Molecular Phylogenetics

Problems 5 2016

**Maximum likelihood phylogenetics #1**

***A primer.***

Here’s how to obtain a probability matrix from a model for any branch length, relative rate matrix, and frequency vector. We’ll use an arbitrary version of the GTR model as an example, which we’ll call JT14. This is an admittedly absurd model, but it should illustrate the math pretty well.

1. To begin, the model consists of a relative rate matrix **R** and a vector of frequencies = {A, C,G,T,}. **R** expresses the relative rate rij at which nucleotides change from any of the four nucleotides (i) to any of the others (j). For our example, let **R**=

- 1 2 3

1 – 4 5

2 4 – 6

3 5 6 -

This is a somewhat silly matrix, but it keeps things clear for teaching purposes. Note it is symmetrical. As you’ll see below, the absolute value of the entries does not matter; the relative rates are what is important.

1. We need to use this information to find the **Q** matrix, which contains the instantaneous rates of change, scaled properly so that it yields total substitution probabilities equal to those predicted by the branch lengths. First, we find the unscaled Q matrix: we multiply **R** by the diagonal vector of state frequencies (that is, we multiply each column by the frequency corresponding to j , the frequency of that column) and then fill the diagonals so that each row sums to 0. (An effect of multiplying **R** by the frequencies is that the probabilities of a state along a very long branch converge on the frequencies. Also, having the rates sum to zero results in the sum of the probabilities in the same row derived from those rates sum to 1, which better be true, because the row represents all the possible outcomes j given the starting state i.)
2. Suppose our frequency vector is =[0.1,0.2,0.3, 0.4]. This gives us **Q\_unscaled** =

-2.0 0.2 0.6 1.2

0.1 -3.3 1.2 2.0

0.2 0.8 -3.4 2.4

0.3 1.0 1.8 -3.1

1. This matrix gives the correct relative instantaneous rates of change, but the absolute substitution probabilities that it will yield won’t correspond to the branch length. To get the proper **Q** that is correctly scaled, we need to find the scaling factor s. To do this, we 1) multiply each diagonal element times the frequency of the nucleotide for its row, and then 2) sum these four terms up and take the reciprocal to get s=-1/(iqii):

s = -1/[(-2.0\*0.1)+(-3.3\*0.2)+(-3.4\*0.3)+(-3.1\*0.4)]=

s = 1/(0.20 + 0.66 + 1.02 + 1.24) = 1/3.12 = 0.3205

1. Now multiply every element in **Q\_unscaled** by s. This gives **Q**, the weighted instantaneous rate matrix. The reason we did this is so that the use of this matrix to compute a P matrix yields a total probability of substitutions equal to the branch length. Here we get **Q**=

|  |  |  |  |
| --- | --- | --- | --- |
| -0.641 | 0.064 | 0.192 | 0.385 |
| 0.0321 | -1.058 | 0.385 | 0.641 |
| 0.0641 | 0.256 | -1.090 | 0.769 |
| 0.0962 | 0.321 | 0.577 | -0.994 |

1. For any branch length b, we can now find **Pb**, the matrix that expresses probability of each type of substitution from state *i* to *j* along a branch of length b, using **Pb**=eb**Q**. This gives the conditional probability of ending up with state *j* along this branch, given that the starting state is i. If the branch length is 0.5, we calculate **Pb**=exp(0.5\***Q**). We can compute this using R, or in Mathematica, or using Mathematica online at http://tinyurl.com/mdhvn6z. If you’re using the latter approach, the notation is as follows. Note that the branch length scalar comes first; the matrix as a whole is contained in a pair of curly brackets, with each row in a pair of curly brackets, and the elements of the row separated by commas. We get **P0.5** =

matrixexp(.5 {{-0.641, 0.064, 0.192, 0.385},{0.0321, -1.058, 0.385, 0.641},{0.0641, 0.256, -1.09, 0.769},{0.0962, 0.321, 0.577, -0.994}})

And that gives us, **P0.5**=

|  |  |  |  |
| --- | --- | --- | --- |
| 0.731 | 0.037 | 0.085 | 0.148 |
| 0.018 | 0.615 | 0.145 | 0.222 |
| 0.028 | 0.096 | 0.624 | 0.251 |
| 0.037 | 0.111 | 0.188 | 0.664 |

I’ll ask you a few questions about this matrix below, but first…

6. A few reminders.

1. The probability of a state *i* a priori, if we know nothing about where it exists on a tree or what its ancestors’ states were, is its frequency i.
2. The probability of a state *j* if we know the branch length *b* and the ancestral state *i* is the element pij in the matrix **Pb**.
3. The probability of a state *j* at the end of a branch length *b* if we don’t know the ancestral state is the probability pij times I, summed over all possible ancestral states i. This makes sense, because *j* can be realized in four different ways from each of the four possible ancestral states; the total probability that any one of the four scenarios would occur is the sum of each scenario. The probability of a scenario involving ancestral A and derived G, for example is the probability that we started with (A) times the probability that we ended up with G given A (pAG)..
4. The probability of having a state *j* at the end of one branch and k at the end of another branch which connect to the same ancestor, which is known to have state i, is pij \* pik . This makes sense, because this scenario requires two independent events to happen – the two transitions along the two independent branches – so the joint probability is the product of the two independent probabilities.
5. The JC69 model is the simplest, unbiased model, in which all rates and frequencies are the same. In this model, the probability pii of starting with state *i* and ending up with state *i* along a branch of length *b* is pii = ¼ + ¾(e-4b/3). You can figure out any pij for yourself now that you know that.
6. The probability of a sequence of sites is the product of the probability of each site.

**Homework 5 – Likelihood calculations – Molecular Phylogenetics 2014**

1. Given the simplest possible Jukes-Cantor model and a single sequence from a single species – but no tree, no sequences of relatives, no alignment -- what is the probability of the sequence AGCC in that species?
2. Using the JC model, what is the probability of going from A in the ancestor to C in the descendant along a branch of length 0.5?
3. On a branch of length 0.5 using the JC model, what is the probability of ending up with a C in the descendant if we don’t know the ancestral state?
4. Now let’s go back to the model JT14, whichwasexplained in the notes above. If you look at any row of **P0.5**, the diagonals are larger than any of the off-diagonals. Explain why this is true in terms of the parameters provided.
5. Among the diagonal elements of **P0.5**, the one at the upper left is the highest. What does this cell represent? Why does this element have the largest value of any of the diagonals?
6. Which is higher, the probability of ending up with C or with T, if you begin with G? Why is this true, in terms of the model?
7. Which is higher, the probability of ending up with C or T if you begin with A? Why is this true, in terms of the model?
8. Use whatever method you prefer to compute the probability matrix for a branch length of 50. Which of the diagonals is highest? Why is the answer different from the answer to question number 2?
9. Going back to **P0.5**, what is the probability of going from A to C along this branch? What is the probability of going from C to A? How can this be true if the model is “reversible”? Please explain why this occurs, given the parameters provided.
10. Compute the probability matrix for this same model, but when the branch length equals 0.8.
    1. Display the matrix here.
    2. What is the probability of ending up with a G if we started with a T? How does this compare to the probability of the same T->G transition if the branch length were 0.5?
    3. Consider a mini-sequence ACCT at the end of the branch with length of 0.8. If the starting sequence were AAAA, what would be the probability of ACCT at the end?
    4. If the starting sequence were CCCC, what would be the probability of ACCT at the end?
    5. Suppose we have no knowledge of the ancestral states. What is the probability of having A at the end of a branch with length 0.8?
    6. What would be the probability of having a C at the end of the branch if we have no knowledge of the ancestral states?
    7. Consider a subtree (X:0.5, Y:0.8), where X and Y are the names of the taxa and the numbers give the branch lengths from their ancestor to each tip. What is the probability of the following mini-alignment, assuming we know nothing of the ancestral states? Show your work.

X ACT

Y CAT

1. Consider a model in which the rate for any transition is twice that of an transversion, and in which the frequency of G is equal to the frequency of C but three times the frequency of A or of T (which are equal to each other).
   1. Write the frequency vector for this model. Remember that its elements must sum to 1, so you might have to do a little basic algebra in order to scale it properly.
   2. Show the relative rate matrix **R**.
   3. Show the unscaled matrix **Q\_unscaled**.
   4. Compute the scaling factor. Show your work.
   5. Show the scaled matrix **Q.**
   6. Calculate and show the probability matrix for a branch length of 0.5.
   7. Calculate and show the probability matrix for a branch length of 20.