View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Sydney eScholarship

A Novel Quartet-Based Method for Inferring Evolutionary Trees from Molecular Data

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Information Technology
Faculty of Engineering and Information Technology

Monther Tarawneh
March 2008

Science is nothing but perception
Plato (Ancient Greek
Philosopher
428-348 BC)

To my mother

To my father

أهداء الـى امي وابي

Acknowledgment

First and foremost, I would like to thank my supervisor Dr. Bing Bing Zhou. Without him this thesis would not have been possible. I am grateful to his kind supports, excellent advice, intensive collaboration throughout the whole time we were working together, for intensively reading this thesis, and the financial support he had gave me during my PhD research. Furthermore, I thank my associate supervisor Prof. Ablert Zomaya for his support and opportunity to work in the environment of advanced network research. Also, for his encouragement and help during my PhD.

Thanks to Chen, Daniel, Penghao, and every one attended our very intensive enjoyable collaboration during the last 3 years, every meeting result in new idea, some key definition and deep sight. I am especially grateful to Daniel Chu who contributed by implementing the distributed version of QB algorithm.

I appreciate very much the unfinished work with Abdur Sikder and Gowrie, it dose open my eyes into many future direction, hope we can finish.

I would like to thank every one in the ANRG group at the school of IT especially, Michael Charleston. Besides, Friday meeting is a good experience for every research student.

Thanks to David London, Greg and the workshop people for their technical support during my time in the school of information technology/University of Sydney.

I wish to thank several colleagues for their continuous support in any way, and creating such friendly environment especially Abdur, Gowrie, Shwen, Penghao, Khalid, Mohammad, Tanveer, and all members in the school of IT at Sydney University.

Finally, I would like to thanks my parent for their ever lasting support of my work and study.

Declaration

I, Monther Tarawneh, do hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, nor material which to a substantial extent has been excepted for the award of any degree or diploma of a university or other institute of higher learning, except where due acknowledgments were made in the text.

……………………..
Monther A Tarawneh

Abstract

Molecular Evolution is the key to explain the divergence of species and the origin of life on earth. The main task in the study of molecular evolution is the reconstruction of evolutionary trees from sequences data of the current species.

This thesis introduces a novel algorithm for inferring evolutionary trees from genetic data using quartet-based approach. The new method recursively merges sub-trees based on a global statistical provided by the global quartet weight matrix. The quarte weights can be computed using several methods. Since the quartet weights computation is the most expensive procedure in this approach, the new method enables the parallel inference of large evolutionary trees.

Several techniques developed to deal with quartets inaccuracies. In addition, the new method we developed is flexible in such a way that can combine morphological and molecular phylogenetic analyses to yield more accurate trees. Also, we introduce the concept of critical point where more than one possible merges are possible for the same sub-tree. The critical point concept can provide information about the relationships between species in more details and show how close they are. This enables us to detect other reasonable trees.

We evaluated the algorithm on both synthetic and real data sets. Experimental results showed that the new method achieved significantly better accuracy in comparison with existing methods.

List of Figures

Figure 1.1 Growth of public database GenBank…………………………………...2

Figure 1.2 Rooted phylogenetic tree, colored according to the three-domain system…...……………………………………………….3

Figure 2.1 Rooted and unrooted phylogenetic trees...………………………………13

Figure 2.2 Basic steps in phylogenetic analysis……………..……………………15

Figure 2.3 Equal substitution rates in JC69 model………………………………….20

Figure 2.4 Representations of GTR Parameters…………………………………….23

Figure 2.5 Computation of a Parsimony score resulting in two possible assignments……………………………………………….…32

Figure 2.6 Computation of a likelihood value………………………………………35

Figure 2.7 Step-wise addition showing all the possible places for the fifth specie…………………………………………………..42

Figure 2.8 Sub-tree Pruning and Regrafting (SPR)…………………………………43

Figure 2.9 A possible bisection and reconnection (TBR)…………………………...44

Figure 2.10 A possible Nearest Neighbour Interchange (NNI) exchanges ……………………………………………….………...45

Figure 3.1 The seven topologies for each quartet…………………………...…51

Figure 3.2 The three possible fully resolved trees for a quartet {a, b, c, d}……………………………………………………………...54

Figure 3.3 A simple example of the Global Weight Matrix……………………...…55

Figure 3.4 Calculating the number of quartet trees using the bi-partitioning technique……………………….……………………..57

Figure 3. 5 Unrooted tree separated into three sub-trees by removing the internal node that connects them………………………...……..59

Figure 3.6 A simple example of tree reconstruction from a global quartet weight matrix ………………………………………..…65

Figure 4.1 An example of tree reconstruction from an inaccurate global quartet weight matrix………………..…………………………74

Figure 4.2 An example of a common critical point……………………………...…77

Figure 4.3 QBNJ workflow graph…………………………………………….….…79

Figure 4.4 Threshold α as a function of confidence value *c*………………………...81

Figure 5.1 The average RF values for the trees generated using QB4-ML and Phyml, where MD=0.3……………………………….93

Figure 5.2 The average RF values for the trees generated using QB4-ML and Phyml, where MD=0.3……………………………….94

Figure 5.3 The average RF values for the trees generated using QB4-ML and Phyml, where MD=0.3………………………………..94

Figure 5.4 The average RF values for the trees generated using QB4-ML and Phyml, where MD=0.3………………………………..95

Figure 5.5 Likelihood values for 1037 trees………………………...…………..98

Figure 6.1 Some of the excavates taxa and their relatives…………………………104

Figure 6.2 Three possible merge patterns at a critical point……………………..…108

Figure 6.3 Three possible combinations for Malawimonas………………………..109

Figure 6.4 Tree with multi-scenarios for Excavate taxa……………………….……..111

Figure 7.1 Master workflow graph…………………………………..……….117

Figure 7.2 Workers workflow graph…………..………………………………..118

Figure 7.3 Speedup vs 32 CPUs for different numbers of Sequences (or taxa)..…121

Figure 7.4 Percentage of Computation Time………………………………………122

Figure 7.5 Speedup of program with 256 CPUs………………………………..…..123

List of Tables

Table 1.1 Amino Acids…………………………………………………………….6

Table 1.2 Number of possible trees for n sequences…………………………………7

Table 5.1 Experimental results for sequence length 300……………………………86

Table 5.2 Experimental results for sequence length 600……………………………97

Table 5.3 Comparison of variable and fixed α in QB for DNA
sequence lengths of 300 .....………………………………………...….89

Table 5.4 Results using QB4 with ML for sequence data set
lengths of 300... ….…………………………………………..................92

Table 5.5 Comparative results for a 24-sequences data set………………………....95

Table 5.6 Comparative results using QBNJ and ML criterion for the
24 sequences……………………………………………………….………..96

Table 5.7 Average RF values for the 24 sequences trees…………………………...97

Table 6.1 The 24 species included in the experiment……………………...……103

Table of Contents

Acknowledgments……………………………………………………………...iv
Declaration...…………………………………………………………………v
Abstract………………………………………………………………………vi
List of Figures………………………………………………………………vii
List of Tables…………………………………………………………………...ix

Chapter 1……………………………………………………………………..1
Overview……………………………………………………………………...1
1.1 Introduction…………………………………...………………….....1
1.2 Genetic Sequences……………………………….……………...…5
1.3 Problem Complexity…………………………………………………7
1.4 Contribution………………………………….………………………8
1.5 Thesis Structure……………………………………………………9

Chapter 2……………………………………………………………………12
Reconstructing Evolutionary Trees……………..…………………12
2.1 Introduction ……………………………………………………………12
2.2 Basic Concepts………………………………………………………14
2.3 Sequence Alignment……………………………………………...16
2.4 Models and Assumptions in phylogenetic.………………………….…18
2.4.1 Choosing the Appropriate Model…………………………………….24
2.5 Tree-Building………………………………………………………..25
2.5.1 Distance-Based Methods…………………………………………….26
2.5.1.1 Least Square Methods……………………………………………...26
2.5.1.2 Minimum Evolution Methods……………………………………...27
2.5.1.3 Clustering Methods…………………………………………….....29
2.5.1.4 Limitations of Distance-Based Methods……………….……..30
2.5.2 Character-Based Methods…………………………………………....30
2.5.2.1 Parsimony Criterion …………………………………...…..31
2.5.2.2 Maximum Likelihood Criterion…………………….……….. 33
2.5.2.3 Bayesian Inference………………………………………………38
2.6 Heuristic Search Methods……………………………………….….40
2.6.1 Stepwise Addition ………………………………………...…41
2.6.2 Tree Rearrangement………………………………………….…42
2.7 Measure of Confidence..……………………………………….…46
2.8 Testing and Comparing Programs………………………………..…47
2.9 Summary…………………………………………………………….48

Chapter 3……………………………………………………………………49
Global Quartet Weight Matrix…………………...……………………49
3.1 Introduction ………………………………………………………...49
3.2 Quartet Trees…………………………………..…………………50
3.3 Quartet Puzzling Algorithm……………………………………….....52
3.4 Global Quartet Weight Matrix……………………………………………54
3.5 Reconstructing QWM from A given Tree…………………………..…56
3.5.1 Efficient Algorithm ……………………………………………………58
3.6 Tree Reconstruction from QWM……………………………………………62
3.7 Summary…………………………………………………………….…67

Chapter 4……………………………………………………………………68
Design and Implementation of a Novel Quartet based Algorithm………………………………………………………………………68
4.1 Introduction………………………………………………………………68
4.2 Average Confidence Value………………………………………………70
4.3 Quartet Weight Correction……………………………………………..71
4.4 Example………………………………………………………………..73
4.5 Computational Cost……………………………………………………75
4.6 Critical Points and Multiple Trees...…………………………..……….…76
4.7 Invariable Threshold α ……………………………………………...80
4.8 Summery……………………………………………………………….81

Chapter 5……………………………………………………………………83
Performance Evaluation …..…………………………………………………83
5.1 Introduction……..……………………………………………………83
5.2 Data set...……….……………………………………………………84
5.3 Methods and Parameters……………………..……………………….…85
5.4 Result………………………………………………………………….…85
5.5 Invariable Threshold α……………………..…………………..…….89
5.6 QB with Maximum Likelihood (QB+ML)...…………………………….90
5.6.1 Experiment on 12 sequences……...…………………..…………91
5.6.2 Experiment on 24 Sequences.......……………………………..95
5.7 Multiple Maximum Likelihood Points…..…………………………..97
5.8 Summery………..……………………..……………………98

Chapter 6…………………………………………………………………..100
Testing the Relationships among Excavate Taxa……………………..100
6.1 Introduction…………………………………………………………...100
6.2 Material and Methods……………………………………………...102
6.2.1 Data Source……………………………………………………102
6.2.2 Phylogenetic analysis…………………………………………105
6.3 Experimental Result……………………………………………..106

6.3.1 Trimastix……………………………………………………………106
6.3.2 Parabasalids, Diplomonads, and Carpediemonas………………………107
6.3.3 Malawimonas……………………………………………………….107
6.3.4 Jakobids…………………………………………………………..110
6.3.5 Euglenozoa and Heterolobozea………………………………………..110
6.4 Summery…………………………………….................................112

Chapter 7………………………………………………………………..114
Parallel Implementation…..…………………………………………...114
7.1 Introduction.………………………………………………………… 114
7.2 The Parallel Algorithm…………………………………………………115
7.3 Experimental results…………………………………………………..119
7.4 Summary……………………………………………………………...124

Chapter 8……………………………………………………………..125
Discussion and future work……………………….…………………...125
8.1 Conclusion…………….……………………………………………...125
8.2 Future work.…………………………………………………………..128

References…..………………………………………………………….130