20

Assessing substitution saturation with DAMBE

THEORY

Xuhua Xia

20.1 The problem of substitution saturation

The accuracy of phylogenetic reconstruction depends mainly on (1) the sequence quality, (2) the correct identification of ***homologous*** sites by sequence alignment, (3) regularity of the substitution processes, e.g. ***stationarity*** along different lineages, absence of ***heterotachy*** and little variation in the substitution rate over sites, (4) ***consistency***, ***efficiency*** and little ***bias*** in the estimation method, e.g. not plagued by the ***long-branch attraction*** problem, and (5) sequence divergence, i.e. neither too conserved as to contain few substitutions nor too diverged as to experience substantial ***substitution saturation***. This chapter deals with assessing substitution saturation with software DAMBE, which is a Windows program featuring a variety of analytical methods for molecular sequence analysis (Xia, 2001; Xia & Xie, 2001).

Substitution saturation decreases phylogenetic information contained in sequences, and has plagued the phylogenetic analysis involving deep branches, such as major arthropod groups (Lopez *et al.*, 1999; Philippe & Forterre, 1999; Xia *et al.*, 2003). 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 (Lockhart *et al.*, 1992; Steel *et al.*, 1993; Xia, 2001, pp. 49–58; Xia *et al.*, 2003), which often does not reflect phylogenetic relationships.

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.) Published by Cambridge University Press.

Other than simple suggestions such as avoiding sequences with many pairwise JC69 (Jukes & Cantor, 1969) distances larger than 1 (Nei & Kumar, 2000, p. 112) and plotting ***transitions*** and ***transversions*** against a corrected ***genetic distance***, as implemented in DAMBE (Xia, 2001; Xia & Xie, 2001)(see also practice in Chapter 4), there are currently five main approaches for testing 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). These approaches, in addition to having difficulties with sequences with divergent nucleotide frequencies (Steel *et al.*, 1993), suffer from the problem that, as long as we have two closely related species, the tests will lead us to conclude that significant phylogenetic information is present in the data set even if all the other sequences have experienced full substitution saturation. This problem is also shared by the third approach, a tree-independent measure based on relative apparent ***synapomorphy***, implemented in the RASA program (Lyons-Weiler *et al.*, 1996). The forth approach (Steel *et al.*, 1993, 1995) is based on the ***parsimony*** method, proposed specifically to alleviate the problem of sequence convergence due to similarity in nucleotide frequencies. The convergence would become increasingly serious with increasing substitution saturation. Indeed, sequence similarity will depend entirely on similarity in nucleotide frequencies with full substitution saturation. The fifth is an ***information entropy***-based index of substitution saturation (Xia *et al.*, 2003). DAMBE (Xia, 2001; Xia & Xie, 2001) implements the last two approaches.

In what follows, I will (1) outline the method by Steel *et al.* (1993) to highlight its potential problems and its implementation in DAMBE with extensions, and (2) introduce the entropy-based index (Xia *et al.*, 2003) and its implementation in DAMBE with extensions not covered in the original paper. This is followed by a "Practice" section on how to use DAMBE to carry out the assessment of substitution saturation with molecular sequences by using one or both of the implemented methods. For simplicity, I will refer to the two implemented methods as *Steel's method* and *Xia's method*.

20.2 Steel's method: potential problem, limitation, and implementation in DAMBE

Steel *et al.* (1993) presented two statistical tests for testing phylogenetic hypotheses in a maximum parsimony (MP) framework for four species. The first evaluates the relative statistical support of the three possible unrooted topologies. The second

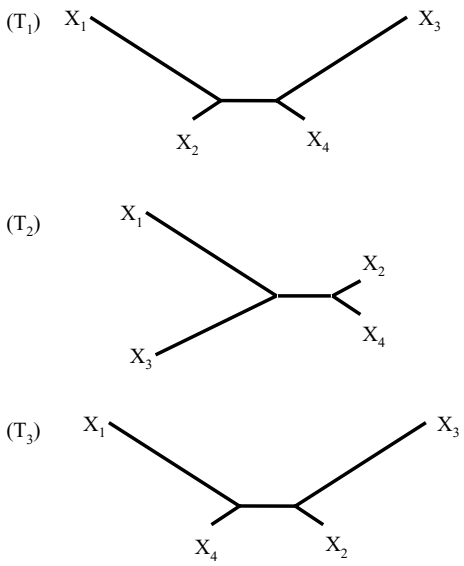

Fig. 20.1 The true topology (T_1) and the two possible alternative topologies: T_2 and T_3.

tests whether the distribution of the phylogenetically informative sites differs from the null model in which sequence variation is generated randomly, i.e. equivalent to sequences having experienced full substitution saturation.

Steel's method has a serious problem involving long-branch attraction mediated by substitution saturation. Long-branch attraction refers to the estimation bias (or tendency) of grouping highly divergent taxa as ***sister taxa*** even when they are not. While it is historically associated with the MP method (Felsenstein, 1978), other methods, such as distance-based methods with the ***minimum evolution*** criterion and with a distance index that severely underestimates the true genetic distance between divergent taxa, also suffer from the problem (Felsenstein, 2004).

To illustrate the problem in the MP context, let us focus on the topology in Fig. 20.1, with four species (four nucleotide sequences) designated X_i ($i = 1, 2, 3, 4$) and three possible unrooted topologies designated T_i ($i = 1, 2, 3$, Fig. 20.1, with T_1 being the true topology). Let X_{ij} be the nucleotide at site j for species X_i, and L be the sequence length. For simplicity, assume that nucleotide frequencies are all equal to 0.25. Suppose that the lineages leading to X_1 and X_3 have experienced full substitution saturation, so that

$$\Pr(X_{1j} = X_{ij,i\neq 1}) = \Pr(X_{3j} = X_{ij,i\neq 3}) = 0.25 \tag{20.1}$$

The lineages leading to X_2 and X_4 have not experienced substitution saturation and have

$$\Pr(X_{2j} = X_{4j}) = P \tag{20.2}$$

where $P > 0.25$ (Note that P approaches 0.25 with increasing substitution saturation). For simplicity, let us set $P = 0.8$, and L = 1000.

We now consider the expected number of informative sites, designated n_i (i = 1, 2, 3) as in Steel *et al.* (1993), favoring T_i. Obviously, site j is informative and favors T_1 if it meets the following three conditions: $X_{1j} = X_{2j}$, $X_{3j} = X_{4j}$, $X_{1j} \neq X_{3j}$. Similarly, site j favors T_2 if $X_{1j} = X_{3j}$, $X_{2j} = X_{4j}$, $X_{1j} \neq X_{2j}$. Thus, the expected numbers of informative sites favoring T1, T2 and T3, respectively, are

$$\begin{aligned} E(n_1) &= \Pr(X_{1j} = X_{2j}, X_{3j} = X_{4j}, X_{1j} \neq X_{3j})\, L \\ &= 0.25 \times 0.25 \times 0.75 \times 1000 \approx 47 \\ E(n_2) &= \Pr(X_{1j} = X_{3j}, X_{2j} = X_{4j}, X_{1j} \neq X_{2j})\, L \\ &= 0.25 \times 0.8 \times 0.75 \times 1000 = 150 \\ E(n_3) &= E(n_1) \approx 47 \end{aligned} \tag{20.3}$$

Designating c as the total number of informative sites, we have $c = \Sigma n_i = 244$. Equation (20.3) means that, given the true topology T_1, $P = 0.8$, L = 1000, and the condition specified in equations (20.1)–(20.2), we should have, on average, about 47 informative sites favoring T_1 and T_3, but 150 sites supporting T_2. Thus, the wrong tree receives much stronger support (150 informative sites) than the true topology (T_1) and the other alternative topology (T_3). This is one of the many causes for the familiar problem of long-branch-induced attraction (Felsenstein, 1978), popularly known as long-branch attraction, although short-branch attraction would seem more appropriate.

Suppose we actually have such sequences and observed $n_1 = n_3 = 47$, and $n_2 = 150$. What would Steel's method tell us? Steel *et al.* (1993) did not take into account the problem of long-branch attraction. They reasoned that, if the sequences are randomly generated, then some sites will be informative by chance and it is consequently important to assess whether the number of informative sites supporting a particular topology exceeds the number expected by chance. They designate the number of informative sites favoring T_i by chance alone as N_i ($i = 1$, 2, 3). The probability that N_i is at least as large as n_i, according to Steel *et al.* (1993), is

$$\Pr(N_i \geq n_i) = \sum_{k \geq n_i}^{c} \binom{c}{k} s_i^k (1 - s_i)^{c-k} \tag{20.4}$$

where s_i is defined in Steel *et al.* (1993) as the expected proportion of informative sites supporting T_i by chance. Because the nucleotide frequencies are all equal to 0.25 in our fictitious case, $s_i = 1/3$, so that

$$\begin{aligned} \Pr(N_2 \geq n_2) &= \sum_{k\geq 150}^{244} \binom{244}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{244-k} \approx 0 \\ \Pr(N_1 \geq n_1) = \Pr(N_3 \geq n_3) &= \sum_{k\geq 47}^{244} \binom{244}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{244-k} \approx 1 \end{aligned} \tag{20.5}$$

These equations mean that, by chance alone, it is quite likely to have $N_1 \geq 47$ and $N_3 \geq 47$. That is, there is little statistical support for T_1 and T_3. However, there is significant support for T2 since it is quite unlikely to have $N_2 \geq 150$ by chance alone. The same conclusion is reached by using the normalized values of n_i as recommended in Steel *et al.* (1993). So T_2, which is a wrong topology, is strongly supported by Steel's method.

In addition to the test above for evaluating the relative support of alternative topologies, Steel *et al.* (1993) also presented a statistic for testing whether the distribution of the informative sites differs from what is expected from random sequences (the null model):

$$X^2 = \sum_{i=1}^{3} \frac{(n_i - \mu_i)^2}{\mu_i} \tag{20.6}$$

where $\mu_i = cs_i = c/3 \approx 81.3$ and X2 follows approximately the χ^2 distribution with two degrees of freedom. In our case,

$$X^2 = \sum_{i=1}^{3} \frac{(n_i - \mu_i)^2}{\mu_i} \approx \frac{2(47 - 81.3)^2}{81.3} + \frac{(150 - 81.3)^2}{81.3} \approx 87.17 \tag{20.7}$$

With two degrees of freedom, the null model is conclusively rejected ($p = 0.0000$). In applying the tests, one typically would test the null model first by the χ^2-test above to see if there is any phylogenetic information left in the sequences. If the null model is rejected, then one proceeds to evaluate the relative statistical support for the three alternative topologies. In our case, we would conclude that there is a very strong phylogenetic signal in the sequences and, after the next step of evaluating the statistical support of the three alternative topologies, reject T_1 and T_3, and adopt T_2. This seemingly flawless protocol would lead us to confidently reject the true tree (T_1) and adopt the wrong tree (T_2).

Steel's method is limited to four OTUs, and its extension to more than four species (Steel *et al.*, 1995) is not well described for efficient computer implementation. One way to circumvent the problem is to take a heuristic approach by sampling all possible combinations of four OTUs (quartets), performing Steel's

test by calculating X^2, and checking which species are most frequently involved in tests that fail to reject the null hypothesis of no phylogenetic information. For example, with five OTUs, there are five possible combinations of four OTUs, i.e. {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {1, 3, 4, 5}, and {2, 3, 4, 5}. In general, given N OTUs, the number of possible quartets are

$$N_{quartet} = \frac{N(N-1)(N-2)(N-3)}{4 \times 3 \times 2 \times 1} \tag{20.8}$$

For each quartet, we apply Steel's method to perform the χ^2-test. If the null hypothesis is rejected in all five tests, then we may conclude that the sequences have experienced little substitution saturation. On the other hand, if some tests fail to reject the null hypothesis, then one can check the OTU combination in the quartet and which OTU is involved most often in such cases. An OTU that is involved in a large number of tests that fail to reject the null hypothesis (designated as $N_{\text{insignificant}}$) may be intuitively interpreted as one with little phylogenetic information useful for resolving the phylogenetic relationships among the ingroup OTUs. $N_{\text{insignificant}}$ does not include tests with $c \leq 15$ because, with a small c, the failure to reject the null hypothesis is not due to substitution saturation but is instead due to lack of sequence variation. However, it is important to keep in mind that such intuitive interpretation may be misleading given the long-branch attraction problem outlined above.

DAMBE generates two indices to rank the OTUs. The first is $N_{\text{insignificant}}$. The second is

$$\phi = \sqrt{\frac{\chi^2}{c}} \tag{20.9}$$

which is often used in contingency table analysis as a measure of the strength of association. The value of ϕ ranges from 0 to 1 in contingency table analysis, but can be larger than 1 in a goodness-of-fit test when χ^2 is calculated according to (20.6). However, the scaling with c renders ϕ relatively independent of the number of informative sites and consequently more appropriate for inter-quartet comparisons. With five OTUs, each OTU is involved in four quartets and associated with four ϕ values. The mean of the four φ values for an OTU should be correlated with phylogenetic information of the OTU.

The interpretation of both $N_{\text{insignificant}}$ and ϕ are problematic with the long-branch attraction problem mentioned above. For illustration, suppose we have four sequences that have experienced substitution saturation (designated as group 1 sequences) and four sequences that are conserved with few substitutions among them (designated as group 2 sequences). Define a bad quartet as the combination of two group 1 sequences and two group 2 sequences, i.e. where long-branch attraction will happen. Such a bad quartet will always generate a large χ^2 and ϕ.

The total number of bad quartets is 36 out of a total of 70 possible quartets in this fictitious example. This means that group 1 sequences will often be involved in tests rejecting the null hypothesis with a large χ^2 and ϕ and be interpreted as containing significant phylogenetic information according to the two indices. On the other hand, if there are eight group 1 sequences and eight group 2 sequences, then a large number of quartets will be made of group 1 sequences only to allow substitution saturation among group 1 sequences to be revealed. Based on my own unpublished computer simulation, the indices are useful when there are more group 1 sequences than group 2 sequences or when there are at least four group 1 sequences.

20.3 Xia's method: its problem, limitation, and implementation in DAMBE

Xia's method (Xia *et al.*, 2003) is based on the concept of entropy in information theory. For a set of N aligned sequences of length L, the entropy at site i is

$$H_i = -\left(\sum_{j=1}^{4} p_j \log_2 p_j\right) \tag{20.10}$$

where j = 1, 2, 3, 4 corresponding to nucleotide A, C, G and T, and p_j is the proportion of nucleotide j at site i. 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 simply

$$\overline{H} = \frac{\sum_{i=1}^{L} H_i}{L}, \quad Var(H) = \frac{\sum_{i=1}^{L} (H_i - \overline{H})^2}{L-1} \tag{20.11}$$

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. 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:

$$H_{FSS} = -\left(\sum_{N_A=0}^{N}\sum_{N_C=0}^{N}\sum_{N_G=0}^{N}\sum_{N_T=0}^{N} \frac{N!}{N_A!N_C!N_G!N_T!} P_A^{N_A} P_C^{N_C} P_G^{N_G} P_T^{N_T} \sum_{j=1}^{4} p_j \log_2 p_j\right) \tag{20.12}$$

$$Var(H_{FSS}) = \sum_{N_A=0}^{N}\sum_{N_C=0}^{N}\sum_{N_G=0}^{N}\sum_{N_T=0}^{N} \frac{N!}{N_A!N_C!N_G!N_T!} P_A^{N_A} P_C^{N_C} P_G^{N_G} P_T^{N_T} \left(\sum_{j=1}^{4} p_j \log_2 p_j - H_{FSS}\right)^2 \tag{20.13}$$

P1: SJT
9780521877107c20 CUUK273-Lemey ISBN: 978 0 521 87710 7 Top: 0.47916in Gutter: 0.75in August 4, 2008 12:12

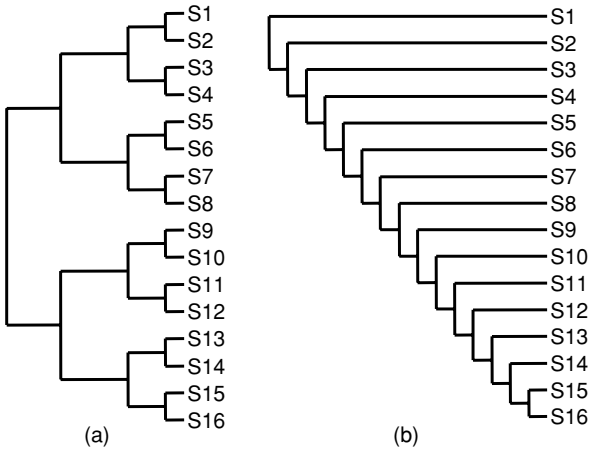

Fig. 20.2 Two extreme topologies used in simulation. (a) – symmetrical, (b) – asymmetrical.

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 $p_j = N_j/N$. 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 $\overline{H}$ value in (20.11) is significantly smaller than H_{FSS}. If $\overline{H}$ is not significantly smaller than H_{FSS}, then the sequences have experienced severe substitution saturation. This leads to a simple index of substitution saturation is defined as

$$I_{ss} = \overline{H}/H_{FSS} \tag{20.14}$$

We can see intuitively that the sequences must have experienced severe substitution saturation when I_{ss} approaches 1, i.e. when $\overline{H}$ equals H_{FSS}. However, the test of $\overline{H} = H_{FSS}$ is only theoretically useful because the sequences often 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.

Computer simulation (Xia *et al.*, 2003) suggests that $I_{ss.c}$ depends on N, L, and the topology, with larger $I_{ss.c}$ associated with more symmetrical topologies (Fig. 20.2). The ability of phylogenetic methods in recovering the true tree decreases with the degree of substitution saturation, but the effect of substitution saturation

P1: SJT
9780521877107c20 CUUK273-Lemey ISBN: 978 0 521 87710 7 Top: 0.47916in Gutter: 0.75in August 4, 2008 12:12

is alleviated by increasing L. The relation between P_{true} (the probability of the true tree being recovered) and the tree length (TL) appears to be sufficiently described by the following equation:

$$P_{true} = 1 - e^{-e^{(a-TL)/b}} \tag{20.15}$$

which was graphically plotted for various combinations of N and L and for symmetric and asymmetric topologies (Xia *et al.*, 2003). The last term in (20.15), with an exponential of an exponential, is a special form of the ***extreme value distribution (EVD)*** or Gumbel distribution (Gumbel, 1958). For the symmetrical topology, the fit of the equation to the data is almost perfect in all cases, with r^2 values greater than 0.965.

Defining the critical tree length (TL_c) as the tree length when $P_{true} = 0.95$, $I_{ss.c}$ is naturally the I_{ss} value corresponding to TL_c. When an observed I_{ss} value is significantly smaller than $I_{ss.c}$, we are confident that substitution saturation is not serious. This will be illustrated later with the elongation factor-1α sequences.

The computer simulation in Xia *et al.* (2003) is limited to $N \leq 32$. Because the $I_{ss.c}$ is based on simulation result, there is a problem with more than 32 species. To circumvent this problem, DAMBE will randomly sample subsets of 4, 8, 16 and 32 OTUs multiple times and perform the test for each subset to see if substitution saturation exists for these subsets of sequences.

PRACTICE

Xuhua Xia and Philippe Lemey

The results in this section are generated with DAMBE version 4.5.56, which better reflects the content in this chapter than previous versions. The new version of DAMBE can be found at `http://dambe.bio.uottawa.ca/dambe.asp`. The installation of DAMBE involves only a few mouse clicks.

Three sets of sequences will be used for practice: (1) 8 aligned cytochrome oxidase subunit I (COI) sequences from vertebrate mitochondrial genomes in the VertebrateMtCOI.FAS file, (2) 16 aligned EF-1α sequences (Xia *et al.*, 2003) from major arthropod groups and putative outgroups in the InvertebrateEF1a.FAS file, and (3) 41 aligned simian immunodeficiency virus (SIV) genomes, restricted to a single, nonoverlapping reading frame for the coding genes that could be unambiguously aligned, in the SIV.fas file (Paraskevis *et al.*, 2003). These files come with DAMBE installation and can be found at the DAMBE installation directory (C:\Program Files\DAMBE by default) or they can be downloaded from www.thephylogenetichandbook.org.

Start DAMBE and Click "Tools|Options" to set the default input and output directories to the directory where you have downloaded and saved these files. Set the default input file format to the FASTA format (DAMBE can read and convert sequence files in almost all commonly used sequence formats).

20.4 Working with the VertebrateMtCOI.FAS file

The VertebrateMtCOI.FAS file contains the mitochondrial COI sequences from *Masturus lanceolatus* (sunfish), *Homo sapiens* (human), *Bos taurus* (cow), *Balaenoptera musculus* (blue whale), *Pongo pygmaeus* (Bornean orangutans), *Pan troglodytes* (chimpanzee), *Gallus gallus* (chicken), and *Alligator mississippiensis* (American alligator). The third codon position will be analyzed separately from the other two codon positions.

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

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.

(1) Click 'File|Open standard sequence file' to open VertebrateMtCOI.FAS. When prompted for sequence type, choose "Protein-coding Nuc. Seq", select "VertMtDNA (Trans_Table = 2)" in the dropdown box (DAMBE has implemented all known genetic codes), and click the "Go!" button. The sequences will be displayed, with identical sites indicated by a "*".

(2) Click "Sequence|Work on codon position 1 and 2". Codon positions 1 and 2 are highly conserved, with many "*'s" below the sequences indicating many identical sites.

(3) Click "Seq.Analysis|Measure substitution saturation|Test by Xia *et al.*" A dialog appears where you can specify the proportion of invariant sites (P_{inv}) with the default being 0. P_{inv} is important for sequences with very different substitution rates over sites. For example, the first and second codon positions of functionally important protein-coding genes are often nearly invariant relative to the third codon position. The effect of substitution saturation at highly variable third codon positions may therefore go unrecognized without specifying P_{inv} because, at nearly two thirds of the sites, hardly any substitutions may be observed. In DAMBE, one can estimate P_{inv} by clicking "Seq.Analysis|Substitution rates over sites|Estimate proportion of invariant sites". So, "cancel" the test for the moment, and estimate P_{inv} using this analysis option. By specifying to "Use a new tree", a window appears providing a choice of tree-building algorithm and options. Choose the Neighbor-Joining algorithm, keep the default settings, click "Run" and then "Go!". At the end of the text output, the estimated P_{inv} is shown (P(invariate) = 0.73769). So, go back to the Test by Xia *et al.* and enter "0.74" as proportion of invariant sites. Clicking "Go!" results in the following text output:

```
Part I. For a symmetrical tree.
===============================================
Prop. invar. sites                0.7400
Mean H                            0.5550
Standard Error                    0.0298
Hmax                              1.6642
Iss                               0.3335
Iss.c                             0.7873
T                                15.2523
DF                              261
Prob (Two-tailed)                 0.0000
95% Lower Limit                   0.2749
95% Upper Limit                   0.3920
```

P1: SJT
9780521877107c20 CUUK273-Lemey ISBN: 978 0 521 87710 7 Top: 0.47916in Gutter: 0.75in August 4, 2008 12:12


```
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
============================================
Iss.c                              0.6817
T                                 11.7056
DF                               261
Prob (Two-tailed)                  0.0000

95% Lower Limit                    0.2749
95% Upper Limit                    0.3920
```

In this example, we obtain $I_{ss} = 0.3335$, much smaller than $I_{ss.c}$ (= 0.7873 assuming a symmetrical topology and 0.6817 assuming an asymmetrical topology). The sequences obviously have experienced little substitution saturation.

(4) We will build a tree to serve as a reference against the tree built with the third codon position. Click "Phylogenetics|Maximum likelihood|Nucleotide sequence|DNAML" and have a look at the options that you can specify. Click the "Run" button and you will see the tree topology shown in Fig. 20.3. You may use distance-based methods such as the ***neighbor-joining*** (Saitou & Nei, 1987), *Fitch-Margoliash* (Fitch & Margoliash, 1967) or *FastME* method (Desper & Gascuel, 2002; Desper & Gascuel, 2004) implemented in DAMBE and generate exactly the same topology with virtually any genetic distances. To obtain a distance-based tree from aligned nucleotide sequences with DAMBE, click "Phylogenetics|distance method|nucleotide sequence", optionally set of the options, and click the "Run" button.

(5) Click "Sequence|Restore sequences" to restore the sequences to its original form with all three codon positions (or just re-open the file). Click "Sequence|Work on codon position 3". The 3rd codon positions in vertebrate mitochondrial genes evolve extremely fast (Xia *et al.*, 1996).

(6) Click "Seq.Analysis|Measure substitution saturation|Test by Xia et al.". For the 3rd codon position, P_{inv} can be left at its default value since

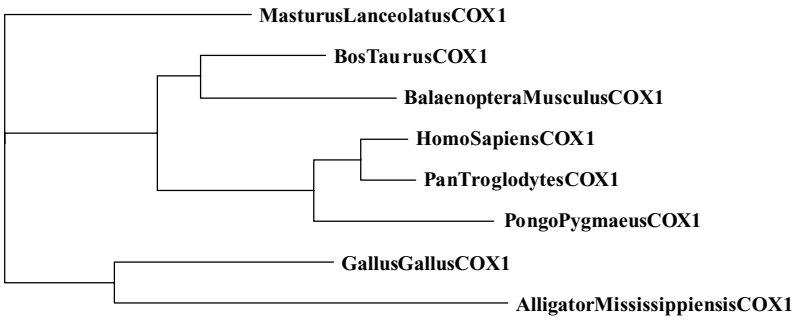

Fig. 20.3 Maximum likelihood tree from DNAML based on the first and second codon positions of the vertebrate mitochondrial COI sequences.

there are very few invariable sites. The resulting $I_{ss} = 0.7145$ is only marginally smaller than $I_{ss.c}$ (= 0.7518) assuming a symmetrical topology and substantially larger than $I_{ss.c}$ (= 0.6401) assuming an asymmetrical topology. This means that the sequences may still be useful if the true topology is not very asymmetrical. To verify this, Click "`Phylogenetics|Maximum likelihood|Nucleotide sequence|DNAML`" and click the "`Run`" button. The resulting tree has exactly the same topology as in Fig. 20.3. Note that, for this set of sequences consisting of the 3rd codon positions only, genetic distances based on substitution models more complicated than K80 (Kimura, 1980) cannot be calculated for all pairs of OTUs. For example, TN93 distance (Tamura & Nei, 1993) cannot be computed if $Q/(2\pi_Y) + P_1\pi_Y/(2\pi_T\pi_C) \geq 1$, or $Q/(2\pi_R\pi_Y) \geq 1$, or $Q/(2\pi_R) + P_2\pi_R/(2\pi_A\pi_G) \geq 1$, where Q, P_1 and P_2 are proportions of transversions, T↔C transitions and A↔G transitions, respectively and π_Y, π_R, π_T, π_C, π_A, and π_G are the frequencies of pyrimidines, purines, T, C, A and G, respectively. This highlights one of the limitations for distance-based methods involving highly diverged sequences. For the ***p-distance***, the Poisson-corrected *p*-distance and K80-distance that can still be calculated with this set of sequences, only the ***UPGMA*** method, but not the neighbor-joining (Saitou & Nei, 1987), Fitch-Margoliash (Fitch & Margoliash, 1967) or FastME method (Desper & Gascuel 2002; Desper & Gascuel 2004), will recover the topology in Fig. 20.3. Interestingly, the maximum parsimony tree from this set of sequences is the same as that in Fig. 20.3. It is therefore important to keep in mind that establishing the existence of phylogenetic information in the sequences does not mean that the true tree can be recovered by any tree-building algorithm, and that we still know poorly as to which method is better than others.

(7) To perform a test using Steel's method, click "`Seq.Analysis|Measure substitution saturation|Test by Steel et al.`" and click the "`OK`" button. The output is in three parts. The first is the nucleotide frequencies. The second is the output for individual test of each quartet. The third part shows which OTU might be problematic:

```
Sequences ranked from the best to the worst.
================================================
Seq_Name                    Mean_Phi   Num_Insignif
------------------------------------------------------------
PanTroglodytesCOX1            0,1584             10
HomoSapiensCOX1               0,1495             12
PongoPygmaeusCOX1             0,1262             12
GallusGallusCOX1              0,1131             20
AlligatorMississippiensis     0,1124             20
BosTaurusCOX1                 0,1096             20
BalaenopteraMusculusCOX1      0,1085             20
MasturusLanceolatusCOX1       0,0897             22
================================================
Num_Insignif conditional on c > 15.
```

The output for the third codon positions only shows that *Masturus lanceolatus* (sunfish) has the smallest mean ϕ value and is involved in the largest number of tests that fail to reject the null hypothesis of no phylogenetic information, indicating that it might be too divergent from the rest of the OTUs. One may be interested to learn that the JC69 distances between *M. lanceolatus* and other OTUs are all greater than 1. This reminds us of the suggestion (Nei & Kumar, 2000, p. 112) to avoid such sequences.

20.5 Working with the InvertebrateEF1a.FAS file

The elongation factor-1α (EF-1α) 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 conserved (Walldorf & Hovemann, 1990), and consequently has been used frequently in resolving deep-branching phylogenies (Cho *et al.*, 1995; Baldauf *et al.*, 1996; Regier & Shultz, 1997; Friedlander *et al.*, 1998; Lopez *et al.*, 1999).

The InvertebrateEF1a.FAS file contains the EF-1α from four chelicerate species (CheU90045, CheU90052, CheU90047, CheU90048), four myriapod species (MyrU90055, MyrU90053, MyrU90057, MyrU90049), two branchiopod species (BraASEF1A, BraU90058), two hexapod species (HexU90054, HexU90059), one molluscan species (MolU90062), one annelid species (AnnU90063) and two malacostracan species (MalU90046, MalU90050). The phylogenetic relationship among major arthoropod taxa remains controversial (Regier & Shultz 1997).

(1) Click "`File|Open standard sequence file`" to open InvertebrateEF1a.FAS as before, choose the default "`standard`" genetic code and click the "`Go!`" button.
(2) Click "`Sequence|Work on codon position 3`".
(3) Click "`Seq.Analysis|Measure substitution saturation|Test by Xia et al.`" P_{inv} can be left at its default value. The resulting $I_{ss} = 0.6636$, not significantly ($p = 0.1300$) smaller than $I_{ss.c}$ ($= 0.7026$) assuming a symmetrical topology and substantially larger than Iss.c ($= 0.4890$) assuming an asymmetrical topology. This means that the sequences consisting of 3rd codon positions only have experienced so much substitution saturation that they are no longer useful in phylogenetic reconstruction. To verify this, click "`Phylogenetics|Maximum likelihood|Nucleotide sequence|DNAML`" and click the "`Run`" button. The resulting tree, with no consistent clustering of EF1-α from the same species, is absurd and totally different from the tree one would obtain by using the 1st and 2nd codon positions.

(4) To apply Steel's method to the analysis of the 3rd codon positions, click "`Seq.Analysis|Measure substitution saturation|Test by Steel et al`" and click the "`OK`" button. The last part of the output show the mean ϕ values ranging from 0.028 to 0.0376, in dramatic contrast to the mean ϕ values for the 3rd codon position of the mitochondrial COI gene (between 0.0897 and 0.1584). An OTU with a mean ϕ value smaller than 0.04 may be taken as lacking phylogenetic information based on computer simulation. The mean ϕ values range from 0.0774 to 0.1061 when Steel's method is applied to the 1st and 2nd codon positions of the EF-1α sequences, but range from 0.2296 to 0.3133 when applied to the 1st and 2nd codon positions of the vertebrate mitochondrial COI gene. In short, all indications suggest that the set of invertebrate EF-1α sequences have experienced much greater substitution saturation than the set of vertebrate mitochondrial COI sequences.

20.6 Working with the SIV.FAS file

The test with Xia's method involving more than 32 OTUs is different from those with 32 or fewer OTUs, and is illustrated with this set of 41 SIV sequences obtained from various African primates.

(1) Click "File|Open standard sequence file" to open SIV.FAS as before. Since this file has unresolved bases, a window pops up asking you how to deal with them. DAMBE presents three options for dealing with ambiguous codes. The first is to explicitly mark them as unresolved. The second will treat them in a probabilistic manner depending on what computation is involved. Take R (coding for either A or G) for example: if 80% of the purines in the input sequences are As, then a R is counted as 0.8 A and 0.2 G in computing frequencies. In computing nucleotide substitutions, such a R facing a homologous A on another sequence will be treated as identical with a probability of 0.8 and a transition with a probability of 0.2. The final option keeps the ambiguities in the sequences. Choose option 2 by entering "`2`" and clicking the "`Go!`" button.

(2) Click "`Seq.Analysis|Measure substitution saturation|Test by Xia et al.`" and set P_{inv} to "`0.17`" before performing the analysis. The output table shows that the average I_{ss} for subsets of 4, 8, 16 and 32 are significantly smaller than the corresponding $I_{ss.c}$ if the true topology is symmetrical:

NumOTU	Iss	Iss.cSym	T	DF	P Iss.	cAsym	T	DF	P
4	0.573	0.850	35.922	5001	0.0000	0.845	35.283	5001	0.0000
8	0.558	0.847	36.663	5001	0.0000	0.767	26.534	5001	0.0000
16	0.575	0.832	33.524	5001	0.0000	0.680	13.686	5001	0.0000
32	0.576	0.814	31.387	5001	0.0000	0.568	1.058	5001	0.2902

Note: two-tailed tests are used.

While substitution saturation becomes a problem when the true topology is extremely asymmetrical and when the number of OTUs (N) is greater than 16 (e.g. $P = 0.2902$ for $N = 32$), such asymmetrical trees are probably not realistic for these SIV sequences. We can conclude that there is still sufficient phylogenetic information in the complete SIV data set. However, analyzing only the 3rd codon position (keeping the $P_{inv} = 0$) will reveal that the average I_{ss} values are already considerably higher.

Acknowledgment

I thank Stephane Aris-Brosou, Pinchao Ma, and Huiling Xiong for comments, and NSERC Discovery, RTI and Strategic grants for financial support.