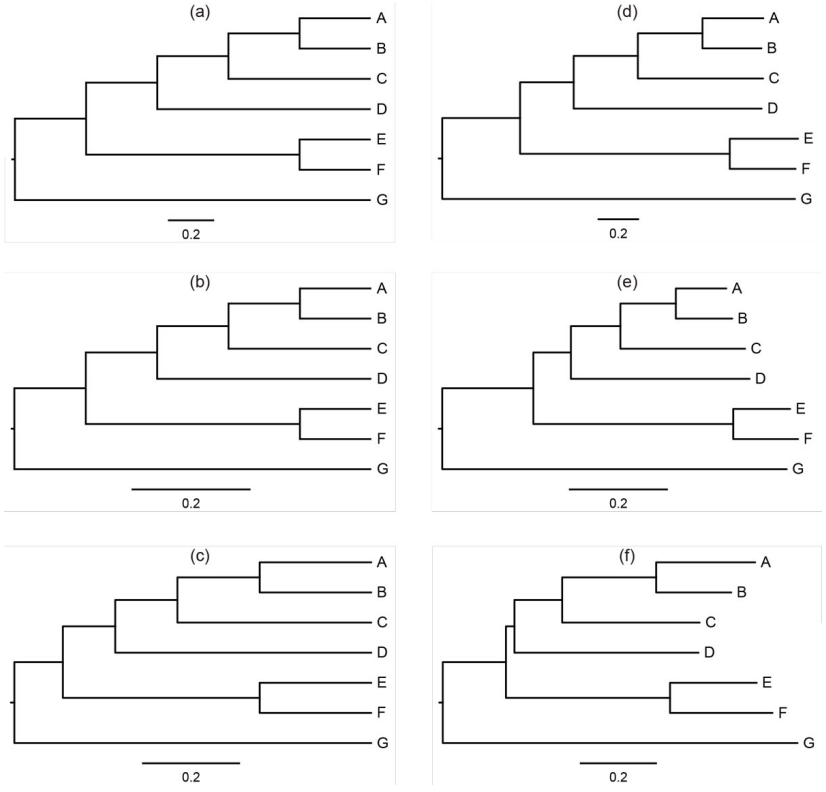

Figure 5 Effect of recoding on phylogenetic estimates. Panel (a) displays the tree that was used to generate alignments of 3000 nucleotides under a time-reversible 4-state Markov model. Panel (b) shows the corresponding tree with edge lengths adjusted according to the *RY*-recoding scheme. Panel (c) shows the corresponding tree with edge lengths adjusted according to the *KM*-recoding scheme. Panels (d), (e), and (f) present the corresponding results obtained by analysis of the data generated on the tree in panel (a). The scale bar corresponds to the expected number of substitutions $\left(\text{i.e.}, -\sum_{i=1}^{4} \pi_i r_{ii}\right)$.

Every edge in the tree obtained from the *KM*-recoded data is also shorter than the corresponding edge in the tree obtained from the original data, but the relative length of each edge in the two trees differ, the reason being that the process generating the original data was not lumpable with respect to *KM*-recoding.

Figures 5d, 5e, and 5f show the corresponding results for the data generated by Monte Carlo simulation. The three trees display the same characteristics as those shown in Figures 5a, 5b, and 5c, while also showing some variation in the edge lengths that is due to the random nature of the data and the finite sample size. Hence, although recoding of nucleotides might be useful for a variety of reasons, using recoded data, without having tested for lumpability first, might lead to biased phylogenetic estimates.

Example 1 – Primate mitochondrial DNA

In a previous study [37], a set of mitochondrial nucleotide sequences of hominoid origin were found to to fit the GTR model [49], implying that the data are consistent with evolution under globally SRH conditions. We applied the likelihood-ratio test to these data. Figure 6 shows the *PP*-plots from tests for lumpability for *RY* recoding, indicating non-lumpability, and *AGY* recoding, indicating lumpability.

Example 2 – Primate nuclear DNA

In a previous study [50], a ~9.3kb fragment of X chromosomal DNA was obtained from 26 species of primates and analysed phylogenetically using the HKY model. In so doing, the authors implicitly assumed evolution under globally SRH conditions. We wanted to apply the likelihood-ratio test to these data, so we obtained the same 26 sequences from GenBank [51], aligned them using MAFFT [52] (with the linsi option invoked), and, using SeaView [53], removed all columns with gaps and/or ambiguous characters. The resulting alignment contained 6913 sites from the 26 species. We then applied the matched-pairs test of symmetry [38] to the data to determine whether the sequences were consistent with evolution under globally SRH conditions. The *PP*-plot in Figure 7 clearly shows that the data are consistent with evolution under these conditions. Hence, it is appropriate to use our likelihood-ratio test to determine whether any

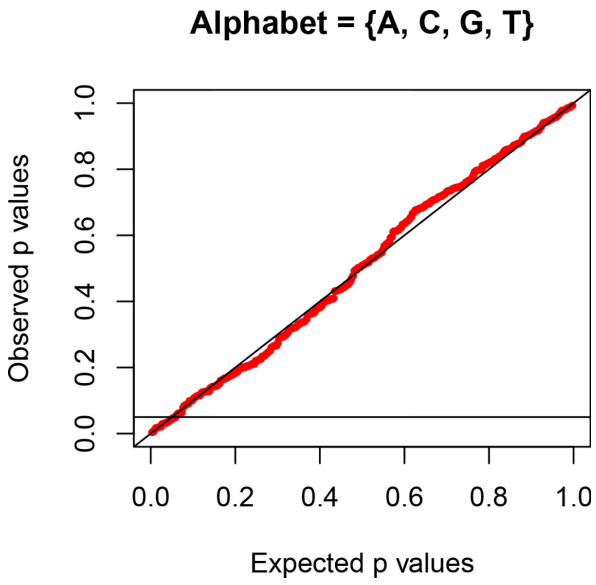

Figure 7 *PP*-plot for matched-pairs test of symmetry for primate nuclear data. *PP*-plot demonstrating consistency with evolution under globally SRH conditions for the 4-state process.

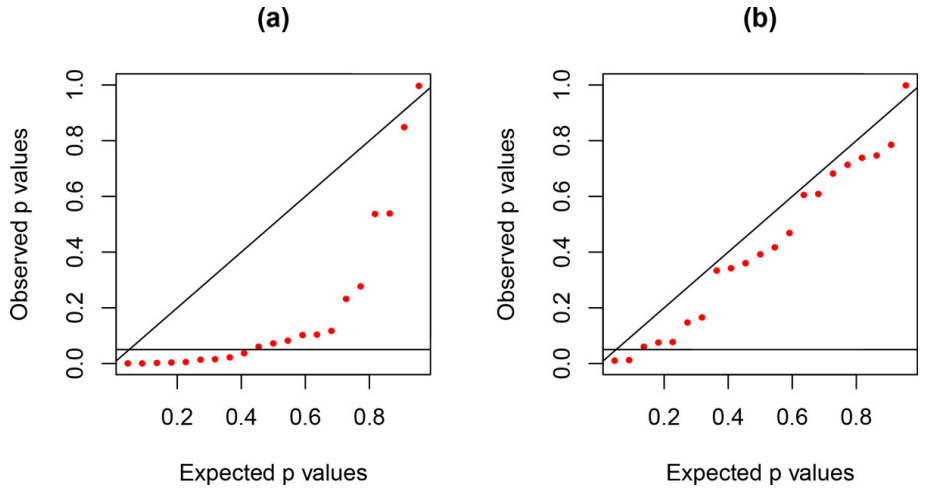

Figure 6 *PP*-plots for lumpability tests for hominoid mitochondrial data. *PP*-plots for the likelihood-ratio tests for lumpability for hominoid data for *RY*-recoding (a) and for *AGY*-recoding (b), respectively.

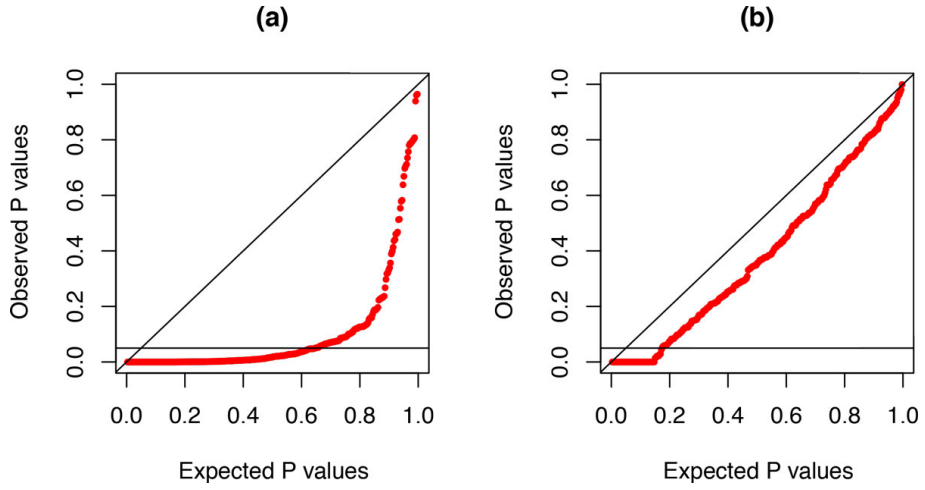

Figure 8 *PP*-plots for tests of lumpability for primate nuclear data. *PP*-plots for lumpability for *RY*-recoding (a) and *SW*-recoding (b), respectively.

of the recoding schemes would retain the Markovian properties of the original data. Figure 8 presents the *PP*-plots from the likelihood-ratio test for lumpability for *RY*-recoding, showing strong evidence against lumpability, and for the *SW*-recoding, which provided the least evidence against lumpability. It is evident that no recoding should be applied to these data.

Conclusions

Bias in estimates of phylogenetic parameters can occur when recoding of nucleotides or amino acids is used to transform data associated with models of evolution, which are not lumpable with respect to the recoding scheme used. A test proposed in this paper, which is based on a likelihood-ratio test, can yield an indication of whether the same results for estimable parameters can be expected from fitting a given model of evolution and its recoded version to the data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LSJ and JR conceived the project. KWL carried out the pilot study, showing that the Index test was feasible. VAVR and JR developed the tests and wrote the manuscript with input from LSJ and KWL. VAVR wrote the codes and generated the numerical results. All authors have read and approved the manuscript.

Acknowledgements

VAVR was supported by the University of Sydney World Scholars scheme. LSJ, KWL, and JR were supported by a Discovery Grant (DP0453173) from the Australian Research Council. We thank DL Lovell and TKF Wong for their constructive suggestions to this manuscript.

Declarations

The publication costs for this article were funded by resources made available to LSJ by CSIRO and JR by the School of Mathematics and Statistics, University of Sydney.

This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 2, 2014: Selected articles from the Twelfth Asia Pacific Bioinformatics Conference (APBC 2014): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S2.

Authors' details

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia. [2]School of Biological Sciences, University of Sydney, NSW 2006, Australia. [3]CSIRO Computational Informatics, Floreat, WA, 6014, Australia. [4]CSIRO Ecosystem Sciences, Canberra, ACT 2601, Australia.

Published: 24 January 2014

References

1. Irwin DM, Kocher TD, Wilson AC: **Evolution of the cytochrome b gene in mammals.** *Journal of Molecular Evolution* 1991, **32**:128-144.
2. Adkins RM, Honeycutt RL: **Molecular phylogeny of the superorder Arconta.** *Proceedings of the National Academy of Science of the United States of America* 1991, **88**:10317-10321.
3. Adkins RM, Honeycutt RL: **Evolution of the primate cytochrome c oxidase subunit II gene.** *Journal of Molecular Evolution* 1994, **38**:215-231.
4. Woese CR, Achenbach L, Rouviere P, Mandelco L: **Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts.** *Systematic and Applied Microbiology* 1991, **14**:364-371.
5. Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mithocondrial genomes.** *Molecular Phylogenetics and Evolution* 2003, **28**:171-185.
6. Cavender JA, Felsenstein J: **Invariants of phylogenies in a simple case with discrete states.** *Journal of Classification* 1987, **4**:57-71.
7. Gibson A, Gowri-Shankar V, Higgs PG, Rattray M: **A comprehensive analysis of mammalian mithochondrial genome base composition and improved phylogenetic methods.** *Molecular Biology and Evolution* 2005, **22**:251-264.
8. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH: **Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus.** *The Plant Cell* 2001, **13**:645-658.
9. Phillips MJ, Lin YH, Harrison GL, Penny D: **Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials.** *Proceedings of the Royal Society London Series B* 2001, **268**:1533-1538.
10. Kosiol C, Goldman N, Buttimore NH: **A new criterion and method for amino acid classification.** *Journal of Theoretical Biology* 2004, **228**:97-106.

11. Kosiol C: **Markov models for protein sequence evolution.** *PhD thesis* University of Cambridge; 2006.
12. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases.** *Molecular Biology and Evolution* 2004, **21**:1455-1458.
13. Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS: **SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides.** *Bioinformatics* 2006, **221**:2162-2163.
14. Susko E, Roger AJ: **On reduced amino acid alphabets for phylogenetic inference.** *Molecular Biology and Evolution* 2007, **24**:2139-2150.
15. Anisimova M, Kosiol C: **Investigating protein-coding sequence evolution with probabilistic codon substitution models.** *Molecular Biology and Evolution* 2004, **26**:255-271.
16. Masta SE, Longhorn SJ, Boore JL: **Arachnid relationships based on mitochondrial genomes: asymmetric nucleotide and amino acid bias affects phylogenetic analyses.** *Molecular Phylogenetics and Evolution* 2009, **50**:117-128.
17. Phillips MJ, Gibb GC, Crimp EA, Penny D: **Tinamous and moa flock together: mitochondrial genome sequence analysis reveals independent losses of flight among ratites.** *Systematic Biology* 2010, **59**:90-107.
18. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW: **Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences.** *Nature* 2010, **463**:1079-1083.
19. Criscuolo A, Gribaldo S: **Large-scale phylogenomic analyses indicate a deep origin of primary plastids within Cyanobacteria.** *Molecular Biology and Evolution* 2011, **28**:3019-3032.
20. Regier JC, Zwick A: **Sources of signal in 62 protein-coding nuclear genes for higher-level phylogenetics of arthropods.** *PLoS ONE* 2011, **6**:e23408.
21. Cho S, Zwick A, Regier JC, Mitter C, Cummings MP, Yao J, Du Z, Zhao H, Kawahara AY, Weller S, Davis DR, Baixeras J, Brown JW, Parr C: **Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)?** *Systematic Biology* 2011, **60**:782-796.
22. White NE, Phillips MJ, Gilbert MTP, Alfaro-Nunez A, Willerslev E, Mawson PR, Spencer PBS, Bunce M: **The evolutionary history of cockatoos (Aves: Psittaciformes: Cacatuidae).** *Molecular Phylogenetics and Evolution* 2011, **59**:615-622.
23. Zwick A, Regier JC, Cummings MP, Mitter C: **Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera).** *Systematic Entomology* 2011, **36**:31-43.
24. Niehuis O, Hartig G, Grath S, Pohl H, Lehmann J, Tafer H, Donath A, Krauss V, Eisenhardt C, Hertel J, Petersen M, Mayer C, Meusemann K, Peters RS, Stadler PF, Beutel RG, Bornberg-Bauer E, McKenna DD, Misof B: **Genomic and morphological evidence converge to resolve the enigma of Strepsiptera.** *Current Biology* 2012, **22**:1309-1313.
25. Regier JC, Brown JW, Mitter C, Baixeras J, Cho S, Cummings MP, Zwick A: **A molecular phylogeny for the leaf-roller moths (Lepidoptera: Tortricidae) and its implications for classification and life history evolution.** *PLoS ONE* 2012, **7**:e35574.
26. Regier JC, Mitter C, Solis MA, Hayden JE, Landry B, Nuss M, Simonsen TJ, Yen S-H, Zwick A, Cummings MP: **A molecular phylogeny for the pyraloid moths (Lepidoptera: Pyraloidea) and its implications for higher-level classification.** *Systematic Entomology* 2012, **37**:635-656.
27. Zwick A, Regier JC, Zwickl DJ: **Resolving discrepancy between nucleotides and amino acids in deeplevel arthropod phylogenomics: differentiating serine codons in 21-amino-acid models.** *PLoS ONE* 2012, **textbf7**:e47450.
28. Gibb GC, Kennedy M, Penny D: **Beyond phylogenetics and evolution: pelecaniform and Ciconiiform birds, and long-term niche stability.** *Molecular Phylogenetics and Evolution* 2013, **68**:229-238.
29. Regier JC, Mitter C, Zwick A, Bazinet AL, Cummings MP, Kawahara AY, Sohn J-C, Zwickl DJ, Cho S, Davis DR, Baixeras J, Brown J, Parr C, Weller S, Lees DC, Mitter KT: **A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (Moths and Butterflies).** *PLoS ONE* 2013, **8**:e58568.
30. Rota-Stabelli O, Lartillot N, Philippe H, Pisani D: **Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study.** *Systematic Biology* 2013, **62**:121-133.
31. Sohn J-C, Regier JC, Mitter C, Davis D, Landry J-F, Zwick A, Cummings MP: **A molecular phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and its implications for classification, biogeography and the evolution of host plant use.** *PLoS ONE* 2013, **textbf8**:e55066.
32. Lau KW: **Studies of methods used to infer molecular phylogeny: Dealing with the effect of compositional heterogeneity.** *PhD thesis* University of Sydney, School of Biological Sciences;; 2009.
33. Guédon Y, d'Aubenton-Carafa Y, Thermes C: **Analysing grouping of nucleotides in DNA sequences using lumped processes constructed from Markov chains.** *Journal of Mathematical Biology* 2006, **52**:343-372.
34. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM: **Phylogenetic inference.** In *Molecular Systematics*. Sunderland: Sinauer Associates;Hillis DM, Moritz C, Mable BK 1996:407-514.
35. Nomenclature Committee of the International Union of Biochemistry, (NC-IUB): **Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984.** *Proceedings of the National Academy of Sciences of the United States of America* 1986, **83**:4-8.
36. Bryant D, Galtier N, Poursat MA: **Likelihood calculation in molecular phylogenetics.** In *Mathematics evolution and phylogeny*. New York: Oxford University Press;Gascuel O 2005:33-92.
37. Jayaswal V, Jermiin LS, Robinson J: **Estimation of phylogeny using a general Markov model.** *Evolutionary Bioinformatics* 2005, **1**:62-80.
38. Ababneh F, Jermiin LS, Ma C, Robinson J: **Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences.** *Bioinformatics* 2006, **22**:1225-1231.
39. Jermiin LS, Jayaswal V, Ababneh F, Robinson J: **Phylogenetic model evaluation.** In *Bioinformatics: Data, sequence analysis, and evolution – Volume 1*. Humana Press. Totawa;Keith J 2008:331-363.
40. Ababneh F, Jermiin LS, Robinson J: **Generation of the exact distribution and simulation of matched nucleotide sequences on a phylogenetic tree.** *Journal of Mathematical Modelling and Algorithms* 2006, **5**:291-303.
41. Iosifescu M: *Finite Markov processes and their applications* Chichester: John Wiley and Sons, Ltd; 1980.
42. Kemeny JG, Snell JL: *Finite Markov chains* New York: Springer-Verlag; 1983.
43. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian Protein Metabolism*. Academic Press. New York;Munro HN 1969:21-132.
44. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *Journal of Molecular Evolution* 1981, **17**:368-376.
45. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *Journal of Molecular Evolution* 1980, **16**:111-120.
46. Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *Journal of Molecular Evolution* 1985, **22**:160-174.
47. Jernigan RW, Baran RH: **Testing lumpability in Markov chains.** *Statistics and Probability Letters* 2003, **64**:17-23.
48. Schweder T, Spjotvoll E: **Plots of P-values to evaluate many tests simultaneously.** *Biometrika* 1982, **69**:493-502.
49. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *Journal of Molecular Evolution* 1984, **20**:86-93.
50. Tosi AJ, Detwiler KM, Disotell TR: **X-chromosomal window into the evolutionary history of the guenons (Primates: Cercopithecini).** *Molecular Phylogenetics and Evolution* 2005, **36**:58-66.
51. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **Genbank.** *Nucleic Acids Research* 2013, **41**:D36-D42.
52. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.** *Molecular Biology and Evolution* 2013, **30**:772-780.
53. Gouy M, Guindon S, Gascuel O: **SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building.** *Molecular Biology and Evolution* 2010, **27**:221-224.

doi:10.1186/1471-2105-15-S2-S8
Cite this article as: Vera-Ruiz *et al.*: Statistical tests to identify appropriate types of nucleotide sequence recoding in molecular phylogenetics. *BMC Bioinformatics* 2014 **15**(Suppl 2):S8.