



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

THE SUPERMATRIX APPROACH ON INFERRING PHYLOGENY THROUGH BIOINFORMATICS FRAMEWORK

WONG EE BHEI



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

SULTAN IDRIS EDUCATION UNIVERSITY

2024



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

THE SUPERMATRIX APPROACH ON INFERRING PHYLOGENY THROUGH
BIOINFORMATICS FRAMEWORK

WONG EE BHEI

DISSERTATION PRESENTED TO QUALIFY FOR A MASTER OF SCIENCE
(RESEARCH MODE)

FACULTY OF SCIENCE AND MATHEMATICS
SULTAN IDRIS EDUCATION UNIVERSITY

2024



Please tick (✓)

Project Paper

Masters by Research

Master by Mixed Mode

PhD

/

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on **09 JUL 2024**.

i. Student's Declaration:

I, WONG EE BHEI, M20182002208, FACULTY OF SCIENCE AND MATHEMATICS hereby declare that the work entitled THE SUPERMATRIX APPROACH ON INFERRING PHYLOGENY THROUGH BIOINFORMATICS FRAMEWORK is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

Signature of the student

ii. Supervisor's Declaration:

I, RAJA FARHANA RAJA KHAIRUDDIN hereby certifies that the work entitled THE SUPERMATRIX APPROACH ON INFERRING PHYLOGENY THROUGH BIOINFORMATICS FRAMEWORK was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a ~~*partial~~/full fulfillment for the conferment of MASTER OF SCIENCE (BIOLOGY), and the aforementioned work, to the best of my knowledge, is the said student's work.

10 JUL 2024

Date

Signature of the Supervisor



**INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES**

**BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM**

Tajuk / Title: THE SUPERMATRIX APPROACH ON INFERRING PHYLOGENY
THROUGH BIOINFORMATICS FRAMEWORK

No. Matrik / Matric's No.: M20182002208

Saya / I: WONG EE BHEI

(Nama pelajar / Student's Name)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-

acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (✓) bagi pilihan kategori di bawah / Please tick (✓) for category below:-


☐ **SULIT/CONFIDENTIAL**

Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / Contains confidential information under the Official Secret Act 1972


☒ **TERHAD/RESTRICTED**

Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / Contains restricted information as specified by the organization where research was done.

☐ **TIDAK TERHAD / OPEN ACCESS**


(Tandatangan Pelajar/ Signature)

Tarikh: 10 JUL 2024


(Tandatangan Penyelia / Signature of Supervisor)
& (Nama & Cop Rasmi / Name & Official Stamp)

Raja Farhana R. Khairuddin (PhD)
Senior Lecturer
Faculty of Science & Mathematics
Universiti Pendidikan Sultan Idris

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGEMENT

I extend my heartfelt gratitude to the individuals who have played pivotal roles in shaping my academic journey and contributing to the successful completion of this dissertation. First and foremost, I owe an immense debt of gratitude to my supervisor, Dr. Raja Farhana. Spending a decade under her guidance has been a transformative experience. She does not only introduce me to the captivating world of the field but also provided unwavering support, encouraging me to challenge my limits. Transitioning to bioinformatics from a social science background has been really challenging. Her trust in my abilities, often surpassing my own, has been a driving force throughout this journey. A true exemplar of the proverb, "Nothing is impossible," she consistently presented solutions even in the face of challenges. Beyond academia, she imparted invaluable life lessons that will stay with me always. I would like to express my profound appreciation to my Pn. Marina. Despite the occasional intimidation, I deeply respect her, having learned not only academically but also about the art of being a commendable teacher and student. Her guidance has been instrumental in my growth. My sincere thanks go to my examiners, Dr. Remmy, Dr. Nurbaya (UPM), Mr. Zahid, and Prof. Madya Dr. Annie. Their constructive feedback and insightful comments have significantly enhanced the quality of my writing and scientific communication skills. A heartfelt acknowledgment is extended to my labmates and friends for their unwavering emotional and physical support, especially during critical times, celebrating small successes and infusing my study life with vitality. I am also grateful to the university and faculty staff for their assistance in navigating procedural matters throughout my academic endeavors. Lastly, I dedicate a special note of appreciation to my family especially my parents. Their enduring trust and freedom have been the bedrock of my achievements. To mom and dad, thank you for always being there during my ups and downs. I am happy to share that I have completed my studies, and this accomplishment is as much yours as it is mine.

ABSTRACT

The objective of the study was to develop a supermatrix framework by incorporating the unique evolutionary signals of each gene that can improve phylogenetic inferences. The aim was to address the issue of the supermatrix approach, which potentially disregards individual gene properties. The study involved two main experiments: i) Experiment 1 focused on the effect of evolutionary signals across housekeeping genes on phylogeny inference using Chlorellaceae species. ii) Experiment 2 involved the development of a bioinformatics framework using the supermatrix approach. The framework incorporates unique properties of each gene in phylogenetic tree construction, such as sequence heterogeneity, sequence informative sites, taxonomic conflicts at the kingdom level, and tree distance between the gene tree and species tree. The genes were concatenated, based on their similar gene properties, into a supermatrix for phylogeny inference. Generated phylogenetic tree was compared with tree-of-life, which was used as a benchmark dataset, to validate the usability of the supermatrix framework. Our findings revealed that each individual housekeeping gene has different evolutionary signals, and ignoring these signals would affect the inferred phylogeny. The developed bioinformatics framework demonstrated an improvement in the accuracy of the inferred phylogenetic trees compared to the conventional tree inference approach based on Robinson-Foulds tree distance and Shimodaira Hasegawa test. This framework also corroborates with previous studies, which suggest that incorporating more genes in the supermatrix approach can enhance phylogenetic inference. Analysing individual gene properties by considering the unique evolutionary signals in gene concatenation through the supermatrix approach could improve phylogenetic inferences. The improvement of using the supermatrix approach could enhance the understanding of the evolutionary relationships between species, which further could be applied in various fields such as biodiversity conservation, medicine and healthcare.



PENGUNAAN KAEDAH SUPERMATRIK TERHADAP INFERENS FILOGENI MELALUI KERANGKA BIOINFORMATIK

ABSTRAK

Objektif kajian ini adalah untuk membangunkan kerangka kerja supermatrik dengan menggabungkan isyarat evolusi unik setiap gen bagi menambahbaik inferens filogeni. Tujuan utama kajian ini adalah untuk menangani isu kaedah supermatrik yang berpotensi mengabaikan sifat individu gen. Kajian ini melibatkan dua eksperimen utama: i) Eksperimen 1 fokus kepada kesan isyarat evolusi terhadap gen rujukan dalam inferens filogeni dengan menggunakan spesies Chlorellaceae. ii) Eksperimen 2 melibatkan pembangunan kerangka kerja bioinformatik melalui pendekatan supermatrik. Kerangka kerja ini menggabungkan kaedah supermatrik dengan mengambil kira sifat unik setiap gen seperti sifat keheterogenan jujukan, tapak maklumat jujukan, konflik taksonomi di peringkat alam, dan jarak filogeni antara pokok gen dengan pokok spesies. Gen yang mempunyai sifat yang serupa telah digabungkan menjadi satu supermatrik untuk inferens filogeni. Pokok-pokok filogenetik yang dihasilkan telah dibandingkan dengan pokok kehidupan yang menggunakan set data rujukan utama bagi mengesahkan kebolegunaan kerangka kerja supermatrik. Penemuan kajian kami mendedahkan bahawa gen rujukan individu mempunyai isyarat evolusi yang berbeza, dan pengabaian kepelbagaian isyarat evolusi akan menjejaskan filogeni yang dihasilkan. Kerangka kerja bioinformatik yang telah dibangunkan menunjukkan peningkatan terhadap ketepatan pokok filogenetik yang dihasilkan berbanding dengan pokok yang terhasil melalui kaedah inferens filogeni konvensional berdasarkan jarak pokok Robinson-Foulds dan ujian Shimodaira Hasegawa. Kerangka kerja ini juga menyokong kajian lepas dimana penggunaan lebih banyak gen dalam pendekatan supermatrik dapat mengukuhkan inferens filogenetik. Analisa sifat gen dengan pertimbangan terhadap isyarat evolusi yang unik pada setiap gen bagi kaedah supermatrik dapat meningkatkan kualiti inferens filogeni. Penambahbaikan kaedah supermatrik boleh meningkatkan pemahaman tentang hubungan evolusi antara spesies, yang boleh diaplikasi dalam pelbagai bidang, seperti pemuliharaan biodiversiti, perubatan dan penjagaan kesihatan.



CONTENTS

	Page
DECLARATION OF ORIGINAL WORK	ii
DECLARATION OF DISSERTATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
ABSTRAK	vi
CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiv
CHAPTER 1 INTRODUCTION	
1.1 Background Study	1
1.2 Problem Statement	5
1.3 Research Objectives	9
1.4 Research Questions	9
1.5 Significance of Study	10
CHAPTER 2 LITERATURE REVIEW	
2.1 Approaches in phylogenetic inference	12
2.1.1 Phylogenetic tree	13
2.1.2 Sequence alignment for phylogenetic analysis	15
2.1.3 Tools for sequence alignment	18

2.1.4	Evolutionary model for phylogenetic analysis	19
2.1.5	Phylogenetic analysis	22
2.1.6	Reliability of phylogenetic analysis	28
2.2	Phylogenetics and evolution	30
2.2.1	Homology in evolution	30
2.2.2	Tree of life	34
2.2.2.1	Gene tree and species tree	36
2.2.3	Measuring the diversity of life and species	38
2.2.4	Supermatrix and supertree approaches in the phylogenetic analysis	39
2.3	Sequence dataset for phylogenetic analysis	42
2.3.1	A case study on Chlorellaceae family	44
2.3.2	COGs for tree of life	45
2.3.3	Characteristics of the sequences	45
2.3.3.1	GC content	46
2.3.3.2	Genetic distance	47
2.3.3.3	Transition and transversion ratio	48
2.3.3.4	Disparity index	49
2.3.3.5	Sequence informative site	50
2.3.3.6	Tree distance	51

CHAPTER 3 ASSESSING SEQUENCE HETEROGENEITY IN CHLORELLACEAE HOUSEKEEPING GENES FOR PHYLOGENETIC INFERENCE

3.1	Introduction	523
3.2	Methodology	56
3.2.1	Sequence retrieval and filtration on Chlorellaceae species	56

3.2.2	Inference of the reference gene trees	58
3.2.3	Inference of the single gene trees	58
3.2.4	Gene sequence characteristics analysis	59
3.2.4.1	GC content	60
3.2.4.2	Genetic distance and Ts/Tv ratio	60
3.2.4.3	Disparity index	61
3.2.4.4	Tree topology incongruence	61
3.2.4.5	Tree congruency test between single gene trees	62
3.2.5	Supermatrix analysis	62
3.2.6	Supermatrix analysis with alternate gene arrangement	63
3.3	Result and Discussion	65
3.3.1	Identification of the housekeeping genes	65
3.3.2	Sequence characteristics of 18S, ITS and <i>rbcL</i> genes	67
3.3.3	Discordance between Chlorellaceae gene trees	72
3.3.4	Effect of supermatrix approach in accommodating the sequence heterogeneity for the Chlorellaceae species phylogeny	77
3.4	Conclusion	80
CHAPTER 4 THE EFFECT OF THE SUPERMATRIX FRAMEWORK IN INFERRING THE TREE OF LIFE		
4.1	Introduction	81
4.2	Benchmark dataset for the tree of life	83
4.2.2	Designing bioinformatics framework	83
4.2.2.1	Sequence filtration	83
4.2.2.2	Parameter for COG clustering based on phylogenetic analysis	85
4.2.2.3	Inferring supermatrix tree	87

4.2.2.4	Evaluation of the supermatrix trees	89
4.2.2.5	Supermatrix tree with alternate arrangement	91
4.3	Result and Discussion	92
4.3.1	Sequence characteristics	92
4.3.1.1	Sequence heterogeneity	92
4.3.1.2	Informative site	94
4.3.1.3	Tree distance	96
4.3.1.4	Conflict between Kingdom taxonomy	98
4.3.2	Supermatrix framework	100
4.3.3.1	Phase I: Sequence retrieval and filtration	102
4.3.3.2	Phase II: COG clustering for supermatrix	103
4.3.3.3	Phase III: Supermatrix tree inference	109
4.3.3.4	Phase IV: Supermatrix analysis	112
4.3.4	Effect of gene concatenation on tree inference	113
4.3.4.1	COG clustering level	113
4.3.4.2	Parameter level	120
4.3.5	Best supermatrix tree	132
4.4	Conclusion	133
CHAPTER 5	CONCLUSION	135
REFERENCE		139
APPENDIX		



LIST OF TABLES

Table No.		Page
2.1	Comparison of phylogenetic analysis methods	27
3.1	K2P genetic distance and disparity index (I_D) between and within genus for 18S, ITS and <i>rbcL</i> genes	70
3.2	Assessment of the congruency among single gene trees and supermatrix trees of 43 sequences of 14 Chlorellaceae species using normalized Robinson-Foulds distances (nRF) and Shimodaira Hasegawa test (SH-test)	75
3.3	Normalised Robinson Foulds and SH-test supermatrix genes with alternative gene arrangement	79
4.1	Cluster of orthologous genes (COGs) included in the study	84
4.2	Details of the dataset and model used to infer reference trees	90
4.3	Percentage range of each parameter for the COGs	104
4.4	Distribution of COG genes in each cluster for the sequence heterogeneity parameter	105
4.5	Model used for each COG cluster in parameter supermatrix tree	110
4.6	COG clusters across all parameters	114
4.7	Tree distances between each COG cluster tree and their parameter tree	121
4.8	Percentage of normalized Robinson-Foulds distance between the parameter supermatrix trees and the reference trees	132





LIST OF FIGURES

Figure No.		Page
2.1	Structure of a phylogenetic tree	14
2.2	Rooted tree and unrooted tree	14
2.3	Phylogenetic tree showing type of paralogs	31
2.4	Phylogenetic tree showing orthologous relationships	31
2.5	Detection of orthologous group	33
3.1	Two phases of sequence filtration for 18S, ITS and <i>rbcL</i> genes	57
3.2	Schematic diagram of gene concatenation matrix	63
3.3	Schematic diagram of gene concatenation matrix with alternate gene arrangement	64
3.4	Presence and absence of 18S, ITS and <i>rbcL</i> in all Chlorellaceae species	66
3.5	Distribution of GC content (%) across the Chlorellaceae	67
3.6	Transition/transversion ratio (Ts/Tv) against K2P distance	68
3.7	Maximum likelihood (ML) trees for all single and concatenated genes	73
4.1	Sequence characteristics employed for COG clustering	86
4.2	Sequence heterogeneity for each COG	93
4.3	Informative sites across the COG	95
4.4	Tree distance of each COG	97
4.5	Conflict between Kingdom taxonomy for each COG tree	99



4.6	Overview of the supermatrix framework	101
4.7	Flowchart I: Sequence retrieval and filtration	102
4.8	Flowchart II: COG clustering for each parameter	103
4.9	Flowchart III: Supermatrix tree inference	112
4.10	Flowchart IV: Supermatrix analysis	112
4.11	Tree topologies for each COG cluster based on tree distance parameter	115
4.12	Tree topologies for each COG cluster based on sequence heterogeneity parameter	116
4.13	Tree topologies for each COG cluster based on informative site parameter	117
4.14	Tree topologies for each COG cluster based on kingdom conflict parameter	118
4.15	Tree topology differences between four-parameter supermatrix trees according to their Kingdoms	122
4.16	Tree topology differences between reference trees of life	124
4.17	Comparison of tree distances for (i) parameter supermatrix tree for P1-P4, (ii) cluster supermatrix tree, and (iii) individual COG trees to the reference trees	126
4.18	SH-test result for the (i) reference trees; (ii) parameter supermatrix trees; (iii) cluster supermatrix trees; (iv) individual COG trees	127
4.19	Distribution of tree distances between the reference tree and C30P tree with (i) individual COG trees, (ii) cluster supermatrix trees, (iii) parameter supermatrix trees and (iv) reference trees	129
4.20	The frequency and presence of the COG in Q2 and Q3 regions among all cluster supermatrix tree in Figure 4.21	130
4.21	Conflicting branches between P2 & P3 and P1 & P4 parameter supermatrix trees	133

LIST OF ABBREVIATIONS

AICc	Akaike Information Criterion
AT	Adenine-Thymine
C30	Reference tree from 30 concatenated genes
C30P	C30 with partitioned model
CC	Reference tree from Ciccarelli's study
COG	Cluster of Orthologous Group
COI	Cytochrome oxidase subunit I
COVID-19	Coronavirus disease 2019
FBP	Felsenstein Bootstrap Proportion
GC	Guanine-Cytosine
GTR	Generalised Time-reversible
I _D	Disparity Index
ITS	Internal Transcribed Spacer
JTT	Jones-Taylor-Thornton
K2P	Kimura-2-parameter
LG	Le-Gascuel

ML	Maximum Likelihood
MP	Maximum Parsimony
NCBI	National Centre of Biotechnology Information
NJ	Neighbour Joining
NNI	Nearest Neighbour Interchange
nRF	normalised Robinson-Foulds
OTU	Operational taxonomic unit
P1	Parameter tree based on tree distance
P2	Parameter tree based on sequence heterogeneity
P3	Parameter tree based on informative site
P4	Parameter tree based on kingdom conflict
rbcL	ribulose biphosphate carboxylase large chain
RNA	Ribonucleic acid
SH test	Shimodaira-Hasegawa test
SPR	Subtree Pruning and Regrafting
STRING	Search Tool for Recurring Instances of Neighbouring Genes
TBE	Transfer Bootstrap Expectation
TIM	Transitional Model
Ts/Tv ratio	Transition/Transversion ratio



APPENDIX

- A Accession number and GC content of 18S, ITS and rbcL markers.
- B Distribution of the species included in the study at each taxonomic rank



CHAPTER 1

INTRODUCTION

The total number of species on Earth is estimated to range from millions to billions (Larsen, Miller, Rhodes, and Wiens, 2017; Mora, Tittensor, Adl, Simpson, and Worm, 2011). However, more than 80% of these species remain unidentified (Mora, Tittensor, Adl, Simpson, and Worm, 2011). The theory of evolution explains how one life form changes over time and how life diversifies from the origin of the species. Charles Darwin sketched the first phylogenetic tree in 1837, which modelled the relatedness of one species with another. A phylogenetic tree depicts the concept of evolutionary relatedness between species, with the trunk of the tree representing the common ancestor of the species and branches showing the diversification or evolution of the parental species to its offspring. The phylogenetic tree, which portrays the evolutionary relationship and relatedness among different organisms and is universal, where it is



capable of inferring the evolutionary relationship of all cellular organisms, is called the tree of life.

Phylogenetics is the study of phylogeny or evolutionary relationships between organisms. Previously, when molecular data were not widely available, researchers investigated the evolution of species' morphological characteristics. With the advent of sequencing technology, enormous molecular sequence data has been produced. The availability of this sequence data has resulted in the increasing importance of phylogenetic studies to make sense of sequence data. Moreover, evolutionary history can be better inferred, especially with the use of molecular sequence data compared to morphological data (Betancur et al., 2013; Yang and Rannala, 2012).



relatedness between species. For example, the chimpanzee and bonobos have been known to be the closest living relative of human in the animal kingdom (Prüfer et al., 2012). This relationship could not be deciphered solely based on morphological information, as the chimpanzee and bonobos do not look or act like humans but share 99% of their DNA. Understanding the exact evolutionary relationship has always been a challenge in phylogenetic research. Thus, inferring an accurate phylogenetic tree is important to illustrate the correct evolutionary relationship between species.

To infer the evolutionary relationships between organisms, phylogeny was inferred using phylogenetic analysis. The common flow in the phylogenetic analysis was initiated with the retrieval of the sequences of interest, followed by sequence alignment before inference of phylogeny. Selection of the sequence of interest is vital





in phylogenetic analysis as it helps to infer a better evolutionary relationship between species. Housekeeping and orthologous genes are ideal genetic information used as inputs for phylogenetic analysis. Housekeeping genes are responsible for the maintenance of basic cellular functions and evolve slower than other genes (Joshi, Ke, Drangowska-Way, O'Rourke, and Lewis, 2022). On the other hand, orthologous genes, which are derived from a common ancestor, separated only by a speciation event, and are conserved, are also favourable in phylogenetic analysis (Koonin, 2005a).

Generally, a phylogenetic tree inferred from single-gene information is known as a gene tree. In contrast, when multiple gene information is accommodated in tree inference, the output tree is known as a species tree. A gene derived from a common ancestral gene in different species carries useful information to infer a phylogenetic tree; however, previous studies have emphasised that a gene tree is not a reliable estimate of species evolutionary history (Forterre, 2015; Low, Džunková, Chaumeil, Parks, and Hugenholtz, 2019; Thiergart, Landan, and Martin, 2014). Therefore, phylogeny should be constructed using multiple genes to generate a more reliable and accurate evolutionary relationship between species (Dong et al., 2022).

However, considering multiple gene information in tree inference can be challenging, as each gene has evolved independently and possesses conflicting evolutionary stories throughout the genome. The conflicts in the evolutionary story of the gene tree and species tree are known as phylogenetic incongruence. Incongruence could possibly occur due to errors during tree inference, for example, insufficient sequence length, sampling error, or failure of the reconstruction method to account for the properties of the sequence data. Evolutionary events such as incomplete lineage



sorting, horizontal gene transfer, gene loss, and gene duplication, which occur independently throughout the genome, could also contribute to incongruence. Thus, selecting suitable genes and the tree inference method can be used to infer the evolutionary relationships between species more accurately.

The inference of evolutionary species between the Chlorellaceae species is commonly carried out based on the standard housekeeping genes, that is, the 18S ribosomal RNA (Khaw, Khong, Shaharuddin, and Yusof, 2020). Since using a single gene to infer the phylogenetic tree could miss out important evolutionary signals, other housekeeping genes such as ITS and *rbcL* were also included in the phylogenetic analysis of the Chlorellaceae species. Chlorellaceae is one of the most prominent taxonomic families of microalgae with 222 species across 56 genera. Given its desirable characteristics in offering new possibilities for more effective and affordable alternative energy resources, the evolutionary relationships between the species inferred from previous studies conflicted with one another, resulting in an obscure taxonomy and phylogenetic relationship between the species.

Supermatrix approach is one of the methods to reconstruct evolutionary relationship between species, by combining sequence data from multiple gene or genomic regions into a single, concatenation matrix known as a supermatrix. Unlike the conventional tree inference method that focus only on individual genes, it analyses data from the concatenation matrix, as a combination. Thus, the approach is capable of leveraging larger amount of information, for more robust phylogenetic inference. However, the phylogenetic signals in each gene could have been mixed or averaged up when being analysed as a whole in the concatenation matrix.

Therefore, in this study, the effect of the supermatrix approach by incorporating gene characteristics to improve phylogenetic inference was investigated. The study is divided into two experiments. A preliminary study initially explored the impact of the supermatrix approach using housekeeping genes on the phylogenetic inference of Chlorellaceae species. However, limitations in available Chlorellaceae sequence data restricted further analysis. Thus, the study pivoted to the COG dataset use in a prior established tree of life study. Leveraging these benchmark orthologous sequences, a bioinformatics framework was designed to infer the Tree of Life. From these experiments, the gene characteristics between the housekeeping genes of the Chlorellaceae species and the tree of life could be assessed, and the effect of the bioinformatics framework incorporating these gene characteristics could be identified. This study could provide insights into the use of gene characteristic information in

1.2 Problem Statement

Although incorporating multiple gene information in inferring phylogeny can be more convincing, it has been hampered by phylogenetic discordance, which arises from the conflict between different gene information. The conflicting signal carried by each gene makes inferring a species tree challenging to accommodate the heterogeneous information from these genes. The common practice for handling phylogenetic discordance was to employ the supermatrix approach in phylogenetic analysis.



The supermatrix approach has been proven to resolve the conflict between heterogeneity and to increase phylogenetic resolution (Dong et al., 2022). This approach considers all gene information and analyses it as a whole. Although accommodating more genes in the supermatrix approach can improve the resolution of the inferred trees, this approach tends to ignore differences or genetic variation by randomly averaging conflicting signals (Lartillot, Brinkmann, and Philippe, 2007; Smith, Walker-Hale, Walker, and Brown, 2020). Although averaging the conflicting phylogenetic signals can improve the resolution of the tree, each of the signals might carry important evolutionary information, and they are worthy of attention for elucidating the comprehensive story between species.

Research has explored various evolutionary complexities that pose challenges to supermatrix-tree inference (Gatesy et al., 2019). To accommodate genetic variation, evolutionary models have been developed and implemented in tree inference to capture the substitution process pattern from heterogeneous sequence data. Selecting an appropriate evolutionary model is an important procedure in phylogenetic tree inference, as it can affect the accuracy of the phylogenetic tree, especially in certain character-based tree inference methods that rely on evolutionary models, such as maximum likelihood and Bayesian models.

The challenge of model usage in tree inference is that it is expected to account for all variations between different genes. Partitioned models were introduced decades ago to improve the usage of a homogenous model (single model for tree inference of entire supermatrix alignment) in heterogeneous datasets by allowing the use of independent models at each gene partition for multiple gene datasets (Bull and



Swofford, 1993; Yang, 1996). However, the model setting has been claimed to cause over-partitioning and poor parameter estimation (Redmond and McLysaght, 2014; Redmond and McLysaght, 2021). Despite the importance of evolutionary model usage in phylogenetic tree inference, the optimal method for selecting the best partition scheme remains unclear.

Accurate identification and classification of species are crucial for effective conservation efforts and understanding their potential applications. The Tree of Life (TOL) is a metaphor used to describe a branching diagram that could represent evolutionary history of living organisms on Earth. However, constructing TOL as a benchmark for all phylogenetic studies presents several challenges. The evolutionary history is not always a simple and direct branching process. Evolutionary events such as horizontal gene transfer, coalescent and incomplete lineage sorting could lead to complex relationships that are difficult to capture accurately in a single tree.

Currently, there have been two established TOL studies carried out, which covers all the three domains of life (Ciccarelli et al., 2006; Hug et al., 2016). Both of the TOLs included the cluster of orthologous groups (COG) to infer the evolutionary relationships of species. Ciccarelli's dataset consisted of 30 COGs from hundreds of organisms, covering the three domains of life; Eukaryote, Archaea and Bacteria. Ciccarelli employed supermatrix approach to randomly concatenated all the identified COGs for maximum likelihood tree inference using JTT+I+ Γ (Jones et al., 1992) model. This method aimed to capture the maximum amount of evolutionary signal, but it also introduced potential biases associated with concatenating sequences with varying evolutionary rates.



While, Hug's dataset covered thousands of organisms from 16 COGs, but focusing more on the bacterial and archaeal domains. Similarly, the 16 COGs were concatenated randomly to form a super-matrix of sequences for maximum likelihood tree inference. The best model employed for the dataset was LG+ Γ (Le and Gascuel, 2008). This study aimed to provide a more detailed understanding of relationships within these specific domains, acknowledging the limitations of a universal TOL for capturing the intricacies of bacterial and archaeal evolution.

The conventional supermatrix approach employed often lack of precision needed, particularly for intricate groups, such as the Chlorellaceae family. The complex evolutionary history and morphological similarities within this family pose significant challenges to their classification, hindering their potential as sustainable biofuel sources.

The evolutionary history of Chlorellaceae species was deemed complex and cryptic, possibly because of rapid diversification and morphological similarities. Studies have struggled to accommodate the diversity of gene properties and intricate relationships, leading to conflicting interpretations. Different genes within the Chlorellaceae genome evolve at varying rates and respond differently to evolutionary pressures, contributing to heterogeneity among the genes. This confounds conventional tree-building approaches, leading to incongruent results based on the chosen genes.

Therefore, the critical challenge lies in reconciling the ambiguity present in the current phylogenetic reconstructions and revealing the true evolutionary relationships within the Chlorellaceae family. This necessitates a novel approach that accounts for the complexities of evolutionary history, diverse sequence characteristics, and functional adaptation. Addressing this challenge is not only essential for maximising



the biofuel potential of Chlorellaceae but also for improving understanding of their role in the intricate tree of life.

1.3 Research Objectives

This study has three main objectives.

- a. To identify suitable housekeeping genes as the backbone of the phylogeny.
- b. To design a bioinformatics framework for supermatrix analysis to infer phylogeny.
- c. Implementing the proposed framework for inferring phylogeny.

1.4 Research Questions

- a. Which housekeeping genes are most suitable for inferring phylogeny?
- b. What is the bioinformatics framework of the supermatrix approach for inferring phylogeny?
- c. How does the proposed framework of the super-matrix approach affect the resulting phylogeny?



1.5 Significance of Study

The accurate identification of species is important for conservation and extinction assessments. Phylogenetic trees, visualised as tree-like diagrams with branching structures, offer crucial insights into the evolutionary relationships between species by revealing their shared ancestry and evolutionary history. This information empowers researchers to understand the characteristics of each species, their relationships with other organisms, and their potential vulnerability. Furthermore, by tracing the branches back to the shared ancestors, phylogenetic trees provide valuable insights into the evolutionary history of each species, including their adaptations, shared traits, and potential threats (Pinto-Ledezma, Diaz, Halpern, Khoury, and Cavender-Bares, 2023). This allows them to distinguish between closely related species, particularly in intricate groups where traditional methods fall short, thereby ensuring accurate identification and classification (Shahzad et al., 2020).

The Chlorellaceae family, with rapid growth on non-arable land and high lipid content, is an attractive candidate for sustainable biofuel production (Feng et al., 2020; Lal, Banerjee, and Das, 2021). However, their complicated evolutionary history, due to their cryptic diversity and morphological similarities, poses challenges (Krivina, Temraleeva, and Bukin, 2021; Malavasi, Škvorová, Němcová, and Škaloud, 2022; Yuan et al., 2020). Distinguishing between closely related species can be a daunting task, hindering our understanding of their potential. In such situations, phylogenetic trees can serve as an invaluable tool to offer perspectives on the evolutionary history of a species.





Phylogenetic trees serve as more than just evolutionary trees; they can be powerful tools to unravel the relationships between species, guide conservation efforts, and unlock the potential hidden within the diverse tapestry of life. However, the accuracy and reliability of trees rely heavily on inference methodologies. The inability of the tree inference method to account for varying evolutionary rates and processes across different genes introduces uncertainties that affect the accuracy of tree inference. This study incorporated gene characteristics in the supermatrix approach, where genes are grouped based on shared properties, resulting in smaller and more homogeneous supermatrices. Different evolutionary models were applied to each gene cluster to account for their specific characteristics and reduce the impact of inappropriate model assumptions.



more accurate phylogenetic inferences than the conventional single-model supermatrix method. In addition, grouping genes with similar properties minimises conflicting signals between gene trees, thereby improving the overall robustness of the inferred tree. This framework could be beneficial for analysing complex groups with high heterogeneity and provides an alternative for exploring the evolutionary dynamics of the diversity of genes within complex groups of species, offering better insights into their evolutionary history and relationships.

