GENE CLUSTERING BASED ON CO-OCCURRENCE WITH CORRECTION FOR COMMON EVOLUTIONARY HISTORY

by

Chaoyue Liu

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

Table of Contents

List of Tables

List of Figures

Abstract

As the number of sequenced genomes increases rapidly, new approaches are needed for the computational annotation of protein functions and to better understand the ecological roles of genomes.

In this thesis, a gene clustering approach based on the correlated evolution method (Pagel) and hierarchical clustering is proposed to find sets of co-occurring genes according to their weighted phylogenetic profiles. Hierarchical clusters can be cut at many different levels of similarity; since our primary interest is the evaluation of functional associations, we used the semantic similarity of Gene Ontology terms to optimize the choice of cuts in the hierarchy, and to evaluate our clustering outcomes. The results can be used to predict the functions of the unannotated genes and to discover candidate sets of lateral gene transfer events.

We applied this approach to the gene set of the large clostridial genome "*Lachnospiraceae* bacterium 3-1-57FAA-CT1", and generated informative clusters of genes with correlated evolutionary histories, which in many cases shared functional similarity as well. We compared the results of our method to the recently described approach, Clustering by Inferred Modules of Evolution (CLIME), and found considerable similarity between the two sets of predictions. However, our hierarchical clustering approach allows the exploration of degrees of protein similarity, and the generation of smaller or larger clusters as appropriate. In both cases, we found strong evidence that clusters of genes having similar phylogenetic histories also tend to be functionally linked.

List of Abbreviations and Symbols Used

ARI	Adjusted Rand Index
CAFA	Critical Assessment of Functional Annotation
CLIME	Clustering by Inferred Models of Evolution
DAG	Directed acyclic graph
ECMs	Evolutionarily Conserved Modules
GO	Gene Ontology
HMM	Hidden Markov Model
IC	Information Content
Lb-CT1	*Lachnospiraceae* bacterium 3-1-57FAA-CT1
LGT	Lateral Gene Transfer
LR	likelihood ratio
RI	Rand Index
Uniprot	Universal Protein Resource

Acknowledgements

I would never have been able to finish my thesis without the guidance of my committee members, help from friends, and support from my family.

First and foremost I offer my sincerest gratitude to my supervisors, Dr.Hong Gu and Dr.Robert Beiko for the continuous support of my study. Their patience, motivation and immense knowledge helped me in all the time of research and writing of this thesis.

I would also like to acknowledge the rest of my thesis committee, Dr.Tobias Kenney and Dr.Joseph Bielawski and I am gratefully indebted to their insightful comments on this thesis.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general.

Chapter 1

Introduction

1.1 Background

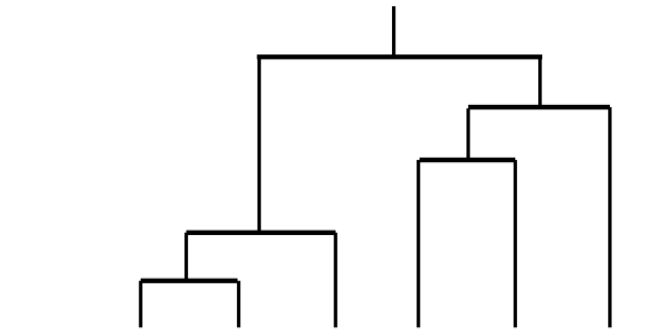
As the number of sequenced genomes increases rapidly, new approaches are needed for the computational annotation of protein functions and to better understand ecological roles of genomes [36, 27]. Gene clustering analysis can be a useful and fast tool for identifying biologically relevant groups of genes, which is the process of partitioning a given gene set into groups based on some specified features [44, 39, 35, 13].

The aim of our project is to cluster genes based on their evolutionary histories. We seek to group the genes with correlated occurrence patterns, which can be represented by the distribution of genes across a group of genomes, and such genes may also be involved in related biological processes. The hypothesis behind this approach is that functionally linked genes evolve in a correlated fashion, and should also present in the same genomes. A reasonable proposition is to compare phylogenetic profiles, which encode the presence and absence of genes across a set of genomes, as a basis for detecting correlated genes [36, 26, 20].

However, in a comparative analysis of phylogenetic profiles, to treat each species (genomes in our case) as the unit of analysis assumes that the traits under investigation (genes in our case) should evolve independently in each of the species [24]. 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 being independent [9, 11]. Therefore, if we want to use co-occurrence information of genes across many genomes to make evolutionary and functional inferences, we must develop approaches that take phylogenetic correlations into account.

Figure 1.1, which represents the distributions of 3 genes across 6 genomes, is an example to illustrate the effect of the phylogeny. If we want to calculate the similarity between phylogenetic profiles in the usual way, by considering 6 genomes

as independent units, Gene 2 and Gene 3 will have the same similarity with Gene 1, since they both co-occur with Gene 1 in exactly three genomes (Gene 2 in Genome 1, 2, 4; Gene 3 in Genome 3, 4, 6). However, if we hierarchically structure these 6 genomes with a tree, which represents the evolutionary relationships among them, the presences of genes in each genome should not be equally weighted, since closely related genomes tend to possess the same genes due to common descent. Since Genomes 1 and 2 are closely related, it is not surprising that Gene 1 and Gene 2 are found in both; conversely, the distribution of co-occurrence for Genes 1 and 3 could be viewed as more informative. So the objective of this project is to explore the removal of phylogenetic correlations in order to identify genes with strongly correlated distributions, and in addition, to build clusters for further examination.



	Genome 1	Genome 2	Genome 3	Genome 4	Genome 5	Genome 6
Gene 1	✓	✓	✓	✓		✓
Gene 2	✓	✓		✓		
Gene 3			✓	✓		✓

Figure 1.1: The distributions of 3 genes across 6 genomes: "$\surd$" indicates that the gene exists in this genome; the tree diagram shows the evolutionary relationships among 6 genomes.

In this thesis, a gene clustering approach based on the correlated evolution method of Pagel [24], and hierarchical clustering is proposed to find sets of co-occurring genes according to their phylogenetic profiles. The semantic similarity of Gene Ontology (GO) terms is also introduced as a criterion to find the optimal cutting height on the created hierarchical dendrogram and also to evaluate our clustering outcomes based

on functional similarity.

There are three kinds of information needed for this project: a phylogenetic tree, the phylogenetic profiles and the Gene Ontology terms of the target genes. The phylogenetic tree provides the distance between genomes, based on the similarities and differences in their genetic characteristics [34]. The phylogenetic profiles summarize the presence and absence of genes across the genomes on the phylogenetic tree. The phylogenetic tree and profiles are the input data of the Pagel method to calculate the evolutionary similarities between genes, from which the distance matrix used in the hierarchical clustering method is derived [24]. The information about gene ontologies are used to find the optimized height cutting on the hierarchical clustering dendrogram, and also to evaluate our clustering outcomes.

The next parts of this chapter will introduce phylogenetic trees and phylogenetic profiles in greater detail, and present an outline of the remainder of the thesis.

1.2 Phylogenetic Tree

The phylogenetic tree represents the evolutionary relationships among organisms, species or genomes (in our case), which is produced on the basis of sequenced genes or genomic data [40]. Several terms are used to describe features of phylogenetic trees:

- *The root*, if present, represents the common ancestor of all entities in the tree.

- *Nodes* in the tree can be *terminal* if they represent observations, or *internal* if they represent the ancestor of two or more observed entities.

- *Branch lengths* represent the amount of evolutionary change over time.

Take the simple phylogenetic tree in Figure 1.2 as an example, there are 3 existing species denoted as A, B and C; the numbers on top of each branch are the branch length which is usually expressed as the average number of nucleotide or amino-acid substitutions per site.

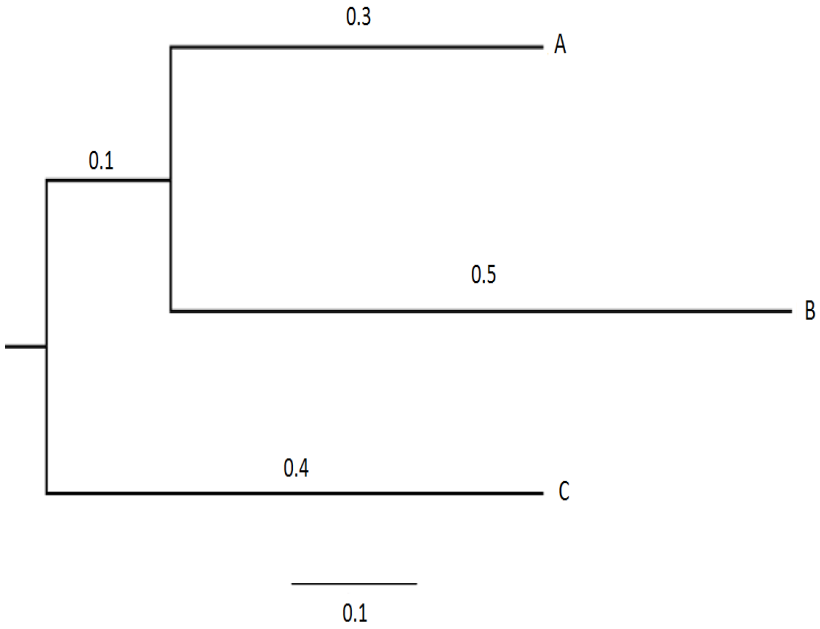


Figure 1.2: An example phylogenetic tree with 3 genomes: A, B, C are 3 existing species; the numbers represent the length of the branches; the line segment at the bottom is the distance scale for genetic change.

1.3 Phylogenetic Profiling

Phylogenetic Profiling is a bioinformatics technique to describe a gene's distribution across a set of genomes on the basis of homology information. The protein encoded by each gene in a fully sequenced genome can be assigned to a specific set of proteins based on homology relationships. The presence or absence of this protein across genomes can then be represented by its phylogenetic profile. So the phylogenetic profile can be considered as a long binary sequence encoding the presence or absence of the gene with digits corresponding to a given set of genomes, as Table 1.1 shows.

The theory of using phylogenetic profiles to predict functional relationships between genes is that functionally linked proteins tend to evolve in a correlated pattern [26]. The genetic changes between related species can be affected by many factors,

	GI:488233145	GI:488629332	GI:488629959
Eubacterium rectale DSM 17629	0	0	0
Faecalibacterium prausnitzii SL3	1	0	0
Ruminococcus obeum A2-162]	1	1	1
Thermincola potens JR	0	0	0
Peptoclostridium difficile NAP07	0	0	1
Clostridium sp. 7_3_54FAA	1	0	1

Table 1.1: Examples of Phylogenetic Profile: the column labels are the names of genomes; the row labels represent the GI numbers of genes; each column is a binary vector which represents the presence (1) and absence (0) of genes.

including gene transfer and gene loss, and proteins that are involved in similar biological processes may be gained and lost together, leading to similar phylogenetic profiles. Since the species in a phylogenetic profile are not actually independent, especially closely related species which are expected to have similar sets of genes, we must also take phylogenetic correlations into account when inferring the similarity of profiles.

1.4 Thesis Structure

The remainder of this thesis is organized into 5 chapters. Chapter 2 describes the Pagel method and hierarchical clustering. Chapter 3 introduces the definition of GO, the semantic similarity of GO terms and how to evaluate the clustering outcomes. The application on the gene set of "*Lachnospiraceae* bacterium 3-1-57FAA-CT1" is given in Chapter 4, and more specific results are also illustrated in this chapter. Chapter 5 introduces the other evolution based approach, CLIME and includes the comparison of the results from two approaches. Finally, the conclusion of this thesis and the ideas for future work are given in Chapter 6.

Chapter 2

Clustering of Genes based on Evolutionary Correlation

In this chapter, we will introduce the Pagel method, which was developed to test for correlated evolutionary change in traits (in our case, genes) on a given phylogenetic tree [24]. Then we implement the hierarchical clustering on the evolutionary correlation of genes obtained from Pagel's method to generate gene clusters, and the optimized tree cutting on the hierarchical dendrogram is also discussed more specifically.

2.1 Estimating Evolutionary Correlation Among Genes

Pagel (1999) developed a statistical method for identifying significant evolutionary correlations between two discrete characters across a phylogenetic tree. To characterize the evolution across a phylogenetic tree, two continuous-time Markov models are built: one model where 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. Then, the tree pruning algorithm is introduced to simplify the calculation of likelihoods, and the hypothesis of correlated evolution is eventually tested by comparing the fit of two different models to the observed data set [24].

2.1.1 Continuous-time Markov Process Assumptions

A continuous-time Markov model is used to measure the evolutionary changes between two characters along the branches of a phylogenetic tree. In our case, the length of the branch represents the evolutionary time t, and the discrete characters are the genes, which have two states at any interior node on a given phylogenetic tree: 0 (absence) and 1 (presence). A continuous-time Markov process has the following features [23, 33, 29]:

- Markov property

Given the time point t, and time interval h, a continuous-time Markov process $\{X(t), t \geq 0\}$ has the Markov property that,

$$P[X(t+h) = j | X(t) = i, X(s) = x(s) \text{ for } 0 \leq s \leq t] = P[X(t+h) = j | X(t) = i]. \tag{2.1}$$

The Markov property implies that given all the states of the process prior to time t, $X(t+h)$ only depends on the most recent $X(t)$.

- Transition Probability

  Given the time interval $h \to 0$, the probability of transition for a continuous-time Markov chain is defined as,

$$P_{ij}(h) = hq_{ij} + o(h) \tag{2.2}$$

$$P_{ii}(h) = 1 - hv_i + o(h) \tag{2.3}$$

  where $q_{ij} = \lim_{t \to 0} \frac{P_{ij}(t)}{t}$ is the transition rate of changing from state i to state j, and v_i is the rate at which the process leaves state i, which is the summation of the rates of leaving to all other states, $v_i = \sum_{j \neq i} q_{ij}$.

  Consider a simple Markov process with two states i and j, we have the following equality,

$$P_{ij}(t + dt) = P_{ii}(t) q_{ij} dt + P_{ij}(t)(1 - q_{ji} dt) + o(dt), \tag{2.4}$$

  The meaning of Equation (2.4) is that the character can change its state from i to j in two different ways, 1) stay in state i before time t, then transition to state j in a small time interval dt; 2) transition to state j in time t, and stay in state j in following dt.

- Solution of $P(t)$

  To solve $P(t)$, we use Kolmogorov's Backward Equation, which is a more general form of Equation (2.4),

$$\frac{\mathrm{d}P_{ij}(t)}{\mathrm{d}t} = \sum_{k \neq i} q_{ik} P_{kj}(t) - v_i P_{ij}(t). \tag{2.5}$$

Or, in the matrix form,

$$P(t+dt) = P(t)(I + Qdt) \tag{2.6}$$

where Q is the matrix of transition rates with entry $Q_{ij} = \begin{cases} q_{ij}, i \neq j \\ -v_i, i = j \end{cases}$. If we take the derivative respect to t, we have

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \frac{P(t+dt) - P(t)}{dt} = P(t)Q \tag{2.7}$$

which gives the solution,

$$P(t) = e^{Qt} \text{ , where } e^{Qt} = \sum_{k=0}^{\infty} Q^k t^k / k! \tag{2.8}$$

Since the time t is given, the transition probability $P(t)$ is only determined by the transition rates in Q. Under the Markov process assumption, the probability of change from state i to state j in a branch, depends only on the state at the beginning of the branch, and the time t which is represented by the length of the branch [24, 9].

When considering the evolution of a pair of characters at the same time, two characters can evolve in either an independent or correlated fashion. We first fit a Markov model to the data in which two characters are assumed to evolve independently, and then compare the goodness of fit to a more complicated model which allows for the correlated change. If the model of correlated evolution fits the data significantly better than the independent evolution model, there is evidence of correlated evolution of the two characters.

Independent Evolution Model

In this model, the states (presence or absence) of two discrete characters A and B are assumed to change independently, which means that variable B evolving across a given phylogenetic tree would not be affected by the state of variable A. Since the two characters evolve independently, their transition rates are also independent, and only 4 rate parameters (Table 2.1) are needed in this model to represent all possible state transitions.

There is also the dual transition scenario, where both characters change at the same time. For example, the transition rate for the case where A and B both changed

Parameter	Character	Transition of state
α_1	A	$0 \to 1$
α_2	A	$1 \to 0$
β_1	B	$0 \to 1$
β_2	B	$1 \to 0$

Table 2.1: 4 parameters in the independent model

from 0 to 1, would be simply the product of α_1 and β_1, since the two characters evolve independently. In the matrix form, the transition rates look like below.

$$\begin{array}{c} (\text{A,B}) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} - & \beta_1 & \alpha_1 & \alpha_1\beta_1 \\ \beta_1 & - & \alpha_1\beta_2 & \alpha_1 \\ \alpha_2 & \beta_2\alpha_1 & - & \beta_1 \\ \alpha_2\beta_2 & \alpha_2 & \beta_2 & - \end{pmatrix} \end{array}$$

In Pagel's paper [24], all the dual state transition rates are set to zero, since a dual transition can be decomposed into two successive events, for example, a dual transition $(0,0) \to (1,1)$ can be reached via two transitions,$(0,0) \to (0,1)$, and then $(0,1) \to (1,1)$. It is also possible based on the properties of the Markov model, that according to Equation (2.2), the probability of this dual state transition happens, is $o(h) \to 0$, so the transition matrix can be simplified as below,

$$\begin{array}{c} (\text{A,B}) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} - & \beta_1 & \alpha_1 & 0 \\ \beta_2 & - & 0 & \alpha_1 \\ \alpha_2 & 0 & - & \beta_1 \\ 0 & \alpha_2 & \beta_2 & - \end{pmatrix} \end{array}.$$

Dependent Evolution Model

In the dependent evolution model, A and B are assumed to interact in evolution, resulting in correlated transition rates. In other words, the current state of character A can impact the transition change of character B. There are 8 parameters instead of 4 parameters (independent model) to represent all possible state transitions (Table 2.2).

Parameter	Character	Dependence	Transitions
q_{12}	B	A = 0	$0 \to 1$
q_{13}	A	B = 0	$0 \to 1$
q_{21}	B	A = 0	$1 \to 0$
q_{24}	A	B = 1	$0 \to 1$
q_{31}	A	B = 0	$1 \to 0$
q_{34}	B	A = 1	$0 \to 1$
q_{42}	A	B = 1	$1 \to 0$
q_{43}	B	A = 1	$1 \to 0$

Table 2.2: 8 parameters in the dependent model

In the matrix form,

$$\begin{array}{c} \text{(A,B)} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} - & q_{12} & q_{13} & 0 \\ q_{21} & - & 0 & q_{24} \\ q_{31} & 0 & - & q_{34} \\ 0 & q_{42} & q_{43} & - \end{pmatrix} \end{array}$$

Since the dependent model has more parameters than the independent model, it is trivially expected to fit better and have the greater likelihood. The likelihood ratio test is used to assess the significance of the increase in likelihood, given the increased number of parameters.

2.1.2 Likelihood Calculation

Figure 2.1 is an example of a simple phylogenetic tree, where $S = \{s_1, s_2, ...s_9\}$ denote the states on each node and $T = \{t_1, t_2, ...t_8\}$ represent the branch lengths. To represent the likelihood function, we define $P(s_i, s_j, t_j)$ as the probability that a branch begins in state s_i and ends in state s_j in time period t_j. Considering the evolution of a character, A, across this phylogenetic tree, given all the states of each tip and node on the tree in Figure 2.1, the likelihood of this particular realization is the product over all branches of the tree with the probabilities derived from the continuous-time Markov process (in Section 2.1.1),

$$\begin{aligned} L(A|S,T) =& P(s_9, s_8, t_8)P(s_8, s_3, t_3)P(s_8, s_7, t_7)P(s_7, s_5, t_5) \\ & \cdot P(s_7, s_4, t_4)P(s_9, s_6, t_6)P(s_6, s_1, t_1)P(s_6, s_2, t_2) \end{aligned} \tag{2.9}$$

The overall likelihood for A, is the summation of the likelihood in Equation (2.9) of all possible assignments of states on each interior node,

$$\begin{aligned} L(A) = &\sum_{s_9=0}^{1}\sum_{s_8=0}^{1}\sum_{s_7=0}^{1}\sum_{s_6=0}^{1} P(s_9,s_8,t_8)P(s_8,s_3,t_3)P(s_8,s_7,t_7) \\ &\cdot P(s_7,s_5,t_5)P(s_7,s_4,t_4)P(s_9,s_6,t_6)P(s_6,s_1,t_1)P(s_6,s_2,t_2) \end{aligned} \tag{2.10}$$

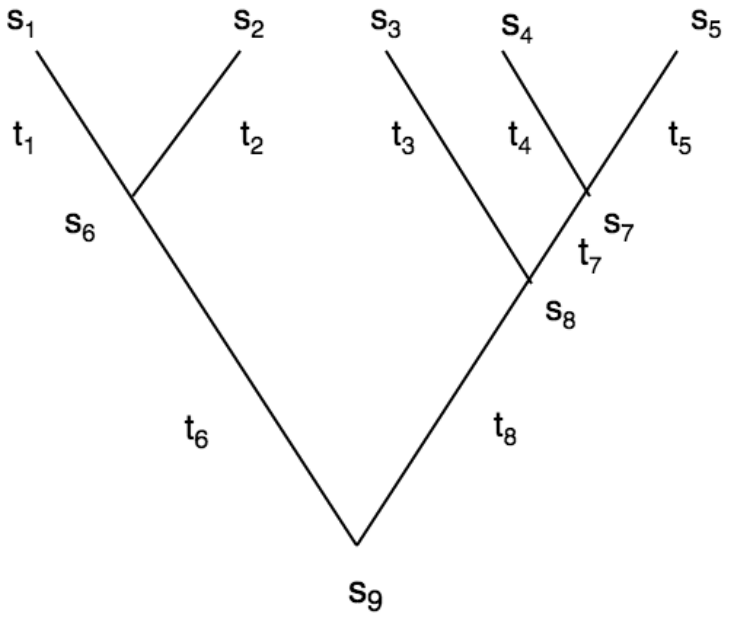


Figure 2.1: A phylogenetic tree of five species adapted from Pagel (1999) [24]: s denotes the state on each node, and t represents the time period on each branch.

It is the same procedure to find the likelihood of B, and the likelihood for the independent evolution model is simply the product of A and B,

$$\begin{aligned} L_{Independent}(A,B) = &\prod^{\text{A and B}} \Big[\sum_{s_9=0}^{1}\sum_{s_8=0}^{1}\sum_{s_7=0}^{1}\sum_{s_6=0}^{1} P(s_9,s_8,t_8) \\ &\cdot P(s_8,s_3,t_3)P(s_8,s_7,t_7)P(s_7,s_5,t_5)P(s_7,s_4,t_4) \\ &\cdot P(s_9,s_6,t_6)P(s_6,s_1,t_1)P(s_6,s_2,t_2)\Big]. \end{aligned} \tag{2.11}$$

In the dependent evolution model, changes in these two characters A and B are correlated, and their joint changes can be described by 4 states, which are denoted

as 1 = (0,0); 2 = (0,1); 3=(1,0); 4 = (1,1). So for the phylogenetic tree in Figure 2.1, each node has 4 possible states, except the tips $(s_1, s_2, ...s_5)$, which are observed in the phylogenetic profiles. So the likelihood for the dependent evolutionary model is given by,

$$\begin{aligned} L_{Dependent}(A,B) = &\sum_{s_9=1}^{4}\sum_{s_8=1}^{4}\sum_{s_7=1}^{4}\sum_{s_6=1}^{4} P(s_9,s_8,t_8) \\ &\cdot P(s_8,s_3,t_3)P(s_8,s_7,t_7)P(s_7,s_5,t_5)P(s_7,s_4,t_4) \\ &\cdot P(s_9,s_6,t_6)P(s_6,s_1,t_1)P(s_6,s_2,t_2) \end{aligned} \tag{2.12}$$

2.1.3 Pruning Algorithm

Directly calculating the above likelihoods can be computationally expensive, especially when the phylogenetic tree is large. Take the phylogenetic tree in Figure 2.1 for example, there are 5 leaves $(s_1, s_2, s_3, s_4, s_5)$ and 4 interior nodes (s_6, s_7, s_8, s_9). The state of each leaf is observed, but for each interior node, there are 4 possible states in the dependent evolution model, and 4^4 possible terms in Equation (2.12). So if given a phylogenetic tree with n leaves, there would be $(n-1)$ interior nodes and we need $4^{(n-1)}$ terms for calculating the likelihood. The cost of computing is exponential in n, which is technically impossible for a big tree (even for n = 11, $4^{10} = 1,048,576$).

Felsenstein (1981) developed a computationally feasible method for calculating the likelihood over a phylogenetic tree, which is called the "pruning" algorithm [8]. The pruning algorithm moves the summations to the right in the equation to reduce the computing task. One assumption of the pruning algorithm is that the evolution over different branches on the tree is independent [8]. So for instance, in Figure 2.1, the node s_6 on the left branch would not be affected by the nodes on the right branch but only the root s_9, and the pattern of parentheses for these 5 tips can be expressed as $(s_1, s_2)(s_3, (s_4, s_5))$. Therefore, given the state of the root s_9, the likelihood for the left branch can be written as,

$$\sum_{s_6=1}^{4} P(s_9,s_6,t_6)P(s_6,s_2,t_2)P(s_6,s_1,t_1). \tag{2.13}$$

And in the same way, the likelihood for the right branch is

$$\sum_{s_8=1}^{4} P(s_9,s_8,t_8)P(s_8,s_3,t_3)(\sum_{s_7=1}^{4} P(s_7,s_4,t_4)P(s_7,s_5,t_5)) \tag{2.14}$$

Since the left branch is independent of the right, the likelihood of both branches is the simple product of Equation (2.13) and 2.14, and also because the root s_9 is the base, it is finally added to the front,

$$\begin{aligned}\sum_{s_9=1}^{4}(\sum_{s_6=1}^{4} & P(s_9,s_6,t_6)P(s_6,s_2,t_2)P(s_6,s_1,t_1)\\ & \cdot(\sum_{s_8=1}^{4} P(s_9,s_8,t_8)P(s_8,s_3,t_3)\\ & \cdot(\sum_{s_7=1}^{4} P(s_7,s_4,t_4)P(s_7,s_5,t_5)))).\end{aligned} \tag{2.15}$$

After applying the pruning algorithm, Equation (2.11) and (2.12) can be rewritten as

- Independent evolution model:

$$\begin{aligned}L_{Independent}(A,B) = \prod^{\text{A and B}} \Big[\sum_{s_9=0}^{1}[(\sum_{s_8=0}^{1} & P(s_9,s_8,t_8)P(s_8,s_3,t_3)\\ & \cdot\sum_{s_7=0}^{1} P(s_8,s_7,t_7)P(s_7,s_5,t_5)P(s_7,s_4,t_4))\\ & \cdot\sum_{s_6=0}^{1} P(s_9,s_6,t_6)P(s_6,s_1,t_1)P(s_6,s_2,t_2))]\Big]\end{aligned} \tag{2.16}$$

- Dependent evolution model:

$$\begin{aligned}L_{Dependent}(A,B) = \sum_{s_9=1}^{4}\Big[(\sum_{s_8=1}^{4} & P(s_9,s_8,t_8)P(s_8,s_3,t_3)\cdot\\ & \sum_{s_7=1}^{4} P(s_8,s_7,t_7)P(s_7,s_5,t_5)P(s_7,s_4,t_4))\\ & \cdot\sum_{s_6=1}^{4} P(s_9,s_6,t_6)P(s_6,s_1,t_1)P(s_6,s_2,t_2))\Big]\end{aligned} \tag{2.17}$$

2.1.4 Statistical Test for Dependence of Traits

A likelihood ratio test is used for the model selection between the independent evolution model and the dependent evolution model. Under the assumption of the null

hypothesis that the two characters evolve independently, the null model (independent evolution model) is actually a special case of the alternative model (dependent evolution model), because the null model can also be reformed into the model with eight parameters but four pairs of the parameters are the same. For example, in Table 2.2, $q_{12} = q_{34}$ since the state transition of B would not be influenced by the current state of A when assumption that A and B are independent is true. In the same way, we have the following equalities:

- $q_{12} = q_{34}$ for B: $0 \rightarrow 1$; $q_{21} = q_{43}$ for B: $1 \rightarrow 0$
- $q_{13} = q_{24}$ for A: $0 \rightarrow 1$; $q_{31} = q_{42}$ for A: $1 \rightarrow 0$

So the independent evolution model and the dependent evolution model are two nested models, which satisfies the requirements of the likelihood ratio test condition. The dependent evolution model is a more complex model with more parameters, which is guaranteed to fit the data at least as well as the simpler model, and will almost certainly have a higher likelihood. The likelihood ratio statistics is calculated as follows:

$$\begin{aligned} LR &= -2ln(L_{H_0}/L_{H_1}) \\ &= -2ln(L_{Indepedent}(A,B)/L_{Dependent}(A,B)). \end{aligned} \tag{2.18}$$

The likelihood ratio follows a χ^2 distribution, with degrees of freedom equal to the difference in the number of parameters, which in our model is $8 - 4 = 4$. The resulting P-value expresses the evolutionary relatedness between two characters, and a smaller P-value suggests a stronger evidence for correlated evolution.

2.2 Clustering of Genes

In this section, we will use the hierarchical clustering method to group phylogenetic profiles based on their similarity, which is the evolutionary dependency of genes given by Pagel's method, and to discover the functional connections among genes. We also discuss the way of optimizing the tree cutting on the hierarchical dendrogram to discover more cohesive clusters.

2.2.1 Review of Hierarchical Clustering

Hierarchical clustering is a commonly used method of cluster analysis, which groups data by creating a binary tree or dendrogram. Hierarchical clustering is also a good method for exploring the gene clusters because the hierarchical dendrogram provides visual information about the relative similarity of gene clusters, and the clusters obtained can be adjusted according to a threshold parameter (height).

Hierarchical clustering generally includes two types of strategies:

- Agglomerative clustering is a "bottom-up" approach, which starts with every data point in its own cluster and then merges sets of clusters with the smallest dissimilarity progressively until all data points are in one cluster.

- Divisive clustering is a "top-down" approach, which starts with all data points in one cluster and splits clusters into smaller groups with maximal dissimilarity progressively until all data points are in individual clusters.

The agglomerative method, which is the method we used in this project, is more popular than the divisive method, since the divisive method has the problems of selecting a cluster to split and finding the optimal sub-division of the chosen cluster, and has the exponential time complexity $O(2^n)$, whereas the agglomerative method has the polynomial complexity $O(n^3)$ [16, 7]. So here, we only introduce the procedure of the agglomerative hierarchical clustering in detail. The algorithm of the agglomerative clustering can be described as:

- Assign each data point into an individual cluster.

- Find the pair of clusters with the shortest distance, and merge them into a single cluster.

- Calculate the distances between each of the old clusters and the new cluster.

- Iteratively merge two closest clusters, until all data points are represented in a single cluster.

In the step of merging the closest pair of clusters, the measure of dissimilarity between sets of data points, which is called the linkage, is required to decide which

clusters should be combined first. Given the pairwise dissimilarities d_{ij} between data points, the linkage can be regarded as a function $d(G, H)$ which determines how the distance between two clusters G and H is measured. Commonly used options include [15, 1, 16]:

- single linkage

  The single linkage is also known as the closest-neighbor linkage, and the distance between two clusters G and H is the distance between the nearest neighbors in each cluster,

  $$d_{single}(G, H) = \min_{i \in G, j \in H} d_{ij}.$$

- complete linkage

  The complete linkage is also called the furthest-neighbor linkage, and the distance between G,H is the maximum distance between the data points of each cluster,

  $$d_{complete}(G, H) = \max_{i \in G, j \in H} d_{ij}.$$

- average linkage

  The distance between two clusters is the average of the distances between all the points in each cluster,

  $$d_{average}(G, H) = \frac{1}{n_G \cdot n_H} \sum_{i \in G, j \in H} d_{ij},$$

  where n_G and n_H are the sizes of clusters G and H.

- Ward linkage

  The Ward linkage uses the incremental sum of squares, which will merge the two clusters with the minimum increase in the total within-cluster sum of squares, and the linkage function is defined as,

  $$d_{ward}(G, H) = \frac{n_G \cdot n_H}{n_G + n_H} d^2(\hat{g}, \hat{h})$$

  where the $\hat{g}$, $\hat{h}$ is called the "clustroid", which is the point that has the smallest sum of squares of distances to other points,

  $$\hat{g} = \underset{g \in G}{argmin} \sum_{i \in G} d^2(i, g),\ \hat{h} = \underset{h \in H}{argmin} \sum_{j \in H} d^2(j, g).$$

These approaches use different strategies to find the clusters with some featured properties; for example, the single linkage method adopts a "friends of friends" clustering strategy, which means that in order to merge two clusters, only one pair of data points needs to be close. By contrast, the complete linkage method finds clusters based on the worst-case dissimilarity among pairs. We will choose the Ward method, which aims to find compact and spherical clusters, to perform the hierarchical cluster analysis on our data in Chapter 4. We tried other methods as well, but the Ward method tends to find small-sized clusters which are more appropriate to make the predictions, and also generated the largest number of informative clusters.

Table 2.3 is the matrix of Euclidean distances between the data points in Figure 2.2, and Table 2.4 shows the procedure of the agglomerative hierarchical clustering on these points step by step. Figure 2.3 is the final results of the dendrogram using the Ward linkage, and the height on the "y-axis" equals to the distance at which the clusters are merged.

Table 2.3: The matrix of Euclidean distances between the points in Figure 2.2.

	a	b	c	d	e	f
a	0	0.71	5.66	3.61	4.25	3.91
b	0.71	0	4.95	2.92	3.54	3.20
c	5.66	4.95	0	2.24	1.41	1.80
d	3.61	2.92	2.24	0	1.41	1.80
e	4.25	3.54	1.41	1.00	0	0.50
f	3.91	3.20	1.80	1.12	0.50	0

Table 2.4: The steps of the agglomerative clustering on the data set in Figure 2.2.

Step 1:	{a}, {b}, {c}, {d}, {e}, {f}
Step 2:	{a}, {b}, {c}, {d}, {e, f}
Step 3:	{a, b}, {c}, {d}, {e, f}
Step 4:	{a, b}, {c}, {d, e, f}
Step 5:	{a, b}, {c, d, e, f}
Step 6:	{a, b, c, d, e, f}

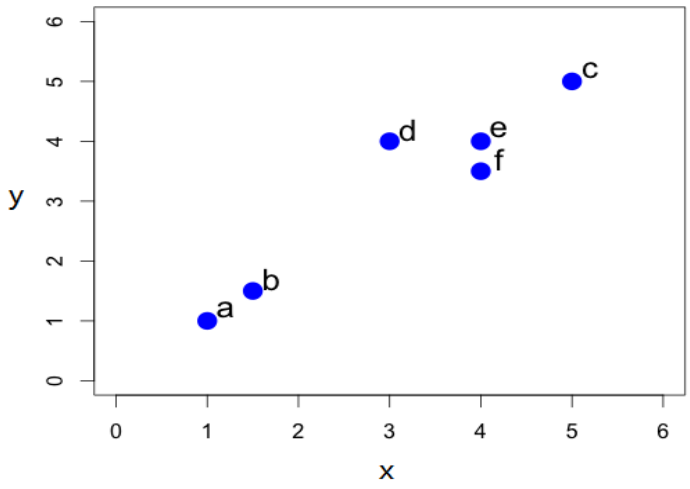


Figure 2.2: A simple example of the agglomerative hierarchical clustering with six data points.

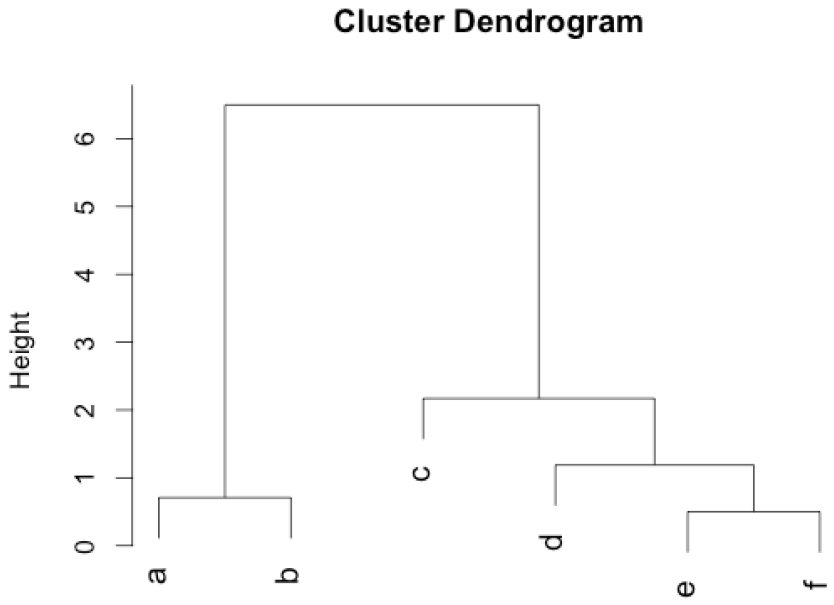


Figure 2.3: The hierarchical clustering dendrogram using the Ward linkage.

2.2.2 Optimized Tree Cutting

Hierarchical clustering is represented as a dendrogram and we can obtain a set of clusters by cutting the dendrogram at a certain height. However, the final cutting

height is not decided since cutting at a different height will give a different set of clusters, so that it is necessary to find an optimized height, and we can use external information to guide the cutting strategy.

In our application, we can measure the functional similarity among genes in clusters to develop an optimal cutting strategy using GO terms in Chapter 3. We can first use the cohesiveness to score clusters and the weighted mean score of a set of clusters at a certain height can be used to evaluate the performance of cutting at this height.

So for finding the appropriate height, we first calculated the weighted mean score of the set of clusters generated by cutting at each height, and chose the height with the best score. Figure 2.4a shows an example of fixed height tree cutting on a simple dendrogram, with several candidate cutting heights and a globally optimal height.

However, fixed-height tree cutting ignores the possibility that maximally cohesive clusters may only be obtained by cutting different branches at different heights. Dynamic tree cutting uses a similar approach to find the best height for each branch, rather than for the entire tree. So dynamic tree cutting may give different heights, and each cut is the best height at that branch with the maximum score. Figure 2.4b is an example of dynamic tree cutting, every final tree cutting (solid line segments) should be the best position of cutting its branch, and the final solution of tree cutting consists of the optimized heights for all branches.

Figure 2.4: An example of tree cutting: Figure (a) shows the fixed height tree cutting, where the dotted lines are possible examples of cuttings and the solid line is the best cutting with the maximum weighted mean of scores; Figure (b) shows dynamic height cutting, where every solid line segment is the optimized height cutting at its branch.

Chapter 3

Evaluation of Gene Clusters

Since we have described the clustering approach in the previous chapter, in this chapter we will develop a framework to evaluate the generated gene clusters for assessing the performance of our approach. Under the hypothesis that the genes with similar phylogenetic histories also tend to be involved in the related biological processes, we use the Gene Ontology framework for assessment. GO is a widely used classification scheme that was used in the Critical Assessment of Functional Annotation (CAFA) large-scale evaluation experiment [27].

3.1 Introduction to Gene Ontology

The Gene Ontology project, founded in 1998, is a collaborative effort of bioinformatics resource integration, which aims to provide structured, controlled vocabularies for the consistent descriptions of genes and gene products [6]. The Gene Ontology Consortium began as a joint project of three organism databases: FlyBase (Drosophila), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). Since then, the GO consortium keeps growing and expanding more major repositories for plant, animal, and microbial genomes [2].

The GO project contains three domains of ontologies - biological processes, cellular components and molecular functions [2]:

- Biological Process: operations or sets of molecular events with a defined beginning and end.
- Cellular Component: the parts of a cell or its extracellular environment.
- Molecular Function: the elemental activities of a gene product at the molecular level.

Since we expect that genes with correlated evolutionary patterns will often be evolved in related biological processes, we will use the GO biological process terms

for the assessment of gene clusters. Some examples of biological process GO terms are "DNA-templated transcription initiation [GO:0006352]", "carbohydrate metabolic process [GO:0005975]" and "ATP synthesis coupled electron transport [GO:0042773]".

GO is structured as a directed acyclic graph (DAG) where each GO term has defined relationships to one or more other terms, and there are two basic semantic relations between terms, "is a" and "part of" (Figure 3.1, for example). If we define A "is a" B, we mean that node A is a subtype of node B, such as "mitotic cell cycle [GO:0000278]" is a subtype of "cell cycle [GO:0007049]". The relation "part of" is used to represent part-whole relationships: if A is a necessary component of B, then A is part of B, such as "cell communication [GO:0007154]" is part of "single-organism cellular process [GO:0044763]" [6, 2].

Figure 3.1 (QuickGo, www.ebi.ac.uk/QuickGO/) shows the DAG for the GO term "signal transduction [GO:0007165]". From the GO graph, we can see that the GO term "signal transduction [GO:0007165]" is a subclass of the GO term "regulation of cellular process [GO:0050794]" and also a part of the GO terms "cellular response to stimulus [GO:0051716]", "cell communication [GO:0007154]", and "single organism signaling [GO:0044700]".

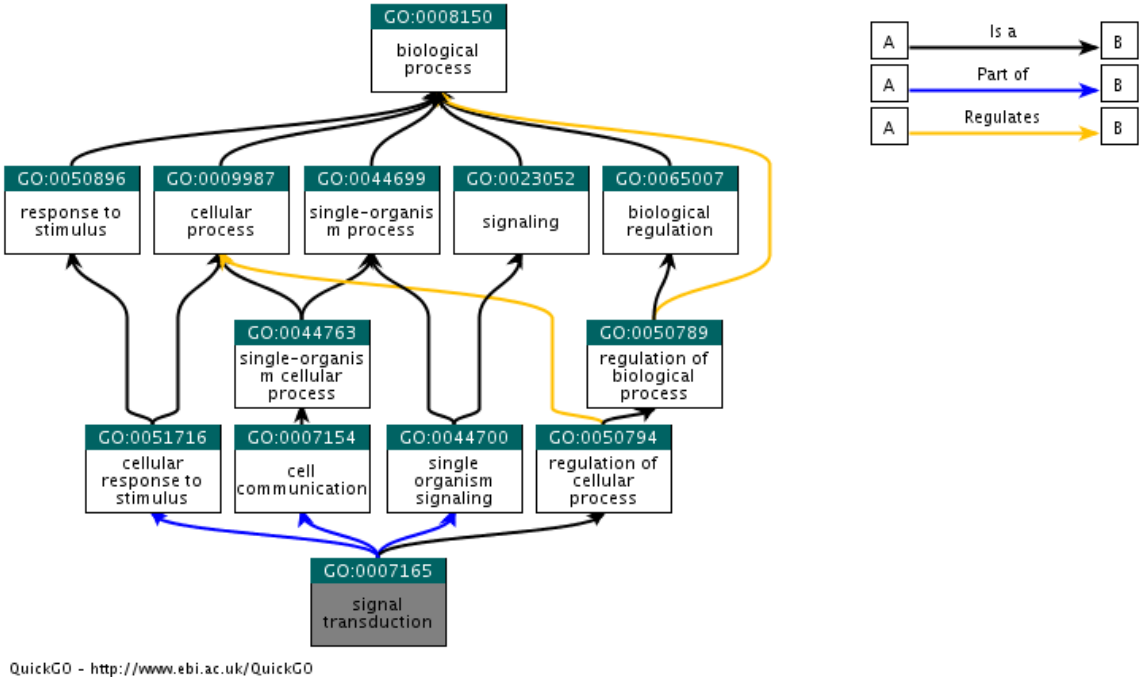


Figure 3.1: DAG for GO term "signal transduction [GO:0007165]": each box represents a GO term; lines represent the relationships between two connected GO terms.

3.2 Semantic Similarity of GO Terms

From the Uniprot Knowledgebase (www.uniprot.org), we can acquire all the genes annotated with GO terms, but we still need to quantify the extent of connection between GO terms in order to further measure the functional similarity between genes.

Semantic similarity approaches are used to produce numeric values that express the similarity of a pair of GO terms, which are the basis for the further assessment of gene clusters. There are multiple types of approaches to calculate the similarity between GO terms, broadly classified as graph-based and information-content-based (IC) methods. The IC-based approach mainly uses term probabilities to compute the information content, that is the frequency of the GO term in the whole GO database; commonly-used IC based approaches include Resnik [28], Jiang and Conrath [14], and Lin [18]. The graph based approaches use the topology of the GO graph to measure the similarities between terms. G-SESAME [37] is a graph based approach, which is also the approach we used in this project.

To describe G-SESAME, we first represent the DAG of a given GO term A by three components, denoted as $DAG_A = (A, T_A, E_A)$:

- A is the GO term itself
- T_A is a set of GO terms, including A itself and all its ancestors.
- E_A is the set of edges in the GO DAG, linking A with its ancestors in T_A.

The semantic value of the GO term A is the aggregate contribution of all GO terms in DAG_A. The weights of the contribution depend on how far the terms are from term A; terms farther from term A are more general and contribute less. So the contribution of a GO term t to GO term A, which is called S-value and denoted as $S_A(t)$, is defined as:

$$S_A(t) = \begin{cases} 1 & \text{if } t = A \\ max\left\{w_e * S_A(t') | t' \in \text{children of t}\right\} & \text{if } t \neq A \end{cases} \tag{3.1}$$

The w_e in the equation is called the semantic contribution factor for edge e from GO term t to term t'. Since the GO term A is the most specific term to itself, its

contribution is 1, and w_e is always between 0 and 1. The semantic contribution factor w_e depends on the relations of the edge, either "is a" or "part of", with the values of these two relations commonly defined as 0.8 and 0.6 respectively. Since there are potentially many edges connecting term A with term t, we take the maximum score of all paths connecting these terms. For example, in Figure 3.1, there are two paths from "signal transduction [GO:0007165]" to "single-organism process [GO:0044699]". One is "GO:0007165" $\rightarrow$ "GO:0044700" $\rightarrow$ "GO:0044699", and the other is "GO:0007165" $\rightarrow$ "GO:0007154" $\rightarrow$ "GO:0044763" $\rightarrow$ "GO:0044699". Then we calculate the S-values for both options and choose the maximum $S_A(t)$.

After calculating the $S_A(t)$ for all t in T_A, the final semantic value of GO term A is the sum of all S-values:

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{3.2}$$

To calculate the semantic similarity of a pair of GO terms A and B, we first extract the DAGs for both terms, $DAG_A = (A, T_A, E_A)$ and $DAG_B = (B, T_B, E_B)$. Then the semantic similarity between GO terms A and B is defined as

$$S_{GO}(A, B) = \frac{\sum_{t \in T_A \cap T_B}(S_A(t) + S_B(t)))}{SV(A) + SV(B)} \tag{3.3}$$

An example of calculating the semantic similarity between GO terms "cell communication GO:0007154" and "single organism signaling GO:0044700" is shown in Table 3.1.

Figure 3.2 shows the subgraph for the GO term "single organism signaling GO:0044700", which includes 4 GO terms in total. According to Equation (3.1), we can calculate all the S-values of the GO terms in the DAG for term "GO:0044700", and the results are in Table 3.1. Then, $SV_{(GO:0044700)}$ can be calculated as the sum of all the S-values in Table 3.1, $(1 + 0.8 + 0.8 + 0.64) = 3.24$.

GO terms	GO:0044700	GO:0023052	GO:0044699	GO:0008150
S-value	1	0.8	0.8	0.64

Table 3.1: S-values for GO terms in DAG for term "GO:0044700"

In the same way, we can calculate all the S-values for GO term "cell communication GO:0007154", given its DAG as shown in Figure 3.3 below. All the S-values are listed in the Table 3.2, where $SV_{(GO:0007154)} = (1 + 0.8 + 0.64 + 0.64 + 0.512) = 3.592$.

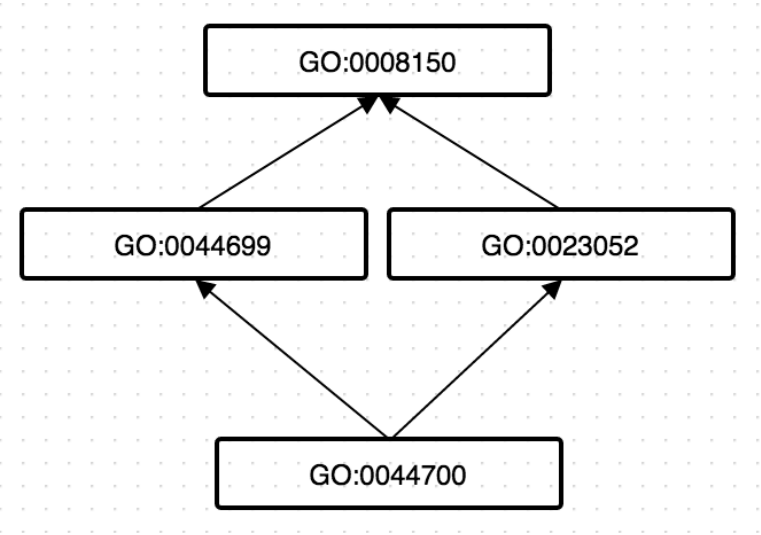


Figure 3.2: Subgraph of the GO graph for term "single organism signaling GO:0044700": all edges represent the relationship, "is a".

GO terms	GO:0007154	GO:0044763	GO:0044699	GO:0009987	GO:0008150
S-value	1	0.8	0.64	0.64	0.512

Table 3.2: S-values for GO terms in DAG for term "cell communication GO:0007154"

From Figure 3.2 and Figure 3.3, the common terms for both DAGs are $T_A \cap T_B =$ {"GO:0044699", "GO:0008150"}. According to Equation (3.3),

$$S_{GO}(\text{"GO:0044700"}, \text{"GO:0007154"}) = \frac{(0.64 + 0.8) + (0.512 + 0.64)}{3.24 + 3.529} = 0.379.$$

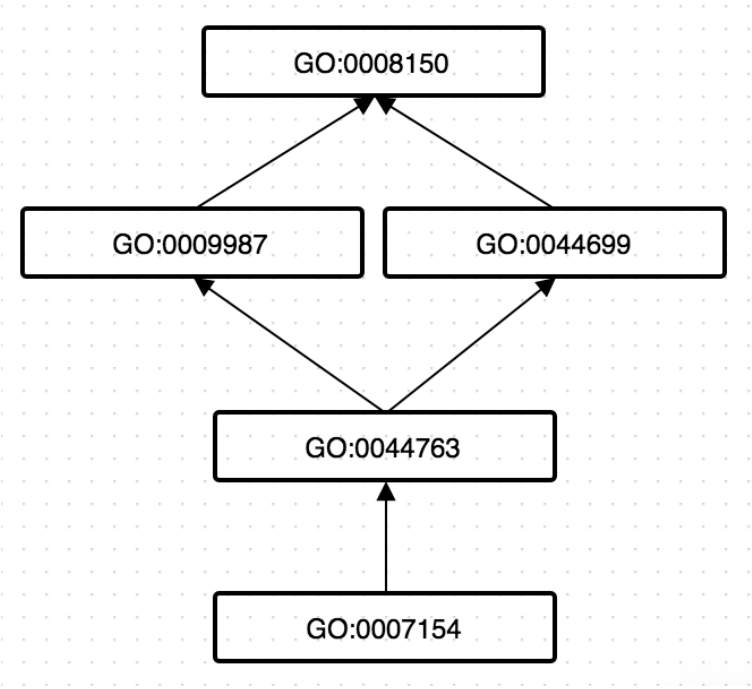


Figure 3.3: Subgraph of the GO graph for "GO:0007154": all edges represent the relationship, "is a".

Since the semantic similarity is specific to the pair of GO terms, and a cluster consists of two or more genes, we eventually score the performance of a cluster, by computing the mean of all combinations of pairs of GO terms involved in the cluster. So given a cluster C in size n, the GO score of C is defined as,

$$S_{GO}(C) = \frac{\frac{1}{2}\sum_{i,j \in C, i \neq j} S_{GO}(i,j)}{\binom{n}{2}}. \tag{3.4}$$

3.3 Resampling and Statistical Test

From Equation (3.4), any given cluster of genes can be scored using a single numeric value, which we call the GO score of that cluster. However, these GO scores are dependent on the sizes of clusters, since larger clusters are less likely to have a high GO score. For instance, a cluster with a GO score of 1 must have member genes that are all annotated with the exact same term, but a large cluster is much less likely to satisfy this criterion than a small cluster. Therefore we cannot use GO scores as the uniform standard for all clusters of different sizes.

To avoid this cluster size bias, we used a resampling method to assign significance scores to each inferred cluster by randomly generating the same-sized cluster many times to estimate the sampling distribution of GO scores, and to calculate P-values of clusters. We also considered the influence of resampling with replacement and resampling without replacement on the distribution of GO score:

- Resampling without replacement, is to draw the genes randomly and exactly once from the whole gene set, so that each gene can appear at most once in a sampled cluster. However, this method may not work well for very large clusters that contain most proteins in the data set, because the members of that cluster would be very similar each time and the GO scores would vary only a small amount across replicates, which will cause a extremely small variance in the sampling null distribution of GO scores. Therefore, even though the GO score is only slightly greater than the mean, the P-value would still be small and would suggest statistical significance. Taking an extreme example, if we want to sample n out of n genes, we will get the exact same set of genes every time, and the variance of GO scores in this scenario will be 0.

- Resampling with replacement is equivalent to drawing GO terms rather than genes from our gene pool, because the same gene can not appear more than once in the resulting clusters, whereas GO terms can. Resampling with replacement has two features listed below, making it more appropriate in our project:
  - Samples are drawn independently.
  - Resampling with replacement also considers the frequency of GO terms in our collection of GO terms. For example, if one cluster whose size is small and all genes have GO term "transport [GO:0006810]", with the semantic similarity guaranteed to be equal to 1. However, it doesn't mean it is absolutely a good cluster, because "transport [GO:0006810]" is a very common GO term.

So compared to resampling without replacement, the sample mean estimated by resampling with replacement will increase because if one gene was resampled into a cluster twice, it will contribute two equal GO terms, with the semantic similarity between them as 1.

Figure 3.4 shows the difference between results from resampling a cluster of size 30 without replacement and with replacement, on a subset of 100 genes from the dataset we used in Chapters 4 and 5. We can see that the distribution for resampling with replacement has a greater mean (0.3915) than without replacement (0.3845), and an obviously more flat distribution. As a result, the resampling with replacement method will produce a smaller number of significant clusters, but more justifiable than resampling without replacement. Therefore, only resampling with replacement method is used in our project.

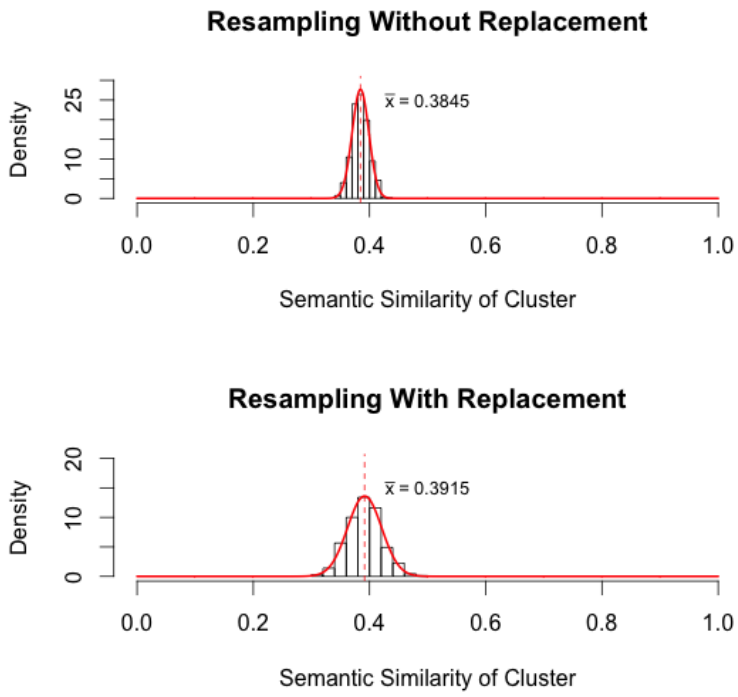


Figure 3.4: Comparison of resampling with and without replacement: the red line is the estimated normal distribution of GO score by resampling 1000 times; the red dotted vertical line is where the sample mean is located.

Chapter 4

Application of the Correlated Evolution Approach

In this chapter, we will apply our method to calculate the evolutionary correlation among genes of the bacterium "*Lachnospiraceae* bacterium 3-1-57FAA-CT1" using a set of other genomes from the same taxonomic order to build phylogenetic profiles. These profiles will then be used to infer hierarchical clusters, which will be split using a functional criterion. Finally, we will examine the functional cohesion of our clusters in relation to randomly resampled clusters.

4.1 Data Description

The bacterium "*Lachnospiraceae* bacterium 3-1-57FAA-CT1", abbreviated as "Lb-CT1", is a family in the order *Clostridiales* and isolated from a human fecal sample. "Lb-CT1" has 6506 genes in total, and we will analyze the phylogenetic profiles of its genes across a phylogenetic tree with 687 genomes from the *Clostridia* family to explore how its genes interact with each other. The phylogenetic tree was built from a core set of genes using the approach of gene selection similar to AMPHORA [38]. The phylogenetic tree construction performed using RAxML [32] and the GTRCAT model was used for sequence substitution probabilities. Phylogenetic profiles were constructed by comparing conceptually translated proteins of "Lb-CT1" to those of all other genomes using Rapsearch2 [43]. A maximum e-value threshold of 10^{-20} to the "Lb-CT1" proteins was used to include genes in phylogenetic profiles. All the GO information used in this application is obtained from the Uniprot Knowledgebase (www.uniprot.org).

Pagel's method is computationally expensive and there will be $\binom{6502}{2} = 21,160,765$ pairs if we want to apply this method on the whole gene set, so we must reduce the dataset under consideration to a more tractable size. From the initial tree reconstructed from 687 genomes, we subsampled "Lb-CT1" and an additional 83 genomes at random, which reduced the number of the phylogenetic profiles and allowed us to

consider only 2892 distinct profiles.

4.2 Quantifying Evolutionary Correlation Among Genes

The "Discrete" program implements the models described in Pagel (1999) and enables the analysis of binary characters on phylogenetic trees [25]. The "Discrete" program will output the likelihoods of independent and dependent evolutionary models for each pair of genes based on their phylogenetic profiles, from which we can calculate the log-likelihood ratio and P-value to represent the similarity between genes.

To illustrate how the Pagel method takes the phylogenetic effect into consideration while measuring the similarity between phylogenetic profiles, Figure 4.1 is a typical example from our real data, and the specific values of similarity calculated by the Pagel method are shown in Table 4.1. Using Gene R as the reference profile, intuitively Gene 2 should be more similar to the base profile than Gene 1, if we simply count the number of unmatched presences (Gene 2 is missing only genome a, while Gene 1 has two additional represented genomes near b). However, if we consider the phylogenetic effect, the two additional representatives near b are less critical than the single miss at a, because the genomes around b are closer relatives and more likely to carry the same genes due to shared descent. The last 2 profiles (Gene 3 and Gene 4), both of which have an extra representative relative to the base profile, also show that the extra presence at e is a more significant difference than the one at d due to its greater phylogenetic distance.

Table 4.1: Similarity between Gene R and the other four genes calculated by the Pagel method: inside of parentheses are the corresponding GI numbers.

	Gene 1 (GI:496546121)	Gene 2 (GI:496545868)	Gene 3 (GI:496549699)	Gene 4 (GI:488634388)
Gene R (GI:488634373)	17.29659	13.13266	12.23021	10.48783

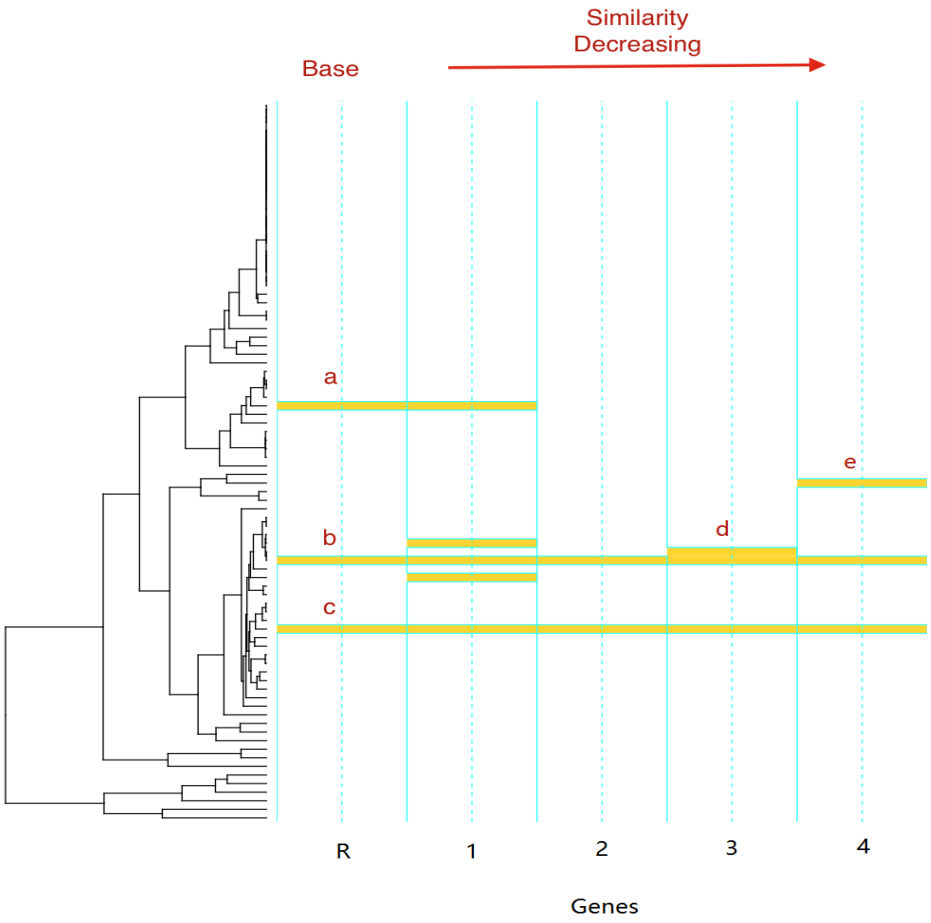


Figure 4.1: Comparison of Phylogenetic Profiles: the dendrogram on the left side is the phylogenetic tree with 84 tips; the columns are the phylogenetic profiles of 5 genes showing their presence and absence in each genome; Gene R is the reference object; the other 4 genes are displayed in order of the similarity to Gene R from large to small, calculated by the Pagel method.

4.3 Hierarchical Clustering of Genes

From the previous section, we obtained the likelihoods for all pairs chosen from these 2892 genes, and we can further compute log-likelihood ratio statistics and the P-values. The P-values correspond to monotonic non-linear transformation of the log-likelihood ratio statistics. However, the hierarchical clustering procedure is not scale invariant for this non-linear transformation, thus does not output the same clustering results. We need to choose which statistics to work on. Our hierarchical clustering procedure is based on the Ward linkage which aims to minimize the within cluster sum of squared distances between all pairs. The property of squared distances are more similar to the chi-squared statistics from the log-likelihood ratio test. Thus we base our clustering on the log-likelihood ratio statistics. Since our log-likelihood ratio statistics are positively correlated with similarity, we convert these values to dissimilarity scores simply by subtracting them from the maximum value.

Figure 4.2 is an overview of the generated hierarchical clustering dendrogram based on the likelihood ratios, and the number of clusters is shown on the right side reflecting its changes along with heights. Due to the hierarchical nature of the dendrogram, we can analyze the interactions among genes and associations between clusters in different levels by recovering and splitting the clusters along with height.

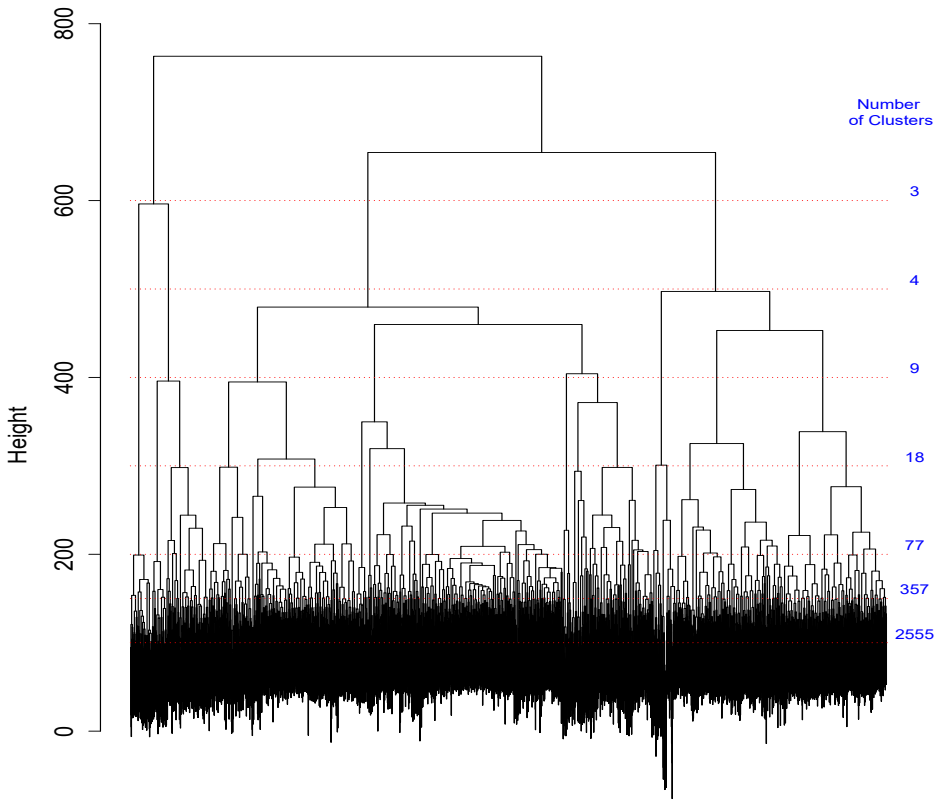


Figure 4.2: Overview of the Hierarchical Clustering Dendrogram: each dotted line is a possible realization of tree cutting, and the value on the right margin is the total number of clusters induced by cutting branches at that height.

4.4 Optimizing the Cutting Height

After we generate the hierarchical dendrogram, we first use fixed-height tree cutting to identify clusters, and then compare with the results from the dynamic tree cutting. Figure 4.3 shows the weighted mean of GO scores of the clusters at different heights, which is defined as $W_{GO} = \sum_{i=1}^{k} \frac{n_i}{N} S_{GO}(c_i)$, given N genes grouped into k clusters $\{c_1, c_2 ... c_k\}$ with corresponding sizes $\{n_1, n_2 ... n_k\}$. We observe a steady increase of W_{GO} from a height of 600, to a maximum score of 0.512 at a height of 130, followed by a sharp drop to 0. We therefore choose $h = 130$ as the optimal height.

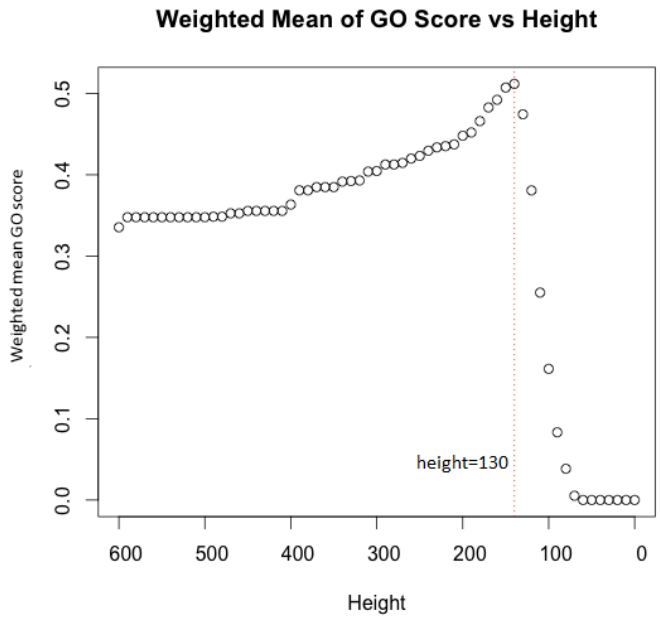


Figure 4.3: Weighted Mean GO score vs Height: the maximum weighted mean GO score happens at height 130.

Since dynamic tree cutting has the ability to choose different h values for different clusters, we expect the dynamic approach to produce better functional cohesion. Figure 4.4 compares the results of the two types of cutting approach, which shows that the results are as expected that, fixed-height hierarchical clustering (blue bar) creates 22 more clusters with GO score below 0.5, than dynamic hierarchical clustering (green bar), which generates 29 more "good" clusters whose GO scores are greater than 0.5.

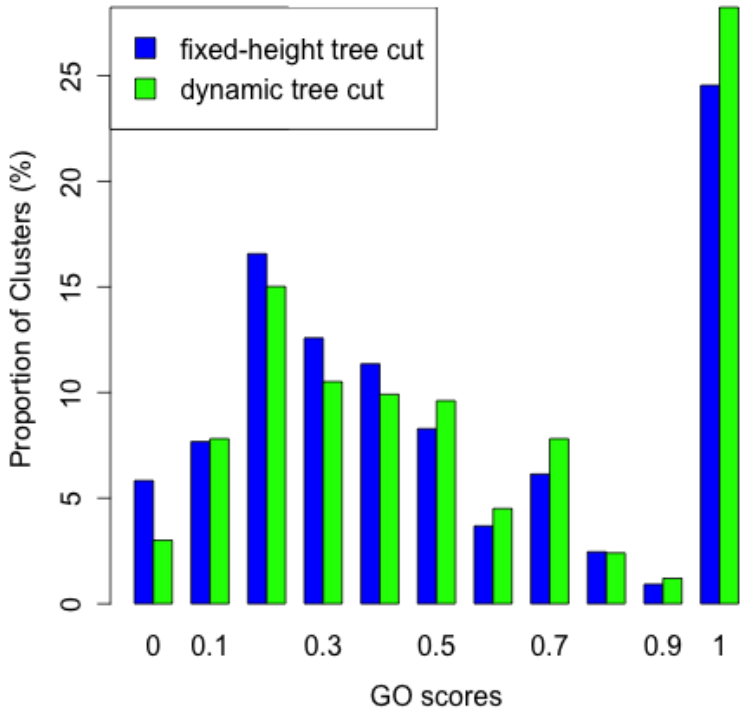


Figure 4.4: Comparison of the performance in fixed-height cutting and dynamic cutting: blue bars represent the results of fixed-height tree cutting; green bars represent the results of dynamic tree cutting.

4.5 Evaluation of Clustering Results

To demonstrate that the genes within a cluster are more likely to be functionally linked in a biological process, we generated a thousand random clusters with resampling to estimate the distribution of the GO scores for each size of clusters, from which we can calculate the significance of our clusters.

Since the dynamic approach yielded more functionally coherent clusters than the fixed-height method on our dataset, we evaluated the dynamically generated clusters in greater depth. Figure 4.5 shows the significance of GO scores of our 24 largest clusters against 1000 of their same sized random clusters. Nineteen out of 24 performed better than the randomly generated clusters, which means that the genes in these clusters have high chance of being associated in biological process.

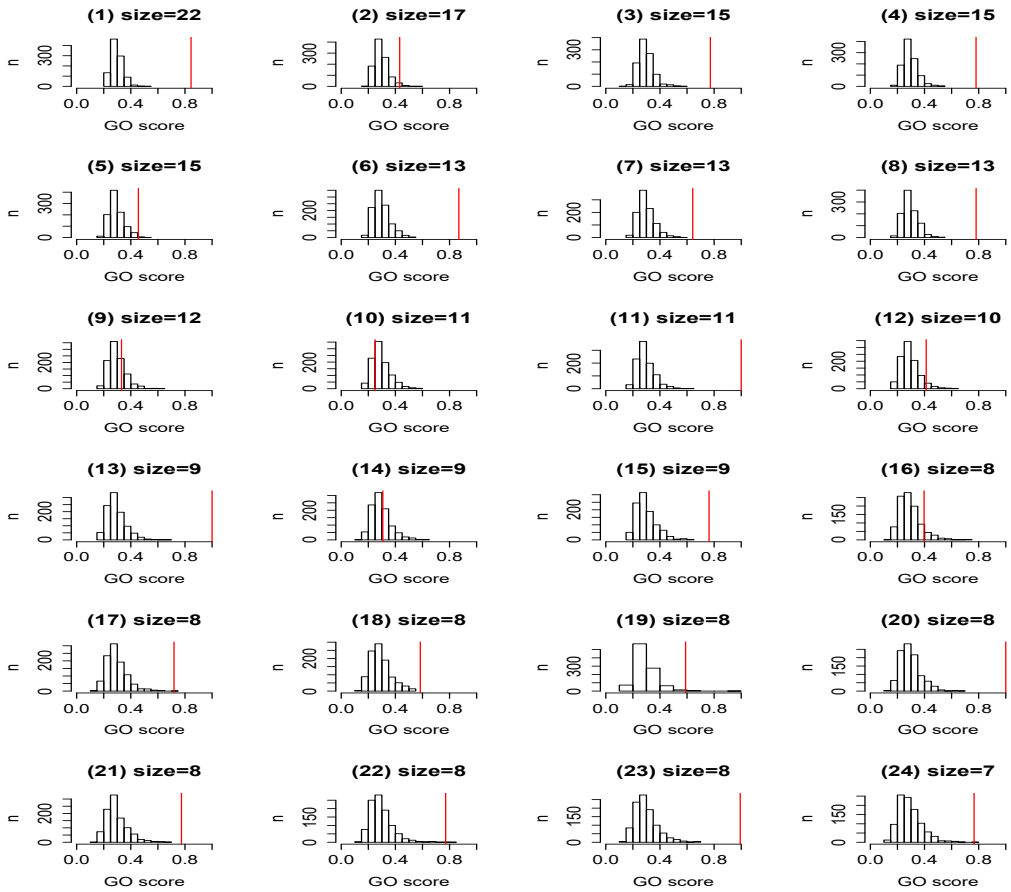


Figure 4.5: Performance of largest 24 Clusters Measured by GO score: in each figure, the red line is the GO score of the cluster we acquired, and the bars show the GO score distribution from 1000 randomly resampled clusters of the same size.

4.6 Examples of Gene Function Prediction

To make use of our clustering outcomes, we could try to predict the functions of the unannotated genes, since within a functionally cohesive cluster, we might expect unannotated genes with similar phylogenetic distributions may be involved in the same biological process as the genes annotated with GO terms. However, two factors may generate inaccurate predictions on the unannotated genes based on our method:

- Low GO score clusters

  Although some low-scoring clusters may be informative as well, we cannot make the predictions for the genes in those clusters precisely, since a low GO score means that there are various GO terms in those clusters.

- Large proportion of unannotated genes

  If only a small percentage of genes in the cluster have the GO terms, we might still get accurate predictions in some cases, but overall we have less confidence if the GO term coverage is poor.

So, we only focus on the clusters avoiding both of these two conditions and here we provide two examples of these types of clusters.

The cluster, denoted as "235H150", contains 10 genes and 8 of them have the exact same GO term, "phosphorelay signal transduction system [GO:0000160]", which means this cluster is very cohesive. So we can make the prediction that the remaining two genes ("GI:496544772", "GI:496545797") have the same "phosphorelay signal transduction system" term. Figure 4.6 also shows that these 10 genes have similar phylogenetic profiles.

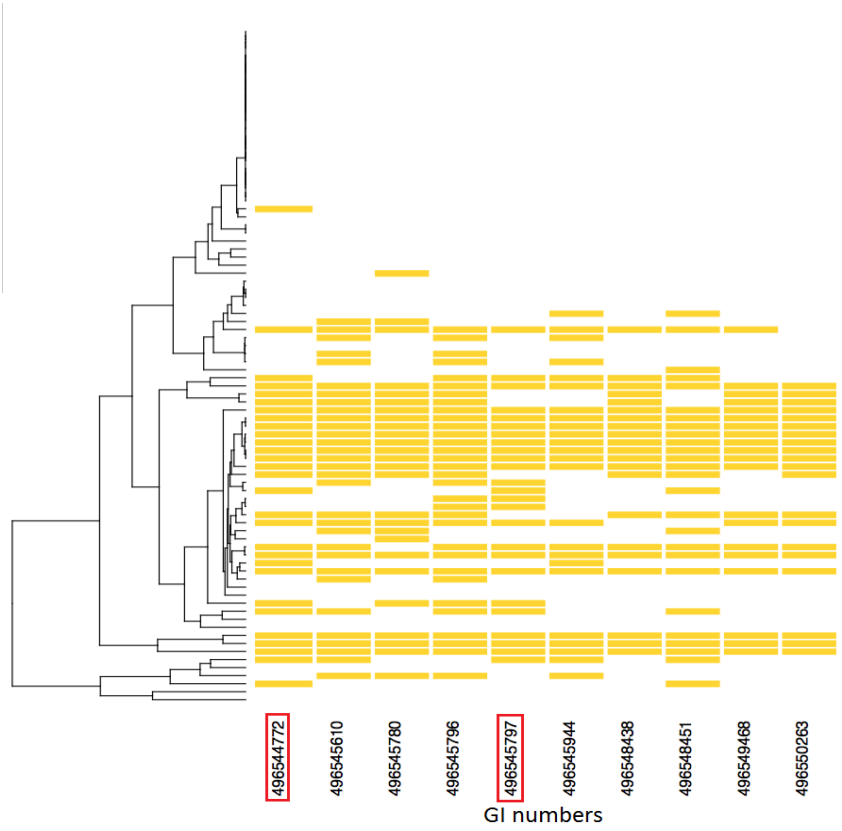


Figure 4.6: Phylogenetic Profiles of Genes in Cluster "235H150": the columns are the phylogenetic profiles of 10 genes; the dendrogram on the left is the phylogenetic tree of 84 genomes; two unannotated genes are marked with red rectangles.

The cluster, denoted as "115H130", consists of 4 genes and has a mean GO score of 0.918. Figure 4.7 shows the GIs and GO terms of these 4 genes and "GI:496544038" has no matching GO terms, and is therefore a good candidate for functional prediction.

GI	Gene Ontology(biological process)
496544038	Unknown
496546564	branched-chain amino acid biosynthetic process [GO:0009082]
496546568	leucine biosynthetic process [GO:0009098]
496546569	leucine biosynthetic process [GO:0009098]

Figure 4.7: Composition of the Cluster "115H130"

As we can see that "GI:496546568" and "GI:496546569" have the same GO term, but different from "GI:496546564". According to the information from the online GO database (http://www.ebi.ac.uk/QuickGO/), as shown in Figure 4.8, "leucine biosynthetic process" is one of the child terms of "branched-chain amino acid biosynthetic

process". So we can predict conservatively that the GO term of "GI:496544038" is "branched-chain amino acid biosynthetic process", since it is a broader term than "leucine biosynthetic process".

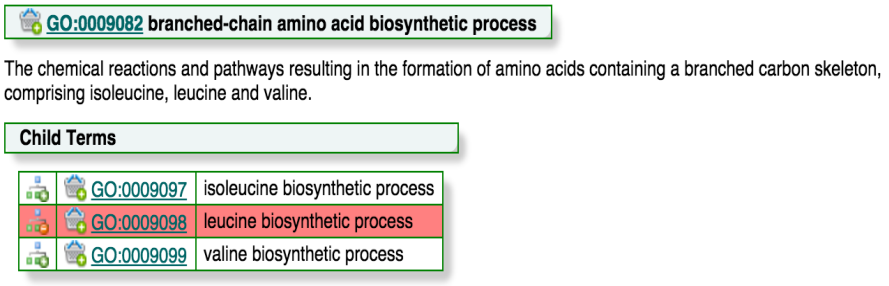


Figure 4.8: Child Terms of "GO:0009082"

Figure 4.9 displays the phylogenetic profiles of genes in cluster "115H130", and there is a dramatic difference at the top of the phylogenetic tree marked by the red rectangle. However, although the difference between the first two profiles and the last two profiles is very striking, the phylogenetic correction makes the profiles very similar and points to potentially a single lateral gene transfer (LGT) event (at location a) distinguishing the distribution of the two sets of genes.

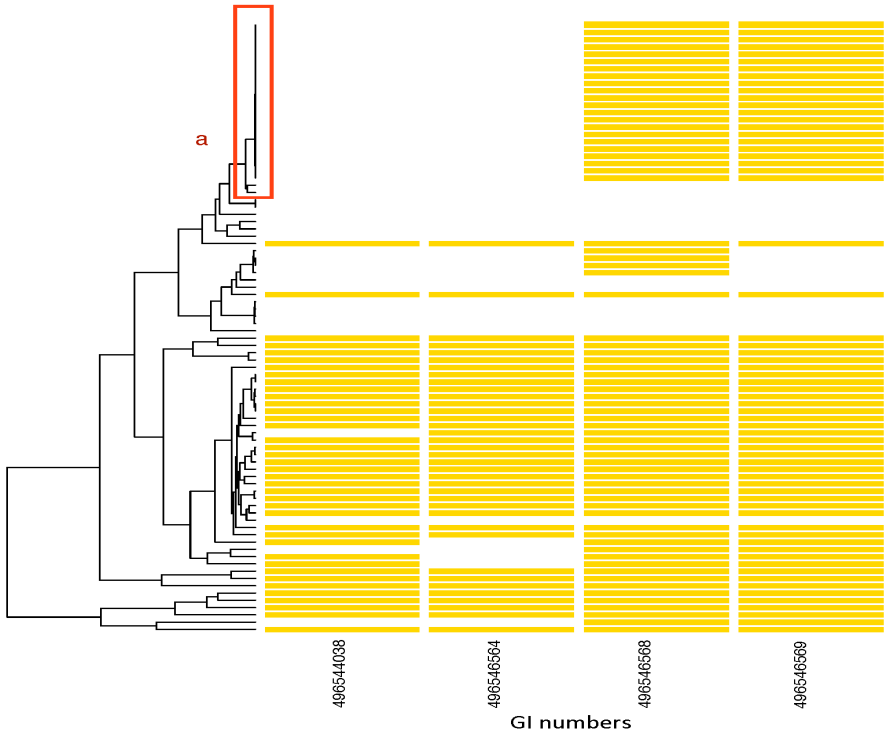


Figure 4.9: Phylogenetic Profiles of Genes in Cluster "115H130": the columns are the phylogenetic profiles of 4 genes; the dendrogram on the left is the phylogenetic tree of 84 genomes; a potential LGT event occurred at a.

Chapter 5

Comparison with Clustering by Inferred Modules of Evolution (CLIME)

In this chapter, we will introduce another evolution based approach, CLIME, and apply this method on the same data set to generate a new set of clusters. The Rand index is also introduced to measure the similarity between clusters resulting from two approaches.

5.1 Introduction to CLIME

CLIME, short for Clustering by Inferred Models of Evolution, is a computational method of predicting gene function published by Li and Calvo in 2014 [17]. Briefly, CLIME is a clustering algorithm based on a hidden Markov model (HMM), to group genes into evolutionarily conserved modules (ECMs) according to the inferred gene gain and loss events with the assumption that each gene has a single gain event and zero or more loss events [17]. It has been applied to the human mitochondrial proteome and proteomes of yeast and red algae to explore the evolutionary modularity.

Figure 5.1 shows the schematic overview of CLIME as presented in the original paper [17]. The complete procedure of CLIME includes Input, Partition, Expansion, and Output:

- Input

  Three data files are input by users,

  - a phylogenetic tree
  - a matrix containing the gene phylogenetic profiles
  - a gene set

  of which, the input gene set is just a subset of genes included in the phylogenetic profile matrix.

- Partition

  A Bayesian mixture model of HMMs is applied on the input gene set to generate the initial ECMs, and at the same time, the number of ECMs and the evolutionary model of each ECM are also learned. The evolutionary model is represented by a single gain branch and a vector of loss probabilities for each branch.

- Expansion

  The ECM expansion set, which is denoted as ECM+, is created to include all other genes in the profile matrix by calculating the likelihood of each gene against the evolutionary model of each ECM and assigning the gene to the best-fitting ECM.

- Output

  The disjoint ECM clusters and their associated ECM+ expansions are the outputs of CLIME.

Since CLIME is also based on phylogenetic profiles and common evolutionary history and its objectives are similar, we applied CLIME to the same data set in Chapter 4 using the program developed by the authors, and contrasted its predictions with our results from Chapter 4.

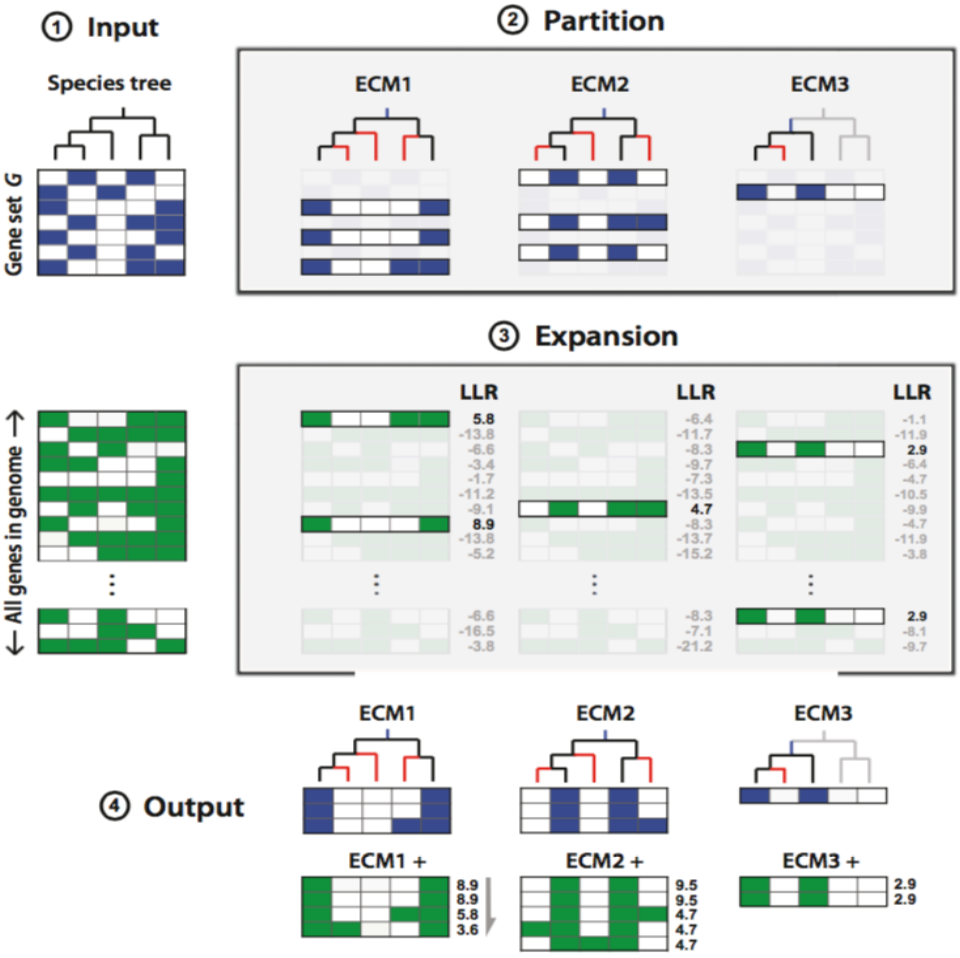


Figure 5.1: Overview of CLIME: (1) input the phylogenetic profiles of a set of genes across the given phylogenetic tree; (2) partition the input set of genes into disjoint ECMs and each ECM is modeled with a gain branch (blue) and branch-specific probalilities of gene loss (red); (3) assign other genes to the best-fitting ECM, scored by the log-likelihood ratio; (4) output the disjoint ECM clusters and associated ECM+ expansions. Figure reproduced with the permission of the copyright holder.

5.2 Rand Index

To compare the clustering outcomes of CLIME and our method, the Rand index is applied to measure the similarity of the clustering outcomes generated by two approaches. To calculate the Rand index, given a set of N elements $S = \{s_1, s_2..., s_n\}$, and two sets of clusters, $X = \{x_1, ..., x_r\}$ and $Y = \{y_1, ..., y_s\}$, we can have the following notations [12, 31]:

- n_1, the number of pairs in S that are in the same cluster of X and in the same cluster of Y;
- n_2, the number of pairs in S that are in different clusters of X and in different clusters of Y;
- n_3, the number of pairs in S that are in the same cluster of X and in different clusters of Y;
- n_4, the number of pairs in S that are in different clusters of X and in the same cluster of Y.

Then the Rand index, RI, is defined as,

$$RI = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4} = \frac{n_1 + n_2}{\binom{N}{2}}$$

which is a number between 0 and 1 [12].

For example, in Figure 5.2, 10 genes are grouped into a set of clusters A, which has 3 clusters of size 4, 3, and 3 respectively. The same 10 genes are grouped into the other set of clusters B, based on a different method, and the set B has two clusters with sizes of 6 and 4.

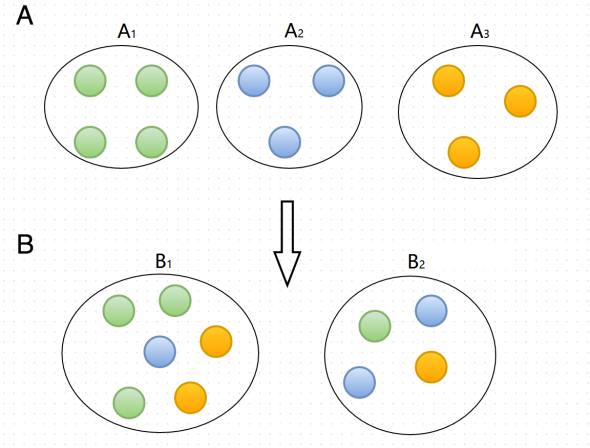


Figure 5.2: An example of comparison between two clusters

From the Figure 5.2, in cluster B_1, there are 3 nodes from A_1 and 2 nodes from A_3, and there are 2 nodes from A_2 in cluster B_2. Table 5.1 shows the mapping of pairs for the purpose of calculating the Rand index.

Table 5.1: The Rand index table between A and B

	same cluster in X	different clusters in X
same cluster in Y	$n_1 = 5$	$n_4 = 16$
different clusters in Y	$n_3 = 7$	$n_2 = 17$

and the Rand index between A and B is,

$$RI = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4} = 0.488.$$

As we can see the Rand index gives the same weight to n_1 and n_2, but the situation of n_2, which can happen more often by chance, contributes more to the Rand index.

To get a more accurate comparison between two clustering outcomes, the adjusted Rand index is also introduced, which is the corrected-for-chance version of the Rand index [41]. Table 5.2, which is called the contingency table, introduces the notations for the computation of the adjusted Rand index between two sets of clusters $X = \{x_1, ..., x_r\}$ and $Y = \{y_1, ..., y_s\}$ of sizes $\{a_i, ..., a_r\}$ and $\{b_1, ..., b_s\}$, where n_{ij} represents the number of objects in common between two clusters x_i and y_j, The contingency table summarizes the overlap between X and Y, and the adjusted Rand index is

Table 5.2: The contingency table of X and Y

$X\backslash Y$	y_1	y_2	$\cdots$	y_s	Sums
x_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
x_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

defined as,

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}) - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}. \tag{5.1}$$

Still taking the example in Figure 5.2, Table 5.3 shows the contingency table of A and B.

Table 5.3: The contingency table of the example in Figure 5.2

$A\backslash B$	b1	b2	Sums
a1	3	1	4
a2	1	2	3
a3	2	1	3
Sums	6	4	

and the adjusted Rand index between A and B is,

$$ARI = \frac{(3+1+1) - \frac{(6+3+3)(15+6)}{45}}{\frac{1}{2}((6+3+3)+(15+6)) - \frac{(6+3+3)(15+6)}{45}} = -0.055.$$

The negative ARI tells that the index is less than the expected index, and it is much less than the RI of the same data set, which also shows that ARI is a more strict measurement than general RI.

To compare our predictions with those of CLIME, we calculated the adjusted Rand index between the set of clusters at each height of our inferred hierarchy with the set of clusters generated by CLIME. The maximum Rand index occurred at the height = 140, which is very close to the optimized height of the fixed height tree cutting (height = 130) in Figure 4.3.

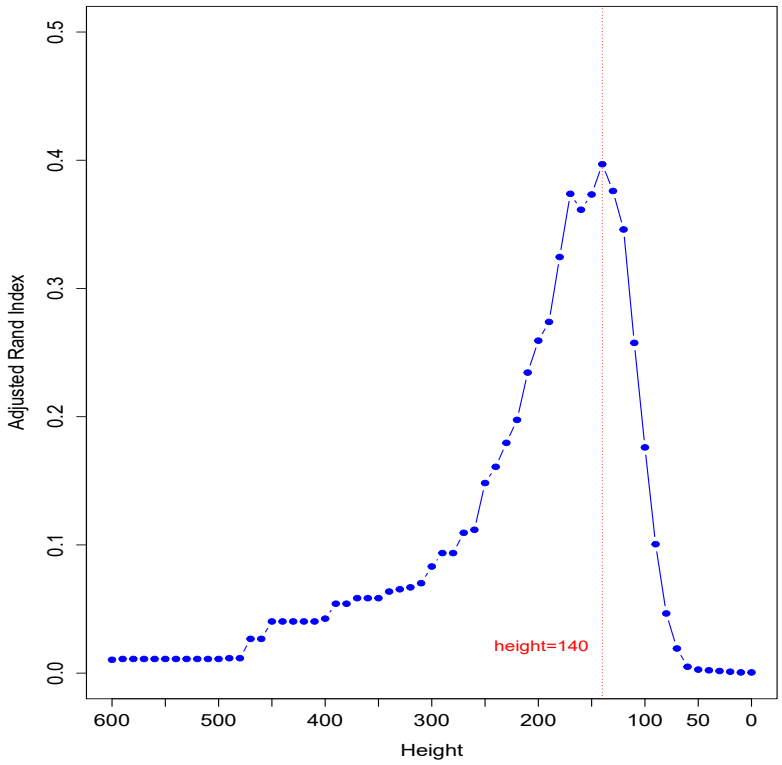


Figure 5.3: Change of adjusted Rand index between CLIME clusters and the hierarchical dendrogram, cut at different heights: the maximum ARI is indicated with a dotted red line.

5.3 Comparisons of the Results

CLIME generated 330 clusters from our test data set involving 1211 out of the total of 2892 genes, whereas our method clustered 2429 genes. So for comparison purposes, we compared only the 1022 genes which are clustered by both methods. Figure 5.4 shows the distribution of the sizes of the clusters from our method and these 1022 genes are clustered into 207 clusters, without counting 38 singletons.

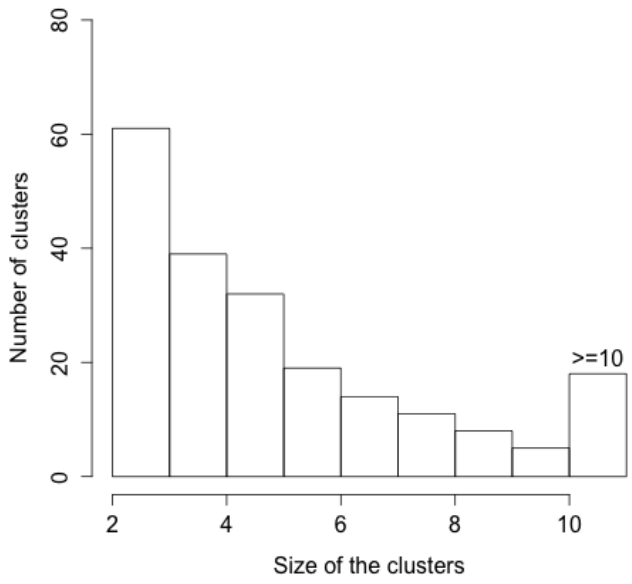


Figure 5.4: Size distribution of clusters generated by hierarchical clustering.

The distribution of clusters by size for CLIME is shown in Figure 5.5 below. These 1022 genes are arranged into 272 clusters and 18 singletons.

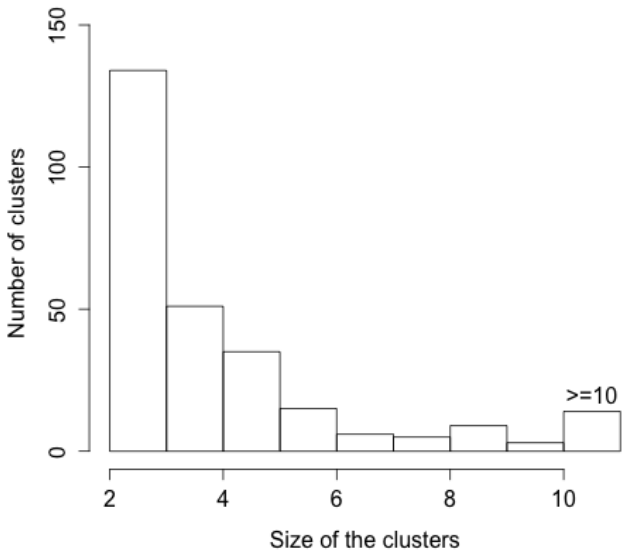


Figure 5.5: Size distribution of clusters generated by the CLIME software.

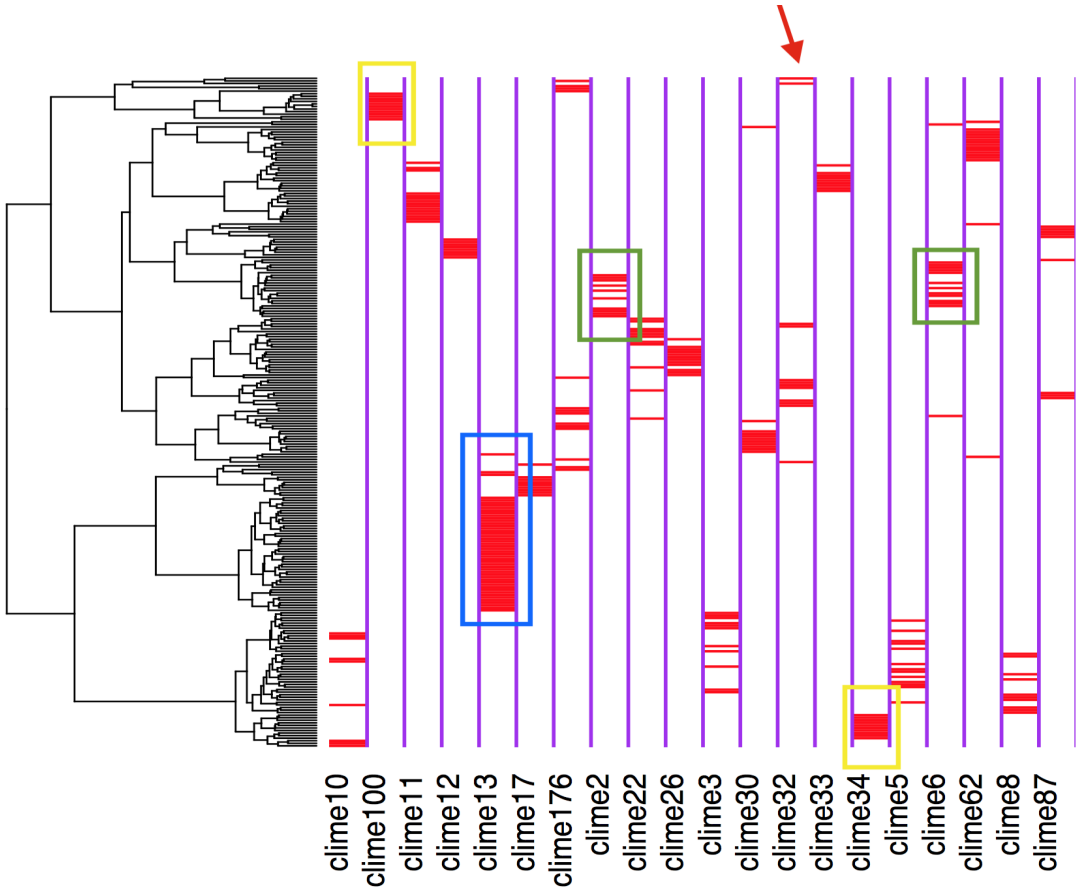


Figure 5.6: Comparison between 20 CLIME clusters with hierarchical clustering dendrogram: the tree diagram on the left is the subtree of the hierarchical clustering dendrogram (Figure 4.2) for a subset of 262 genes; each column shows the member genes of a CLIME cluster and the red bar represents that the gene is a member; symbols in the graph identify the types of clusters: well matched clusters (yellow rectangle), similar cluster (blue rectangle), complementary clusters (green rectangle) and dissimilar cluster (red arrow).

Figure 5.6 above compares 20 CLIME clusters with the dendrogram generated by our method. Because the hierarchical clustering was based on this dendrogram, the structure of this dendrogram can represent roughly the clusters from the hierarchical clustering. If the CLIME clusters are adjacent, it means the results of CLIME and our method are similar, like the area bordered by blue in the graph. If the CLIME clusters are the whole pieces with no blank, like the two blocks in the yellow borders, they are probably well matched clusters as ours. Some clusters seem to be complementary, like the green ones, which might be affected by the different heights of tree cuts, and

also some clusters are quite different from ours, like the one directed by the red arrow.

Our cluster "29H260", which contains eight proteins, is exactly the same as the CLIME cluster "clime16". The GO terms of genes in this cluster are shown in Table 5.7 below, according to the Uniprot database. These 8 genes have the same GO terms, which are all related to "cobalamin biosynthetic process [GO:0009236]", though some of them have different secondary terms. This cluster is supported by both approaches, and shows high functional cohesion.

GI	Gene ontology (biological process)
496547196	cobalamin biosynthetic process; glutamine metabolic process
496547199	cobalamin biosynthetic process
496547205	cobalamin biosynthetic process; glutamine metabolic process
496547208	cobalamin biosynthetic process
496547209	cobalamin biosynthetic process
496547210	cobalamin biosynthetic process; porphyrin-containing compound biosynthetic process
496547212	anaerobic cobalamin biosynthetic process
511538146	cobalamin biosynthetic process; corrin biosynthetic process

Figure 5.7: GO information about a Pagel cluster,"29H260"

Figure 5.8 below gives information about the composition of a cluster from our method, "64H200". This cluster has 9 elements, which are divided by CLIME into two clusters (denoted as "clime263" and "clime25"), plus three unclustered proteins. According to the GO information on the Uniprot website, the GO terms of these 9 genes are all the same, "carbohydrate metabolic process [GO:0005975]", and Figure 5.9 also shows that these 9 genes have similar phylogenetic profiles. According to their shared GO terms, it is reasonable to group these genes into one rather than two clusters. This example also shows that the height of tree cutting on this cluster is very reasonable, which is successfully detected by our approach.

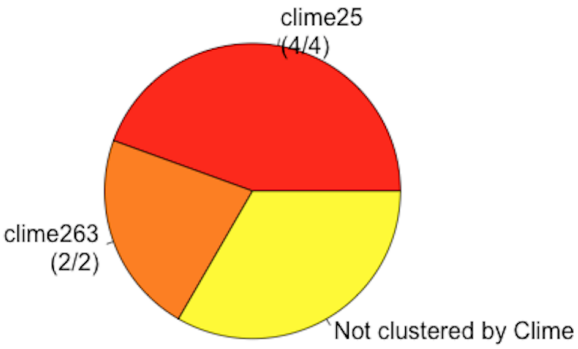


Figure 5.8: Composition of a Pagel cluster, "64H200"

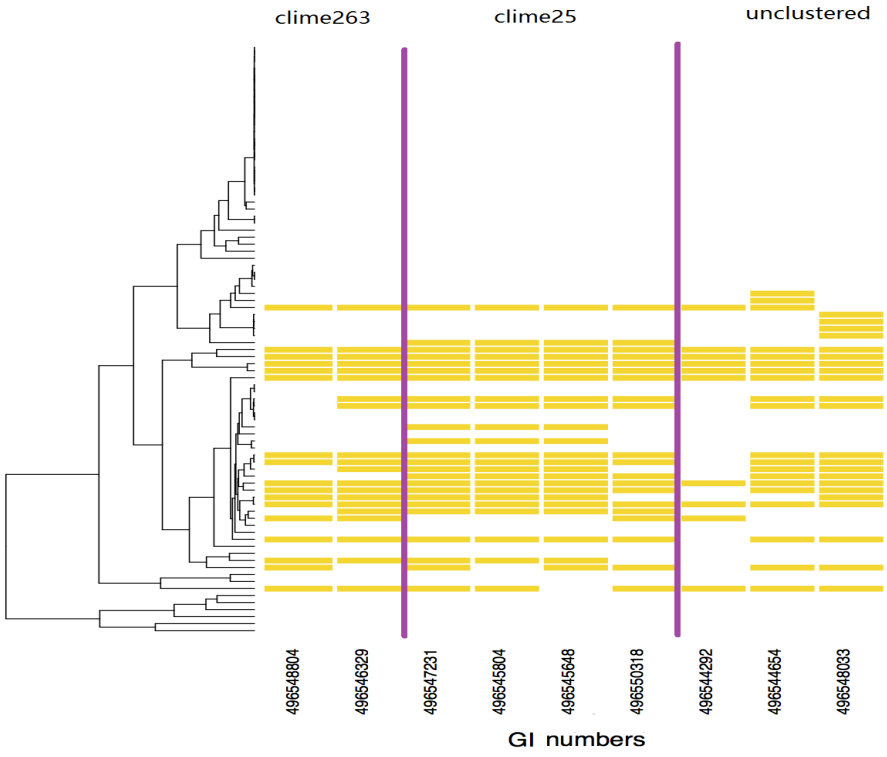


Figure 5.9: Phylogenetic Profiles of Genes in Cluster "64H200": the columns are the phylogenetic profiles of 9 genes; the blue borders separate the genes into 3 groups ("clime263", "clime25" and unclustered proteins); the dendrogram on the left is the phylogenetic tree of 84 genomes.

Chapter 6

Conclusion and Future Work

In this thesis, we described the complete process of a gene clustering approach, which covers calculation of evolutionary correlation, gene clustering, and evaluation of clusters. We applied our approach to the gene set of "*Lachnospiraceae* bacterium 3-1-57FAA-CT1", and found informative clusters which can be used to predict the functions of the unannotated genes and to discover the possible gene gain and loss events occurring in the genomes during their evolution. In this application, our approach shows two key advantages:

- From the example in Figure 4.9, the striking difference in the phylogenetic tree, which could be caused by a single LGT event, will heavily influence the phylogenetic profiles, but our method will be minimally influenced by these highly correlated genes.

- The hierarchical structure of our resulting dendrogram (Figure 4.2) allows us to analyze the interactions among genes and associations between clusters in different levels by recovering and splitting the clusters along with height.

Furthermore, we compared the results to another evolution based gene clustering approach, CLIME. There are two main differences between the basic principles of two approaches,

- CLIME has the constraint on the phylogenetic history that a gene can be gained only once.

- CLIME only considers the topology but not the branch lengths of the phylogenetic tree.

In spite of dissimilarities between the two methods, the two sets of predictions have the considerable similarity, and both successfully showed that genes with similar histories

of evolution on the phylogenetic tree also tend to be functionally linked and predicting functions of genes on the basis of phylogenetic profiles is a practical approach.

However, our approach can be improved in several ways, which are expected to be done in future work:

- Enhance the evaluation of gene clusters by permutating the hierarchical clustering dendrogram to generate the clusters using the same tree cutting strategy rather than only resampling the genes.
- To predict the function more precisely, we need to quantify the effect of the proportion of unannotated genes in the clusters and the different heights of cutting on the hierarchical clustering dendrogram.
- We must optimize the algorithm of Pagel's method to make it feasible to run on the whole phylogenetic tree.
- Pagel's method assumes a fixed evolutionary rate along the branches of the tree, and we can improve the method to adjust to changing evolutionary rates.

Bibliography

[1] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] David Baum. Reading a phylogenetic tree: The meaning of monophyletic groups. *Nature Education*, 1(1):190, 2008.

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

[5] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[6] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

[7] Matthieu Cord and Pádraig Cunningham. *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer Science & Business Media, 2008.

[8] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[9] Joseph Felsenstein. Phylogenies and the comparative method. *American Naturalist*, pages 1–15, 1985.

[10] Jesse Gillis and Paul Pavlidis. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (cafa). *BMC bioinformatics*, 14(3):1, 2013.

[11] Brian K Hall. Descent with modification: the unity underlying homology and homoplasy as seen through an analysis of development and evolution. *Biological Reviews*, 78(3):409–433, 2003.

[12] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[13] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[14] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[15] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[16] Igor Kononenko and Matjaž Kukar. *Machine learning and data mining: introduction to principles and algorithms*. Horwood Publishing, 2007.

[17] 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.

[18] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[19] Phillip W Lord, Robert D Stevens, Andy Brass, and Carole A Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Pacific Symposium on Biocomputing*, volume 8, pages 601–612, 2003.

[20] Edward M Marcotte, Matteo Pellegrini, Michael J Thompson, Todd O Yeates, and David Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, 1999.

[21] Conor J Meehan and Robert G Beiko. A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. *Genome biology and evolution*, 6(3):703–713, 2014.

[22] John F Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.

[23] James R Norris. *Markov chains*. Number 2008. Cambridge university press, 1998.

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

[25] Mark Pagel. Users manual for discrete. university of reading. *Reading, UK*, 2000.

[26] 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.

[27] 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.

[28] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[29] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[30] Naruya Saitou and Tadashi Imanishi. Relative efficiencies of the fitch-margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol*, 6(5):514–525, 1989.

[31] Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial neural networks–ICANN 2009*, pages 175–184. Springer, 2009.

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

[33] Williams J Stewart. *Introduction to the numerical solutions of Markov chains*. Princeton Univ. Press, 1994.

[34] Naoko Takezaki and Masatoshi Nei. Genetic distances and reconstruction of phylogenetic trees from microsatellite dna. *Genetics*, 144(1):389–399, 1996.

[35] Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

[36] Jean-Philippe Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18(suppl 1):S276–S284, 2002.

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

[38] Martin Wu and Jonathan A Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):R151, 2008.

[39] Xiang Xiao, Ernst R Dow, Russell Eberhart, Zina Ben Miled, and Robert J Oppelt. Gene clustering using self-organizing maps and particle swarm optimization. In *Parallel and Distributed Processing Symposium, 2003. Proceedings. International*, pages 10–pp. IEEE, 2003.

[40] Jin Xiong. *Essential bioinformatics*. Cambridge University Press, 2006.

[41] Ka Yee Yeung and Walter L Ruzzo. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[42] 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.

[43] Yongan Zhao, Haixu Tang, and Yuzhen Ye. Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, 2012.

[44] Xiaobo Zhou, Xiaodong Wang, Edward R Dougherty, Daniel Russ, and Edward Suh. Gene clustering based on clusterwide mutual information. *Journal of Computational Biology*, 11(1):147–161, 2004.

Appendix A

Copyright Permission of the Figure from CLIME

ELSEVIER LICENSE TERMS AND CONDITIONS

This is a License Agreement between Chaoyue Liu ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Chaoyue Liu
Customer address	5252 Tobin Street Halifax, NS B3H4K2
License number	3854320510079
License date	Apr 22, 2016
Licensed content publisher	Elsevier
Licensed content publication	Cell
Licensed content title	Expansion of Biological Pathways Based on Evolutionary Inference
Licensed content author	Yang Li,Sarah E. Calvo,Roee Gutman,Jun S. Liu,Vamsi K. Mootha
Licensed content date	3 July 2014
Licensed content volume number	158
Licensed content issue number	1
Number of pages	13
Start Page	213
End Page	225
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 1

Title of your thesis/dissertation	GENE CLUSTERING BASED ON CO-OCCURRENCE WITH CORRECTION FOR COMMON EVOLUTIONARY HISTORY
Expected completion date	Apr 2016
Estimated size (number of pages)	60
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 CAD
VAT/Local Sales Tax	0.00 CAD / 0.00 GBP
Total	0.00 CAD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be

deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the

copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).
If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript
    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not