Gene Placement

osabary

Mar 2022

1 Introduction

The problem of gene placement in phylogenetic trees is a well known problem in which the goal is to place a new species on a given phylogenetic tree that represents an evolutionary chain of some given species.

2 Gene Rates - Multipliers

Let T be a species phylogenetic tree, consists of M species and their genomes. We denote by g_i the genome of the i-th organism from the tree. Furthermore, let S, be a set of N gene sequences $S = s_1, s_2, \dots S_N$. Each gene sequence $s_i \in S$ can be aligned to the species' genome using an MSA algorithm. Furthermore, each gene has evolution rate which we denote by r_i the evolution rate of the i-th gene s_i. Note that the topology of T is define by the average over all the genes in S. We define the matrix A as an $M \times N$ binary matrix, such that the j-th entry of the i-th row of the matrix is 1 if the j-th gene, s_j, appears for the i-th specie of the backbone and 0 otherwise.

Given a hamming distance of h, our (asymptotically additive) estimate of the phylogenetic distance under the Jukes Cantor (69) model is $d = \frac{3}{4}\log(1 - \frac{4}{3}h)$.

Given a query sequence, that consist of $k < N$ genes, APPLES minimizes the global (least square) hamming distance between the genes distances. Let $h_{i,j}$ denote the hamming distance for gene i to species j. Let $(t1, \dots, t_j, \dots t_M)$ be the distanced on *species* tree if you add query q to a particular location on that tree. Let L_i be the length of the i-th gene and define:

$$d_j = \frac{3}{4}\log(1 - \frac{4}{3}\frac{1}{\sum_i L_i}\sum_i L_i h_{i,j})$$

Then, APPLES finds the position on the tree (thus, t_j values) that minimizes:

$$\sum_j \left(\frac{1}{d_j}\right)^2 (d_j - t_j)^2 \ .$$

Our goal in this project, is to add sort of normalization/weighted version of the phylogenetic distance calculation, such that the evolution rate of each gene will be considered as well. That is, to define a weight function, which receives (an estimator of) the evolutionary rate, the length of the gene, and the matrix A, and estimates the weight of the gene. Thus, $w_i = f(r_i, L_i, A)$ represents the weight of the i-th gene, in the phylogenetic distance calculation.

$$d_j = \frac{3}{4}\log(1 - \frac{4}{3}\frac{1}{\sum_i w_i}\sum_i w_i h_{i,j})$$

3 Site Rate Multipliers

In this section, since the results for the gene rate multipliers did not show a significant improvement we decided to extend our approach, and to consider site rates. The site rate were calculated based on the site entropies. Given M species, and Q sites, for the j-th site, we first calculate p_X for $X \in \{A, C, G, T, -\}$, which is the normalized probability to see the base X in the j-th site. Then, the entropy of the j-th site is defined as:

$$H_j = \sum_{X \in \{A,C,G,T\}} p_X log(p_X),$$

where the entropy is ∞ if the site is gap in all species, that is if $p_- = 1$.

The site entropies approximate the level of preservation of each site. Hence, the next step is to perform a discretization on the site entropies in order to define k site multipliers on the sites, where k is hyperparameter.

1. The *naive site multipliers* are the k values $x_1, x_2, \ldots, x_k$ in the range between 0.5 to 1.5 that partition the segment to k equal parts.

2. Optimized k values using regression. The optimization is done in the following way.

   (a) Inital values are: $x_1, x_2, \ldots, x_k$.

   (b) The constraints are: $x_1 < x_2 < \ldots < x_k$, and the median **or** the expected multiplier is 1.

   (c) Given $d_{ti,j}$, the tree distance between the i-th and the j-th species, which their sequences is given by s_i and s_j. The target function is:

   $$\arg \min_{x_1, x_2, \ldots x_k} \sum_{s_j, s_i} (JC^{-1}(d_t) - d_H(s_j, s_i, (x_1, x_2, \ldots, x_k))^2,$$

   where JC^{-1} denotes the inverse function of the JC transformation, and $d_H(s_j, s_i, (x_1, x_2, \ldots, x_k)$ denotes the hamming distance bewtween the species s_i and s_j with site multipliers $(x_1, x_2, \ldots, x_k)$.