A phylogenetic model to predict the patterns of presence and absence of genes in bacterial genomes and estimate the frequency of horizontal gene transfer

A phylogenetic model to predict the patterns of presence and absence of genes in bacterial genomes and estimate the frequency of horizontal gene transfer

By

Seyed Alireza Zamani Dahaj

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2014) McMaster University

(Physics and Astronomy) Hamilton, Ontario

TITLE: A phylogenetic model to predict the patterns of presence and absence of genes in bacterial genomes and estimate the frequency of horizontal gene transfer

AUTHOR: Seyed Alireza Zamani Dahaj

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: XII, 90

Abstract

For a group of bacterial genomes, the core genome is the set of genes present in all the individual genomes and the pangenome is the set of genes present in at least one of the genomes. Typically, a relatively small fraction of genes is in the core, and many other genes are only found in only one or a small number of genomes. This indicates that there is a wide range of time scales of genome evolution, with rapid insertion and deletion of some genes and long-term retention of others. Here, we study the full set of the genes in a group of 40 complete genomes of Cyanobacteria. Genes are clustered using sequence similarity measures, and for each cluster we obtain the pattern of presence and absence of the genes across the 40 species. We use evolutionary models of gene insertion and deletion to calculate the likelihood of each of the observed patterns. One important case we consider is the infinitely many genes model (IMG) in which each gene can only originate once but can be deleted multiple times. In contrast, the finitely many genes model (FMG) allows more than one insertion of the same type of gene in different genomes, which would be the case if there were horizontal gene transfer (HGT).

The maximum likelihood model allows us to predict which genes have a presence-absence pattern that is best explained by horizontal transfer. We find that about 15% of the genes experienced HGT in their history of evolution. It is found that there is a broad range of rates of insertion and deletion of genes, which explains why there are a large number of genes that follow a typical treelike pattern of vertical inheritance, despite the presence of a significant minority of genes that undergo HGT. We also estimate the

ancestral genome size of Cyanobacteria. It is found that that the inferred frequency of HGT and the size of the ancestral genome both depend on the ratio of insertion to deletion rates of genes. However, the variation in the estimated ancestral genome size is much less than in previous treatments that used parsimony.

As the phylogenetic tree of Cyanobacteria is not completely specified, we test our models on ten different trial species trees that differ by small rearrangements of species. It is found that the estimated frequency of HGT and the maximum likelihood values of the insertion and deletion rate parameters are not very sensitive to small changes in the tree. However, the likelihood of the gene presence/absence patterns on different trees differs significantly among the trees. Therefore, these patterns can be used for phylogenetic inference. This kind of phylogenetic inference makes use of all the genes present on the genomes. In contrast, phylogenetic methods using protein sequence evolution only make use of the relatively small number of genes that are present in all of the genomes in the set. We compare the likelihood ranking of trial trees using the presence/absence data with the ranking of the same trees using protein sequence evolution with conserved common genes (present in all cyanobacteria and a large proportion of other genomes) and signature genes (present in all cyanobacteria and no other species).

Acknowledgements

I would like to thank my supervisor Dr. Paul Higgs for his supportive advice, invaluable direction and help. I would also like to thank Dr. Brain Golding for his comments and Dr. Radhey Gupta for providing us with cyanobacteria sequence data and his suggestions during developing this work.

I would also like to thank my wife for her patience and my family and friends for their support.

dedicated to Rojin

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Evolution of Bacterial Genomes

Since Lamarck introduced bifurcating trees as a metaphor to describe the evolution of species, the concept of a tree of life has been a powerful tool for taxonomic classification (de Lamarck, 1839). For large multicellular organisms, there are many observable morphological features that make useful phylogenetic characters, and for animals with hard body parts, these characters can also be observed in extinct species in the fossil record. However, micro-organisms have very few easily observable characters and almost no fossil record. Phylogenetic studies of micro-organisms were therefore very difficult until the advent of molecular sequence data.

Among the first authors to carry out molecular phylogenetics with micro-organisms were Woese and Fox (1977), who used ribosomal RNA (rRNA) as a taxonomic marker (Fig 1-1). They showed that there are three domains of life: Bacteria, Archaea and Eukarya, whereas previously, only the distinction between eukaryotes and prokaryotes had been made. However, in the rRNA phylogenetic tree, many of the taxonomic groupings and the relationships between them remain poorly resolved, especially in the domain of bacteria. In recent decades, protein sequence data has become available for large numbers of genes, and complete genomes are also now available for thousands of bacteria. The availability of this sequence data, instead of reducing the ambiguity of the tree of life, as was originally hoped, has created more

uncertainty in the tree like picture for the evolution of bacteria (Hilario & Gogarten, 1993). Different genes were shown to have different histories, and gain and loss of genes from genomes was found to be relatively rapid. Even close strains of the same species were discovered to have very diverse gene repertoires. For instance, study of three strains of *Escherichia coli* found that less than 40% of their genes were present in all three genomes (Fig 1-2) (Welch *et al*, 2002).

Fig 1-1. *Tree of life of all known organism based of universal highly conserved proteins (Ciccarelli et al. 2006); Bacteria (purple), Archaea (green) and Eukaryotes (red).*

The differences between gene trees and the differences between the genome contents of related species are both attributable to a large extent to horizontal gene transfer (HGT). This means that prokaryotes have the ability to exchange genes among distant species and to pick up genetic material from the environment and insert it into their chromosomes. The evolution of prokaryotic genomes includes different processes of gain, duplication and HGT (Syvanen, 1985). Depending on which set of genomes is studied, the relative importance of these processes could vary (Lobkovsky *et al*, 2013).

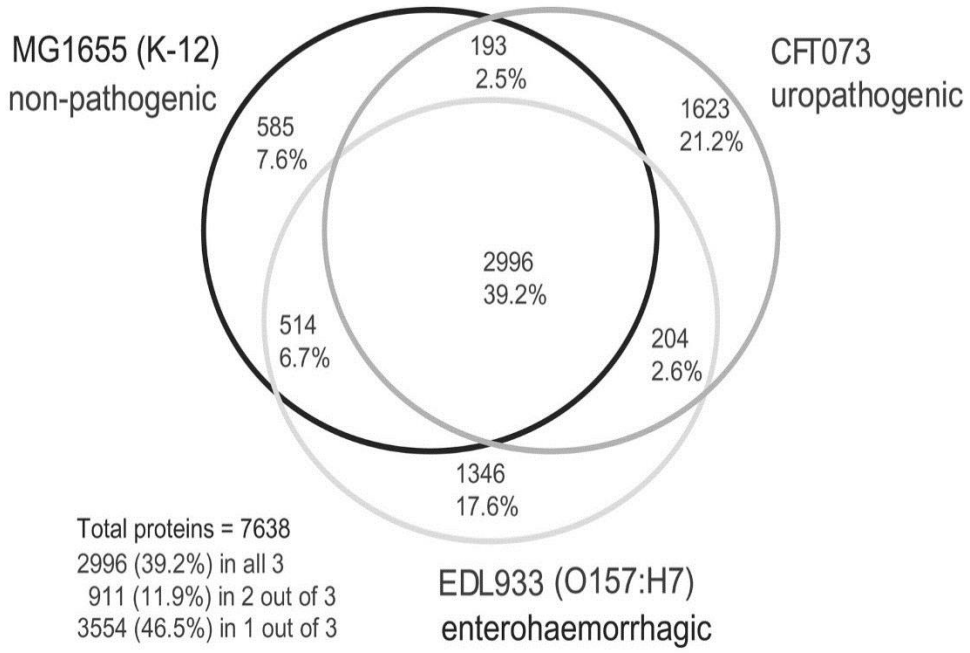


Fig 1-2. *Venn diagram of three E. coli genomes. Fewer than 40% of genes are common to these three strains of E.coli. (Welch et al. 2002).*

For a set of related genomes, such as the three *E. coli* genomes in Fig 1-2, the core genome is the set of genes present in all the individual genomes, and the pangenome is the set of genes present in at least one of the genomes. Typically, a relatively small fraction of genes are in the core, and many other genes are only found in just one or a small number of genomes. For example, analysis of 17 different *E. coli* strains genomes showed that the core genome decreased to around 2,200 genes, whereas the pangenome size grows to more than 13,000 (Rasko *et al,* 2008). If the pangenome continues to increase indefinitely when new genomes are added, it is said to be open, whereas if it tends to a limit, it is said to be closed. In most of the cases that have been studied, the pangenome is found to be open, which means that new genes arise continuously in each lineage (Collins & Higgs, 2012).

The plot of gene frequencies versus the number of species that have that gene is called gene frequency distribution (Fig 1-3). *G(k)* is the number of genes (or gene families) that is found in *k* genomes.. *G(k),* regardless of the group of genomes, always has an asymmetrical U-shape. *G(k)* can be characterized with a "core" of (closely) universal genes, a "shell" of relatively common genes and a "cloud" of very fast evolving genes. A variety of mathematical models has been developed to fit the shape of the *G(k)* distributions (Baumdicker *et al.* 2010; Collins and Higgs, 2012; Lobkovsky *et al.* 2013).With very closely related groups, such as the Prochlorococcus in Fig. 1-3 (Baumdicker *et al.* 2010) there is a relatively large fraction of core genes. For more diverse groups, the core fraction goes down but the U shape is retained (Collins and Higgs, 2012; Lobkovsky *et al.* 2013). This nontrivial fact shows that there is a wide range

of time scales for gene gain and loss, with rapid insertion and deletion of some genes and long-term retention of others.

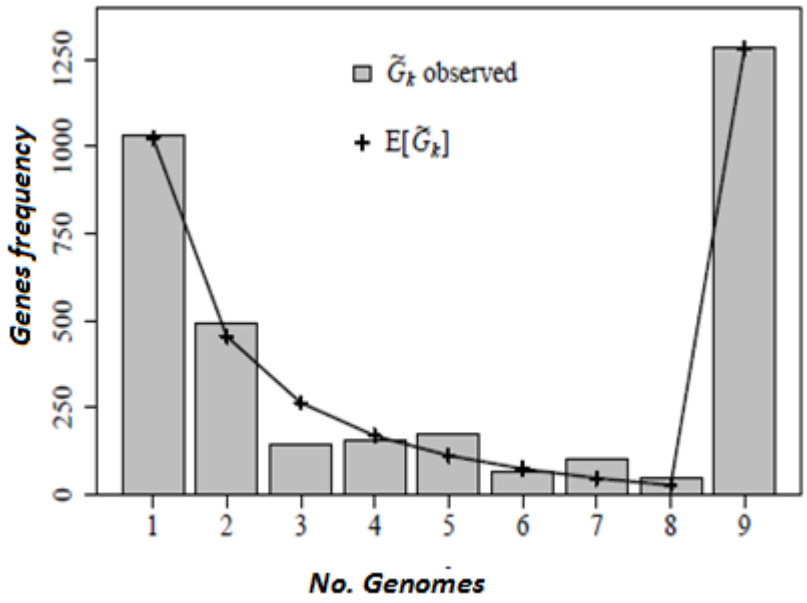


Fig 1-3. *The gene frequency distribution for nine closely related strains of Prochlorococcus and expected gene frequency distribution from "Infinitely Many Genes" model (Baumdicker et al, 2010).*

Horizontal Gene Transfer Mechanisms

The process by which bacteria gain unrelated DNA sequences and insert them into their genome is called "Horizontal Gene Transfer", or sometimes "Lateral Gene Transfer", because these genes are not inherited from the parent DNA (Syvanen, 1985). Some people argue that the rate of HGT is very high, and these events dilute the tree of life of prokaryotes. Thus, a single bifurcating tree is not enough to explain the evolution of prokaryotes (Kunin *et al.* 2005, Hilario & Gogarten, 1993, Zhaxybayeva *et al.* 2006,

Bapteste *et al.* 2007). On the other hand, others argue that there are strong signs of a tree that are visible despite the noise of HGT events and that very few of the genes of a typical genome exhibit HGT (Wang & Kim, 2005, Pere Puigbò *et al.* 2009). Thus, there remains a lack of consensus in the molecular evolution community of the frequency of HGT and the extent to which the tree of life picture remains valid in the light of modern genome data.

There are three primary mechanisms for horizontal gene transfer (Gyles & Boerlin, 2013): transformation, conjugation, and transduction..

Transformation was observed in 1928, by the British bacteriologist Fredrick Griffith, who discovered that a strain of *Streptococcus pneumoniae* could be infectious after being exposed to heat-killed virulent strains. Griffith deduced that there was some kind of transformation mechanisms from the heat-killed strain that made the harmless strain infectious. This process was identified in 1944 as DNA transformation (Avery *et al*, 1944). This process involves uptake of a piece of alien DNA from their environment and adding it to the genomic material of the recipient cell. Transformation occurs when a cell is in a state of competence, in which the organism gains the ability to take up extracellular naked DNA due to the presence of specific membrane proteins that import the DNA (Chen *et al*, 2004).

Conjugation is a process in which one cell is a donor and the other is a recipient. In this process, two cells make a direct cell-to-cell contact or a bridge-like connection between themselves (Gyles & Boerlin, 2013). Through this connection, they exchange

their genetic material which is usually an F-plasmid with a length of about 100 kb. F-plasmids (where the F stands for Fertility factor) could have the genes that are responsible for bacterial antibiotic resistance. So, this process could be very beneficial for the bacteria. Sometimes this process is regarded as sexual activity by bacteria because it contains cell to cell contact and genetic material exchange between them. It was first discovered by Tatum and Lederberg (1947) in their studies of gene recombination of *Escherichia coli.*

In transduction, bacteriophages (bacterial viruses) transfer genetic material from one bacterium to another. When a phage infects a bacterium, it hijacks the host replication machinery and forces it to replicate its DNA or RNA. It is possible that occasionally the phage picks up some part of the host DNA instead of its own genetic material. So, this allows host genes to be transferred to another bacterium in the next round of infection (Gyles & Boerlin, 2013). This process could be used for genetic engineering for introducing specific genes into different bacterial genomes (Kiel *et al*, 2010).

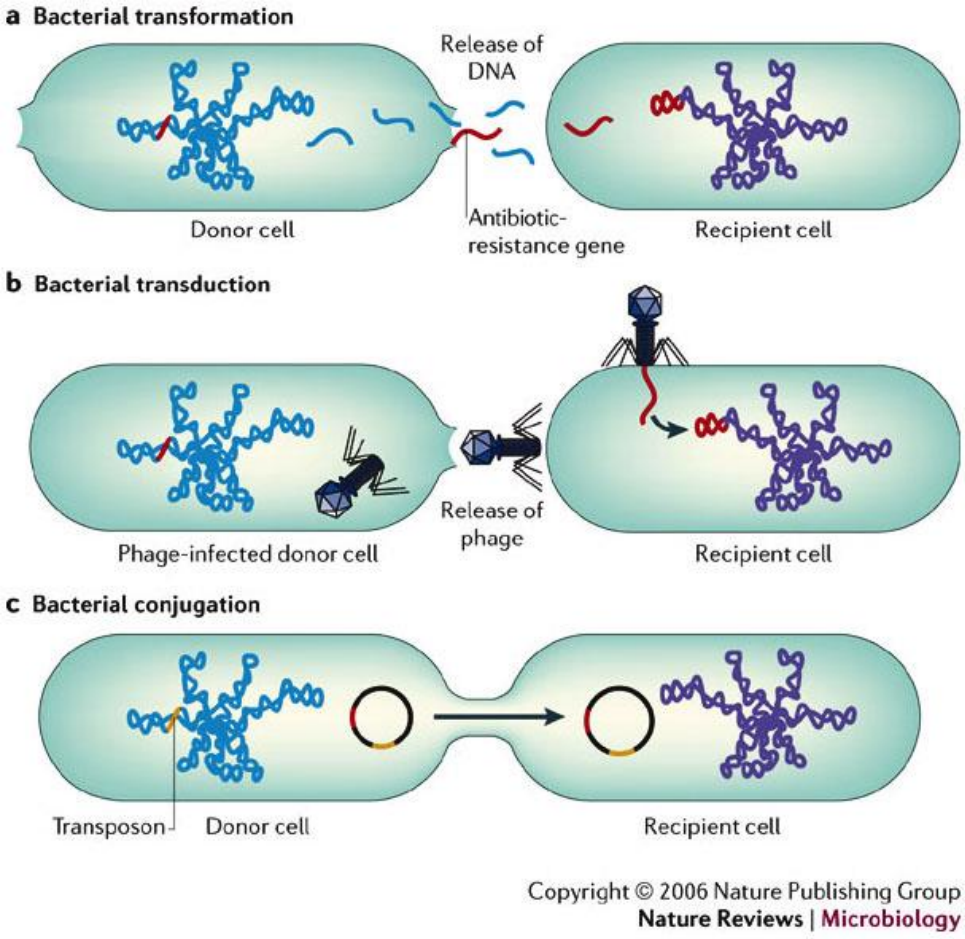


Fig 1-4. *Schematic of three process of Horizontal Gene Transfer.*

Quantifying Horizontal Gene Transfer

As mentioned above, there is still substantial debate about the rate of HGT and its significance in the evolution of prokaryotes. Consequently, there are lots of efforts to resolve this problem and to quantify it, either theoretically or experimentally (Karberg *et*

al. 2008). Several computational methods for analyzing HGT problem exist today. Here, we discuss three of the main approaches to estimate the frequency of HGT.

The aim of the first approach is to look for regions of a genome that differ significantly from the surrounding region of the genome in which they are found. In particular, GC content and codon usage frequencies differ substantially between genomes and are dependent on the particular nature of mutation and selection that occurs in each genome. Recently inserted genes retain the sequence properties that are characteristic of the originating genome and thus stand out from their surroundings (Azad & Lawrence, 2007). However, once inserted, the transferred genes are subject to the same set of mutational processes as the rest of the genes in the genome. Therefore they ameliorate over time to reflect the sequence composition of the recipient genome (Lawrence & Ochman, 1997). Also, if the inserted genes come from an organism with similar sequence composition and characteristics it is impossible to detect them with this method (Wang, 2001).

The second approach for detecting HGT events is comparison between the gene trees of different genes and a reference tree thought to represent the species evolution. In nearly all cases where large numbers of gene trees are studied, there are some conflicts between the species tree and single gene trees. These conflicts can be indicative of HGT, although it should be borne in mind that single genes are sometimes insufficient to resolve a phylogeny. Phylogenetic trees are in any case sensitive to methods and evolutionary models used even in cases where no HGT is thought to have occurred. Thus

it remains unclear how much of the conflict between gene trees is due to insufficiency of phylogenetic methods and how much is due to HGT.

Several methods have been developed to reconcile gene trees with species trees (Lyubetsky & V'yugin , 2003; David & Alm. 2011; Szöllősi *et al.* 2012). These methods use the parsimony principle, which means the preferred scenarios are the ones that require the smallest number of evolutionary events to explain the observed data. The events include horizontal gene transfer, gene duplication, gene loss, speciation and gene birth or genesis events. A unique penalty corresponds to each event and the goal is to minimize the total penalty. David and Alm (2011), with their version of this method, found that the rate of HGT has been nearly constant after a huge expansion in the genome materials of prokaryotes around three billion years ago.

Bansal *et al.* (2011), mentioned some of the most important problems of this method. This method is an NP-hard problem and it is very hard to solve exactly under most formulations. Secondly, there may be multiple alternative optimal answers for the HGT inference problem. And finally, there isn't any reason that HGT events should follow a parsimony principle, especially when the HGT rate is relatively high. So, Bansal *et al.* (2011) try to treat the problem statistically and develop a method that does not infer individual HGT events. Instead, they try to detect the highways of HGT, in which many different genes were horizontally transferred between two species. Because these highways should affect the history of many genes, the main idea behind their program is to combine lots of gene tress and use that to infer the most important HGT events. One of

the biggest advantages of their method is that they solved the HGT inference problem in polynomial time.

The last approach to the HGT problem is by studying the presence and absence patterns of gene families across different species. Early versions of this approach used maximum parsimony to infer the HGT events. In this method, the aim is to find the minimum number of evolutionary events (origination, loss and HGT) that leads to the observed pattern, given the species tree (Mirkin *et al*, 2003). Recent studies using these patterns are using maximum likelihood (Hao W & Golding GB, 2006; Cohen O *et al*, 2008; Szöllősi GJ *et al*, 2012; Kannan L *et al*, 2013; Sjöstrand J *et al*, 2014). For instance, Cohen and Pupko (2010) studied a data set of 4873 different gene families across 66 species by annalyzing their presence-absence patterns. They developed likelihood based evolutionary models for explaining the dynamics of the patterns. One of the main assumption of their model is that all gene families are evloving on the same phylogenetic tree as the species. Also, all gene families are assumed to evolve independently and through a Markov process over two state (presence/absence). In their model, each gene family had a deletion and insertion rete. The insertion and deletion rates were sampled from two independent discrete gamma distributions (Yang, 1993). Each gamma distritribution was partitioned into four different rates, so there were sixteen different combinations of the insertion-deletion rates. They first calculated the posterior probability that each gene family got inserted in each branch. If the two highest insertion probabilities both passed a given thresold (which was taken to be 0.85), it was assumed that this gene family exhibited HGT, because at least one of these insertions must

represent the transfer of a gene that already existed elsewhere The thresold was set to 0.85 because this was found to minimize the number of false positive inferences of HGT that was found in simulated data. Using tis method, they proposed that at least 34% of the analyzed gene families at least once in their history have experienced HGT.

Dagan and Martin (2007) tried to find the minimum rate of HGT by calculating the ancestor genome size under different restricted HGT rates. If HGT is not permitted or if it is permitted at too low a rate, then genes found in present day species must have descended from early common ancestors. This means that the estimated ancestral genome size is much larger than the size of current genomes, which seems unlikely. On the other hand, if the HGT rate is too large, almost all gene presence absence patterns are explained via HGT and the estimated ancestral genome size is very much smaller than modern genomes, which also seems unrealistic. They found that the rate of HGT at which the ancestral genome size was comparable to modern genome sizes was around 1.1 event per gene family.

Aims of This Thesis

In order to estimate the rates of genes origination, deletion and horizontal gene transfer (HGT) in the bacterial genomes, we developed a phylogenetic method based on maximum likelihood criteria. One important case we consider is the infinitely many genes model (IMG) in which each gene can only originate once, but can be deleted multiple times (Baumdicker *et al*, 2010). In contrast, the finitely many genes model (FMG) allows more than one insertion of the same type of genes in different genomes,

which would be the case if there were horizontal gene transfer. Previously, Collins and Higgs (2012) studied the capabilities of different evolutionary models based on IMG to predict the properties of the pangenomes, core genomes and gene frequency distribution of different monophyletic subsets of 172 completely sequenced genomes of *Bacili*. They found that a model that consist of a set of essential genes, one slow evolving and a fast evolving IMG classes can explain most of the data better than other tested models. In this work, we go one step beyond the gene frequency distribution and looking at the patterns of absence and presence of each gene family. These patterns have more information than simple distribution graph and make a distinction between the gene families that are present in the same number of genomes but not necessary the same genomes. So, it is possible to identify genes whose profile is inconsistent with vertical inheritance and which are candidates for HGT by using evolutionary models of gene insertion and deletion to calculate the likelihood of each of the observed patterns.

In this work, we studied the set of the gene families in a group of 40 fully sequenced cyanobacterial genomes. We estimate the number of gene families that are the strong candidates to exhibit HGT and gene families' insertion and deletion rate. We test our method on ten distinguishable phylogenetic trees to see their effect on HGT amount as well as the impact of changing the root.

Chapter 2 Phylogeny of Cyanobacteria

Cyanobacteria

We chose cyanobacteria as a test case with which to try out the new methods for genome evolution that we developed in this thesis because a large number of well annotated complete genomes of this group are available. Moreover, a variety of studies of genome evolution have been made in this phylum by other methods (Shi & Falkowski, 2008; Szöllősi *et al*, 2012; Dagan *et al*., 2013; Sjöstrand *et al.*, 2014; Howard-Azzeh *et al*, 2014). Cyanobacteria are the only known phylum of bacteria that is capable of oxygenic photosynthesis. In fact, they are responsible for one of the major events in the history of life on Earth by inventing oxygenic photosynthesis and converting the Earth's atmosphere from reducing to oxidizing approximately 2.4 billion years ago. This milestone ultimately provided sufficient conditions for development of more complex life forms that depended on aerobic metabolism (Shi & Falkowski, 2008). Cyanobacteria also contributed significantly to the evolution of eukaryotes by conferral of the photosynthesis cycle to plants and other eukaryotes through endosymbiosis in the form of chloroplasts. Moreover, they still have an important role in today geochemical cycles through N_2-fixation, and sequestration of phosphorous and trace metals such as iron, magnesium, zinc and so on (Moisander *et al*, 2010). Because of these unique capabilities they attract a lot of interest as subjects for green biotechnology and biofuel production (Shih *et al.* 2013). They are also among the largest and most diverse groups of bacteria in their

genetic material, cellular differentiation, physiological capabilities and choice of habitat (Larsson, *et al*, 2011; Howard-Azzeh *et al*, 2014).

Despite their crucial role in the evolution of life in this planet, comprehensive studies of their evolution and biology only began recently. In one of the recent studies Shi *et al* (2012) sequenced the genomes of 54 diverse cyanobacterial strains which doubled both the amount and the phylogenetic diversity of cyanobacterial genome sequence data. They found 21,000 novel proteins in cyanobacteria with no detectable similarity to known proteins. Horizontal gene transfer is also considered to be a major player in the evolution of cyanobacteria. Dagan *et al*. 2013, studied the evolution of cyanobacteria and divide their genetic material into three different classes: *(i)* ordinary components or vertically inherited genes, *(ii)* Horizontally transferred genes, and *(iii)* bestowal genes to plastids of eukaryotes. The last class consists of both the first and the second one. They find that around 60% of cyanobacterial gene families have been affected at least once by horizontal gene transfer events in their history. Also, by studying the vertical components, they inferred that water-splitting photosynthesis should have been originated in freshwater.

Zhaxybayeva *et al* (2006) argued that because of complex evolution of cyanobacterial genomes and the high rate of HGT events, it is impossible to represent their phylogeny by a strictly bifurcating tree. They analysed 1128 protein-coding gene families from 11 completely sequenced cyanobacterial genomes available at that time and tried to find genes affected by HGT events within cyanobacteria and other phyla. They

compared single gene trees with the species reference tree and attempted to find cases of conflict with single gene trees. They found that the genes from all functional categories had experienced HGT in their history. Two years later, Shi and Falkowski (2008) commented that despite relatively high HGT rate in cyanobacteria, there is a strong sign of a species phylogenetic tree. In their study, they showed inconsistency among the history of 682 orthologous shared within 13 cyanobacterial genomes. However, they found that there is a core of 323 gene families with nearly indistinguishable evolutionary history. These highly conserved core genes consist of genes that code for the vital components in the oxygenic photosynthesis cycle and ribosomal apparatus. They suggest that because these two fundamental components are related by macromolecular interactions in complex protein structures and sophisticated metabolic pathways, it is very hard for them to evolve independently. Also, as a side effect of their complexity, there is a strong selection force against horizontal transfer of just a few of the genes in the core genome. Although, this core is highly conserved, other genes in cyanobacterial genomes are more likely to be affected by HGT events. Finally, they propose that the most recent common ancestor of cyanobacterium phylum lacked the ability to do nitrogen fixing and probably was a thermophilic organism.

In this chapter, we will use protein sequence data from completely sequenced cyanobacterial genomes to obtain phylogenetic trees. We will also consider several trees obtained by previous studies. Our aim is to produce a set of plausible trial trees that are supported by gene sequence evolution. In chapter 3, we will then study the evolution of the gene presence-absence patterns on these trial trees.

Maximum Likelihood Methods for Phylogenetic

One of the most useful methods for finding phylogenetic trees is "Maximum Likelihood". We will describe it here because we make use of it for obtaining the phylogeny of cyanobacteria from protein sequences in this chapter, and also for studying gene presence/absence patterns in the following chapter.

Routinely in probability problems, the probability of a set of data X is calculated, given a set of model parameters values θ. This is written as $P(X|\theta)$. The Likelihood method may be thought of as the reverse process; it starts with a specific set of data and tries to calculate the probability of explaining that data with different models. The aim of the maximum likelihood method is to find the parameters values that maximize the probability of observing the data. So, we define the likelihood function as

$$L(\theta|X) = P(X|\theta) \qquad (2\text{-}1)$$

It should be remembered that the $P(X|\theta)$ is a normalized probability distribution over all the outcomes for the data, whereas $L(\theta|X)$ is not normalized, because often there is not a well-defined complete set of possible models and model parameters. Nevertheless, the relative likelihood of the data according to alternative sets of model parameters is still a meaningful quantity. In phylogeny, the data X are the set of aligned sequences of a group of species and the parameters θ are the tree topology, length of branches, the background evolutionary model and so on (Gascuel, 2005). In other words, likelihood shows how "likely" it is to observe the data for a given model, so, a higher likelihood means better model. The maximum likelihood criterion is to find the set of parameters (here the best

tree) that maximize the likelihood function. The algorithm for calculating the likelihood of a tree for a specific evolutionary model and set of aligned sequences was first introduced by Felsenstein (1981).

Here is a brief description of Felsenstein algorithm for calculating the maximum likelihood for a tree with a certain topology and fixed branch lengths. The function $P_{ij}(t)$ is the transition probability from state i to state j given the evolutionary distance t (branch length) between them (Higgs & Attwood, 2005). This function should be calculated from an appropriate evolutionary model. For instance for a model of gene presence and absence, there are two states, 0 and 1 and four transition probability functions $P_{00}(t)$, $P_{01}(t)$, $P_{10}(t)$, and $P_{11}(t)$. These probabilities are given in chapter 3 when we discuss presence absence data. For a model of protein sequence evolution, there are 20 possible states corresponding to the 20 amino acids, and there are 400 transition probabilities. There are a variety of standard models of protein sequence evolution (see Higgs and Attwood (2005) and references therein).

Fig 2-2 shows part of a phylogenetic tree, with an internal node n that has state i, and two descendent nodes n_1 and n_2, having states j and k. The times on the branches leading to these two nodes are t_1 and t_2. Let $L_i(n)$ be the likelihood of the part of the tree that descends from node n, given that the state of node n is i. The likelihood can be calculated via a recursion relation.

$$L_i(n) = \sum_j \sum_k P_{ij}(t_1) L_j(n_1) P_{ik}(t_2) L_k(n_2) \qquad (2\text{-}2)$$

To initiate the recursion, we use the fact that if either of the descendent nodes is a tip node, then its state is given in the pattern, and the likelihood is 1 for that state, and zero for all other states. For example, if n_1 is a tip, and it has state $j = 1$, the $L_1(n_1) = 1$ and $L_0(n_1) = 0$.

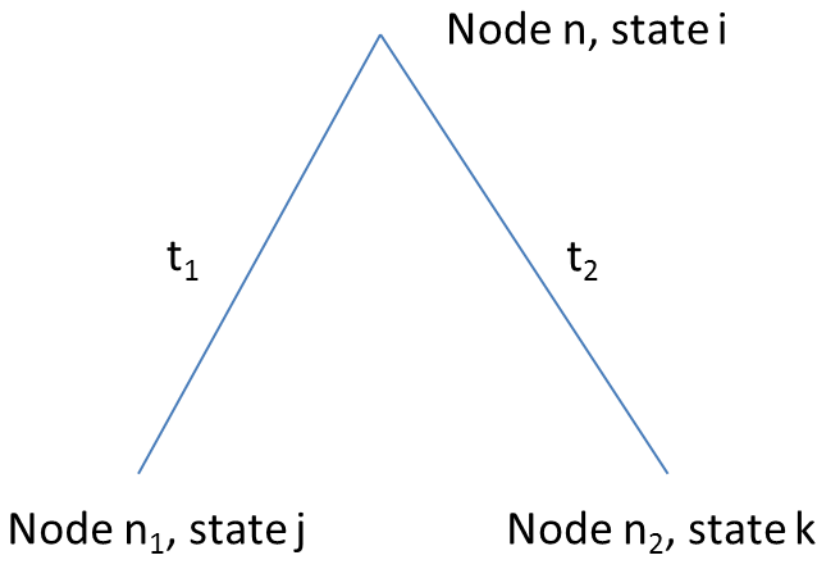


Fig 2-2. *Schematic of likelihood calculation for a single site on a given fixe tree.*

This recursion can be repeated until we reach the root node, and the likelihoods L_i(root) are obtained. For calculating the total likelihood of the pattern, we need to sum over all possible values of the state of the root node, with each possibility being weighted by the equilibrium frequency of that state, π_i according to the evolutionary model that we

used to calculate the probability functions, $P_{ij}(t)$. Hence the likelihood for one gene family is equal to:

$$L_{pat} = \sum_i \pi_i L_i(root) \qquad (2\text{-}4)$$

One of the interesting property of the likelihood of a pattern, L_{pat}, is that if the substitution model is time reversible, L_{pat} is independent of the position of the root within the tree. The time reversibility condition is true for most of the models of protein and DNA evolution. So, when we are calculating likelihood, we do not have to worry about the root and could choose any internal node as a root. However, if we wish to draw a phylogeny as a rooted tree, we need to use information from outside the current data set to define where the root is.

The calculated likelihood is for one pattern in the data (*i.e.* a particular gene presence/absence pattern or a particular site in a protein sequence alignment). Total likelihood of the data (L_{tot}) is equal to the product of likelihood of all sites. The likelihood is by its definition less than one. Hence, L_{tot} become astonishingly small and it is more practical to use the *ln* of likelihoods. Then $ln\, L_{tot}$ is equal to sum of the *ln-likelihood* of each pattern:

$$ln\, L_{tot} = \sum_{pat} ln\, L_{pat} \qquad (2\text{-}5)$$

Usually, the maximum likelihood programs quoted the *lnL* values as a final result. Because total likelihood is a product of so many numbers that are very smaller than one then, the *lnL* is large negative number.

L_{tot} depends on the values of the parameters in the evolutionary model, the branch lengths, and the topology of the tree (*i.e.* the relationships between the species). It is necessary to search over these parameters using some kind of optimization procedure to find a configuration that maximizes the likelihood. For any given tree topology, the branch lengths and the model parameters are continuous variables. Optimization algorithms make small changes in these variables and determine whether the likelihood goes up or down. The topology itself is a discrete change. The simplest kinds of tree arrangements are known as nearest neighbour interchange and subtree pruning and regrafting (Felsenstein, 2004). A maximum likelihood tree-search program tries many alternative tree topologies and values for the continuous parameters in a heuristic way, keeping track of the best configuration found so far. It is hoped that the will find a configuration that is the maximum likelihood configuration or very close to it if the program is run for long enough.

It is often found that many alternative trees and parameter values give likelihoods that are only slightly lower than the maximum likelihood value. In Bayesian phylogenetic methods, the aim is to draw conclusions from a set of high likelihood trees, rather than from one single maximum likelihood tree. It is possible to use a "Monte Carlo Markov Chain" (MCMC) process to generate a set of high likelihood trees. The probability that a configuration appears in the output of an MCMC program is proportional to its likelihood. The consensus tree of all the trees visited during an MCMC run gives information about the ensemble of high likelihood trees. There are several methods for building the consensus tree. The essence of all of them is to establish the common

elements between the trees or take the lowest common denominator. In other words, the main idea is to identify common sub-trees in the collection of input trees and represent them as a single tree. By exclusion, it is possible to recognize the areas of conflict (Gascuel, 2005). For the consensus tree of an MCMC run, the most important quantities are the posterior probabilities of each possible clade of species. If a set of species always forms a clade in every tree visited by the MCMC run, then the posterior probability for this clade is 100%. For parts of the tree that are less well specified by the data, there will be conflicting arrangements of species in the set of MCMC trees. The consensus tree will show the most frequent of these possible rearrangements, and the posterior probability of this clade can be indicated in the tree drawing at the appropriate node of the tree.

Cyanobacteria in the HOGENOM Database

In this work, we construct the phylogenetic trees of all cyanobacterial genomes present in HOGENOM database (Penel *et al*, 2009). This database contains homologous genes from 1470 fully sequenced organisms from all three domains of life. It has clusters of gene families of complete genomes of forty different species of cyanobacteria. Each organism is assigned its unique four or five letter code (Table 2-1). Gene clusters in HOGENOM must have at least two genes in them. Single genes that do not cluster with any other sequence are not included in the database. In some cases, however, two or more duplicate genes in a single genome form a cluster with no members from outside this genome. For our work, we excluded these single-genome clusters and included only clusters that had gene members from at least two separate genomes. For the 40 fully sequenced cyanobacterial genomes in HOGENOM, there are 10304 unique homologous

gene families present in at least two genomes, and 3510 different patterns of presence and absence.

This group of cyanobacteria is very diverse in their genome size, habitat and physiological systems. The largest genome size in this group belongs to *Acaryochloris marina MBIC11017* which has 8383 protein-coding genes. On the other hand, *CYANOBACTERIUM UCYN-A* has only 1199 protein-coding genes. The average genome size of this group is 3580 genes while the average number of genes per genome is 2878 in HOGENOM because it doesn't have genes that are unique to a single species (Fig 2-2).

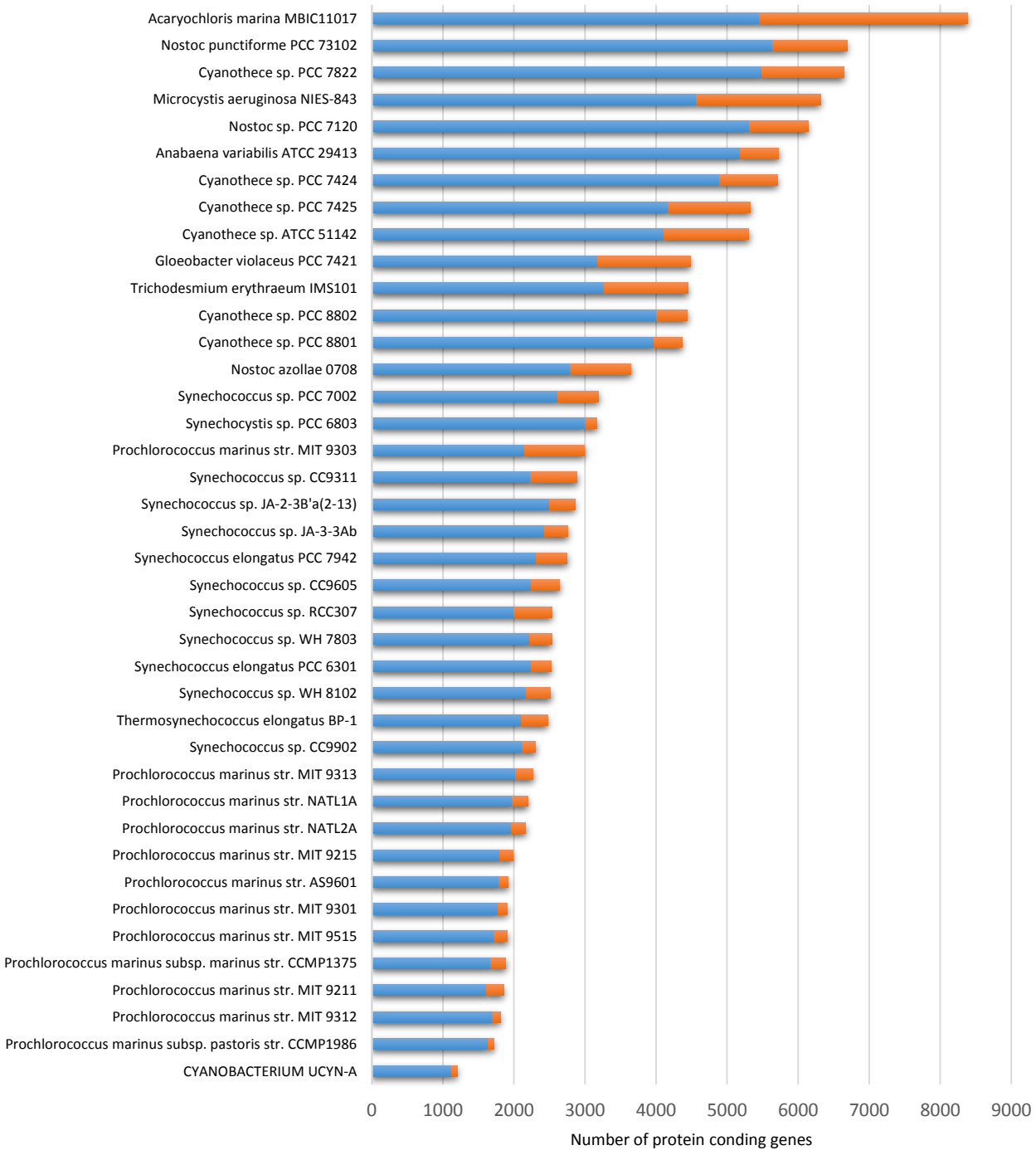


Fig 2-2. *Genome size among 40 cyanobacteria. The bars indicate the number of the genes in each genome. The blue and orange portions of them show number of the genes in the HOGENOM database and unique genes of each genome, respectively.*

***Table 2-1**. Full name of studied cyanobacteria and their HOGENOM code name and clade.*

Name	HOGENOM Code	Clade
Acaryochloris marina MBIC11017	ACAM1	Chroococcales
Cyanothece sp. PCC 7822	CYAP2	Chroococcales
Cyanothece sp. PCC 7424	CYAP7	Chroococcales
Cyanothece sp. PCC 7425	CYAP4	Chroococcales
Cyanothece sp. ATCC 51142	CYAA5	Chroococcales
Cyanothece sp. PCC 8802	CYAP0	Chroococcales
Cyanothece sp. PCC 8801	CYAP8	Chroococcales
Synechococcus sp. PCC 7002	SYNP2	Chroococcales
Synechocystis sp. PCC 6803	SYNY3	Chroococcales
Synechococcus sp. CC9311	SYNS3	Chroococcales
Synechococcus sp. JA-2-3B'a(2-13)	SYNJB	Chroococcales
Synechococcus sp. JA-3-3Ab	SYNJA	Chroococcales
Synechococcus elongatus PCC 7942	SYNE7	Chroococcales
Synechococcus sp. CC9605	SYNSC	Chroococcales
Synechococcus sp. RCC307	SYNR3	Chroococcales
Synechococcus sp. WH 7803	SYNPW	Chroococcales
Synechococcus elongatus PCC 6301	SYNP6	Chroococcales
Synechococcus sp. WH 8102	SYNPX	Chroococcales
Thermosynechococcus elongatus BP-1	THEEB	Chroococcales
Synechococcus sp. CC9902	SYNS9	Chroococcales
CYANOBACTERIUM UCYN-A	UCYNA	Chroococcales
Gloeobacter violaceus PCC 7421	GLVIO1	Gloebacter
Nostoc punctiforme PCC 73102	NOSP7	Nostocales
Nostoc sp. PCC 7120	NOSS1	Nostocales
Anabaena variabilis ATCC 29413	ANAVT	Nostocales
Nostoc azollae 0708	NOSA0	Nostocales
Microcystis aeruginosa NIES-843	MICAN	Oscillatoriales
Trichodesmium erythraeum IMS101	TRIEI	Oscillatoriales
Prochlorococcus marinus str. MIT 9303	PROM3	Prochlorales
Prochlorococcus marinus str. MIT 9313	PROMM	Prochlorales

Prochlorococcus marinus str. NATL1A	PROM1	Prochlorales
Prochlorococcus marinus str. NATL2A	PROMT	Prochlorales
Prochlorococcus marinus str. MIT 9215	PROM2	Prochlorales
Prochlorococcus marinus str. AS9601	PROMS	Prochlorales
Prochlorococcus marinus str. MIT 9301	PROM0	Prochlorales
Prochlorococcus marinus str. MIT 9515	PROM5	Prochlorales
Prochlorococcus marinus subsp. marinus str. CCMP1375	PRMAR1	Prochlorales
Prochlorococcus marinus str. MIT 9211	PROM4	Prochlorales
Prochlorococcus marinus str. MIT 9312	PROM9	Prochlorales
Prochlorococcus marinus subsp. pastoris str. CCMP1986	PROMP	Prochlorales

Alignments of homologous genes were downloaded from the HOGENOM database. Phylogenetic analysis was performed on two different sets of genes; signature and common genes. Here, we define signature genes as gene families which are present in a single copy in all of the 40 cyanobacteria species and not present in any other organism in the database. We define Common genes as gene families that are present in a single copy in every one of the 40 cyanobacteria and at least 1000 out of 1470 species in the database. There were 27 and 60 gene families in signature class and common class respectively. Concatenated alignments were made for these two sets of genes and phylogenetic analysis was carried out separately on these two concatenated alignments.

For doing the phylogeny we used the popular MrBayes software package (Huelsenbeck & Ronquist, 2001; Ronquist *et al*, 2012). This software is very powerful for Bayesian Phylogenetic analysis using the Markov chain Monte Carlo (MCMC) method (Larget & Simon, 1999). Before the main analysis started, the posterior

probabilities of eight different protein evolutionary models were calculated by MrBayes in order to find which one fits our data better that the others. These models included "Dayhoff" (Dayhoff & Schwartz, 1978), "Poisson"(Jukes & Cantor, 1969), "equalin" (Felsenstein, 1981), "GTR" (Tavaré, 1986), "Blosum" (Henikoff & Henikoff, 1992), "mtmam" (Coa *et al*, 1998), "Rtrev" (Dimmic *et al*, 2002) and finally "WAG" (Whelan & Goldman, 2001). For both the signature genes and the common genes, it was found that the "WAG" model has posterior probability more than 0.95. So, for the final analysis, we used only the WAG model of substitution. Variation in rates across sites was accounted by using a discrete gamma distribution with four rate categories. The shape parameter was initially set to 1 and got optimized during the runs. Each MCMC run consisted of four parallel chains of 5×10^5 iterations and sampling is done every 2000 iteration. The first 25% of the iterations were treated as a burn-in period that was excluded from analysis. The trees are shown in Fig 2-3 (Signature gene tree) and 2-4 (Common gene tree).

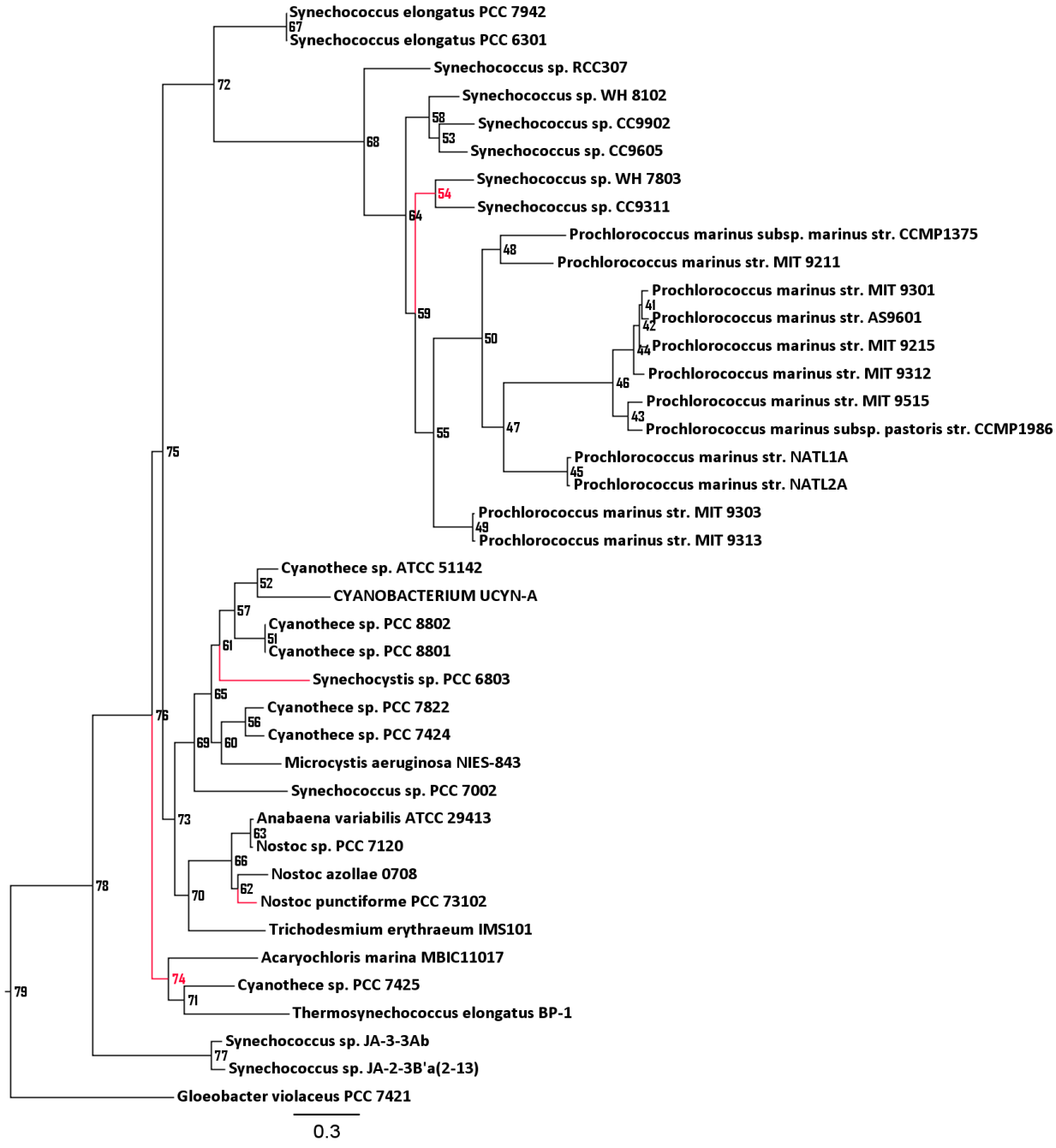


***Fig 2-3**. Consensus tree of signature genes: the red branches are the difference between common genes and signature genes tree. The scale bar for branch lengths is in units of amino acid substitutions per site. Node labels are arbitrary numbers that are used to distinguish tree arrangements. All nodes in this tree have 100% posterior probability according to MrBayes.*

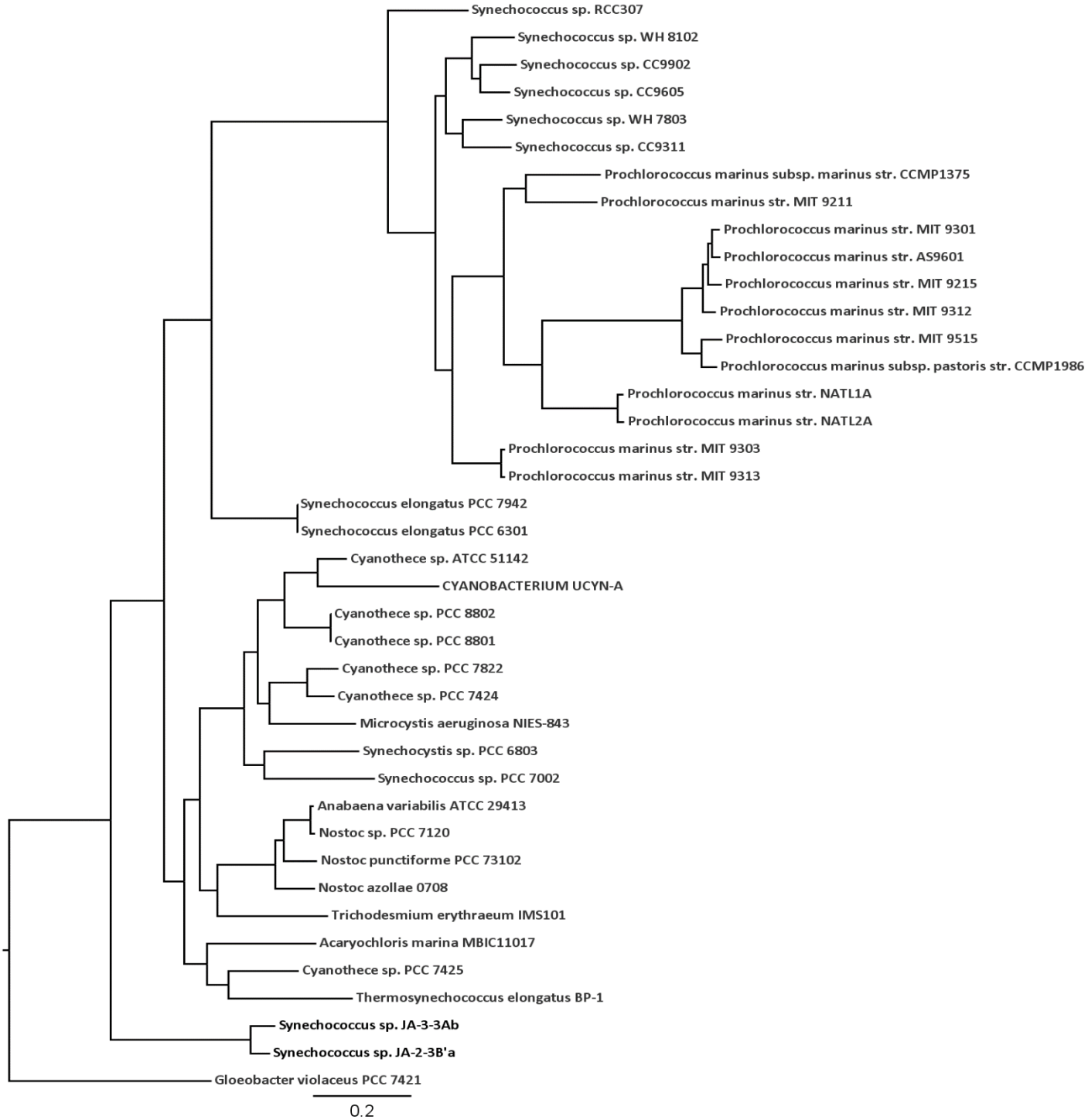


Fig 2-4. *Consensus tree of common genes. The scale bar is in unites of amino acid substitutions per site. All nodes in this tree have 100% posterior probability according to MrBayes.*

The two trees are very similar, but there are four branches highlighted in red in Fig. 2-3 whose position is different in Fig 2-4. The most important difference between the trees is in the position of *Acaryochloris marina MBIC11017*, *Cyanothece sp. PCC 7425*, and *Trichodesmium erythraeum IMS101* (Node 74). These three species also form a clade in other publications, although their position relative to their groups is not always the same (Shih *et al*, 2013; Howard-Azzeh *et al*, 2014). The common gene tree is consistent with some of the recent studies (Larsson *et al*. 2011; Szöllősi *et al*, 2012; Dagan *et al*. 2013), except in the emerging position of *Synechocystis sp. PCC 6803*. On the other hand the signature gene tree is similar to the results of Shih *et al*, 2013, except that there is difference in the position of *Trichodesmium erythraeum IMS101* between them. In nearly all previous mentioned works, the trees are rooted with *Gloeobacter violaceus PCC 7421* as the outgroup, as shown in Fig 2-3 and 2-4. So, we chose to root all our tested trees like that.

Obtaining a Set of 10 Trial Trees

One of the primary goals of chapters 3 and 4 is to investigate the effect of different tree topologies in the likelihood of the different models and HGT. Hence, in addition to our two computed trees, we also used eight alternative trees that are slight variations of these two (shown in Appendix). Two of extra trees are from other literature (Shih, *et al*, 2010 – tree 5, Gupta & Mathews, 2010-tree 10). Three of them are nearest neighbor interchanged trees of the Signature genes tree in nodes 46 (tree 8), 50 (tree 7), 64 (tree 2) and 66 (tree 3). For the last two, we removed *Synechocystis sp. PCC 6803* from its place and put in as out-group of node 65 (tree 9) and regrouped with

Synechococcus sp. PCC 7002 (tree 6). All these changes appeared in several studies about cyanobacteria phylogenies (Gupta & Mathews, 2010; Dagan *et al*, 2012; Szöllősi *et al.* 2012). Also, all these changes are within strongly supported clads of cyanobacteria and none of them is contain moving species from one clade to another major clade (Howard-Azzeh *et al*, 2014). The numbering of trees 1 – 10 is according to their likelihood ranking using the gene presence/absence data that will be discussed in Chapter 3 (Table 3-4).

We calculated the likelihood of each set of gene in all tree to see whether they are ruled out by the different set of genes that we have. We fixed the topology of the trees and used MrBayes for computing likelihood for each tree with the same conditions and parameters that we used to calculate the first two signature genes and common genes trees. MrBayes calculates the harmonic mean of lnL of the last 25% trees of the MCMC chains. Table 2-2 compares these *lnL* values for the 10 trial trees.
A log difference between the likelihood in the range of 3-5 can be considered strong evidence in favor of the better tree, while a log difference more than 5 is very strong evidence (Kass & Raftery, 1995). Therefore, according to signature genes data, all tree are ruled out compare to the signature genes trees execpt the second tree (Table 2-2a). In the case of common genes, all trees has much worst likelihood compare to the common genes tree (Table 2-2b).

Table 2-2. *lnL of common genes and signature genes data and on different trees. (a)The second column shows the difference between the lnL of the common genes tree and the others. (b) The second column shows the difference between the log likelihood of the signature genes tree and the others.*

(a)

Tree-Number	ΔlnL
4	0
2	-133.9
9	-176.6
6	-176.6
3	-183.3
1	-189.7
5	-342.1
7	-631.0
8	-2796.9
10	-3347.0

(b)

Tree-Number	ΔlnL
1	0
2	-2.6
9	-20.5
3	-20.7
5	-95.3
7	-123.6
6	-132.4
4	-227.5
8	-2364.2
10	-3471.1

Branch lengths comparison of Signature gene and Common gene trees

Of the 79 branches in the phylogenetic trees, 73 branches are in equivalent positions in the two trees. For these 73 branches, we compared the branch length in the signature gene tree with the length in the common gene tree (Fig 2-5). The two sets of branch lengths are very strongly correlated, and from the slope of this graph we deduce that the signature genes of cyanobacteria evolve 1.76 times faster than the common genes.

As the common genes are present in more than 1000 species in addition to the 40 cyanobacteria, these are probably important genes that have a function in a majority of species. We would therefore expect them to be under stabilizing selection. Although the signature genes are not found outside cyanobacteria, the fact that they are present in all 40 cyanobacteria suggests that these genes also have an important function in this group of species. The observation that the signature genes evolve slightly faster than the common genes suggests that these are under slightly relaxed stabilizing selection relative to the common genes. However, an alternative explanation could be that they are under positive selection for rapid amino acid change. In future work, this could be investigated by comparison of the rates of synonymous and non-synonymous substitutions (d_S and d_N) in these different groups of genes.

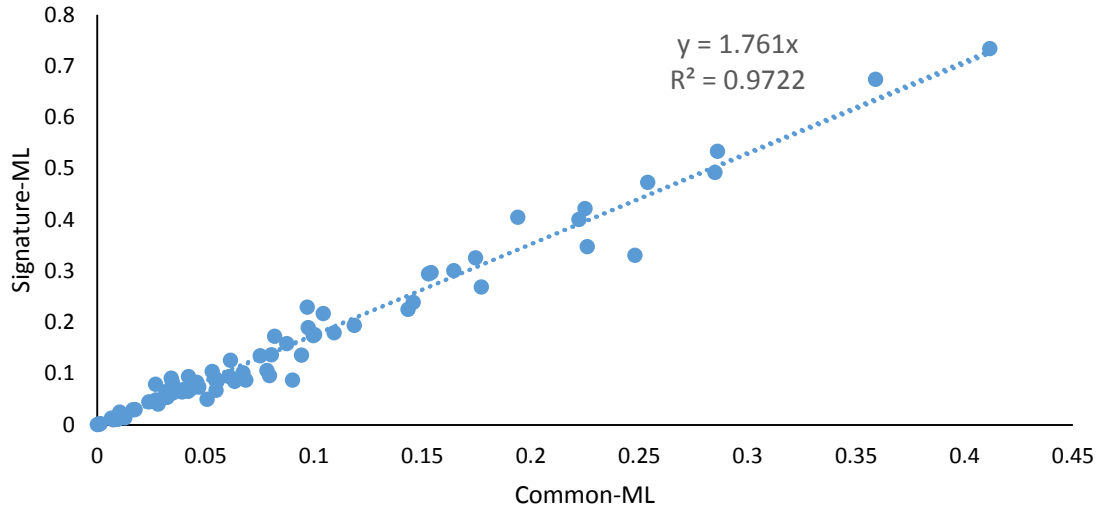


Fig 2-5. *Comparison of branch lengths between equivalent branches of the common gene and signature gene trees. In total, the signature gene tree has longer branches compared to the common gene tree which indicates that the signature genes are evolving somewhat faster.*

Chapter 3 Likelihood of patterns of gene presence and absence

Evolutionary Models for the Gene Presence and Absence Patterns

As discussed in Chapter 1, there have been several previous studies that use models of gene gain and loss to analyze the evolution of gene presence/absence patterns using maximum likelihood methods (e.g., Hao and Golding 2008, Cohen *et al,* 2008). The basic ingredient for most of these methods is a two state model, where 1 means the gene is present and 0 means it is absent. A transition from 0 to 1 means that a gene is gained. This could represent the insertion of a gene via HGT or the gain of a gene via molecular evolution within the lineage of the organisms in the group being studied. In the simplest formalism of these models, a transition from 0 to 1 can occur independently more than once in different parts of the tree. If a gain occurs a second time, this must be due to HGT, because it is assumed to be impossible that the same gene sequence could evolve from scratch more than once. It is important to consider the pool of gene sequences from which a newly inserted gene might arise. This pool consists of all the genes on other bacteria, on viruses and plasmids *etc.* If 0 to 1 transitions are allowed multiple times, this corresponds to the assumption that the pool of genes available for HGT has finite diversity. If a gene is chosen from this pool, it is possible to pick the same gene on separate insertion events. For this reason, we call this the Finitely Many Genes (FMG) model.

Here we are specifically interested in estimating the frequency of HGT. Therefore it is useful to consider a phylogenetic model in which HGT is excluded. This acts as a null model for with which we can compare cases where HGT is allowed. The null model we will use here is called the Infinitely Many Genes (IMG) model. It was introduced by Baumdicker *et al* (2010) and was used by Collins and Higgs (2012) for describing the gene frequency spectrum. Here we wish to use this as a phylogenetic model. The IMG model corresponds to the assumption that the pool of genes available for HGT is infinitely diverse. If we pick a gene from this pool to insert, then we will never pick the same gene more than once. Hence the 0 to 1 transition can only occur once for any particular gene. The IMG model can also be interpreted in a second way. If no HGT is possible at all, then each new gene arises as a result of sequence evolution within the lineage in which it is observed – e.g. a burst of rapid mutations, or gene scrambling event that makes the gene different enough to no longer cluster with other genes. As gene sequence space is so huge, it is supposed that there is always a certain rate of creating new sequences by this means. This means that new sequences arise continuously without HGT, but the same sequence can never arise more than once. This is the IMG limit that we wish to consider here.

To calculate the likelihood of gene patterns with the IMG requires a reformulation of the likelihood algorithm. To explain how this is done, we will discuss the likelihood calculation in the FMG case, and then show how the IMG limit can be derived.

Likelihood Calculation with the FMG Model

In the FMG model, each gene family has an insertion rate a and deletion rate v. We need to calculate the probability, $P_{ij}(t)$ that a transition occurs from state *i* to state *j* in time *t*, where *i* and *j* are 0 or 1. Because each of i and j can have two states then we have to calculate four different probabilities. The probability that a gene family is present is $\pi_1 = a/(a+v)$ and the probability that it is absent is $\pi_0 = v/(a+v)$. We also call the size of the pool that a gene families belongs to M. In other words, *M* is the number of types of genes in the universe that have insertion rate *a* and deletion rate *v*. So the mean number of gene families from that pool per genome is

$$G = M\frac{a}{a+v} = M\frac{a/v}{1+a/v} \tag{3-1}$$

In other words, for a set of *M* gene families with equal insertion and deletion rates, *G* is the number of gene families from *M* that is expected to be observed in a typical genome.

Four transition probabilities for specific insertion and deletion rates are equal to (Ross, 1996):

$$P_{00}(t) = \frac{v}{v+a} + \frac{a}{a+v}e^{-(a+v)t} \tag{3-2}$$

$$P_{01}(t) = \frac{a}{a+v}(1 - e^{-(a+v)t}) \tag{3-3}$$

$$P_{10}(t) = \frac{v}{v+a}(1 - e^{-(a+v)t}) \tag{3-4}$$

$$P_{11}(t) = \frac{a}{a+v} + \frac{v}{a+v} e^{-(a+v)t} \quad (3\text{-}5)$$

Using these probabilities, we can calculate the likelihood of the patterns via the recursion algorithm described in Chapter 2. The expected number of occurrences of a pattern is equal to:

$$N_{pat} = ML_{pat} \quad (3\text{-}6)$$

The sum of the likelihoods over all possible patterns is equal to 1; hence, the sum of the expected number of occurrences is equal to *M*. However we cannot observe the null pattern (genes that are not present in any of the species). Therefore, the expected total number of observable patterns is:

$$N_{tot} = M(1 - L_{null}) \quad (3\text{-}7)$$

Although *M* is not directly observable, we can choose *M* so that the expected number of observed patterns is equal to the number of patterns observed in the data. Hence, the expected number of occurrences of a pattern becomes

$$N_{pat}^{\exp} = N_{tot} \frac{L_{pat}}{1 - L_{null}} \quad (3\text{-}8)$$

and the frequency of the pattern in the observed data is

$$q_{pat} = \frac{L_{pat}}{1 - L_{null}} . \quad (3\text{-}9)$$

To fit the data to the model, we need to maximize the log likelihood of the whole data set, which is

$$\ln L = \sum_{pat} N_{pat}^{obs} \ln q_{pat}. \qquad (3\text{-}10)$$

This modification of the pattern likelihoods to account for the in observability of the null pattern has also been done by Hao & Golding (2008) and Cohen *et al* (2008). In the analysis below, we also decided to exclude patterns that are only present in a single species, firstly, because gene clusters with only one gene are not counted in the HOGENOM database, and secondly, because there may be some doubt as to whether sequences that have no homologs in other organisms are really expressed genes. To exclude the single-genome patterns we calculate the total likelihood, L_{inv}, of all the invisible patterns, *i.e.* the sum of L_{null} and the likelihood of all the patterns with a single 1. The frequency of the observable patterns is then

$$q_{pat} = \frac{L_{pat}}{1 - L_{inv}} \ , \qquad (3\text{-}11)$$

and ln$L$ is as in equation (3-10), but the sum over patterns excludes the patterns with a single 1.

Likelihood Calculation with the IMG Model

In the IMG model, each gene can only originate once but can be deleted multiple times. This is the limit of FMG where $M \to \infty$, and $a \to 0$ with the total rate of insertion of all genes, u, kept constant:

$$u = Ma \tag{3-12}$$

The transition probabilities are:

$$P_{00}(t) = 1 \tag{3-13}$$

$$P_{01}(t) = 0 \tag{3-14}$$

$$P_{10}(t) = 1 - e^{-vt} \tag{3-15}$$

$$P_{11}(t) = e^{-vt} \tag{3-16}$$

For any one type of gene, the equilibrium probabilities of being present and absent are $\pi_1 = 0$ and $\pi_0 = 1$, *i.e.* the gene is always absent! However, there is a continued rate of insertion and deletion of new genes, so the mean genome size is finite: $G = u/v$. In this limit $L_{null} = 1$ and $L_{pat} = 0$ for all other patterns. So, the expected number of occurrences of each pattern cannot be calculated from Equation (3-8). In fact, $N_{pat}^{\exp}$ has a finite value, but it must be calculated in a different way. Suppose there is a set of gene families evolving according to the IMG model with deletion rate v per family and overall insertion rate u. The number of genes that are present in the root is $N_{root} = {}^{u}/_{v}$. Also,

for any non-root node, *n*, the number of gene families arising on the branch leading to this node, N_n, satisfies:

$$\frac{dN_n}{dt} = u - vN_n \tag{3-17}$$

Hence

$$N_n = \frac{u}{v}(1-e^{-vt}) \tag{3-18}$$

In the IMG model, each gene can only originate once. We call call a node an origin node if it is possible that the gene originated on the branch leading to this node. A node is an origin node if it is an ancestor of all the nodes that have a 1 in the pattern, as illustrated in Fig 3-1. Using the notation of chapter 2, $L_1(n)$ is the likelihood of the data pattern arising given there was a 1 on node *n*, and *n* is one of the origin nodes. The expected number of each pattern $N_{pat}^{\exp}$ has finite value and can be calculated as below:

$$N_{pat}^{\exp} = \sum_{n=origin} N_n L_1(n) \tag{3-19}$$

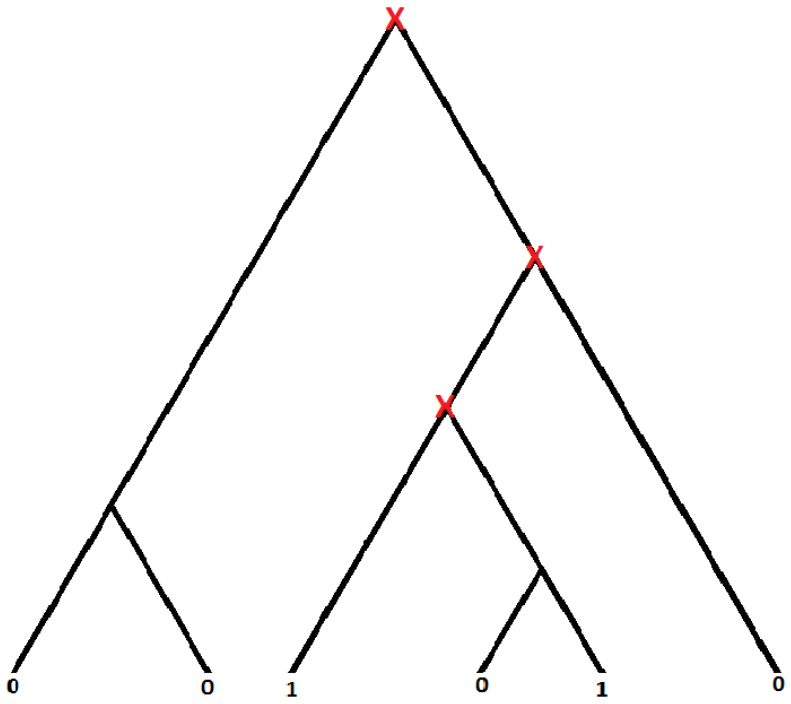


Fig 3-1. *The nodes labelled X are ancestors of all the 1's in the pattern. These are the only possible origin nodes for the gene in the IMG model.*

From this, it is possible to calculate q_{pat} and $\ln L$ as for the FMG.

It is known from previous studies that there are genes that are rapidly inserted and deleted and also genes that are conserved across broad evolutionary distances (Hao and Golding, 2008, Cohen *et al*, 2008). Therefore it is necessary to consider models with more than one category of genes. In all the models considered here, we will use five categories, each with its own deletion rate v. We will choose the v's for the categories so that they are a discrete approximation to a gamma distribution, following the method of Yang (1993). The mean v is constrained to be 1; hence a single shape parameter, α, is required to define all five v's.

For each category, the parameter G is the mean number of genes per genome belonging to this category. In the in the IMG model, the insertion rate is $u = Gv$, *i.e.* u is known, once G and v are specified. In the simplest case, all five categories are constrained to have the same G. The u is different for each rate class because the v is different. Thus, for the model of five IMG classes with equal numbers of genes per class, there are only two parameters: α and G. We also consider relaxing the assumption of equal numbers of genes per class. In this case there are five G parameters, but still only one α parameter. It should be realized that if different G parameters are allowed for each class, then the resulting distribution of deletion rates is no longer an approximation to a gamma distribution, and it is much more general than this. We have retained the method of Yang (1993) to calculate the deletion rates because this allows all five v's to be specified by one shape parameter α, and it means that all five v's change smoothly when α is changed, which helps the program during the optimization to find the maximum likelihood.

For the FMG model, we also need to specify the ratio of insertion to deletion rates, *a/v*. in addition to *v* and *G*. Once these quantities are specified, the number of genes in the finite pool, *M*, is known. Equation (3-1) can be rearranged to give

$$M = G(1+\frac{1}{a/v}). \qquad (3\text{-}20)$$

We use *G* as a parameter rather than *M* because *G* has the same meaning in both FMG and IMG models, and *G* is finite in both models, whereas *M* comes infinite in the IMG limit. We use the *a/v* ratio as a parameter, rather than *a*, because we begin with the assumption that this ratio is constant for each rate class. The simplest FMG model we consider has five rate categories, specified by one α parameter, one *G* and one *a/v*. We also consider generalizations where there is a different *G* for each category and/or a different *a/v* ratio for each category. The most general case has 11 parameters – α, 5 *G'*s and 5 *a/v* ratios.

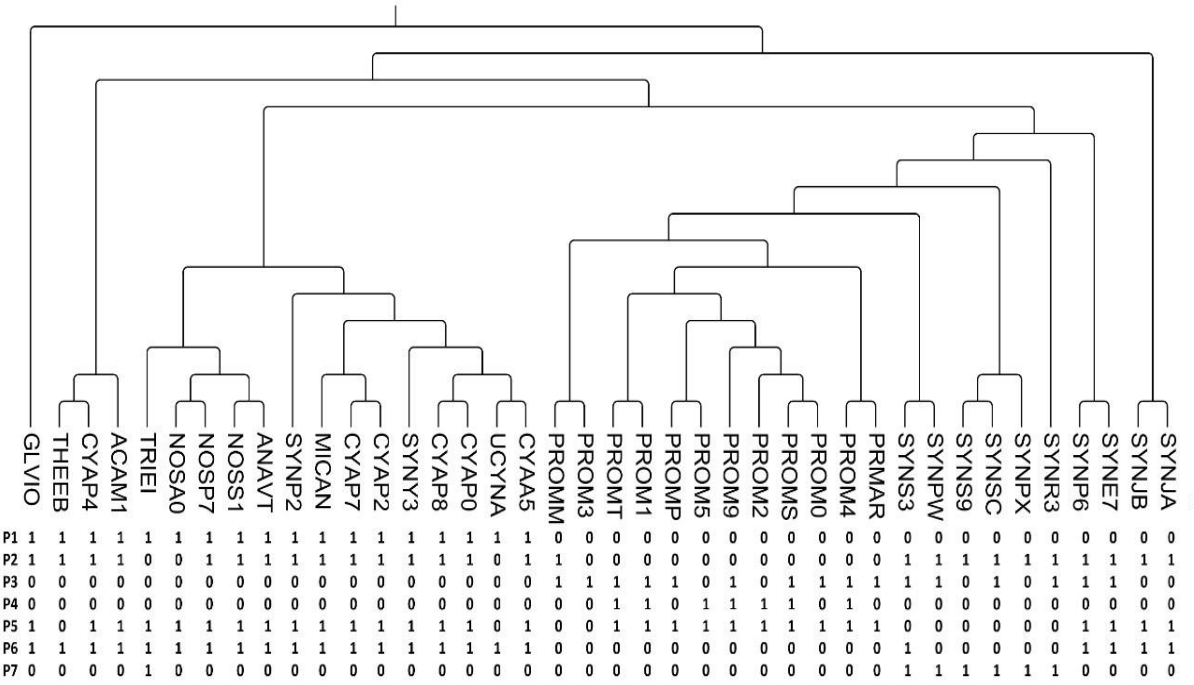


Fig 3-2. *Example of patters of presence and absent of gene families. The patterns that are patchier than the others (P5, P6, P7), have more probability of being subjected to HGT events.*

The ability of the models to fit the data were compared using the Akaike Information Criterion (AIC) (Akaike, 1974).

$$AIC = 2k - 2\ln L \qquad (3\text{-}21)$$

Where k is the number of parameters in the model. The AIC criterion selects models that give high likelihoods but penalizes models that use too many parameters. We can use this

to compare models with different combinations of IMG and FMG rate classes and investigate which one gives the best fit to the data.

We use an optimization program in which the tree topology is specified and does not change. The branch lengths and model parameters are optimized for the fixed tree by making repeated small changes until no further improvement can be found. Repeat runs are made from different starting configurations in order to check that the program is finding the optimal solution.

Model selection for the Cyanobacteria Gene Family Data

We now apply the above methods to the analysis of the Cyanobacteria gene family data obtained from the HOGENOM database (Penel *et al*, 2009). Table 3-1 shows the likelihood of the models defined above using the signature gene tree (Fig 2-3) as the specified phylogeny. The models are listed in order of increasing complexity. In each case, increasing the model complexity leads to a much higher likelihood and a much lower AIC. Thus we conclude, that the FMG models fit much better than the IMG models, that allowing five different G parameters is much better than fixing all G's to be equal (both for the IMG and FMG cases) and that allowing 5 different a/v ratios in the FMG model is much better than fixing all these ratios to be equal. We also tested the models on the common gene tree (Fig 2-4) and found the same ranking of the models.

Table 3-1. *The AIC results for Signature genes tree with 5F and 5I models and different constraint. The model with the least constraint (allowing both Gamma categories and insertion rates to vary independently) has the best AIC.*

Model	**No. parameters**	**Ln(L)**	**AIC**
IMG, same Gs	2	-91818.2	183642.5
IMG, diff Gs	6	-90809.4	181630.9
FMG, same Gs & *a/v*	6	-90416.6	180839.3
FMG, diff Gs, same *a/v*	7	-90240.8	180495.6
FMG, same Gs, diff *a/v*	7	-89858.6	179731.3
FMG, diff Gs & *a/v*	11	-89487.2	178996.4

It is interesting to consider in more detail the way that the likelihood depends on the insertion-deletion ratio. Fig 3-3 shows the log likelihood as a function of *a/v* for the FMG model in which *G* is the same in each category. For each point on the curve, the *a/v* ratio is fixed, and *G* is the same for each category. This was done for both the common gene and signature gene trees. The optimal a/v is nearly the same -0.00704 for the common-gene tree and 0.00709 for the signature-gene tree (Fig 3-3). The common-gene tree is found to have a slightly better likelihood across the whole range of *a/v*, but this does not change the main conclusion regarding the value of *a/v*. The limit of *a/v* = 0 is the IMG model. Thus the figure shows that small but non-zero values of *a/v* are a significant improvement over the IMG model. It is thus clear from these results that the presence of

HGT shows up in the likelihood analysis of the gene family data. By comparing FMG and IMG models in this way, we can reject the null model that excludes HGT. However, it should also be noted that the optimal *a/v* ratio is very small.

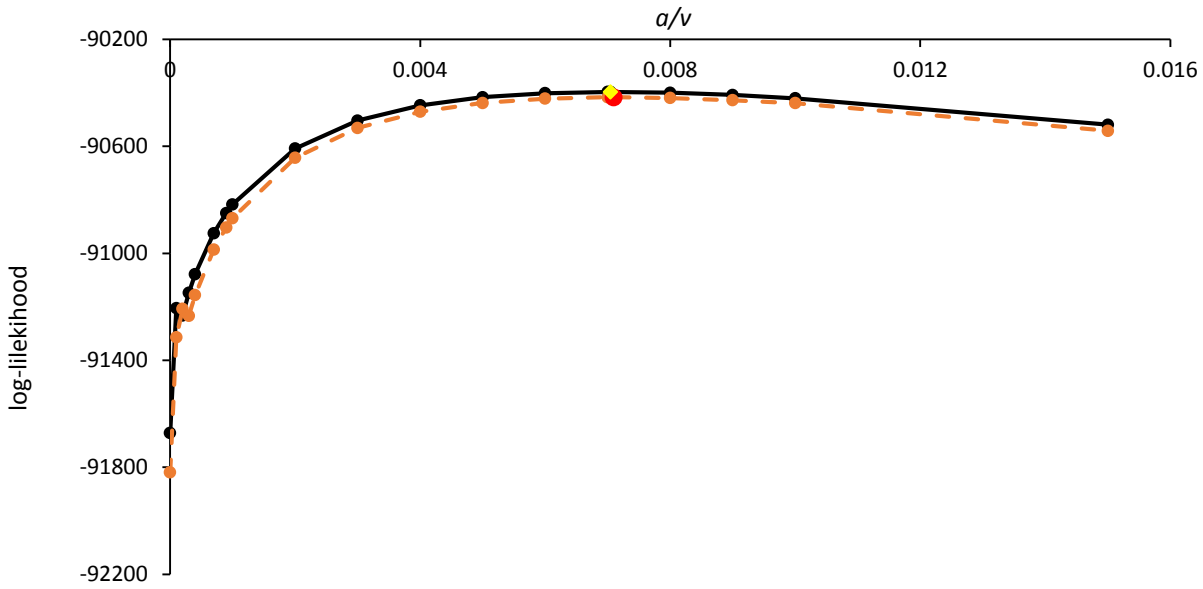


Fig 3-3. *The likelihood as function of a/v rate for the common genes tree (black solid line) and signature gens tree (orange dashed line) when all categories was forced to have equal size. The a/v was changed to different values (black and orange points) and likelihood was computed. The red circle and yellow square are the optimum value for the a/v that maximize the likelihood calculated by program itself which is in agreement with the manual results.*

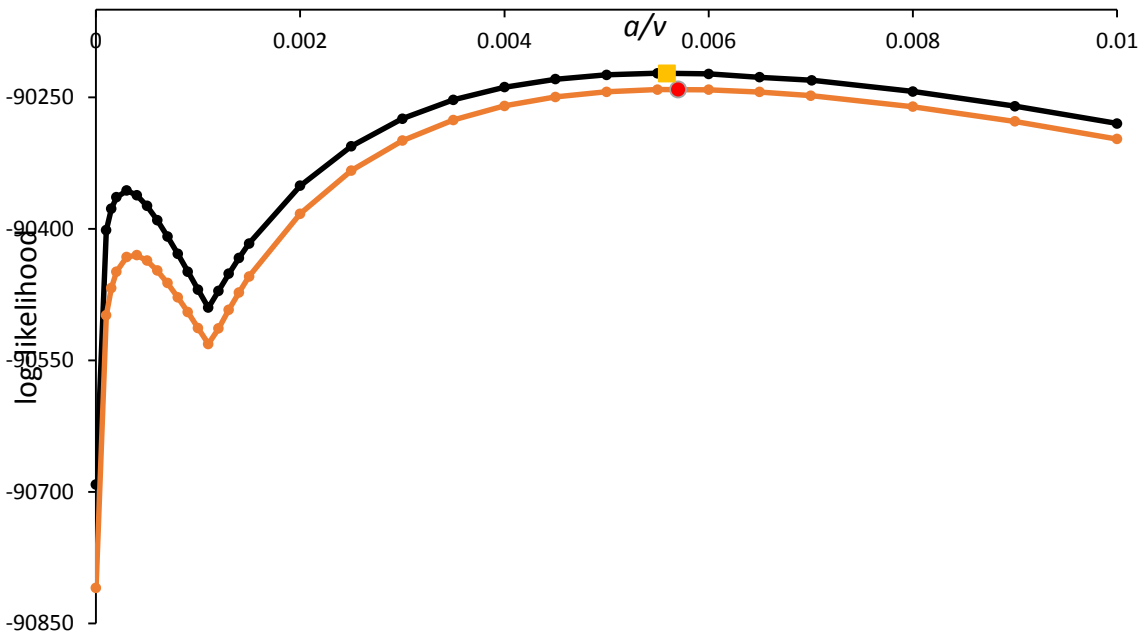


Fig 3-4. *The likelihood under different a/v rate for the common gens tree (black solid line) and signature gens tree (orange dashed line) when the categories could have different size. The a/v was changed to different values (black and orange points) and likelihood was computed. The red circle and yellow square are the optimum value for the a/v that maximize the likelihood calculated by program itself which is in agreement with the manual results.*

Figure 3-4 is similar to Figure 3-3, except that the five rate categories are allowed to have different *G's*. The trends are more or less the same as the previous case. There is a second bump for very small values of a/v which occurs because the optimal values of *G* switch between two alternate states at this point.

From Table 3-1 we concluded that the most general FMG model, where both the *G* and *a/v* parameters are different in each category, gave the best fit. Table 3-2 shows the full set of maximum likelihood parameters for this model on the signature genes tree. It can be seen that the optimal *a/v* ratios are always less than 1. This means that the size of the gene pool *M* for each category, is always much greater than the mean number of genes (*G*) present in a genome. It can also be seen that *a/v* differs substantially between classes. For the second and fifth class, *a/v* is extremely low, and *M* is extremely high as a consequence. For these classes, the FMG model is approaching the IMG limit.

Table 3-2. *Example of the solution for the free parameters in the most general case (five G, gamma distribution shape parameter* α *and five a/v). The Gs have a U-shape distribution. Also, the insertion rate for the second and fifth class are very small* ($\alpha = 0.607$).

cat	***G***	***v***	***a/v***	***M***
1	749.1	0.03	0.01	2857.6
2	583.2	0.20	<1E-7	5.7 x 10^9
3	228.3	0.50	0.37	528.2
4	458.4	1.05	0.03	1.5 x 10^4
5	918.5	2.40	0.0007	2.9 x 10^6

In order to test whether these individual classes are better explained by IMG than FMG, we modified the optimization program so that each gene class could be specified as either FMG or IMG independently. These results are shown in Table 3-3 for both the signature gene tree and the common gene tree. 5F means that all classes were FMG. FIFFF means that only the second class is set to IMG, and FIFFI means that both second and fifth classes are set to IMG. For both these combinations, there is a slight reduction of AIC, *i.e.* a slightly better fit according to this model selection criterion. Thus the *a/v* ratio for these two categories is not significantly different from zero, and the IMG is not rejected as a null model for these categories. The result is the same for the two trees. We also tested other combinations of FMG and IMG categories and found that these were all

worse than the 5F model according to the AIC. Thus the *a/v* ratio is significantly different from zero for the other three categories. All other trial trees have same behaviour.

***Table 3-3**. AIC statistics under different combinations of IMG and FMG for signature genes and common genes trees. Both trees have the same trend. For all of them in the five FMG model, the second category has very tiny insertion rate. Also the FIFFF model has the lowest AIC result which is lower from five FMG case. 5F model is chosen as reference model. For all other models their AIC difference with the 5F is showed.*

Tree-Number	**5F**	**FIFFF**	**FIFFI**
Signature genes	178996.4	-2.0	-2.2
Common genes	179012.2	-2.0	-0.3

Comparison of the Trial Trees Using Presence/Absence Data

We also implemented our models on all other trees and to investigate to what extant the behaviour of the models are depended on the background species tree. We first calculate the likelihood of all trees with the most general model (five FMG with unequal Gs and insertion rate) and numbered the trees according to their likelihood rank (Table 3-4(a)). Although, their likelihoods are very different, all trees showed the same behaviour in the terms of gamma categories distribution and insertion deletion rates as the signature and common genes trees.

Moreover, the likelihood of all trees for more restricted models were calculated and compared with each other (Table 3-4(b), 3-4(c), 3-4(d)). The ranking of the trees' likelihood is not the same as the table 3-4(a). However, tree number 10 has the least likelihood in all cases and trees 4 and 1 are always in the three best trees. These results suggest that choosing the best tree according to the presence absence patterns depends on the constraints of the models.

Table 3-4. *lnL comparison of different trees under different models. The second column shows the difference between the log likelihood of each tree and the best result for each model. (a) Unequal gamma categories and a/v. (b) Unequal gamma categories but same a/v for all genes. (c) Same size gamma categories but unique a/v for each category. (d) Same size gamma categories and a/v.*

(a)

Tree-Number	ΔLL
1	0
2	-1.1
3	-6.4
4	-7.9
5	-20.5
6	-21.9
7	-29.2
8	-34.0
9	-51.8
10	-267.2

(b)

Tree-Number	same Gs & a/v
4	0
6	-5
1	-21.7
3	-25.3
2	-26.8
9	-34.4
7	-48.2
5	-62.4
8	-85.4
10	-298.3

(c)

Tree-Number	diff Gs, same a/v
4	0
6	-3.1
1	-20.5
3	-25.0
2	-26.2
9	-34.5
7	-47.8
5	-55.7
8	-79.3
10	-303.7

(d)

Tree-Number	same Gs, diff a/v
4	0
1	-3.1
2	-6.2
6	-9.9
3	-13.4
7	-25.8
5	-34.9
9	-45.8
8	-51.2
10	-259.9

Gene Frequency Distribution

In previous work with the IMG model (Collins and Higgs 2012) the model parameters were estimated by fitting to the gene frequency distribution *G(k)*. In this work we used the likelihood of the pattern data for fitting the model. Nevertheless, once the parameters are determined, we can calculate the predicted *G(k)* for the optimum parameters. Figure 3-5 shows the results for the five FMG and five IMG models on the signature gene tree. The five IMG model fits the data noticeably less well than the five FMG model. The variations such as FIFFI that have very similar likelihoods to the five FMG model also have very similar *G(k)* curves (not shown here).

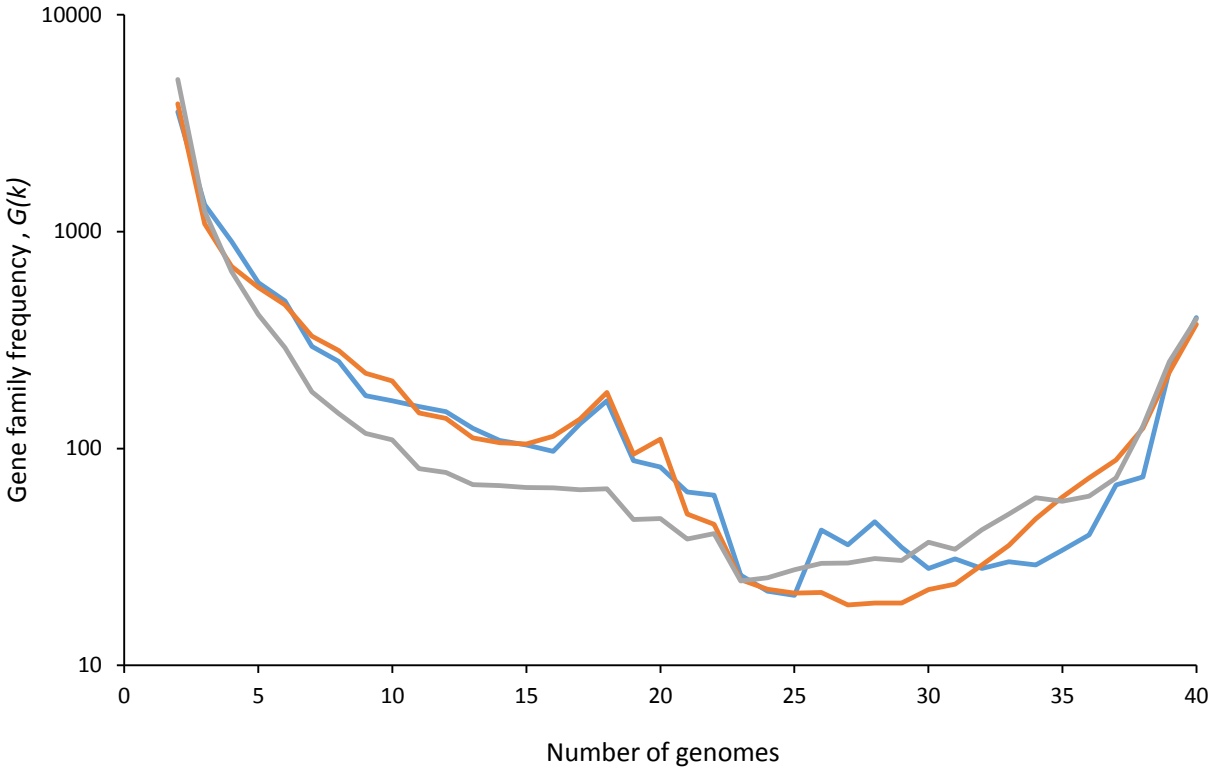


Fig 3-5. *The gene frequency distribution for five IMG and five FMG models. The frequency spectrum was calculated from two different models. Observed data are shown as blue line. The red line is for the five FMG model. Five IMG model is represented as gray line.*

Chapter 4 Horizontal Gene Transfer in Cyanobacteria

Different Scenarios for Gene Evolution

We define three scenarios for evolution of each gene family; scenarios 0, 1 and 2 (Fig 4-1). Scenario 0 consists of the gene families that were present at the root and continually present in all the internal nodes leading from the root to the species that possess that gene (as in Fig 4-1(a)). Scenario 1 consists of gene families that were not present at the root and were gained only once at some internal node of the tree. The requirement that the gene was only inserted once is maintained by insisting that there is an unbroken pathway of 1's on the internal nodes that connect the tips that have that gene family (as in Fig 4-1(b)). These two scenarios can arise even if there is no HGT, and they can arise in the IMG model. Scenario 2 consists of all other possible cases. This includes genes that were not present at the root and were inserted at least twice within the tree, and genes that were present at the root and were inserted at least once at some other point on the tree. In Scenario 2, there is no unbroken pathway of 1's that connects all the 1's at the tips (as in Fig4-1(c)). Gene histories in Scenario 2 can only arise if HGT is occurring. They can arise with the FMG model but not the IMG model.

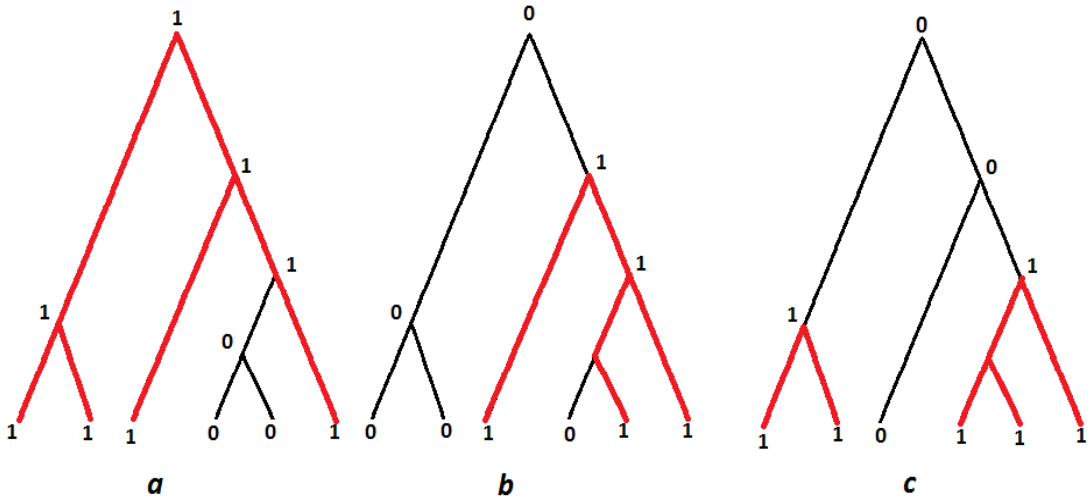


Fig 4-1. *Example of scenarios: (a) Scenario zero, there is pathway of ones that connect all the ones in the tips up to the root. (b) Scenario one, there is pathway of ones that connect all the ones in the tips up to one of the internal nodes (Single insertion). (c) Scenario two, there isn't any pathway of ones to connect all the one in the tips (multiple insertions).*

After finding the maximum likelihood parameters for the data (as in Chapter 3), for each pattern, we calculated the posterior probability of being in each scenario. The three posterior probabilities p_0, p_1 and p_2 add up to one for each pattern. If the posterior probabilities are summed over all the patterns, this gives the expected number of gene families whose patterns are explained by each of the scenarios, as shown in Tables 4-1 and 4-2. The sum of the number of gene families in the three scenarios is equal to the total number of gene families in the data, 10304, in each case.

Table 4-1 shows the effect of changing models on the number of gene families in each scenario. Likewise, table 4-2 reflects the expected number of gene families in each scenario for different background tree. The number of genes families in each scenario is nearly independent of model and tree.

Table 4-1. *Expected number of gene families in each scenario for the different models evaluated on the signature gene tree. The underlying tree is the signature genes tree.*

Scenario	**same Gs & a/v**	**Same Gs, diff a/v**	**diff Gs, same a/v**	**FIFFI**
0	2315.4	2128.9	2448.2	2141.6
1	6464.3	6550.7	6482.9	6613.1
2	1524.2	1624.3	1371.9	1549.2

Table 4-2*. Expected number of gene families being in each scenario for the different trees with 5F model. This results shows that the distribution of gene families in three scenario is nearly independent of the underlying tree.*

Tree No.	***1***	***2***	***7***	***10***
scenario 0	2186.1	2164.6	2168.2	2139.8
scenario 1	6559.5	6601.9	6582.5	6584.3
scenario 2	1558.5	1537.5	1553.3	1579

One interesting thing to investigate is the contribution of each scenario in the different part of gene frequency distribution. We calculated the sum of the posterior probabilities for all genes present in *k* genomes. These sums were then divided by *G(k)* to obtain the probabilities that genes in *k* genomes are in scenarios 0, 1 and 2, as shown in Fig 4-2.

It is found that genes present in almost all genomes ($k \geq 35$) have a high probability of being in Scenario 0. This is what we would expect, because genes present in large numbers of species are likely to have been present at the root. It is found that genes present in small numbers of genomes ($k \leq 5$) have a high probability of being in scenario 1. Genes that were inserted once in the fairly recent past will be found in a small group of closely related species. On the other hand, if a gene were found in a small number of genomes that are widely separated on the tree, this would be explained by Scenario 2, and would require HGT. Thus the fact that most genes with small k fall into Scenario 1 tells us that the presence and absence patterns are quite consistent with a tree-like picture of evolution, and that HGT is not frequent enough to destroy this signal.

It can also be seen in Fig 4-2 that the frequency of Scenario 2 genes is highest in the intermediate part of the spectrum k = 20-30. Thus there is a significant probability of HGT for these intermediate genes. However, it should be remembered that there are rather few genes at these k values because $G(k)$ is U-shaped (as in Fig 3-5).

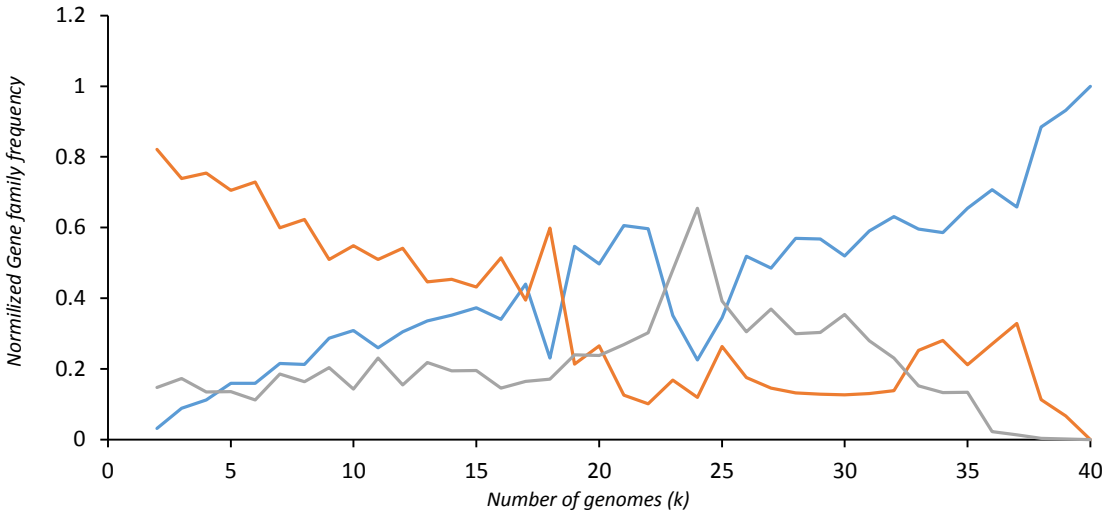


Fig 4-2. *Probability distribution of each part of gene frequency distribution being explained by each scenario; Scenario zero (Blue line), Scenario one (Orange line) and Scenario two (gray line).*

Branch Lengths Measured via Gene Gain and Loss

In standard phylogenetic trees obtained from protein sequence data, branch lengths are measured in terms of numbers of substitutions per site. Hao and Golding, 2006 used maximum likelihood method to infer insertion/deletion rates on each branch of a phylogenetic tree of 13 closely related bacteria. They found that genes insertion/deletion happens at the same or greater than the rate of nucleotide substitution. They also suggest that the branches that lead to the tips have higher rate of insertion/deletion compare to interior branches. In our model, the average gene deletion

rate is set to 1; hence the time unit for branch lengths is the mean time for deletion of an average gene. So, here we fixed the gene deletion rates for the whole tree and try to find the optimized branch lengths. Figure 4-3 shows the signature genes tree with the optimized branch lengths with the five FMG model.

Among the 40 species, *Cyanobacterium UCYN-A* has very long branch which could be due to its very small genome size compare to its neighbors. Because it has an unrealistically long branch, to make comparison between these branch lengths and protein evolution branch lengths, we exclude *Cyanobacterium UCYN-A*. From the slope of Fig 4-4, the gene deletion branch lengths are about 1.4 times longer than the branch lengths on the signature gene tree for protein evolution for the same tree which means that gene deletions are fast - comparable to rate of amino acid substitutions.

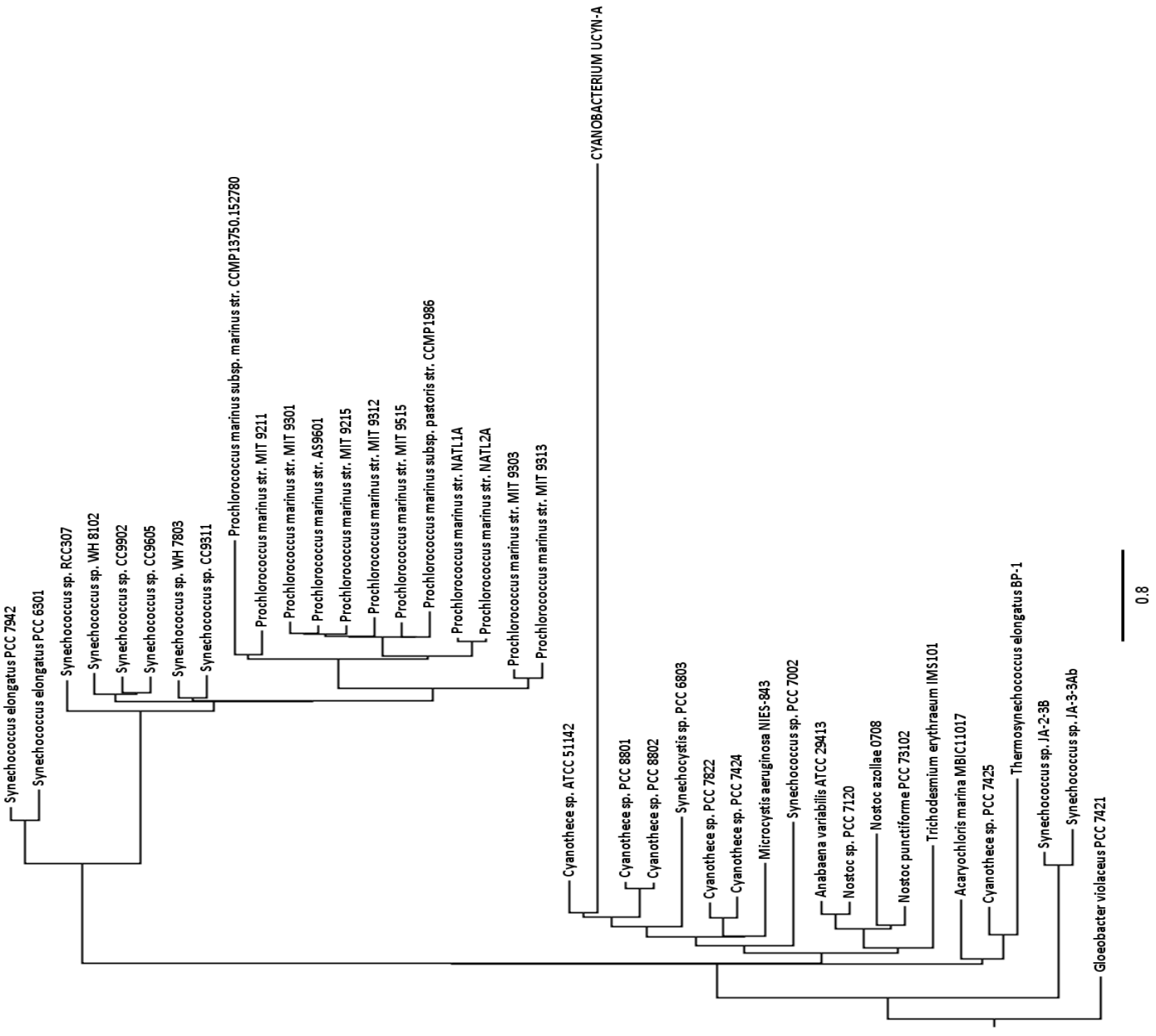


Fig 4-3. *Signature genes tree with branch lengths in the unit of gene insertion and deletion.*

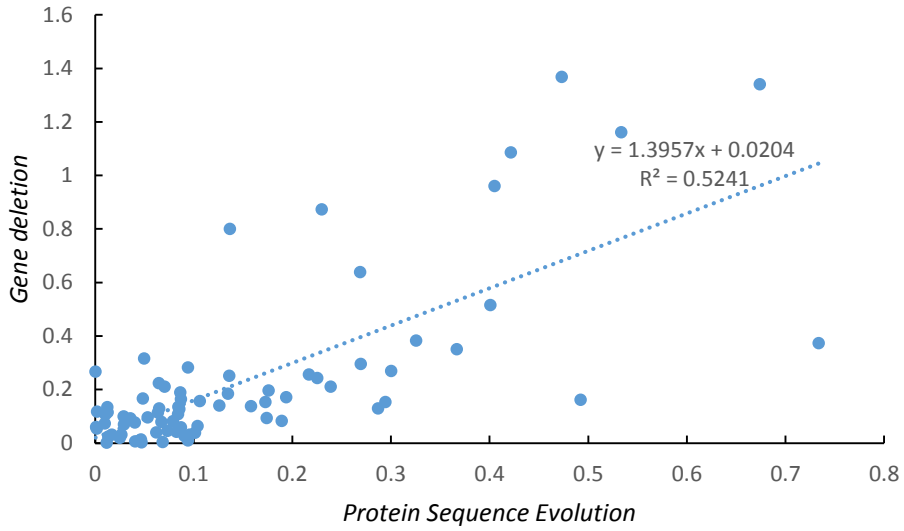


Fig 4-4. *Comparison between gene deletion and protein sequence evolution branch lengths excluding UCYNA.*

One limitation of our model is that it assumes the *G* parameters are constant in all parts of the tree. It is clear from Fig (2-2) that genome sizes differ widely among species, and this is not captured by our model. Among all tested genomes, *Cyanobacterium UCYN-A* has the shortest genome size while its neighbors have relatively large genome. It is also, the only cyanobacteria that lose its oxygenic photosynthetic cycle. Also, its branch length is extraordinary long when branch lengths get optimized (Fig 4-3).

To test whether the maximum likelihood parameters of the model are sensitive to the presence of outliers of small genome size, so, we decide to exclude it from the best tree and input data to test what is the effect of this species in distribution of the gamma rates and scenarios. Removal of *CYANOBACTERIUM UCYN-A* has not any meaningful

effect on the portion of the gene families in each gamma category, however, it increased the number of the gene families in the zeroth and second scenarios by 10 and 1.6 percent of the total gene families, respectively (Table 4-3). Although this does not make much difference to the main conclusions of this work, it does suggest that significant quantitative improvements could be made to the ability to fit the data if we account for variation of genome sizes across the tree in a more complicated model in the future.

Table 4-3. *Expected number of the gene families in each scenario in the signature genes tree with UCYN-A and without it.*

No. Species	**40**	**39**
Scenario 0	2186.1	3227.7
Scenario 1	6559.5	5350.7
Scenario 2	1558.5	1716.6

Predicting the Number of Gene Families with Only One Member

HOGENOM does not include gene families with only a single gene. Therefore, in all of the previous analysis, we only fit the model to gene families present in two or more genomes. Nevertheless it is possible to calculate how many single gene families there are in each genome by subtracting the number of genes in the clusters in the HOGENOM data from the total number of protein coding genes in each genome. From this the number of clusters present in a single species, *G(1)*, can be obtained.

When fitting the models to the data, the expected number of genes in each gamma category of the model are normalized so that the expected total number of observable patterns is equal to the number of observed patterns in the data, *i.e.* $\sum_{k\geq 2} G(k)$ is the same in the data and the model fit. Even though the single-species clusters are excluded from the fitting process, the model gives a prediction of what *G(1)* should be. The orange curve in Fig 4-5 has been extended to include *G(1)*. Surprisingly this fits the data almost exactly. Having calculated the numbers of single gene clusters in each genome, as described above, it is possible to estimate the effect of these clusters on the fitting process. We repeated the analysis treating single gene clusters as observable data. This produced very little change in the predicted *G(k)* (gray line in Fig 4-5).

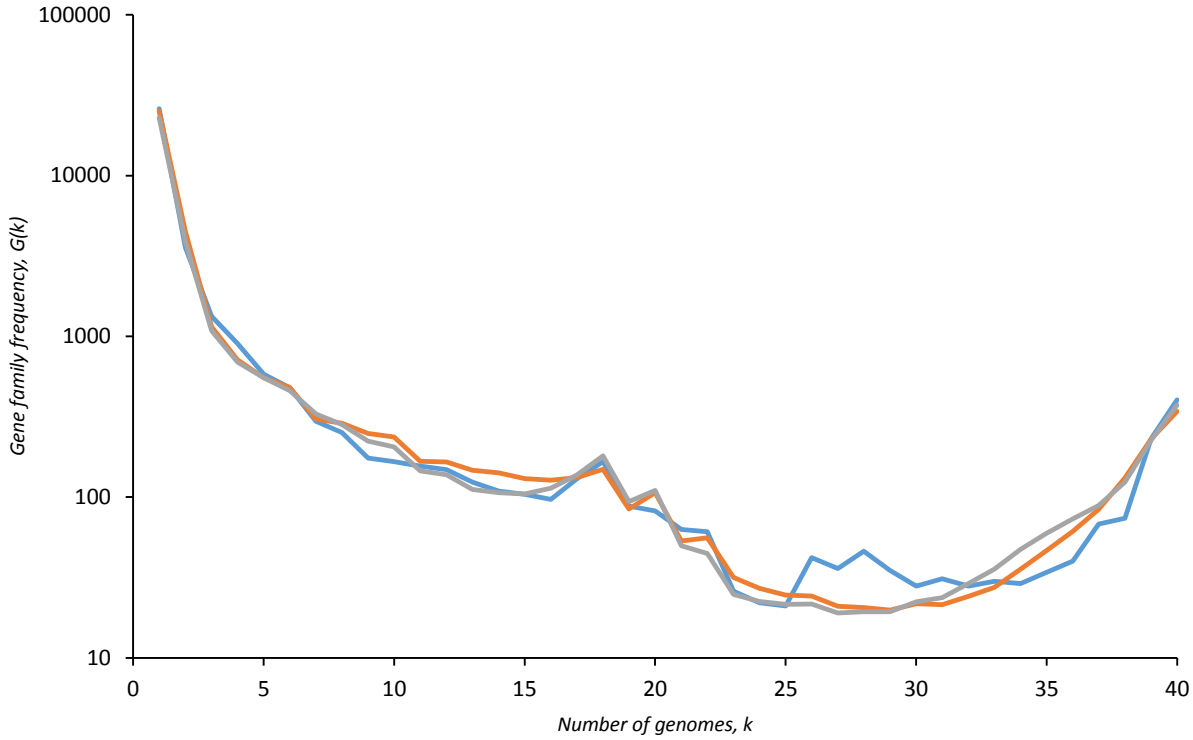


Fig 4-5. *Gene frequency distribution for observed data (Blue line) prediction with single member families (Orange line) and without them (Gray line).*

The Relationship between Ancestral Genome Size and HGT

Using a parsimony analysis of gene presence and absence, Dagan and Martin (2007) showed that the estimated sizes of the ancestral genomes of various groups of bacteria were strongly sensitive to the relative weighting of HGT and gene deletion events. If HGT events were too few, the ancestor genome size became unrealistically large compared to modern genomes, whereas if HGT was too frequent, the ancestral genome size became much smaller than modern genomes. We therefore wished to see whether this effect was also visible using our method.

In order to find the ancestral genome size of cyanobacteria, we calculated the posterior probability of each gene family being present at the root. Adding all of these gives the estimated ancestral genome size. This number includes Scenario 0 genes and also a certain fraction of scenario 2 genes. To test the effect of HGT on this quantity, we fixed the insertion-deletion ratio to be equal for all gamma categories and changed that ratio from zero to 0.02. This was done with the *G* parameter equal and the *G* parameters independent (in the same way as Figures 3-3 and 3-4). As this ratio increases the number of gene families in the second scenario increases too (Fig 4-6 & 4-7). On the other hand, the size of scenario zero decreased as well as the ancestral genome size. The sum of the posterior probability of being in the zeroth scenario is less than the ancestor genome size because it is more restricted than the ancestor genome size.

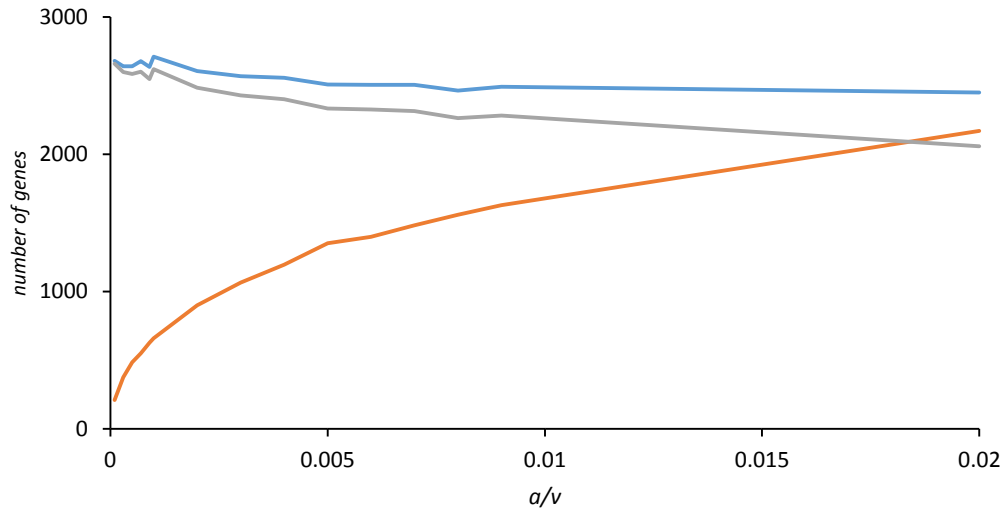


Fig 4-6*. Expected number of gene families in second scenario (Orange line), zeroth scenario (Gray line) and the ancestral genome size (Blue line) as function of a/v when the size of the gamma categories is equal.*

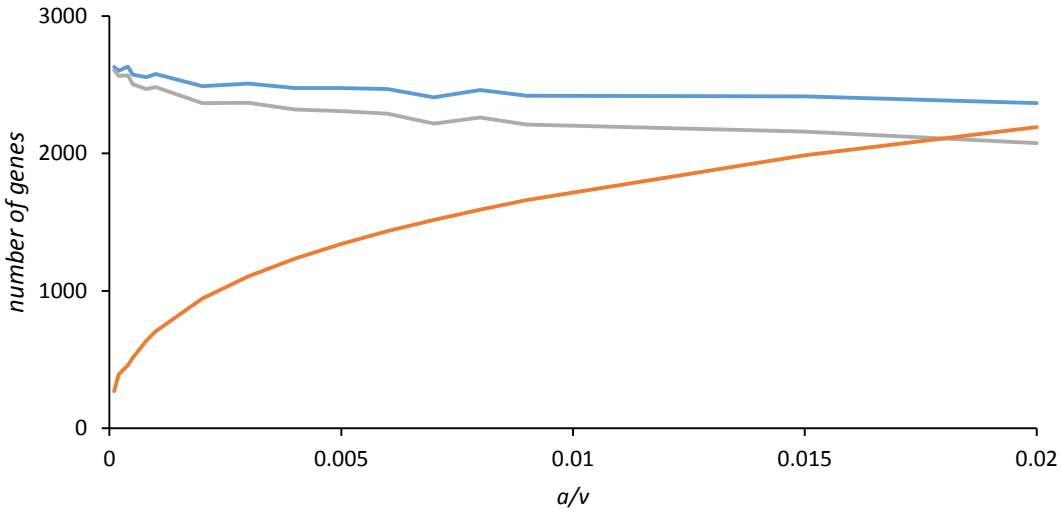


Fig 4-7. *Expected number of gene families in second scenario (Orange line), zeroth scenario (Gray line) and the ancestral genome size (Blue line) as function of a/v when the size of the gamma categories is not equal.*

However, the decrease in the size of ancestor genome is not very significant while the number of gene families in the second scenario almost becomes ten times bigger. One possible explanation for this is that the program assumes the genomes through the tree are all in equilibrium which means that the program assumed that the frequency of gene family category is constant all over the tree. Hence, it tries to keep the ancestor genome size as close as possible to the average of genome size of the 40 cyanobacteria.

Effect of Changing the Root on HGT

We have found the candidate gene families that display HGT by looking at their patterns of presence and absence. But, these patterns depend on where the root is. Hence, we also tested the effect of changing the root in the number of gene families in each possible scenario. First the best tree is rooted on the *Gloeobacter violaceus* (node 79) branch based on the previous papers ((Larsson *et al*, 2011; Dagan *et al,* 2012; Shih *et al*, 2013). Also, we rooted the trees from nodes 55, 66, 74, 75 (mid root) and 78 (Gupta & Mathews, 2010).

Table 4-4. *Expected number of gene families in each scenario for different roots. The underlying tree is the signature genes tree. The numbering of the nodes is as in Fig 2-3. Number of genes in the second scenario is not sensitive to the position of the root.*

Root node	**55**	**66**	**74**	**75**	**78**	**79**
Scenario 0	1823.7	4169.3	2954.7	1865.9	2575.7	2186.1
Scenario 1	6929.9	4583.7	5794.8	6887.4	6146	6559.5
Scenario 2	1550.4	1550.9	1554.5	1550.7	1582.2	1558.5

Although, the number of the gene families in the zeroth and first scenario were changed, their sum was nearly independent of the root (Table 4-4). These results show that the rate of HGT does not depend on where the root is. There is a significant change in the size of first and zeroth scenarios when the tree is rooted from nodes 66 and 76. Node 66 is in the middle of *Nostocales* that have lots of genes in common with each other. So, when the tree is rooted from node 66, all these genes are moved to the root. Therefore, size of the zeroth scenario increased while the size of the first one decreased.

Testing a Reshuffled Tree

One of the arguments made against the use of a bifurcating tree for prokaryotic evolution is that the rate of HGT is too high that it totally dilutes this metaphor (Kurland *et al,* 2008). So, different trees should not show more difference it the number of HGT events and consequently likelihood of the genes absence and presence patterns. Although our tested tree showed meaningful dissimilarity in terms of likelihood and other measured

parameters, we create totally wrong tree by moving all species to a different place. The results from this tree are much worse than the other trees. For example, the number of genes families in the second scenario growth dramatically or the gene frequency distribution become obviously incorrect (Fig 4-8).

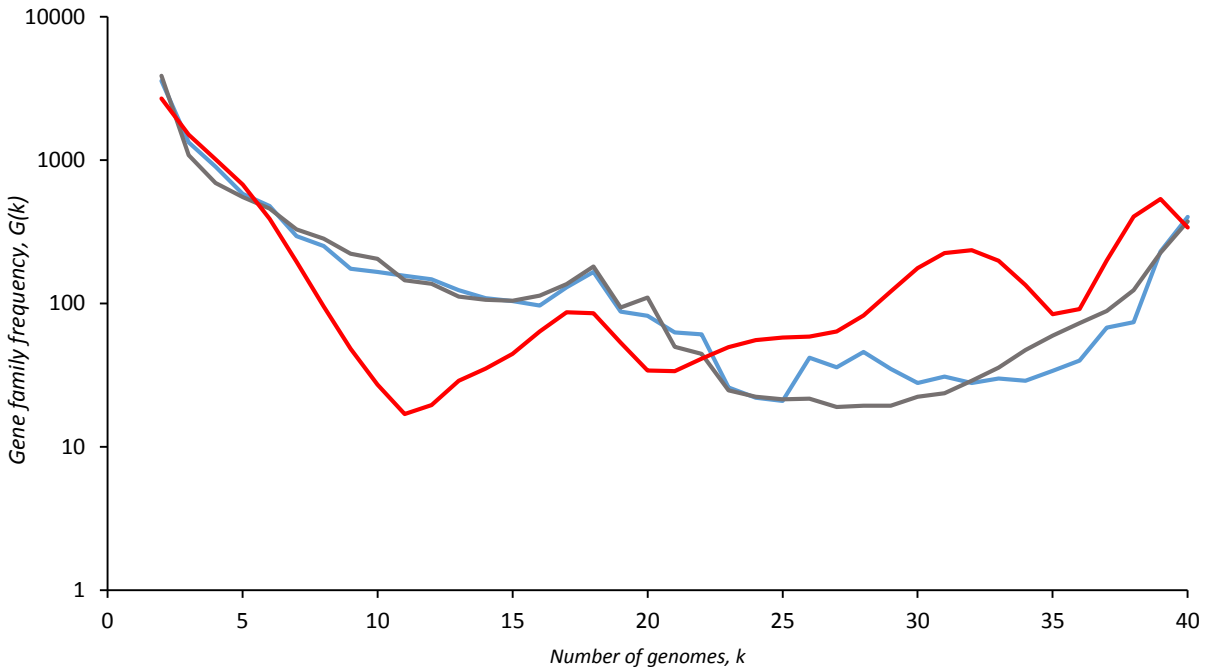


Fig 4-8. *Gene frequency distribution for observed data (Blue line) prediction with underlying wrong tree (Red line) and with signature genes tree as background (Gray line). Five FMG model was used to calculate the distributions.*

These discrepancies between wrong tree and other trees results suggests that although it is tough to find the real species tree, not any tree could be used as a background. One of the interesting results of the wrong tree is the gene family distribution between the scenarios. There is meaningful growth in the number of gene

families in the second scenario which means that there should be enormous amount of HGT (Table 4-5).

Table 4-5. *Comparison between the scenarios' size for the two different trees with five FMG model.*

Scenario	***Reshuffled tree***	***Signature genes***
0	1235.1	2186.1
1	224.3	6559.5
2	8844.6	1558.5

Discussion

We obtained the patterns of presence and absence of the gene families across the 40 species of cyanobacteria from HOGENOM database. The phylogeny of these 40 species was obtained by using Bayesian maximum likelihood criteria for two different sets of genes: Signature and Common. Signature genes are the gene families that are present only in these 40 species and nowhere else in the database. Common genes are the gene families that are present in all 40 genomes and at least 1000 organisms' other than cyanobacteria in HOGENOM database. Most of the signature genes are responsible for the secondary functions of the cells or doesn't have known function (Appendix, Table 1). However, their presence in all 40 cyanobacteria indicates that also have considerable functions in this group of cyanobacteria (Appendix, Table 2). On the other hand, nearly all of the common genes have primary function in the cell and are vital for the cell to survive. Despite the fact that the trees from the signature genes and the common genes data are not very different from each other, according to the likelihood test they are not consistent with each other which means that the difference between the likelihood of the same input data (signature genes or common genes) on the two trees is very large and one of them is ruled out by the other one.

For prediction of presence-absence patterns of gene families we used IMG and FMG models and tried to combine them to find which fraction of a typical genome in our group of species could be described better with IMG or FMG. We also investigated the effect of different constraints on the likelihood of the model. Our model has five different deletion rates that are calculated from a discrete gamma distribution. So, a typical

genome is partitioned into the five different categories with different deletion rates and each category could follow IMG or FMG (which has an insertion rate too). We used these evolutionary models to calculate the likelihood of each observed patterns and estimate HGT. Some of the well stablished works in this field used maximum parsimony to find the HGT events, which has several disadvantages. The most important drawback of the maximum parsimony is that there isn't any way to find the best cost function unless you have the real answer. So, you can't optimize your evolutionary model and cost function. However, maximum likelihood criteria is a very powerful tool for comparing different models and optimize the free parameters during the run time.

First we compared five IMG versus five FMG models under different constraints. In all cases, five FMG models were by far better than five IMG model. However, that doesn't mean that all gene families have exhibited HGT in their history. Although most of the genes tend to be described better with FMG model, about half of them have a small insertion rate that is practically zero. Our best model is the model with three FMG and two IMG categories. Also, our analysis indicates that about 15% of gene families have experienced HGT in their history. Some sequence composition studies of cyanobacteria suggest that between 9.5% and 17% of the genes in their genome is gained through HGT process (Nakamura *et al,* 2004). Our method prediction about the portion of genes that have been affected by HGT is lower than most of previous analysis with different methods (e.g., Zhaxybayeva *et al,* 2006; Shi and Falkowski, 2008; Dagan et al, 2013; Sjostrand *et al,* 2014). We do not see HGT events that lead to a second member of a family or replace an old member by a new member. Also, because, our method is

working in the level of gene families and deals only with presence/absence patterns, it cannot see any case of conflicts between individual genes sequence characterization and host genome. Therefore, we only see events that introduce a new family to a genome. Taken together, we infer no HGTs from phylogenetic and sequence conflicts; hence, our approach delivers conservative HGT rate during cyanobacteria genome evolution.

Besides our two main trees, we also tested eight different trees with slightly different topologies to test the impact of the background tree on selection of the best model as well as frequency of HGT. All the trees showed the same behaviour and have a similar optimum value of *a/v* and a similar fraction of genes belonging to scenario 2. Thus, in determining the number of genes that exhibit HGT, small rearrangements of the background species tree are not very important. We also find the fraction of scenario 2 genes was almost independent of the root. These two results are similar and in agreement with previous studies (Dagan and Martin, 2007; Cohen & Pupko, 2010). However when all the species are reshuffled around the signature genes tree, the portion of gene families that showed HGT increased dramatically.

In general, this work suggests that although it is not possible to ignore HGT, there is still strong signal of an underlying tree. Also, because there is a meaningful difference between the likelihood of different trial trees, presence-absence patterns can be used for phylogenetics by using maximum likelihood criteria. However, selecting the model and its constraint are more important than the tree in determination of the fractions of genes in the three scenarios.

There are two main directions that can be pursued in the future works; First, one important issue that we didn't check in this work is the effect of different clustering methods on the selection of the best model as well as HGT. It has been suggested that different clustering methods could affect the frequency of HGT because they directly alter the size of clusters (Dagan & Martin, 2007). The other possible direction is to test our method with different group of species with different evolutionary distance and taxonomy depth to investigate HGT rate changes during the history of bacterial evolution.

Appendix

List of Signature and Common Genes

Table 1. List of signature genes with their definition

HOG ID	definition
24480	Cell division protein sepF
34199	ATP synthase F ATP synthase subunit b 2
69499	NAD
69835	NAD
232399	HOG000232399 40 40
232400	N utilization substance protein B homolog
232425	HOG000232425 40 40
232461	UPF0367 protein A9601_01421 UPF0367 protein AM1_1885 UPF0367 protein Aazo_3777 UPF0367 protein Av
232464	HOG000232464 40 40
232468	HOG000232468 40 40
232503	Global nitrogen regulator
232616	HOG000232616 40 40
232667	HOG000232667 40 40
232810	HOG000232810 40 40
232833	HOG000232833 40 40
232857	Acetazolamide conferring resistance protein zam
233073	Inner membrane protein oxaA
233121	HOG000233121 40 40

233124	Drug sensory protein A EC=2.7.13 3
233180	ATP synthase protein I
233186	HOG000233186 40 40
233194	Ribonuclease E homolog Short=RNase E EC=3.1.26 12
233206	HOG000233206 40 40
233211	HOG000233211 40 40
233213	HOG000233213 40 40
233235	HOG000233235 40 40
233244	tRNA

Table 2. List of common genes with their definition

HOG ID	Definition
8988	2-oxoglutarate carboxylase small subunit EC=6.4.1 30S ribosomal protein S15 Acetyl-/propionyl-coe
9628	30S ribosomal protein S19 SsrA-binding protein 1 SsrA-binding protein 2 SsrA-binding protein 4 S
10093	ATP-dependent Clp protease ATP-binding subunit ClpX 1 ATP-dependent Clp protease ATP-binding subuni
10898	Cardiolipin synthase Short=CLS EC=2.7.8 - AltName: Probable cardiolipin synthase 1 Short=CLS EC
16242	50S ribosomal protein L10 Peptidyl-prolyl cis-trans isomerase EC=5.2.1 8 tRNA
19780	39S ribosomal protein L17 mitochondrial Short=L17mt 50S ribosomal protein L17 4 50S ribosomal pr
19802	5-tetrahydropyridine-2 6-dicarboxylate 30S ribosomal protein S9 2 30S ribosomal protein S9 4
20624	Elongation factor 4 1 Short=EF-4 1 EC=3.6.5 n1 AltName: Elongation factor 4 2 Short=EF-4 2 EC=3

35046	39S ribosomal protein L20 mitochondrial Short=L20mt 50S ribosomal protein L20 2 50S ribosomal pr
36164	Signal recognition particle 54 kDa protein chloroplastic Signal recognition particle protein AltNa
38087	Carbamoyl-phosphate synthase arginine-specific small chain Carbamoyl-phosphate synthase pyrimidine-
39067	2-C-methyl-D-erythritol 2 4-cyclodiphosphate synthase 30S ribosomal protein S7-1 30S ribosomal pro
39879	30S ribosomal protein S13 2 30S ribosomal protein S13 4 30S ribosomal protein S13 AltName: Full=BS
39903	50S ribosomal protein L6 3 50S ribosomal protein L6 4 50S ribosomal protein L6
40063	28S ribosomal protein S12 mitochondrial Short=MRP-S12 30S ribosomal protein S12 1 30S ribosomal
40097	30S ribosomal protein S15 1 30S ribosomal protein S15 2 30S ribosomal protein S15 3 30S ribosomal
40108	50S ribosomal protein L35 1 50S ribosomal protein L35 3 50S ribosomal protein L35 4 50S ribosomal
44954	Open rectifier potassium channel protein 1 AltName: Pseudouridine synthase EC=5.4.99 - Flags: Frag
47187	Uridylate kinase

	Short=UK EC=2.7.4 22 AltName:
49094	Cell division protein ftsZ Flags: Precursor Cell division protein ftsZ homolog 1 chloroplastic
71892	28S ribosomal protein S2 mitochondrial Short=MRP-S2 30S ribosomal protein S2 1 30S ribosomal pro
72595	30S ribosomal protein S5 1 30S ribosomal protein S5 4 30S ribosomal protein S5 5 30S ribosomal pr
73580	Diaminopimelateepimerase Short=DAP epimerase EC=5.1.1 7 Pseudouridine synthase EC=5.4.99 - Uvr
74815	eptide chain release factor 1 1 Short=RF-1 1 Peptide chain release factor 1 4 Short=RF-1 4 Pept
100368	50S ribosomal protein L3-1 chloroplastic Flags: Precursor 50S ribosomal protein L3-2 chloroplasti
109568	Probable tRNAthreonylcarbamoyladenosine biosynthesis
111560	30S ribosomal protein S19 2 30S ribosomal protein S19 4 30S ribosomal protein S19
111597	30S ribosomal protein S11 2 30S ribosomal protein S11 AltName: Full=RRP-S11

	30S ribosomal protein
111610	39S ribosomal protein L27 mitochondrial Short=L27mt 50S ribosomal protein L27 2 50S ribosomal pr
154887	FK506-binding protein 1A
164573	50S ribosomal protein L16 2 50S ribosomal protein L16 AltName: Full=RRP-L16 50S ribosomal protein
173604	Dihydrodipicolinate synthase 1
183702	50S ribosomal protein L14 2 50S ribosomal protein L14 AltName: Full=RRP-L14 50S ribosomal protein
204095	30S ribosomal protein S8 2 30S ribosomal protein S8 4 30S ribosomal protein S8
205046	50S ribosomal protein L22 1 50S ribosomal protein L22 2 50S ribosomal protein L22 4 50S ribosomal
207015	50S ribosomal protein L1 4 50S ribosomal protein L1 5 50S ribosomal protein L1
218168	Protein translocase subunit SecA 1 Short=tbSecA Protein translocase subunit SecA 2 Protein transl
218325	Defective chorion-1 protein FC125 isoform Flags: Probable ribosome-binding factor A chloroplastic
210610	30S ribosomal protein S3 2 30S ribosomal protein S3 5 30S ribosomal protein S3
218466	28S ribosomal protein S18c mitochondrial Short=MRP-S18-c

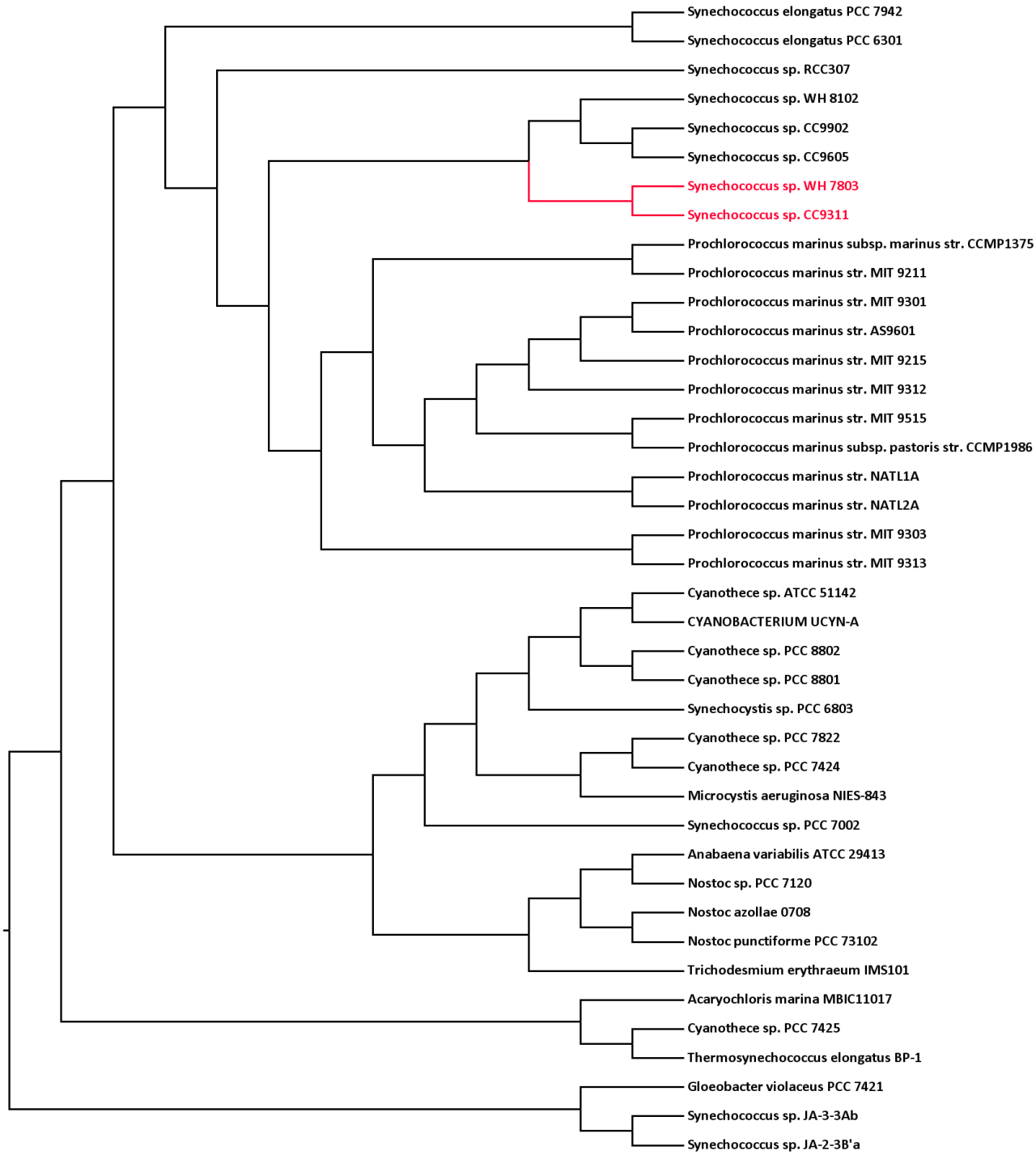
	30S ribosomal protein S18 1 30S ribosom
218798	23S rRNA Putative TrmH family tRNA/rRNAmethyltransferase
223473	4-hydroxy-3-methylbut-2-enyl diphosphate reductase Probable queuinetRNA-ribosyltransferase
224767	CDK5RAP1-like protein CDK5 regulatory subunit-associated protein 1 AltName: Putative methylthiotra
224767	CDK5RAP1-like protein CDK5 regulatory subunit-associated protein 1 AltName: Putative methylthiotra
225286	50S ribosomal protein L13 1 50S ribosomal protein L13 2 50S ribosomal protein L13
226717	Chaperone protein DnaJ 1 Chaperone protein DnaJ 2 Chaperone protein DnaJ 3 Chaperone protein DnaJ
227962	50S ribosomal protein L27 5-methyltetrahydropteroyltriglutamate--homocysteine Probable ribosomal R
229290	DNA polymerase EC=2.7.7 7 Flags: Fragment Elongation factor Tu 1 Short=EF-Tu 1 AltName: Full=P-43
231262	50S ribosomal protein L15 2 50S ribosomal protein L15 4 50S ribosomal protein L15
231311	50S ribosomal protein L5 2 50S ribosomal protein L5 4 50S ribosomal protein L5
232982	30S ribosomal protein S19 chloroplastic 50S ribosomal protein L2 3

	50S ribosomal protein L2 5 50
237875	2-aminoethylphosphonate--pyruvate transaminase EC=2.6.1 Bifunctional 3-dehydroquinate dehydratase/
239392	Glycine cleavage system H protein 1 AltName: Glycine cleavage system H protein 1 Flags: Precursor
242360	6-phosphofructokinase 1 Short=Phosphofructokinase 1 Ribosome maturation factor rimP 1 Ribosome ma
242675	3-isopropylmalate dehydratase large subunit EC=4.2.1 Phenylalanyl-tRNAsynthetase alpha chain 1 E
248742	50S ribosomal protein L18 2 50S ribosomal protein L18 4 50S ribosomal protein L18 AltName: Full=RR
268118	Protein late bloomer UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--L-lysine ligase UDP-N-acetylmuramy
276708	Dihydrolipoamide dehydrogenase EC=1.8.1 4 AltName: Full=E3 Dihydrolipoyl dehydrogenase 1 mitochon
277829	Uncharacterized protein ycsD Uncharacterized thioesterdehydrase BH1850 EC=4.2.1 -
282351	GTP-binding protein TypA/BipA AltName: Full=Tyrosine GTP-binding protein TypA/BipA homolog GTP-bin

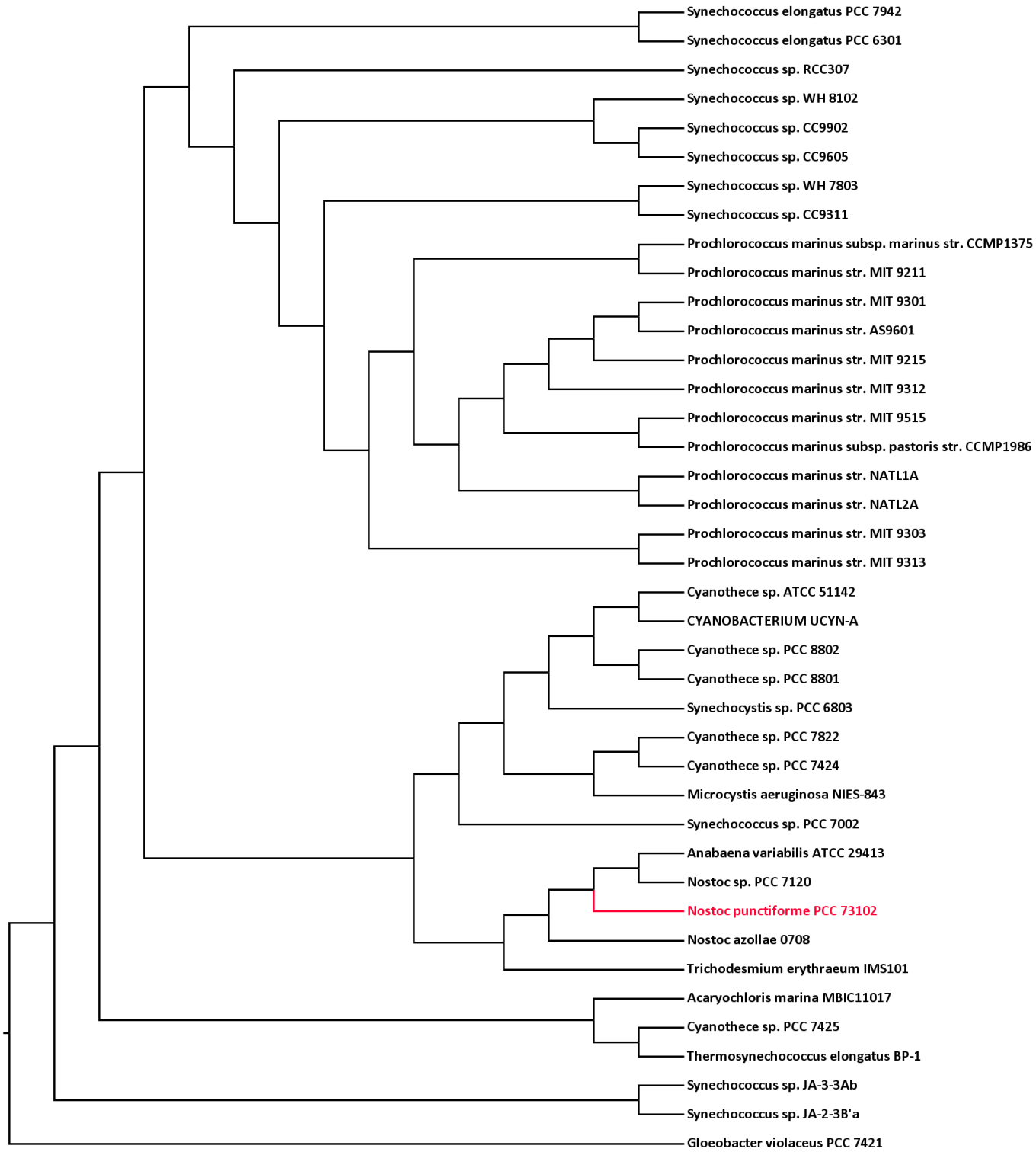
List of Trial Trees

The red branched are showing the differences between the trees and signature genes tree.

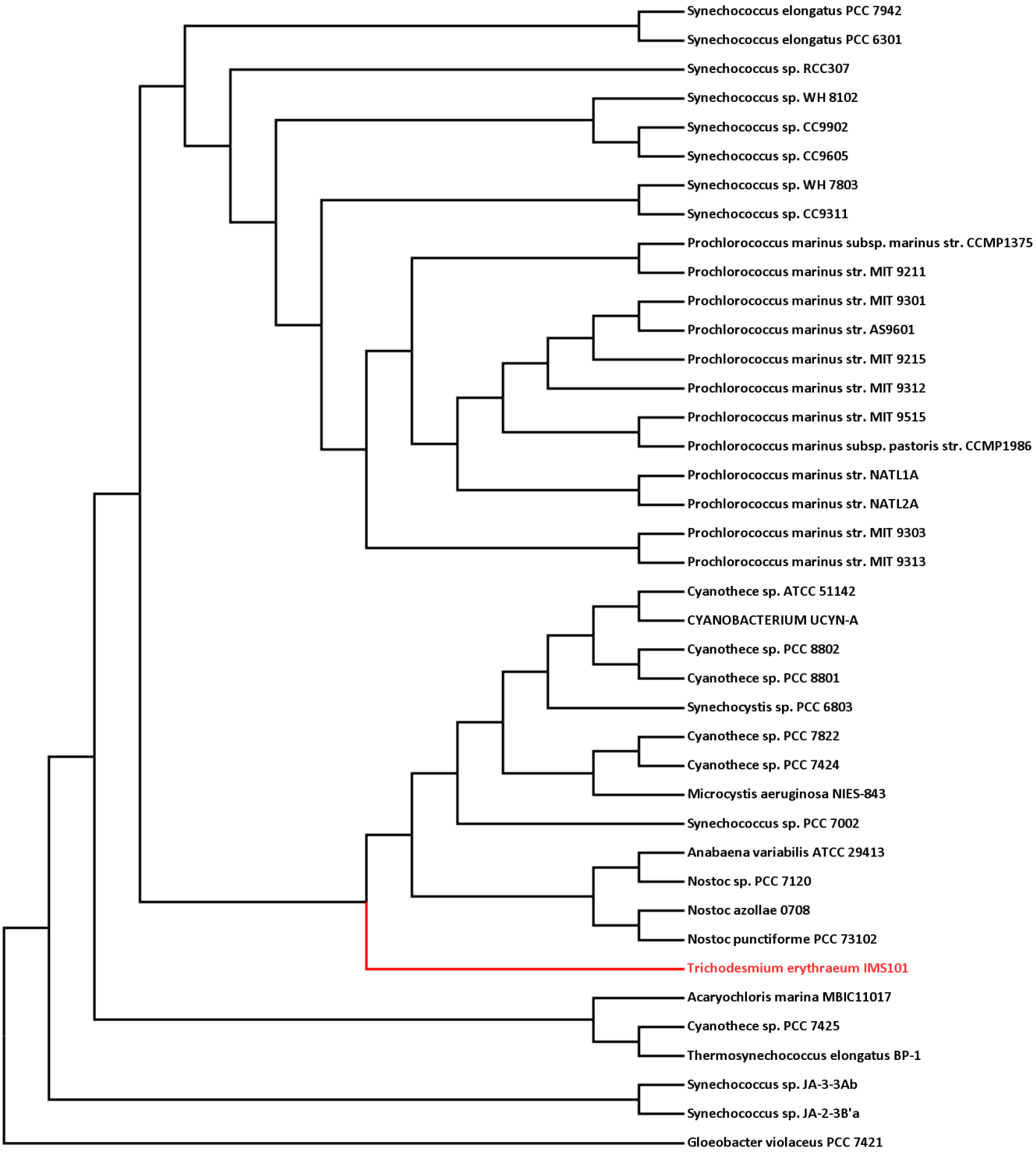
Tree Number 2



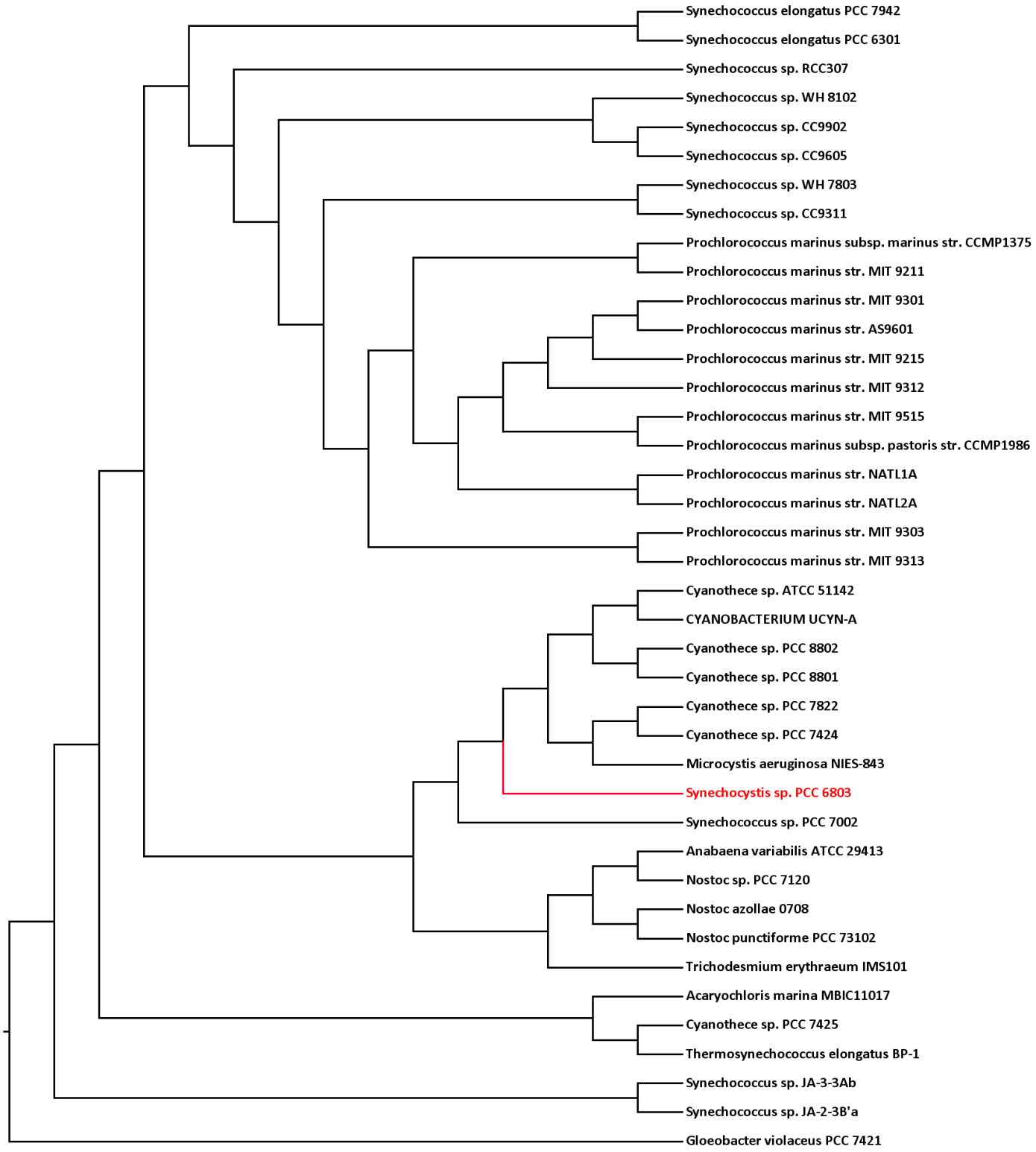
Tree number 3



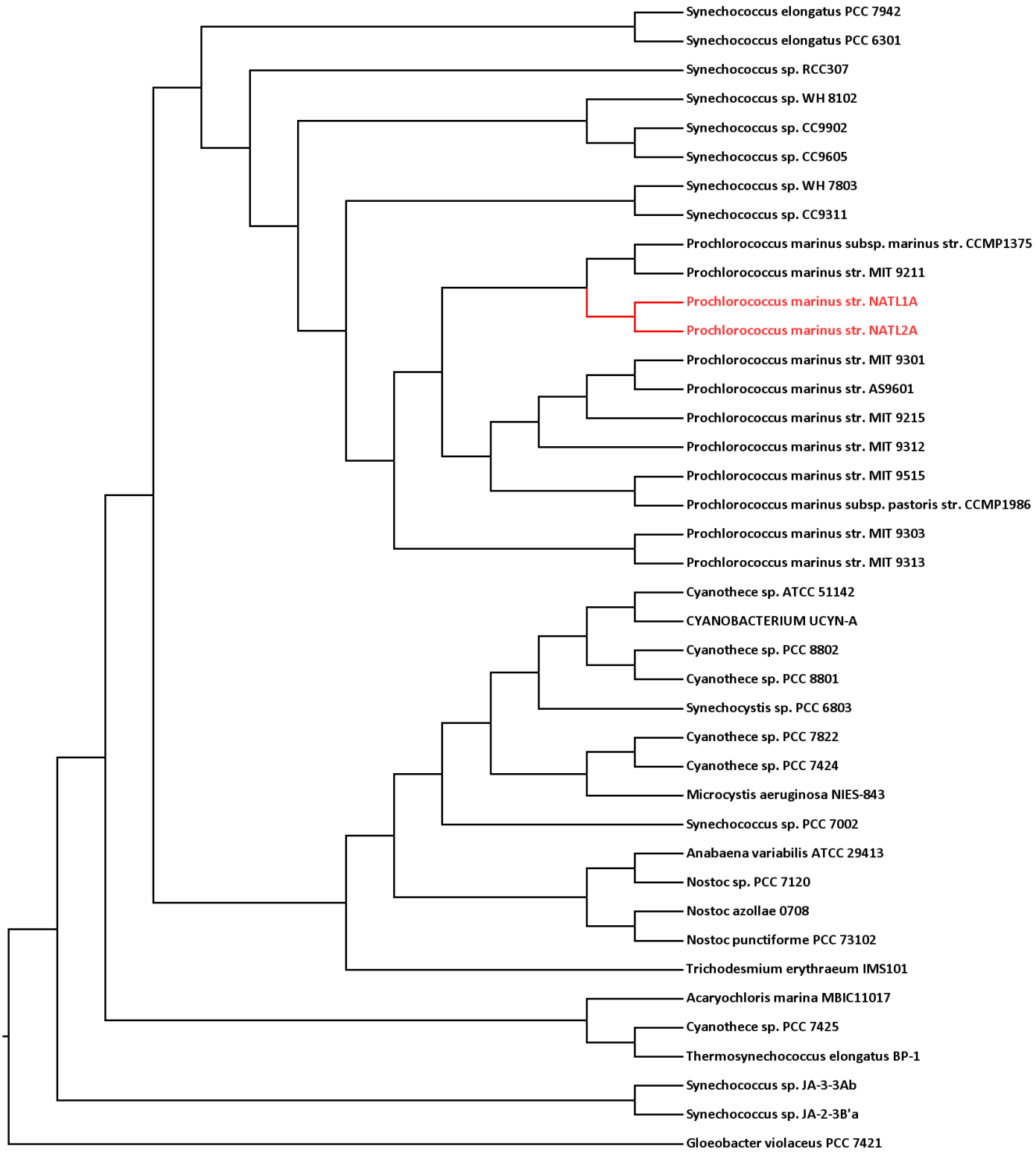
Tree number 5



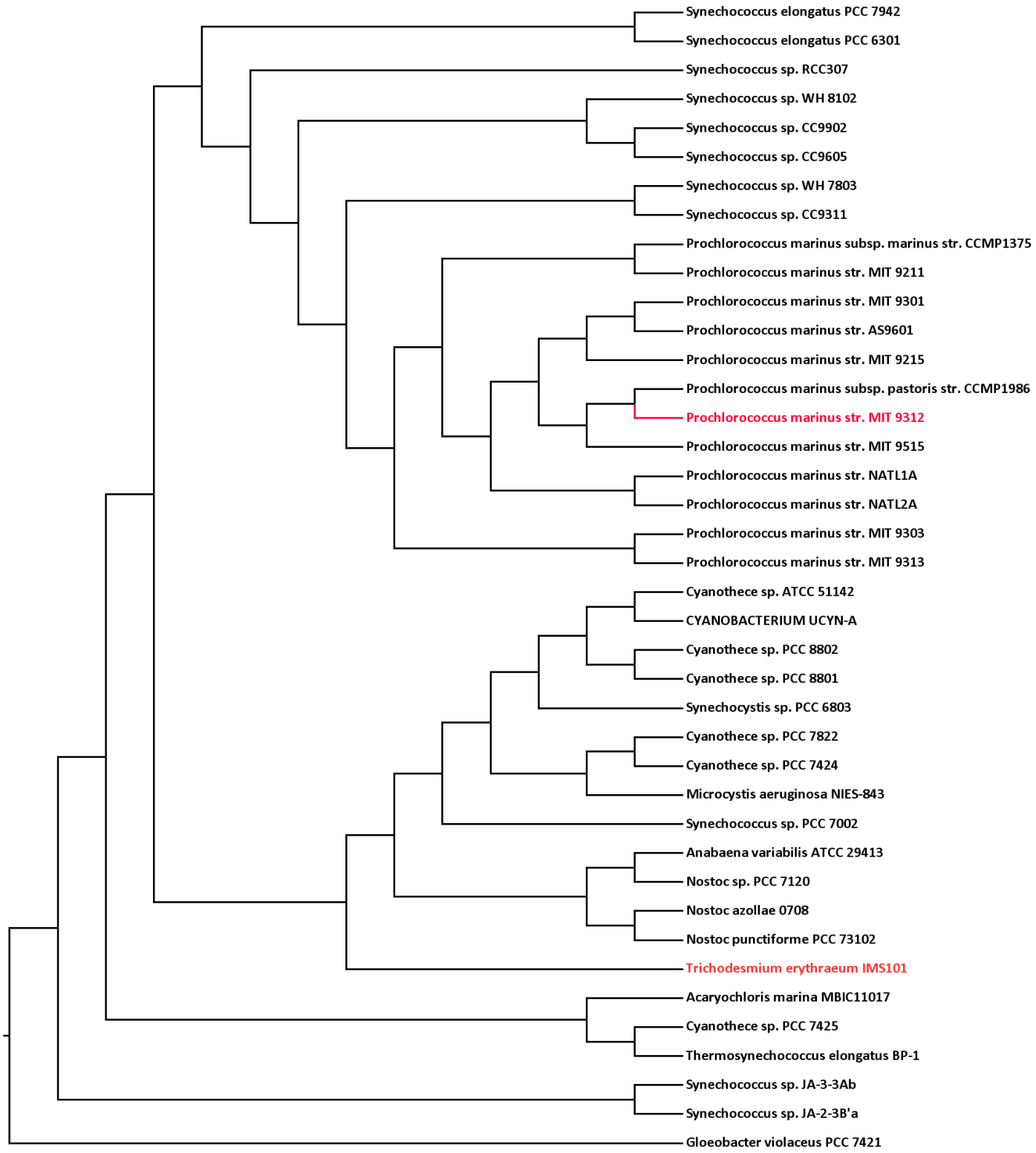
Tree number 6



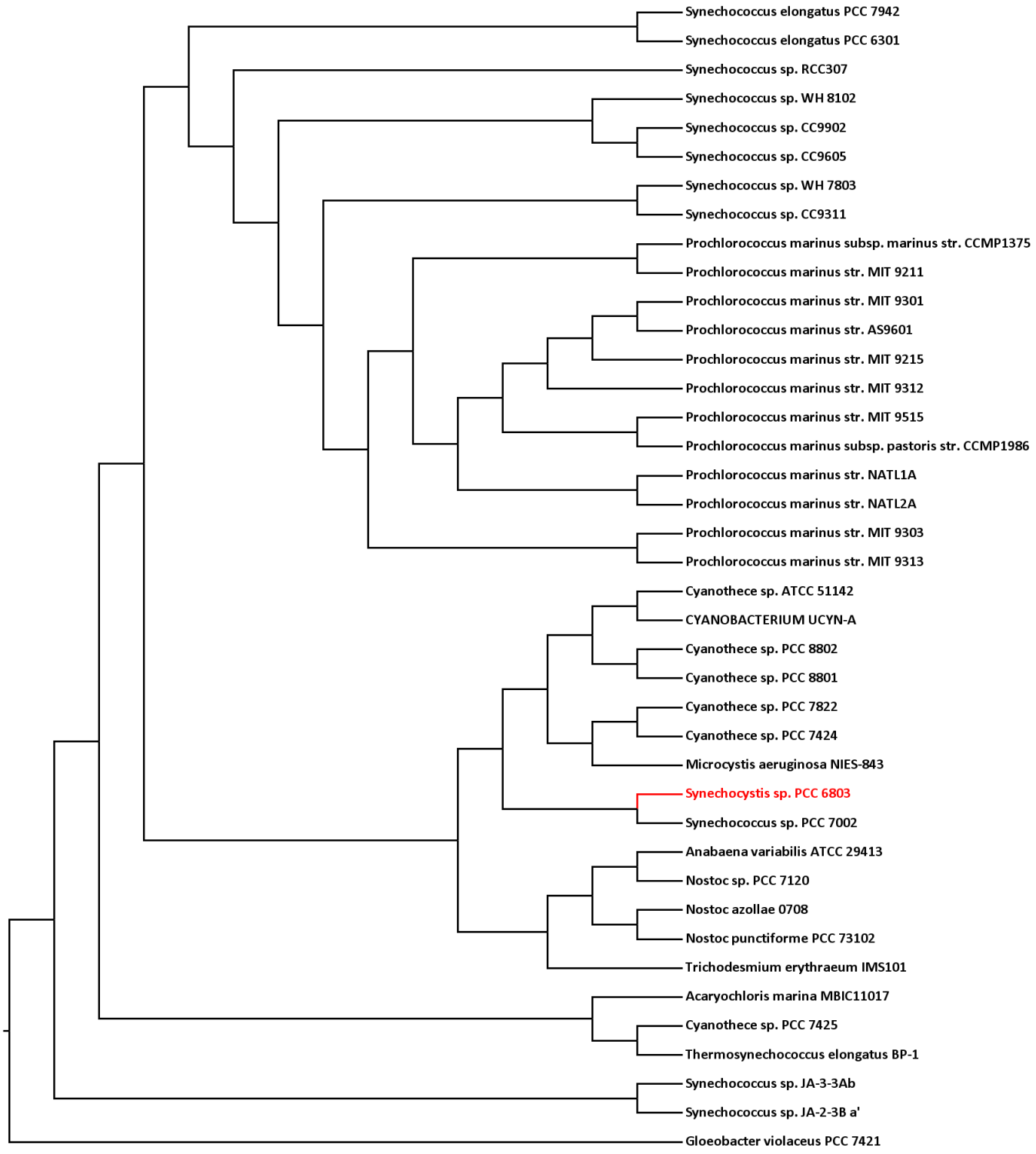
Tree number 7



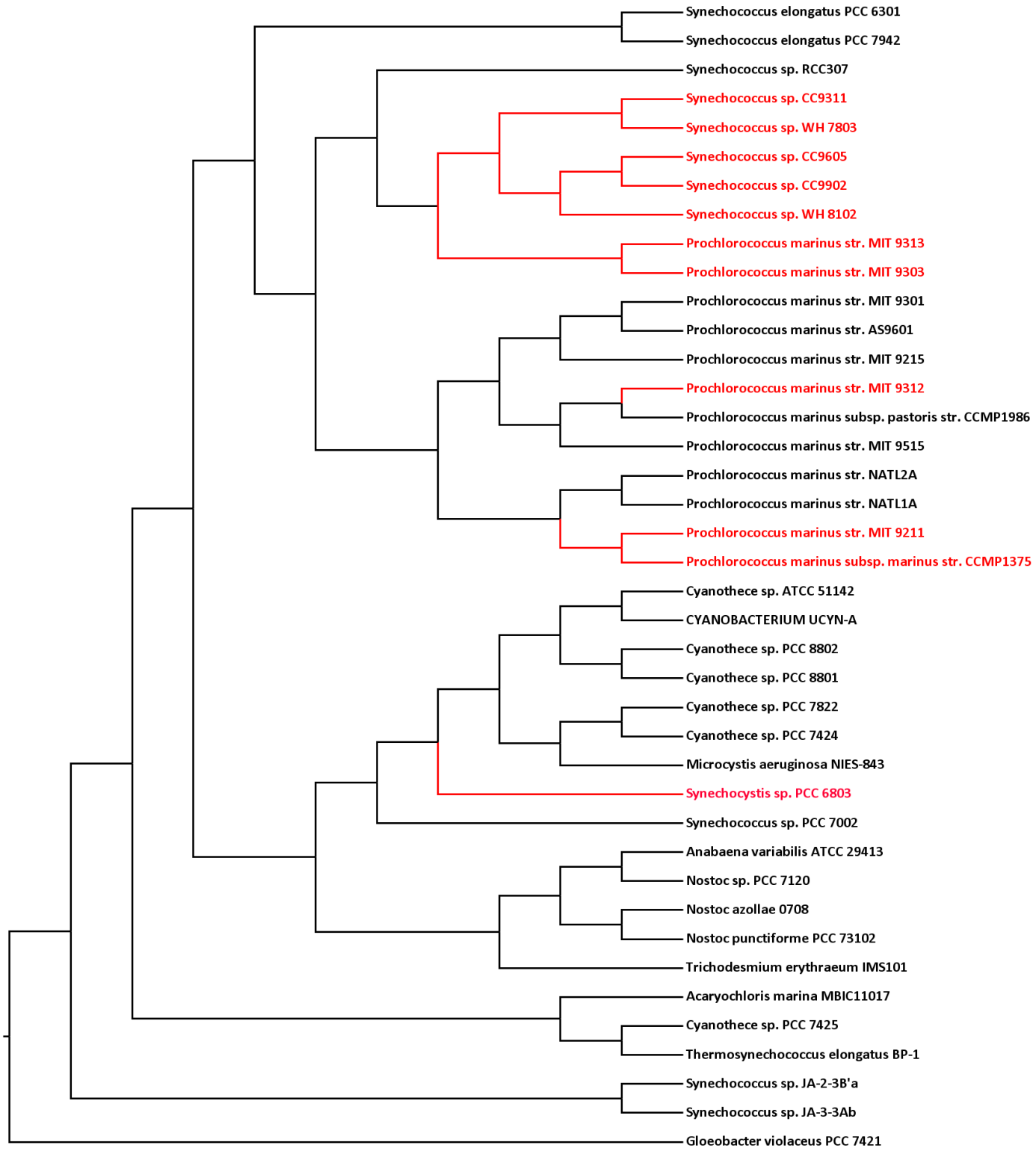
Tree number 8



Tree number 9



Tree number 10



Bibliography

(n.d.).

Akaike H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*, 716-723.

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics, 20*, 407-415.

Avery OT, Colin MM, McCarty M. (1944). Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med, 79*, 137-158.

Azad RK & Lawrence JG. (2007). Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res, 35*, 4629–4639.

Bansal MS, Banay G, Gogarten JP, Shamir R. (2011). Detecting highways of horizontal gene transfer. *J Comput Biol, 18*, 1087-1114.

Bapteste E, O'Malley MA, et al (11 co-authors). (2009). Prokaryotic evolution and the tree of life are two different things. *Biol Direct, 4*, 34.

Baumdicker F, Hess WR, Pfaffelhuber P. (2010). The diversity of a distributed genome in bacterial populations. *Ann Appl Prob, 20*, 1567-1606.

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *The ISME J, 8*, 2530-2542.

Campbell DT, Fiske DW. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*, 81.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Pääbo S, Hasegawa M. (1998). Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J Mol Evol, 47*, 307-322.

Cavalli-Sforza LL, Edwards AW. (1967). Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet, 19*, 233.

Chen I, Dubnau D. (2004). DNA uptake during bacterial transformation. *Nat Rev Microbiol, 2*, 241-249.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science, 311*, 1283-1287.

Cohen O & Pupko T. (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Mol Biol Evol, 27*, 703-713.

Cohen O, Rubinstein ND, Stern A, Gophna U & Pupko T. (2008). A likelihood framework to analyse phyletic patterns. *Philos Trans R Soc Lond B Biol Sci, 363*, 3903-3911.

Collins RE, Higgs PG. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol, 29*, 3413-3425.

Dagan T, Artzy-Randrup Y, Martin W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A, 105*, 10039-10044.

Dagan T, Martin W. (2006). The tree of one percent. *Genome Biol, 7*, 118.

Dagan T, Martin W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A, 104*, 870-875.

Dagan T, Roettger M, et al (18 co-authors). (2013). Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evo, 5*, 31-44.

David LA, Alm EJ. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature, 469*, 93-96.

Dayhoff MO, Schwartz RM. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (Vol. 5, pp. 345–352). Washington, D.C: National.

de Lamarck, JBDM. (1839). *Histoire naturelle des animaux sans vertebres* (Vol. Vol. 3). Meline: Cans et Compagnie.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol, 55*, 65-73.

Felsenstein J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol, 17*, 368-376.

Felsenstein J. (2004). *Inferring phylogenies.* Sunderland: Sinauer Associates.

Gascuel O. (2005). *mathematics of evolution and phylogeny.* New York: Oxford University Press.

Ge F, Wang LS, Kim J. (2005). The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol, 3*, e316.

Gupta RS, Mathews DW. (2010). Signature proteins for the major clades of Cyanobacteria. *BMC Evol Biol, 10*, 24.

Gyles C, Boerlin P. (2013). Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. *Vet Pathol, 51*, 328–340.

Hao W, Golding GB. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res, 16*, 636-643.

Hao W, Golding GB. (2008). Uncovering rate variation of lateral gene transfer during bacterial. *BMC Genomics, 9*, 235.

Henikoff S, Henikoff JG. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A, 89*, 10915-10919.

Higgs PG, Attwood TK. (2005). *Bioinformatics and molecular evolution.* Blackwell Science Ltd.

Hilario E, Gogarten JP. (1993). Horizontal transfer of ATPase genes—the tree of life becomes a net of life. *Biosystems, 31*, 111-119.

Howard-Azzeh M, Shamseer L, Schellhorn HE & Gupta RS. (2014). Phylogenetic analysis and molecular signatures defininga monophyletic clade of heterocystous cyanobacteriaand identifying its closest relatives. *Photosynth Res, 122*, 171-185.

Huelsenbeck JP, Ronquist F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*, 754-755.

Jukes TH, Cantor CR. (1969). *Evolution of Protein Molecules.* New York: Academic Press.

Kannan L, Li H, Rubinstein B & Mushegian A. (2013). Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biol Direct, 8*, 1-12.

Karberg KA, Gary JO, James JD. (2011). Similarity of genes horizontally acquired by Escherichia coli and Salmonella enterica is evidence of a supraspecies pangenome. *Proc Natl Acad Sci U S A, 108*, 20154-20159.

Kass RE & Raftery AE. (1995). Bayes factors. *J Am Stat Assoc, 90*, 773–795.

Kiel C, Yus E, Serrano L. (2010). Engineering signal transduction pathways. *Cell, 140*, 33-47.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Res, 15*, 954-959.

Kurland CG, Canback B, Berg OG. (2003). Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A, 100*, 9658-9662.

Larget B, Simon DL. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol, 16*, 750-759.

Larsson J, Nylander JA, Bergman B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evol Biol, 11*, 187.

Lawrence JG & Ochman H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol, 44*, 383-397.

Lobkovsky AE, Wolf YI, Koonin EV. (2013). Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol Evo, 5*, 233-242.

Lyubetsky VA & V'yugin VV. (2003). Methods of horizontal gene transfer determination using phylogenetic data. *In Silico Biol, 3*, 17–31.

Mirkin BG, Fenner TI, Galperin MY & Koonin E V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in. *BMC Evol Biol, 3*, 2.

Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS, Carlson CA, Montoya JP & Zehr JP. (2010). Unicellular Cyanobacterial Distributions Broaden the Oceanic N2 Fixation Domain. *Science, 327*, 1512-1514.

Nakamura Y, Itoh T, Matsuda H & Gojobori T. (2004). Biased. *Nat Genet, 36*, 760–766.

Olsen GJ, Woese CR. (1993). Ribosomal RNA: a key to phylogeny. *The FASEB J, 7*, 113-123.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perrière G. (2009). Databases of homologous gene families for comparative genomics. *BMC bioinformatics, 10*, S3.

Puigbò P, Wolf YI, Koonin EV. (2009). Search for a 'Tree of Life'in the thicket of the phylogenetic forest. *J Biol, 8*, 59.

Rasko DA, Rosovitz MJ, Myers GS, et al (13 co-authors). (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol, 190*, 6881-6893.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol, 61*, :539–542.

Shi T, Falkowski PG. (2008). Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci U S A, 105*, 2510-2515.

Shih PM, Wu D, et al (23 co-authors). (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A, 110*, 1053-1058.

Sjöstrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B & Lagergren J. (2014). A Bayesian method for analyzing lateral gene transfer. *Syst Biol*, syu007.

Ross SM. (1996). *Stochastic processes.* New York, NY: Wiley.

Syvanen M. (1985). Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol, 112*, 333-343.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A, 109*, 17513-17518.

Tatum EL, Lederberg, J. (1947). Gene recombination in the bacterium Escherichia coli. *J Bacteriol, 53*, 673.

Wang, B. (2001). Limitations of compositional approach to identifying. *J Mol Evol, 53*, 244–250.

Welch RA, Burland V, et al (19 co-authors). (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A, 99*, 17020-17024.

Whelan S, Goldman N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol, 18*, 691-699.

Woese CR, Fox GE. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A, 74*, 5088-5090.

Yang Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol, 10*, 1396–1401.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res, 16*, 1099-1108.