Ziheng Yang
14 March 2010

# Answers to selected exercises in Yang (2006)

Many of the questions in the book ask you to confirm certain results so that the answers are written in the questions. Most of such questions are not dealt with here. The answers below are possibilities, as the answers may not be unique. As in the book, I added verbal explanations around the mathematical derivations, so please excuse the verbosity.

## Chapter 2

Equation 2.19 (page 56)

This equation states that $d_S$ is the average mutation/substitution rate before selection acts on the protein, averaged over the three codon positions. This equation underlies my claim that the often made statement that the use of $d_N/d_S$ to detect selection on the protein requires the assumption of neutral evolution of synonymous mutations is incorrect. For some discussions of this issue, see Yang and Nielsen (2008 Mol. Biol. Evol. 25: 568-579: right column on page 576) and a post at the PAML discussion page: http://gsf.gc.ucdavis.edu/viewtopic.php?f=1&t=201&hilit=neutral+synonymous&sid=403aef5fd1dac0c6abedd0afb777b63d

The reader is invited to confirm the equation in the book. Below is a possible proof. All the symbols are defined in the book.

$$(d_{1B} + d_{2B} + d_{3B})/3 = t\sum_{i\neq j} \pi_i q_{ij}^1 \Big/ 3.$$

$$\rho_S^1 = \sum_{i\neq j,\ \mathrm{aa}_i = \mathrm{aa}_j} \pi_i q_{ij}^1 \Big/ \sum_{i\neq j} \pi_i q_{ij}^1,$$
$$\rho_N^1 = \sum_{i\neq j,\ \mathrm{aa}_i \neq \mathrm{aa}_j} \pi_i q_{ij}^1 \Big/ \sum_{i\neq j} \pi_i q_{ij}^1.$$

Note that $\rho_S^1$ and $\rho_N^1$ are the proportions of synonymous and nonsynonymous sites and the scaling above is to ensure that they sum to 1. (The proportions of substitutions $\rho_S$ and $\rho_N$ in eq 2.14 sum to 1, by construction of the codon model.)

$$d_S = S_d/S = t\rho_S/(3\rho_S^1) = \frac{t\sum_{i\neq j,\ \mathrm{aa}_i = \mathrm{aa}_j} \pi_i q_{ij}}{3\sum_{i\neq j,\ \mathrm{aa}_i = \mathrm{aa}_j} \pi_i q_{ij}^1 \Big/ \sum_{i\neq j} \pi_i q_{ij}^1} = t\sum_{i\neq j} \pi_i q_{ij}^1 \Big/ 3,$$

using the result that $\sum_{i\neq j,\ \mathrm{aa}_i = \mathrm{aa}_j} \pi_i q_{ij} = \sum_{i\neq j,\ \mathrm{aa}_i = \mathrm{aa}_j} \pi_i q_{ij}^1$.

□

## Chapter 4

4.3 Calculate the probabilities of sites with data *xxyy*, *xyyx*, and *xyxy* in four species for the unrooted tree of figure 4.13, using two branch lengths $p$ and $q$ under a symmetrical substitution model for binary characters (Exercise 1.3). Here it is more convenient to define the branch length as the proportion of different sites at the two ends of the branch. Show that $\Pr(xxyy) < \Pr(xyxy)$ if and only if $q(1-q) < p^2$. With such branch lengths, parsimony for tree reconstruction is inconsistent (Felsenstein 1978a).

**Solution.** The calculation of the probability of a site pattern is described in §4.2.1. In this simple example the tree is small so that there is no need to apply the pruning algorithm of §4.2.2. Note that with the JC69-like symmetrical substitution model for binary characters (see equation 1.72), the matrix of transitions probabilities for a branch of length $q$ (where $q$ is the proportion of different sites) is

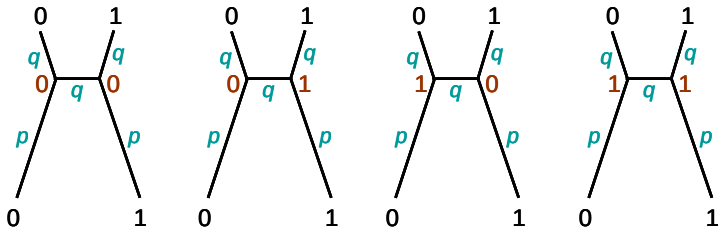
$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}. \qquad (1)$$

Similarly, the transition probability for a branch of length $p$ is $p$ for a difference or $1-p$ for an identity. Also when the number of changes per site goes from 0 to ∞, the proportion of differences goes from 0 to ½, so that $0 < p, q < ½$ (see equation 1.72).

For four species, there are $2^4 = 16$ site patterns, and some of them have identical probabilities of occurrence due to the symmetry of the problem. Thus

$$\Pr(xxyy) = \Pr(0011) + \Pr(1100) = 2 \times \Pr(0011).$$
$$\Pr(xyxy) = \Pr(0101) + \Pr(1010) = 2 \times \Pr(0101).$$

Consider Pr(0011). We have to average over four possible character configurations at the two ancestors (or ancestral reconstructions): 00, 01, 10, and 11, as shown in the figure below.

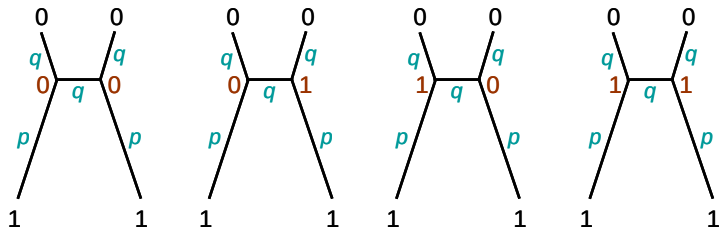



Place the root at the left ancestral node, say. The character at the root is 0 or 1, each with probability ½. Given the state at the root, the evolutionary process of the character on the tree is described by five transition probabilities along the five branches. Thus

$$\begin{aligned} \Pr(0011) = {} & \tfrac{1}{2}(1-q)(1-q)q(1-p)p \\ & + \tfrac{1}{2}(1-q)q(1-q)(1-p)(1-p) \\ & + \tfrac{1}{2}qqqpp \\ & + \tfrac{1}{2}q(1-q)(1-q)p(1-p). \end{aligned} \qquad (2)$$

The probability is a sum of four terms, corresponding to the four trees above. Each term is the product of ½, the prior probability for the root state, times five transition probabilities for the five branches. I have written them in the order top left, middle, top right, bottom left, and bottom right.

Similarly, the probability for the site pattern 0101 is a sum over four ancestral reconstructions: 00, 01, 10, and 11, as shown in the figure below.




Thus

$$\begin{aligned}\Pr(0101) = {} & \tfrac{1}{2}(1-q)(1-q)(1-q)\,pp \\ & + \tfrac{1}{2}(1-q)qqp(1-p) \\ & + \tfrac{1}{2}qq(1-q)(1-p)\,p \\ & + \tfrac{1}{2}q(1-q)q(1-p)(1-p).\end{aligned} \qquad (3)$$

We now have, after some simplifications (which I was lazy enough to use Mathematica to do)

$$\Pr(xxyy) - \Pr(xyxy) = 2\Pr(0011) - 2\Pr(0101) = (2q-1)(p^2 - q + q^2). \qquad (4)$$

Because $q < ½$, we have $\Pr(xxyy) < \Pr(xyxy)$ if and only if $q(1-q) < p^2$.

**Notes**

- You can use the same strategy to calculate $\Pr(xyyx) = 2\Pr(0110)$, but it is $< \Pr(xyxy)$ and so won't affect our analysis of the performance of parsimony.
- Felsenstein (1978 Syst. Zool. 27:401-410, equation 11) used the following reasoning to calculate $\Pr(xxyy)$. Having 0 or 1 at the root won't affect $\Pr(xxyy) = \Pr(0011) + \Pr(1100)$ although it affects the relative contributions of the two terms. Thus one may fix the root state (at the left ancestor) at 0 and consider the two possible states (0 and 1) at the right ancestor, with two terms for Pr(0011) and two terms for Pr(1100), and four terms in total. The result is of course the same.

## Chapter 5

*5.3 Suppose the target density is $N(\theta, 1)$, and the MCMC uses the sliding-window proposal with normal proposals, with the jump kernel $x^* \sim N(x, \sigma^2)$. Show that the acceptance proportion (the proportion at which the proposals are accepted) is (Gelman *et al.* 1996)

$$P_{\text{jump}} = \tfrac{2}{\pi}\tan^{-1}\left(\tfrac{2}{\sigma}\right). \qquad (5.48)$$

**Solution.** The true value of the location parameter $\theta$ is arbitrary, so without loss of generality we fix it at 0, and the target density is

$$\phi(x) = \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}. \qquad (5)$$

Suppose the current value is $x$, and the new proposed value is $y$. The proposal density is

$$q(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-x)^2\right\}. \qquad (6)$$

There are two ways by which a proposal $y$ is accepted. First, if $y$ has a greater height, $\phi(y) > \phi(x)$, it is accepted straightaway. Second, if $y$ has a smaller height, it is accepted with probability $\phi(y)/\phi(x)$. The acceptance proportion is a sum over those two situations

$$\begin{aligned} P_{\text{jump}} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(x)q(y \mid x)\,\mathrm{I}_{\phi(y)>\phi(x)}\,\mathrm{d}x\,\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(x)q(y \mid x)\,\mathrm{I}_{\phi(y)\le\phi(x)}\,\frac{\phi(y)}{\phi(x)}\mathrm{d}x\,\mathrm{d}y \\ &= I_1 + I_2. \end{aligned} \qquad (7)$$

Here the indicator $\mathrm{I}_A = 1$ if $A$ is true and 0 otherwise. The first integral $I_1$ is the probability of accepting the proposal when $\phi(y) > \phi(x)$ and $I_2$ is the probability when $\phi(y)/\phi(x) < 1$. By switching symbols $x$ and $y$ in $I_2$, it is clear that $I_2 = I_1$. (This equality is due to the fact that the Markov chain generated by the MCMC algorithm is time-reversible and holds for any target density and for any Metropolis-Hastings algorithm.)

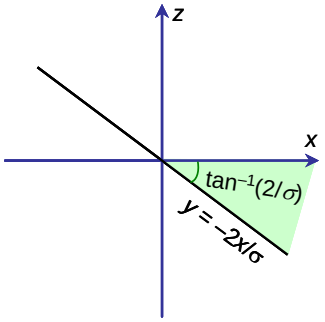
$$P_{\text{jump}} = 2I_1 = 2\iint_{\phi(y)>\phi(x)}\phi(x)\,f(y \mid x)\,\mathrm{d}y\,\mathrm{d}x = 2\times 2\int_0^{\infty}\int_{-x}^{x}\tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\tfrac{1}{2\sigma^2}(y-x)^2\right\}\mathrm{d}y\,\mathrm{d}x. \qquad (8)$$

Note that the integrand is the density of a bivariate normal distribution centered at the origin. Because of the symmetry of the density, the volume (or integral) left of the $y$ axis is the same as that on the right, so the integral is twice that on the right (where $x$ goes from 0 to $\infty$). Furthermore the region of integration $\phi(y) > \phi(x)$ is equivalent to $|y| < |x|$ or $-x < y < x$ when $x > 0$.

Change variables from $y$ to $z = (y - x)/\sigma$, with $\mathrm{d}z = \mathrm{d}y/\sigma$.

$$P_{\text{jump}} = 4\int_0^{\infty}\int_{-2x/\sigma}^{0} \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2} \tfrac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz\, dx\,. \qquad (9)$$

Here the integrand is the density of a standard bivariate normal distribution, so that the volume over the whole $x$-$z$ plane is 1. The region of integration is shown in the graph on the right. The line $y = -2x/\sigma$ has slope $-2/\sigma$ and forms an angle $\tan^{-1}(2/\sigma)$ with the $x$ axis. The integral in (9) is thus equal to the angle in the graph divided by $2\pi$ (the angle for the whole plane).




$$P_{\text{jump}} = 4 \times \frac{\tan^{-1}(2/\sigma)}{2\pi} = \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right). \qquad (10)$$

Alternatively changing variables from the $x$-$z$ Cartisian system in equation (9) to the polar system ($x = r\cos\alpha$, $z = r\sin\alpha$) gives the same result.

5.4 Write a program to implement the MCMC algorithm of subsection §5.3.2 to estimate the distance between the human and orangtutan 12S rRNA genes under the JC69 model. Use any programming language of your choice, such as BASIC, Fortran, C/C++, Java, or Mathematica. Investigate how the acceptance proportion changes with the window size $w$. Also implement the proposal of equation (5.34). (Hint: use the logarithms of the likelihood and prior in the algorithm to avoid numerical problems.)

A C program named MCMCJC2s.c is in the data archive, posted at the book web site.

## Chapter 9

9.1 Write a small simulation program to study the *birthday problem*. Suppose that there are 365 days in a year and that one's birthday falls on any day at random. Calculate the probability that at least two people out of a group of $k = 30$ people have the same birthday (that is, they were born on the same day and month but not necessarily in the same year). Use the following algorithm. (The answer is 0.706.)

1. Generate $k = 30$ birthdays, by taking 30 random draws from 1, 2, …, 365.
2. Check whether any two birthdays are the same.
3. Repeat the process $10^6$ times and calculate the proportion of times in which two out of 30 people have the same birthday.

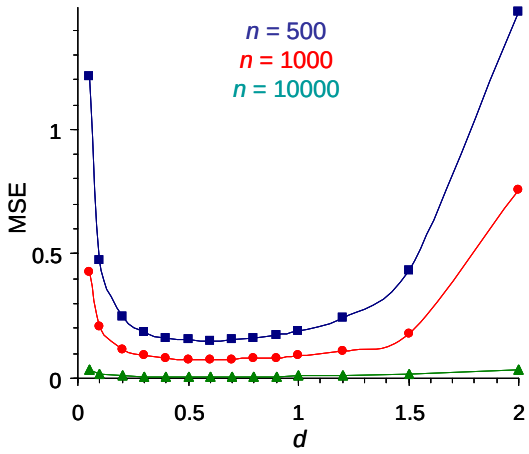
A C program named birthday.c is in the data archive, posted at the book web site.

9.1. Long-branch attraction by parsimony. Use the JC69 model to simulate data sets on a tree of four species (fig. 9.3a), with two different branch lengths $a = 0.1$ and $b = 0.5$. Simulate 1000 replicate data sets. For each data set, count the sites with the three site patterns *xxyy*, *xyxy*, *xyyx*, and determine the most parsimonious tree. To simulate a data set, reroot the tree at an interior node, as in, say, figure 9.3b. Generate a sequence for the root (node 0) by random sampling of the four nucleotides, and then evolve the sequence along the five branches of the tree. You may also use the approach of multinomial sampling. Consider a few sequence lengths, such as 100, 1000, and 10,000 sites.

If you don't want to write your own simulation program, you can use an existing program such as evolver in paml or seq-gen to generate datasets.

9.2. Write a small simulation program to study the optimal sequence divergence when two sequences are compared to estimate the transition/transversion rate ratio $\kappa$ under the K80 model. Assume $\kappa = 2$ and use a sequence length of 500 sites. Consider several sequence distances, say, $d = 0.01, 0.02, \ldots, 2$. For each $d$, simulate 1000 replicate data sets under the K80 model and analyze it under the same model to estimate $d$ and $\kappa$ using equation (1.11). Calculate the mean and variance of the estimate $\hat{\kappa}$ across replicate data sets. Each data set consists of a pair of sequences, which can be generated using any of the three approaches discussed in subsection §9.5.1.

The results should be like those in the graph. The MSE is large for both large and small $d$s, indicating that it is hard to estimate $\kappa$ reliably with both very divergent and very similar sequences. For all three sequence lengths examined here, the optimal sequence divergence is at about 0.6, although the MSE is small for the range $0.3 < d < 1$. Such optimal distances seem to be larger than most biologists think. Also note that when the sequence length ($n$) is large and the datasets informative, the MSE is approximately proportional to $1/n$, so that the MSE for $L = 1000$ is about 10 times as large as that for $n = 10000$.



In large datasets, the bias is nearly 0, and the variance is proportional to $1/n$.

There are a variety of ways of conducting this simulation. The method of using the exponential waiting times is time-consuming for long sequences and large distances. The method of generating one sequence and then using the transition probability matrix to evolve the other is faster. An even faster method is to sample from the trinomial distribution with the three categories corresponding to the following three site patterns: sites that are identical between the sequences, sites with a transitional difference, and sites with a transversional difference.