STAT682_Summary_Report

Yuzhuo Kang

4/10/2020

Background

1.Phylogenetic inference

In order to estimate the likelihood from the genetic sequence, the first step is to build the substituion rate matrix, Q-matrix. The substituion rate is calculated based on the observed matrix of site pattern counts.

The matrix Q is parameterized by stationary distributions $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ and transition rate vector ρ. π is the observed base relative frequencies. ρ is the transition rate from one nucleobase to another nucleobase.It represents the observed relative frequencies of changed sites. For instance, ρ_{AC} is the transition rate from A to C per unit time. $\rho_{i+} = \sum_{j \neq i} \rho_{ij}$. The elements in the Q-matrix is derived by:

$$Q(\pi, \rho)_{ij} = \begin{cases} \frac{\rho_{ij}}{2\pi_i} & \text{if } i \neq j \\ \frac{\sum_{k \neq i} \rho_{ik}}{2\pi_i} & \text{if } i = j \end{cases}$$

The next step is to get the probability transition matrix, whihc is defined as $P(t) = e^{Qt}$ for a non-negative t. t represents the branch length. If Q matrix has eigendecomposition $Q = V\Lambda V^{-1}$, then the probability transition matrix over a time t is defined as $P(t) = Ve^{\lambda t}V^{-1}$. V is a matrix of eigenvalues of Q, and Λ is a diagonal matrix of the eigenvalues λ of Q. V^{-1} is the matrix inverse of V. For the $P(t)$ matrix, the sum of each row is equal to 1.

In summary, the parameters include:

- $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ where $\sum_i \pi_i = 1$
- $\rho = \{\rho_{ij}\}$ for $1 \leq i < j \leq 4$ where $\sum_{i=1}^{3} \sum_{j=i+1}^{4} \rho_{ij} = 1$ and for $i > j$, $\rho_{ij} = \rho_{ji}$

The indices $1 \ldots, 4$ with the DNA bases $\{A, C, G, T\}$. π is the stationary distribution of the continuous-time Markov chain governed by Q.ρ_{ij} represents the long-run proportion of transitions between states i and j in either direction.

Put all these free parameters in one vector: $\theta = (t, \pi_A, \pi_C, \pi_G, \rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT})$

where t>0, all other parameters between 0 and 1.

let

- $\pi_T = 1 - (\pi_A + \pi_C + \pi_G)$
- $\rho_{GT} = 1 - (\rho_{AC} + \rho_{AG} + \rho_{AT} + \rho_{CG} + \rho_{CT})$

2. Models of molecular evolution

All the free parameters in section 1 could be estimated using the Jukes-Cantor model.In the JC model, it is assumed that each base in the sequence has an equal chance of changing. If the evolution follows J-C model, the evolutionary distance (JC1969 distance) between the two sequences given the alignment could be calculated as follows:

$$d = -\frac{3}{4} ln(1 - \frac{4}{3}p)$$

where p is defined by the following equation:

$$p = \frac{3}{4} - \frac{3}{4}e^{-\mu t}$$

where μ is the transition rate of all pairs of nucleobases in the Q matrix, and here t refers to the elapsed time. p is the sum of probabilities that the site is different at the two end of the branch. The value of μt is the product of th rate of change and time. This d is used to find the MLE with all the parameters in θ.

d is the maximum likelihood estimate (MLE) of the distance. d is infinite if the difference between sequences is larger than 0.75, so this model could only be applied for finite-length sequences (Felsenstein, 2004).

3.Simulation of data sets

The purpose is to simulate DNA sequence data on a tree given a sequence length n and a nucleotide substitution model parameterized by a 4×4 time-reversible infinitesimal rate matrix Qparameters. Q matrix is parameterized by simplex vectors $\pi \in \Delta_4$ and $\rho \in \Delta_6$.

The first step is to determine the total number of internal nodes and tips. The index of the root has the smallest number among the internal nodes. The sequence of the root is generated using the stationary distribution $\pi = \{\pi_i\}$ and the total length of the sequence. The second step is to traverse through the tree and generate sequences at each child. Different edges in the tree have different lengths. The probability transition matrix $P(t)$ for each edge could be calculated with the edge length, list of parameters from time-reversible infinitesimal rate matrix Q, π, and ρ. Then the child sequence is deduced from the parent sequence at each edge. The purpose is to generate the simulated sequences at the tip of the tree. The simulation result changes a lot because it is dependent on the probability transition matrix and the root sequence is not fixed.

4.Maximum likelihood estimation

The likelihood at a specific time given the Q matrix, t branch length and sequence data is

$$L = \prod_{i=1}^{n} \pi_{x[i]} P_{x[i],y[i]}(t)$$

The total length of each sequence is n. π is the stationary distribution of Q. x and y are the two sequences which only contain A,C,G,T.

To estimate the maximized likelihood at the branch length t, the likelihood function is converted to the log form.Taking logarithms addresses transforms products of very small positive numbers to sums of moderate numbers. The log-likelihood is the sum of the logs of the likelihoods calculated over each site. The formula is

$$\sum_{i=1}^{n} log(\pi_{x[i]}) + \sum_{i=1}^{n} log(P_{x[i],y[i]})(t)$$

The optimize function could find the optimal t which maximize the log-likelihood in the given interval of branch length. When all the parameters are estimated, the optim estimation with L-BFGS-B method is used to find the optimal numeric value of each parameter $(t, \pi_A, \pi_C, \pi_G, \rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT})$that maximized the log-likelihood function.

5.Bootstrap

The bootstrap is the application to sample all nucleobases from the set of n nucleobases with replacement by n times. A data matrix with the same number of species and same number of characters as in the original data. Bootstrap is a resampling analysis that constructs the new tree and tests if the same nodes are recovered in each replication. The bootstrap value indicates whether the node is well supported after resampling and simulations. Similarly, if a fraction P of the bootstrap replicates have the branch present, the the branch is assigned with probability P. Each replication shows different construction of the tree. The resulting sample of phylogenies shows approximately the same variation as the estimated sequence by sampling n new sites for each tree (Felsenstein, 2004).

The bootstrap for the given data set is conducted using RAXML. It calculates the likelihood of each repication and saves the estimated best tree with maximized likelihood.

The bootstrap results using Raxml also contain the stationary distribution value $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ and a substitution rate vector of $r = \{r_{ij}\}$ for $1 \leq i < j \leq 4$ from the GTR model. The last element in the r vector r_{GT} is always equal to 1. The r vector is converted to $\rho = \{\rho_{ij}\}$ to build the list of Q-matrix parameters. The ρ value is derived by the following formula:

$$\rho_{ij} = \frac{2\pi_i r_{ij} \pi_j}{\mu}$$

where μ normalize the ρ_{ij} to make $\sum_{i=1}^{3} \sum_{j=i+1}^{4} \rho_{ij} = 1$.

Research Statement

The goal of this research is to investigate the relationship between central branch length and bootstrap support values of best tree topology when there are four taxas. The hypothesis is that when the central branch length increases, the bootstrap support value of the best tree topology from bootstrap estimation increases correspondingly.

Data Collection

The data set is art4.fasta. It contains the sequences of four species, Sheep, Giraffe, Hippo, and Pig. The total length of each sequence is 1140, and there is no missing genome.

Simulation plan

First, the bootstrap is applied to the art4 data set to get the best estimated phylogenetic tree and other relevant parameters using RAXML. The model is the GTR+G4m model of nucleotide substitution, which is based on likelihood estimation method. The total iteration time is 1000. The starting tree contains 10 parsimony trees.

The plan is to simulate data on best tree from the art4.fasta, but with shorter central branch length. For other parameters, the stationary distriution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ and substitution rate r_{ij} are same as the best Model from bootstrap result. The vector of r_{ij} is converted to the vector of ρ_{ij} as shown in previous section.

The bootstrap results show that the central branch length of the best tree is 0.076413 in the art4.fasta data. The selected shorter central branch lengths of the tree are from 0.001,0.002,0.003,0.004,0.005,0.006,0.007,0.008,0.009,0.010. For each of the 10 branch lengths, 100 datasets will be simulated. There are total 1000 datasets. The genetic sequences for all the species are different for each data set. The goal is to find the bootstrap support percentage of each possible topology for each central branch length.

Next, for each dataset, the bootstrap analysis with total 1000 iteration time is conducted to calculate the total bootstrap support counts value of all the three possible topology among 1000 simulations. The bootstrap model is still GTR+G4m model of nucleotide substitution.

Results

First, the results of bootstrap on the art4 dataset with 1000 iterations shows stationary distribution $\pi_A, \pi_C, \pi_G, \pi_T$ are 0.310608,0.302900,0.132237, and 0.254255.The long-run proportion of transitions rate $\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT}, \rho_{GT}$ is 0.2647720752,0.1613423888 0.0335659908,0.0123514054,0.5277369051,and 0.0002312347.

After conducting bootstrap analysis on all the 1000 simulated datasets, for each branch length, the total bootstrap support counts value for all the topology are added together to calculate the bootstrap support percentage of every possible topology.

The following table shows the bootstrap support percentage of all the possible three topology at different branch lengths. In the table, each row represents different central branch length. The column is the corresponding bootstrap support values for the best tree topology estimated by bootstrap and remaining two possible topology:

	topology of best tree	topology 2	topology 3
0.001	0.40890	0.28268	0.30842
0.002	0.39312	0.31756	0.28932
0.003	0.45193	0.26352	0.28455
0.004	0.45554	0.28760	0.25686
0.005	0.52805	0.22208	0.24987
0.006	0.59059	0.18975	0.21966
0.007	0.60246	0.20867	0.18887
0.008	0.63080	0.18091	0.18829
0.009	0.60321	0.17972	0.21707
0.010	0.67021	0.15531	0.17448

In the best tree topology, the sheep and giraffe are closely related than other two species.The table shows that when the central branch length of the tree increases from 0.001 to 0.01, the bootstrap support percentage of the best tree topology increaes from 0.409 to 0.670.

Conclusion

In summary, using the art4 dataset which contains the sequences of four taxas, when the central branch length of the tree increases, the bootstrap support percentage of the best tree topology from the bootstrap estimation has substantial improvement.

This study could be furtuer improved by increasing the number of simulations in the bootstrap analysis. The current total simulation time is 1000. In addition, other possible central branch lengths could be simulated to investigate if the bootstrap support value still increases as the central branch length increases.