INFERENCE OF GENE COEVOLUTION BASED ON PHYLOGENETIC PROFILES

by

Chaoyue Liu

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
September 2022

To my parents, for their endless love, support, and encouragement.

To my beloved, for having you on this journey.

Table of Contents

List of Tables

List of Figures

Abstract

Phylogenetic profiles, which summarize the presence and absence patterns of genes in a set of genomes, can be used to identify genes that have correlated evolutionary histories. However, comparative analysis of phylogenetic profiles should take into account the phylogenetic effect under consideration. In this study, we developed phylogenetic comparative methods to infer the gene coevolution.

We first proposed an approach that uses Pagel's correlation model to infer the evolutionary similarities between genes and a hierarchical-clustering approach to define sets of genes with correlated distributions across the organisms. The results support the assumption of our work that the genes with correlated evolutionary histories tend to be functionally linked.

However, Pagel's method is computationally expensive and tends to overestimate the signal of coevolution. We developed a new coevolutionary model - the Community Coevolution Model (CCM), which has the additional advantage of being able to examine multiple genes as a community to reveal a more complete picture of the dependency relationships. We also developed a simulation procedure to generate phylogenetic profiles of gene sets with correlated evolutionary trajectories and adjustable strength of interactions. The results show that the CCM is more accurate than Pagel's method and other heuristic tree-aware methods and provides more biological insights such as the evolutionary rates, significance levels and directions (positive/negative) of interactions.

We also developed a matrix decomposition-based method (Chisq-PLR), especially for large-scale analysis. Our method not only has computational speed that is competitive with other heuristic methods but also gives support to better biological explanations. This fast method can be used to pre-process large data sets to reduce the number of computations that need to be carried out by CCM or other mechanistic model based methods.

Acknowledgements

The completion of this work could not have been possible without the guidance of my supervisors, help from friends, and support from my family.

First, I would like to express my sincerest gratitude to my supervisors, Dr. Hong Gu and Dr. Robert G. Beiko for their patience, encouragement, and knowledge throughout this research. I would also like to extend my gratitude to my committee members, Dr. Toby Kenney and Dr. Lam Ho for their valuable insights and time.

Last but not least, I would like to thank my family and all the friends I have met here for their support and encouragement.

Chapter 1

Introduction

Organismal traits can evolve in a coordinated way, with correlated patterns of gains and losses reflecting important evolutionary associations. Discovering these associations can reveal important information about the functional and ecological linkages among traits. Comparative studies can provide useful insights into selection and adaptation of organismal traits in concert with their evolutionary history [82]. The types of traits that can be assessed in this framework are broad and can include morphology, behaviour, physiology, and ecology [70].

In this study, we treat an individual gene's presence and absence as a trait distributed across a set of genomes. Genes can exhibit similar patterns of presence and absence [72] due to correlated evolutionary processes of gains (resulting from lateral gene transfer) and losses, for reasons such as participation in a common biochemical pathway, physical linkage, or co-localization on a mobile genetic element such as a plasmid [27, 12, 17]. Lateral gene transfer (LGT) refers to the movement of genetic information between organisms other than by standard vertical transmission from parent to offspring [65] and is an important force in microbial evolution enabling processes such as adaptation to extreme environments [65, 32], and acquisition of new metabolic functions [68]. Examination of these patterns can reveal important information about related functions and common pathways of LGT. A well-established approach to represent presence and absence patterns among genes is the construction of phylogenetic profiles, binary vectors that summarize the presence and absence of genes across a set of genomes, effectively treating each gene as a separate trait [71, 64, 60].

The success of phylogenetic profiling depends on the use of appropriate measures to express the distance and similarity between profiles. However, because of phylogenetic similarity, closely related species are likely to share many traits as a result of the process of descent with modification, which means that the gene distributions,

which are part of a hierarchically structured phylogeny, cannot be regarded as independent [26, 16, 54]. Figure 1.1 illustrates how the phylogenetic effect could impact the comparative results. In Figure 1.1.a, two flagellar genes show substantial dissimilarities in the clade which consists of closely related *Clostridioides difficile*, but have strongly correlated patterns across the remainder of the tree. However, from the evolutionary perspective, the drastic dissimilarity in the clade of *C.difficile* could be potentially attributed to a single LGT event. The phylogenetic tree in this example is highly skewed due to sampling bias and is obviously problematic for phylogeny-naïve methods.

The phylogenetic effect could impact ordinary trees as well. Figure 1.1.b gives another example with a more balanced phylogenetic tree compared to the tree in Figure 1.1.a. Both pairs of genes in Figure 1.1.b have the exact same number of co-occurrences, but pair 1 has the co-occurrences concentrated in one clade of the tree, while the co-occurrences of pair 2 are spread across the tree. Both pairs will show the same level of similarities for phylogenetic-naïve methods, but as other comparative studies suggested [16, 91], pair 1 indicates a "within-clade pseudoreplication" scenario which occurs when two traits have a single origin on the same lineage, and are then inherited by nearly all species in the descendant clade, resulting in (almost) perfectly co-distribution. On the other hand, pair 2 shows replicated co-occurrences across the tree and suggests that the two genes share multiple gain and loss events and therefore are more likely to have a genuine evolutionary association. Both examples suggest that the phylogenetic effect should be taken into account while comparing the phylogenetic profiles and one solution is to compare the phylogenetic profiles from an evolutionary perspective.

In the remainder of this chapter, I will introduce in detail the three components of this comparative study: phylogenetic tree, phylogenetic profiles, and comparative methods for binary traits.

1.1 Phylogenetic Trees

A phylogenetic tree is a branching diagram depicting the evolutionary relationships among different entities (such as genes, genomes, or species) constructed on the basis of sequenced genomics data [25]. For the phylogenetic tree of five taxa shown in

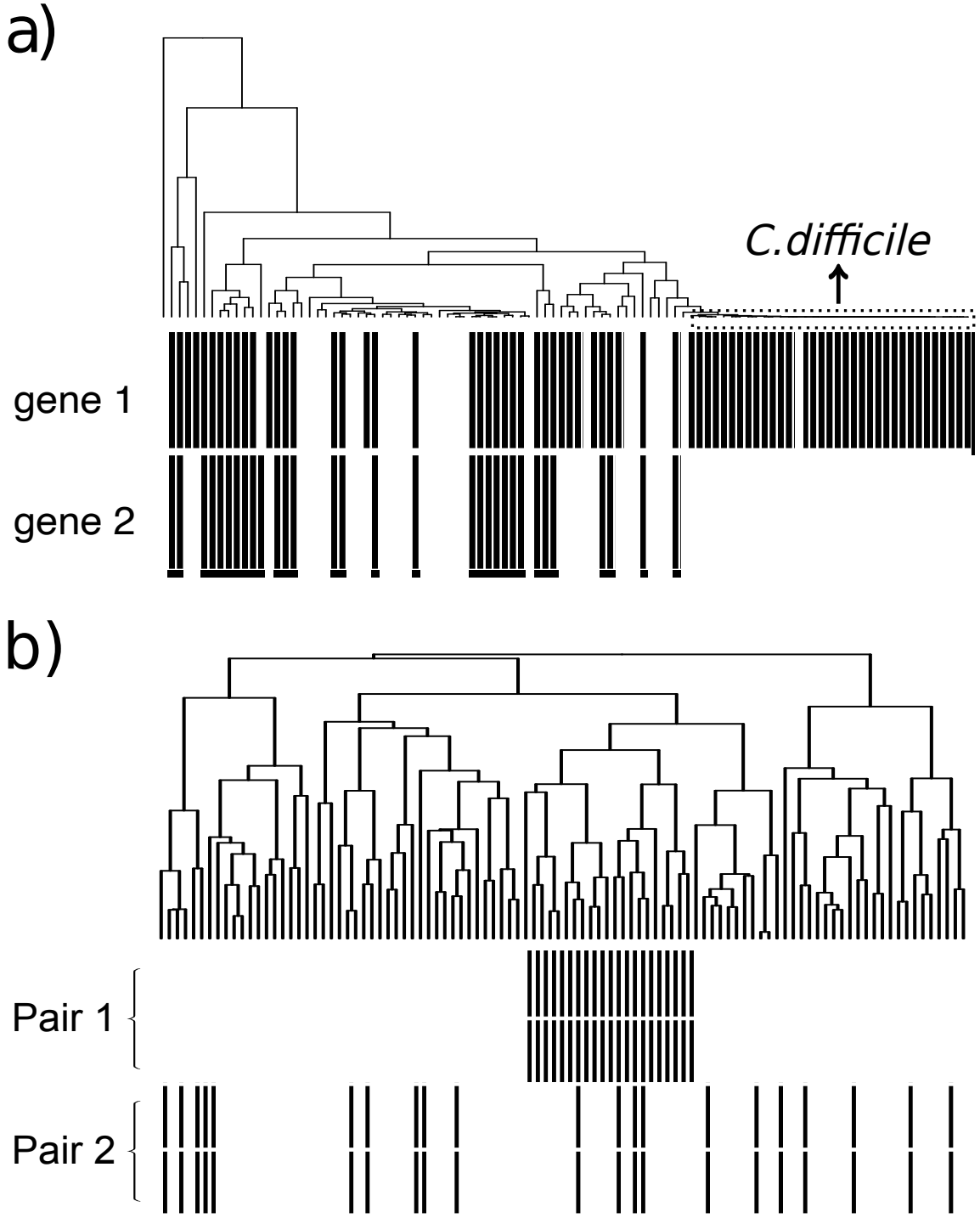


Figure 1.1: Illustration of phylogenetic effects in the comparison of phylogenetic profiles. a) Two correlated genes with a highly skewed phylogenetic tree. b) A pair of genes (Pair 1) that have co-occurrences concentrated in one clade, and another pair of genes (Pair 2) that have co-occurrences across the tree. Each row indicates a phylogenetic profile where black bars represent presences.

Figure 1.2, three basic components of the tree are the root (s_0), nodes ($s_1, s_2, \ldots, s_6$), and branch lengths ($t_1, t_2, \ldots, t_6$).

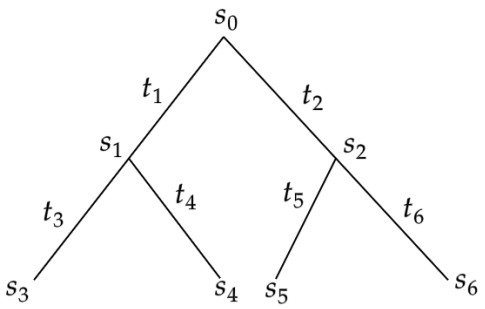


Figure 1.2: A phylogenetic tree with 4 tips: s_0 represents the root of the tree, s_1, s_2 are two internal nodes and s_3, s_4, s_5, s_6 are 4 existing species. $t_1, t_2, ..., t_6$ on each branch indicate the branch lengths.

A phylogenetic tree can be rooted or unrooted. The root of the tree represents the common ancestor of all entities in the tree. A rooted tree is often required by model-based comparative methods since the rooted tree indicates the direction of evolution. Unrooted trees only describe the relatedness between species but do not make assumptions about the ancestral root. The most common way for rooting the tree is to include distantly related organisms as outgroups in the data for constructing the tree so that the root can be determined between the outgroups and the other taxa [33].

The terminal nodes, also called the tips of the tree represent the observed entities (genes, genomes, species, etc) that are used to construct the phylogeny. In a rooted tree, each internal node is the branch point that indicates a divergence event and represents the common ancestor of all taxa descended from the branch point.

Branch lengths represent the amount of evolutionary change over time, which are usually expressed as the number of nucleotide or amino-acid substitutions per site. Thus, longer branches indicate that more changes have occurred or more time has elapsed.

1.2 Phylogenetic Profiles

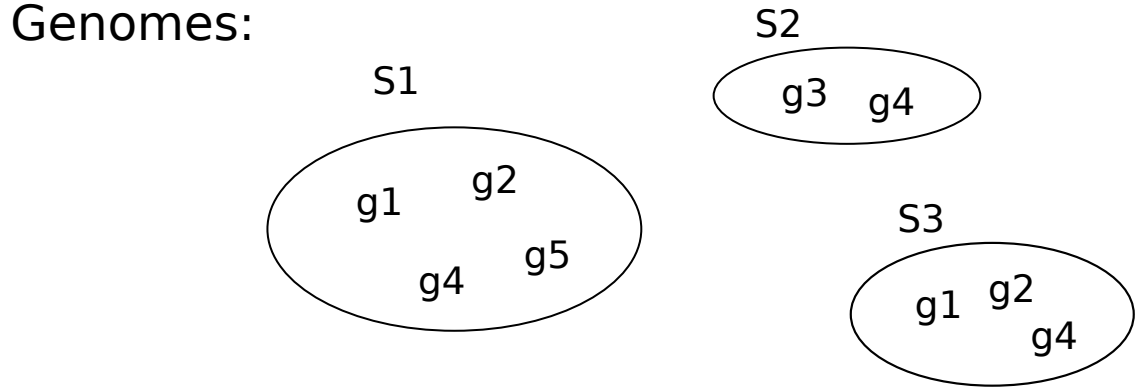
As the number of fully sequenced genomes increases rapidly, the phylogenetic profiling approach based on patterns of gene presence and absence across genomes has become a promising strategy for the computational annotation of gene functions and for better understanding their ecological roles [72, 66, 14, 54]. The underlying hypothesis of the phylogenetic profiling approach is that functionally linked genes are gained and lost together from genomes during evolution, and therefore, they could have homologs in the same set of genomes which results in a correlation of their phylogenetic distributions [72, 44].

The protein encoded by each gene in a fully sequenced genome can be assigned to a specific set of proteins based on homology relationships. Homologs can be classified into paralogs and orthologs, depending on whether they arose by duplication or speciation [47]. Orthologous genes that originated by vertical descent from a single ancestral gene are considered to be more likely to share similar functions [47, 29], and are the type of homologs we focused on in this study. The presence or absence of this protein across genomes can then be represented by its phylogenetic profile, which can be considered as a long binary sequence encoding the presence or absence of the gene across a given set of genomes. As the illustration of constructing phylogenetic profiles shown in Figure 1.3, homologous genes are first searched within each genome and then phylogenetic profiles of genes are constructed on the basis of homology information across genomes.

1.3 Comparative Methods for binary traits

1.3.1 Phylogeny-naïve Methods

In the application of using phylogenetic profiles to predict the functions of proteins, Pellegrini et al.(1999) applied Hamming distance to measure the similarities between two phylogenetic profiles [72, 44]. Hamming distance, also equivalent to Manhattan distance for binary traits, is simply the number of bits that are different between two binary vectors, which in this case, is the number of species that do not have the same absence/presence patterns. Given the profiles of two genes across p genomes $A_{p\times 1}$



Phylogenetic Profiles:

	g1	g2	g3	g4	g5
S1	1	1	0	1	1
S2	0	0	1	1	0
S3	1	1	0	1	0

Figure 1.3: An illustration of phylogenetic profiles of five genes (g_1, g_2, g_3, g_4, g_5) across three genomes (S_1, S_2, S_3).

and $B_{p\times 1}$, the Hamming distance can be calculated as

$$d(A,B) = \sum_{i=1}^{p} |A_i - B_i|.$$

Besides being phylogeny naïve, the other obvious drawback of Hamming distance is the lack of scaling, which does not take into account the total number of occurrences present in the species.

The Jaccard index is a measure of similarity for two vectors with a range from 0 and 1. In comparing phylogenetic profiles, the Jaccard Index is the ratio of the number of genomes that have both genes divided by the number of genomes that have either of the genes. Given the profiles of two genes, the 2×2 contingency table can be constructed as below

		B	
		1	0
A	1	a	b
	0	c	d

and the Jaccard Index is defined as

$$J(A,B) = \frac{a}{a+b+c}.$$

The Jaccard Index only focuses on the similarity in the co-occurrences of two genes and ignores genomes that do not contain any genes.

Mutual Information is a measure of the mutual dependence between two variables in information theory and can be used to quantify the similarities between profiles based on how much information can be gained from the knowledge that one gene is present about the presence of another gene [97]. The mutual information for a pair of genes A and B is defined as

$$I(A,B) = \sum_{i,j\in\{0,1\}} p_{ij}(A,B) \log_2 \frac{p_{ij}(A,B)}{p_i(A)p_j(B)},$$

where p_0 and p_1 denote the fraction of presences and absences respectively in profiles.

In order to calculate the statistical significance of the similarities observed between two phylogenetic profiles, the Hypergeometric test can be applied based on the combinations of presence and absence as shown in the contingency table above

[16, 80]. The hypergeometric distribution is a discrete probability distribution that is often used to describe the probability of number of successes within a fixed number of independent draws without replacement [61]. Based on the contingency table above, the probability of observed patterns can be calculated as

$$Pr(\text{observed}) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{p}{a+c}},\ p = a+b+c+d.$$

The Hypergeometric test assesses the extremeness of observed similarities between profiles, so the P-value is calculated as the sum of the probabilities of all possible realizations of more extreme patterns (more co-occurrences), which is the same as a one-sided Fisher's exact test. There are other metrics that can be used to score similarity between two binary traits. The above four common ones are reviewed here and are also compared in our study.

The standard metrics do not take into account phylogenetic effects and weight each genome in the phylogenetic profile equally. Since closely related genomes are more likely to share similar gene content, they are likely to have an outsized influence on profile comparisons relative to their phylogenetic diversity as shown in Figure 1.1.a. The independence assumptions on genomes that these standard metrics are based on are violated by the fact that genomes are connected by a phylogenetic tree, thus application of these methods could lead to the biased inference results.

1.3.2 Heuristic Methods

Several heuristic approaches have been developed to account for phylogenetic effects in profiles based upon the phylogeny-naïve metrics. Random sampling and shuffling techniques can be used to form a null distribution of standard metrics to estimate the statistical significance of the observed similarities as the example shown in Figure 1.4.a [42, 81]. A restrictive shuffling strategy that takes into consideration the lineages, which shuffles only the taxa specific to a lineage, can further improve the accuracy of the estimated null distribution for lineage-specific proteins [42]

von Mering et al. (2003) attempt to account for the phylogenetic effect by collapsing the tree into a subtree based on a tree-guided selection of species. As illustrated in Figure 1.4.b, if the species within the same clade have the same presence/absence pattern for a pair of genes, this clade is collapsed into one single node and substituted

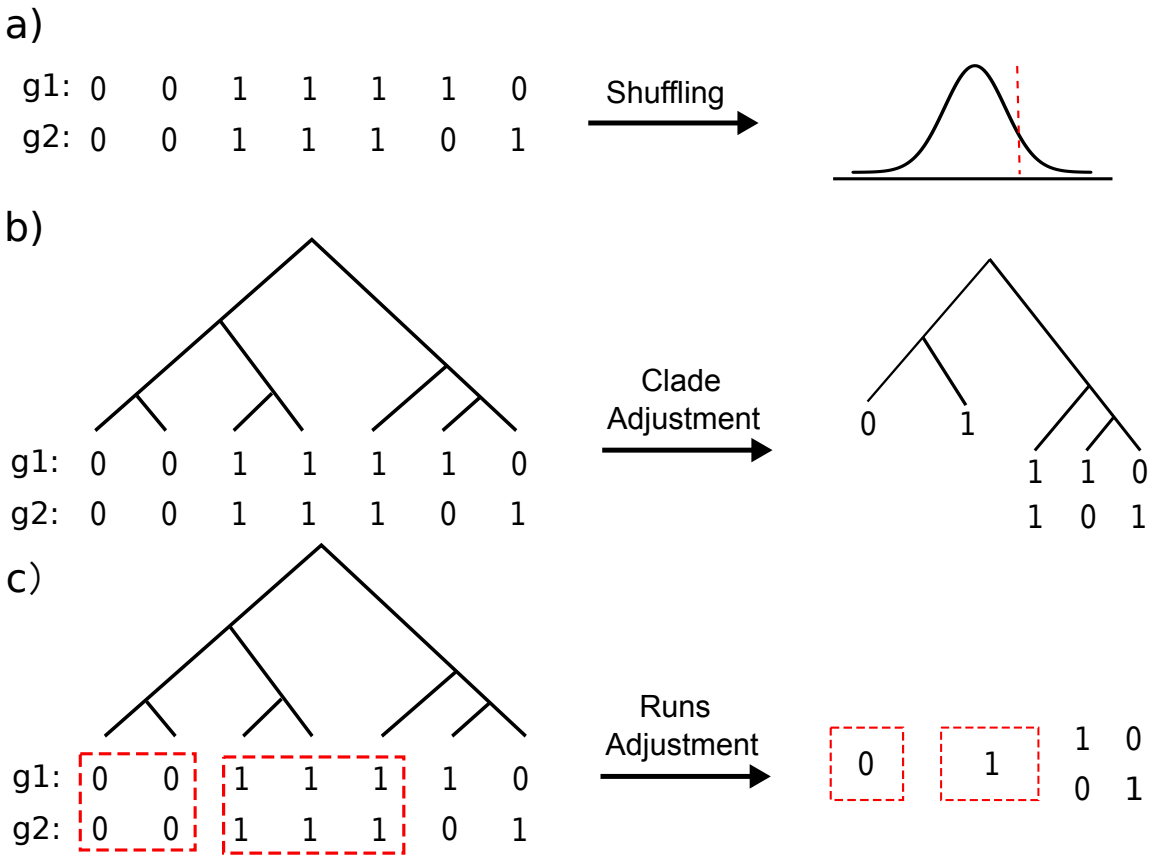


Figure 1.4: An illustration of three heuristic methods. a) Shuffling strategy to create null distribution. b) Clade-adjustments method c) Runs-adjustment method.

by their ancestral state.

Cokus et al. (2007) account for the underlying phylogeny by treating consecutive matches between profiles as a run, which are collapsed to one single state, as shown in Figure 1.4.c, then computing the similarity between phylogenetic profiles based on the enumerated runs of consecutive matches [16]. The hypothesis is that the profiles with many runs are more likely to evolve in a correlated fashion (e.g. Pair 1 in Figure 1.1.b) than the profiles in which all co-occurrences are concentrated in one lineage of the tree (e.g. Pair 2 in Figure 1.1.b).

The shared underlying idea of these methods is the application of a weighting scheme to the genomes in order to counteract phylogenetic effect. However, they are *ad hoc* approaches that do not properly model the underlying evolutionary processes. For example, counting the three consecutive co-occurrences as one run in Figure 1.4.c might be lacking in biological explanation, as these three genomes are relatively distant with their last common ancestor being the root. The choice of the standard metrics could also affect these heuristic methods and produce different results.

1.3.3 Probabilistic Methods based on Evolutionary Models

Evolutionary models aim to explain the distribution of genes by modeling the correlation patterns of gain and loss on a phylogenetic tree. For discrete traits, a continuous-time Markov chain with a finite state space is commonly used for modeling the evolutionary history along the phylogenetic tree. A Markov chain is a stochastic process in which the transition to the next state only depends on the current state.

Under the assumption of a Markov process, the transition probability of traits on one branch only depends on the starting state of the branch and the evolutionary time which is given by the branch length. In addition, the transition on each branch is also assumed to be independent from other branches. To construct the likelihood function of the tree for one realization of the evolutionary process, the straightforward way is to take the product of the transition probabilities of all branches from the root to the tips. However, the computation cost could be expensive as we need to sum over all the possible combinations of the states at each internal node. The equation

below shows the likelihood function for a simple tree in Figure 1.2:

$$L = \sum_{s_0}\sum_{s_1}\sum_{s_2} P_{s_0,s_1}(t_1)P_{s_1,s_3}(t_3)P_{s_1,s_4}(t_4) \times P_{s_0,s_2}(t_2)P_{s_2,s_5}(t_5)P_{s_2,s_6}(t_6), \tag{1.1}$$

where $P_{s_i,s_j}(t_j) = e^{Qt}$ denotes the transition probability from state i to state j via a branch of length t and Q is known as the instantaneous transition rate matrix.

The amount of computation using this straightforward way will increase exponentially as the tree increases in size, but it can be reduced by applying Felsenstein's pruning algorithm [24]. The pruning algorithm is a dynamic programming approach that takes advantage of the nested structure of the tree and computes the likelihood of the tree recursively. By applying the pruning algorithm, the likelihood function L can be reformatted as

$$L = \sum_{s_0}[(\sum_{s_1} P_{s_0,s_1}(t_1)P_{s_1,s_3}(t_3)P_{s_1,s_4}(t_4)) \times (\sum_{s_2} P_{s_0,s_2}(t_2)P_{s_2,s_5}(t_5)P_{s_2,s_6}(t_6))]. \tag{1.2}$$

In this way, the likelihoods for subtrees can be reused and the computation complexity is reduced to linear.

The CoPAP (Coevolution of Presence-Absence Patterns) method models the evolutionary process for each gene independently with a 2 by 2 transition rate matrix Q (two states: 0 and 1) and then uses a stochastic mapping procedure to infer the expected number of gain and loss events for each branch. Coevolutionary correlation between genes is then calculated by computing Pearson's correlation between the inferred evolutionary histories based on simulations [14, 15]. Both CoPAP and Pagel's correlation test methods are based on this maximum likelihood framework with an underlying continuous-time Markov process.

Pagel's correlation approach specifically tests the evolutionary correlations between pairs of binary traits [66]. To characterize the discrete-trait evolution in Pagel's method, two continuous-time Markov models are contrasted: One model where the two characters are assumed to evolve independently, and a second model where two characters are assumed to evolve in a correlated way, possibly due to interactions.

The hypothesis of correlated evolution is tested by comparing the fit of the two different models to the observed data set. Under the assumption that the two traits evolve independently, the null model (independent evolution model) is a special case of the alternative model (dependent evolution model), and the two models can be assessed using a likelihood-ratio test. The dependent evolution model (8 parameters) which has more degrees of freedom in the Q matrix will almost certainly have a higher likelihood than the independent evolution model (4 parameters). Thus, the likelihood ratios, which follow a χ^2 distribution with four degrees of freedom (difference in parameters), will express the relative strength of the evolutionary dependencies between genes.

CLIME (clustering by inferred models of evolution), another probabilistic method, uses a different way to model the evolutionary process [51]. CLIME is a clustering algorithm based on a hidden Markov model (HMM), to group genes into evolutionarily conserved modules (ECMs) with the assumption that each gene has a single gain event and zero or more loss events. The evolutionary model of each ECM is represented by a single gain branch and a vector of loss probabilities for each branch. CLIME first infers initial ECMs from the input gene set and then expands to other genes in the data set by calculating the likelihood of each gene against inferred ECMs and assigning the gene to the best-fitting ECM.

CLIME is specifically designed for eukaryotic data as it allows only one single gain event, so it is likely less suitable for prokaryotic data that have high rates of gene transfers. CoPAP assumes that the gain rate and loss rate independently vary among genes rather than explicitly modeling the interactions during evolution. Although Pagel's correlation model outperformed CLIME and CoPAP in detecting functionally linked genes in our study (Chapter 2), it still suffers from high computational cost and has been criticized for its tendency to overestimate the signal of coevolution in other studies [56, 95]. The features and computation information of the methods reviewed together with the CCM developed in this thesis are summarized in Table 1.1.

Table 1.1: A summary table of comparative methods based on phylogenetic profiles.

	Methods	Description	Tree information	Computation
Phylogeny-naïve	Hamming Distance	The number of differences in two genes' profiles.	No	Low
	Jaccard Index	A measure of similarities between two genes' profiles based on the percentage of co-occurrences among the total number of presences.	No	Low
	Mutual Information	A measure of mutual dependency between two variables.	No	Low
	Hypergeometric	Use the Hypergeometric distribution to calculate the chances of observed co-occurrence patterns between a pair of genes.	No	Low
Heuristic methods	Sampling methods	Based on the simple metrics, create a null distribution of the metric by randomly sampling a subset of genomes and calculate the significance level.	random subsets of species	Low
	Run-adjusted methods	Simulate runs along two genes' profiles, count the consecutive co-occurrences as one run and calculate the simple metrics (mutual information, Hypergeometric) based on the simulated runs.	tree tips (species) order	Low
	Clade-adjusted methods	Collapse the clades in which the two genes have exactly the same patterns, and calculate the standard metrics (mutual information) using the collapsed profiles.	Tree clade groupings	Low
Probabilistic models	CoPAP	Estimate the 2 by 2 transition rate matrix for each gene, and use a simulation procedure with inferred rates to account for the correlation of expected gain and loss events between a pair of genes.	Tree topology and branch lengths	Medium
	CLIME	Infer the ECMs which consist of only one single gain event on the tree followed by none or more loss events, and cluster genes into ECMs.	Tree topology	Medium
	Pagel's correlation test	Model the evolution process of a pair of genes in independent and correlated fashions and implement a likelihood ratio test to detect the significant correlations.	Tree topology and branch lengths	High
	CCM	Model the transition rates of a community of genes using the assumed intrinsic rate and pairwise interaction coefficients.	Tree topology and branch lengths	High

1.4 The Structure of the Thesis

In Chapter 2, we firstly apply Pagel's coevolutionary model and hierarchical clustering to analyze the gene set of the bacterium "*Lachnospiraceae* bacterium 3-1-57FAA-CT1" (abbreviated as LZ), which was isolated from a biopsy retrieved from the transverse colon of a female Crohn's Disease patient at the time of colonoscopy. LZ is of interest because its genome size is very large compared to most of its immediate neighbours (6505 protein-coding genes as compared with a median of 3124 in our complete data set of Clostridia). The results of this study further support the assumption of our work that the genes with correlated phylogenetic profiles also tend to be functionally linked. This work has been published on *Genome Biology and Evolution* (10, no. 9 (2018): 2255-2265).

In Chapter 3, we develop a new model-based comparative method - the Community Coevolution Model (CCM) to analyze the evolutionary associations among genes based on phylogenetic profiles. In the CCM, genes are considered to evolve as a community with interactions, and the transition rate for each trait depends on the current states of other traits. Surpassing other comparative methods for pairwise trait analysis, CCM has the additional advantage of being able to examine multiple traits as a community to detect the conditionally independent links and reveal more dependency relationships. For pair-wise comparisons, our method is more efficient and approximately 5 times faster than Pagel's method. We also develop a simulation procedure to generate phylogenetic profiles with correlated evolutionary patterns that can be used as benchmark data for evaluation purposes. This work has been published online by *Systematic Biology* (10.1093/sysbio/syac052).

In Chapter 4, we propose another novel method based on matrix decomposition to test the dependency between binary profiles conditioning on the tree topology. Although our CCM model is computationally more efficient than Pagel's method, it cannot scale to larger datasets containing thousands of gene profiles due to the quadratic scaling of pairwise comparisons. Our motivation for this study is to develop a fast and accurate method that can pre-process a large data set to reduce the number of computations that need to be carried out by CCM or another probabilistic model based method. The existing heuristic methods such as the runs-adjustment method, and the clade-adjustment method that empirically apply a weighting scheme

to the genomes in order to counteract phylogenetic effects, are weak at biological interpretations. Our approach considers the phylogenetic profile as containing two components of information: one is the underlying phyletic pattern (P) driven by the phylogeny among species as the close related genomes that inherit from the close common ancestors will tend to share similar gene content and thus the same genes are more likely to be found in closely related genomes (row-wise); the other component is unique information about individual genes (S) caused by their own gain/loss events during evolution (column-wise). Then we can test the dependency between a pair of profiles by conditioning on their predicted underlying phyletic pattern (P) such that the genes are considered related only when their individual components (S) show significant patterns. Our method not only has computational speed that is competitive with other heuristic methods but also gives support to better biological explanations to the data.

In Chapter 5, we summarize all the methods and analyses we have developed in this thesis and discuss possible future work.

Chapter 2

Phylogenetic clustering of genes reveals shared evolutionary trajectories and putative gene functions

2.1 Introduction

Lateral gene transfer (LGT) is an important force in microbial evolution, enabling processes such as adaptation to extreme environments [65, 32], acquisition of new metabolic functions [68], and defense against antimicrobial agents [7]. Coordinated transfers of genes can enable rapid ecological shifts, and identification of sets of genes implicated in these shifts can highlight the events that took place during the diversification of prokaryotic groups, and suggest functional linkages between the implicated genes. Given the large phylogenetic diversity of microorganisms that inhabit the human microbiome [93], and in many cases the uncertainty associated with their precise ecological roles [19], augmenting comparative genomic and metagenomic analysis with examination of LGT can produce more information about the capabilities of a given microorganism. Although strong evidence exists for preferential patterns or "highways" of gene sharing among specific groups of prokaryotes [8, 85], small amounts of LGT connect many different taxonomic groups and the overall pattern of sharing resembles a web rather than a clear reticulated tree [46]. These diffuse patterns, coupled with the methodological challenges associated with LGT inference [74, 46], make it difficult to identify sets of genes with similar evolutionary trajectories.

Many approaches have been used to identify sets of genes with similar evolutionary histories. Puigbò et al. (2009) performed a comparative analysis of the topological similarity among 6901 phylogenetic trees built for clusters of orthologous groups of proteins representing a set of 100 prokaryotes, showing that LGT did not obscure a significant central tendency so that a consistent phylogenetic signal still exists [75]. Kunin et al. (2005) applied tree reconstruction methods to infer vertical evolutionary inheritance and then detect LGT events by using an ancestral-state inference

algorithm and estimated the number of genes exchanged across organisms using a weighting scheme [49]. The results suggest genes might propagate across a microbiome rapidly, with certain organisms functioning as hubs in a broader LGT network. Phylogenetic profiles [72] summarize the presence and absence of homologous genes across a set of organisms, and have been used to identify laterally transferred genes [65] and to predict the functions of hypothetical genes [71]. Differences in gene content between related species result from processes such as gene loss, duplication and LGT, and proteins that are involved in similar biological processes may be gained and lost together, leading to similar phylogenetic profiles. However, taxonomic sampling can pose a serious challenge to the interpretation of phylogenetic profiles. Calculating Manhattan or bit distances between pairs of profiles, for example, can be overly simplistic because it weights each contributing genome equally. In a scenario where the phylogenetic sampling of genomes is non-uniform, these distances will be unduly influenced by closely related genomes that have similar profiles due to common descent. The most informative profiles will be those that are widely but sporadically dispersed across very distantly related genomes, as their distribution will not be readily explained solely through a hypothesis of common descent. Vert (2002) [96] and Barker and Pagel (2005) [6] applied phylogenetic reweighting schemes to the assessment and comparison of phylogenetic profiles showing that the genes evolving in a correlated fashion, also tend to be functionally linked.

CLIME [51], short for Clustering by Inferred Models of Evolution, was developed to explicitly consider phylogenetic relationships among genes by inferring evolutionarily conserved modules (ECMs) using a Bayesian mixture model. Each ECM represents a tree-structured hidden Markov model with a single gain branch and branch-specific gene loss probabilities. Then CLIME assigns genes within the genome to the most likely ECM or a new ECM by comparing with a background null model using the likelihood-ratio test. However, CLIME is based on models in which genes can emerge only once and then be lost multiple times. Since the effect of LGT is to create multiple apparent gene gains throughout a tree, the model of CLIME may not be appropriate for modeling prokaryotic genes that are subject to significant amounts of LGT.

Here we propose a phylogenetic-profile-based method that uses phylogenetic modeling to identify pairs of genes with similar historical patterns of gain and loss. Our

approach uses the method of Pagel (1994) [66] to test hypotheses of correlated evolution between pairs of genes. The statistics generated by this approach are used to generate clusters using a hierarchical approach based on average linkage. The resulting clusters can be evaluated in terms of the similarity of their phylogenetic distributions, the functional similarity of the proteins in each cluster, and the phylogenetic trees built from different sets of proteins contained within the cluster, which can be further used to detect LGT events and infer genomes evolution via tree reconstruction methods. The main work flow of this study is shown in Figure 2.1.

One taxonomic group that has shown evidence for high levels of LGT is the class Clostridia. As part of the Firmicutes phylum, the class is a significant component of the human microbiome and contains an ecologically diverse set of organisms, including the pathogens *Clostridioides difficile* (previously *Clostridium difficile*; a notorious cause of nosocomial diarrhea), commensals from genera such as *Roseburia* and *Faecalibacterium*, and organisms that are less well understood [73, 3]. Commensurate with its clinical importance, over 1000 genomes from class Clostridia have been sequenced at least in draft form, providing a rich resource for comparative genomics. Here we apply our new method to a set of 687 genomes from class Clostridia, with a particular focus on Lachnospiraceae bacterium 3-1-57FAA-CT1, a micro-organism which was isolated from a patient with Crohn's disease and has an abnormally large genome for the group. Our method successfully recovers phylogenetically and functionally cohesive clusters of genes, and highlights probable highways of gene sharing that have shaped this genome and its close neighbours.

2.2 MATERIAL AND METHODS

2.2.1 Datasets

The bacterium "Lachnospiraceae bacterium 3-1-57FAA-CT1" lacks a formal taxonomic designation, and we refer to in this paper as "LachnoZilla" or LZ. LZ was isolated from a biopsy retrieved from the transverse colon of a female Crohn's Disease patient aged 22 years at the time of colonoscopy. The patient was suffering a flare at the time of colonoscopy and biopsy material recovered was taken from an

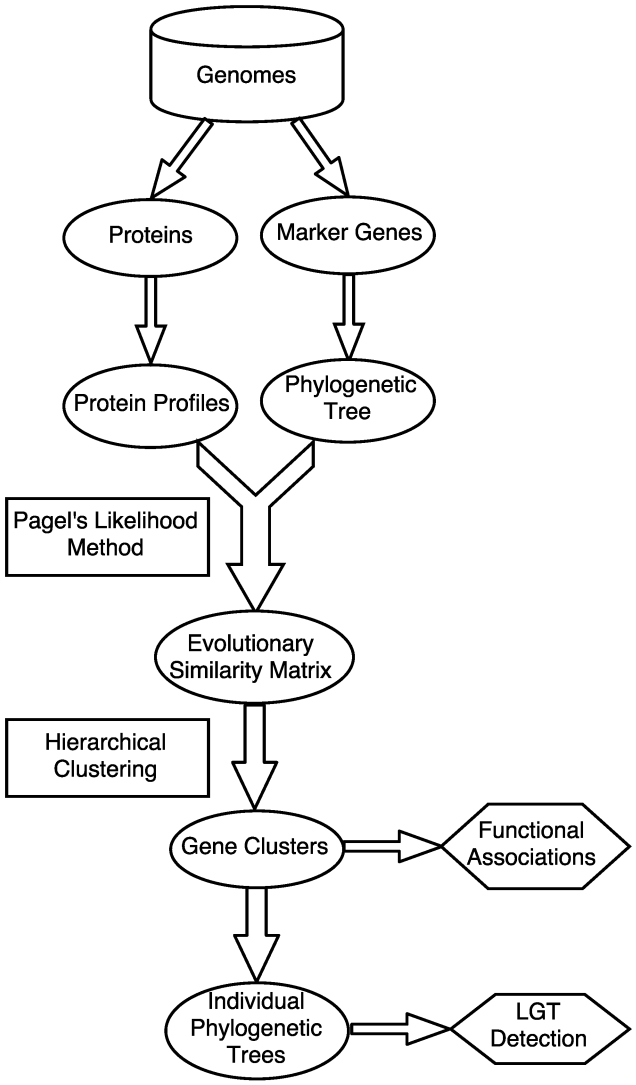


Figure 2.1: Workflow of our weighted phylogenetic profile approach. Genome-sequences are collected to construct the protein profiles and phylogenetic tree. Then, Pagel's likelihood method is implemented to calculate the evolutionary similarities among genes and the hierarchical clustering approach is used to define sets of genes with common distributions across the given genomes. An evaluation framework based on the GO terms (biological process) is also developed to study the function associations in the clustering results. In addition, the individual trees of the members within a same cluster are compared with detect potential LGT events.

inflamed site. Isolation was carried out through serial dilution and culture on fastidious anaerobe agar (Acumedia) containing 5% defibrinated sheep's blood (Hemostat Laboratories) with streak purification. gDNA was extracted using a Maxwell 16 instrument (Promega) according to manufacturer's instructions. Sequencing was performed at the Broad Institute, as part of the Human Microbiome Project Reference Genomes effort (`http://www.hmpdacc.org/reference_genomes/reference_genomes.php`), generating sequence data to 140x coverage. The protein-coding genes were predicted with Prodigal [40] and filtered to remove genes with $\geq 70\%$ overlap to tRNAs or rRNAs. The tRNAs were identified by tRNAscan-SE [55]. The rRNA genes were predicted using RNAmmer [50]. The gene-product names were assigned based on top BLAST hits against the UniProtKB/SwissProt protein database ($\geq 70\%$ identity and $\geq 70\%$ query coverage), and protein family profile search against the TIGRfam HMMer equivalogs.

We retrieved all available completed and draft genomes from class Clostridia (687 genomes including LZ; Table S1 available at `https://doi.org/10.1093/gbe/evy178#supplementary-data`), and a set of eight outgroup genomes from class Bacilli and phyla Actinobacteria and Proteobacteria which are used to root the phylogenetic tree. All genome information used in this work were retrieved from the National Center for Biotechnology Information on August 22, 2014.

2.2.2 Phylogenetic analysis and profile construction

We used a customized version of the AMPHORA2 pipeline [103] to construct a reference phylogeny based on concatenated, conserved protein sequences encoded by the set of genomes. Complete protein sequences were searched against the set of hidden Markov models (HMMs) specified by AMPHORA2, yielding a maximum of 31 protein sequences per genome. Each set of homologous proteins was aligned using the corresponding HMM, then trimmed to remove any column that had a scaled alignment confidence score less than 7. Trimmed alignment files were then concatenated into a single alignment, with any missing genes represented in the alignment using missing-data (i.e., gap) characters. Maximum-likelihood phylogenetic analysis of this supermatrix was performed using RAxML-HPC version 7.2.5, using all default

parameters and the "PROTCATLG" model of sequence substitution [87]. One hundred bootstrap replicate alignments were generated using the SEQBOOT package of PHYLIP version 3.695, and the resulting bootstrap support values mapped to the appropriate bipartitions in the tree. The tree was rooted arbitrarily among the 8 outgroup taxa, providing a defined rooting of the clostridial subtree.

Phylogenetic profiles were constructed using rapsearch version 2.14 [104]. The complete set of predicted LZ proteins was compared against all other genomes in the data set, with an expectation-value threshold of 10^{-20}. Profiles were interpreted as presence/absence matrices, with no weighting of profiles by the number of matching proteins in a given reference genome. Given the computational demands of the Pagel method, we uniformly subsampled 73 random taxa in addition to LZ from the full tree (Figure A.1), to produce a more tractable data set for cluster construction.

2.2.3 Modeling correlated patterns of evolution among sets of proteins

We used the BayesTraits software that implements the statistical approach of Pagel (1994) to correct profiles for shared evolutionary history. This method aims to identify significant evolutionary correlations between two binary characters, which in our case correspond to the presence or absence of two different homologous gene families, as represented by their phylogenetic profiles across a phylogenetic tree. To characterize the discrete-trait evolution in this method, two continuous-time Markov models are contrasted: one model where the two characters are assumed to evolve independently, and a second model where two characters are assumed to evolve in a correlated way, possibly due to interactions. The hypothesis of correlated evolution is tested by comparing the fit of the two different models to the observed data set. Under the assumption that the two characters evolve independently, the null model (independent evolution model) is a special case of the alternative model (dependent evolution model), and the two models can be assessed using a likelihood-ratio test. The dependent-evolution model with more parameters will almost certainly have a higher likelihood. Thus, the likelihood ratios, which follow a χ^2 distribution with four degrees of freedom (difference in parameters), will express the relative strength of the evolutionary dependencies between genes.

We used the resulting likelihood ratios as the basis for a hierarchical clustering

of all profiles. The likelihood ratios for all pairs of profiles were subtracted from the largest such ratio to generate a symmetrical 2697×2697 distance matrix. Clustering of this distance matrix was performed using the method of between-group average linkage (UPGMA). Specific clusters for analysis were generated by cutting the resulting dendrogram at different heights h.

2.2.4 Evaluation of profiles based on phylogenetic and functional similarity

Gene Ontology (GO) is a widely used classification scheme that was also used in the Critical Assessment of Functional Annotation (CAFA) large-scale evaluation experiment [78]. To measure the performance of the clustering methods, we developed a framework based on the biological process (BP) category from GO to evaluate the clustering results. All the available GO annotation of the proteins in this study are acquired from the Uniprot Knowledgebase (`www.uniprot.org`). To measure the biological significance, we evaluate the clusters from two directions: the quality of the clustering and the enrichment of GO terms. To evaluate the performance of the hierarchical clustering at different cutting heights, we calculate the mean of GO semantic scores weighted by the sizes of clusters according to the G-SESAME method [99] which accounts for the fraction of the aggregate contribution of all GO terms up to the closest shared ancestor term. In order to quantify the extent to which the clusters of co-evolved genes are functionally related, we performed a GO enrichment test for the distribution of each GO term across the clusters of co-evolved genes. We adopted the Pearson's Chi-square statistic as our test statistic. However, the Chi-square distribution is not appropriate for this test because there are many gene clusters relative to the number of members in each GO term, which will result in many clusters with zero count of the considered GO term. To address this limitation, we used a resampling technique to estimate the null distribution of Pearson's Chi-square statistic by randomly assigning 100,000 times all the GO terms to the clusters of genes with the sizes given by the sizes of our clusters of co-evolved genes from the hierarchical clustering methods. This test is similar to the hypergeometric tests but faster in computation.

2.3 RESULTS

2.3.1 Genome phylogeny and profiles

All profiles were constructed from a set of 687 genomes, including LZ. A total of 21 genera were represented, with 38 genomes from genus *Clostridium* including 20 genomes of *C. difficile*. Seven genomes including LZ were not taxonomically assigned at the genus level, although the SILVA taxonomy [77, 105] assigned LZ to the genus *Eisenbergiella*. A total of 6505 profiles were constructed with these genomes, including 2814 proteins unique to LZ (Figure 2.2a). A total of 2697 distinct profiles were obtained (Figure 2.2b) based on the uniformly subsampled 74 genomes. Figure 2.3 illustrates the differences between the non-phylogenetic Manhattan distance and Pagel's likelihood based co-evolutionary method in contrasting the similarity of three phylogenetic profiles to a reference profile.

2.3.2 Robustness of Pagel's statistics

Different tree rootings and the randomness in computing the maximum-likelihood estimators in Pagel's software may affect the calculation of coevolutionary similarities, which can result in inconsistent likelihood-ratio statistics for the same pair of genes. To evaluate this instability, we reran the full data set using the other 2 tree-rooting methods: MAD which is based on the minimum ancestor deviation [92], and the naïve midpoint-rooting method. The likelihood-ratio statistics computed from three different ways of tree-rooting all showed correlation scores > 0.9 (Figure 2.4. a-c). The correlations are high despite the instability introduced by the errors involved in the maximum-likelihood computation process. We can conclude from these results that Pagel's statistics are robust relative to different tree-rooting methods and the errors introduced in the computation of MLEs.

2.3.3 Properties of clusters generated by the hierarchical method and CLIME

In spite of the similar cluster-size distributions produced by our method and CLIME, there are substantial differences in the clusters produced. We first used the weighted

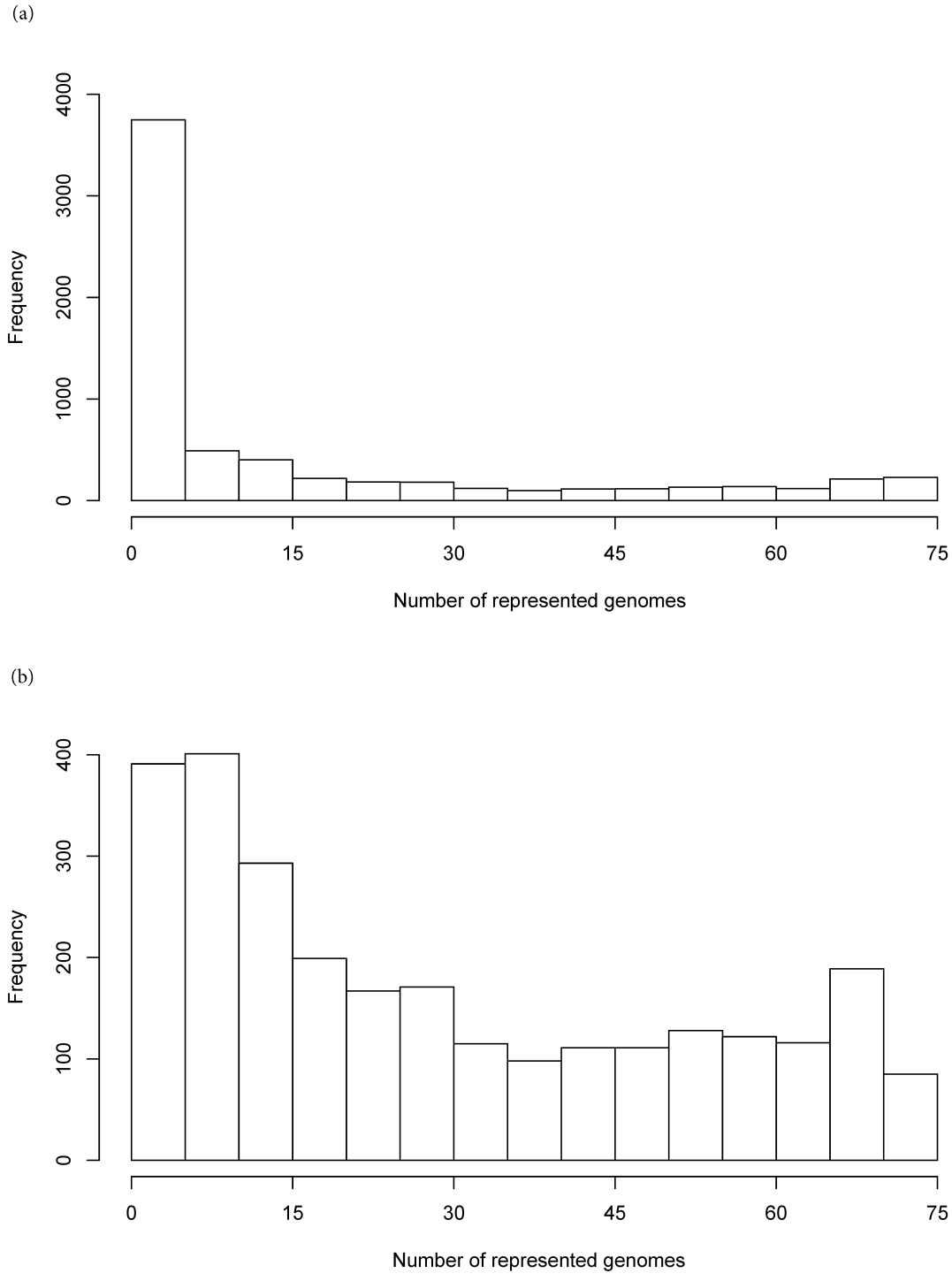


Figure 2.2: (a) Size distribution (number of presences) of 6505 profiles. (b) Size distribution (number of presences) of 2697 unique profiles.

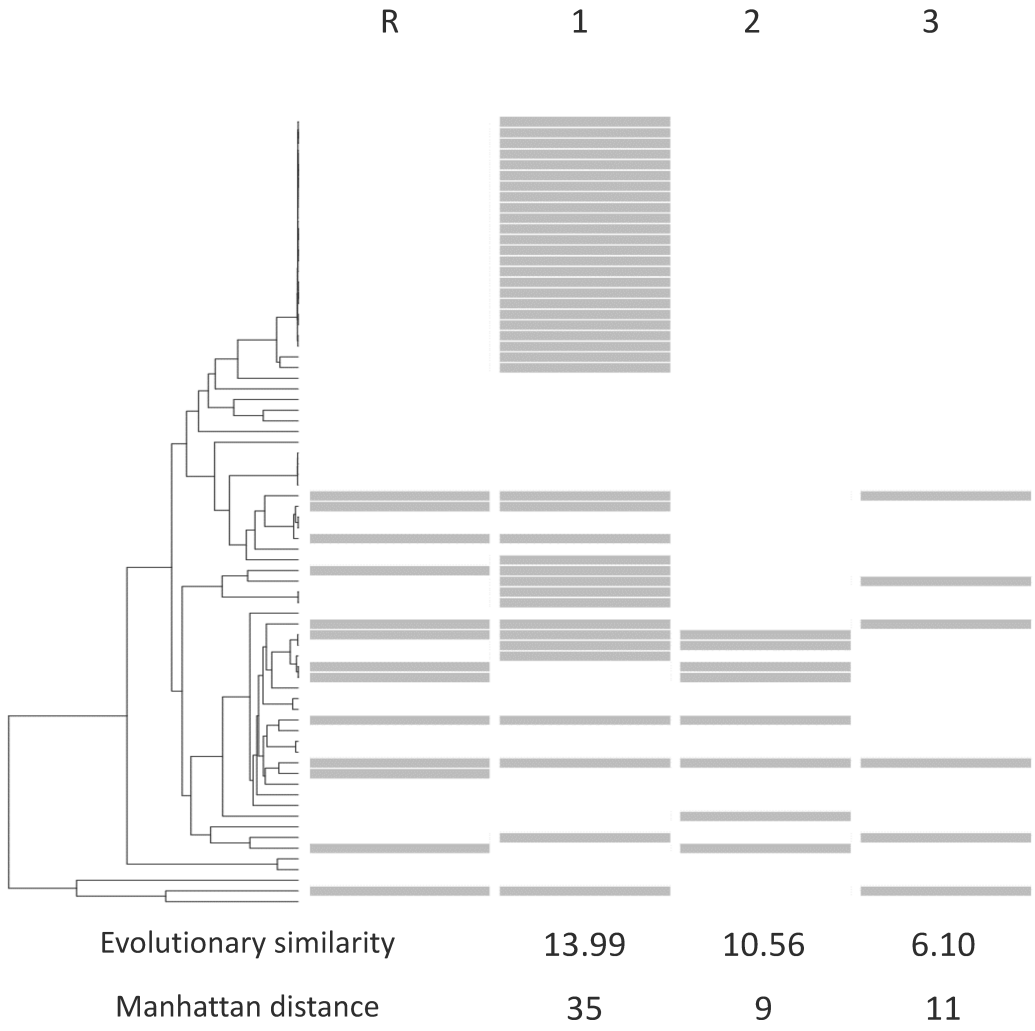


Figure 2.3: Similarity of phylogenetic profiles (1, 2, 3) to a reference profile (R) according to Pagel's likelihood-ratio statistics and Manhattan distance scores. Gray bars indicate the presence of a given gene in the genome that corresponds to the phylogenetic tree on the left. When considering the Manhattan distance, Gene 2 is the most similar to the reference profile, whereas Gene 1 has a very large Manhattan distance due to its representation in a closely related set of *Clostridioides difficile* genomes that do not contain genes in profile R. However, this drastic dissimilarity can be attributed to a single LGT event, and Gene 1 is most similar to the reference profile according to the likelihood-ratio statistic.

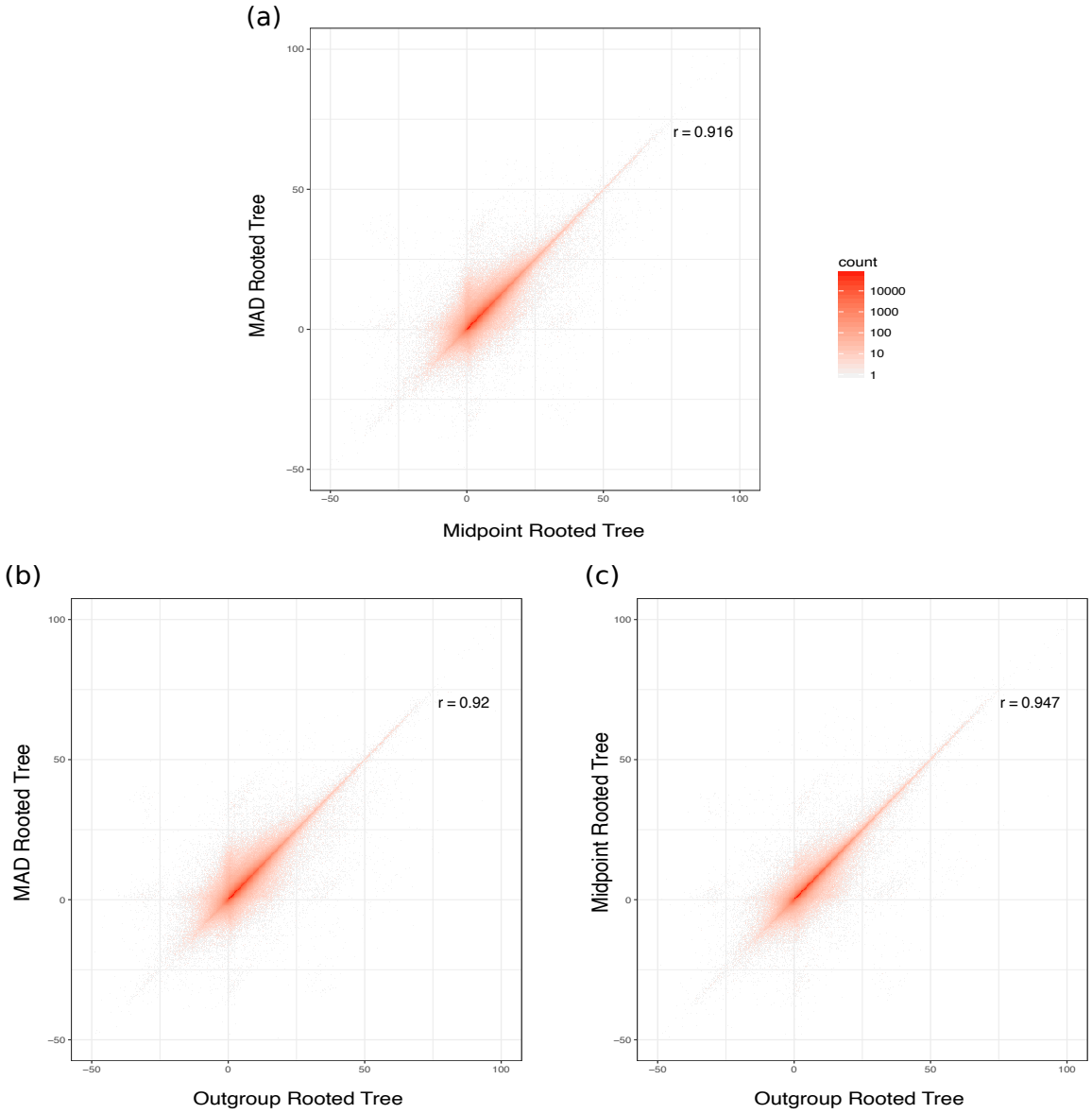


Figure 2.4: The comparisons of the Pagel's likelihood ratio statistics computed from three different methods for tree rooting: (a) Midpoint-rooted tree vs MAD-rooted tree; (b) Outgroup-rooted tree vs MAD-rooted tree; (c) Outgroup-rooted tree vs Midpoint-rooted tree.

average of the GO semantic similarity to compare the overall clustering results between two methods. The CLIME approach, which generates a single set of clusters, yielded an average GO similarity within clusters of 0.61 (Figure 2.5). The hierarchical approach generated a large range of similarity values depending on the choice of threshold, from 0.3 when the cutting height h = 100 to 0.85 when h = 60. Both approaches generated similarity scores that were greater than random. The increase in GO similarity with decreasing h is reasonable, since lower values of h produce clusters with higher overall profile similarity. However, the cost of lower h is that fewer profiles are assigned to nonsingleton clusters.

A key distinction between the hierarchical approach and CLIME is the treatment of gene gain-and-loss events. CLIME allows only a single gain of a trait on the phylogenetic tree. In cases where a gene has a sparse distribution due to LGT, the CLIME model will be a poor fit. This is illustrated in Figure A.2, where a hierarchical cluster containing four profiles (Figure A.2a), all annotated with the mannose metabolic (alpha-mannosidase activity) GO functional category, is split into four singleton clusters by CLIME. In spite of their similar phylogenetic distribution, CLIME's single-gain constraint assigns events to different parts of the tree (Figure A.25b-e), yielding different historical inferences for these four profiles. We also compared to another novel probabilistic evolutionary model CoPAP [14, 15] at different cut-offs (Figure 2.6) and showed that our method performs significantly better than the other two methods (Figure 2.7) via a permutation test.

However, the runtime of CoPAP is much faster than that of CLIME and our method. The most time-consuming step of our method is calculating the Pagel statistics. Each pairwise comparison of genes requires only a few seconds, but over 3 million comparisons are needed to do comparisons over all 2600 genes. However, this step is easy to run in parallel because the pairs of genes are independent. In our case, we spent about 7 days by running 30 jobs in parallel. For the same data set, CoPAP took around 3 hours using the web service developed by its authors and CLIME took around 1 day running on a local computer.

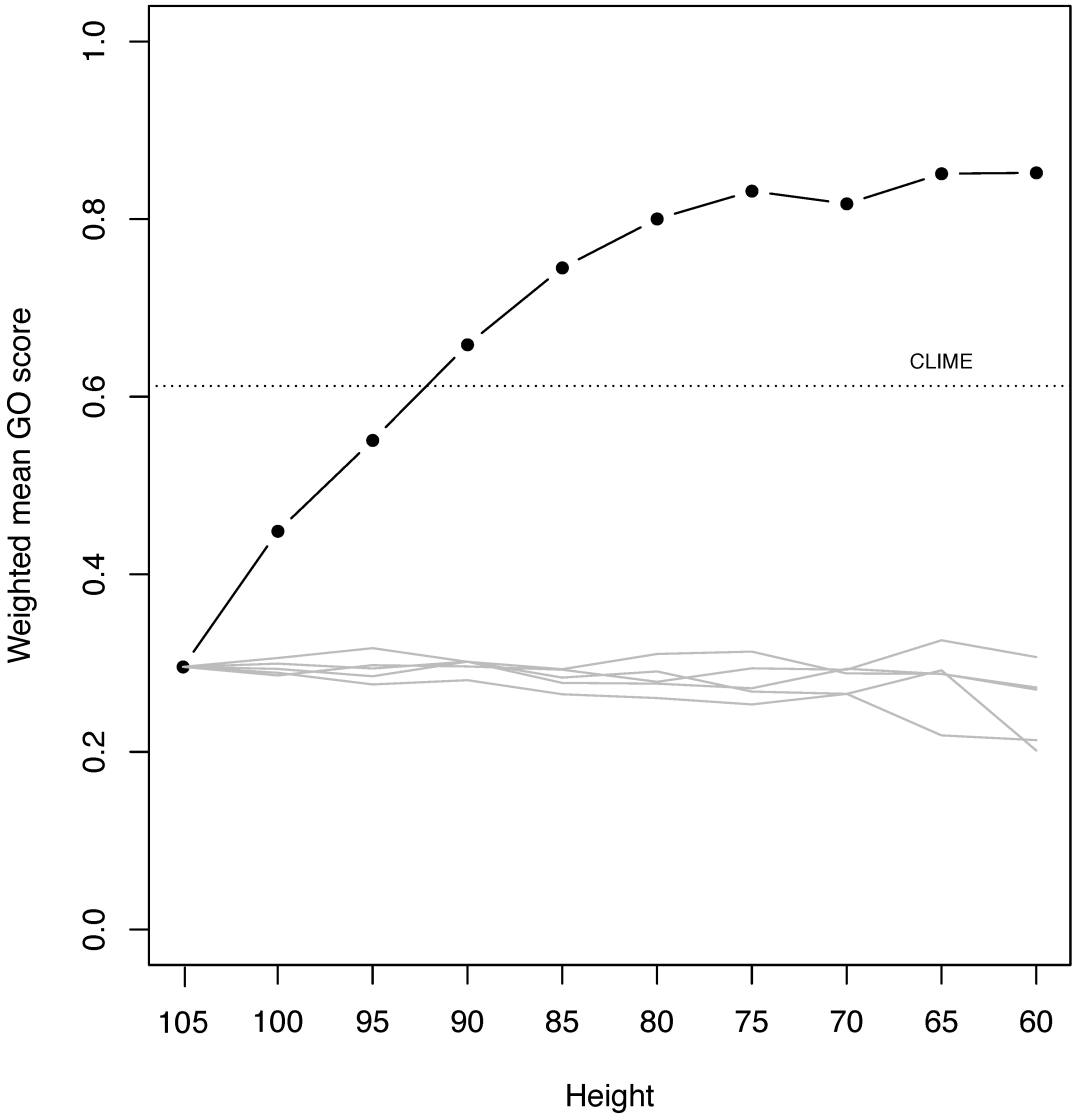


Figure 2.5: Comparison of functional-similarity scores within clusters. The similarity was evaluated using Gene Ontology (GO) terms, for CLIME (dashed line) and the hierarchical approach at different h thresholds (solid line). Gray lines show the distribution of similarities obtained for five sets of clusters obtained by randomly reassigning tip labels in the cluster tree.

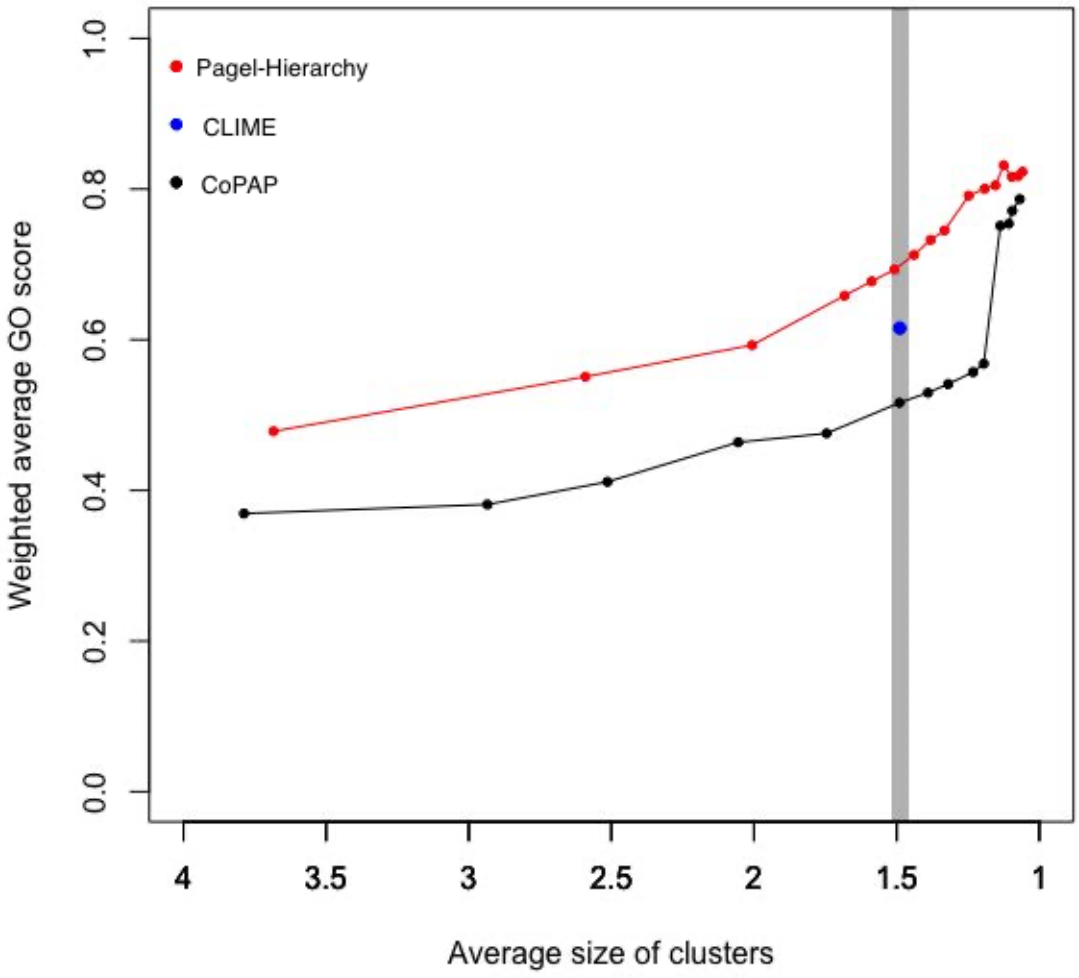


Figure 2.6: The performance of our method and CoPAP at different cut-offs. The x-axis represents the average size of clusters generated by the corresponding cut-offs and the y-axis is the weighted average GO score. The red, blue and black dots represent our method, CLIME and CoPAP respectively. The three data points in the shaded area are further studied using a significance test in Figure 2.7

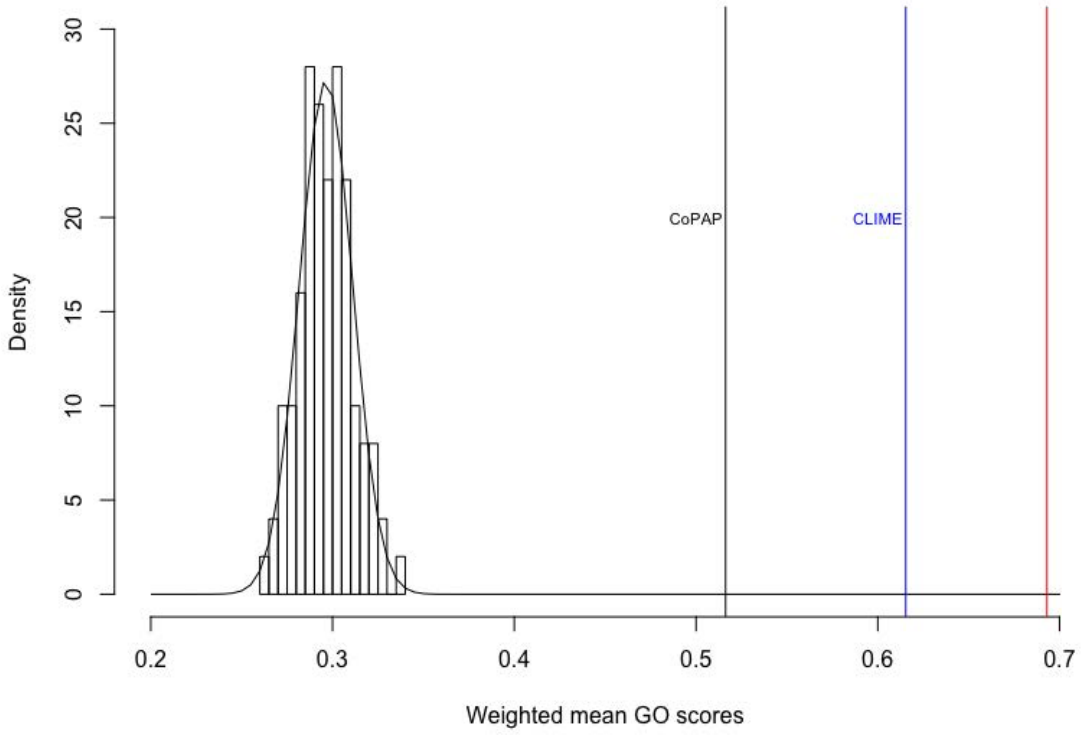


Figure 2.7: The significance of the performance of three clustering methods. The histogram is generated by 100 randomized assignments of genes corresponding to the size disbribution of CLIME's clustering results. The three vertical lines represent CoPAP, CLIME and our method at the cut-offs in the shaded area of Figure 2.6

2.3.4 Biological significance

To test the biological significance of our clustering results, we implemented a GO enrichment test by comparing the Pearson's chi square statistics between the observed gene clusters and 100,000 randomly generated clusters in the same size distributions. We used the Benjamini-Hochberg FDR procedure to control for multiple tests. Figure 2.8 shows the significance of the non-rare GO terms (frequency ≥ 5) in our gene set at different cutting heights of the hierarchical dendrogram (the details of the test are provided in Table S2 available at `https://doi.org/10.1093/gbe/evy178#supplementary-data`). Among those interesting gene clusters, we take the amino-acid biosynthesis and motility-associated gene clusters as illustrative examples.

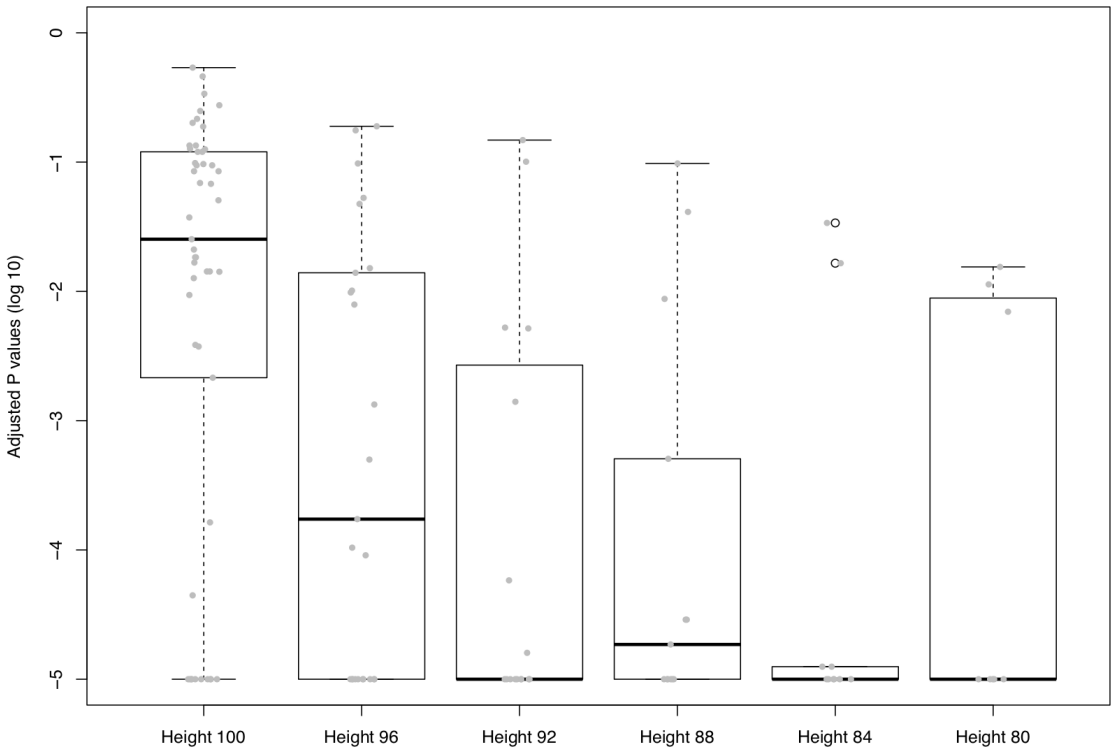


Figure 2.8: Adjusted P values of the nonrare GO terms obtained at different cutting heights of the hierarchical dendrogram. Each dot represents a GO term with frequency > 5 in our gene set at different cutting heights.

2.3.5 Pathway mapping of an amino-acid biosynthesis cluster

From Figure 2.5, it is anticipated that many clusters will have a high degree of functional cohesion based on the GO scores. However, examination of clusters with even relatively low GO similarity scores still showed strong functional similarities in spite of annotations with different terms. We examined in detail a cluster containing 28 distinct profiles, with a height of 100 and a GO score of 0.62 (Figure 2.9). A striking property of this cluster is that it contains many profiles that either include or exclude all of the 20 C. difficile genomes, whereas a phylogenetically naïve approach would assign a great deal of significance to this difference. Although the GO terms in this cluster are not identical, the majority of profiles assigned to the cluster are associated with amino-acid biosynthesis. Several amino-acid biosynthesis pathways are represented, including leucine, isoleucine, histidine, valine, tryptophan, glutamate / glutamine, and cysteine. Many of these pathways are tightly interconnected, notably valine and leucine, but some pathways, in particular histidine, are more distant. Figure A.4 shows how the proteins in this cluster connect in the corresponding "Valine, Leucine and Isoleucine biosynthesis" and "Phenylalanine, Tyrosine and Tryptophan biosynthesis" pathways (KEGG database).

2.3.6 Phylogenetic analysis of a motility-associated cluster

Figure 2.10 shows a functionally cohesive cluster consisting of the proteins related to flagellar assembly and motility, and their corresponding phylogenetic profiles. In addition to predicting the functions for unannotated genes in this cluster, we can also infer LGT events based on the evolutionary pattern we found. The patchy distribution of flagellar gene profiles supports a history that includes many LGT events. To assess the phylogenetic cohesion of the genes in this cluster, we constructed individual phylogenetic trees and compared them with the reference full tree. All flagellar proteins from LZ were grouped with almost the same set of other genomes (Table S3 available at `https://doi.org/10.1093/gbe/evy178#supplementary-data`): *Clostridium hylemonae* DSM15053, Clostridiales bacterium VE202-28, Clostridiales bacterium 1_7_47FAA, *Clostridium bolteae* 90B7, *Clostridium bolteae* 90B8, none of which was a close neighbour in the reference tree or the SILVA taxonomy.

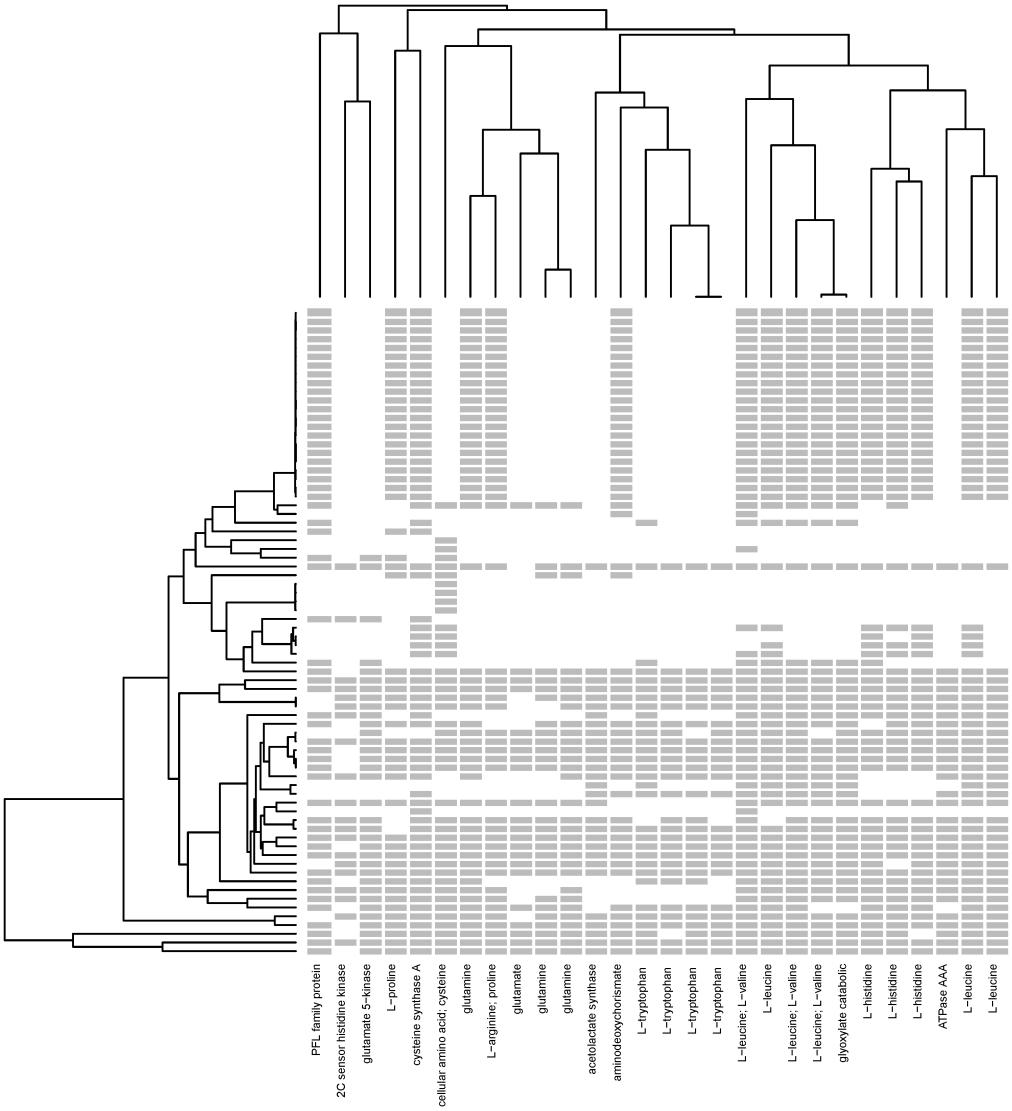


Figure 2.9: Structure, phylogenetic distribution and functional categories of a hierarchical cluster enriched in amino-acid biosynthesis proteins. Each column represents a gene profile; the gray bars indicate the presence of genes and blanks indicate the absence. The dendrogram on the left side is the phylogenetic tree of 74 genomes and the dendrogram on the top is computed by the distance between the profiles. The labels on the x-axis are the genes' functional annotations retrieved from the UniProt database.

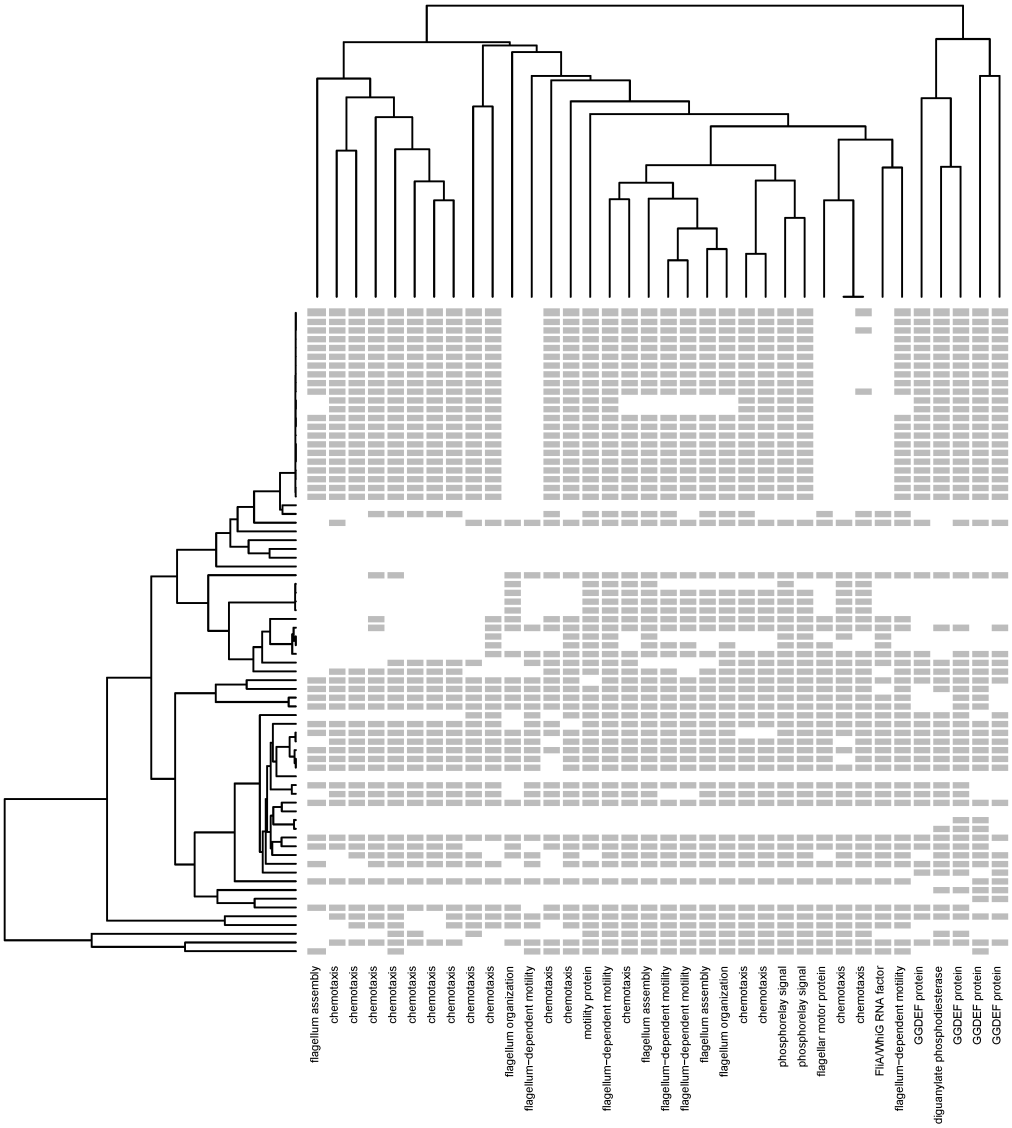


Figure 2.10: Structure, phylogenetic distribution and functional categories of a hierarchical cluster enriched in flagellar motility proteins. Each column represents a gene profile; the gray bars indicate the presence of genes and blanks indicate the absence. The dendrogram on the left side is the phylogenetic tree of 74 genomes and the dendrogram on the top is computed by the distance between the profiles. The labels on the x-axis are the genes' functional annotations retrieved from the UniProt database.

2.3.7 Connections between LZ and *C. bolteae*

LZ has a large genome containing 6887 protein-coding genes, whereas the median genome size in our data set is 3728 genes, and the mean genome size is 3580.4. One possible explanation for this disparity is a genome expansion due to gene duplication and LGT. The two strains of *C. bolteae* whose flagellar genes were proximal to those of LZ may show similar evidence of LGT with LZ and its close relatives. To address this possibility, we examined the homology-search results of LZ versus all genomes, and defined criteria to represent "unexpected similarity" between LZ proteins and their corresponding homologs in *C. bolteae*. We set threshold criteria that required the e-values of the match between the LZ protein and the *C. bolteae* protein sequence be less than 10^{-20}, and also required that the *C. bolteae* protein rank within the top 20 of all matches from a given LZ protein to the entire database of 687 genomes. We then identified the corresponding profiles in our cluster tree, and implemented a binomial test to identify clusters in which these proteins were significantly overrepresented.

Individual profiles of proteins belonging to the flagellar-associated cluster were subjected to phylogenetic analysis. Each of the 36 profiles was first aligned using MUSCLE version 3.7 [23] with default parameters. The resulting alignments were used to infer phylogenetic trees using RAxML version 7.2.5, with the PROTGAMMALG model of sequence substitution. Using this approach, we identified the flagellum / motility cluster and three additional clusters. Although each of these clusters showed significant overrepresentation of the identified proteins, none was completely homogeneous in this regard: however, a majority of profiles did contain representatives from at least one of the two *C. bolteae* genomes (strain 90B7, 97.8%; strain 90B8, 89.6%). The predominant GO annotations within the three additional identified clusters comprised (i) relaxase/mobilization nuclease and ParB-like partition proteins (Figure A.3a); (ii) sequence-specific DNA binding, ATPase activity, and methyltransferase activity (Figure A.3b); and (iii) PAS domain S-box protein, two-component system hybrid sensor kinases, response regulators, and diguanylate cyclase (GGDEF) domain-containing proteins (Figure A.3c). Each cluster covers different subsets of the sampled genomes, but in each case, it is difficult to explain the distributions with only gene-loss events, suggesting an important role for LGT. Many of the functional annotations of these proteins are very general, but cluster (i) suggests

a possible role for plasmid-based transfer, and (ii) suggests that environmental sensitivity and response may be adaptive in a niche occupied by LZ (and/or *C. bolteae*).

2.4 Discussion

Phylogenetic profiles were initially developed at a time when the relatively few sequenced genomes available were phylogenetically very diverse, and their similarity due to common descent was not explicitly incorporated into profile-similarity calculations. However, the intensive focus on sequencing many strains of some named species, notably those species that contain pathogenic isolates, has led to highly uneven sampling across the breadth of microbial diversity. An example in our dataset is the > 100 sequenced genomes of *C. difficile*, of which 21 were retained in our subsampled dataset. Our new phylogenetic-profile-based clustering approach successfully addresses these phylogenetic correlations using Pagel's method for the comparative analysis of discrete characters. The success of our approach is most striking in the many instances we show where clustered profiles can differ in the presence or absence of all 21 strains of *C. difficile*, whereas a phylogenetically naïve approach would assign a very large distance between such profiles. Furthermore, by explicitly allowing multiple gains of proteins in the tree rather than a single common ancestor followed by potentially many gene losses, our method is more suitable than CLIME in the analysis of LGT-prone prokaryotic genomes.

LZ has a very large genome relative to most other clostridia, and elucidating its ecological role will be challenging. However, by examining a subset of its clusters, we can identify not only specific functions that appear to be present in the genome of LZ, but also identify sets of proteins with similar (but not identical) distributions. Our analysis of even a small subset of LZ clusters shows a complex set of relationships with other genomes, but also highlights the functional cohesion of our recovered clusters. In the cases of amino-acid metabolism and flagellar / motility genes, complementary evidence from pathway diagrams, genetic linkage, and phylogenetic trees supports our inferred connections. We were also able to focus on a small subset of clusters in which LZ appeared to have unusual patterns of similarity to *C. bolteae*; the identified genes provided clues to transfer mechanisms, environmental adaptation, and potentially (in the case of flagella) pathogenicity [88].

Protein functional prediction is one of the greatest challenges in bioinformatics [28, 84, 78]. Although we did not explore the effectiveness of this method in functional prediction of proteins, the functional cohesion of many of our recovered clusters suggests that it may have value as a predictive tool. Since we use the techniques that are based only indirectly on homology search, it may prove to be complementary to homology-based (PSI-BLAST) and other approaches.

Since more genomes provide more opportunities to differentiate profiles and give further resolution to clusters, a method that can consider all available genomes would be desirable. One significant limitation of our method is the heavy computational cost of applying Pagel's coevolutionary method to all pairs of distinct phylogenetic profiles: although our full dataset included 687 genomes, computational time limitations restricted us to the analysis of a set of 74 genomes. Even this reduced computation required a total of 13,000 CPU hours approximately on a Linux system. Our future work will consider alternatives to Pagel's method including phylogenetic regression and HMMs that also take phylogenetic correlations into account, and heuristics to subdivide the full tree into tractable subsets of taxa to perform the analysis, then merge the results to obtain a full set of distances. However, our results on even a small subset of available genomes demonstrate that our phylogenetic-profile-based clustering method has the capacity to identify sets of genes with similar distributions and evolutionary histories, with the potential to represent genomes as distinct combinations of these sets, thereby highlighting the important genetic and environmental connections between them.

2.5 Author Contribution

This study extends the work done during my Master's thesis by implementing an improved framework for evaluating the performance of methods, adding result comparisons with other two methods: CLIME and CoPAP, testing the influence of tree rooting and adding more detailed phylogenetic analysis.

Sequencing of the LZ genome was undertaken on behalf of the Human Microbiome Project Consortium and generously supported by the NIH, NHGRI, and NIAID (U54-HG004969 to the Broad Institute). We thank the members of the Broad Institute's Genome Sequencing, Assembly and Annotation Teams, including Sarah Young, Peg

Priest, and Qiandong Zeng. In this study, I have participated in the design of the work, implemented the methods, conducted the analysis and wrote the manuscript. Thanks to Benjamin Wright and Dr. Emma Allen-Vercoe for providing the LZ data sets (the phylogenetic tree and phylogenetic profiles).

Chapter 3

The Community Coevolution Model with Application to the Study of Evolutionary Relationships between Genes based on Phylogenetic Profiles

3.1 Introduction

Comparative-genomic techniques can be used to identify homologous genes that underpin a multitude of traits [48, 35]. Genes can exhibit similar patterns of presence and absence [72] across a set of genomes for reasons such as participation in a common biochemical pathway, physical linkage, or co-localization on a mobile genetic element such as a plasmid [27, 12, 17]. Examination of these patterns can reveal important information about related functions (e.g., participation in a common biochemical pathway) and common pathways of lateral gene transfer (LGT). A well-established approach to represent presence and absence patterns among genes is the construction of *phylogenetic profiles* [71, 64, 60].

The success of phylogenetic profiling depends on the use of appropriate measures to express the distance and similarity between profiles. Approaches include the Hamming distance [72], mutual information [39], Pearson correlation [81], and the hypergeometric test [101]. Although effective, these approaches do not take phylogenetic effects into account. Since closely related genomes are more likely to share similar gene content, they are likely to have an outsized influence on profile comparisons relative to their phylogenetic diversity. Thus, the genomes connected by the phylogenetic tree are not independent [76], which will violate the assumptions of these methods, and therefore skew results. Large genomic datasets are often imbalanced due to high relative abundance or oversampling of certain genomes; for example, the over-representation of pathogen isolates in the set of complete prokaryotic genome sequences [2].

Several heuristic approaches have been developed to account for phylogenetic effects in profiles. For example, [42] and [81] used a null distribution of the similarity scores inferred by sampling the genomes to estimate the impacts of phylogenetic correlation; while [97] corrected for biases in the number of sequenced genomes by collapsing the genomes within the same clade into one single node if a specific gene pair has the same phyletic pattern in these genomes. [16] first ordered the genomes within profiles and enumerated runs of consecutive matches so that the co-occurrences concentrated in part of the tree will be considered as only one run. The shared underlying idea of these methods is the application of a weighting scheme to the genomes in order to counteract phylogenetic effects. These methods can be computationally efficient and feasible for large-scale analysis [91, 89], but they are *ad hoc* approaches that do not properly model the underlying evolutionary processes.

In contrast with weighting approaches, evolutionary models aim to explain the distribution of genes by modeling the correlation patterns of gain and loss on a phylogenetic tree. Model-based approaches include CoPAP which uses a stochastic mapping approach to detect co-evolving gene families [14, 15], the CLustering by Inferred Models of Evolution (CLIME) algorithm that was developed to model gene evolution in eukaryotes [51], and the Count software package for the analysis of numerical profiles using phylogenetic birth-and-death models [18]. However, Count was not specifically developed for binary traits; CLIME assumes that each gene has a single gain event in evolution which is not suitable for prokaryotes which have high rates of gene transfers [98]; and CoPAP assumes that the gain rate and loss rate independently vary among genes rather than explicitly modeling the interactions during evolution. Pagel [66] developed a likelihood-based co-evolutionary method that specifically tests the evolutionary correlations between pairs of binary traits. In Pagel's method, each pair of binary traits is evaluated under both independent and dependent models, and a likelihood ratio test is applied to infer whether there is significant evidence suggesting two traits evolved dependently. Although Pagel's correlation model performed well in previous studies at detecting functionally linked genes [54, 6], Pagel's method is computationally expensive and it can not directly infer the direction (positive / negative) of the correlation. In addition, there is a more general concern regarding the phylogenetic comparative methods represented by Pagel's correlation model raised by

[56] and [95] that comparative methods may overestimate the evidence for correlation of the patterns caused by singular events which they refer to as Darwin's scenario. Darwin's scenario occurs when two traits have a single origin on the same lineage, and are then inherited by nearly all species in the descendant clade, resulting in (almost) perfectly co-distribution. This "within-clade pseudoreplication" could lead to dubious conclusions such as a significant association between fur and middle earbone [56].

Furthermore, most of the existing methods such as Pagel's correlation method and distance-based methods can only be applied to pairs of genes. However, studying phylogenetic profiles in higher-order groups (such as triplets and quadruplets) can offer a more-sensitive approach to detecting complex patterns of correlation [102]. Direct-coupling analysis (DCA) is a class of methods often used to infer direct relationships between residues in biological sequences that can deal with conditional dependency by taking the inverse of the covariance matrix, but it is mainly for continuous data and phylogeny naïve, so it is highly dependent on the sampling of the genomes [62, 4].

Here we propose the Community Coevolution model, a new method that directly infers the strength and direction of the interactions among genes during the evolutionary process. For a pair of genes, CCM is more efficient in that it fits only one model instead of three (one separate model for each gene and one dependent model) in Pagel's method and is approximately five times faster than Pagel's method when tested on phylogenetic trees with 500 tips (performed on a server running Linux with a 2.67 GHz CPU and 18 GB RAM). Although maximum-likelihood estimation is still a time-consuming procedure compared to other heuristic methods based on standard metrics, our method provides more biological insights such as the evolutionary rates, significance levels and directions (positive/negative) of interactions, and more importantly, our method can be extended to model multiple genes as a community to discover more-complex associations. We also develop a simulation procedure to generate phylogenetic profiles with adjustable extents of evolutionary interactions that can be used as benchmark data for evaluating comparative methods.

3.2 Materials and Methods

3.2.1 The Community Coevolution Model

In our community coevolution model (CCM), we consider whether sets of two or more genes have potential associations on a given phylogenetic tree, in particular whether the transition (gain or loss) of any gene within the community is affected by the current states of other members. Associations between genes can be positive if genes tend to be gained and lost together, and negative if the gain of some genes in a set appears to be associated with the loss of others in the same set. Gene sets that show evidence of associations are termed as a "community" in our model.

We formulate the transition rate τ for one gene as a function of its intrinsic rate of gain and loss μ, and the association factor ω depending on the current states of all other genes in the community,

$$\tau = \mu \times \omega. \tag{3.1}$$

To further specify our model, we use the following notation:

n is the total number of genes in the community;

$\mathcal{S} = \{S_1, S_2, ..., S_{2^n}\}$ is the state space of a community consisting of n genes; $S_i = \{x_{i,k}; k = 1, ..., n\}$ is a vector of size n. $x_{i,k}$ denotes the state of the specific kth gene when the community state is at S_i. We define $x_{i,k} = -1$ when the kth gene is absent and 1 when the kth gene is present;

$\mathcal{B} = \{\beta_{hk}; k, h = 1, ..., n\}$ is a symmetric $n \times n$ matrix whose off-diagonal entries are the coefficients of interaction between every pair of genes and diagonal entries indicate half the difference between the gain and loss rates of each gene;

$\mathcal{A} = \{\alpha_1, \alpha_2, ..., \alpha_n\}$ is a vector of size n containing the intrinsic rate which is the mean of gain and loss rates for each gene;

The instantaneous transition rate for a specific gene k when the whole gene community is in state S_i is defined in the log scale as

$$\log(\tau_{i,k}) = \alpha_k - \beta_{kk}x_{i,k} - \sum_{h \neq k}^{n} \beta_{hk}x_{i,k}x_{i,h} \tag{3.2}$$

Positive β_{hk} means the kth and hth genes are positively associated, thus if the current states of kth and hth genes are the same ($x_{i,k}x_{i,h} = 1$), the last term tends to reduce the rate of change for gene k; and vice versa for negative values of β_{hk}.

By taking the exponential of equation (3.2), we have the final model as

$$\tau_{i,k} = \underbrace{\exp(\alpha_k - \beta_{kk}x_{i,k})}_{\mu} \cdot \underbrace{\exp\left(-\sum_{h \neq k}^{n} \beta_{hk}x_{i,k}x_{i,h}\right)}_{\omega} \tag{3.3}$$

where the first part represents the intrinsic gain / loss rates of gene k and the latter part represents the influence from the community.

We model the gene state changes along a phylogenetic tree as a continuous-time Markov process, and assume the instantaneous rate for all transitions involving more than one gene is 0. The transition rate matrix $Q = \{q_{ij}; i, j = 1, ..., 2^n\}$, where element q_{ij} denotes the rate of the community departing from state S_i and arriving in state S_j, can be constructed in accordance with the following rule,

$$q_{ij} = \begin{cases} \tau_{i,k}, & i \neq j \text{ and } S_i - S_j = \pm 2\mathbf{e}_k \\ -\sum_{i' \neq i} q_{ii'}, & i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

where $\{\mathbf{e}_k; k = 1, ..., n\}$ denote the standard basis vectors of all 0's except the kth element as 1 so that $S_i - S_j = \pm 2\mathbf{e_k}$ indicates that only the kth gene changes state.

3.2.2 Constructing the likelihood function given the tree

In addition to the Markov assumption, we also assume that transitions on separate branches are independent. This means that the distribution of the state at the end of a given branch depends only on the starting state of that branch. The computational cost of constructing the likelihood function could be expensive as we need to sum over all the possible combinations of the states at each internal node, but it can be

reduced by applying Felsenstein's pruning algorithm [24]. The pruning algorithm is a dynamic-programming approach that takes advantage of the nested structure of the tree and computes the likelihood for the given tree recursively. By applying the pruning algorithm, the likelihood function L of the tree in Figure 3.1a can be formatted as

$$L(\Theta;\mathcal{T},X)=\sum_{s_0}\left[\left(\sum_{s_1}P_{s_0,s_1}(b_1)P_{s_1,s_3}(b_3)P_{s_1,s_4}(b_4)\right)\times\left(\sum_{s_2}P_{s_0,s_2}(b_2)P_{s_2,s_5}(b_5)P_{s_2,s_6}(b_6)\right)\right]. \tag{3.5}$$

In this way, the likelihoods for subtrees can be reused and the computational complexity is reduced to linear in the number of leaves in the tree. Then the negative log-likelihood function $-\log(L(\Theta;\mathcal{T},X))$ is minimized to acquire the maximum likelihood (ML) estimates of the parameters by using a quasi-Newton optimizer (*nlminb* in *R*, version 4.0.2)[69].

3.2.3 Inference and Regularization of the Maximum Likelihood Estimates

Due to the complexity of the likelihood function, it is necessary to assess whether the optimizer is going to provide acceptable estimates. We examined two ways to obtain the standard error of estimates: first, the parametric bootstrap method that simulates a large number of profiles using the estimated parameters and calculates the standard deviation of the estimates using the bootstrap samples; second, the analytical approach based on likelihood theory which utilizes the numerically approximated Hessian matrix H of the objective function $-\log(L(\Theta))$, with the standard error given as $\text{se}=\sqrt{\text{diag}(H^{-1})}$. Using this estimated standard error, a Z-test is conducted to obtain the p-value for the hypothesis $\beta_{hk}=0$. The bootstrap method is obviously more time-consuming, but we can use the bootstrap to assess the accuracy of the likelihood asymptotic for finite samples. The performance of these two methods is compared in the Results section.

As tree and community size increases, the likelihood function can become extremely complicated. A potential problem with the MLE procedure is overshooting, which happens when the parameter estimators diverge substantially from the true

values due to a flat likelihood surface. To avoid the overshooting problem, we apply l_2-regularization on the parameters and have the penalized objective function

$$-\log\left(L(\Theta;\mathcal{T},X)\right)+\lambda\cdot\left(||\Theta||_2^2\right), \tag{3.6}$$

where λ is the tuning parameter. Unlike setting a boundary for parameters or allowing a large error tolerance, which could cause an early stop of the optimization process to avoid overshooting, the regularization approach leads to more stable estimations by adding smoothness to the surface of the likelihood function.

The l_2-regularization is not meant for model sparsity, but only for dealing with computational issues of likelihood singularity in some occasional cases, and therefore it is not needed in most analyses which avoids unnecessary bias. When it is needed, a reasonable λ is desired to avoid the overshooting problem but without introducing too much bias into the estimation. The condition number, which is the ratio of the largest eigenvalue to the smallest of the Hessian matrix, describes the rate of convergence of the optimization [90], and can be used as a guide to find a proper λ. To provide a rule of thumb, we find that a condition number below 200 generally indicates a successful convergence without the overshooting problem.

3.2.4 Simulation Procedures

To simulate the coevolutionary relationships among genes, we use the framework of CCM that the transition rates of one gene depend on the states of other genes in the community. The procedure for simulating the evolution of a gene community of size n on one branch can be summarized as below:

Input: the starting state of the community $\mathcal{S}$, a user-defined coefficient matrix $\mathcal{B}_{n\times n}$, user-defined intrinsic rates $\mathcal{A}_0$ and branch length b.

1. Substitute the current states and user-defined parameters into Formula 3.3 to calculate the current transition rate for each gene $\tau_h, h=1,...,n$.

2. Sample the transition time for each gene from the exponential distribution, $t_h \sim \text{Exp}(\tau_h), h=1,...,n$.

3. Find the gene k with the minimum transition time, $T_{\min}=\min\{t_1,t_2,...,t_n\}$.

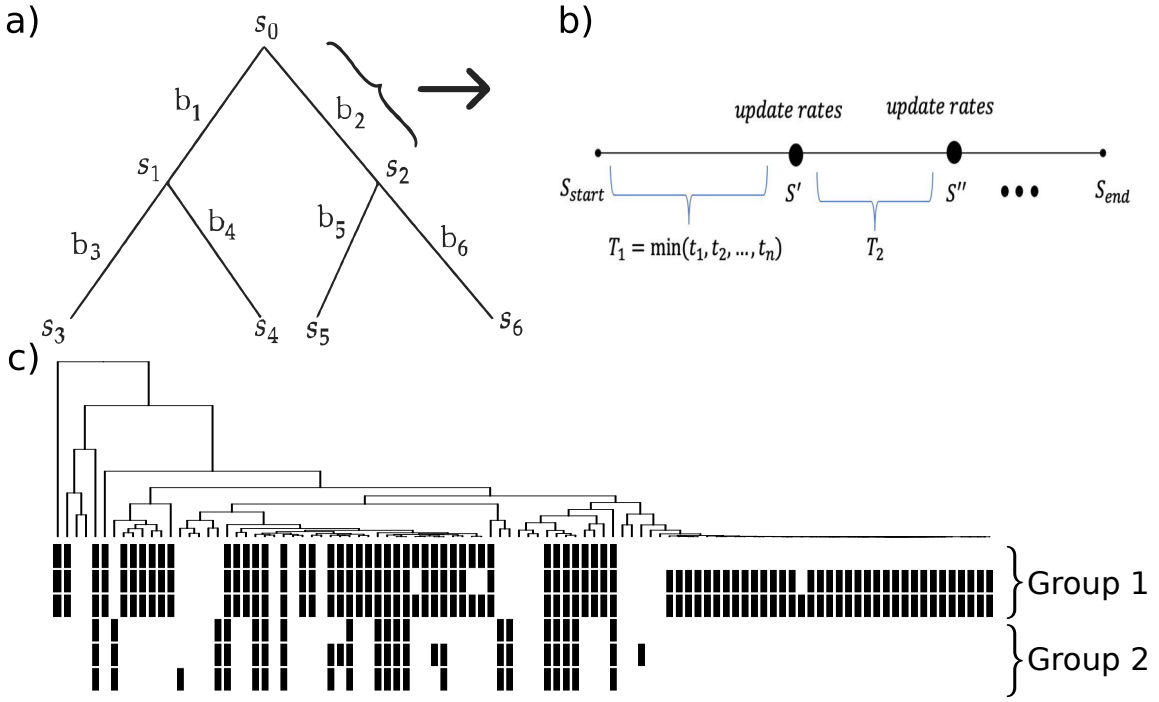


Figure 3.1: (a) A phylogenetic tree with 4 tips: s_i represents the state at each node and b_i denotes the branch length. (b) An illustration of the simulation process on one branch. S denotes the community states and T indicates the time that there is a transition out of the current state. The process ends when the total transition time is beyond the branch length. (c) A random realization of two groups of correlated profiles of size 3 generated by our simulation procedure. The interaction coefficient is set to be 1.5 within a group and 0 between groups. Each row is a profile and each gray bar denotes presence of the gene.

4. If $T_{\min} \leq b$, update the state of gene k in $\mathcal{S}$ with the opposite state and if $T_{\min} > b$, do not update the gene state. Then update the branch length $b = b - T_{\min}$.

5. Repeat steps 1-4 until $b \leq 0$.

Output: the new state of the community $\mathcal{S}'$.

An illustration of the evolutionary process on one branch is shown in Figure 3.1b. Then the end state $\mathcal{S}_{\text{end}}$ will be the starting state for the next adjacent branches. The same procedure will be applied to all branches sequentially from the root to the tips. Figure 3.1c shows a simulation example of 6 genes in two groups of size 3 using the interaction matrix which has within-group interaction coefficients equal to 1.5, indicating strong relationships and between-group interaction coefficients equal to 0, indicating independent evolution.

3.2.5 Analysis of genomes from class Clostridia

We applied our method to the draft assembly of the bacterium "Lachnospiraceae bacterium 3-1-57FAA-CT1" (abbreviated as LZ), which was isolated from a biopsy retrieved from the transverse colon of a female Crohn's Disease patient at the time of colonoscopy [54]. LZ is of interest because its genome size is very large compared to most of its immediate neighbours (6505 protein-coding genes as compared with a median of 3124 in our complete data set of Clostridia) and identifying sets of genes with shared patterns of gain and loss may yield insights into its ecological role in the host. 658 completed and draft genomes from class Clostridia were retrieved from the National Center for Biotechnology Information (NCBI) for the comparative analysis of LZ. The phylogenetic tree was built through the AMPHORA2 pipeline [103] and RAxML-HPC [87] using their concatenated, conserved protein sequences and another set of eight outgroup genomes from class Bacilli and phyla Actinobacteria and Proteobacteria were used for rooting. The phylogenetic profiles were constructed by comparing the complete set of LZ (6505 predicted genes) against all other genomes using rapsearch [104]. Before our analysis, we firstly filtered out the genes that are very rare (present in $< 1\%$ genomes) or very common (present in $> 99\%$ genomes) and obtained the final data set of 3786 profiles. The Markov Clustering algorithm (MCL) was also used to obtain clusters of genes. MCL is a graph-based clustering

method that simulates random walks within the graph to reflect the cluster structure based on the idea that random walks are more likely to stay in one natural cluster than to move across clusters [22].

3.3 Results

3.3.1 Results on Simulated Data

Evaluation of model estimates

We first evaluated the performance of CCM on simulated data. We used the parameters estimated from one pair of flagellar genes in the real data set, with profiles shown in Figure 3.2a. We simulated 100 pairs of genes using the parameters estimated from these two genes. The results are shown in Figure 3.2b. We see that the estimates are distributed around the true parameter values. Furthermore, we compared the estimates of the standard error based on the Hessian matrix and the parametric bootstrap and Table 3.1 shows that the estimation results of two methods are consistent.

Table 3.1: Comparison of estimated standard error using the parametric bootstrap and analytical Hessian methods based on the simulations in Figure 3.2.

	α_1	α_2	β_1	β_2	β_{12}
Bootstrap SE	0.282	0.215	0.107	0.099	0.090
Hessian SE	0.382	0.255	0.104	0.105	0.085

Detection of significant interactions between genes

We next used our simulation approach to examine the ability of the CCM to distinguish genes with associations from those that do not interact. To evaluate the performance of CCM, we simulated 500 gene pairs with no interaction ($\beta_{12} = 0$) as negatives and 500 gene pairs with interactions (β_{12} uniformly drawn between 0.2 and 0.5) as positives. Figure 3.2c shows the distributions of the estimated coefficients of interaction in two groups: the mean value for the "no interaction" group is 0.0046 ($\pm$ 0.0807) and for the "interaction" group is 0.3497 ($\pm$ 0.1173). We also compared the performance of Pagel's correlation test method, the Jaccard Index ($= \frac{\text{number of genomes that have both genes}}{\text{number of genomes that have either of genes}}$) and two heuristic methods: hypergeometric

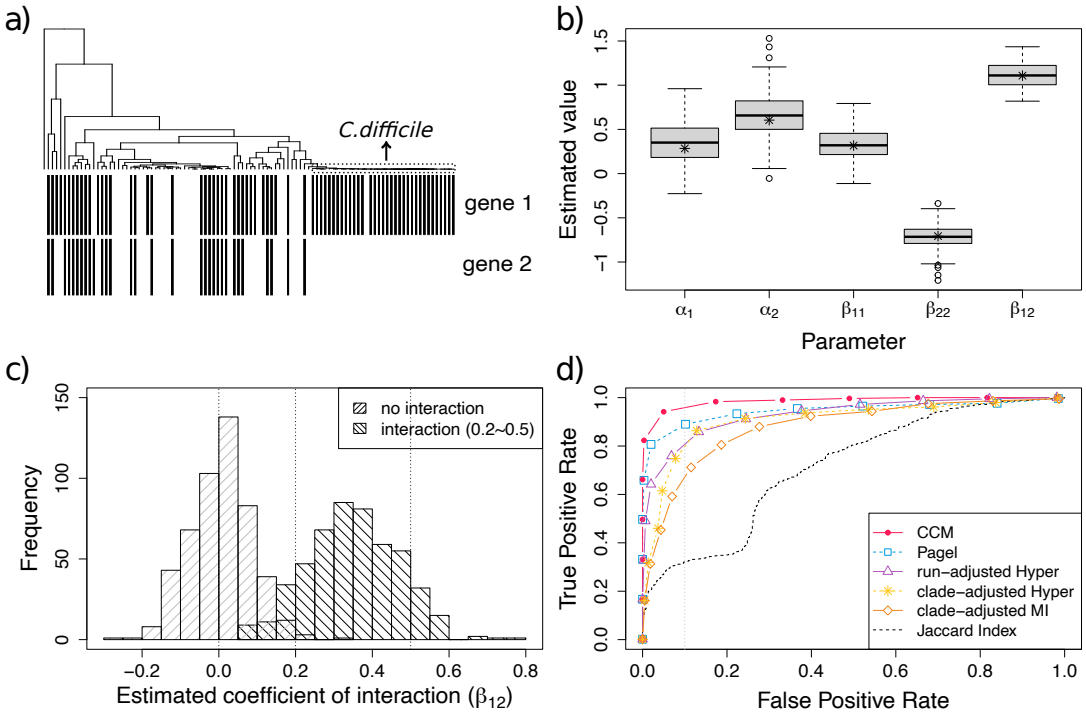


Figure 3.2: Estimation of the parameters using simulated pairs: (a) Two phylogenetic profiles from the real data set; (b) Estimated parameter values from CCM based on 100 simulated pairs using the parameters estimated from the two profiles in (a). The "*" represents the true parameters used in simulation. Evaluation of the interaction in pairs: (c) The distributions of the estimated coefficients of interaction of the "no interaction" group and the "with interaction" group; (d) the ROC curves of detecting the significant linkages by Jaccard Index, Pagel's correlation method, clade-adjusted mutual information and hypergeometric, run-adjusted hypergeometric and our CCM model.

with consecutive runs method [16] and mutual information with clade adjustment method [97]. We evaluated both clade-adjusted and runs-adjusted methods with four different metrics (Hamming, Jaccard Index, Hypergeometric test, and Mutual Information) using simulated data and we found that the hypergeometric test works best. Thus we also included the clade-adjusted Hypergeometric test for comparison since it performed better than the method proposed in their original paper (clade-adjusted mutual information). From the ROC curve shown in Figure 3.2d, our CCM method obtained the highest AUC score of 0.9521 followed by Pagel's correlation model (0.8968), run-adjusted Hypergeometric (0.838), clade-adjusted Hypergeometric (0.7665), and clade-adjusted Mutual Information (0.7215). The Jaccard Index, being a non-phylogenetic method had the lowest AUC score (0.609).

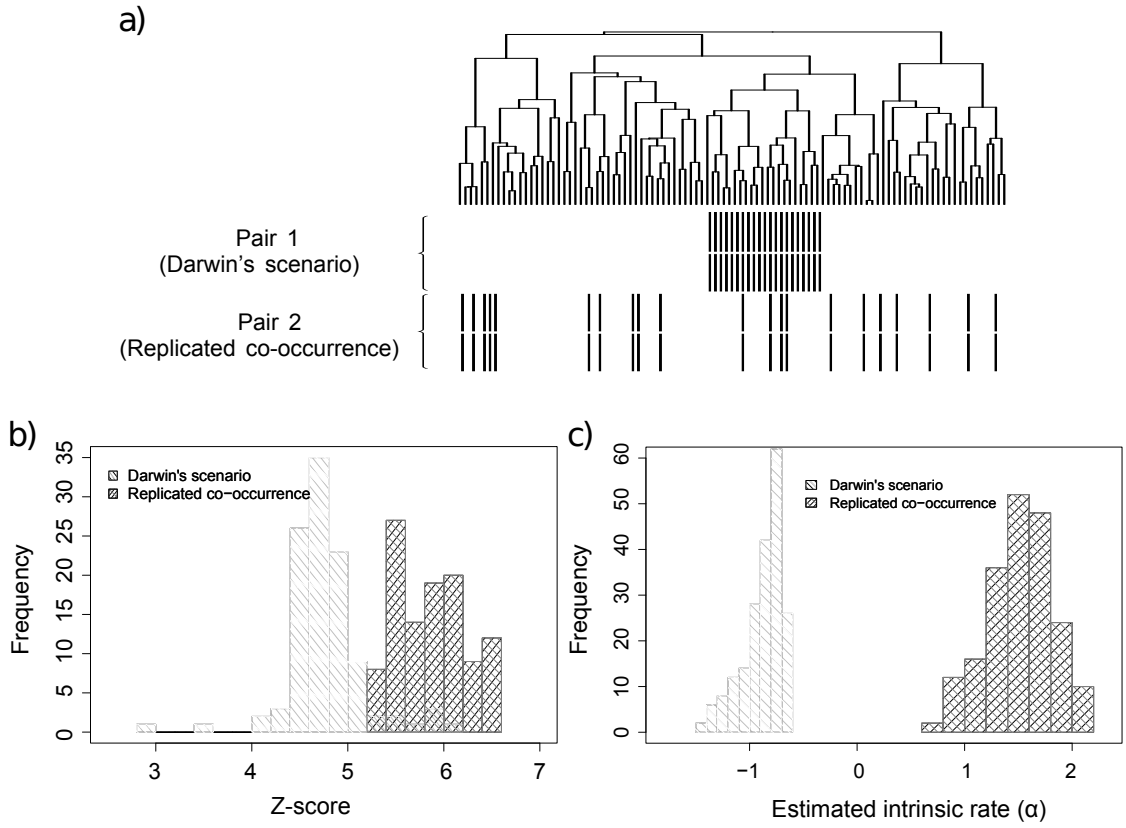


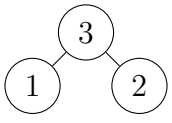
Figure 3.3: Comparison between Darwin's scenario and replicated co-occurrence: (a) An example of Darwin's scenario (Pair 1) and replicated co-occurrence (Pair 2); (b) The distributions of the Z-scores ($\frac{\beta_{12}}{se(\beta_{12})}$) for the two scenarios; (c) The distributions of the estimated intrinsic rates for the two scenarios.

Identifying Darwin's scenario of co-distribution

Under Darwin's scenario, there is a single concurrent origin for two genes leading to the perfect co-distribution across all species within a clade as in the example shown in Figure 3.3a. As Darwin's scenario provides little evidence for coevolution, it is of interest to distinguish such scenarios from a replicated co-evolution scenario that has multiple disjoint instances of a given trait. Both scenarios are considered significantly correlated by CCM due to their perfect co-distribution, but the replicated co-evolution scenario yields stronger significance scores and has much higher intrinsic rates. We demonstrate this difference using 100 simulated data sets. In each simulation, we randomly generate a 100-tip tree, one pair of genes that have co-occurrences concentrated in one random clade chosen uniformly across all clades as Darwin's scenario and another pair of genes that have same number of co-occurrences spreading across the tree as the replicated co-distribution scenario. Although both scenarios produce significant results (p-value< 0.005), replicated co-occurrence tends to achieve greater significance scores (Fig. 3.3b). Very few of the Darwin scenarios would be deemed significant in a real data analysis with correction for multiple testing. Another distinguishing feature between these scenarios is the estimated intrinsic rate α, with gene pairs simulated under Darwin's scenario having much lower estimates of α (-0.88 ± 0.18) than under replicated co-evolution scenario (1.51 ± 0.29) as shown in Figure 3.3c.

Modeling multiple genes as a community to reduce pairwise false-positive links

Most comparative methods (e.g. Pagel's method and all the methods based on distance or similarity metrics) use pairwise comparisons. By modeling more than two genes as a community, the CCM can screen out false-positive links that can be caused by genes that show pairwise evidence for co-evolution, but are conditionally independent when other genes are taken into consideration. For example, consider a community with the following structure:



where gene 3 is directly related to both 1 and 2, but there is no direct connection between gene 1 and gene 2. Pairwise methods will often falsely identify a significant connection between genes 1 and 2. We simulated 100 triplets of genes with this structure where gene 3 is moderately linked to gene 1 ($\beta_{13} = 0.5$), and is strongly linked to gene 2 ($\beta_{23} = 0.8$), but gene 1 and gene 2 are independent conditional on the presence of gene 3 ($\beta_{12} = 0$). As shown in Figure 3.4a, CCM correctly estimates the true parameters. We also ran Pagel's model over the same simulation data in pairs (3 pairs for each group of 3 genes, so 300 pairwise comparisons in total). From Figure 3.4b which shows the distribution of estimated p-values on the conditionally independent linkage between gene 1 and gene 2, we can see that our method resulted in the desired uniform distribution of p-values while Pagel's method shows 76% of estimates had p-value < 0.05.

Recovery of community structures

To evaluate the ability of CCM at recovering the relationships among more genes within a community, we simulated four basic network topologies of a 5-node community: 1) a line structure; 2) a partially connected network where node 3 acts as the bridge that connects two subgroups; 3) a star structure where the node 3 acts as the hub; and 4) a fully connected network. For each structure we simulated 100 datasets with an interaction coefficient of 0.5 for all links. As shown in Figs. 3.4c-f, CCM successfully reveals the linkages within the community. Unlike the pairwise methods that will tend to result in a densely connected network due to false positives among the conditionally independent pairs, our community model provides clear insights in finding the importance of the members (e.g. hub genes), and complex dependency structures within the community.

As community size increases, the Q matrix dimension increases quickly which dramatically slows down the evaluation of the log-likelihood values. Thus our current method can only handle a small number of genes in a community. Table 3.2 provides the approximate running time of our program as a reference for different community and tree sizes. For large groups, one strategy is to run all-vs-all pairwise comparisons first to construct a gene interaction network, which is usually very densely linked at this stage. We then run all the triplets within the network to remove the conditionally

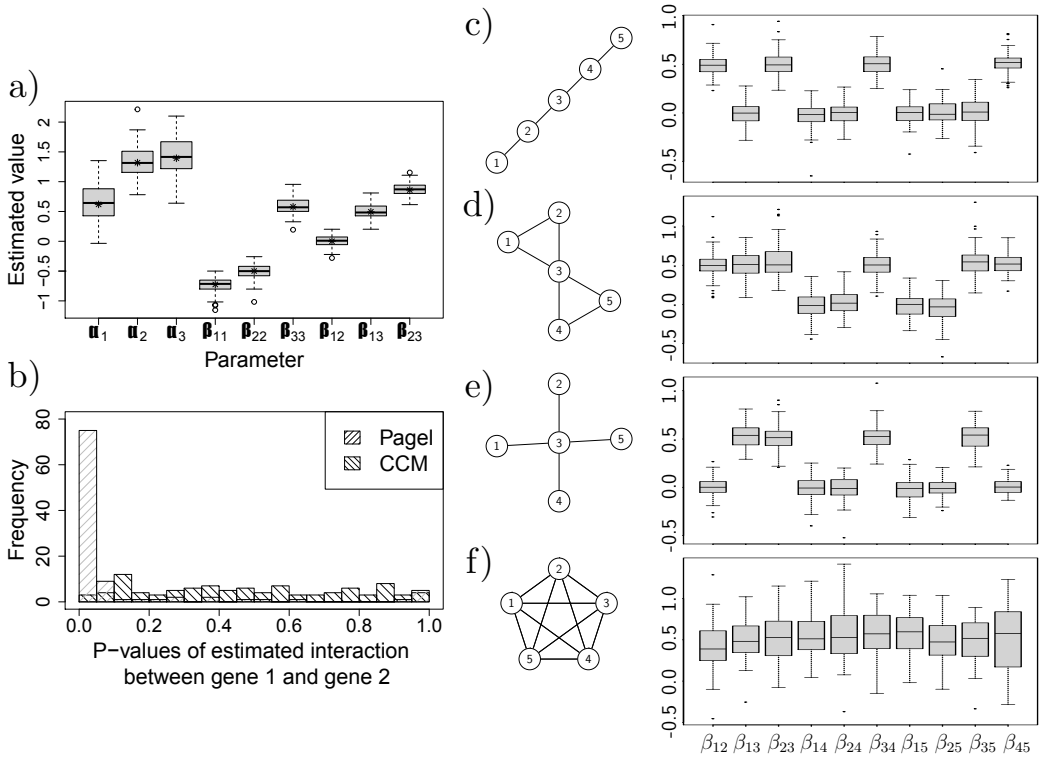


Figure 3.4: Evaluation of the conditionally independent links in the simulated triplets: (a) Estimated parameter values from CCM based on 100 simulated triplets. The sign "*" indicates the true parameters. (b) The p-values of the conditionally independent pairs (gene 1 and 2) inferred by our CCM model and Pagel's model. Simulation of four association-network structures: c) line, d) partially connected, e) star and f) fully connected. The networks on the left demonstrate the structures and the box-plots on the right show the estimated coefficients of interactions within the community. All the edges have an interaction coefficient (β_{ij}) of 0.5.

independent links. We can continue to examine all the subnetworks of size 4 or 5 to further prune the network to the desired sparsity.

Table 3.2: The approximate running time of the community coevolution model for different sizes of tree and community performed on a server running Linux with 2.67 GHz CPU and 18 GB RAM. Abbreviations: s (second), m (minute) and h (hour)

		Number of tips in tree				
		50	100	200	500	1000
Community Size	2	0.67 s	1.11 s	1.90 s	3.76 s	8.55 s
	3	2.24 s	2.90 s	5.04 s	11.22 s	22.76 s
	4	7.73 s	9.86 s	17.33 s	40.02 s	1.19 m
	5	31.04 s	52.37 s	1.60 m	4.93 m	6.60 m
	6	3.65 m	7.82 m	12.50 m	22.91 m	40.31 m
	7	15.3 m	26.43 m	37.89 m	1.25 h	2.43 h

Effect of tuning parameters

We also evaluated the influence of the tuning parameters on the MLEs. We simulated 100 gene pairs with random parameters and expected that the overshooting problem may happen to some of the pairs. These problematic cases will have the estimations of parameters far away from the true values, like the outliers in Figure 3.5. Then we added the regularization term and increased the tuning parameter λ gradually. For each simulated pair, the mean squared error (MSE $= \frac{1}{5}\sum_{i=1}^{5}(\hat{\theta}_i - \theta_i)^2$) of the estimators (5 parameters for two genes) is reported. From Figure 3.5a, we can see that the tuning parameters mainly have a large impact on those outliers. The condition number plot (Fig. 3.5b) shows that when we increase the tuning parameter to make the condition number of the Hessian matrix below 200, the overshooting problems with those outliers were effectively solved.

3.3.2 Results on Prokaryotic Data

Model comparisons

The computational cost of running a single pairwise profile comparison using Pagel's model was around 15 seconds (performed on a server running Linux with a 2.67 GHz CPU and 18 GB RAM). Since the entire dataset requires ($\binom{3786}{2} = 7,165,005$) such comparisons, an exhaustive evaluation of Pagel's method is infeasible. Instead, we

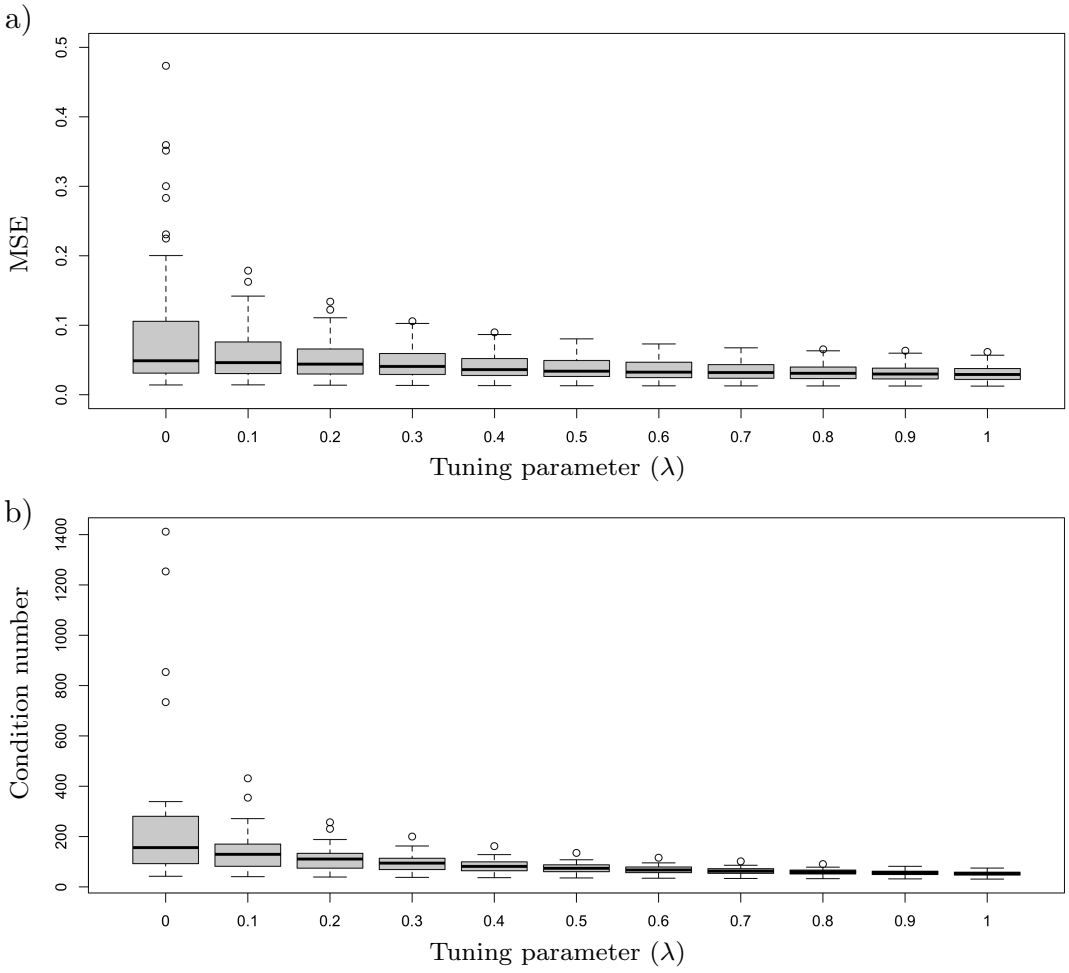


Figure 3.5: Evaluation of the effect of different tuning parameters: (a) MSE of the estimates for different tuning parameters. (b) The condition number of the Hessian matrix which also stands for the largest convergence rate between estimated parameters against different tuning parameters.

focused on a complete pairwise evaluation of 50 adjacent genes to compare against our coevolutionary model. The software we used to implement Pagel's approach is BayesTraitsV3 [59]. A comparison of the negative logarithm of the p-values inferred by the two methods yielded a correlation coefficient of 0.741 as shown in Figure 3.6a (p-values have been adjusted for multiple correlated tests using the Benjamini-Yekutieli (BY) method [9]). Applying a log p-value threshold of 6, we found that both methods agreed on the significance or non-significance of 1188 (96.98 %) comparisons. 29 (2.37 %) comparisons gave a significant result with the Pagel test but not with the coevolutionary model, while the opposite result was seen in the remaining 8 (0.65 %) pairwise comparisons. After examining the discordant pairs in the top-left corner of Figure 3.6a, we found that a common issue for Pagel's model is that most of these pairs reached the default maximum rate of 100 (the mean branch length of the tree has been scaled to 0.1 as suggested by the authors [59]), which indicates that Pagel's dependent model may overestimate the likelihood of correlated evolution because of overshooting and therefore detect more false-positive links. One example of the estimated transition rates by the two methods is compared in Table 3.3. Pagel's dependent model has a strange transition rate matrix where the transition rate from (0,0) to (0,1) is abnormally large (98.575) and the transition rate from (1,0) to (0,0) is 0, which may suggest that Pagel's eight-parameter dependent model may be over-parametrized and therefore over-estimate the likelihood of dependent evolution.

We further evaluated the goodness-of-fit of the two methods to the real data by comparing the likelihood scores. As shown in Figs. 3.6b-c, the CCM model obtains a significantly lower negative log-likelihood than Pagel's independent model (p-value $< 2.2 \times 10^{-16}$) and dependent model (p-value = 0.00217), which suggests that our model generally has better fit to the real data, even though our CCM model (5 parameters) has fewer parameters than Pagel's dependent model (8 parameters).

Gene clustering based on significant pairwise linkages

To discover sets of genes that collectively show evidence of correlated gains and losses, we performed a full pairwise comparison over the genes of LZ using CCM. There are in total 1918 genes annotated with gene ontology (GO) biological process terms, which were used to evaluate the gene functional similarities of the linkages and for

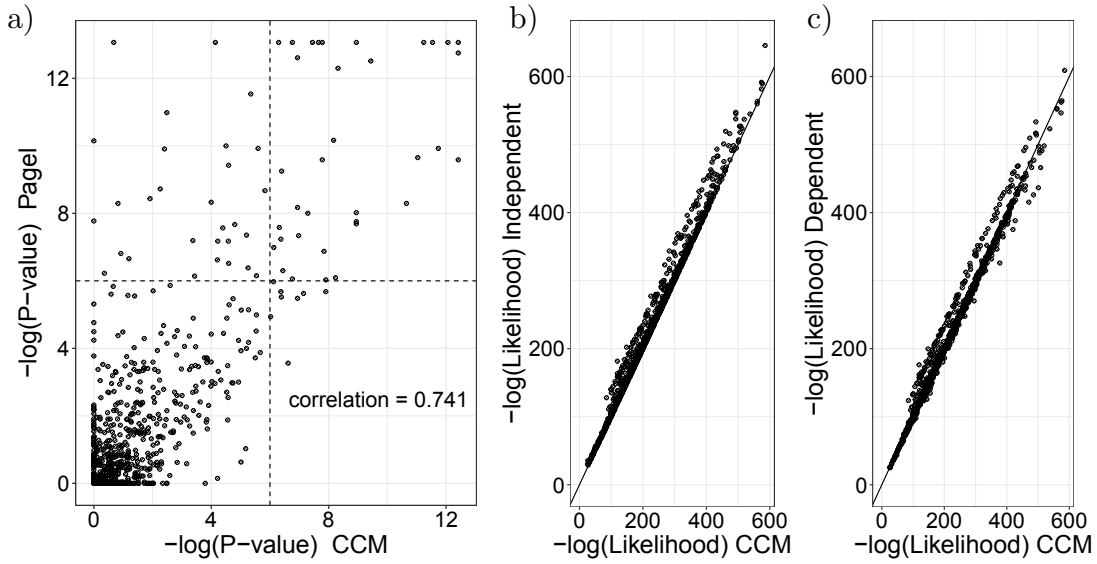


Figure 3.6: (a) The comparison of significance of pairwise linkages by two methods: the horizontal axis is the $-log_{10}$(p-value) of CCM and the vertical axis is the $-log_{10}$(p-value) of Pagel's approach; the correlation between the p-values for the two methods is 0.741. (b) - (c) The comparison of the goodness-of-fit of models to data between the independent model (4 parameters), dependent model (8 parameters) and our CCM model (5 parameters). The independent model and dependent model are the two components of Pagel's approach required for the likelihood ratio test.

Table 3.3: The transition rate matrices inferred by Independent and Dependent models of Pagel's method and the CCM. The gene pair (GI:511537597 and GI:496550319) in this table is considered strongly correlated by Pagel's method (p-value = 0.00011), but not by the CCM (interaction coefficient = 0.0832; p-value=0.283).

(a) Independent Model
-log(likelihood) = 283.183

	0, 0	0, 1	1, 0	1, 1
0, 0	–	2.143	0.506	0
0, 1	0.304	–	0	0.506
1, 0	1.866	0	–	2.143
1, 1	0	1.866	0.304	–

(b) Dependent Model
-log(likelihood) = 271.4941

	0, 0	0, 1	1, 0	1, 1
0, 0	–	98.575	0.375	0
0, 1	10.190	–	0	0.447
1, 0	0	0	–	2.139
1, 1	0	2.267	0.0003	–

(c) CCM
-log(likelihood) = 283.015

	0, 0	0, 1	1, 0	1, 1
0, 0	–	2.158	0.441	0
0, 1	0.335	–	0	0.521
1, 0	2.188	0	–	2.548
1, 1	0	1.853	0.284	–

the GO enrichment analysis. All GO annotations of genes were retrieved from the UniProt database [94]. We use Wang's graph-based method [99] to measure the semantic similarity of GO terms, which produces a score between 0 and 1 for a given pair of GO terms and higher values represent more functional similarity [106, 99]. Figure 3.7 shows that the most significant linkages under our model are between closely functionally related genes. Our results confirm the strong relationship between evolutionary similarity and functional similarity between genes.

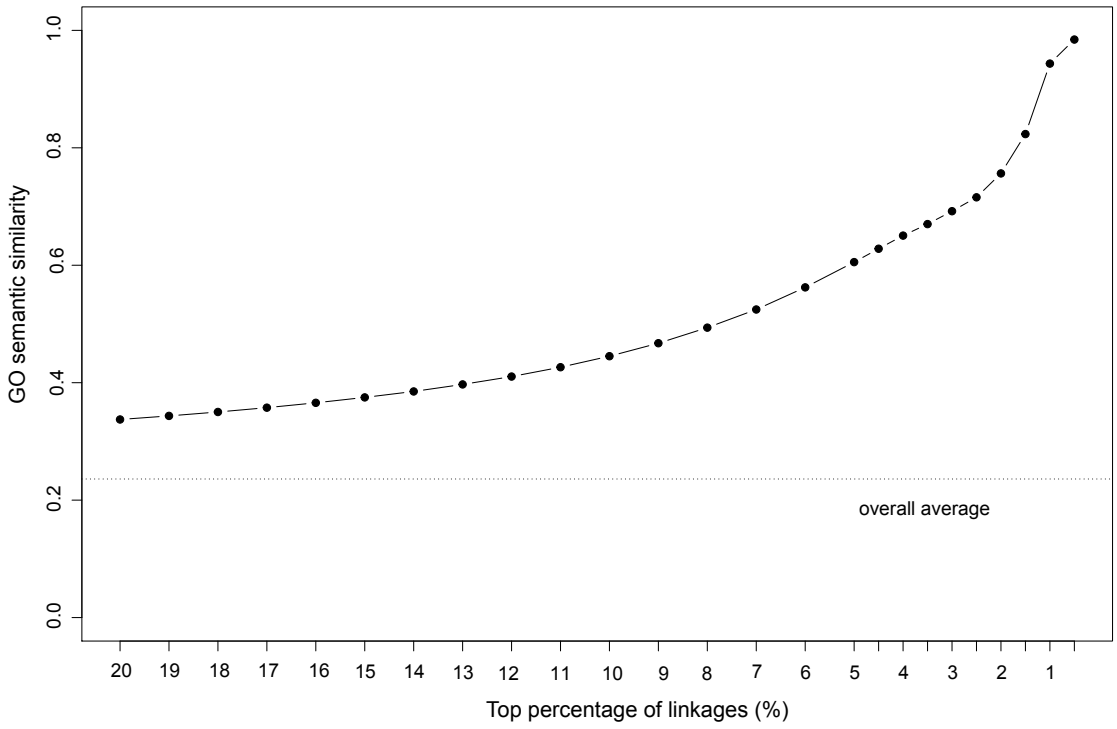


Figure 3.7: The association between functional similarity and the strength of the gene linkages detected by CCM. The horizontal axis shows the percentage of most significant linkages used for evaluation and the vertical axis shows the mean of the GO semantic similarity among the corresponding percentage of linkages. The horizontal line indicates the average functional similarity among all genes.

To obtain the clusters of genes with highly correlated evolution, we firstly applied a strict threshold (coefficient of interaction $\beta_{12} > 0.75$ and Z score > 7.5) on the linkages to obtain a gene network which consists of 1401 vertices and 19,391 highly significant (p-value$< 6.37 \times 10^{-14}$) edges (Fig. 3.8a). We further applied Markov clustering with inflation parameter 1.5 on the network to provide a guidance for

labeling the genes into clusters in the largest component (Fig. 3.8b). We reported the GO enrichment analysis for all the clusters of size at least 5 in Supplementary Table S2 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`.

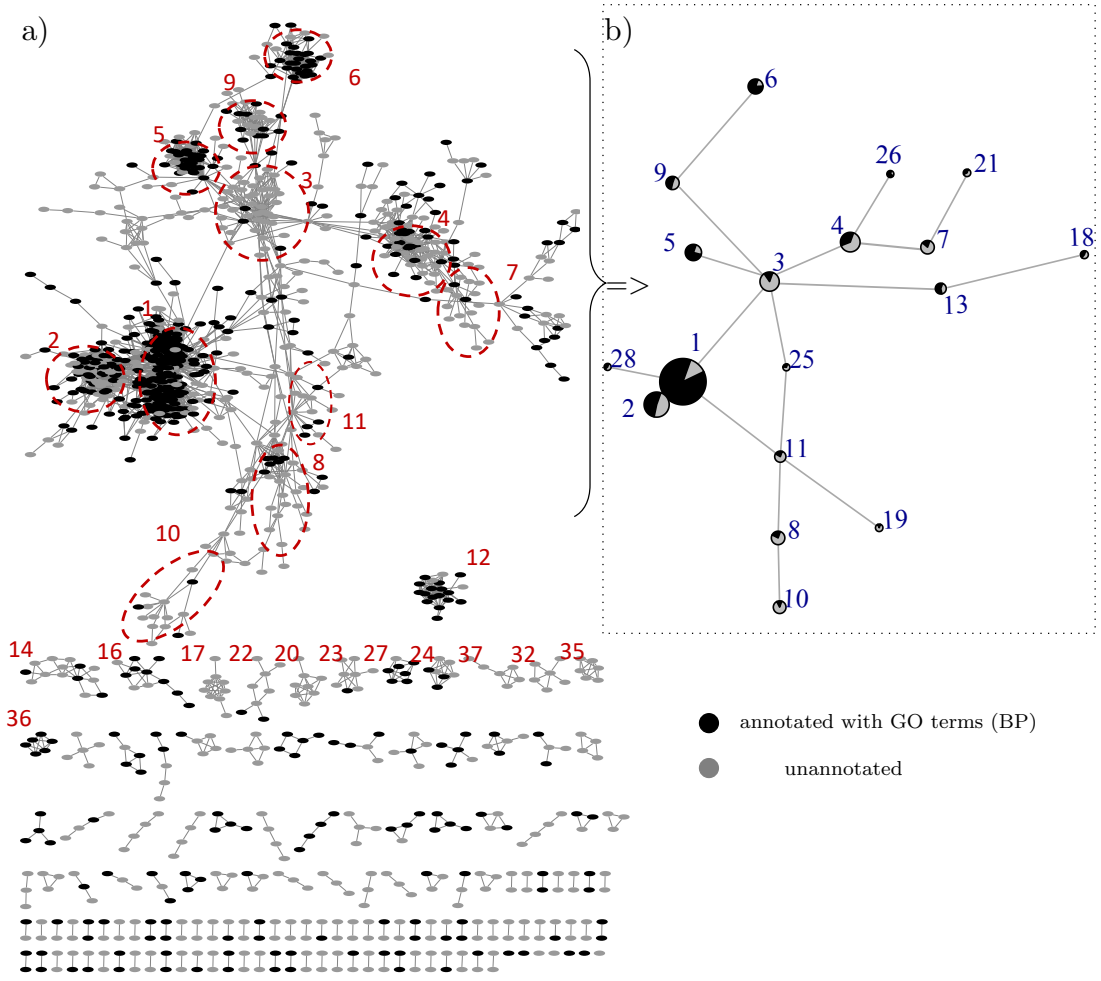


Figure 3.8: Visualization of the gene network: (a) The gene network obtained from the full pairwise comparisons and labeled with the MCL clustering results. Black vertices indicate the genes annotated with GO (BP) terms and gray vertices denote unannotated genes. (b) shows a detailed structure inside the largest component in (a). Each pie chart denotes the percentage of the annotated genes within each cluster. Only the clusters of size > 5 are labeled for a clean visualization.

We do not expect profile similarity and clustering to align perfectly with participation in a common biological process, especially when biological processes are annotated at very low levels of specificity (e.g., 'transmembrane transport'). Nonetheless,

we expect that many genes with common functions (such as transmembrane transport, transcription, and carbohydrate metabolic process) will show similar distributions across genomes, reflecting processes such as hitchhiking on frequently transferred mobile elements and coincidental loss of genes that collectively confer no selective benefit. The flagellum cluster (Cluster 5) and amino-acid biosynthetic cluster (Cluster 6) were also discovered and examined in our previous study using Pagel's correlation method applied on a reduced data set (a 74-tip subtree). It was only possible to analyze a reduced data set because of the computational cost of Pagel's method, and a phylogenetic analysis was also conducted to find potential evidence for lateral gene transfers [54]. In this study, by applying our method to the full data set (659 species), we discovered another candidate group of flagellar genes (Cluster 16) which are much less common (found in only 45 genomes) compared to the genes in Cluster 5 which are found in 396 genomes (Supplementary Table S2 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`).

The intrinsic rates inferred by CCM were consistent with distribution patterns of genes in phylogenetic profiles. For example, the pattern in Cluster 4 appears to be more consistent with Darwin's scenario, which is consistent with its relatively low intrinsic rate (Fig. 3.9). Clusters 29 and 33 have the largest estimated intrinsic rates, and both show patchy distributions in the same very shallow clade in the tree. This rapid gain and loss over a relatively short span in the tree is a possible cause of the high rates. Cluster 36 (profiles in Fig. 3.10b) and Cluster 55 have the largest estimated interaction coefficients (β) and they both show strong functional associations according to their GO annotations as well. More detailed information about clusters can be found in Supplementary Table S2 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`. To complete the analysis, we also provided a list of GO predictions on 823 unannotated genes based on most interacting genes that have known GOs and the results are summarized in Supplementary Table S3 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`.

Examples of inferred evolutionary relationships

The simulation results have shown that the pairwise comparisons could not detect the conditionally independent linkages, so that using all-vs-all pairwise comparisons tends

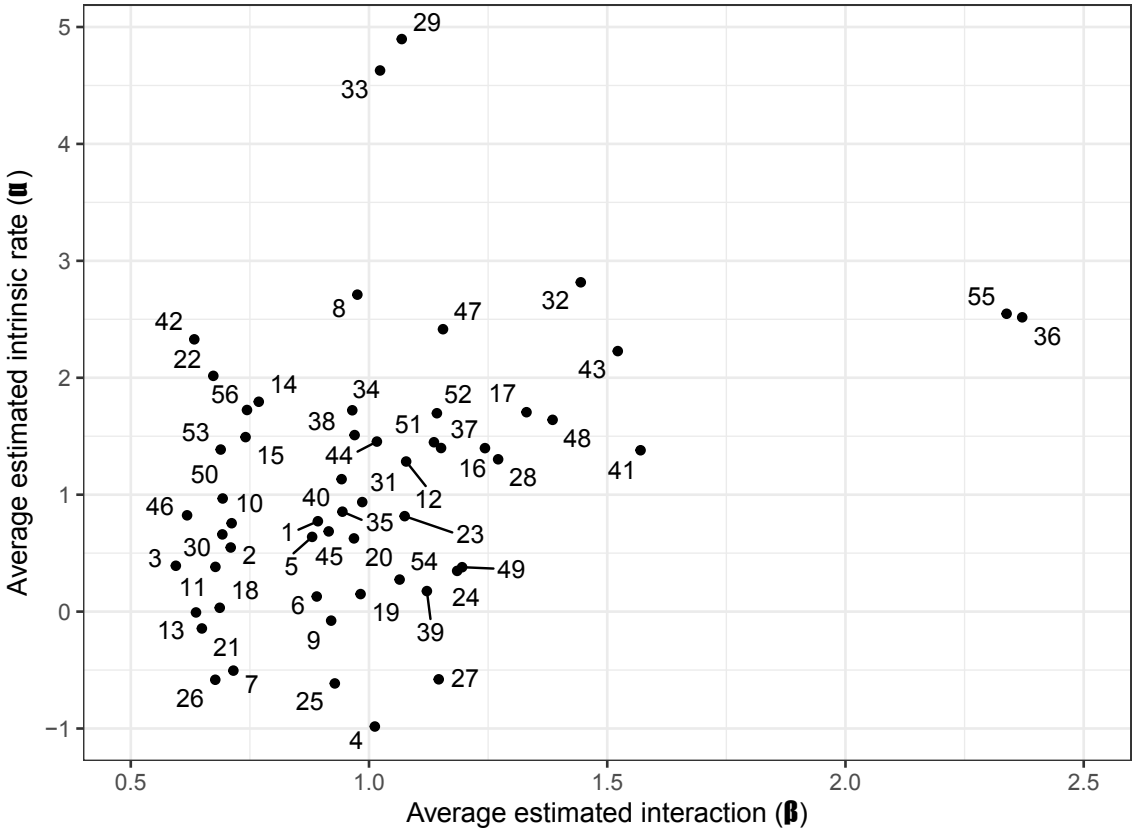


Figure 3.9: Averaged interaction and averaged intrinsic rates within each cluster. The cluster labels are matched to **Supplementary Table S2**.

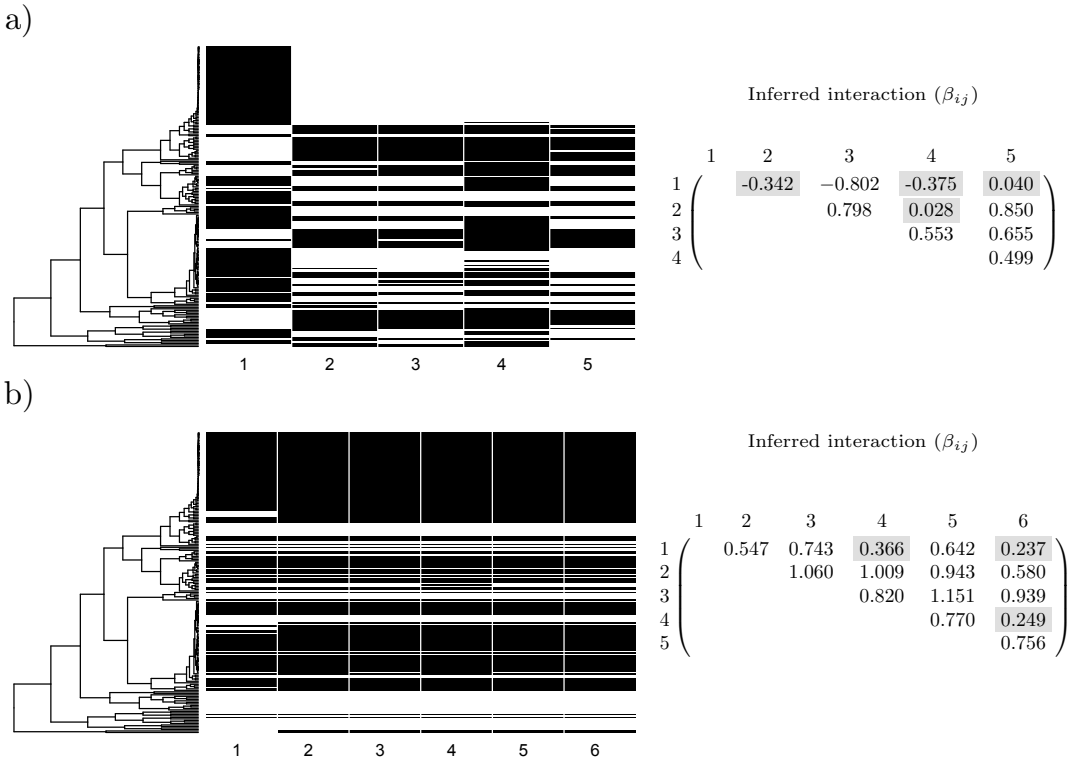


Figure 3.10: (a) The phylogenetic profiles of Cluster 49 are shown on the left. The interaction coefficients estimated by simultaneously modeling five genes as a community are shown on the right. (b) The phylogenetic profiles of Cluster 36 are shown on the left. The interaction coefficients estimated by simultaneously modeling six genes as a community are shown on the right. The gray cells indicate linkages that have p-values > 0.05.

to produce densely connected networks. For example, the five genes in Cluster 49 (Fig. 3.10a) are all related to iron-sulfur (Fe-S) assembly (three are annotated with "iron-sulfur cluster assembly", one is annotated with "cysteine metabolic process" and one has no GO annotation but has the protein name "FeS assembly ATPase SufC"). The pairwise comparisons suggest that the linkages between all five genes are extremely strong (largest p-value 3.13×10^{-9}), which would lead to a fully connected network. However, by modeling these five genes as a community, 4 out of 10 total linkages can be removed as conditionally independent linkages (p-value > 0.05).

In other cases, the pairwise interactions are still significant even when we account for conditional dependence. As an example, Cluster 36 consists of six genes which are all annotated with GO term "alginic acid biosynthetic process". The pairwise comparisons show that all links between the six genes are highly significant (largest p-value 6.43×10^{-12}). By modeling these six genes simultaneously as a community, only 3 out of 15 total linkages have a p-value > 0.05 as shown in Figure 3.10b.

Because the size of the transition matrix, and therefore the computational cost of our method, increases exponentially with the number of genes, it is infeasible to apply our method to large groups of genes. For large clusters, we get around this issue by applying our method to smaller cliques within the network, and using this to detect linkages that are conditionally independent. This is different from directly removing linkages by thresholding, as it aims to only remove the "redundant" linkages conditioning on other genes' presences to reveal the refined structure rather than to break the cluster into smaller groups. For example, we started from the original network of Cluster 6 which consists of 32 amino-acid related genes and 381 highly significant linkages (p-value$< 6.37 \times 10^{-14}$) obtained from all-vs-all pairwise comparisons (Fig. 3.11a). Then we applied CCM over all the triplets within this network and some strong linkages became weakly significant due to the presence of the third gene. We removed 272 such edges (p-value > 0.001 and interaction coefficient (β) < 0.5) and obtained the refined network (Fig. 3.11b). For comparison, we directly deleted the same number of edges from the original graph by increasing the threshold (Fig. 3.11c). This results in a very different network structure consisting of multiple densely connected components, rather than a more sparsely connected network obtained using our method.

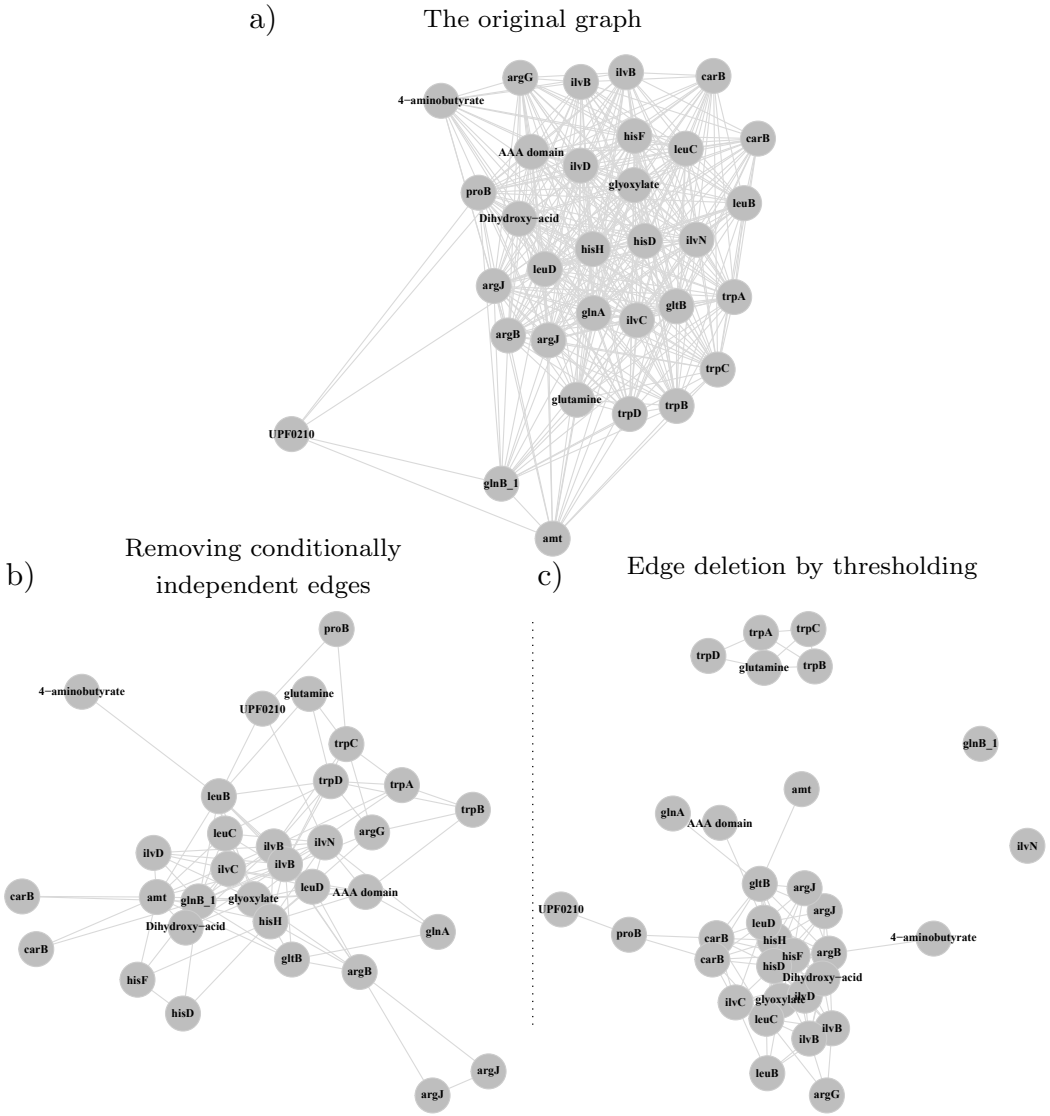


Figure 3.11: Network analysis of the amino-acid gene cluster. (a) The original network (Cluster 6) consists of 32 vertices and 381 highly significant (P-value$< 6.37 \times 10^{-14}$) edges based on the all-vs-all pairwise comparisons (b) Application of the CCM on every triplet from network (a) followed by removal of the conditionally independent edges (p-value > 0.001 and interaction coefficient $\beta < 0.5$). The resulting network consists of 32 vertices and 109 edges. (c) Direct deletion of edges from (a) by thresholding to retain the same number of edges as in (b). The cluster is disconnected into two components and two singletons. The force-directed layout algorithm is used for the network visualization.

3.3.3 Analysis of Mitochondrial Respiratory Complex 1

Eukaryotic genes are less susceptible to LGT [43, 83], and we may therefore expect significant differences in the performance of CCM between prokaryotic and eukaryotic data. To evaluate the performance of CCM on eukaryotic data, we applied our CCM method on a well-studied protein complex which consists of a total of 44 human genes encoding Mitochondrial respiratory complex 1 [51, 34, 5]. The data sets we used are published phylogenetic profiles and a species tree consisting of 138 diverse eukaryotes and a prokaryote outgroup [51, 10]. We performed an all-vs-all comparison using CCM to infer the interactions among 44 genes and illustrate the detailed relationships within the complex with the average linkage hierarchical dendrogram as shown in Fig. 3.12. We also compared our results with CLIME, an approach to infer evolutionary modules specifically for eukaryotic species which assumes that each gene must only have one single gain event in evolution followed by zero or more loss events. CLIME groups 20 of the 44 genes into four evolutionary modules (ECMs) with the remainder as singletons with no assigned group as shown in Fig. 3.12 (the results of CLIME are available at `https://gene-clime.org/`). Comparing our clustering results to the detailed structure of complex I reported by Guo et al. (2017), we find a a single cluster of size 21 encompassing 15 genes that all localize to the matrix arm of C1 including all 7 core subunits (NDUFV1, NDUFV2, NDUFS1, NDUFS2, NDUFS3, NDUFS7, and NDUFS8). The other main cluster includes 20 subunits, 15 of which localize on the membrane arm. We also analyzed the estimated evolutionary rates and find that the loss rates are significantly (p-value $< 2.2\times10^{-16}$) larger than the gain rates (Fig. 3.13), which supports the idea that eukaryotic genes are much less mobile than prokaryotic genes. To further study the structure of complex 1, we first obtained a network consisting of 462 significant (p-value< 0.05) links that were inferred by full pairwise comparisons using CCM. After pruning the network by removing the conditionally independent links (p-value > 0.05) detected from all triplets, we obtained a more sparse network consisting of 101 linkages (Fig. 3.14). We can observe two loosely connected components in this network: one is mainly composed of more densely linked subunits on the matrix arm with higher estimated values for the coefficients of interaction, while the other component is mainly composed of the subunits on the membrane arm. This network representation of the gene-interaction map shows more

comprehensive information about the gene evolutionary cohesiveness than the pure clustering results in Fig. 3.12.

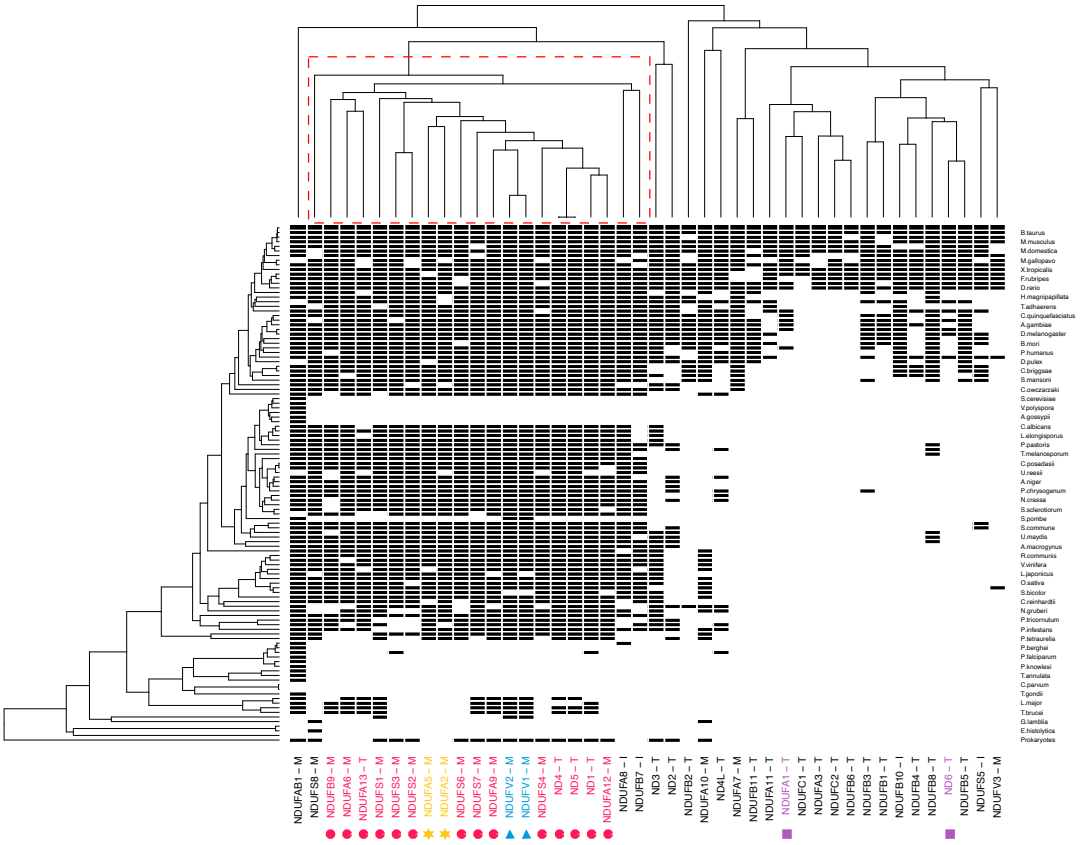


Figure 3.12: Clustering of mitochondrial respiratory complex 1 genes: the heatmap shows the phylogenetic profiles of 44 genes where black bars indicate presence. The column labels give the information of subunits, names - location (M: Matrix, T: Transmembrane, I: Intermembrane). The symbols below the gene names indicate the four components inferred from CLIME and those without symbols below indicate singletons. The dendrogram on the left indicates the eukaryotic tree and the names of species are given on the right as the row labels; the dendrogram above shows the hierarchical structure constructed with the estimated pairwise interactions by CCM.

3.4 Discussion

Identifying associations among traits is an important tool to generate hypotheses about linkages between phenotypic, ecological, and genetic attributes. Often these associations need to be further analyzed or tested experimentally to demonstrate

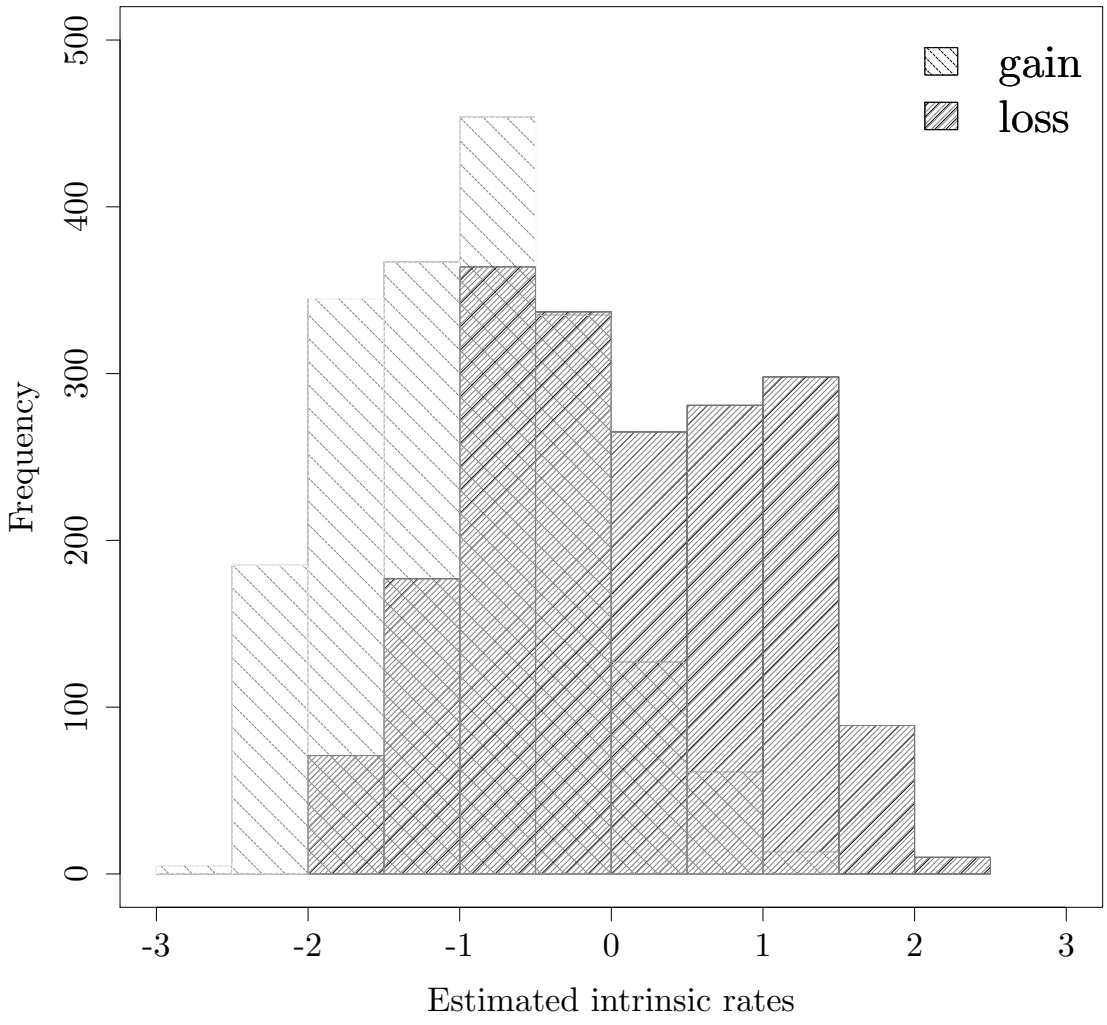


Figure 3.13: The estimated intrinsic gain and loss rates of the complex I genes.

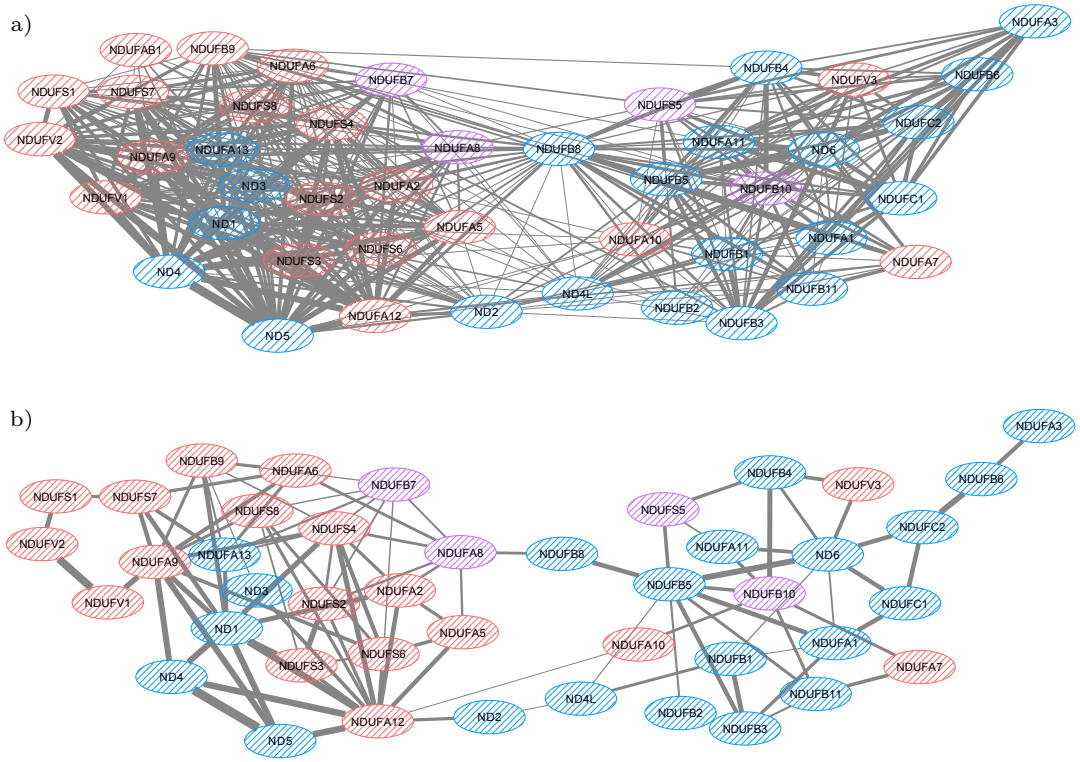


Figure 3.14: Network analysis of the complex I genes. (a) The original network inferred by full pairwise comparisons using CCM, which consists of 462 significant edges (p-value < 0.05). (b) The links that are significantly conditionally dependent (p-value < 0.05) in all triplets from network (a). The resulting network consists of 101 edges. The edge thickness corresponds to the estimated strength of the interaction (βij). Label colors indicate the locations of the subunits: Matrix (red), Transmembrane (blue), and Intermembrane (purple).

whether they have arisen due to selection or other factors [70]. Phylogenetic profiles are a specialized type of trait representation that have been used for over 20 years as a tool to explore and compare genomes; while they can be treated in a similar fashion to other types of traits, the sequences, genetic linkage information, and functional annotations associated with genes in a profile can be used to shed more light on evolutionary hypotheses. Many studies suggest that phylogenetic relationships among source genomes should be taken into account [54, 16, 15, 66]. Chapter 2 demonstrated the utility of Pagel's model [66] in identifying sets of genes with correlated evolutionary trajectories; however, this approach was computationally expensive and could not infer the direction of the relationship. In this chapter, we proposed a new co-evolution model, CCM, to detect genes with correlated evolutionary histories based on phylogenetic profiles. CCM was able to identify correlated genes as well as the direction of the relationship (e.g. Fig. 3.10a) and ran five times faster than Pagel's method when tested on phylogenetic trees with 500 tips. The number of pairwise comparisons increases quadratically with the number of genes to be considered, but the independence of each comparison allows calculations to proceed in parallel. Heuristic methods can be used to quickly subdivide genes into large clusters that can then be refined using the CCM. Our model also has the ability to analyze the evolutionary relationships among sets of genes of size greater than 2. Examining sets of size > 2 can provide a more sparse gene network and greater insights into the complex relationships between genes.

Based on CCM, we also developed a simulation procedure that can generate a set of co-evolved profiles with interactions along a given phylogenetic tree. The strength of the interactions during evolution is also adjustable. A common way to evaluate comparative methods for detecting genes with correlated evolutionary histories is measuring the functional similarities based on gene annotations such as GO terms [78] and KEGG pathways [42]. However, such evaluation is subject to annotation completeness and the correlated patterns may not always reflect shared function as expressed by GO annotations. Our co-evolving simulation procedure provides a way to generate benchmark data for evaluating the comparative methods.

In the simulation study, our method outperformed the non-phylogenetic method (Jaccard Index) and the tree-aware methods (Pagel's correlation model, run-adjusted

methods, and clade-adjusted methods) in detecting the significant links (Fig. 3.2d). We showed that our method can distinguish between Darwin's scenario and the replicated co-occurrence scenario (Fig. 3.3). We also demonstrated that pairwise comparisons can not detect conditionally independent links and further showed the performance of CCM in recovering the community structures (Fig. 3.4).

Finally, we applied our method to 3786 profiles across 659 genomes and the results showed a strong positive relationship between the evolutionary similarity and functional similarity (Fig. 3.9). We also identified the gene clusters with enriched functions (Supplementary Table S2 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`) that can be used to better understand the functional roles of gene groups and predicted 823 unannotated genes based on their most interacting genes with known GO annotations (Supplementary Table S3 available at `https://doi.org/10.5061/dryad.p8cz8w9rd`). We also demonstrated using CCM to refine the network obtained from the pairwise comparisons by removing conditionally independent linkages (Fig. 3.11). In addition to analyzing prokaryotic data, CCM has also been successfully appplied to a eukaryotic data set of the well-studied Human Complex I and the recovered associations mapped well onto the structural associations that exist in the complex (Figs. 3.12, 3.14). The results show that CCM as a general comparative model can also be applied to eukaryotic data. Although our method is specifically used to analyze the phylogenetic profiles in this study, we think it can have wide applications in other fields such as to study phenotypes of species [31], ecological habitats [26], and metagenomic profiling [1].

The uniqueness of the community coevolution model lies in the careful modeling of each gene's instantaneous gain and loss rates dependent on the current states of other genes. In addition to improving our ability to identify related genes, the CCM directly models the dependence between related genes in the evolutionary process. The same idea can possibly be generalized to phylogenetic models to jointly estimate the transition rate matrix of each site based on the current states of its neighbor sites or other related sites. The dependence between different genes, or different sites within a single gene, is an underexplored area in phylogeny and molecular evolution, with the majority of models assuming independence of sites. By developing better-fitting models that incorporate the dependence between different genes, we expect to

gain insights into the mechanisms driving this dependence.

We also met a challenge in extending our method to directly model larger communities. The state space **S** will increase exponentially as we include more genes into the community. Currently, we have successfully tested our method on communities of sizes less than ten, but two problems will arise if we include more genes: the huge memory requirements to store the Q matrix of dimension $2^n \times 2^n$ and the long computation time for eigendecomposition of Q. We have found that if we reorder the rows and columns of the transition matrix, there exists a recursive structure: the Q matrix can be written as a block matrix of the form $Q = \begin{pmatrix} A & B \\ B & A \end{pmatrix}$, where B is an anti-diagonal matrix and A has the same recursive structure as Q, $A = \begin{pmatrix} A' & B' \\ B' & A' \end{pmatrix}$ (B' is still an anti-diagonal matrix and A' is a block matrix). We can solve the first problem by storing the Q matrix as a sequence of small "blocks", but we have not found existing mathematical methods to solve the eigendecomposition of block matrices with such recursive structures. Our future work will explore the possible solutions to decompose the Q matrix more efficiently so that the CCM method is scalable.

3.5 Author Contribution

In this study, I participated in the design of the work, implemented the methods, conducted the analysis and wrote the manuscript.

3.6 Software Availability

The R package evolCCM was written in R v4.0.2 and is available on Github (`https://github.com/beiko-lab/evolCCM`).

Chapter 4

Assessing the Dependency of Phylogenetic Profiles By Conditioning on a Phylogenetic Tree

4.1 Introduction

Genes that show correlated patterns of gain or loss during their evolutionary process can provide great insights in genomic analysis as they are more likely to have similar functions or be involved in identical or related pathways [12, 17, 27]. Phylogenetic profiles, which summarize the presence / absence of genes across a set of genomes, are a commonly used method to study patterns of evolutionary relationships among genes.

Many standard metrics, such as Hamming distance, Jaccard Index, Pearson's correlation, and Hypergeometric test, have been used to measure the (dis)similarities between a pair of phylogenetic profiles [72, 39, 81, 101]. However, such metrics do not consider phylogenetic correlations among species [30, 54]. This *phylogenetic effect* refers to the tendency that the closely related genomes are more likely to share similar gene content due to recent common ancestry [76, 57].

Some heuristic methods have been developed based on standard metrics to utilize empirical techniques to remove the phylogenetic effect [45, 81, 97, 16]. The heuristic methods are fast and scalable to large data sets, but lack biological interpretations. In contrast, evolutionary model-based methods build probabilistic models on phylogenetic trees to describe the evolutionary process [15, 51, 66], and aim to incorporate tree information, including divergence events represented by the tree topology and the evolutionary dissimilarity represented by the branch lengths. These methods provide additional insights but suffer from computational issues due to the complexity of the underlying models. Pagel's correlation test [66] and the Community Coevolution Model (CCM) [53] directly model the interactions between genes and provide the best performance in detecting correlated genes as demonstrated in our previous simulation

and empirical work [54, 53]. However, both approaches rely on complete all-against-all comparisons between all profiles, which makes them impractical for analyzing large datasets. In addition, except for phylogenetic-naïve methods that do not explicitly use any tree information, the phylogenetic comparative methods rely on the given phylogenetic trees to remove the phylogenetic effect, either using only a part of the tree information (e.g. reduced trees, order of species) like heuristic methods or all the tree information (e.g. tree topology, branch lengths). Thus a potential problem with the phylogenetic comparative methods is the phylogenetic error or uncertainty involved in data sampling, choice of substitution models, and incompletely resolved clades [37, 79, 67, 52].

In this chapter, we propose a matrix-decomposition-based method to test the dependency between binary profiles, conditioned on the tree topology. It is computationally efficient for large-scale analyses, gives support to better biological explanations to the data than heuristic methods and also works with or without a provided phylogenetic tree, which makes our method robust to phylogenetic uncertainties. Stemming from similar ideas to phylogenetic inertia estimation methods [13, 21, 11, 20], our approach considers the phylogenetic profile to contain two information components: one is the underlying phyletic pattern (P) driven by the phylogeny among species as closely related genomes that inherit from close common ancestors will tend to share similar gene content and thus a given gene is more likely to be found in closely related genomes (row-wise); the other component is unique information about individual genes (S) caused by their own gain/loss events during evolution (column-wise). Then we can test the dependency between a pair of profiles by conditioning on their predicted underlying phyletic pattern (P) such that the genes are considered related only when their individual components (S) show correlated patterns. We apply this new method on both simulated data from CCM [53] and real data sets to evaluate its ability to correctly discover the correlated genes after correction for phylogeny.

4.2 Methods

Our method consists of three major steps. First, the phylogenetic components are inferred from the given phylogenetic tree. The phylogenetic eigenvector regression (PVR) approach proposed by Diniz-Filho et al. (1998) is commonly used to quantify

the phylogenetic components by applying principal coordinates analysis (PCoA) to a phylogenetic distance matrix and has shown better estimates of phylogenetic intertia than autoregressive methods [21, 13, 20]. We also propose another option to extract phylogenetic components directly from the profiles, which makes our method robust to the errors in estimating the phylogenetic trees. Second, the inferred phylogenetic components are used to predict the expected presence/absence of each gene. Finally, a dependency test based on a modified Pearson-Chisq statistics is implemented between any pair of profiles conditioning on the predicted presences/absences. The method of using the phylogenetic eigenvectors inferred from a reference tree is termed Chisq-PyLR (modified Chi-square test by phylogeny-based logistic regression), while the method of using the eigenvectors directly inferred from profiles is termed Chisq-PrLR (modified Chi-square test by profile-based logistic regression). We will use Chisq-PLR to refer to both methods.

4.2.1 Inferring the phylogenetic vectors from a reference tree or from phylogenetic profiles

A phylogenetic tree describes the evolutionary relationships among a set of species or individuals, where the tips of the tree represent the entities of interest and the branch lengths indicate the extent of genetic divergence. Given a phylogenetic tree $\mathcal{T}$ with m tips, we first calculate the patristic distance matrix $D_{m\times m}$ between every pair of tips using the branch lengths. Then principal coordinates analysis (PCoA) is applied to obtain phylogenetic eigenvectors. PCoA, also known as classical multidimensional scaling (MDS) is a method used to map a distance matrix into a lower k-dimensional Euclidean space so that the distances can be preserved as well as possible [107, 36]. The standard procedure of PCoA can be summarized as

1. Double-centering the distance matrix squared: $B_{m\times m} = -\frac{1}{2}HD^2_{m\times m}H$, where D^2 is the distance matrix squared and $H = I_m - \frac{1}{m}JJ^T$ with I being an $m \times m$ identity matrix and J being an $m \times 1$ vector of all ones.

2. Taking the eigendecomposition of the double-centered matrix $B_{m\times m}$ to acquire the k largest eigenvalues and corresponding eigenvectors $E = \{e_1, e_2, ..., e_k\}$.

Thus, the eigenvectors $E = \{e_1, e_2, ..., e_k\}$ which contain the coordinate information of species in the k-dimensional space preserve most of the phylogenetic relationships among species, and the largest eigenvalues together with their corresponding eigenvectors can be used to recover the major structures of the clades in the tree.

To determine the value of k, a common approach is to use a broken-stick model which assumes that the total variance is randomly divided into pieces and then finds the k components that exceed the expected proportions through the broken-stick distribution [21, 41].

Inferring the phylogenetic vectors without an explicit reference tree

If no reference tree is given, we use an alternative approach that infers the phylogenetic vectors from the profiles directly. Given the binary matrix $Y_{m\times n}$ which consists of the phylogenetic profiles of n genes across m genomes, to infer the phylogenetic eigenvectors, we can apply principal component analysis (PCA) to decompose the profile matrix $Y_{m\times n}$ into a set of k largest eigenvalues $\{\alpha_1, \alpha_2, ..., \alpha_k\}$ and corresponding left eigenvectors $E = \{e_1, e_2, ..., e_k\}$.

This approach provides another option to examine the phylogenetic profiles when the evolutionary tree is not available. However, it requires a sufficient number of profiles to correctly reflect the phylogeny and the results could be impacted by large number of lateral gene transfer events in the data. We suggest choosing a smaller k compared to inferring the phylogenetic eigenvectors from the tree as the first few principal components representing the major tree structures are less likely to be affected by the abundance of unique phyletic patterns caused by lateral gene transfer events.

4.2.2 Predicting the conditional probabilities of gene presence using logistic regression

After extracting the phylogenetic eigenvectors, PVR [21] applies a multiple linear regression of Y (traits) on these eigenvectors (X) to parse out the phylogenetic component from the data. Then the fitted values $\hat{Y}$ represent the phylogenetic component P and the residuals $\epsilon = Y - \hat{Y}$ represent the specific component S. Although the idea of PVR was originally used to analyze continuous traits, it can be extended to

the binary case of phylogenetic profiles using logistic regression,

$$logit(\mathbf{p}_i) = \log\left(\frac{\mathbf{p}_i}{1-\mathbf{p}_i}\right) = X\beta_i,$$

$$\mathbf{p}_i = Pr(Y_i = 1|X)$$

where $\mathbf{p}_i = \{p_{is}, s = 1, \ldots, m\}$ is a vector of size m containing the element p_{is} as the probability of the ith gene present in the sth genome, the predictors X are the inferred phylogenetic eigenvectors and the response variable Y_i $(i = 1, \ldots, n)$ is the phylogenetic profile of the ith gene.

Since the phylogenetic profiles may contain various amounts of phylogenetic information depending on the number of genes present in the profile and the location of the presences across the tree, it is not necessary that all the k significant eigenvectors chosen by the broken-stick model are used. For example, a perfect prediction of the profile's presense/absence could occur by the logistic regression model based on a smaller number of eigenvectors for profiles of simple patterns (e.g., all presences are concentrated in one branch), in which case, it becomes redundant to use more eigenvectors than needed in the logistic regression model. A criterion for model selection such as AIC, BIC and adjusted R^2 can be used to avoid over-fitting.

4.2.3 Testing the dependency between a pair of profiles

Pearson's Chi-square test is a non-parametric statistical test that is commonly used for testing for independence between categorical variables by evaluating how significantly the observed frequency distribution differs from a distribution in which the variables are independent. Given the profiles of two genes i and j, the contingency table between two profiles summarizes the observed presence/absence distribution across m species as below:

		gene j		
		0	1	total
gene i	0	O_{00}	O_{01}	$m_{0\cdot}$
	1	O_{10}	O_{11}	$m_{1\cdot}$
	total	$m_{\cdot 0}$	$m_{\cdot 1}$	m

Pearson's Chi-square test uses the marginal frequencies to calculate the expected frequencies under the null hypothesis of the independence of two variables:

	0	1
0	$E_{00} = \frac{m_{0\cdot} \times m_{\cdot 0}}{m}$	$E_{01} = \frac{m_{0\cdot} \times m_{\cdot 1}}{m}$
1	$E_{10} = \frac{m_{1\cdot} \times m_{\cdot 0}}{m}$	$E_{11} = \frac{m_{1\cdot} \times m_{\cdot 1}}{m}$

Then the Chi-square statistic is calculated as

$$X^2 = \sum_{h,k \in \{0,1\}} \frac{(O_{hk} - E_{hk})^2}{E_{hk}},$$

which follows a χ^2 distribution with degree of freedom $df = (2-1) \times (2-1) = 1$. The problem with directly applying this generic Chi-square test in our phylogenetic profile data is the assumption that the observations are independent is not satisfied, because the species are correlated due to the underlying phylogeny.

Conditioning on the tree $\mathcal{T}$, each entry of the phylogenetic profile of gene i can be considered as an independent Bernoulli trial with probability $p_{is}|\mathcal{T}$, so the total count of presences follows a Poisson distribution with mean $\sum_{s=1}^{m} p_{is}$, which leads to the approximate normal $N(\sum_{s=1}^{m} p_{is}, \sum_{s=1}^{m} p_{is})$ based on the General Central Limit theorem. From the previous logistic regression, the fitted values $\hat{\mathbf{p}}_\mathbf{i}$ and $(1 - \hat{\mathbf{p}}_\mathbf{i})$ denote the predicted probabilities of presence and absence respectively of gene i across m genomes conditioning on the tree $\mathcal{T}$. For a pair of genes, gene i and gene j, the observed frequencies $\{O_{00}, O_{01}, O_{10}, O_{11}\}$ remain the usual counts of four states between two genes across all genomes, but the expected frequencies can be calculated by multiplying the predicted probabilities of the corresponding states under the null hypothesis of independence and then summing the probabilities for all genomes:

	0	1
0	$E_{00} = (\mathbf{1} - \hat{\mathbf{p}}_i)'(\mathbf{1} - \hat{\mathbf{p}}_j)$	$E_{01} = (\mathbf{1} - \hat{\mathbf{p}}_i)'\hat{\mathbf{p}}_j$
1	$E_{10} = \hat{\mathbf{p}}_i'(\mathbf{1} - \hat{\mathbf{p}}_j)$	$E_{11} = \hat{\mathbf{p}}_i'\hat{\mathbf{p}}_j$

The statistic $X^2 = \sum_{h,k \in \{0,1\}} \frac{(O_{hk} - E_{hk})^2}{E_{hk}}$, still follows a Chi-square distribution χ^2 with 1 degree of freedom. This modified Chi-square statistic can be used to test the independence of two genes conditioning on the phylogeny.

4.3 Results

4.3.1 Simulation Results

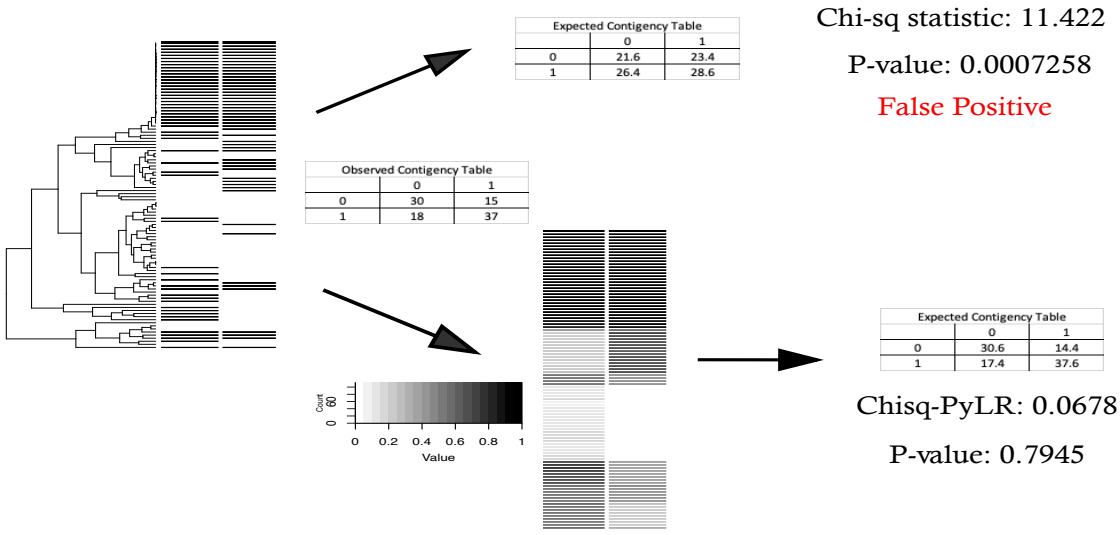
In this simulation study, we used the CCM model [53] to generate phylogenetic profiles to assess the performance of Chisq-PLR methods in correcting for phylogenetic effects on two criteria: whether it reduces the false positive rate and whether it has greater power in comparison with the phylogeny-naïve Pearson's Chi-square test. We also evaluated the impact of potential errors in the given tree on the performance of the method by adding random SPR (Subtree pruning and rearrangement) operations [86].

An illustration of the methods

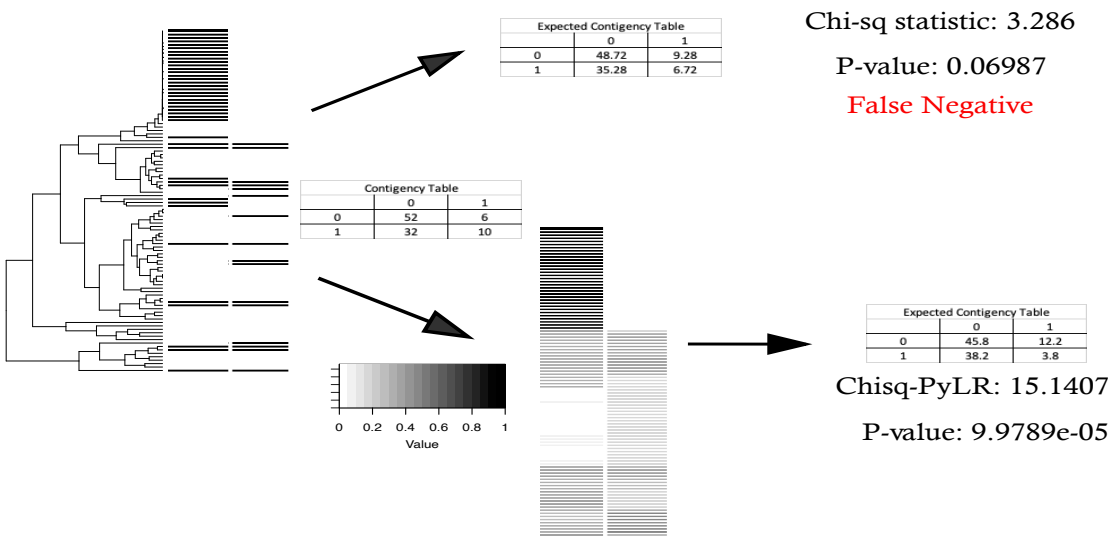
We first generated two separate pairs of phylogenetic profiles to illustrate how the method dealt with the phylogenetic relationships among species. In Fig.4.1a, a pair of independent genes were simulated and the generic Chi-square test wrongly detected it as significant (P-value = 0.0007258) due to the co-occurrences in the closely related genomes at the top region of the tree. Fig.4.1b gives a pair of correlated profiles simulated by CCM with an interaction of strength 0.8, which show highly correlated phyletic patterns across the tree except for the upper region. The Pearson's Chi-square test incorrectly classified this correlated pair (Fig.4.1b) as non-significant ($\alpha = 0.05$) while our method correctly detected it to be strongly significant (P-value = 9.9789×10^{-5}).

Evaluating the Type I error

We further simulated 1000 independent gene pairs and applied Pearson's Chi-square test, Chisq-PyLR, and Chisq-PrLR on the data set to evaluate type I errors. Table 4.1 shows that the generic Chi-square test without correcting for phylogeny has a much higher Type I error than our methods, and Chisq-PyLR and Chisq-PrLR perform similarly. It is noted that with more phylogenetic eigenvectors included, the profile patterns can be predicted better by these eigenvectors, leading to fewer significant pairs and smaller type I errors. This tendency is also shown in the next assessment of statistical power.



(a) A pair of independent genes



(b) A pair of dependent genes

Figure 4.1: An illustration of the method using two simulated pairs. a) An example of simulated independent pairs. b) An example of simulated highly correlated gene pair (interaction coefficient of 0.8 in the CCM model).

Table 4.1: Type I error evaluation using 1000 simulated independent pairs. The number of eigenvectors $k = 2$ (rows marked in red) has the lowest average BIC of all pairs for both methods.

	P-value cutoffs					
	0.001	0.005	0.01	0.05	0.1	0.2
Pearson's Chisq	0.164	0.255	0.297	0.45	0.536	0.623
Chisq-PyLR k = 1	0.062	0.127	0.165	0.315	0.397	0.5
Chisq-PyLR k = 2	0.002	0.007	0.023	0.084	0.16	0.298
Chisq-PyLR k = 3	0.002	0.007	0.018	0.079	0.148	0.278
Chisq-PyLR k = 4	0.001	0.006	0.022	0.075	0.137	0.247
Chisq-PrLR k = 1	0.009	0.03	0.043	0.118	0.191	0.314
Chisq-PrLR k = 2	0.002	0.013	0.025	0.08	0.148	0.287
Chisq-PrLR k = 3	0	0.003	0.013	0.054	0.11	0.238
Chisq-PrLR k = 4	0	0.003	0.007	0.04	0.086	0.193

Evaluating the statistical power

We further simulated 200 gene pairs for each value of the strength of interaction ranging from 0 to 1.5 with a step size of 0.1 respectively (3200 pairs in total) to assess the performance of methods in detecting true positives. We again used CCM to provide a benchmark of ideal results as a comparison. Table 4.2 summarizes the true positive rates of different methods at different P-value cutoffs. Both Chisq-PLR methods (at $k = 2$) have power above 0.8 at P-value cutoff 0.05. The generic Chi-square test has much higher true positive rates than even the CCM (the estimated theoretical optimal method), which is not surprising because it tends to overestimate the correlation signal between profiles as shown in Table 4.1.

We also examined the sensitivity of Chisq-PLR methods for interactions of different strengths at significance level $\alpha = 0.1$. As shown in Fig.4.2, the power of the test increases as the interactions between genes become stronger, with both Chisq-PLR methods achieving power above 0.8 at an interaction strength of 0.6.

Evaluating the effect of tree errors

Unlike other tree-aware methods, the Chisq-PrLR method can directly infer the phylogenetic eigenvectors from the observed profiles so it is immune to possible errors in

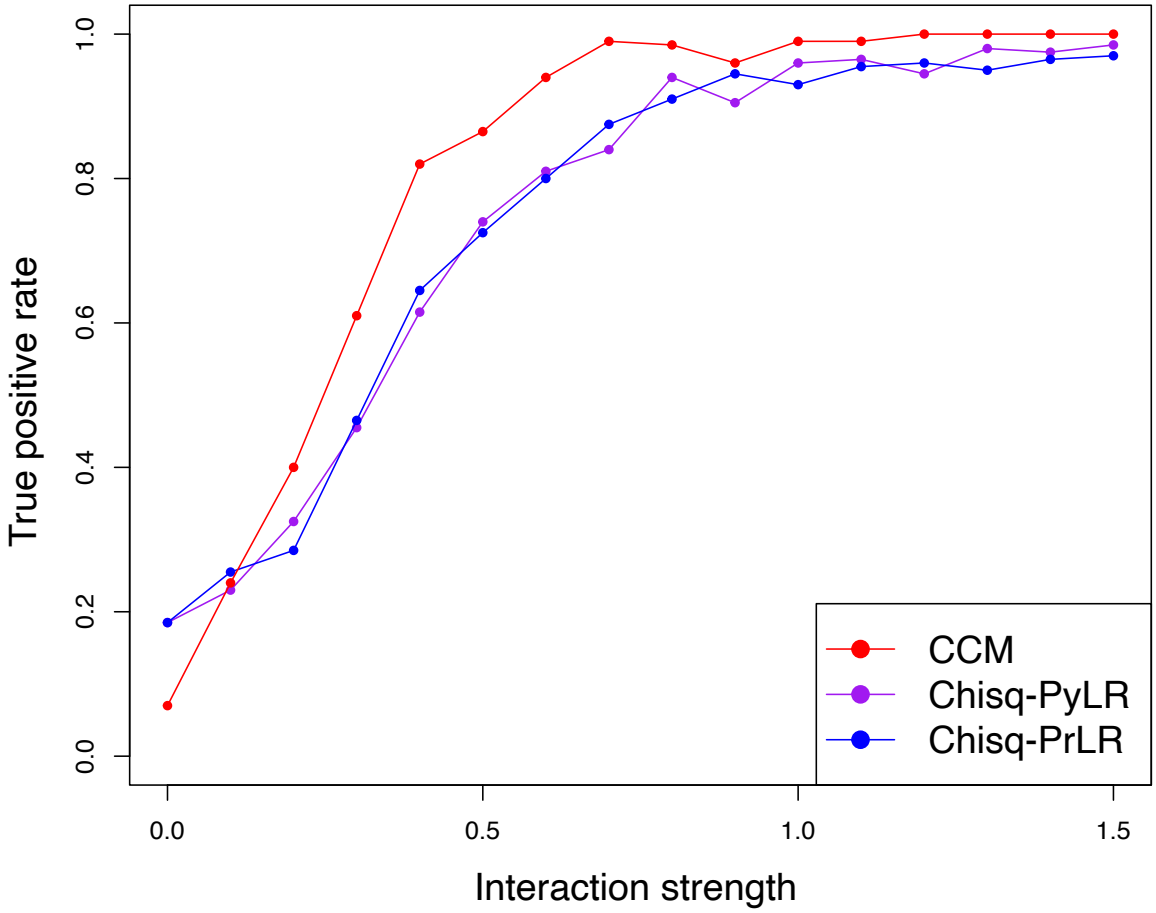


Figure 4.2: Evaluation of the power of Chisq-PLR methods for different interaction strength at significance level $\alpha = 0.1$. The interaction strength of 0 indicates independent gene pairs (true positive rate of 0). The colors of lines indicate three methods: CCM (red), Chisq-PyLR (purple) and Chisq-PrLR (blue).

Table 4.2: Power analysis using simulated data calculated from 3200 pairs of co-evolved gene pairs with different interaction strengths. The number of eigenvectors $k = 2$ (rows marked in red) has the lowest average BIC of all pairs for both methods.

	P-value cutoffs					
	0.001	0.005	0.01	0.05	0.1	0.2
Pearson's Chisq	0.818	0.848	0.863	0.91	0.929	0.945
Chisq-PyLR k = 1	0.678	0.76	0.793	0.885	0.915	0.944
Chisq-PyLR k = 2	0.56	0.675	0.724	0.856	0.89	0.935
Chisq-PyLR k = 3	0.444	0.576	0.636	0.796	0.845	0.904
Chisq-PyLR k = 4	0.307	0.447	0.521	0.701	0.784	0.86
Chisq-PrLR k = 1	0.702	0.808	0.846	0.923	0.947	0.965
Chisq-PrLR k = 2	0.513	0.649	0.699	0.833	0.884	0.936
Chisq-PrLR k = 3	0.395	0.537	0.601	0.768	0.851	0.913
Chisq-PrLR k = 4	0.271	0.414	0.492	0.706	0.795	0.883
CCM	0.608	0.783	0.852	0.959	0.984	0.997

the given tree. In this simulation, we first simulated 100 independent pairs using a tree, then introduced errors by randomly adding SPR operations to that tree, and lastly applied Chisq-PyLR with this altered tree. We examined the impact of different numbers of SPR operations (from 1 to 10) on the false positive rates at significance level $\alpha = 0.1$ and repeated the simulation 20 times for each number of operations, since the location where SPR occurs could have different effects. As shown in Figure 4.3, as more SPRs were introduced into the underlying tree, the false positive rates increased and finally reached a level similar to that of the generic Chi-square test.

4.3.2 Results on the Clostridia data set

Data Sets

We first applied our method to the previously studied data set of the bacterium "Lachnospiraceae bacterium 3-1-57FAA-CT1" (abbreviated as LZ) [54]. 658 completed and draft genomes from class Clostridia were retrieved from the National Center for Biotechnology Information (NCBI) for the comparative analysis of LZ. The phylogenetic profiles were constructed by comparing the complete set of LZ (6505 predicted genes) against all other genomes. Using the same filtering conditions as

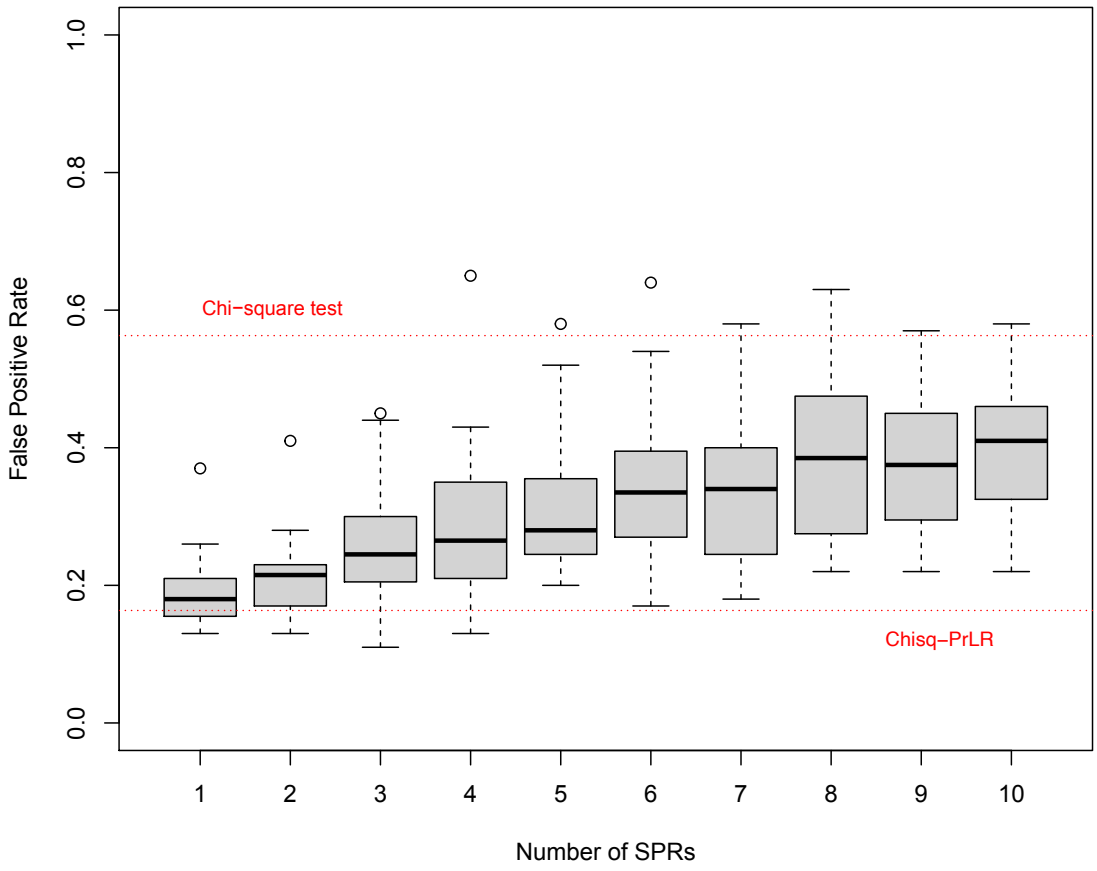


Figure 4.3: Evaluation of the effect of the errors in the tree on the performance of Chisq-PyLR. The x-axis indicates the number of SPRs introduced to the given phylogenetic tree used by Chisq-PyLR. Y-axis indicates the false positive rates detected by the Chisq-PyLR method implemented on the false tree. The horizontal lines indicate the mean false positive rates of the generic Pearson's Chi-square test (0.563 ± 0.054) and Chisq-PrLR method (0.169 ± 0.036) respectively.

for the analysis in CCM, we removed the genes that are very rare (present in $< 1\%$ genomes) or very common (present in $> 99\%$ genomes) as these profiles do not contain much evolutionary information and obtained the final data set of 3786 profiles.

The Chisq-PrLR method with the first principal eigenvector which explains 53.87% of variance, was used to analyze this data set (a 659-tip phylogenetic tree and 3786 phylogenetic profiles). As shown in Fig.4.4, the first principal eigenvector inferred from the profiles correctly extracted the major phylogenetic information such as the close relatives in the *C.difficile* group.

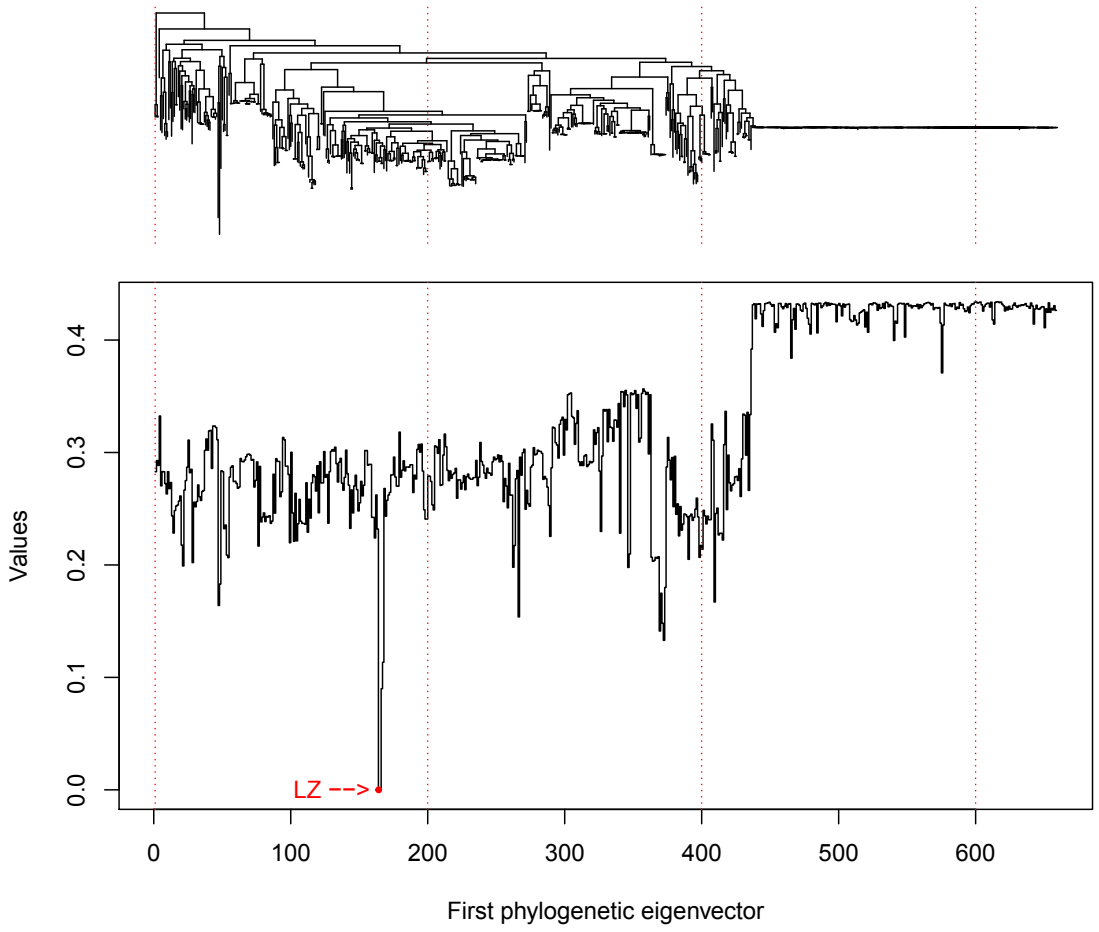


Figure 4.4: The first eigenvector inferred from the profiles, in comparison with the full phylogenetic tree of 659 genomes.

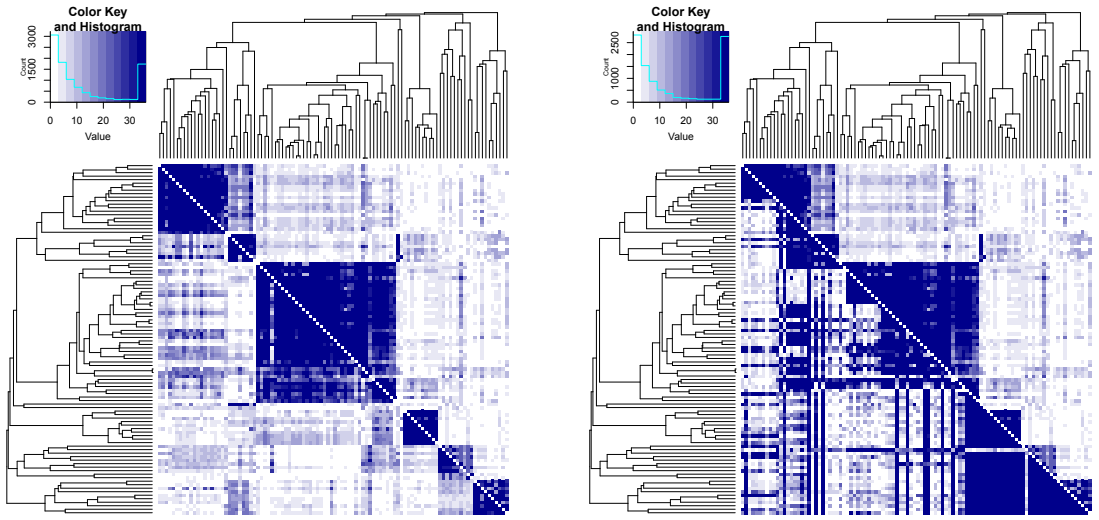
Method comparisons on the LZ data set

Here we used previously studied CCM results to examine the consistency of results between Chisq-PrLR and CCM on the real data and its ability to recover the clustering structure. To provide a clear visual comparison, we selected a random set of 100 genes from previously studied CCM clusters and implemented all vs all pairwise comparisons. As shown in Fig.4.5, Chisq-PrLR identifies the same six major clusters distributed along the diagonal of the heatmap as CCM, but in contrast, the generic Chi-square test detected many significant pairs across the clusters that were not found significant by CCM and did not show consistent clustering structures. A more detailed comparison for all pairs among these 100 profiles between methods, is given in Fig.4.6 and it shows the percentages of most correlated pairs detected by CCM that are also detected by Chisq-PrLR. We calculated the coverage rate as $\frac{|\text{CCM} \cap \text{Chisq-PrLR}|}{|CCM|}$, where $|\cdot|$ represents the number of pairs. The 100% coverage rate (top 10% in CCM vs. top 30% in Chisq-PrLR) compared to 89.9% coverage (top 10% in CCM vs. top 50% in Chi-square) indicates that 10.1% of the strongest links detected by CCM would not be detected by the generic Chi-sq test due to its not taking into account phylogenetic effect. A rate of 0.9697 between the top 20% in CCM and the top 30% in Chisq-PrLR suggests that to recover the most significant 20% of pairs under CCM, instead of implementing an all-vs-all comparison among 100 genes, we can first run Chisq-PrLR which has a running time of around two minutes (performed on a local machine with a 2.5 GHz CPU and 16GB RAM) and then only examine the top 30% of Chisq-PrLR results using CCM and it should produce almost the same result, but could substantially reduce the running time.

4.3.3 Results on the COGs of 678 genomes from the *Lachnospiraceae* family

Data Sets

In this application, we examined the full gene sets of 678 completed and draft genomes from the *Lachnospiraceae* family which were all retrieved from NCBI. In total, 19336 COGs (clusters of orthologous groups) were constructed by searching all the genome



(a) CCM (upper) vs Chisq-PrLR (lower)

(b) CCM (upper) vs Chi-sq (lower)

Figure 4.5: Comparisons of clustering structure recovery between methods. The dendrograms on the left and above are the hierarchical clustering dendrogram using CCM's pairwise comparison scores. The color in the heatmap indicates the significance (-log(P-value)) of dependencies between genes (darker colors indicate stronger dependencies). a) The comparison between CCM and Chisq-PrLR. b) The comparison between CCM and generic Chi-square test.

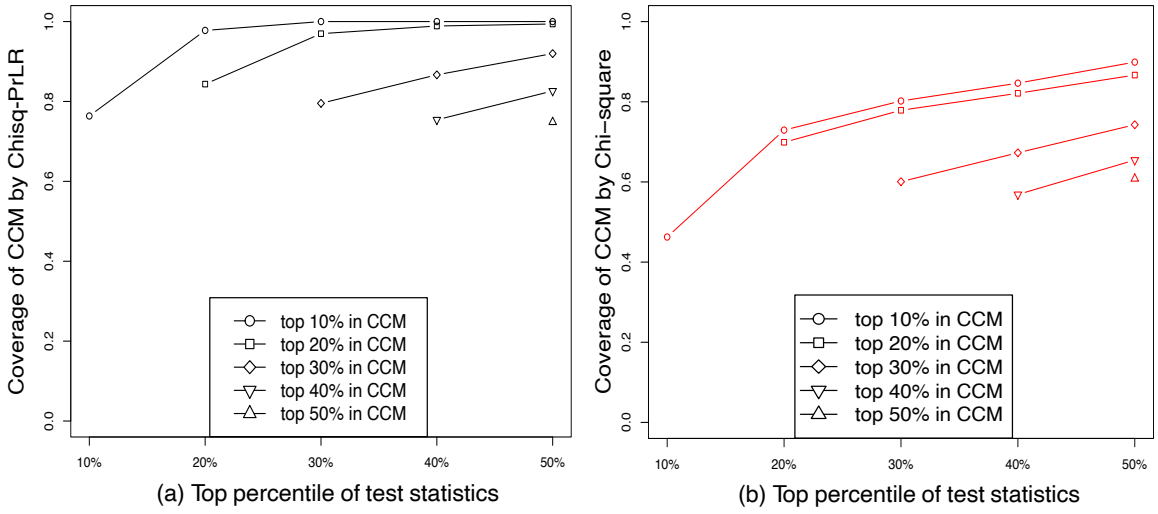


Figure 4.6: Comparisons within highest correlated pairs detected by different methods. The y-axis indicates the coverage rate of pairs with the strongest correlation detected by CCM that were also detected by Chisq-PrLR (a) and generic Chi-sq (b). The coverage rate can be formulated as $\frac{|\text{CCM} \cap \text{Chisq-PrLR}|}{|CCM|}$, where $|\cdot|$ represents the number of pairs. The x-axis indicates the percentages of Chisq-PrLR (a) and generic Chi-sq (b) used to make the comparisons with CCM.

sequences against the eggNOG database which stores the previously studied orthologous groups and functional annotations [38]. The phylogenetic tree was constructed using 46 core genes that are present in all genomes using IQ-TREE [63]. Eight genomes from the *Clostridia* class were used as the outgroup species to root the phylogenetic tree. We further removed the COGs that are either too rare (present in less than 1% of genomes) or too common (present in more than 90% of genomes) as they contain little evolutionary information, to obtain the final data set of $10,755$ phylogenetic profiles.

An all-vs-all comparison was first performed among all profiles using Chisq-PyLR using the first three principal eigenvectors from the tree, for a total of $57,829,635$ pairwise comparisons, and was completed within 3 days on a server running Linux with a 2.67 GHz CPU and 18 GB RAM. The distribution of P-values inferred by Chisq-PyLR and Pearson's Chi-square test is given in Figure 4.7 and shows that the Chisq-PyLR method detected $2,127,944$ more non-significant pairs (P-value > 0.01) than Pearson's Chi-square test. Pearson's Chi-square test also detected $1,172,741$ more strongly significant pairs (P-value $< 1 \times 10^{-10}$).

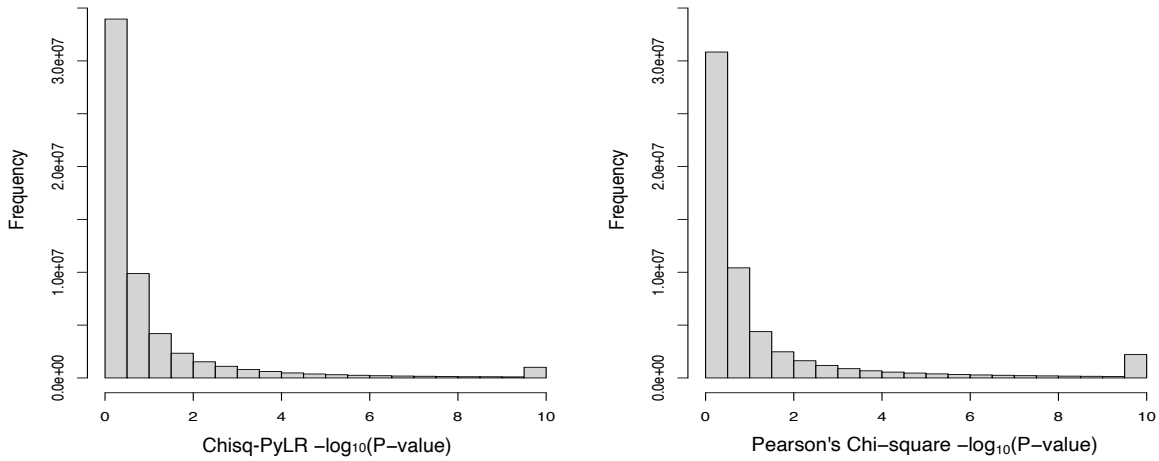


Figure 4.7: Distributions of the P-values ($-\log_{10}$) of all-vs-all comparisons. a) P-values inferred by Chisq-PyLR. b) P-values inferred by Pearson's Chi-square test. All the P-values less than 1×10^{-10} (including 0) are set to be 1×10^{-10}.

To examine the difference between the pairs inferred by the two methods, we applied the CCM method on a randomly sampled subset of 1000 non-significant pairs (P-value > 0.01) inferred by Chisq-PyLR. Figure 4.8 supports the same conclusion with the simulation results that the phylogeny-free Pearson's Chi-square test tends

to generate more false positive results. Within the 1000 randomly sampled non-significant pairs (P-value > 0.01) inferred by Chisq-PyLR, the CCM method showed consistent results with 49 pairs having P-values less than 0.01 and only 7 pairs exceeding the significance level of 0.001, while the Pearson's Chi-square test detected 221 significantly correlated pairs (P-value < 0.01).

Examples of detected functional clusters

Studies [58, 100] have suggested that different functional gene clusters may have different strength of coevolutionary associations. Based on the 2011 COGs that are classified into functional categories from the eggNOG database, Table 4.3 provides a general summary of the functional categories and the corresponding significance of association (the average of all pairwise chi-square statistics within a category). According to the mean and $5th$ percentile of the test statistics, the top three functional categories that show the strongest associations are all related to "N: Cell motility", which is consistent with our previous studies in which the flagellar gene clusters were detected as well as the category of "E: Amino acid transport and metabolism". Note that this summary in Table 4.3, based on functional categories, may be over generalized and the COGs in the same category could be further divided into smaller clusters which would increase the significance level, such as the COGs related to "G: Carbohydrate transport and metabolism" and "T: Signal transduction mechanisms". In our previous study, multiple significant gene clusters related to carbohydrate transport and signal transduction were detected and those clusters also showed distinct phyletic patterns.

To find the functionally associated gene clusters in this large dataset, we first applied hierarchcial clustering with Ward lineage on the chi-square statistics inferred for $57,829,635$ pairs and then extracted 200 clusters. Figure 4.9 shows the distribution of the sizes and cluster compactness (average test statistics) of these 200 clusters. Except for one cluster of 1596 COGs, all other clusters are below 500 in size. The network in Figure 4.10 which consists of 5675 non-singleton COGs and 69840 strong significant links (Chi-square statistics > 100) provides an overview of the clustering structure in the data.

Three examples of mid-sized clusters are given here. The cluster (id:1) in Figure

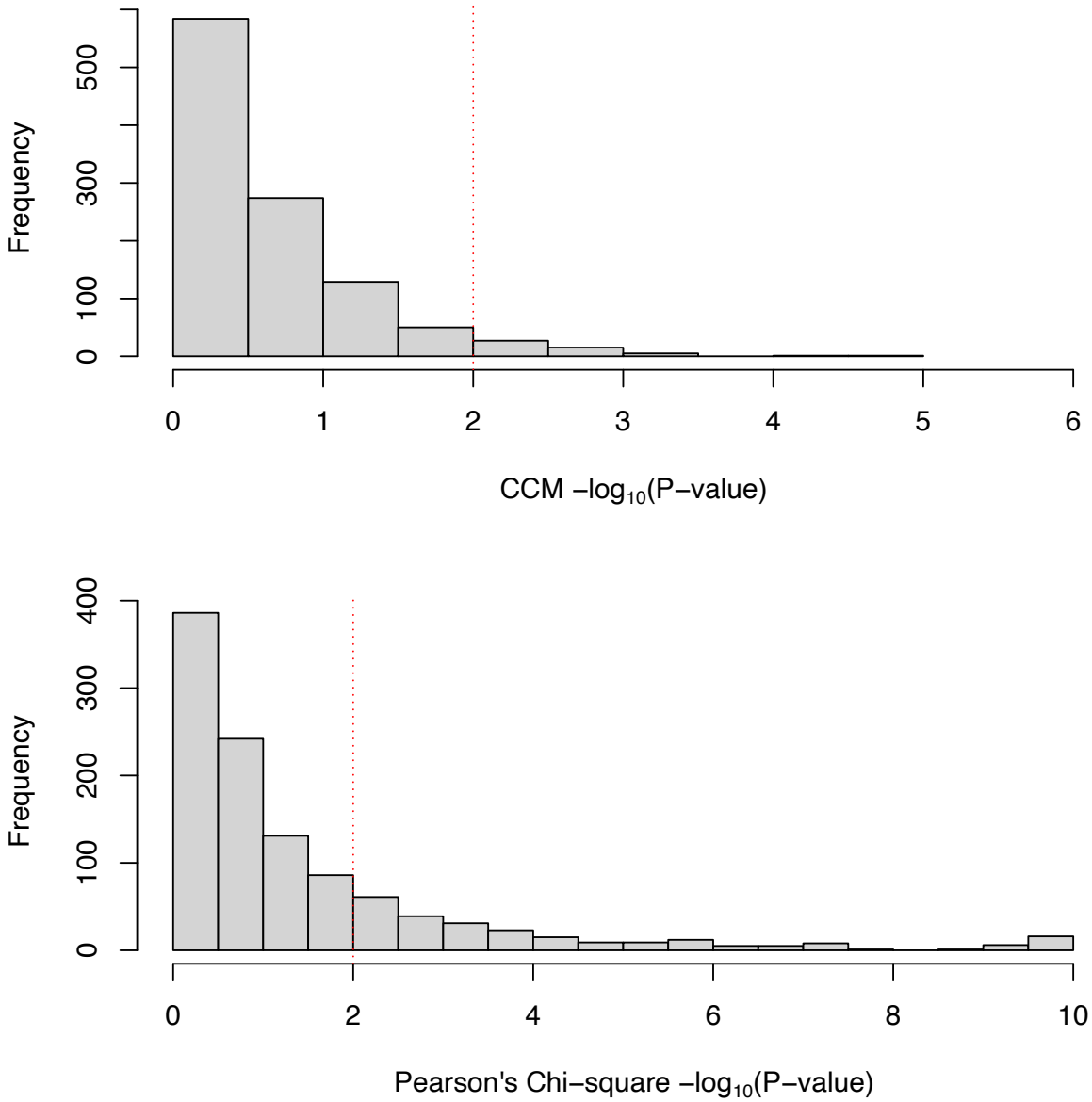


Figure 4.8: Comparisons with CCM and Pearson's Chi-square test using 1000 randomly sampled non-significant pairs (P-value > 0.01) inferred by Chisq-PyLR. a) Distribution of the P-values ($-\log_{10}$) inferred by CCM. b) Distribution of the P-values ($-\log_{10}$) inferred by Pearson's Chi-square test.

Table 4.3: Summary of 2011 annotated COGs in terms of functional categories (ordered by mean test statistics). Frequency: the number of COGs annotated with the corresponding functional category. Descriptive statistics: percentiles (5th and 10th), mean and standard deviation of chi-square statistics inferred by Chisq-PyLR between all COGs in the same functional category. Proportion: the proportion of significant pairs within each functional category at the significance level of 0.001. The functional categories consisting of multiple letters indicate the COGs are assigned into multiple categories. Only functional categories with more than 10 COGs are reported.

Functional Category	Description	Frequency	5th Percentile	10th Percentile	Mean	SD	Proportion
N	Cell motility	47	180.2639	173.5355	51.4452	66.9711	0.519
NT		10	190.7481	184.8269	35.963	68.3033	0.222
NU		21	88.6121	28.8996	15.367	40.2702	0.2238
H	Coenzyme transport and metabolism	111	50.2794	32.048	13.2323	32.694	0.2834
G	Carbohydrate transport and metabolism	180	49.8328	32.4412	12.6163	25.185	0.305
P	Inorganic ion transport and metabolism	155	42.1033	26.8661	11.7659	33.057	0.2552
E	Amino acid transport and metabolism	183	45.5988	29.4346	11.5605	26.8942	0.275
C	Energy production and conversion	142	42.6275	27.602	10.5263	20.0917	0.2672
KT		15	33.1911	24.2891	10.1659	13.4139	0.3429
F	Nucleotide transport and metabolism	60	40.362	23.0598	9.0483	19.9672	0.2056
I	Lipid transport and metabolism	49	26.8406	17.2951	7.5604	22.9695	0.1726
Q	Secondary metabolites biosynthesis, transport and catabolism	61	30.0744	17.4827	7.25164	17.7791	0.1743
T	Signal transduction mechanisms	79	28.714	17.9663	6.9118	12.7692	0.1824
D	Cell cycle control, cell division, chromosome partitioning	46	22.0844	14.9472	6.5338	13.6552	0.1604
J	Translation, ribosomal structure and biogenesis	58	23.7671	14.8102	6.2871	17.1087	0.1488
O	Posttranslational modification, protein turnover, chaperones	76	24.3818	15.7948	6.1246	18.9283	0.1393
M	Cell wall/membrane/envelope biogenesis	130	24.1162	15.1342	5.7598	11.0261	0.1522
V	Defense mechanisms	47	22.42	13.2397	5.6431	14.7632	0.136
U	Intracellular trafficking, secretion, and vesicular transport	28	19.2776	11.6816	5.3919	17.6907	0.119
K	Transcription	190	21.9949	12.601	4.9393	11.7957	0.116
L	Replication, recombination and repair	213	18.8062	10.9203	4.7573	14.5683	0.0991
Total	All annotated COGs	2011	29.4427	18.4821	6.9105	13.7137	0.1628

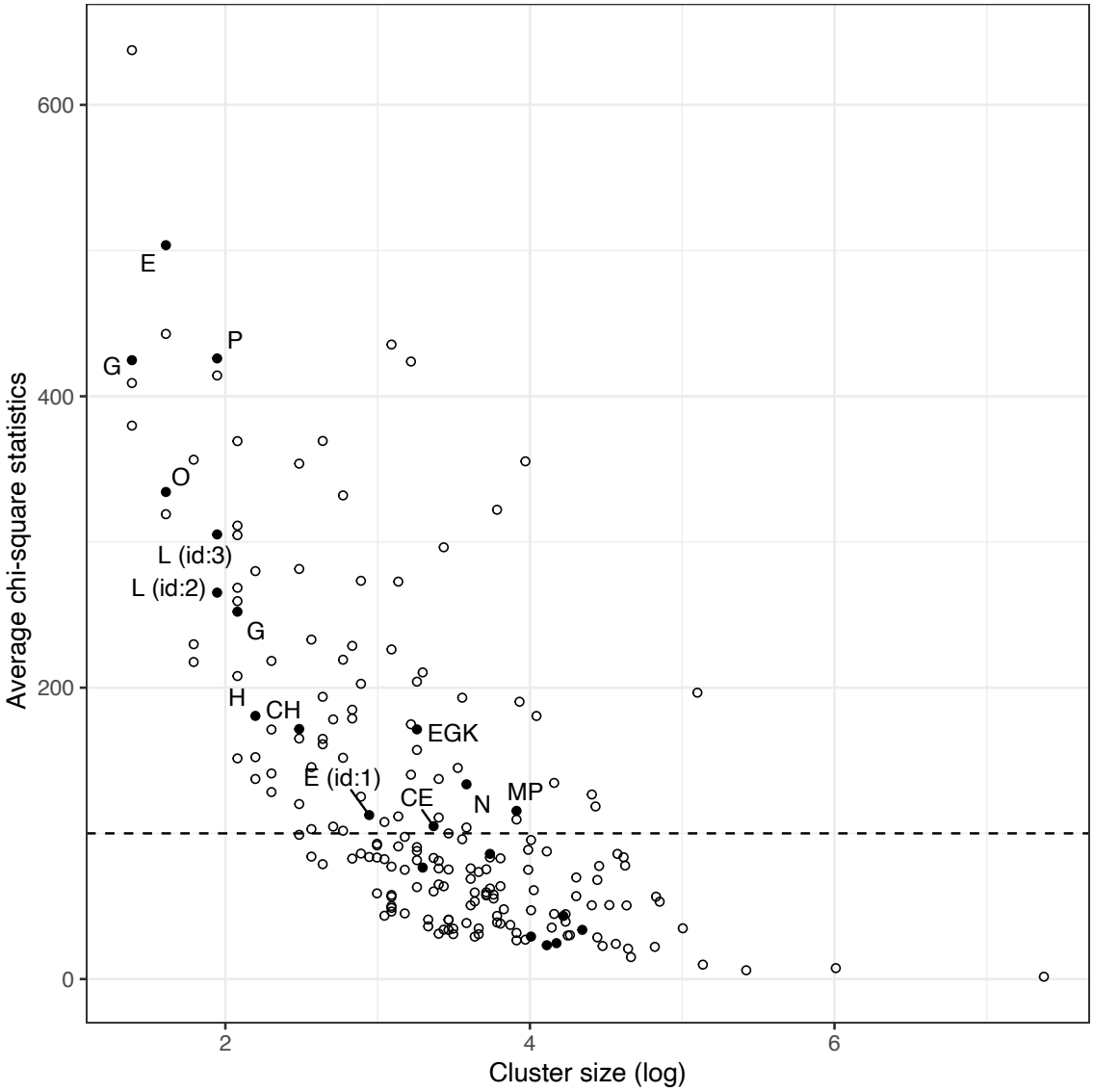


Figure 4.9: 200 clusters generated by applying hierarchical clustering on the all-vs-all pairwise chi-square statistics using Chisq-PyLR. The x-axis indicates the cluster size in log scale and the y-axis indicates the average of chi-square statistics within the cluster. Solid circles indicate that the clusters have more than 50% of COGs classified into known functional categories. The labels of the points indicate the major ($> 80\%$) functional categories within the cluster.

Figure 4.10: The network of 5675 COGs with 69840 strongly significant links (chi-square statistics > 100) inferred by Chisq-PyLR. Blue vertices indicate the COGs annotated with functional category and gray vertices indicate unclassified COGs.

4.9 consists of 19 COGs (12 related to functional category "E: Amino acid transport and metabolism", 5 unknown) and their phylogenetic profiles in Figure 4.11 show a complementary phyletic pattern within the cluster, which indicates a negative correlation and was not often observed for the LZ dataset from the previous section. Figure 4.12 shows the phylogenetic profiles of two clusters (id: 2 and 3 in Figure 4.9) each consisting of 7 COGs. Although 10 out of 14 COGs are classified into the same functional category "L" (3 unknown), the COGs are clustered into two groups and show two distinct patterns: one is relatively rare (present in an average of 79 genomes) and the other one is more common (present in an average of 208 genomes), with no negative correlation shown between these two clusters.

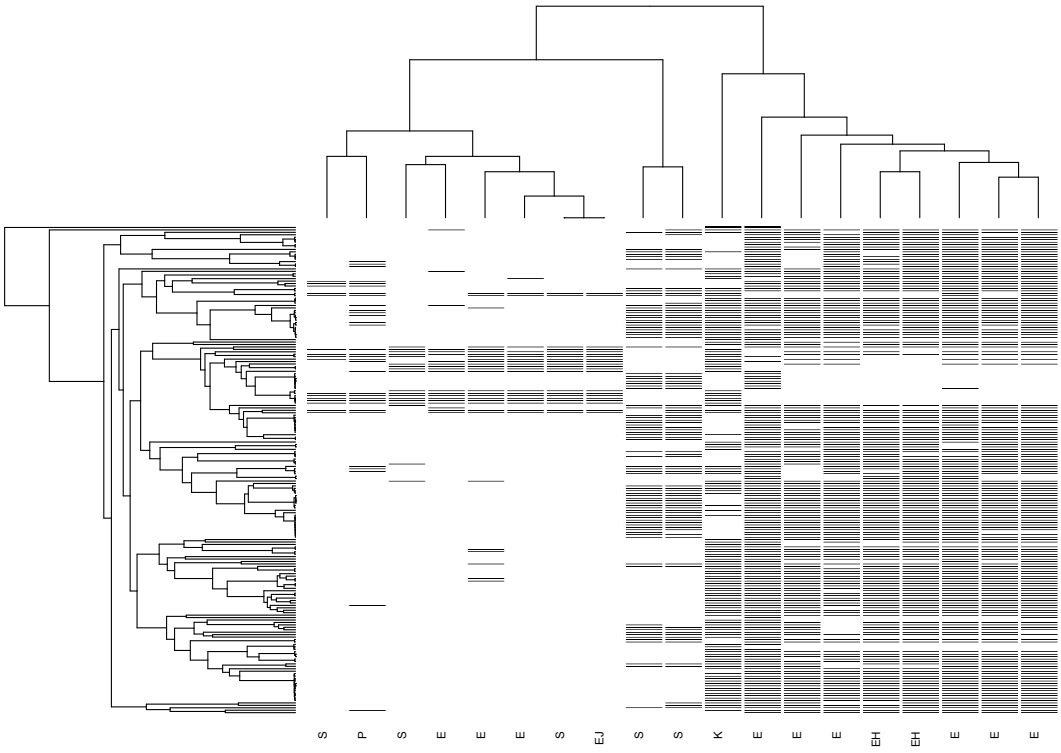


Figure 4.11: Phylogenetic profile of a cluster of COGs (id:1) classified into functional category "E". Functional category "S" indicates unknown annotation. The tree on the left is the phylogenetic tree with 200 random tips for illustration.

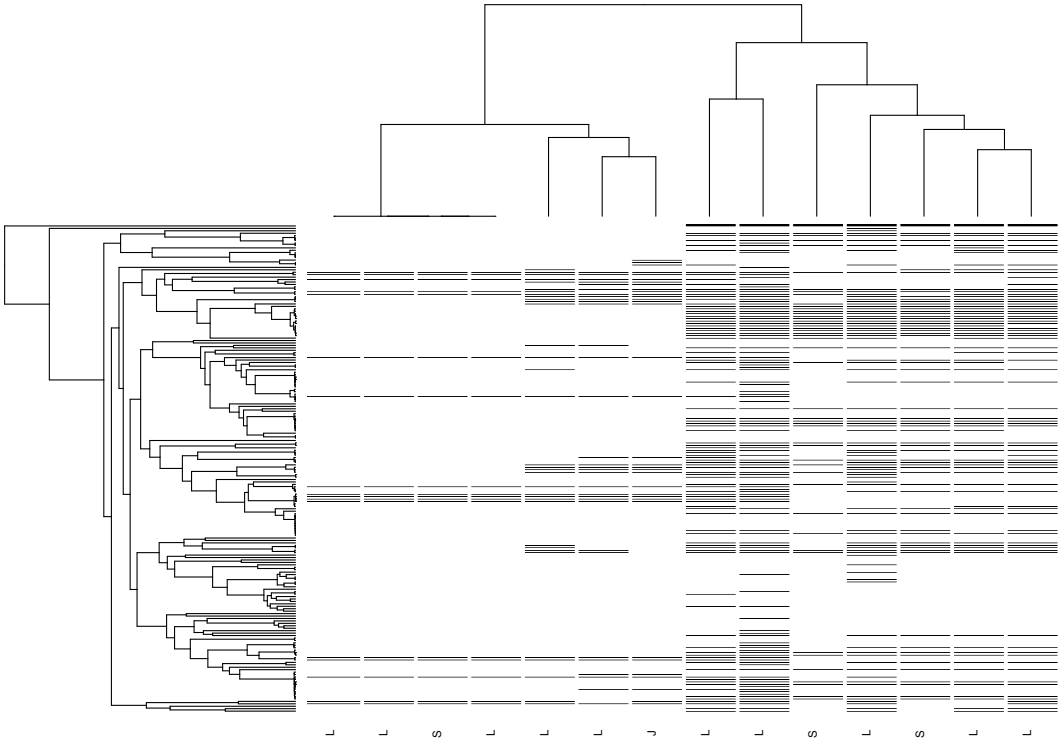


Figure 4.12: Phylogenetic profiles of two clusters (id: 2 and 3) related to functional category "L". Functional category "S" indicates unknown annotation. The tree on the left is the phylogenetic tree with 200 random tips for illustration.

4.4 Discussion

In this chapter, we proposed a matrix decomposition-based method to test the dependency between binary profiles conditioning on the tree topology. Although our CCM model is computationally more efficient than Pagel's method, it still can not handle large data sets since the number of pair-wise comparisons will increase quadratically. Compared to other heursitic methods, our Chisq-PLR methods not only have competitive running speed but also give support to better biological explanations for the data.

We first used simulated data via the CCM framework to evaluate the performance of Chisq-PLR. The results in Table 4.2 and 4.1 show that Chisq-PyLR and Chisq-PrLR performed similarly and the type I error rates of both methods are much lower than that of the phylogeny-naive Pearson's Chi-square test, while still achieving a high statistical power (above 0.8). We also applied the Chisq-PrLR method on the previously studied LZ data using CCM and showed that our method is able to correctly recover a similar clustering structure to CCM. We further applied the Chisq-PyLR method to the $10,755$ COGs of 678 genomes from the *Lachnospiraceae* family to detect the functional gene clusters. All-vs-all comparisons of a total of $57,829,635$ pairs, can be processed by our methods within 3 days on a server running Linux with a 2.67 GHz CPU and 18 GB RAM. We first examined the strength of associations between COGs in terms of annotated functional categories and found that the top three functional categories that shows the most significant associations are all related to "N: Cell motility". We further explored the distribution of the strength of associations within 200 clusters discovered with hierarchical clustering and constructed a network consisting of $5,675$ nodes with $69,840$ edges to provide an overview of the clustering structure of COGs in these 678 Lachnospiraceae genomes.

4.5 Author Contribution

In this study, I participated in the design of the work, implemented the methods, conducted the analysis and wrote the manuscript.

Chapter 5

Conclusions

Phylogenetic profiles, which summarize the presence and absence patterns of genes in a set of genomes, can be used to identify genes that have correlated evolutionary histories. Genes with common distributions are more likely to be functionally linked, and in prokaryotes may highlight shared patterns of lateral gene transfer. However, these distributions are impacted by phylogenetic relationships among genomes. In this thesis, we developed three phylogenetic comparative methods to infer gene coevolution and to discover clusters of genes that have correlated evolutionary relationships based on phylogenetic profiles.

We first proposed an approach in Chapter 2 that uses Pagel's correlation test to infer the evolutionary similarities between genes and a hierarchical-clustering approach to define sets of genes with common distributions across the organisms. We applied this method to the LZ data set. LZ has a very large genome relative to most other clostridia and elucidating its ecological role will be challenging. Our method successfully recovers phylogenetically and functionally cohesive gene clusters and highlights probable highways of gene sharing that have shaped this genome and its close neighbors. The results of this study further support the assumption of our work that the genes with correlated phylogenetic profiles also tend to be functionally linked. One significant limitation of this method is the heavy computational cost of applying Pagel's correlation model to all pairs of distinct phylogenetic profiles: although our full dataset included 687 genomes, computational time limitations restricted us to the analysis of a set of 74 genomes.

In Chapter 3, we proposed the Community Coevolution Model (CCM), a new coevolutionary model to analyze the evolutionary associations among genes. In the CCM, traits are considered to evolve as a community with interactions, and the transition rate for each trait depends on the current states of other traits. CCM has the additional advantage of being able to examine multiple traits as a community to

reveal more dependency relationships. We also developed a simulation procedure to generate phylogenetic profiles of gene sets with correlated distributions and adjustable strength of interactions.

A simulation study demonstrates that CCM is more accurate than other methods including the Jaccard Index and three tree-aware methods including Pagel's correlation test. The parameterization of CCM makes the interpretation of the relations between genes more direct, which leads to Darwin's scenario being identified easily based on the estimated parameters. We showed that CCM is more efficient and fits real data better than Pagel's method resulting in higher likelihood scores with fewer parameters. Our method is more efficient and approximately 5 times faster than Pagel's method and is able to examine the LZ data set with the full phylogenetic tree of 659 genomes. We improved and completed the LZ data analysis by providing a list of predictions on 823 unannotated genes in the LZ data set. We also applied the CCM to 44 proteins in the well-studied Mitochondrial Respiratory Complex I and recovered associations that mapped well onto the structural associations that exist in the complex. The new results showed that our method as a general comparative framework can still work well on eukaryotic data where lateral gene transfers are not as prevalent as in prokaryotic genes.

Although in terms of pair-wise comparisons, the CCM model is more efficient than Pagel's method, it cannot scale to large datasets containing many thousands of profiles due to the quadratic scaling of pairwise comparison. To handle large data sets, we developed a fast matrix decomposition-based method (Chisq-PLR) in Chapter 4 to test the dependency between binary profiles conditioning on the tree topology. It is computationally efficient for large-scale analyses, gives support to better biological explanations to the data than heuristic methods, and also works with (Chisq-PyLR) or without (Chisq-PrLR) a provided phylogenetic tree, which makes our method robust to phylogenetic uncertainties. This fast method can be used to pre-process the large data set to reduce the number of computations that need to be carried out by CCM or another probabilistic model based method.

We first used simulated data via the CCM framework to assess the performance of Chisq-PLR in correcting for phylogenetic effects and the results show that the type

I error rates of the Chisq-PLR methods are much lower than that of the phylogeny-naïve Pearson's Chi-square test, while still achieving high statistical power. We also applied the Chisq-PrLR method on a subset of the previously studied LZ data using CCM and showed that the method is able to correctly recover a similar clustering structure to CCM. We further applied the Chisq-PyLR method to the 10755 COGs of 678 genomes from the Lachnospiraceae family to detect the functional gene clusters. All-vs-all comparisons of a total of 57, 829, 635 pairs, can be processed by our method within 3 days on a server running Linux with a 2.67 GHz CPU and 18 GB RAM. We further explored the distribution of the strength of associations within 200 clusters discovered with hierarchical clustering and constructed a network consisting of 5675 nodes with 69840 edges to provide an overview of the clustering structure of COGs in these 678 Lachnospiraceae genomes.

In this study, we mainly focused on developing the phylogenetic comparative methods with a given reference tree. However, the construction of phylogenetic trees may involve error or uncertainty. Extending our methods to account for the uncertainty of the input phylogenetic tree could be a possible future work. We also met a challenge in extending our CCM model to directly model larger communities. The dimension of the transition rate matrix Q in the CCM model will increase exponentially as we include more genes into the community. However, this Q matrix is highly sparse and we have found that if we reorder the rows and columns of the transition matrix, there exists a recursive structure. Our future work will explore the possible solutions to decompose the Q matrix more efficiently so that the CCM method is scalable.

Bibliography

[1] Kjersti Aagaard, Kevin Riehle, Jun Ma, Nicola Segata, Toni-Ann Mistretta, Cristian Coarfa, Sabeen Raza, Sean Rosenbaum, Ignatia Van den Veyver, Aleksandar Milosavljevic, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PloS one*, 7(6):e36466, 2012.

[2] Luis David Alcaraz, Gabriel Moreno-Hagelsieb, Luis E Eguiarte, Valeria Souza, Luis Herrera-Estrella, and Gabriela Olmedo. Understanding the evolutionary relationships and major traits of bacillus through comparative genomics. *BMC genomics*, 11(1):1–17, 2010.

[3] Vijay C Antharam, Eric C Li, Arif Ishmael, Anuj Sharma, Volker Mai, Kenneth H Rand, and Gary P Wang. Intestinal dysbiosis and depletion of butyrogenic bacteria in clostridium difficile infection and nosocomial diarrhea. *Journal of clinical microbiology*, 51(9):2884–2892, 2013.

[4] Carlo Baldassi, Marco Zamparo, Christoph Feinauer, Andrea Procaccini, Riccardo Zecchina, Martin Weigt, and Andrea Pagnani. Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PloS one*, 9(3):e92721, 2014.

[5] Eduardo Balsa, Ricardo Marco, Ester Perales-Clemente, Radek Szklarczyk, Enrique Calvo, Manuel O Landázuri, and José Antonio Enríquez. Ndufa4 is a subunit of complex iv of the mammalian electron transport chain. *Cell metabolism*, 16(3):378–386, 2012.

[6] Daniel Barker and Mark Pagel. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS computational biology*, 1(1):e3, 2005.

[7] Miriam Barlow. What antimicrobial resistance has taught us about horizontal gene transfer. *Horizontal Gene Transfer*, pages 397–411, 2009.

[8] Robert G Beiko, Timothy J Harlow, and Mark A Ragan. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40):14332–14337, 2005.

[9] Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[10] Alexander G Bick, Sarah E Calvo, and Vamsi K Mootha. Evolutionary diversity of the mitochondrial calcium uniporter. *Science*, 336(6083):886–886, 2012.

[11] Simon P Blomberg and Theodore Garland Jr. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology*, 15(6):899–910, 2002.

[12] Peter M Bowers, Matteo Pellegrini, Mike J Thompson, Joe Fierro, Todd O Yeates, and David Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome biology*, 5(5):1–13, 2004.

[13] James M Cheverud, Malcolm M Dow, and Walter Leutenegger. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution*, 39(6):1335–1351, 1985.

[14] Ofir Cohen, Haim Ashkenazy, David Burstein, and Tal Pupko. Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics*, 28(18):i389–i394, 2012.

[15] Ofir Cohen, Haim Ashkenazy, Eli Levy Karin, David Burstein, and Tal Pupko. Copap: coevolution of presence–absence patterns. *Nucleic acids research*, 41(W1):W232–W237, 2013.

[16] Shawn Cokus, Sayaka Mizutani, and Matteo Pellegrini. An improved method for identifying functionally linked proteins using phylogenetic profiles. In *BMC bioinformatics*, volume 8, pages 1–12. Springer, 2007.

[17] Qian Cong, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449):185–189, 2019.

[18] Miklós Csűös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, 2010.

[19] Huttenhower Curtis, Martin J Blaser, Gevers Dirk, Karthik C Kota, Knight Rob, Bo Liu, Lu Wang, Abubucker Sahar, James R White, Jonathan H Badger, et al. Structure, function and diversity of the healthy human microbiome. *Nature (London)*, 486(7402):207–214, 2012.

[20] Yves Desdevises, Pierre Legendre, Lamia Azouzi, and Serge Morand. Quantifying phylogenetically structured environmental variation. *Evolution*, 57(11):2647–2652, 2003.

[21] José Alexandre Felizola Diniz-Filho, Carlos Eduardo Ramos de Sant'Ana, and Luis Mauricio Bini. An eigenvector method for estimating phylogenetic inertia. *Evolution*, 52(5):1247–1262, 1998.

[22] Stijn van Dongen and Cei Abreu-Goodger. Using mcl to extract clusters from networks. In *Bacterial Molecular Networks*, pages 281–295. Springer, 2012.

[23] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[24] Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

[25] Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[26] Noah Fierer, Christian L Lauber, Kelly S Ramirez, Jesse Zaneveld, Mark A Bradford, and Rob Knight. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME journal*, 6(5):1007–1017, 2012.

[27] Hunter B Fraser, Aaron E Hirsh, Dennis P Wall, and Michael B Eisen. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences*, 101(24):9033–9038, 2004.

[28] Iddo Friedberg. Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, 7(3):225–242, 2006.

[29] Debra L Fulton, Yvonne Y Li, Matthew R Laird, Benjamin GS Horsman, Fiona M Roche, and Fiona SL Brinkman. Improving the specificity of high-throughput ortholog prediction. *BMC bioinformatics*, 7(1):1–16, 2006.

[30] László Zsolt Garamszegi. *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*. Springer, 2014.

[31] Marta Goberna and Miguel Verdú. Predicting microbial traits with phylogenies. *The ISME Journal*, 10(4):959–967, 2016.

[32] J Peter Gogarten and Jeffrey P Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005.

[33] Sean W Graham, Richard G Olmstead, and Spencer CH Barrett. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular biology and evolution*, 19(10):1769–1781, 2002.

[34] Runyu Guo, Shuai Zong, Meng Wu, Jinke Gu, and Maojun Yang. Architecture of human mitochondrial respiratory megacomplex i2iii2iv2. *Cell*, 170(6):1247–1257, 2017.

[35] Bernhard Haubold and Thomas Wiehe. Comparative genomics: methods and applications. *Naturwissenschaften*, 91(9):405–421, 2004.

[36] Claire N Hirst and Donald A Jackson. Reconstructing community relationships: the impact of sampling error, ordination approach, and gradient length. *Diversity and distributions*, 13(4):361–371, 2007.

[37] John P Huelsenbeck, Bruce Rannala, and John P Masly. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475):2349–2350, 2000.

[38] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, et al. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1):D309–D314, 2019.

[39] Martijn Huynen, Berend Snel, Warren Lathe, and Peer Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, 10(8):1204–1210, 2000.

[40] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1–11, 2010.

[41] Donald A Jackson. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214, 1993.

[42] Raja Jothi, Teresa M Przytycka, and L Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC bioinformatics*, 8(1):1–17, 2007.

[43] Patrick J Keeling and Jeffrey D Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.

[44] Philip R Kensche, Vera van Noort, Bas E Dutilh, and Martijn A Huynen. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society Interface*, 5(19):151–170, 2008.

[45] Peter Kharchenko, Lifeng Chen, Yoav Freund, Dennis Vitkup, and George M Church. Identifying metabolic enzymes with multiple types of association evidence. *BMC bioinformatics*, 7(1):1–16, 2006.

[46] Thorsten Kloesges, Ovidiu Popa, William Martin, and Tal Dagan. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular biology and evolution*, 28(2):1057–1074, 2011.

[47] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39(1):309–338, 2005.

[48] Eugene V Koonin, L Aravind, and Alexey S Kondrashov. The impact of comparative genomics on our understanding of evolution. *Cell*, 101(6):573–576, 2000.

[49] Victor Kunin, Leon Goldovsky, Nikos Darzentas, and Christos A Ouzounis. The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, 15(7):954–959, 2005.

[50] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W Ussery. Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic acids research*, 35(9):3100–3108, 2007.

[51] Yang Li, Sarah E Calvo, Roee Gutman, Jun S Liu, and Vamsi K Mootha. Expansion of biological pathways based on evolutionary inference. *Cell*, 158(1):213–225, 2014.

[52] Yang Li, Shaoyang Ning, Sarah E Calvo, Vamsi K Mootha, and Jun S Liu. Bayesian hidden markov tree models for clustering genes with shared evolutionary history. *The Annals of Applied Statistics*, 13(1):606–637, 2019.

[53] Chaoyue Liu, Toby Kenney, Robert G Beiko, and Hong Gu. The community coevolution model with application to the study of evolutionary relationships between genes based on phylogenetic profiles. *Systematic Biology*, 2022.

[54] Chaoyue Liu, Benjamin Wright, Emma Allen-Vercoe, Hong Gu, and Robert Beiko. Phylogenetic clustering of genes reveals shared evolutionary trajectories and putative gene functions. *Genome biology and evolution*, 10(9):2255–2265, 2018.

[55] Todd M Lowe and Sean R Eddy. trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic acids research*, 25(5):955–964, 1997.

[56] Wayne P Maddison and Richard G FitzJohn. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic biology*, 64(1):127–136, 2015.

[57] Oleksandr M Maistrenko, Daniel R Mende, Mechthild Luetge, Falk Hildebrand, Thomas SB Schmidt, Simone S Li, João F Matias Rodrigues, Christian von Mering, Luis Pedro Coelho, Jaime Huerta-Cepas, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME journal*, 14(5):1247–1259, 2020.

[58] Edward M Marcotte et al. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature biotechnology*, 21(9):1055–1062, 2003.

[59] A Meade and M Pagel. Bayestraits v3. 0.1, 2017.

[60] David Moi, Laurent Kilchoer, Pablo S Aguilar, and Christophe Dessimoz. Scalable phylogenetic profiling using minhash uncovers likely eukaryotic sexual reproduction genes. *PLoS computational biology*, 16(7):e1007553, 2020.

[61] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John wiley & sons, 2010.

[62] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[63] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

[64] Yulong Niu, Shayan Moghimyfiroozabad, Sepehr Safaie, Yi Yang, Elizabeth A Jonas, and Kambiz N Alavian. Phylogenetic profiling of mitochondrial proteins and integration analysis of bacterial transcription units suggest evolution of f1fo atp synthase from multiple modules. *Journal of molecular evolution*, 85(5):219–233, 2017.

[65] Howard Ochman, Jeffrey G Lawrence, and Eduardo A Groisman. Lateral gene transfer and the nature of bacterial innovation. *nature*, 405(6784):299–304, 2000.

[66] Mark Pagel. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342):37–45, 1994.

[67] Mark Pagel and Andrew Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6):808–825, 2006.

[68] Csaba Pál, Balázs Papp, and Martin J Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics*, 37(12):1372–1375, 2005.

[69] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

[70] Kathryn S Peiman and Beren W Robinson. Comparative analyses of phenotypic trait covariation within and among populations. *The American Naturalist*, 190(4):451–468, 2017.

[71] Matteo Pellegrini. Using phylogenetic profiles to predict functional relationships. In *Bacterial molecular networks*, pages 167–177. Springer, 2012.

[72] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[73] Kathryn J Pflughoeft and James Versalovic. Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease*, 7:99–122, 2012.

[74] Hervé Philippe and Christophe J Douady. Horizontal gene transfer and phylogenetics. *Current opinion in microbiology*, 6(5):498–505, 2003.

[75] Pere Puigbò, Yuri I Wolf, and Eugene V Koonin. Search for a'tree of life'in the thicket of the phylogenetic forest. *Journal of biology*, 8(6):1–17, 2009.

[76] R Alexander Pyron, Gabriel C Costa, Michael A Patten, and Frank T Burbrink. Phylogenetic niche conservatism and the evolutionary basis of ecological speciation. *Biological Reviews*, 90(4):1248–1262, 2015.

[77] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.

[78] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

[79] Thiago F Rangel, Robert K Colwell, Gary R Graves, Karolina Fučíková, Carsten Rahbek, and José Alexandre F Diniz-Filho. Phylogenetic uncertainty revisited: Implications for ecological analyses. *Evolution*, 69(5):1301–1312, 2015.

[80] Valentín Ruano-Rubio, Olivier Poch, and Julie D Thompson. Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC bioinformatics*, 10(1):1–17, 2009.

[81] Ilyas R Sadreyev, Fei Ji, Emiliano Cohen, Gary Ruvkun, and Yuval Tabach. Phylogene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic acids research*, 43(W1):W154–W159, 2015.

[82] Glenn M Sanford, William I Lutterschmidt, and Victor H Hutchison. The comparative method revisited. *BioScience*, 52(9):830–836, 2002.

[83] Shannon J Sibbald, Laura Eme, John M Archibald, and Andrew J Roger. Lateral gene transfer mechanisms and pan-genomes in eukaryotes. *Trends in Parasitology*, 36(11):927–941, 2020.

[84] Kimmen Sjölander. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2):170–179, 2004.

[85] Elizabeth Skippington and Mark A Ragan. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS microbiology reviews*, 35(5):707–735, 2011.

[86] Yun S Song. On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, 7(3):365–379, 2003.

[87] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[88] Bärbel Stecher, Rémy Denzler, Lisa Maier, Florian Bernet, Mandy J Sanders, Derek J Pickard, Manja Barthel, Astrid M Westendorf, Karen A Krogfelt, Alan W Walker, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal enterobacteriaceae. *Proceedings of the National Academy of Sciences*, 109(4):1269–1274, 2012.

[89] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021.

[90] William Carlisle Thacker. The role of the hessian matrix in fitting models to measurements. *Journal of Geophysical Research: Oceans*, 94(C5):6177–6196, 1989.

[91] Benjamin JM Tremblay, Briallen Lobb, and Andrew C Doxey. Phylocorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. *Bioinformatics*, 2021.

[92] Fernando Domingues Kümmel Tria, Giddy Landan, and Tal Dagan. Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology & Evolution*, 1(1):1–7, 2017.

[93] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[94] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

[95] Josef C Uyeda, Rosana Zenil-Ferguson, and Matthew W Pennell. Rethinking phylogenetic comparative methods. *Systematic Biology*, 67(6):1091–1109, 2018.

[96] Jean-Philippe Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18(suppl_1):S276–S284, 2002.

[97] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

[98] Michiel Vos, Matthijn C Hesselman, Tim A Te Beek, Mark WJ van Passel, and Adam Eyre-Walker. Rates of lateral gene transfer in prokaryotes: high but why? *Trends in microbiology*, 23(10):598–605, 2015.

[99] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[100] Fiona J Whelan, Rebecca J Hall, and James O McInerney. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Molecular biology and evolution*, 38(9):3697–3708, 2021.

[101] Jie Wu, Simon Kasif, and Charles DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530, 2003.

[102] Jie Wu, Joseph C Mellor, and Charles De Lisi. Deciphering protein network organization using phylogenetic profile groups. *Genome Informatics*, 16(1):142–149, 2005.

[103] Martin Wu and Alexandra J Scott. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics*, 28(7):1033–1034, 2012.

[104] Yuzhen Ye, Jeong-Hyeon Choi, and Haixu Tang. Rapsearch: a fast protein similarity search tool for short reads. *BMC bioinformatics*, 12(1):1–10, 2011.

[105] Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic acids research*, 42(D1):D643–D648, 2014.

[106] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

[107] Alain F Zuur, Elena N Ieno, and Graham M Smith. Principal coordinate analysis and non-metric multidimensional scaling. *Analysing ecological data*, pages 259–264, 2007.

Appendix A

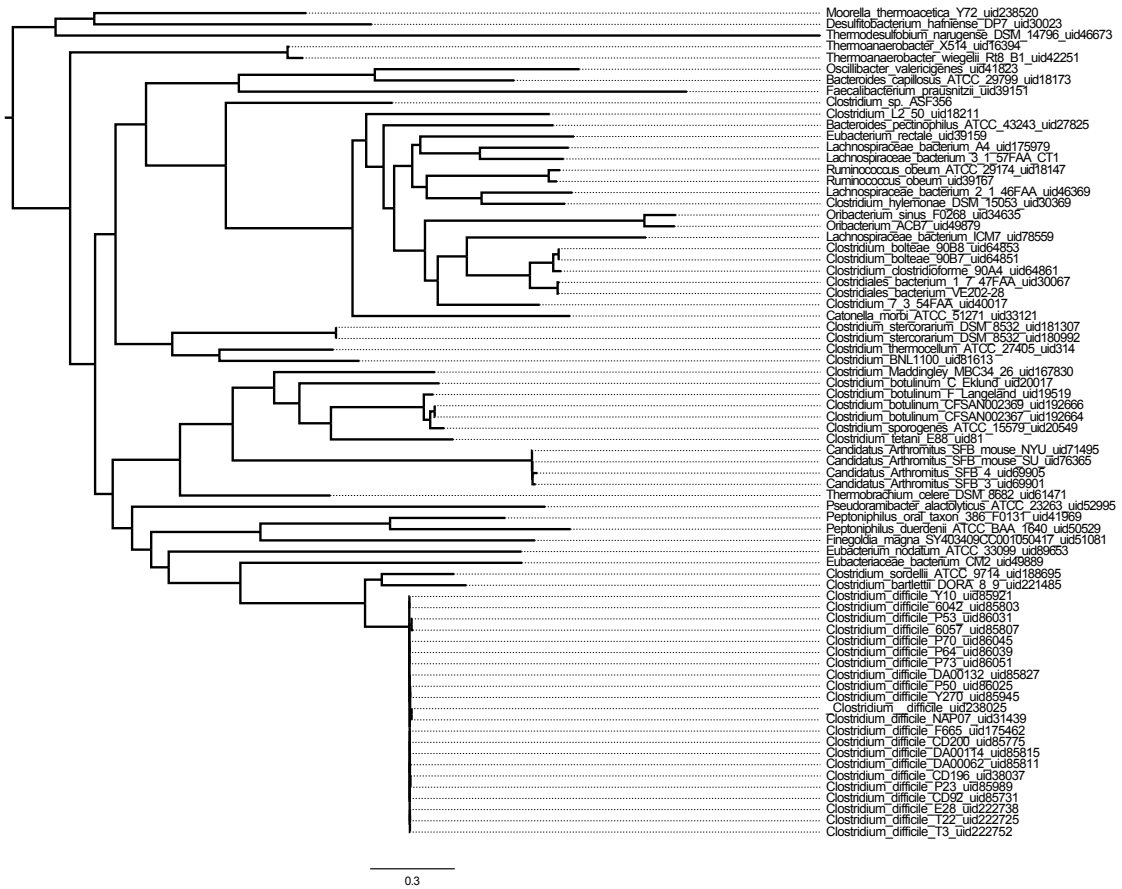
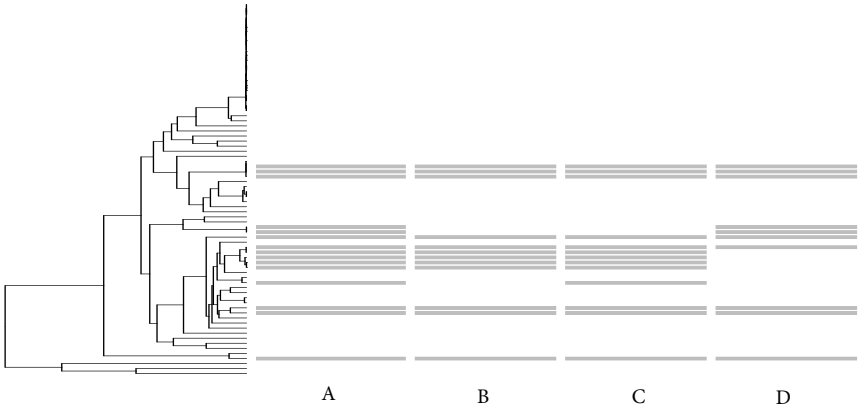
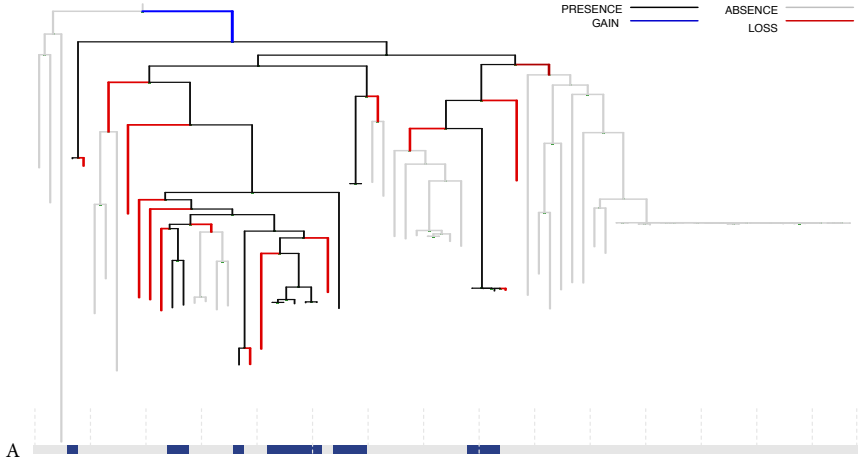


Figure A.1: Phylogenetic tree of 74 genomes used to build profiles, subsampled from the full tree of 687 genomes.

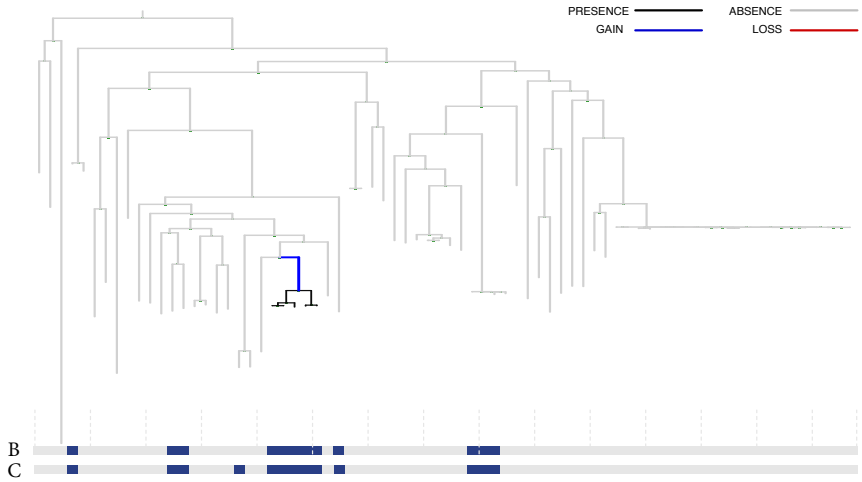
(a)



(b)



(c)



(d)

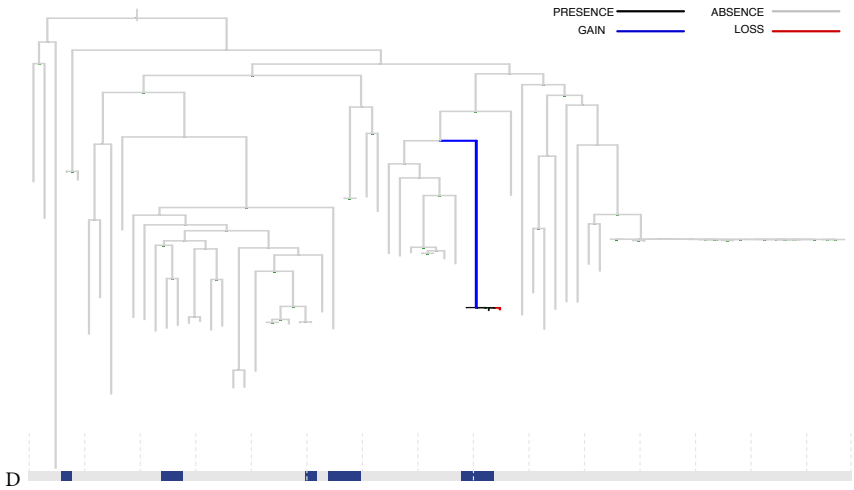
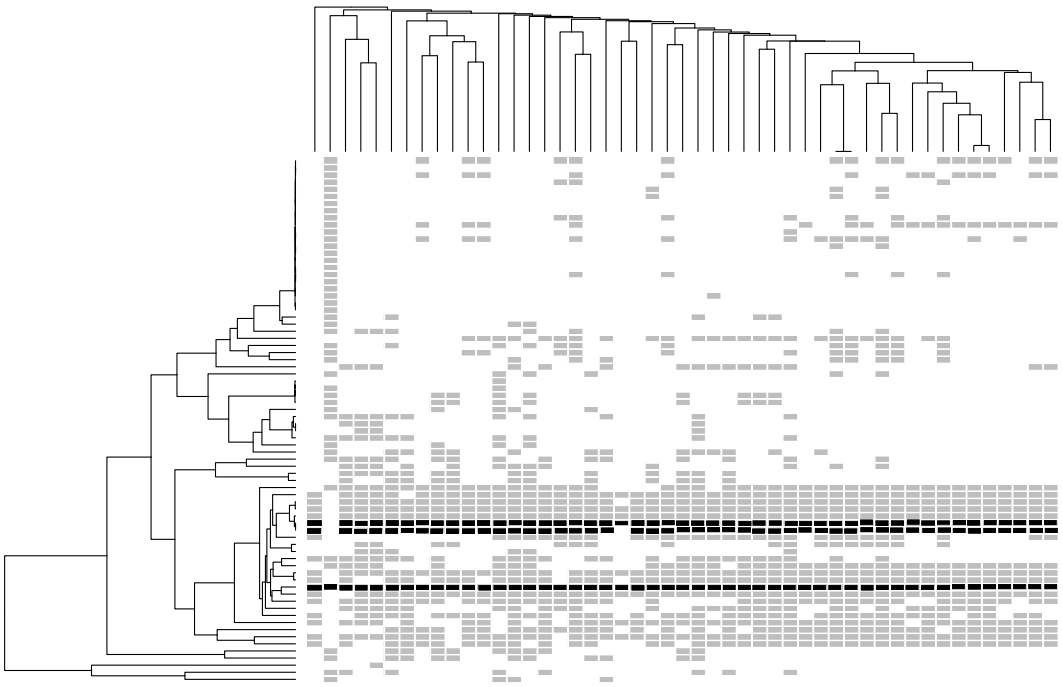
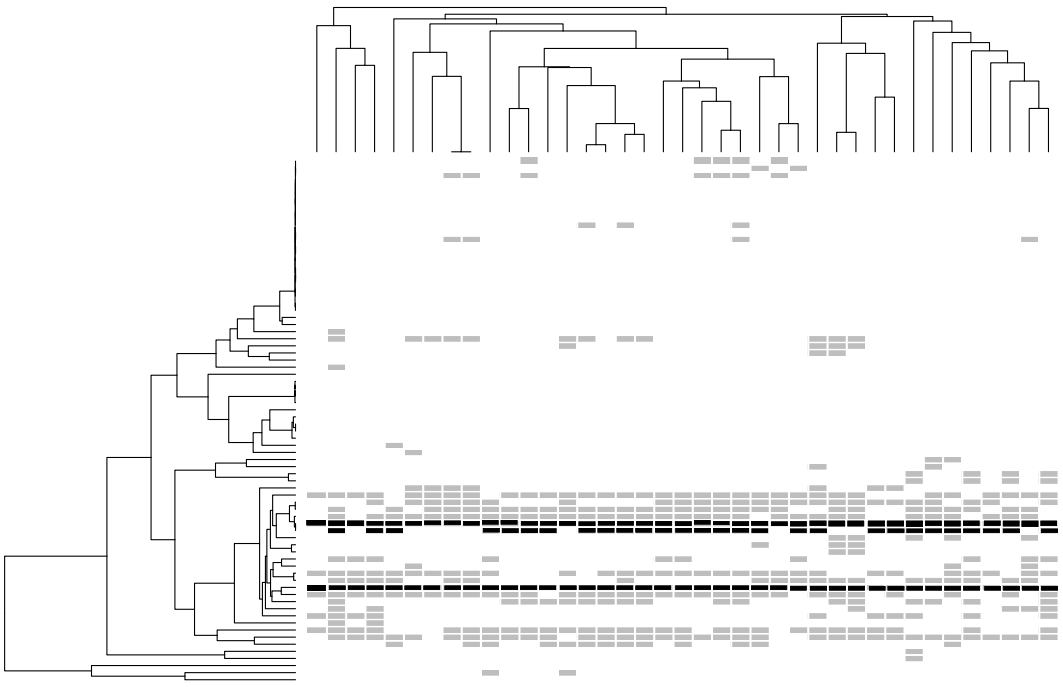


Figure A.2: A hierarchical cluster that is split into singletons by CLIME. (a) Patterns of presence and absence of four phylogenetic profiles across the 74 genomes in the phylogenetic tree. (b-e) Mapping of each gene to the reference tree by CLIME.In (b-e) figures, the tree is the phylogenetic tree of 74 genomes; the blue and red lines represent gene gain and loss respectively; In the profiles, the dark blocks represent the presence and the gray means absence.

(a)



(b)



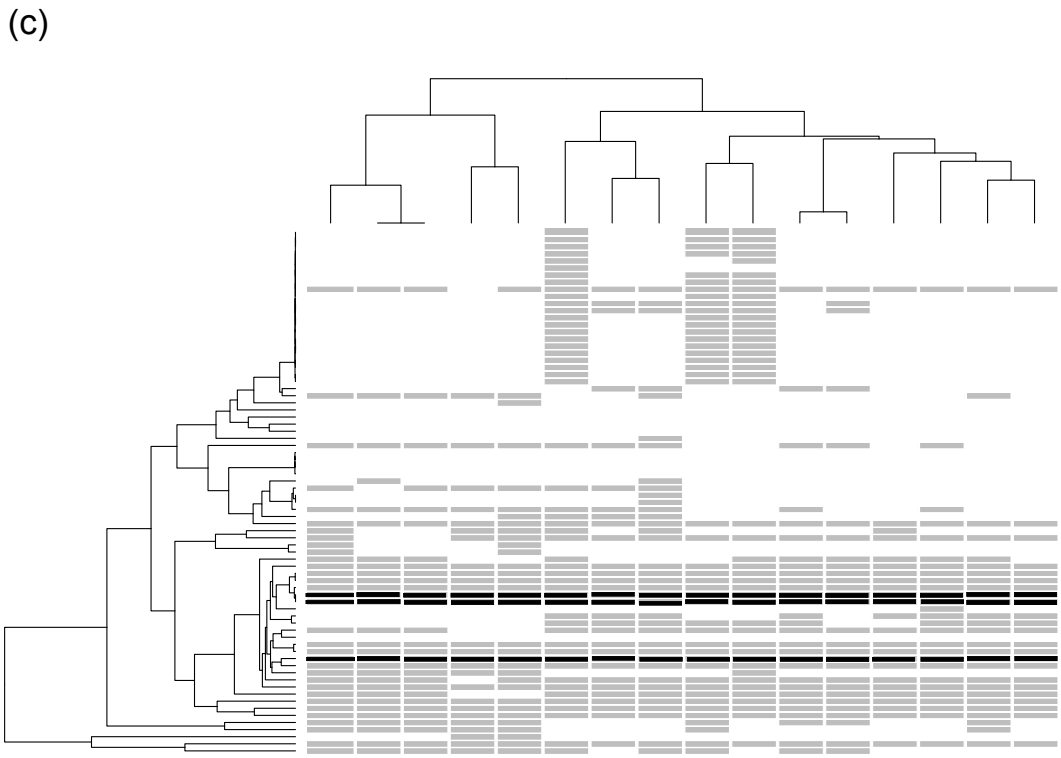
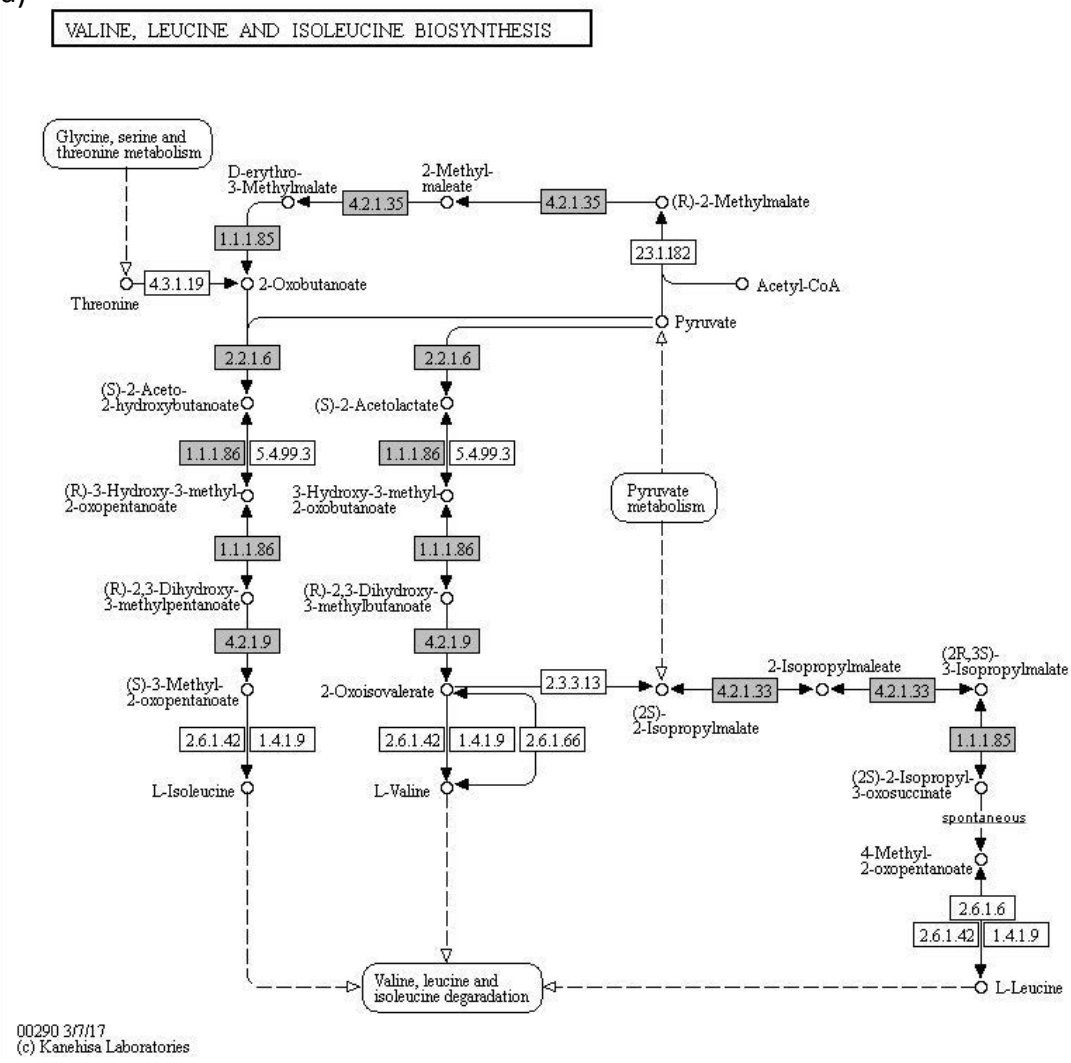


Figure A.3: Structure, phylogenetic distribution and functional categories of three clusters with significant over representation of the identified proteins to the two strains of *C. bolteae*. The three rows of black bars represent *C. bolteae* 90B7, *C. bolteae* 90B8 and LZ from top to bottom.

(a)


VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
D-erythro-3-Methylmalate
2-Methyl-maleate
(R)-2-Methylmalate
4.2.1.35
4.2.1.35
1.1.1.85
2.3.1.182
Acetyl-CoA
Threonine
4.3.1.19
2-Oxobutanoate
Pyruvate
2.2.1.6
2.2.1.6
(S)-2-Aceto-2-hydroxybutanoate
(S)-2-Acetolactate
1.1.1.86
5.4.99.3
1.1.1.86
5.4.99.3
(R)-3-Hydroxy-3-methyl-2-oxopentanoate
3-Hydroxy-3-methyl-2-oxobutanoate
Pyruvate metabolism
1.1.1.86
1.1.1.86
(R)-2,3-Dihydroxy-3-methylpentanoate
(R)-2,3-Dihydroxy-3-methylbutanoate
4.2.1.9
4.2.1.9
(S)-3-Methyl-2-oxopentanoate
2-Oxoisovalerate
2.3.3.13
2-Isopropylmaleate
4.2.1.33
4.2.1.33
(2R,3S)-3-Isopropylmalate
(2S)-2-Isopropylmalate
2.6.1.42
1.4.1.9
2.6.1.42
1.4.1.9
2.6.1.66
1.1.1.85
(2S)-2-Isopropyl-3-oxosuccinate
L-Isoleucine
L-Valine
spontaneous
4-Methyl-2-oxopentanoate
2.6.1.6
2.6.1.42
1.4.1.9
Valine, leucine and isoleucine degaradation
L-Leucine
00290 3/7/17
(c) Kanehisa Laboratories

(b)

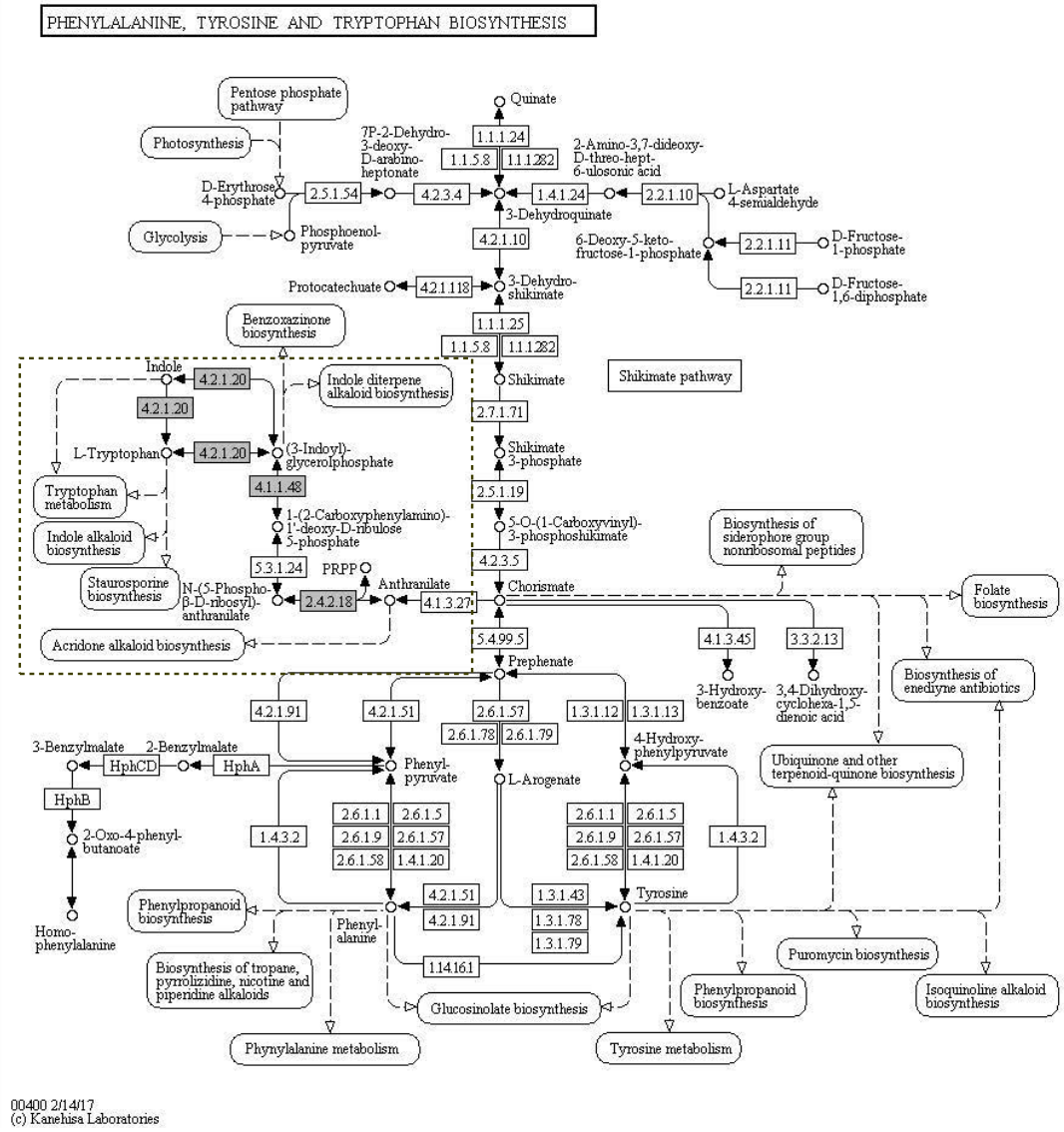


Figure A.4: The Amino-acid biosynthesis pathway map. The shaded boxes are the functions covered in the detected gene cluster shown in Figure 2.9. Source: KEGG PATHWAY (`https://www.genome.jp/`).