Gene Rearrangement and Molecular Evolution

Gene Rearrangement and Molecular Evolution in Animal Mitochondrial Genomes

By

Wei Xu, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2005) McMaster University

(Department of Physics and Astronomy) Hamilton, Ontario

TITLE: Gene Rearrangement and Molecular Evolution in Animal Mitochondrial Genomes

AUTHOR: Wei Xu, B.Sc.

SUPERVISOR: Paul G. Higgs

NUMBER OF PAGES: xi, 82

Abstract

Phylogenetic analysis of gene order data is a developing area, and many questions on gene orders are still unresolved. In this project, we started from the OGRe database, where we obtained the mitochondrial genome information (after some corrections), designed a logarithm correction for breakpoint distance, applied distance matrix methods to both breakpoint distance and the logarithm of breakpoint distance for gene orders, and then focused on Arthropoda phylogeny We tried many phylogenetic methods to infer Arthropod phylogeny; however, no method yielded a satisfying result. We constructed an Arthropod phylogenetic tree based on both molecular and morphological evidences. After we estimated the phylogenetic tree, we used maximum likelihood methods to estimate branch lengths for tRNAs and proteins, calculated the breakpoint numbers and inversion numbers for gene orders, and calculated the correlations among these four measures. We found that: when gene order rearrangements and mutations on sequences are small, the changes are independent, and, when the rearrangements and mutations are large, the changes seem to be correlated. The branch lengths in the tRNA and protein trees are highly correlated in low mutation situations and less correlated when mutation rates are larger.

Acknowledgements

I thank my supervisor, Paul Higgs, for his direction and discussions.

I also thank Jon Stone, James Wadsley and Jianping Xu for their discussions and suggestions.

I feel grateful to thank Ziyi Zhang, H.C. Lee and David Sankoff for their encouragement.

I enjoyed playing basketball with Xiaoguang Yang. He always encouraged me to take more shots and never blames me for my terrible performance.

I am indebted to Supratim Sengupta, Asmahan Abuarish, Felix Wong, Daniel Banks and David Cooke for their help all the time.

to the memory of my grandparents

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Although the first complete mitochondrial was only sequenced in 1981 (Anderson et al.), as technology develops, automatic genome sequencing is becoming common. We are now fortunate to have hundreds of complete mitochondrial genomes. This gives a valuable source of information for the research by computational biologists. It can help those researchers to potentially reveal the rules and facts about the processes that govern genome evolution. From these data, one can study from two different aspects: sequence analysis and gene order analysis.

Protein and nucleotide sequences have been used for phylogenetic inference for almost 40 years (Felsenstein 2004). Many researchers from different disciplines have dedicated themselves to design more realistic and powerful methods to make use of this information. However, some problems are difficult to solve. A purely randomly evolving molecule will be ideal for phylogenetic inference. Obviously, molecular sequences, either DNA sequences or proteins, are affected by selective pressure, which undermines the power of sequence analysis. Methodologically, many other problems make inference not as convincing as we have thought, long branch attraction, base and codon frequency bias as so on. One may either suffer from the problems or invest huge effort in trying to overcome them.

The gene order information can be a useful phylogeny inference resource. Many studies demonstrate that gene order can be used to infer phylogenetic questions. Some are based on intuitive descriptions and some are based on systematic comparisons. (Sankoff et al. 2000; Keogh et al. 1998) However, here are lots of questions that remain unanswered. Are gene orders really as useful as we have hoped? Why and how do gene orders change? What kind of rules govern their change? Are changes in gene orders neutral? If not, does selection pressure only have a limited influence on gene order, in which case we might have a better method than sequence analysis.

In this project, we want to investigate, in a systematic way, the correlation between sequence and gene order information. As molecular information has succeeded greatly, if gene order information also can be used in phylogenetic inference, we expect a positive correlation between results obtained from using these two different data types.

1.2 Methods to Infer the Phylogeny

1.2.1 Introduction to Phylogeny

Phylogeny is a hypothesis showing the evolutionary relationships between species with tree-like diagrams. The leaf nodes represent species at the terminals of the evolution (current day species and fossil taxa which have no descendants) and the inner nodes represent their ancestors. The lengths of the branches separating two species denote how distantly these two species are related. Sometimes branch lengths are drawn proportional to the time since divergence of the species. In other cases, the lengths are proportional to the amount of evolutionary change that has occurred on the branches.

The phylogenetic tree can be rooted, where there is a common ancestral species (or some characteristics of this common ancestor that we can know) for all the other species; or be unrooted, where the common ancestral species is unknown and we don't know the direction of the evolution either. If we neglect lengths of the branches and only consider how those nodes and lines are connected, the tree turns to a topology tree.

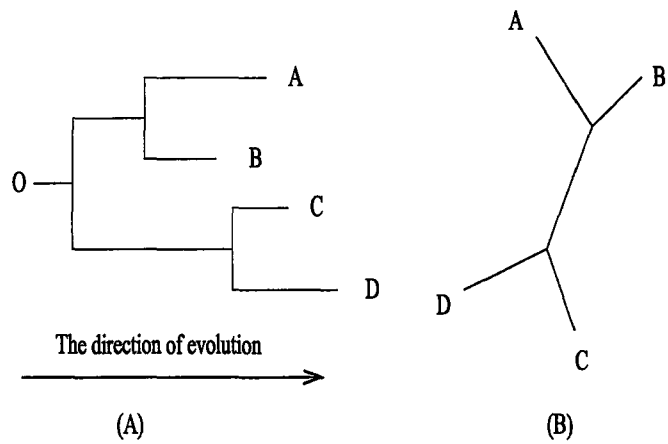

Figure 1.1: Examples of rooted tree (A) and unrooted tree (B)

The phylogeny tree can be bifurcate, where all leaf nodes are only connected by one branch and all inner nodes are connected by exact 3 branches (in rooted tree, the common ancestral node may be connected by 2 branches, and this is the only exception). For a bifurcate tree with n leafs (rooted or unrooted), the number of topology trees is $(2n-5)!!$ and $(2n-3)!!$ for unrooted and rooted trees correspondingly. The sign "!!" here is called the double factorial by definition of $(2i+1)!! = (2i+1) \times (2i-1) \times \ldots \times 3 \times 1$. When the number of leafs is large, the number of topology trees is huge.

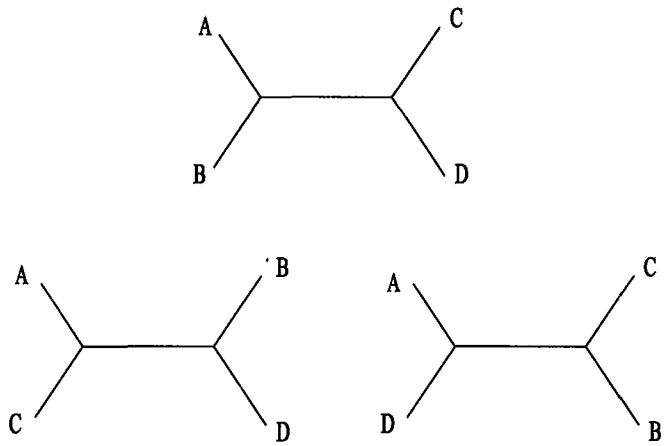

Figure 1.2: There are 3 unrooted topologies for 4 species

1.2.2 Phylogenetic Methods

Parsimony methods were the earliest developed. Edwards and Cavalli-Sforza (1963) stated that the evolutionary tree to be preferred is the one that involves "the minimum net amount of evolution". Parsimony methods search the tree space to find the tree with minimum number of total changes. There are many discussions of parsimony methods on statistics and philosophy levels (Felsenstein 2004). When evolutionary events rarely happen, the result from parsimony methods are accurate. However, when evolutionary changes are not so rare. Parsimony methods will underestimate the total number of changes. Nevertheless, currently, many of the algorithms to compare multiple gene orders still use parsimony methods, because in this situation the combinations of states for gene order are extreme huge and parsimony methods are one of the quick methods.

Distance matrix methods were introduced by Cavalli-Sforza and Edwards (1967) and by Fitch and Margoliash (1967). The main idea is to calculate a certain kind of distance measure for pairs of species and find the phylogenetic tree that predicts the previously calculated distances as close as possible. By defining different criteria

for closeness, we obtain different distance matrix methods. The-least squares and Fitch-Margoliash criteria are popular.

Unlike the other distance matrix methods, Neighbor Joining (Saitou and Nei 1978) uses a clustering algorithm to find phylogeny tree. Previously described distance matrix methods need to search tree space to find the tree that best meets the criterion, however, Neighbor Joining does not need to search the tree space– hence, it has a very big speed advantage. The main idea of Neighbor Joining is to join the closest nodes sequentially. Suppose D_{ij} is the distance between nodes i and j. The steps of the algorithm are:

1. calculate for each node i, $u_i = \frac{1}{n-2} \sum_{j:j \neq i}^{n} D_{ij}$
2. find the pair of nodes i and j that have smallest value of $D_{ij} - u_i - u_j$
3. join node i and j through a new node (ij). Calculate the branch length from node i to the new node $v_i = \frac{1}{2}(u_i - u_j + D_{ij})$ and the branch length from node j to the new node $v_j = \frac{1}{2}(u_j - u_i + D_{ij})$
4. calculate the branch length from the new node (ij) to any other node k as $D_{(ij)k} = \frac{1}{2}(D_{ik} + D_{jk} - D_{ij})$
5. remove items related to nodes i and j and replace them by the items of node (ij)
6. repeat the first step until only one node is left

Distance matrix methods only use low order information by pairwise comparison, losing the high order information which can be achieved by multiple comparison. There are some limitations for the usage of these methods.

Maximum likelihood methods (Felsenstein 1981) are more powerful than distance matrix methods. Given an evolutionary model, for any topology tree with specified branch lengths, we can calculate their likelihood function. This method looks for the tree topology, branch lengths and some other parameters which maximize the likelihood function.

Evolution rates can vary among sites in sequences (nucleotides or amino acids sequences). A Gamma distribution (Yang 1993) of the rates is used to calculate the branch lengths and likelihood function. Due to calculation difficulty, discrete Gamma distribution (Yang 1994) is used to approximate the distribution of real rates. This approximation improves the speed of calculation without loss of much information. Besides the Gamma distribution, the Beta distribution and the invariant model (some sites are invariant and the other sites evolve at the same rate) were proposed.

The Bayesian method also uses a likelihood function. It assumes some prior distributions of related parameters, and calculates the posterior distribution of these parameters using the information in the data. Recent Bayesian methods have adopted Markov Chain Monte Carlo algorithm, which magnifies the usage of Bayesian method greatly (Huelsenbeck et al. 2001).

1.3 The Concept of Gene Order

When we consider gene order, we are referring to the relative gene positions and directions on genomes. To describe gene positions, we need a coordinate system. DNA has two strands, and the sequence information can exist on both strands. We need to define the direction in which we read the genomes. We use the transcription direction of majority of genes as the direction of the genome. For linear chromosomes, it is natural to read from one end. For circular genomes, every gene can be the starting

point. After setting genome directions and a starting point, we can describe gene positions. As genes can be transcribed in two opposite directions, we define the direction of a gene as *positive* if the gene's transcription direction is the same as the genome direction; and *negative* otherwise.

We can represent gene orders by scale graphs. In scale graphs, each gene is represented by a block, whose size is determined by its gene length in the genome. In scale graphs, both strands of DNA are drawn. To describe the direction of a gene, we just need to draw that gene on its corresponding strand on the scale graph. For linear genomes their scale graphs are linear. For circular genomes, their scale graphs can be linear or circular. Figure 1.3 shows one example of human mitochondrial genome from NCBI website.

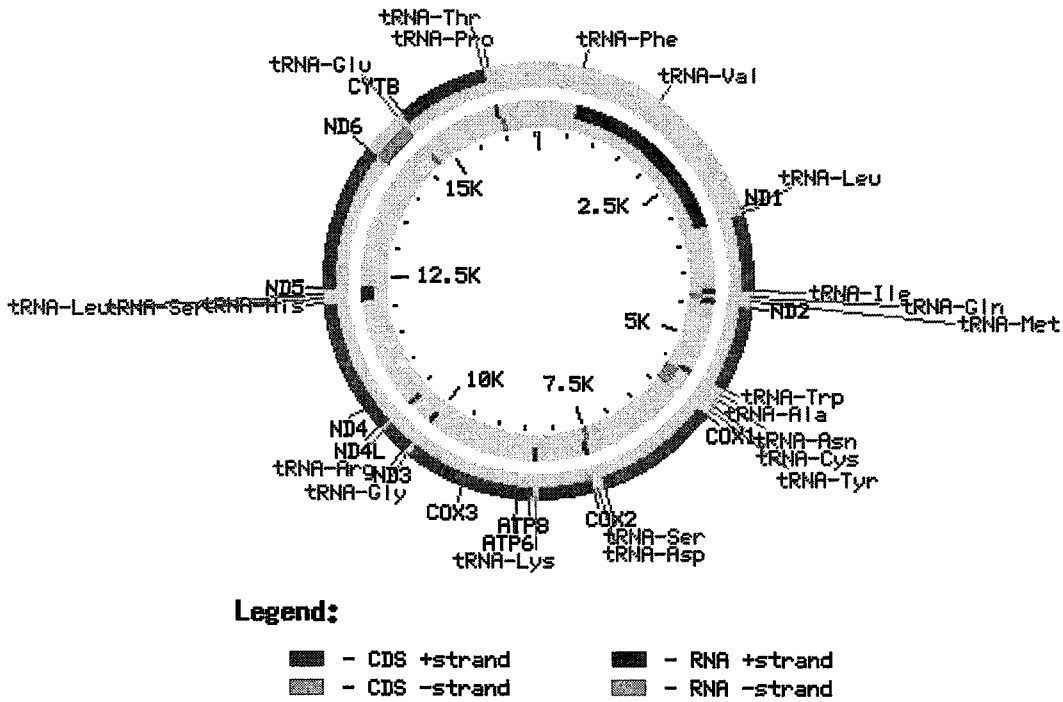

Figure 1.3: The scale graph of human mitochondrial genome

A gene order can be represented by a string of gene names (or abbreviations) separated by commas. A "-" sign can be added for gene on the reverse strand. We also can represent gene orders with numbers instead of names. Using numbers, we can describe gene orders' mathematical nature better. Mathematically, a gene order of N genes can be thought as a signed permutation of the numbers 1 to N.

For some genomes, we might not know genes directions, because of sequencing limitations. Then, these gene orders are called unsigned (or unsigned permutations in the number style). We call gene orders with directions as signed.

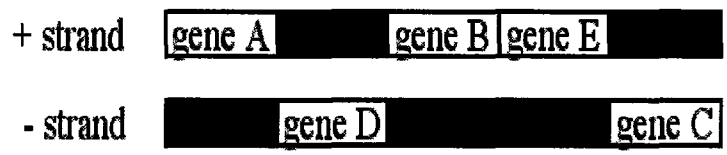

Figure 1.4: An example of gene order

In Figure 1.4, one artificial example of gene order is given. Suppose, the linear genome is composed by 5 genes on 2 strands, beginning from left to right. Gene A is on + strand, and we denote it as "A". Gene D is on - strand, then we denote is as "-D". Let's use "," to separate genes. So the gene order for the artificial genome is "A,-D,B,E,-C".

1.4 Animal Mitochondrial Genomes

The living world may be divided into three domains, Bacteria, Archaea and Eukaryotes. Eukaryote species have a nucleus, distinguishing them from the other two domains. Another feature for Eukaryote is that almost all Eukaryote species have mitochondria (with the exception of a few single celled organisms reported by Embley

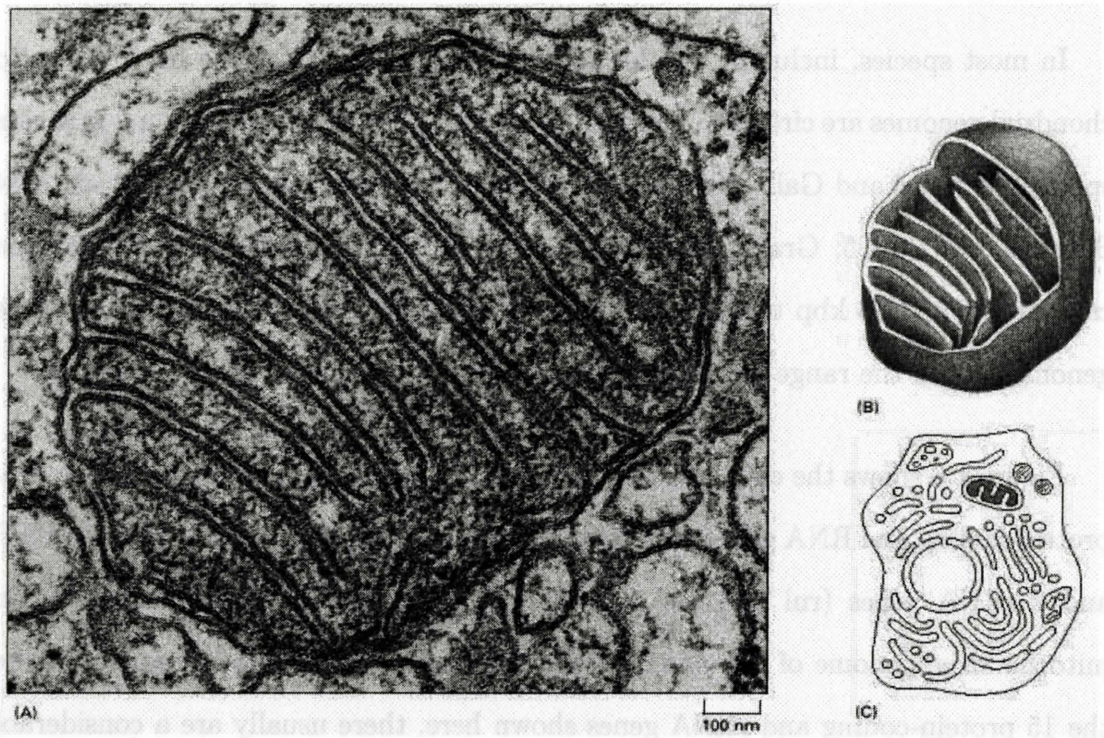

Figure 1.5: An example of mitochondria. (A) A cross section, as seen in the electron microscope. (B) A drawing of a mitochondrion with part of it cut away to show the three-dimensional structure. (C) A schematic eukaryote cell, with the interior space of a mitochondrion, containing the mitochondrial DNA and ribosomes, colored. Note the smooth outer membrane and the convoluted inner membrane, which houses the proteins that generate ATP from the oxidation of food molecules (Alberts et al. 2002).

et al. (1997)). Mitochondria are organelles with two membranes, usually rod-shaped. They are regarded as the powerhouses of the (eukaryote) cells. Sugars, amino acids and fatty acids are oxidized in the mitochondrion, and energy is stored in high energy phosphate bonds in adenosine triphosphate (ATP). Other reactions in the cells can use ATP as energy source directly.

In most species, including the Metazoan species we will analyze here, the mitochondrial genomes are circular. However, some linear mtDNA are also found in several species (Warrior and Gall 1985; Wesolowski and Fukuhara Biol; Kovac et al. 1984; Suyama et al. 1985; Grant and Chiang 1980). The size of mitochondrial genomes can vary from 14.3 kbp to over 2400 kbp (Gray 1989) The metazoan mitochondrial genomes are in the range of 14kbp-20kbp.

Figure 1.6 shows the comparison among mitochondria from different species, with protein-coding and RNA genes only. Only 3 protein genes (COX1, COX3 and CYTB) and 2 rRNA genes (rnl and rns) are shared by all mitochondrial genomes. The mitochondrial genome of humans is a typical animal mitochondrial genome. Besides the 15 protein-coding and rRNA genes shown here, there usually are a considerable number of tRNA genes. In the metazoa, the typical number of tRNAs is 22, which is sufficient to translate the complete genetic code. Some mitochondrial genomes may lack ATP8 or some tRNA genes. Some other mitochondrial genomes may have 2 or 3 copies of certain genes. Metazoa mitochondrial genome always begins with COX1 gene.

Mitochondria are usually inherited through the maternal line. (the exception has been found. *Lampsilis ornata* has both male and female mitotypes. Serb and Lydeard (2003).) The genes are linked and should yield the same evolutionary tree upon analysis without problems arising from recombination or horizontal gene transfer.

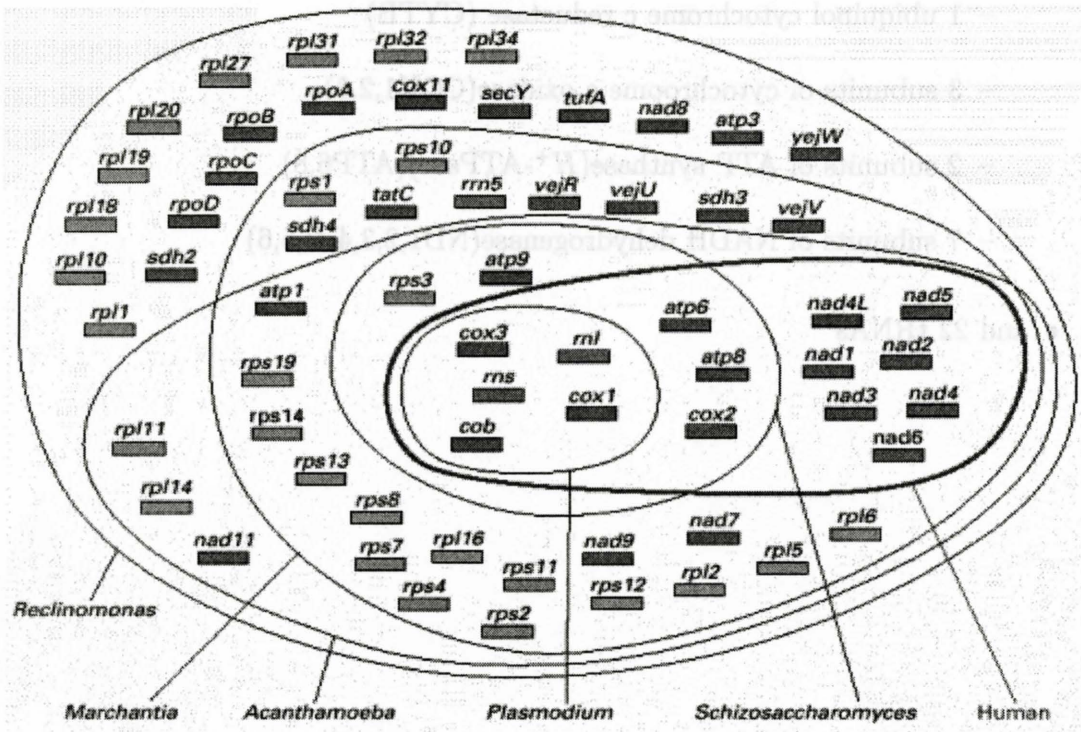

Figure 1.6: Comparison of mitochondrial genomes. Less complex mitochondrial genomes encode subsets of the proteins and ribosomal RNAs that are encoded by larger mitochondrial genomes. The five genes present in all known mitochondrial genomes encode ribosomal RNAs (RNS and RNL), cytochrome b (CYTB), and two cytochrome oxidase subunits (COX1 and COX3) (Alberts et al. 2002).

In summary, a typical animal mitochondrial genome contains:

- 2 rRNA

  large subunit(RNL) and small subunit(RNS)

- 13 proteins

  - 1 ubiquinol cytochrome c reductase (CYTB)
  - 3 subunits of cytochroome c oxidase(COX1,2,3)
  - 2 subunits of ATP synthase(H^+-ATPase)(ATP6,8)
  - 7 subunits of NADH dehydrogenase(ND1,2,3,4,4l,5,6)

- and 22 tRNAs

Chapter 2

The OGRe Database

2.1 The OGRe Database

OGRe (Organellar Genome Retrieval, http://www.ogre.mcmaster.ca) is an object relational database of complete mitochondrial genome information for over 600 Metazoan species. An object relational database has a predicate logic and set theory based model and allows developers to integrate the database with their own custom data types and methods (McClure 1997). The OGRe database implements Postgre SQL, which is a flexible, powerfully open source object-relational SQL database management system. The OGRe gets data from NCBI (http://www.ncbi.nlm.nih.gov) and provides a resource for the comparative analysis of mitochondrial genomes at several levels. You can select organelle genomes from any set of species and display or download certain sequences. You also can view the their base frequencies and codon usage frequencies. OGRe provides several genome visualization tools. Genome Viewer is the tool to display gene orders. Figure 2.1 is one example of Genome Viewer. The legend tells us the color schedule for labeling genes: green for protein, red for tRNA and blue for rRNA. There are 3 species displayed in Figure 2.1. Although genomes are circular, they are represented by linear scale graphs from left to right. The linear graphs can start from any gene in the genome. Without losing generality, the COX1 gene is always chosen as the first. In the graph, each color block represents a gene. The blocks above the central line indicate that the genes are translated from the nega-

tive strand. The ones below the central line indicate they are translated from positive strand. Spaces between blocks indicate non-transcribed regions. There actually are few non-transcribed regions in animal mitochondria. Below the graph, corresponding gene names are presented. The last line is the information line. The group name (e.g., classes, orders, families, please see details in OGRe database) of that species, Latin name of that species, number of genes, proteins, tRNAs and rRNAs and the length of the genome are displayed from left to right.

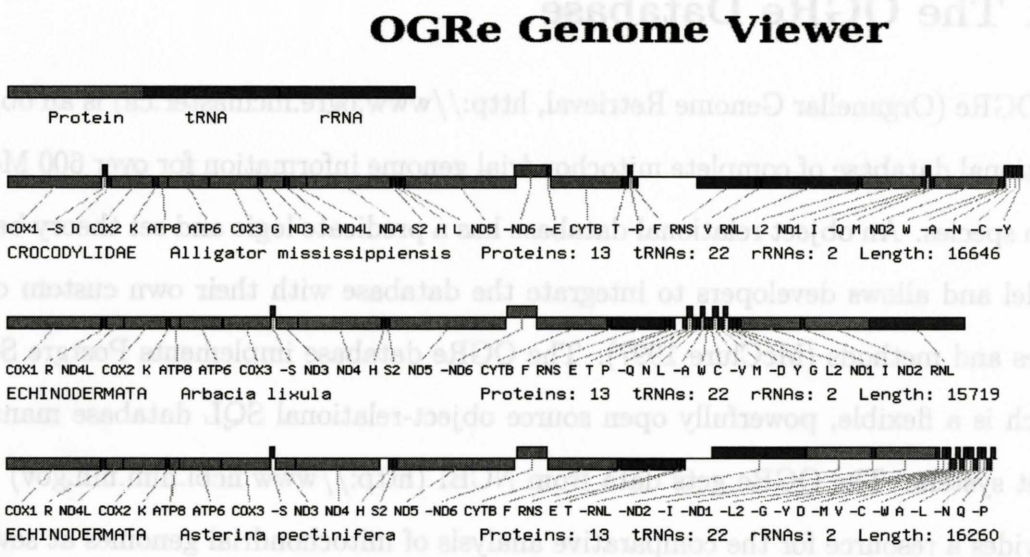

Figure 2.1: OGRe Genome Viewer

Another tool is Genome Comparison. Genome Comparison is very similar to Genome Viewer. It provides an easy and convenient method to compare two genomes by labeling conserved gene clusters in one color. Figure 2.2 is one example of Genome Comparison output. The left end gray gene cluster and the right end gene cluster are actually consecutive, because it's a circular genome.

Daniel Jameson was the designer of original version of OGRe. In that version there were approximated 250 complete mitochondrial genomes. Then, there was an

update in 2004 by Bin Tang, and the size of database became 473 mitochondrial genomes. This is the current public version. Now we are making another update, there will be more than 600 complete mitochondrial genomes after this update by Wenli Jia.

In 2004, I also worked on gene order distance matrix part for OGRe, where one can get a distance matrix for all gene orders contained in database for either breakpoint distance or inversion distance.

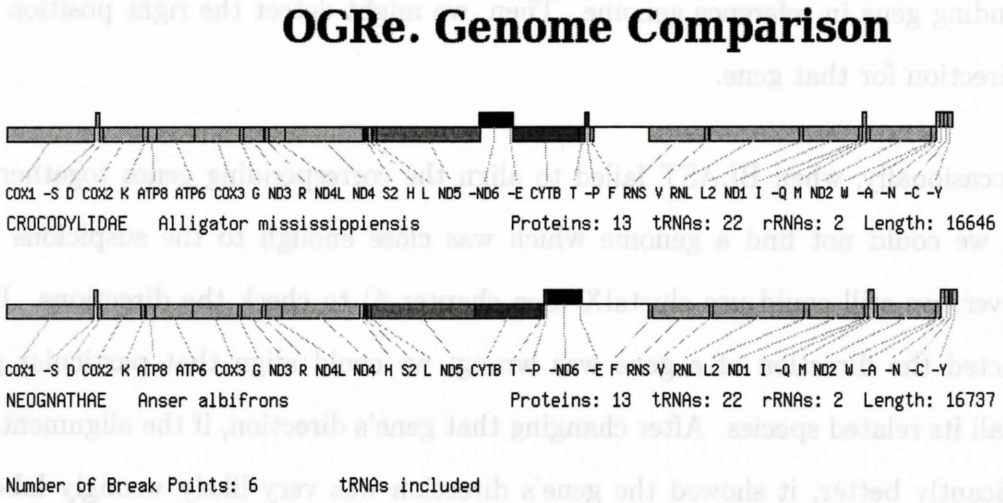

Figure 2.2: OGRe Genome Comparison

2.2 Correcting Information in the OGRe Database

We can observe errors in the original NCBI data. Several types of errors have been found. Some gene starting or ending positions are wrong. Some gene translation

directions are wrongly labeled. Some tRNA genes are unlabeled while their sequences do exist on the genomes.

To check starting and ending positions and translation directions, we used BLAST(Altschul 1991; Altschul et al. 1994). BLAST (standing for Basic Local Alignment Search Tool) is quick and powerful alignment tool. When we suspected one gene was wrongly labeled either for its position or direction, we tried to find another genome (reference genome) which is close phylogenetically to the suspicious genome. Then we ran BLAST for those two genomes. The suspicious gene would be aligned with the corresponding gene in reference genome. Then, we might detect the right position and the direction for that gene.

Occasionally, when BLAST failed to align the corresponding genes together because we could not find a genome which was close enough to the suspicious one. However, we still could use clustalX (See chapter 4) to check the directions. If we suspected the direction of a gene was wrong, we could align that particular gene from all its related species. After changing that gene's direction, if the alignment was significantly better, it showed the gene's direction was very likely wrongly labeled. We have spotted out approximated 100 errors of these two kinds. You can find the detailed information of these corrections on OGRe's website.

To find out the unlabeled genes, there were two applicable methods. The first involved using BLAST. The process was similar to the previous one. The second method involved using tRNASCAN (Lowe and Eddy 1997). After running tRNASCAN for one suspicious genome, it would give a list of possible sequences that might be tRNAs. We first checked the positions of those sequences, if there was one sequence which did not overlap any existing genes, then we checked whether we could find the right anti-codon and the reasonable secondary structure. If all conditions held true,

then we determined that sequence was the missing gene. We have found 10 missing tRNAs; this information is listed in Table 2.1.

gene name	species	method
tRNA-Glu	ACISTEMIT	blast
tRNA-Ser(AGY)	ALEDUGMIT	blast
tRNA-Asp	BOSINDMIT	blast
tRNA-Val	CAEFULMIT	tRNASCAN
tRNA-Met	CAICROMIT	blast
tRNA-Ser(AGY)	MELBICMIT	blast
tRNA-Met	MUNCRIMIT	blast
tRNA-Val	RHYRAPMIT	tRNASCAN
tRNA-Gly	TAPTERMIT	blast
tRNA-Ile	UROTALMIT	tRNASCAN

Table 2.1: The list of 10 unlabeled tRNAs . The first column is the genome code, the second column is the name of that tRNA and the last column is the method by which the missing tRNA was found

In general, finding and annotating genes on large genomes is a difficult task in bioinformatics. However, for animal mitochondrial genomes, the task is easier as we have so many known genomes with which to compare a new sequence, and we have a very good idea which genes we expect to find on a new genome.

Chapter 3

Gene Order and Its Application in Phylogeny Inference

3.1 Introduction to Gene Order Analysis

In this chapter, the interest is not on single gene order itself, but on comparisons of two or more gene orders. Let's ignore genes sizes and consider gene orders just as assigned permutations. When we compare two permutations, if there is a string of genes such that each gene is in the same relative position and direction in both permutations, then we call this string of genes a conserved segment. If one conserved segment has 6 genes in it, any consecutive 5 or few genes also is a conserved segment. Normally, a conserved segment refers to the one with largest number of genes.

For example, in Figure 3.1 the gene orders of two Echinodermata species are compared. This figure was generated by Genome Comparison in OGRe. There are two conserved segments, a left one (in light Gray) and a right one (in dark). The genes in the right conserved segment have the same order and direction relative to the the other segment, which was inverted as a whole.

Combined with the concept of conserved segment is the concept of breakpoint. When two genes are neighbors on the genome, we call this property they have, *neighborhood.* For signed gene orders, when we say neighborhood of two genes, we should consider their directions. Gene Orders like "A,B" and "A,-B" do not have the same

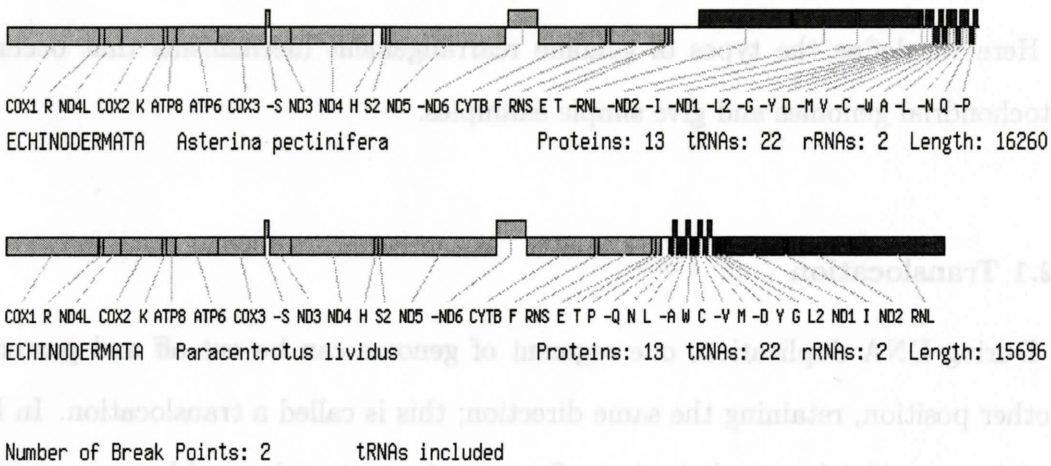

Figure 3.1: An example of real inversion

kind of neighborhood. When the neighborhood of two genes exists in genome I and not in genome II, we say there is one breakpoint in genome II. The relationship of breakpoint number (bp) and conserved segment number (cs) is $bp = cs$ for circular genomes and $bp = cs - 1$ for linear genomes. If the two genomes contain different sets of genes, the two genomes will have different breakpoint numbers. We choose the larger breakpoint number as the breakpoint distance between these two genomes. In Figure 3.1, the breakpoint distance is 2. The concept of breakpoint distance is independent of the mechanism that causes the genome rearrangement. It's very easy to calculate the breakpoint number. Now, when not much information of genome rearrangement mechanisms is known, breakpoint distance is a good and relative robust measure for the distance of genomes.

3.2 Genome Rearrangement Mechanisms

Here we define the types of genome rearrangement mechanisms that occur in mitochondrial genomes and give simple examples.

3.2.1 Translocation

During DNA duplication, one segment of genome can be cut off and pasted to another position, retaining the same direction; this is called a translocation. In Figure 3.2, an artificial example is given. Segment A was translocated between segment B and segment C. 3 or 2 breakpoints are induced for circular or linear genome correspondingly. In Figure 3.3 shows one real example. Between human and chicken gene orders, the difference is caused by a single translocation.

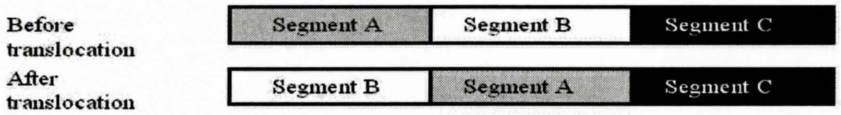

Figure 3.2: An example of artificial translocation

3.2.2 Inversion

A string of genes can also be cut off and pasted to the same position but on the opposite strand, and this is called a inversion. As demonstrated by Figure 3.4, segment B changes its direction but remains in the same location. Figure 3.1 is a real example of inversion. The right dark conserved segment was inverted.

Translocations can only change the locations of genes, but can not change their directions. Inversions can change both the locations and directions. What's more,

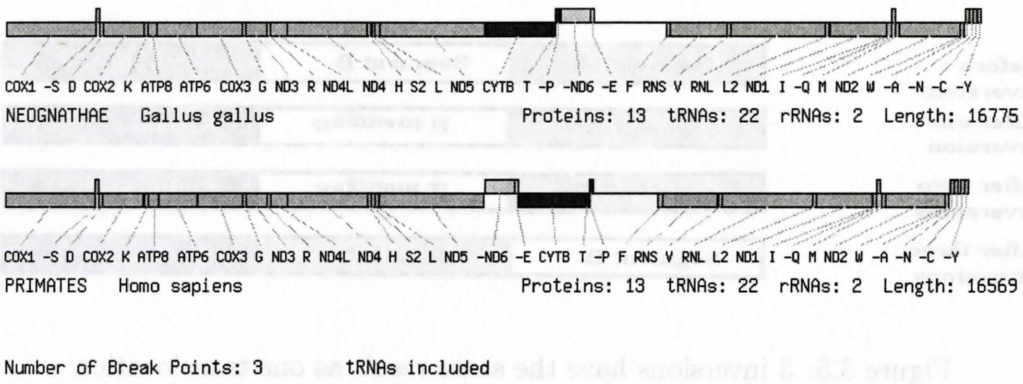

Figure 3.3: An example of real translocation. Between human and chicken gene orders, the difference is caused by a single translocation

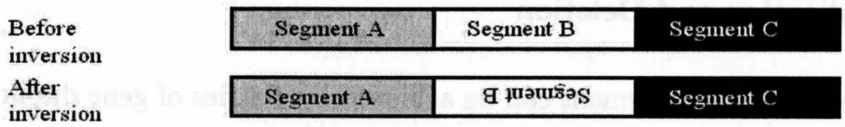

Figure 3.4: An example of artificial inversion

the effect of any kind of translocation can be achieved by a series of inversions. The most simple example is that 3 certain consecutive inversions will have the same effect as one translocation, as illustrated by Figure 3.5.

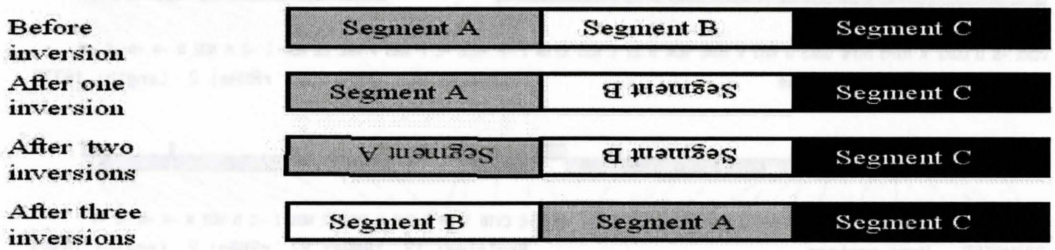

Figure 3.5: 3 inversions have the same result as one translocation

3.2.3 Transversion

A transversion is when, one segment is cut off and pasted to another location and inverted at the same time. It also can be achieved by a translocation followed by an inversion.

3.2.4 Duplication and Deletion

Also, genome rearrangement can be achieved by a series of gene duplications and deletions. During evolution of the mitochondrial genome, we know that the genome was reduced considerably due to gene deletions. Many of these genes have been transferred to the nucleus. However, since the origin of the metazoa, the number of genes remains fairly stable. Duplication followed by deletion can lead to gene order changes. In Figure 3.7, gene A and gene B are duplicated and then the first copy

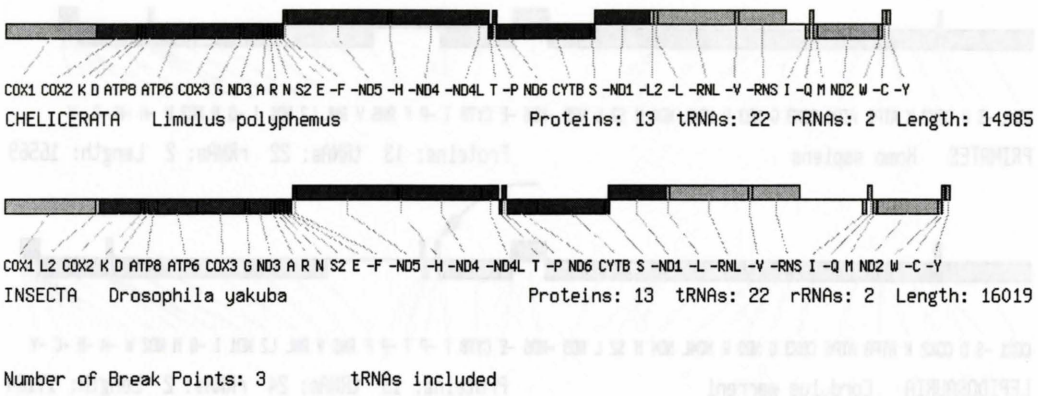

Figure 3.6: An example of real transversion. Between *Limpulus polyphemus* and *Drosophila yakuba* gene orders, the only differece is cause by a single transversion of L2 gene.

of gene A and the second copy of gene B are deleted, thus the gene order has been changed just as a translocation.

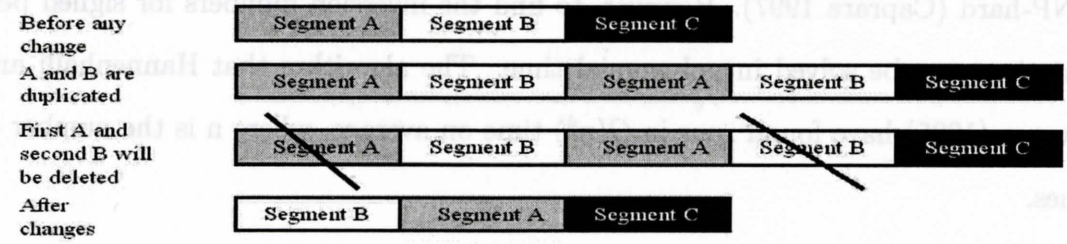

Figure 3.7: An example of duplication and deletion also can change the order of genes

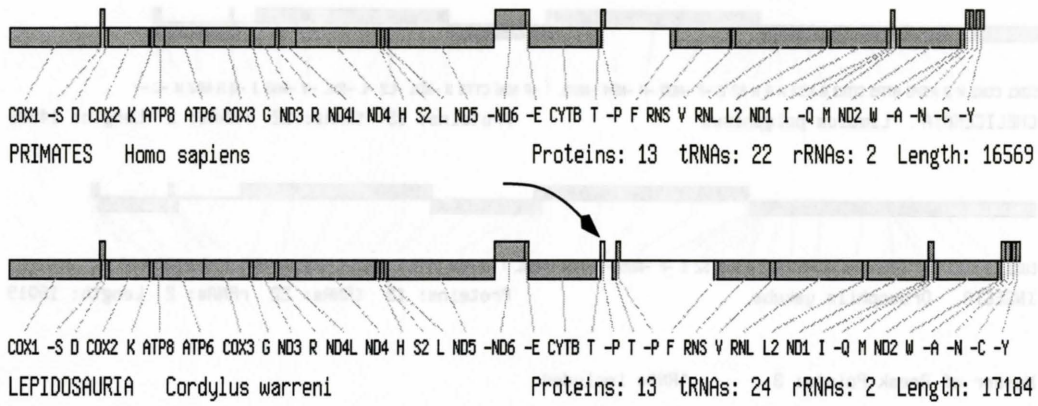

Figure 3.8: An example of real duplication. Compared to human mitochondrial gene order, there are two duplicated genes in *Condylus warreni* mitochondrial gene order. If the first copy of T tRNA and the second copy of P tRNA are deleted in future, the gene order can be changed.

3.2.5 Mathematical Background and Some Software

3.2.5.1 Calculation the Inversion Distance

Calculation of inversion distance has been studied for a long time, and many useful results have been reported. To find the inversion numbers for unsigned permutations is NP-hard (Caprara 1997). However, to find the inversion numbers for signed permutations can be solved in polynomial time. The algorithm that Hannenhalli and Pevzner (1995) have found runs in $O(n^4)$ time on average, where n is the number of genes.

3.2.5.2 Underestimation

The methods mentioned above to find either the minimum number of inversions between two gene orders or the breakpoint number, which also means the shortest

distance in mathematics. The real rearrangement is a stochastic process. It is very possible that the real number of rearrangements is larger than the number we infer. This underestimation will reduce the real distance between genomes and hence, influences the phylogeny which is inferred from those distances.

3.2.5.3 The Correction for Breakpoint Number

In distance matrix methods, the additive property is very useful in generating an accurate phylogenetic tree. The additive property stipulates that the distance measures with the additive property should increase linearly with the number of changes that have happened along a branch. However, the breakpoint measure is not additive. The sum of breakpoint numbers from genome A to B and B to C is not always equal to the breakpoint number from genome A to C. We define an additive distance measure, the evolutionary distance, as the number of real rearrangements happened along that branch. By its definition, this distance is additive.

Suppose there is one rearrangement, after which the whole genome is broken into m segments and m new breakpoints are generated. As we know, one single inversion can produce 2 breakpoints and one single translocation or transversion can produce 3 breakpoints. For a circular genome with n genes, there exist n neighborhoods, and after a series of rearrangements, there will exist at most n breakpoints. For each rearrangement, the genome will be broken into m segment. The chance to remove a breakpoint during one rearrangement is very small (less than $\frac{d}{n^2}$, where d is the edit number); and we assume here, that once a breakpoint is created, it will never be removed. Suppose, that before each rearrangement, there are b breakpoints compared to the original genome. So, on average, $b \times \frac{m}{n}$ new breakpoints will produced and $(n-b) \times \frac{(n-m)}{n}$ original neighborhoods remain (compared to the original genome). So

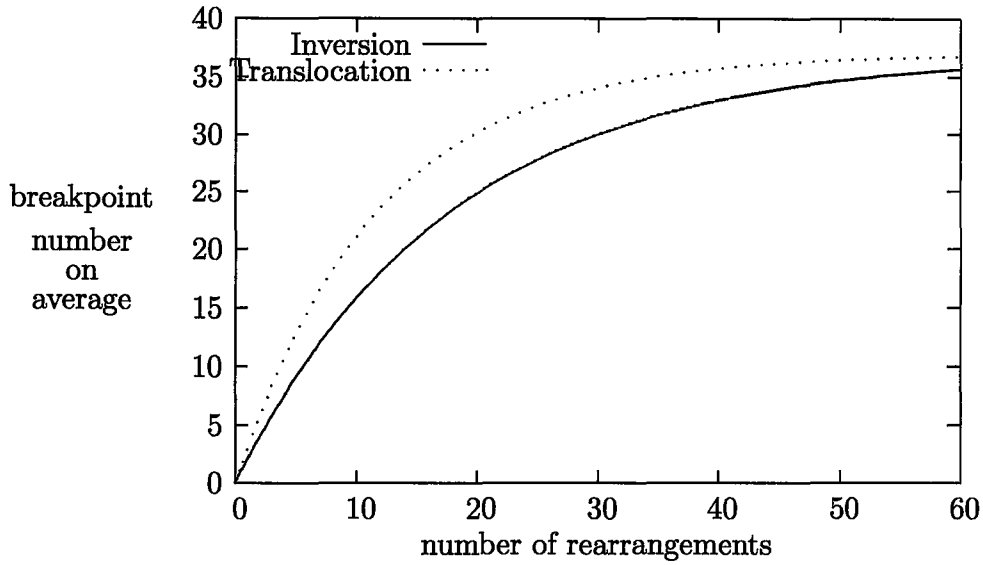

Figure 3.9: This figure shows the relationship of average breakpoint number with translocation number or inversion number. The curves level off when tranlocation or inversion number increases. It shows the possibility of underestimation.

when we start from the original genome, where $b = 0$, the number of breakpoints on average after d of rearrangements is:

$$b = n(1 - (\frac{n-m}{n})^d) \tag{3.1}$$

Figure 3.9 shows examples when m=2 for inversion and m=3 for translocation or transversion. When the number of rearrangements increases, both curves level off where the underestimation happens.

What we can observe directly is breakpoint number. From the observed breakpoint number, we can estimate the real evolutionary distance $\bar{d}$

$$\bar{d} = \log_{1-\frac{m}{n}}(1 - \frac{b}{n}) \tag{3.2}$$

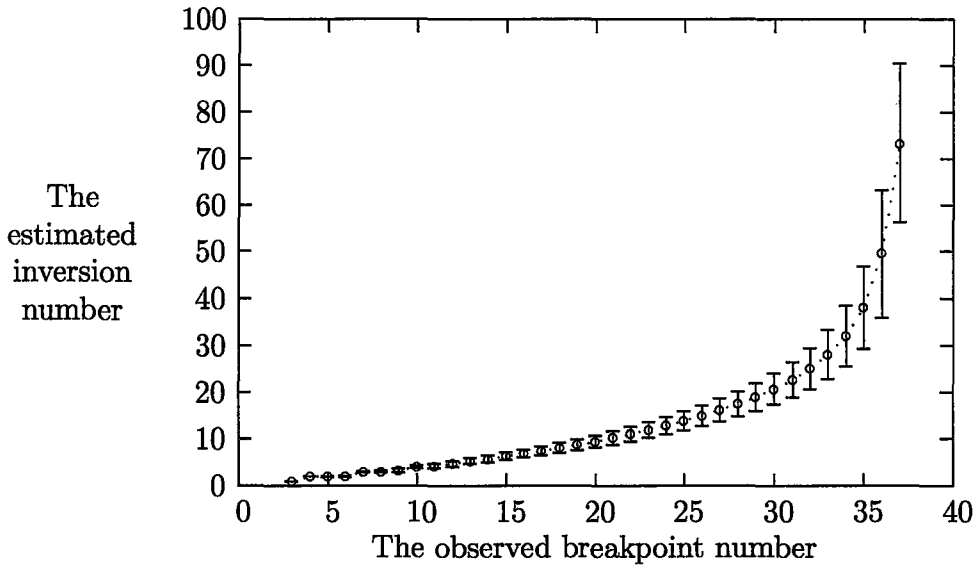

Figure 3.10: This figure shows the relationship between observed breakpoint number and the estimated inversion number. When breakpoint number increases, the uncertainty of estimation increases.

Figure 5.2 shows one example, where the rearrangement are inversions. When breakpoint number increases, the standard deviation becomes larger, and there exists a region where the true rearrangement number is difficult to infer because of the large uncertainty.

However, the purpose here is not to find the relationship of breakpoint number and inversion number but an additive measure for gene orders. We don't know what are the real rearrangements, and the value of m remains unknown. The form of the previous formula may be changed to

$$\bar{d} = \ln(1 - \frac{b}{n}) / \ln(1 - \frac{m}{n}) \tag{3.3}$$

Here, we define our new measure as $\bar{d}_{ln} = ln(1 - \frac{b}{n})$, when n is large, the term $1 - \frac{m}{n}$ will have less influence. I applied this logarithm correction for breakpoint distance to all distinct gene orders in OGRe.

3.2.5.4 Dereange2–The Program to Infer the Path to Change a Gene Order

Derange2 (Blanchette et al. 1996) is a heuristical program designed to find the pathway with the minimum weight between two gene orders with identical gene sets. It allows users to assign weights to different mechanisms (inversion, translocation and transversion). Deciding what weights should be assigned is a difficulty. First, the weights of different mechanisms in reality are unknown and very difficult to estimate. Second, even if we know those real weights, this set of parameters may differ from the one used for those programs adopting the parsimony method.

3.3 Phylogenetic Analysis of Gene Order Information

The methods mentioned before compare only two genomes. When we compare three genomes at once, we want to know the median genome among those 3 genomes. To find the median genome under certain criteria we need to search all possible gene orders. This is a huge task, especially when the number of genes is large. To compare more than 3 genomes, we need to search all the possible topologies and all the median gene orders at same time. Another possible method involves calculating the pairwise gene order distances and applying distance matrix methods to them. In this section, we will introduce two programs which follow the first method and discuss distance matrix methods for gene orders in next section.

3.3.1 BPAnanlysis–The First Attempt

BPanalysis (Sankoff and Blanchette 1998) was the first attempt to infer the phylogeny by comparing multiple genomes directly. The program finds the tree such the total number of breakpoints summed over all branches of the tree is minimized. This is a parsimony criterion. The problem to find the median gene orders can be turned into a Traveling Salesman Problem (TSP) (Sankoff and Blanchette 1997). BPAnalysis solves the median problem gradually and repeatly until one stable state has been arrived at. Although, overall, the comparison of multiple genomes seems NP-hard (L Pei-er and R. Shamir reported for N = 3), it is tractable for moderate genome sizes.

Sankoff and Blanchette (1998) reported the underestimation phenomenon. The underestimation became manifest when the breakpoint number per branch reached half of number of genes; and there was 30% underestimation when the breakpoint number reached two-thirds of number of genes (Sankoff and Blanchette 1998). Under such big underestimation, there is little chance to get the correct phylogeny.

BPAnalysis has another problem, its speed is too slow. I have tried it for 16 genomes with 35 genes, on an x86 pc with a 3.0 GHz P4 CPU. It ran about 2 weeks.

3.3.2 GRAPPA

GRAPPA (Moret et al. 2001) stands for Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms. The program was written in c instead of c++, and many optimizations have been made for its speed and more options (different approximate algorithms for Traveling Salesman Problem and different methods to label the inner gene orders) have been implemented. The authors tried some heuristic algorithms to calculate inversion numbers and extend the parsimony criteria in BPAnalysis to the minimum inversion criterion.

In their papers, the authors claimed their program ran one-million times faster than BPAnalysis. We wanted to know how does GRAPPA perform in our situation. In the following text, we present our test for GRAPPA in a simulated situation which is similar to the case of animal mitochondrial genomes.

3.3.2.1 Test of GRAPPA

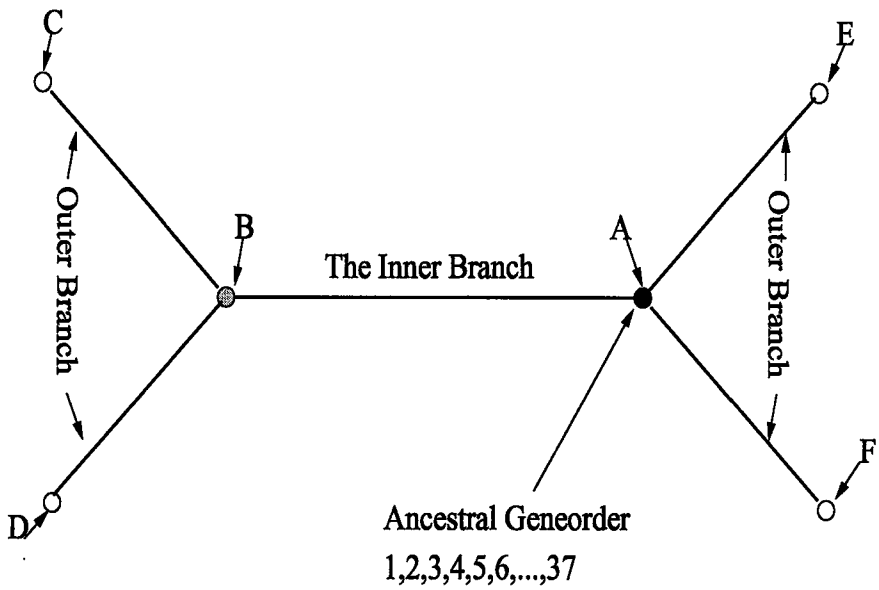

Figure 3.11: the topology for the simulated data. There are 4 present-day gene orders descending from other extinct 2. The right extinct gene orders serves as the ancestral one, with identity permutation $(1, 2, 3 \ldots, n)$

Figure 3.11 shows the topology of the simulated data. There are 4 leaf genomes (C,D,E,F) and 2 extinct genomes (A,B). Genome A, the ancestor, has an identity permutation i.e. $(1, 2, \ldots, n)$. Genome B descends from A. Genomes C, D and E, F descend from B, A correspondingly. The number of genes each genome contains is 37. The rearrangement mechanism we used was inversions. The inner branch lengths vary from 1 to 11 and 13, 15 inversions and the outer branch lengths vary from 1 to 10 and

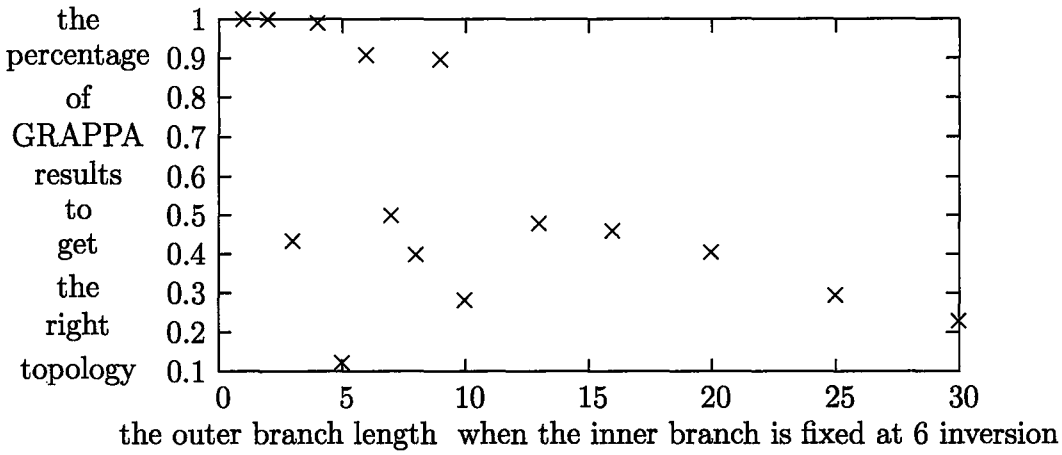

Figure 3.12: The chance for GRAPPA to retain the correct topology when the inner branch length is fixed to 6, while the outer branch lengths vary from 1 to 30 inversions.

13, 16, 20, 25 and 30 inversions. For each each pair of inner and outer branch lengths, we generated 100 random datasets. Then we used GRAPPA to estimate the phylogeny using the minimum inversion distance criterion. The parameters for GRAPPA were "-T4 -t4 -K2 -L" (for details please see their document files). We collected the results and counted how many groups of results retained the correct topology (we use T-rate to represent this percentage) and how many groups of results retained the correct ancestral gene order (G-rate).

Figure 3.12 shows the T-rate when the inner branches are fixed at 6 inversions and the outer branch length varies. T-rate drops quickly when the outer branch length increases to 5 inversions. For four species, there are only 3 unrooted topologies. If we choose the topology tree randomly, the average chance to get the correct one is 33%. Percentages lower than 30% were observed in our test. The test suggested that in those low percentage area, GRAPPA does not work at all. Figure 3.13 shows the G-rate when the inner branches are fixed at 6 inversions and the outer branch length

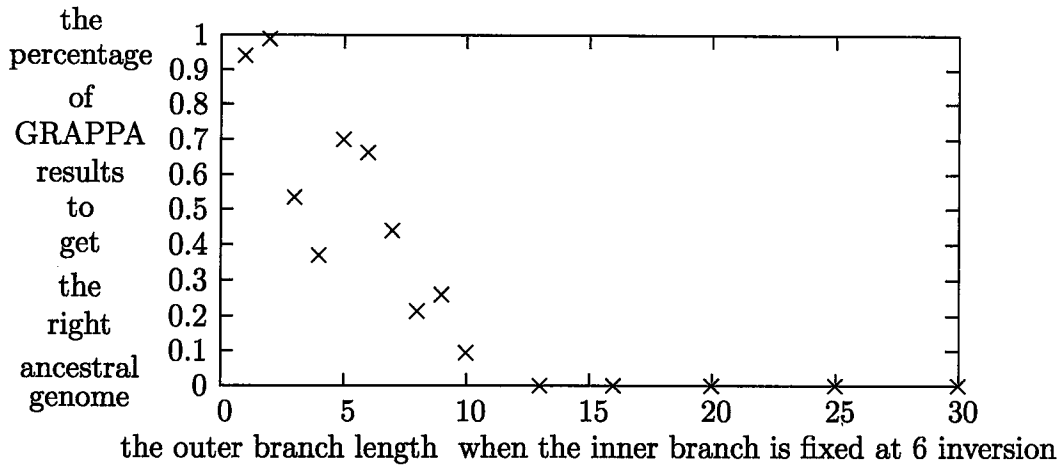

Figure 3.13: The chance for GRAPPA to retain the correct ancestral gene order when the inner branch length is fixed to 6, while the outer branch lengths vary from 1 to 30 inversions.

varies. The percentage also drops quickly when the outer branch lengths increase to 5 inversions. When outer branch length increases to 10 inversions, there is almost no chance for GRAPPA to get the correct ancestral gene order.

Figures 3.14 and 3.15 show the range where the chance for GRAPPA to find the correct topology or right ancestral gene order is larger 50% or 80%. Figure 3.14 suggests that inner branch has less influence on G-rate than the outer branch, except when inner branch length is between 8 and 10 inversions, whereupon percentage has a suddenly drops. Figure 3.15 suggests that the sum of inner branch length and 2 times the outer branch length has a large influenced on G-rate. The critical value for this sum is approximated 16 inversions, below which GRAPPA has over 50% chance to retain the correct ancestral gene order.

The test results show that GRAPPA has a large chance to find the correct topology when the inner branch is within 15 inversions and the outer branch is within 6

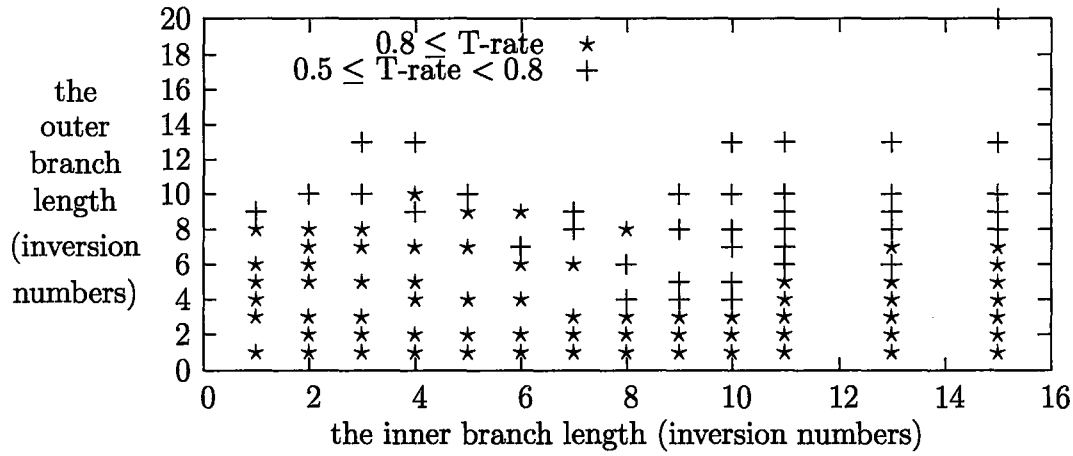

Figure 3.14: Ranges of inner branch lengths and outer branch lengths where the chance for GRAPPA to retain the correct topology (T-rate) is larger than 0.5 and 0.8. The data below 0.5 are not shown. There is no testing data when inner branch length is 12, 14 and 16 inversions or outer branch length is 11, 12 and 14 inversions

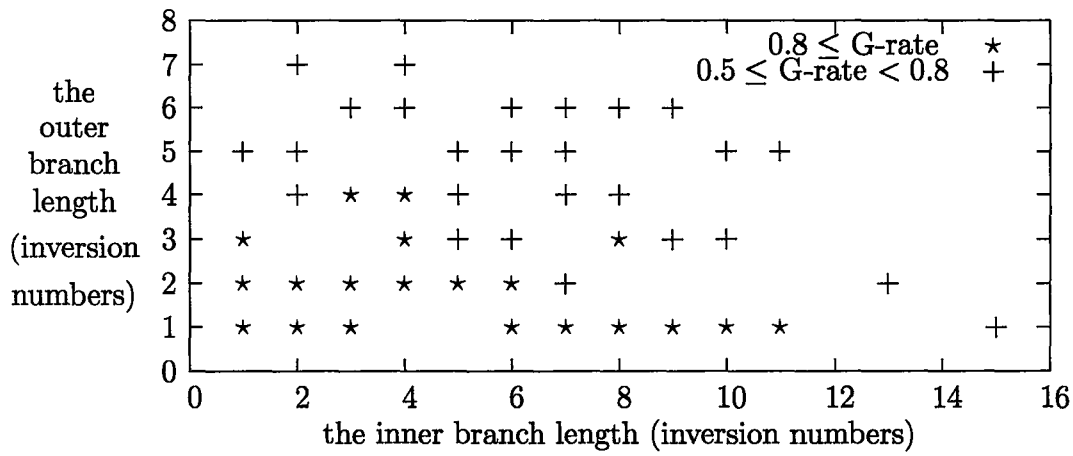

Figure 3.15: Ranges of inner branch lengths and outer branch lengths where the chance for GRAPPA to retain the correct ancestral gene orders (G-rate) is larger than 0.5 and 0.8. The data below 0.5 are not shown. There is no testing data when inner branch length is 12, 14 and 16 inversions or outer branch length is 11, 12, 14 and up to 19 inversions

inversions in a situation of 4 species and has a large chance to find the correct ancestral gene order in a very small range i.e., the outer branch is within 2 inversions and the inner branch length is within 6 inversions. When the topology is complex (more species), and translocations or transversions are involved in gene order rearrangement, the chance to get the right answer will be even smaller.

3.4 Using Distance Matrix Methods on Gene Orders

We wanted to apply distance matrix methods to all distinct animal mitochondrial gene orders in OGRe. There are 98 distinct gene orders in OGRe with 473 species. One can get the list of these distinct gene orders from OGRe website.

First, we applied neighbor joining method to the breakpoint distance matrix of all animal mitochondrial gene orders. The breakpoint distances were normalized by dividing the raw breakpoint number by the maximum number of genes contained in the two genomes being compared. By normalization, the deviation caused by the difference of genome sizes can be eliminated.

Figure 3.16 shows the result of applying distance matrix methods to normalized breakpoint distance. In this figure, the lower part is relatively well defined at the phyla level. At the upper part, the tree is scrambled, except for several groups (Platyhelminthes, Cnidarians) are well defined. All the species in upper part of the tree have a long branch, obviously this could is the case of long branch attraction. Even through some phyla form well-defined groups (e.g. vertebrates), the detailed phylogeny within these groups is not consistent with known relationships, derived from other sources. Chordates are grouped with Echinodermata; a group of Mollusca, Annelids and one Branchiopoda are grouped with Arthropods. Halanych et al.

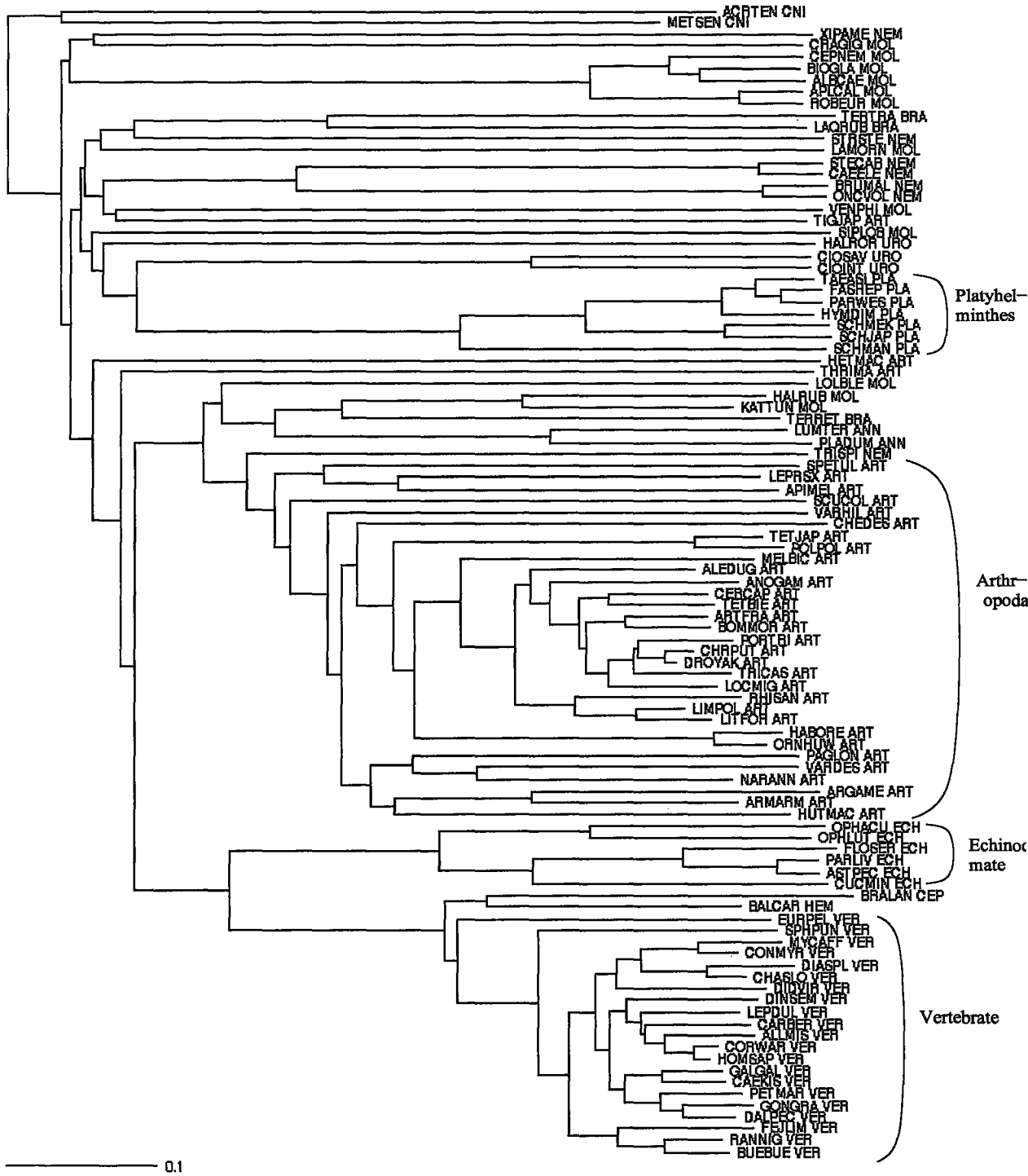

Figure 3.16: The phylogeny tree generated by NJ method with normalized breakpoint distance matrix

(1995) proposed a new phylogeny where the Bilateria is divided between Deuterostomes and Protostomes. In this phylogeny, Vertebrates, Echinoderms and etc. belong to Dueterostomes and Platyhelminthes, Brachiopods, Molluscs, Nematodes and Arthropods belong to Protostomes. In traditional classification, the Bilateria is divided into Coelomates (Vertebrates, Echinoderms, Brachiopods, Molluscs, Annelids and Arthropods), Pseudo-coelomates (e.g. Nematodes and) and Acoelomates (e.g. Platyhelminthes). The Coelomates and Pseudo-coelomates form sister groups and Acoelomates is their outgroup. In Figure 3.16, Platyhelminthes are outgroup of Vertebrates, Arthropods and othergroups. This is inconsistent with the new proposed phylogeny and it partly agrees with the traditional one.

Then, we used logarithm correction for breakpoint. Figure 3.17 shows the resulting tree. The shape of the tree seems better than that in Figure 3.16, and the groups can be recognized by their branch lengths. The species with long breakpoint distances now are in better positions than before. However, Echinoderms are the outgroup of Vertebrates now, which might be the result of large uncertainty of breakpoint distance correction. In upper part of the tree, Platyhelminthes and Nematodes are the outgroup of Vertebrates + Arthropoda + Annelids + Mollusca, which form Coelomates. Although Platyhelminthes and Nematode form one group, it might not be true.

None of the results is satisfying. Beside the limitations of distance matrix methods, there are two other reasons. Firstly, for such small genomes, the breakpoint distances are easily saturated(i.e. the breakpoint numbers are close the their maximum values). Secondly, gene orders evolve highly unequally in some groups. When the gene order distances are close to their maximum values, the underestimation is very large for breakpoint number and the uncertainty becomes very large for the corrected logarithm form of breakpoint distance. In this case, any distance matrix method will fail. The

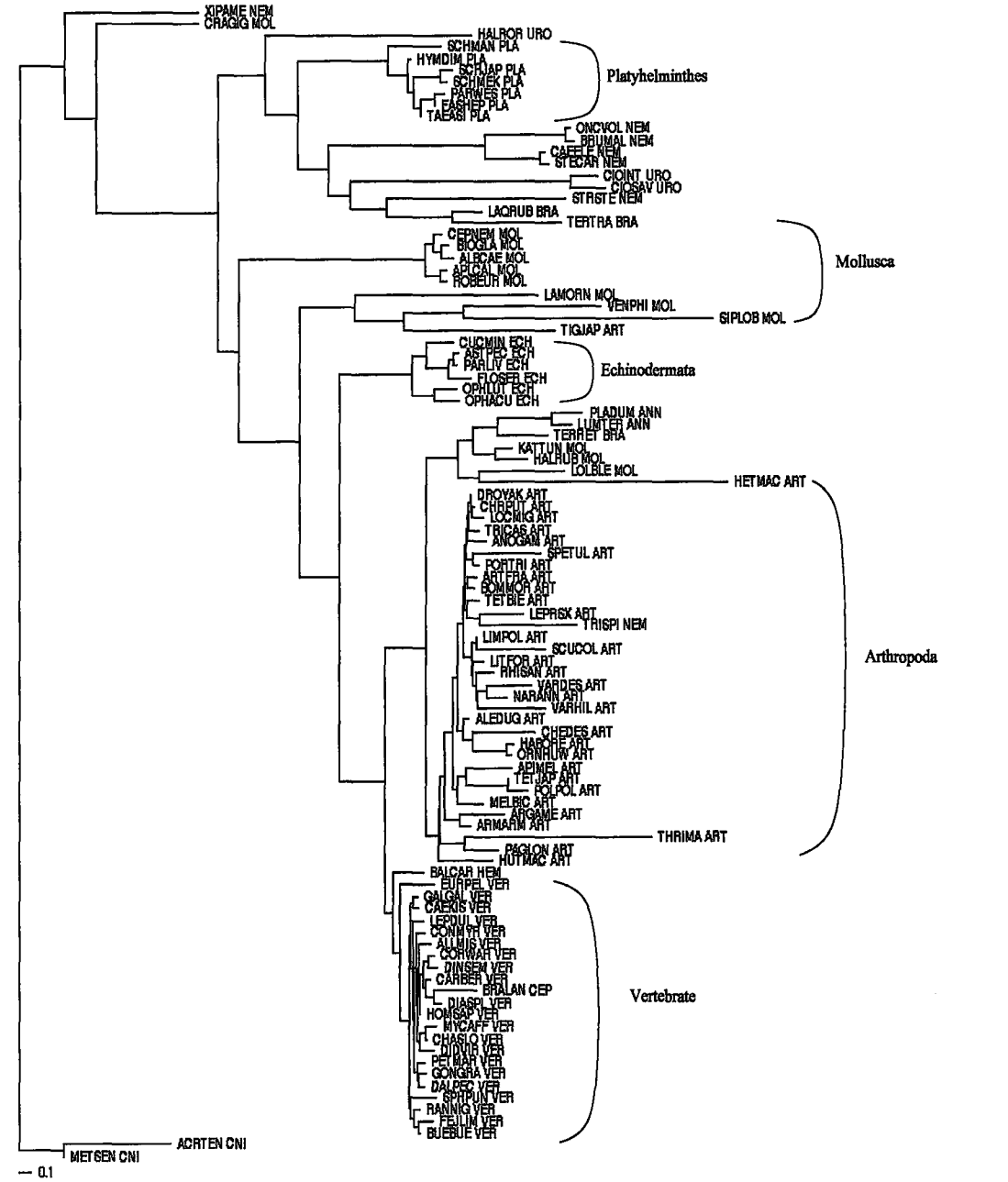

Figure 3.17: The phylogeny tree generated by NJ method with logarithm form of breakpoint distance matrix

fast evolved species have long branches, then due to long branch attraction, the distance matrix method sometimes failed to find the correct topology. For the fast with a saturated distance, they will lose traces of their ancestral gene orders; then, it is impossible to find the their correct phylogeny using gene order information.

Chapter 4

Phylogenetic Analysis of Arthropoda Using Mitochondrial Sequences

4.1 Arthropod Phylogeny

Although some aspects of arthropod phylogeny are well understood, there are some key questions that are not resolved. We, therefore, wished to use the mitochondrial sequence data in OGRe to obtain a molecular phylogeny of arthropods. An additional motivation for looking at the arthropod phylogeny is that we wish to compare information from sequence evolution with information from gene order in the same set of species. The set of arthropod species in OGRe is interesting in this respect because it contains both very conserved gene orders, thought to be similar to those of the ancestral protostome, and highly derived gene orders, almost completely scrambled with respect to the ancestral order and to other existing species (see discussion of the trees derived from breakpoint distances in Chapter 3). The comparison of gene order and sequence information will be the subject of Chapter 5. Before we can do this, we require a best estimate of the phylogeny of the arthropod species in the OGRe data set. The main aim of this chapter is therefore to obtain this best estimate tree.

The species studied here are listed in Table 4.1. Each species belongs to one of four principal taxa labeled in bold font in column one. Important lower-level taxa that

are relevant to the phylogenetic discussion in this chapter are listed in column one. The table also gives reference to the accession numbers of the complete mitochondrial genomes. The group Chelicerata contains the horseshoe crab, *Limulus polyphemus*, several spiders (Araneae), and several ticks and mites (Acari), plus many other groups whose mitochondrial genomes are not available. The group Myriapoda contains representatives of the centipedes (Chilopoda) and millipedes (Diplopoda). The groups Crustacea and Hexapoda are both very diverse and contain representatives of many different subgroups listed in Table 4.1.

The relationships among these groups has been debated for a long time, but evidence is now mounting to support the arrangement ((Chelicerata, Myriapoda), (Crustacea, Hexapoda)). The grouping of Crustacea and Hexapoda is known as Pancrustacea. This grouping is supported by sequence evidence (Shultz and Regier 2000; Giribet et al. 2001). It is also supported by gene order evidence. It was shown (Boore et al. 1998) that a tRNA-Leu gene has been translocated in the common ancestor of Crustacea and Hexapoda. This argument was confirmed by Higgs et al. (2003) using a combination of sequence analysis and gene order data. The pairing of Chelicerata and Myriapoda is less certain, but is suggested by the most recent results using combined 18S and 28S rRNA (Mallatt et al. 2004). An alternative possibility is that Myriapoda is a sister group to Pancrustacea and that Chelicerata branches prior to this, as found by Giribet et al. (2001) and Pisani (2004).

The Hexapoda group, as usually defined, contains the insects and also the springtails (Collembola). There has been recent debate regarding the position of Collembola, with some authors arguing they are not a sister group to insects. In this case Hexapoda would not be monophyletic. It may also be that the Crustacea are polyphyletic within the Pancrustacea group. We will discuss these issues after presenting our own results.

Group name	latin name	common name	NC number
Chelicerata			
Acari	Amblyomma triguttatum	ornate kangaroo tick	NC_005963
Acari	Carios capensis	softbacked tick	NC_005291
Acari	Haemaphysalis flava	hardbacked tick	NC_005292
Acari	Ixodes hexagonus	hedgehog tick	NC_002010
Acari	Ixodes holocyclus	paralysis tick	NC_005293
Acari	Ixodes persulcatus	taiga tick	NC_004370
Acari	Ornithodoros moubata	soft tick	NC_004357
Acari	Ornithodoros porcinus	soft tick	NC_005820
Acari	Rhipicephalus sanguineus	brown dog tick	NC_002074
Acari	Varroa destructor	honeybee mite	NC_004454
Araneae	Habronattus oregonensis	spider	NC_005942
Araneae	Heptathela hangzhouensis	spider	NC_005924
Araneae	Ornithoctonus huwena	Chinese earth tiger	NC_005925
Xiphosura	Limulus polyphemus	Atlantic horseshoe crab	NC_003057
Crustacea			
Branchiopoda	Artemia franciscana	brine shrimp	NC_001620
Branchiopoda	Daphnia pulex	water flea	NC_000844
Branchiopoda	Triops cancriformis	tadpole shrimp	NC_004465
Branchiura	Argulus americanus	fish louse	NC_005935
Cephalocarida	Hutchinsoniella macracantha	cephalocarid crustacean	NC_005937
Cirripedia	Pollicipes polymerus	goose barnacle	NC_005936
Cirripedia	Tetraclita japonica	Japanese acorn barnacle	NC_008974
Copepoda	Tigriopus japonicus	Tigriopus japonicus	NC_003979
Malacostraca	Cherax destructor	Australian freshwater crayfish	NC_011243
Malacostraca	Pagurus longicarpus	long-clawed hermit crab	NC_003058
Malacostraca	Panulirus japonicus	Japanese spiny lobster	NC_004251
Malacostraca	Penaeus monodon	black tiger shrimp	NC_002184
Malacostraca	Portunus trituberculatus	Japanese blue crab	NC_005037
Ostracoda	Vargula hilgendorfii	sea firefly	NC_005306
Pentastomida	Armillifer armillatus	tongue worm	NC_005934
Remipedia	Speleonectes tulumensis	remipede	NC_005938

Group name	latin name	common name	NC number
Hexapoda			
Coleoptera	Crioceris duodecimpunctata	spotted asparagus beetle	NC_003372
Coleoptera	Pyrocoelia rufa	firefly	NC_003970
Coleoptera	Tribolium castaneum	red flour beetle	NC_003081
Diptera	Anopheles gambiae	African malaria mosquito	NC_002084
Diptera	Chrysomya putoria	blow fly	NC_002697
Diptera	Drosophila melanogaster	fruit fly	NC_001709
Hymenoptera	Apis mellifera ligustica	common honeybee	NC_001566
Hymenoptera	Melipona bicolor	stingless bee	NC_004529
Lepidoptera	Antheraea pernyi	Chinese oak silkmoth	NC_004622
Lepidoptera	Bombyx mori	domestic silkworm	NC_002355
Lepidoptera	Ostrinia furnacalis	Asian corn borer	NC_003368
Orthoptera	Locusta migratoria	migratory locust	NC_001712
Paraneoptera	Aleurodicus dugesii	Doogie Howzer whitefly	NC_005939
Paraneoptera	Heterodoxus macropus	wallaby louse	NC_002651
Paraneoptera	Lepidopsocid RS-2001	scaly-winged barklouse	NC_004816
Paraneoptera	Philaenus spumarius	meadow spittlebug	NC_005944
Paraneoptera	Thrips imaginis	plague thrips	NC_004371
Paraneoptera	Triatoma dimidiata	kissing bug	NC_002609
Thysanura	Tricholepidion gertschi	bristletail	NC_005437
Collembola	Gomphiocephalus hodgsoni	springtail	NC_005438
Collembola	Tetrodontophora bielanensis	giant springtail	NC_002735
Myriapoda			
Chilopoda	Lithobius forficatus	centipede	NC_002629
Chilopoda	Scutigera coleoptrata	house centipede	NC_005870
Diplopoda	Narceus annularus	millipede	NC_003343
Diplopoda	Thyropygus sp.	millipede	NC_003344
Outgroups			
Mollusca	Katharina tunicata	black chiton	NC_001636
Brachiopoda	Terebratulina retusa	Terebratulina retusa	NC_000941

Table 4.1: 55 species were selected from 65 Arthropoda species in OGRe

4.2 Models

We have introduced the commonly used phylogenetic methods in Chapter 1. Now we continue to introduce all kinds of models which describe the mutation rates of nucleotides or amino acids in the sequences..

4.2.1 DNA Models

The earliest DNA model is Jukes-Cantor model (Jukes and Cantor 1969). This model simply assumes all 4 nucleotides evolve at the same rate. The Kimura two-parameter model (Figure 4.1) assumes the rates for transition and transversion are not equal, but the ratio of transition/transversion remains constant.

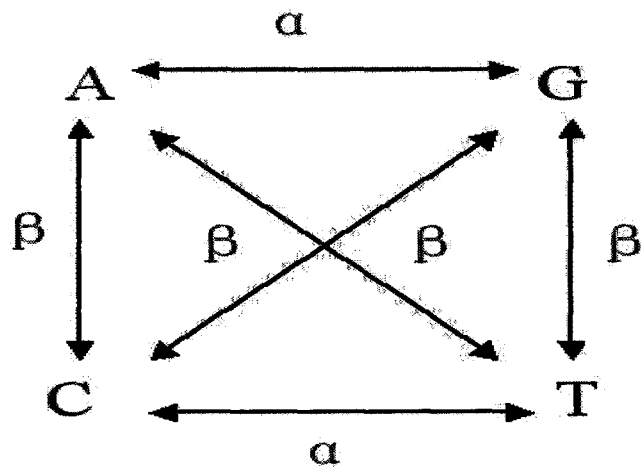

Figure 4.1: Kimura two-parameter DNA model. The transition (between C,T or A,G) rate is α, the transversion rate is β.

If we apply time reversibility restriction (the number of transitions from the first nucleotide to the second one is the same as the number from the second to the first) to the model and relax other restrictions, it will lead to a more flexible model that remains mathematically tractable..

The HKY model (Hasegawa et al. 1985) is one of these models. Table 4.2 shows the HKY model. The probability for a nucleotide A changes to another nucleotide G is $\alpha\pi_G \times \pi_A$, i.e. the rate of A to G ($\alpha\pi_G$) $\times$ the frequency of A nucleotide (π_A). The probability for a G to change to A is $\alpha\pi_A \times \pi_G$. And the two probabilities are equal. HKY model is very successful for its relatively simple mathematical abstraction and good approximation of real situations. It has been well accepted.

The GTR model is the general time-reversible model. Mutation rates between different pairs of nucleotides are different, hence there are 6 free parameters. It's the most flexible model under the time reversibility restriction.

From \ To	A	G	C	T
A	-	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	-	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	-	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	-

Table 4.2: HKY model. It's a time reversible model with transition rate α and transversion rate β.

4.2.2 tRNA Models

For maximum likelihood and Bayesian methods it is possible to use mixed information to infer phylogeny at once. Several genes can be put together while each gene is assigned with a different mutation rate; or some nucleotide or amino acid sequences can be put together; or sequences and secondary structure are considered together.

RNA sequence with secondary structure is another application. Figure 4.2 is a typical tRNA secondary structure. There are 4 stems (receptor stem, D-stem, TΨC-stem and anti-codon stem) and 3 loops (D-loop, TΨC-loop and anti-codon loop.) The anti-codon is located on anti-codon loop. Most tRNAs have the same structure

as illustrated by this figure, except a few tRNAs may lack a D-stem or TΨC stem. In the stem area, in addition to traditional Watson-Crick pairs, there exist another G-T pairs, although the interaction is much weaker. For the pairs, the change of one nucleotide can induce the change of the other one (compensatory substitution). For these area, different models should be applied.

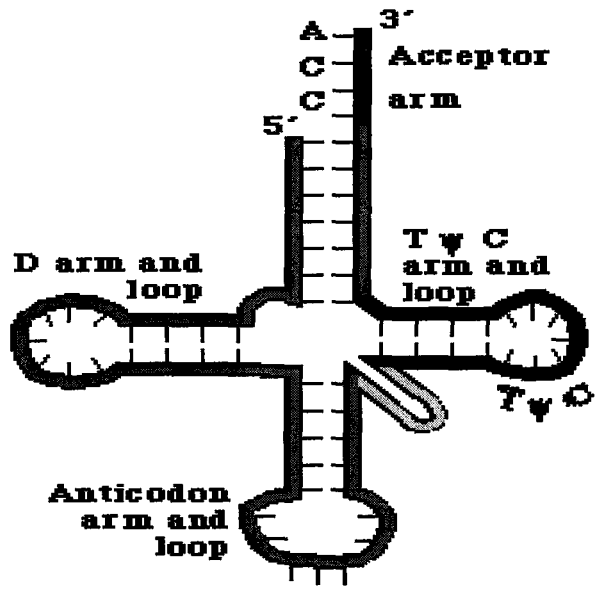

Figure 4.2: A typical tRNA structure.

The 6 states model (Tillier 1994) for paired area proposes that there are 6 states for paired nucleotides. The rate of change from one state to another via two transitions is 1; the rate of change via one transition is α_1 and the rate of change via two transversion is α_2. The 7 states model (Tillier and Collins 1998) is similar to the 6 states model, except one mismatch state is added and the rate related to mismatch state is α_3.

However, beside compensatory mutations, a slide of several nucleotides in a stem area due to insertions or deletions also can change the composition of stems, which we observed when we performed the alignment.

4.2.3 Protein Models

In contrast to the nucleotide models, most protein models are empirical. As different phylogenetic criteria are applied to different sets of data, different kinds of mutation probability matrices are obtained.

Dayhoff and Eck (1968) used parsimony criteria for a set of closely related sequences, and obtained the Dayhoff model. Later on, this model was updated by using 71 sets of closely related proteins in 1979 (Felsenstein 2004). In 1992, Jones, Taylor, and Thornton got their mutation probability matrix by using data containing not so closely related sequences. There are some specific models for mitochondrial and plasmid proteins. Adachi and Hasegawa used the maximum likelihood method and mitochondrial proteins from mammals, chicken, frog, fish, and lamprey and obtained the mtREV24 model in 1996. Then Yang, Nielsen, and Hasegawa, in 1998, used the maximum likelihood method, with variable rates in sites and proteins from 20 mammalian species, which became the *mtman* model (Yang 2004). These two models are time-reversible.

4.3 Software Packages

4.3.1 PHYLIP

PHYLIP (Felsenstein) is one of the earliest packages for constructing phylogenies, dating back to 1980. It is a powerful and comprehensive package, which can deal with many different types of molecular data, such as DNA, protein, restriction sites and gene frequencies, with different kinds of method, parsimony, distance matrix, maximum likelihood. Phylogeny trees also can be edited or drawn in PHYLIP.

We are interested in the maximum likelihood method in PHYLIP package. There are two main maximum likelihood programs in Phylip, dnaml for DNA sequences and proml for proteins. Both programs support the approximation of Gamma distribution for rates. dnaml supports HKY model but not the GTR model; proml supports Dayhoff and Jones-Taylor-Thornton model but without mitochondrial models (mtREV, mtman), which we are mostly interested in.

The proml program in PHYLIP implements maximum likelihood methods for proteins. It supports similar options as dnaml. However, it only supports Dayhoff and Jones-Taylor-Thornton mutation probability models. As we are interested in mitochondrial proteins, we prefer to use mtREV24 model or mtman model or even the general time reversible model.

4.3.2 PHASE

PHASE (Jow et al. 2002; Hudelot et al. 2003) is a MCMC Bayesian method package for analysis RNA sequences. It uses the secondary structure information for tRNAs and rRNAs. As explained before, in the paired areas, the mutation may behave like compensatory mutations and different mutation models must be applied. For compensatory mutations, there are 6-states, 7-states and more general 16-states models available.

4.3.3 Mr. Bayes

Mr. Bayes is another software that utilizes a Bayesian inference method. It has many similar properties to PHASE, however it is more general in some sense. We used it to analysis our data, and the results it retained are almost the same to PHASE.

4.3.4 PAML

PAML (Yang 1994) implements maximum likelihood methods. Compared to PHYLIP, it supports more mutation models (mtREV, mtman, GTR for DNA and protein) and options. One can specify the general time reversible model for amino acids and estimate the mutation probability matrix from the data. In this package, Yang utilized the discrete Gamma distributed heterogeneous mutation rates but also considered the correlation of rates at adjacent sites. This package provides many methods to investigate synonymous and non-synonymous substitution rates and methods to deal with codons.

The algorithm to find the maximum likelihood phylogeny runs slowly in PAML. The feasible method is to research the trees around the user specified tree or optimize the parameter according to user defined tree.

4.4 Analysis of the Data

4.4.1 Data Preparation

As we know, the number of unrooted trees for n species is $(2n-5)!!$. Closely related species may share the same gene order and have very little difference either in protein or DNA sequences. We pick only some species from such closely related groups, to decrease the tree space. Not much information will be lost by doing this. In the group Diptera, we only selected *Anopheles gambiae*, *Chrysomya putoria* and *Drosophila melanogaster* and deleted the other 8 species that were available. *Antheraea pernyi*, *Ostrinia furnacalis* and *Bombyx mori* were selected and another 2 were deleted in the group Lepidoptera. Two non-arthropod species (*Terebratalia transversa* and *Katharina tunicata*) are added as outgroups. The total number of

species is 57. We picked out all the tRNA genes in these 57 species; the number of tRNAs was slightly smaller than 22×57, because several species lack some tNRAs. (*Melipona bicolor* lacks tRNA-Cys and tRNA-Gln; *Aleurodicus dugesii* lacks tRNA-Gln.) Duplicated or pseudo-genes are not considered. For proteins, we used the 4 largest genes, which are COX1, COX2, COX3 and CYTB. All 57 species contain these four genes.

4.4.2 Alignment

Alignment is a basic and important procedure in phylogeny. ClustalW (ClustalX for Windows) (Higgins and Gibson 1994) and T-coffee (Notredame et al. 2000) are two very important alignment programs. Their essential procedure is to align the most closely related sequences first and the most divergent ones last according to a guide tree generated by some easy method. T-coffee is an improvement of ClustalW. It builds a library to weight different sequence patterns automatically, then uses this library to finish the rest of procedures. The result of T-coffee is better than ClustalW; however, there is a big cost of speed.

We used ClustalX (the windows version of ClustalW) to align our tRNA sequences and used some already-aligned sequences (for which the secondary structure has been considered) from Jameson (2004) as a profile. Although T-Coffee is more powerful than ClustalW, it can not take full use of profiles, as does ClustalW. Nevertheless, our tRNA sequences are more divergent than the sequences Jameson (2004) has used.

Except a few tRNAs, the anticodon area was aligned well by ClustalX, for those were not, manual adjustments were needed. The nucleotides in the D-stem area is the second well-aligned area. The alignments for most of the sequences are clear, except several ones which have more extra nucleotides, possibly due to short duplication or

insertion. The number of pairs in the D-stem changed slightly. The nucleotides in the TΨC-stem area were aligned badly. Many tRNAs lack this stem and the number of pairs in this stem changes greatly, from 2 to 5 or more, in addition, the nucleotides in well recognized pairs change greatly, which means these TΨC-stems may not evolve from single ancestral TΨC-stem, and, as mentioned before, there exits evidence that slides of several nucleotides can change the structure (number of pairs) and the content (the nucleotide composition) of TΨC-stem so the compensatory mutation model fails here. The nucleotides in the loop areas diverge a lot. After figuring out the nucleotides in the stem areas, I identified the nucleotides in loop areas and ran ClustalX then pasted the results back to the sequences. Because of the shortness and divergence of the nucleotides in loop areas, the second alignments were not very good either. I also adjusted them manually. When we are not sure about the structures of tRNAs, we refer to papers reporting sequencing of individual genomes, since there often give tables of tRNA structures. (Crozier and Crozier 1993; Spanos et al. 2000; Miller et al. 2004; Lessinger et al. 2000; Stewarta and Beckenbachb 2003; Creas 1999; Ishiwa and Chigusa 1987; O.Clary and R.Wolstenholm 1983; Nardi et al. 2003; Mastal and Boore 2004; Lavrov et al. 2000; Yamauchi et al. 2002; Yamauchi et al. 2004; Stewart and Beckenbach 2005; Bae et al. 2004; Machida et al. 2002; Umetsu et al. 2002; Friedrich and Muqim 2003; Navajas et al. 2002; Ogoh and Ohmiya 2004)

The alignment of the protein sequences were simple. We aligned them using ClustalW and T-Coffee. The result from T-Coffee was slightly better than were those obtained from ClustalW. Unlike DNAs or tRNAs, the alignment of proteins is much less complex. The amino acids sequences were not as divergent as were the tRNAs. The similarities were really high. We used the alignment result from T-Coffee, after deleting the sites with lots of gaps, as the final alignment.

4.5 Discussion of the Protein and tRNA Trees

The consensus tree from the MCMC analysis of the protein sequences is shown in Figure 4.3. Most parts of this tree are well resolved, and many parts make sense according to our expectations from previous studies. There are, nevertheless, several aspects of this tree that are definitely inconsistent with what we observe from other data, and thus it is clear that systematic biases are affecting the position of several species in this tree. From the figure, we see that the two outgroup species fall together and therefore the arthropods are monophyletic, as expected. There is a split between Myriapoda + Chelicerata at the top of the figure and the Crustacea + Hexapoda at the bottom. This agrees with the evidence for the relationship between these four groups, as discussed in section 4.1.

There is however, a group of 7 crustacean and hexapod species running from Speleonectes to Apis that apparently fall within the chelicerates. This is biologically untenable, and must be due to phylogenetic artifacts such as long branch attraction and bias due to variation in base frequencies. This group of 7 contains rather diverse species that are probably not all related. The two Hymenoptera (*Apis* and *Melipona*) belong with the other holometabolous insects at the bottom of the figure (Diptera, Lepidoptera and Coleoptera). The other three insects in this group (*Heterodoxus, Aleurodicus* and *Thrips*) belong with the rest of the Paraneoptera. *Armillifer* and *Speleonectes* are crustaceans of uncertain taxonomic position. It is unlikely that either of these latter two species is closely related to the other species in this group of 7. Another problematic point in Figure 4.3 is the position of *Tricholepidion*, which appears in the middle of the insect group. This is a wingless insect that is almost certainly basal to all the other insects in this study (which are winged) Nardi et al. (2003).

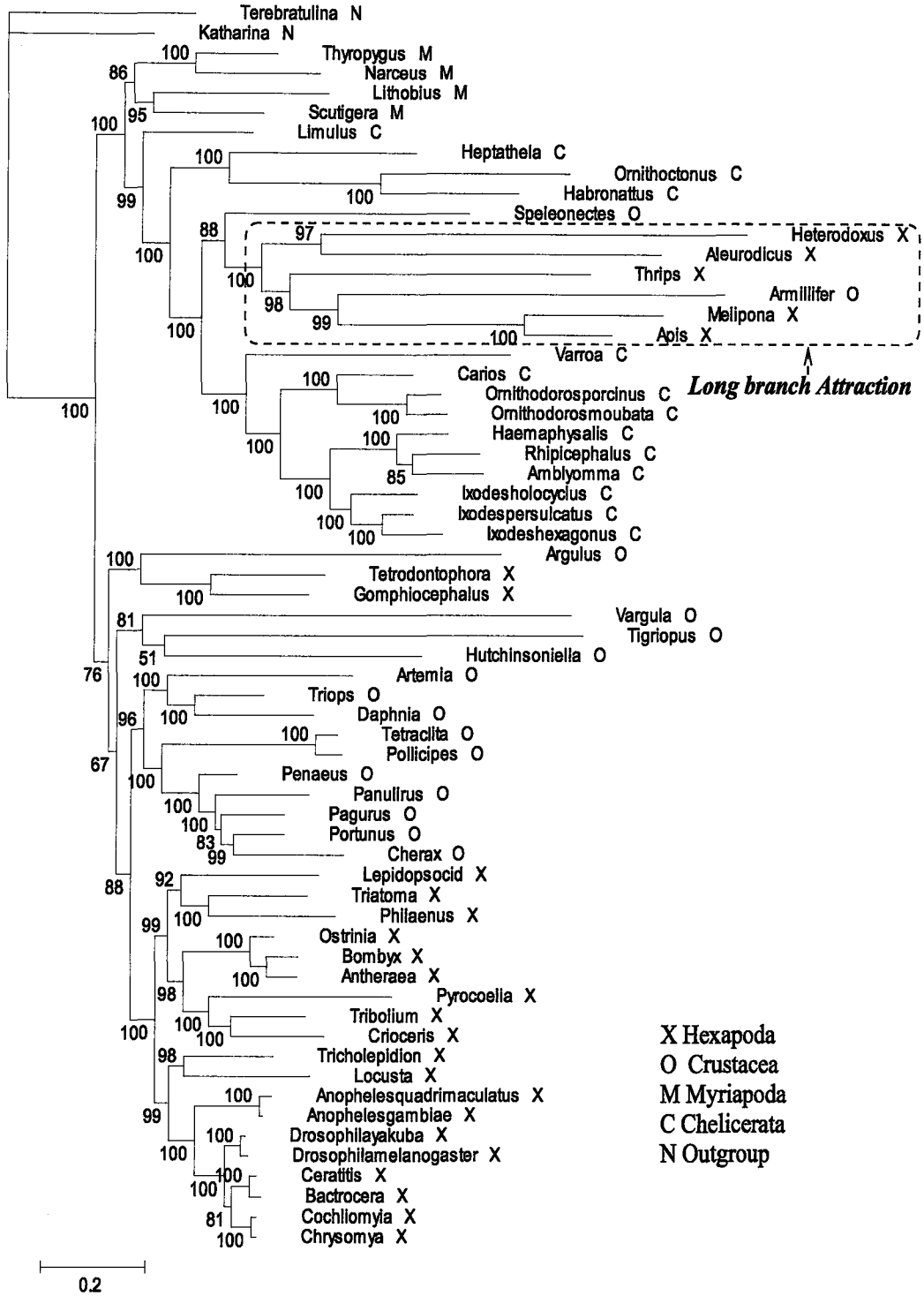

Figure 4.3: The result tree of proteins using PHASE. The number labeled along the branch is the posterior frequency for that branch appearing in all the sampling trees.

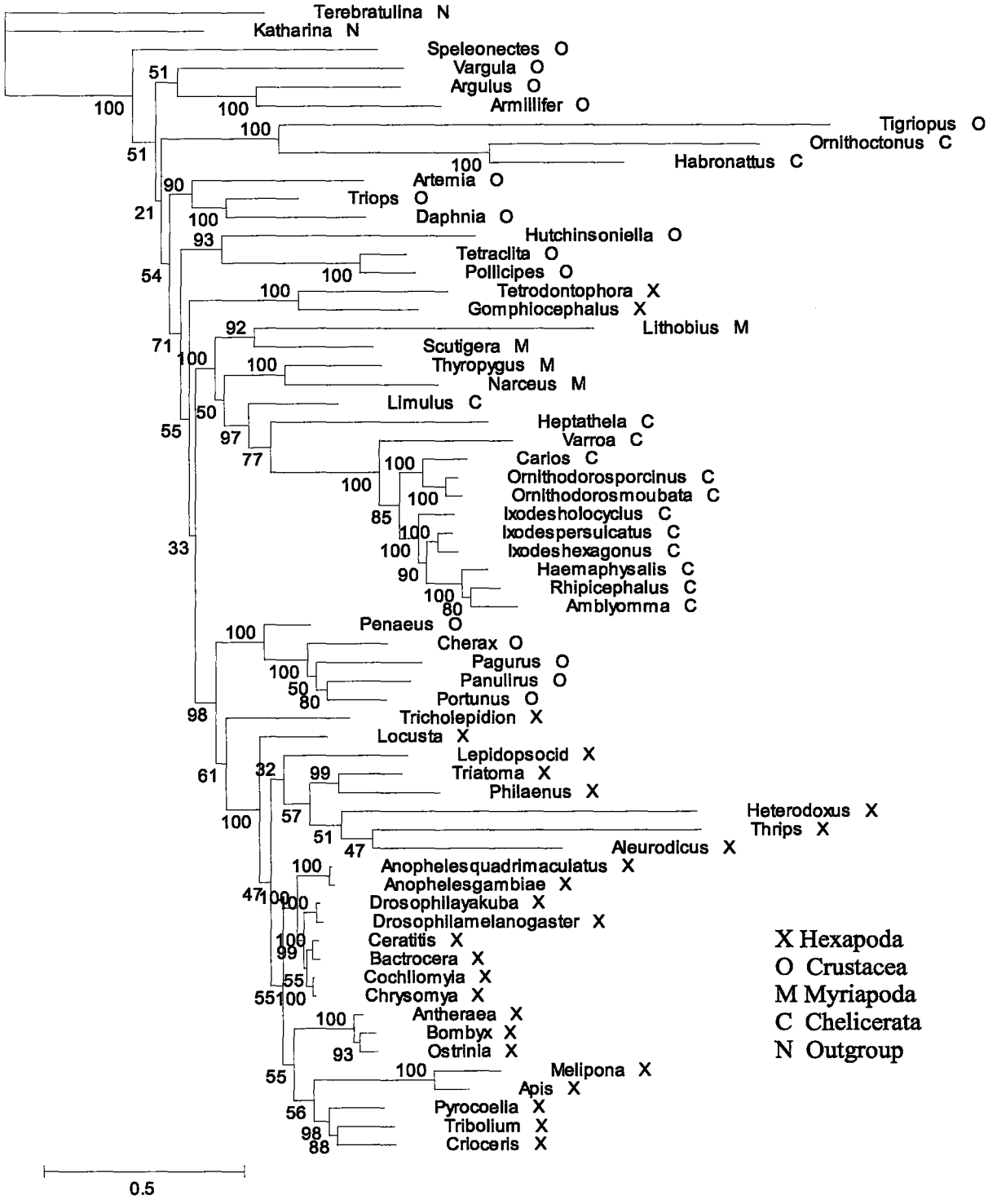

Figure 4.4: The result tree of tRNA using PHASE. The number labeled along the branch is the posterior frequency for that branch appearing in all the sampling trees.

In contrast to these misplaced species, there are also large parts of the tree that appear correct. For example, the myriapods are monophyletic, with the expected split between centipedes and millipedes. Were it not for the group of 7 misplaced species, the phylogeny of the chelicerates would also make sense: *Limulus* is basal, and Araneae and Acari are monophyletic. Several other groups of crustaceans and hexapods are also monophyletic: Collembola, Branchiopoda, Malacostraca, Lepidoptera, Coleoptera and Diptera.

It is useful to compare the protein tree with the tree derived from the concatenated tRNAs see Figure 4.4. The first observation is that the root is among a group of crustaceans. This seems a clear case of long branch attraction between some of the more divergent crustacean sequences (like *Speleonectes*) and the outgroups. If the outgroups are ignored, however, it is possible to reroot this tree so that there is a split between myriapods+chelicerates and crustaceans+hexapods, as in the protein tree. In fact there are some parts of the tRNA tree that appear more reliable than the protein tree. The group of 7 problematic species that were in the middle of the chelicerates in Figure 4.3 are no longer there. The Hymenoptera are with the other holometabolous insects. All six Paraneoptera are monophyletic.

Armillifer is now paired with the crustacean *Argulus*. In fact, *Armillifer* is a member of a highly derived parasitic group known as Pentastomida, whose taxonomic position was very uncertain from morphology. Molecular phylogeny has placed Pentastomids with *Argulus*, both using nuclear 18S rRNA (Abele et al. 1989) and mitochondrial protein sequences (Lavrov et al. 2004). Our tRNA analysis is further confirmation of this result. However, the mitochondrial protein result is obviously sensitive to the set of species included, since we observed above that *Armillifer* was attracted to the group of 7 misplaced species in our own protein tree. We repeated the protein analysis using only the pancrustacean species in our data set. In this

case, *Armillifer* and *Argulus* were again paired. We, therefore, consider this issue to be settled.

One further positive point regarding our tRNA tree is that *Tricholepidion* is basal to the winged insects, whereas this was not true with the protein tree. On the other hand, there are also some points that appear worse in the tRNA tree than the protein tree. The myriapods are no longer monophyletic, and only one of the three spiders (*Heptathela*) remains with the chelicerates. The other two (*Ornithoctonus* and *Habronattus*) have jumped to an evidently false position within Crustacea. These two species have very unusual tRNAs that appear to be incomplete at the DNA level and are formed into functional tRNA molecules only by RNA editing (Masta and Boore 2004). It is, therefore, not surprising that these species are out of place in the tRNA phylogeny. There appears to be a long branch attraction between these sequences from two spiders and the very divergent sequences of *Tigriopus*.

The above discussion compares the ability of the tRNA and protein phylogenies to recover groups for which we already have good evidence. This may give the false impression that the correct tree is known in its entirety. In fact, there are many unresolved issues about which we would like to have more information on, the most important being the relative branching order of the subgroups of crustaceans, and the relationship of these groups to Collembola and the insects. These trees make predictions on these questions; however, our ability to draw conclusions is undermined by the presence of certain species that are clearly misplaced due to bias. We do not know whether to trust the parts of the tree where we have no firm prior expectations.

We have only presented one tree for each of the proteins and tRNAs. In fact, we have carried out many different phylogenetic methods on these data sets. For the tRNAs, we tried using a combination of 7-state models for the paired regions and 4-state models for the unpaired regions. This allows the model to account explicitly

for the occurrence of compensatory substitutions in the RNA helices. Although this method has proved useful in several other problems we have studied in our group, it did not resolve the problems seen in the tRNA tree shown (which uses the 4-state model for all sites and ignores secondary structure). We also tried using a two-state model (that accounts for purines and pyrimidines only) with no noticeable improvement over the four state model. In the case of the proteins, in addition to the MCMC analysis with PHASE shown in Figure 4.3, we also tried maximum likelihood methods with PHYLIP package, the quartet-puzzling methods with Tree-puzzle program. These methods produced trees differing in small respects from those shown, but none was clearly more reliable than the two examples shown in the figures.

Our conclusion is that none of the available methods is able to deal reliably with the problematic features of this data set. The data include species that are very divergent, which make the results prone to long branch attraction, and also contain species with widely varying base compositions.

4.6 A Best Estimated Tree for the Arthropods

In the following chapter, we wish to compare several different measures of evolutionary rates for different species, and we require a best estimate of the arthropod tree on which to carry out this analysis. In this section, we will combine the evidence from our own phylogenetic studies discussed in this chapter with other published evidence to produce our best estimate of the tree. There is reliable evidence to support the majority of nodes in this tree, but, in a few cases where there is conflicting evidence, we have left multifurcations remaining in the tree. The resulting best estimate tree is shown in Figure 4.5. We will now summarize our reasons for selecting this tree.

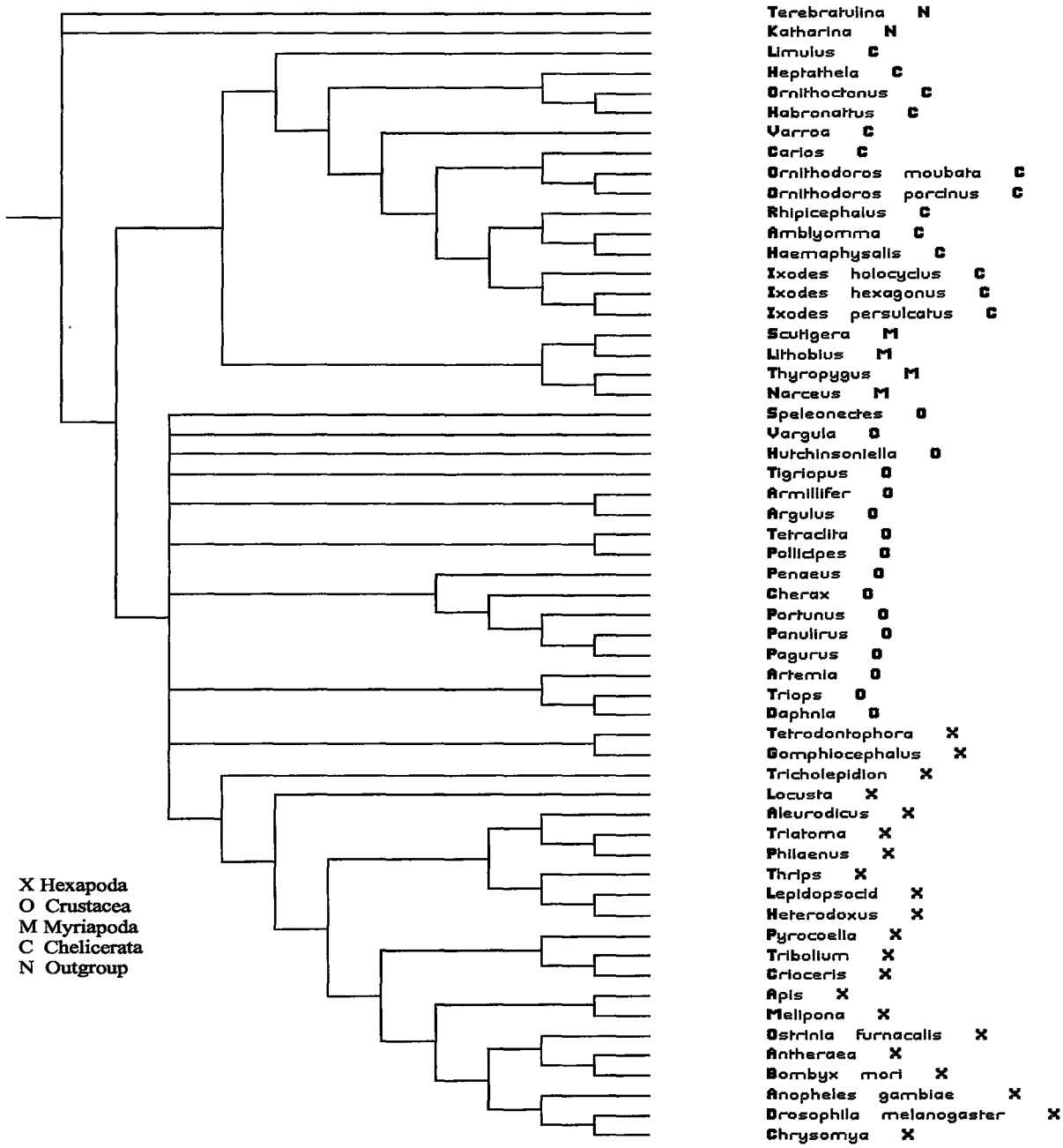

Figure 4.5: The phylogeny of Arthropoda

The relationship of the four principle groups is taken to be ((Chelicerata, Myriapoda), (Crustacea, Hexapoda)). The best evidence for this is the combined 18S and 29S rRNA study of Mallatt et al. (2004). Our own protein results support this (Figure 4.3) although our tRNA results are equivocal due to a problem with the position of the root of the arthropods (Figure 4.4). We note that the alternative (Chelicerata, (Myriapoda, (Crustacea, Hexapoda))) also remains a possibility that cannot be ruled out (Giribet et al. 2001; Pisani 2004).

Although the traditional view is that myriapods are monophyletic, there has been some molecular evidence to the contrary. We will take the myriapods to be monophyletic. This agrees with our protein tree, and with the conclusions of Mallatt et al. (2004). Our tRNA tree would suggest parphyly of the myriapods, but, in our opinion, this is an artifact. Given the monophyly of Myriapoda, the split between Chilopoda and Diplopoda is not controversial.

Within the chelicerates, the positioning of *Limulus* at the base is also non-controversial, and is supported by both tRNA and protein analysis given here. The remaining chelicerates are arachnids, and of the many orders of arachnids only the Acari and Araneae are represented among the complete genomes available. This split is therefore not controversial. We take the detailed phylogeny of the species within these two groups to be as obtained from the protein tree in Figure 4.3. This tree is consistent with classification of these species in the NCBI taxonomy. The tRNA tree appears less reliable here due to the problem mentioned above with *Habronattus* and *Ornithoctonus*, and also because of the slight rearrangement of the three *Ixodes* species, which we assume should really be monophyletic, as in the protein tree.

Although the Pancrustacea group as a whole is well supported, the arrangement of the early branching groups within it is very unclear. In our own studies, the relationships of these groups are not consistent in the tRNA and protein trees, and

they are sensitive to the evolutionary model use and to the set of species included. Several papers that include crustacean phylogenies are: Regier and Shultz (1997); Shultz and Regier (2000); Wilson et al. (2000); Richter (2002); Mallatt et al. (2004); Regier et al. (2005); However there is no consensus of these results and we do consider any of these to be definitive. We have, therefore, left a large number of groups branching simultaneously at this point. The subgroups of Pancrustacea that are well supported by our own data and consistent with previous papers are shown in Figure 4.5. These are the *Armillifer/Argulus* pair, Cirripedia, Malacostraca, Branchiopoda, Collembola and Insecta. The relationship of Collembola and Insecta has been debated in recent papers(Nardi et al. 2003; Delsuc et al. 2003). If these two groups are not sisters, the Hexapoda, as usually defined, is paraphyletic. We do not consider this matter resolved, and we do not believe the available mitochondrial sequence data is sufficient to resolve this large multifurcation.

There are five representatives of Malacostraca in our data. The detailed phylogeny of these species is not quite the same in our protein and tRNA trees. A more detailed study of this group has been given by Morrison et al. (2002). Based on their evidence we take the relationship between these five species to be that shown in Figure 4.5. The relationship between *Artemia*, *Daphnia* and *Triops* is consistent in our protein and tRNA trees, and also agrees with the results of Spears and Abele (2000).

The relationship of the orders within the insects has been widely studied. One of the most complete papers on this is that of Wheeler et al. (2001), and we have followed this. Extracting the relevant groups for our data set from the summary figure 20 of Wheeler et al. (2001) gives: (Thysanura, (Orthoptera, (Paraneoptera, (Coleoptera, (Hymenoptera, (Lepidoptera, Diptera))))))).

The last four listed orders are holometabolous (insects that go through a full metamorphosis). The relationship between these orders is quite hard to reslove, in

particular because of the unusual base composition of the Hymenoptera (*Apis* and *Melipona*). This problem was noted above in our own protein phylogenies. Castro and Dowton (2005) have also addressed this problem with a new genome from the Hymenoptera, *Perga condei*, not contained in our data set. The relationship between the orders depends on the evolutionary model used, but they conclude that when the most realistic models were used, Hymenoptera is a sister to (Lepidoptera + Diptera), as above.

The detailed phylogeny of species within the insect orders is largely non-controversial, with the exception of the six species listed as Paraneoptera (which is a higher level taxon, not a single order). The species in our study are representatives of four different orders: Hemiptera (*Aleurodicus, Triatoma, Philaenus*), Thysanoptera (*Thrips*), Psocoptera (*Lepidosocid*) and Phthiraptera (*Heterodoxus*). These include long-branch species that are problematic in our protein tree. In our tRNA tree, although the six species are monophyletic, the three Hemiptera are not monophyletic. We have again decided to go with the relationships between these orders given by Wheeler et al. (2001), as the sequence-based evidence does not seem very reliable in this group.

Chapter 5

Calculating the Correlation of Gene Order and Sequence Information

The goal of our project is to find the correlation coefficient between sequence information and gene order information. To calculate the correlation of two kinds of molecular information, we have thought about two methods. The first method is to build up a topology tree for the Arthropoda group; then, based on that tree, we use gene orders and sequence information to construct phylogenetic trees with branch lengths. As the trees from different sources have the same topology, we can calculate the correlation of pairwise branch lengths. The second method is to use a distances matrix. The distance matrix contains pairwise distances between all species. We can compare the distances for different kinds of information for the same species pairs. This method does not require any phylogenetic tree.

Figure 5.1 demonstrates the two methods. In figure A, there is a phylogenetic tree on which two different distance measures are shown (the second type of distances are either shown below the branches or in parentheses). The distances are measured from the common ancestor O to the tips. The correlation should be calculated from pairs of distances. In figure B, no phylogenetic tree is required.

If we think abstractly, for one kind of information, there are a set of distances, and we can denote them as a distance vector. For two kinds of information, there are 2 distance vectors. There are two kinds of different correlation questions: self-

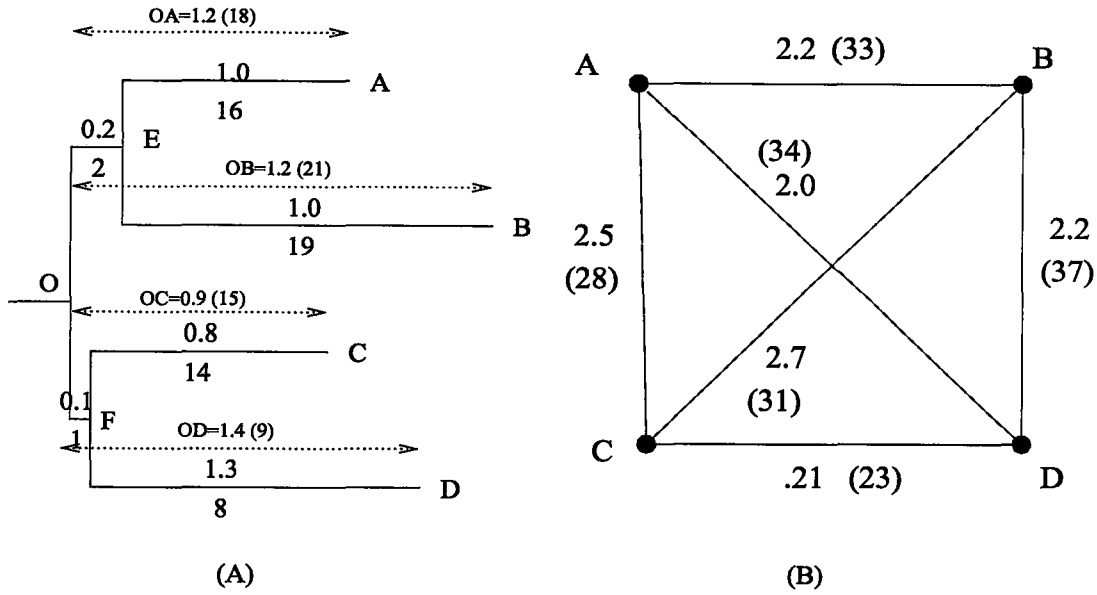

Figure 5.1: Two methods to calculate the correlation: A uses a phylogenetic tree; B uses a distance matrix

correlation and cross correlation. Self-correlation means that changing one element in one distance vector will affect other elements in the same vector. Cross-correlation means changing the value of one distance vector will affect the value of the other vector. Cross-correlation is what we want to calculate and self-correlation is what we want to avoid.

However in both methods, self-correlation is difficult to avoid. In Method One, distances OA and OB both contain distance OE and distances OC and OD both contain distance OF. In Method Two, given n species, there are $2n-3$ number of independent variables. However, in distance matrix there are $\frac{n(n-1)}{2}$ distances, which are highly correlated.

From the above discussion we can see that Method One has less self-correlation than Method Two, especially when inner branch lengths like OE and OF are small.

Hence, we prefer Method One, though this method depends the predefined topology tree.

There is a modified method from Method One. We can compare each branch length (OE,EA,EB, etc.) instead of branch lengths like OA, OB. However, this method has a strong dependence on tree topology, and this method is not applicable for gene order information. Another reason is we do not have a reliable estimate of the gene order of every internal node of the tree. To get a phylogenetic tree with each median gene order labeled for Arthropoda species may require more than several years' computation time.

5.1 Method

The ancestral gene order for Arthropods is assumed to be that of *Limulus polyphemus* (Lavrov et al. 2002). Figure 5.2 shows that the gene order of *Limulus polyphemus* is closest to the ones of non-Arthropod in Arthropod. The gene order of *Drosophila* which is regarded as the ancestral gene order of insects is nearly identical to *Limulus polyphemus*, except that there is a translation of one Lys tRNA(L2). Compared with other gene orders in Arthropod, like *Narceus annularus* and *Pagurus longicarpus*, the gene order of *Limulus polyphemus* is closer to gene orders of *Kathrina tunicata* and *Homo sapiens*. We calculated breakpoint distance and inversion distance (number of duplicated and deleted should be considered before calculation). The cases where the two gene orders have different set of genes should be paid attention. For breakpoint calculation, we took the distance to be the number of breakpoints in the larger of the two genomes. This includes the breakpoints caused by the duplication or deletion. However, for the calculation of inversion, the two gene orders should have same set of genes. In unidentical situation, our method is first to calculate how many genes

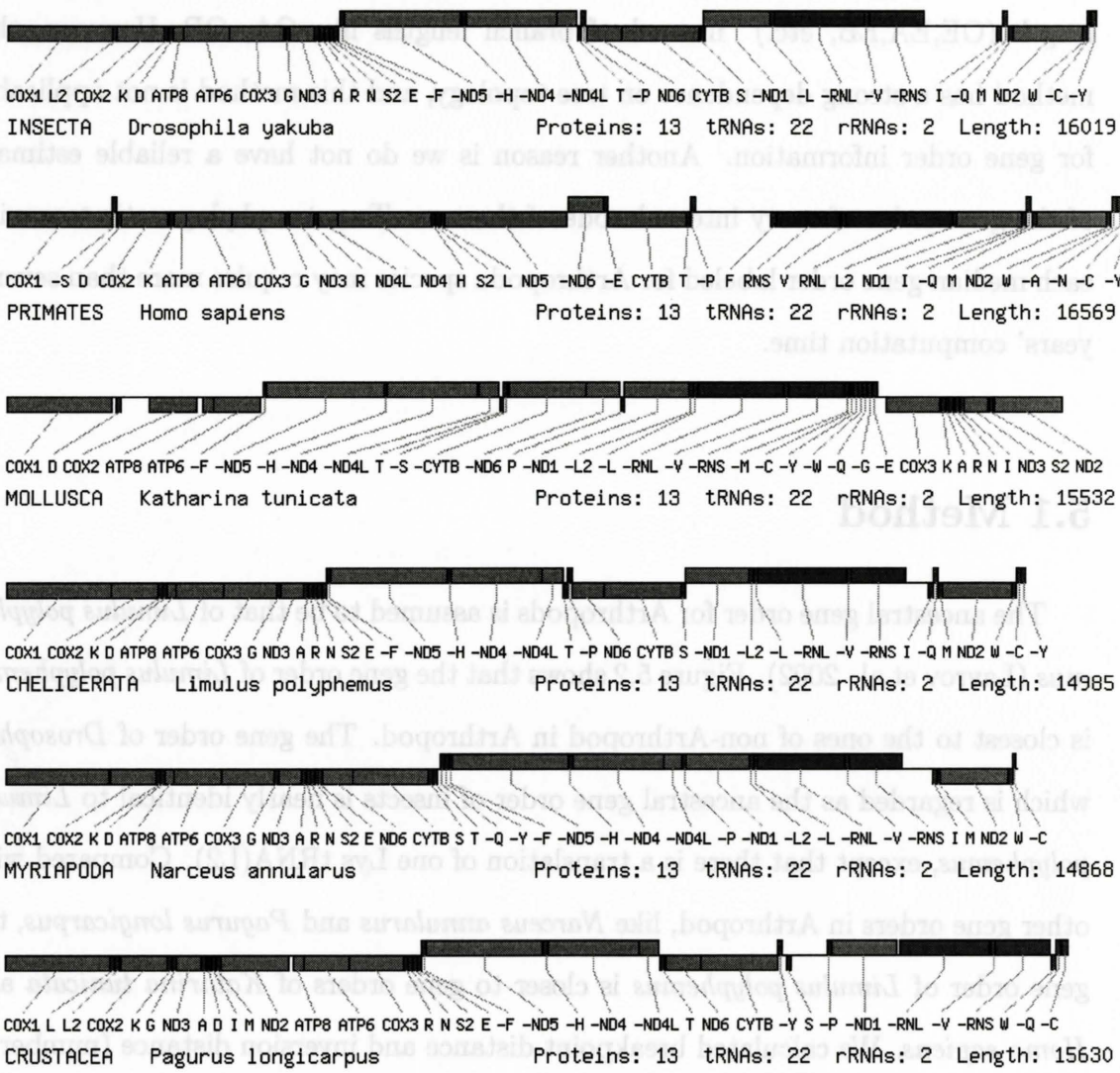

Figure 5.2: The ancestral gene order of Arthropod. The gene order of *Limulus polyphemus* is closest to the ones of non-Arthropod among Arthropod and its gene order is regarded as the ancestral gene order of Arthropod.

involved duplication and deletion. After deleting those genes, the two gene orders have identical gene set. And we report the inversion distance together with the number of duplications and deletions (dd number). The calculation of breakpoint number is straight forward. For the inversion number calculation, we used the GRAPPA program.

Using the best estimate of the Arthropod tree derived in Chapter 4, we calculated branch lengths with Yang's PAML package, because of its flexibility and power. The tree space searching algorithm is not very efficient, however, in our case; we need to optimize the other parameters for a given guide tree. In both protein and tRNA cases, we choose the hierarchical model of mutation rates approximated by a discrete Gamma distributions with 8 categories.

For the tRNA sequences, we did not consider the secondary structure information, because the branch lengths have a more straightforward interpretation when the single site model is used. From our several trials with second structure information, we should admit that the secondary structure information does not help us at all.

5.2 Results

For a given data, there might be several models applicable. The model with more free parameters will lead to a better performance (for maximum likelihood method, it will have a higher likelihood value). The question is whether the better performance of parameter-rich model is because of its ability to capture more characteristics of reality or just because of overestimation. Goldman (1993) established the likelihood ratio test (LRT) for phylogeny methods adopting likelihood functions. Given two models, model A and model B, and model A has n more number of free parameters than model B, and the pivotal quantity σ is defined as two times of ratio of likelihood

functions of model A and B, ie. $\sigma = 2/times([likelihoodofA/likelihoodofB])$ or $sigma = 2 \times ([logarithmlikelihoodofA - logarithmlikelihoodofB])$. The pivotal quantity is χ^2 distributed with n number of freedom where n is the difference of the numbers of free parameters in two models. So, the question about the significance of parameter-rich model turns into a likelihood ratio test.

For tRNA, we tested both HKY model and the general time-reversible model (REV). In HKY model, there are 3 parameters for base frequencies and 2 parameters for mutation rates (α and β) and In REV model, there are 3 parameters for base frequencies and 6 parameters of mutation rates. So the number of degrees of freedom for σ is 4. After running PAML package, the logarithm likelihood value (for a problem with lots data, the likelihood is very small, the logarithm form is convenient despite its negative sign) for HKY model is -43882.128102 and -43802.215971, so σ is 159.8243. For χ^2 distribution with 4 degrees of freedom, the 95% confidence, the value is 9.487729. And σ value is dramatic larger than the 95% confident value. The conclusion is that REV model is significant better than HKY model for Arthropoda tRNA data.

For proteins, we tested both the mtREV model and the general time-reversible model (labeled as REVaa in PAML package). REVaa model has 189 free parameters, and the optimization requires a very long computation time. So, we followed the suggestion given by Yang (Yang 2005). Firstly, we used the mtREV model with 8-categories of Gamma distribution rate to estimate initial branch length and α (the only parameter for Gamma distribution model). Secondly, we fixed the branch lengths and α (the Gamma distribution model was switched off) value estimated in step one and used the REVaa model to estimate an initial value for the amino acid mutation matrix. Thirdly, we continued to use REVaa model and switched Gamma distribution

model on and started from the branch lengths, α value and mutation matrix estimated previously.

In both the mtREV and REVaa models, the frequencies of amino acid are not parameters. They can be either calculated from the distribution in the original data or copied from the specific models. Here, we chose to calculate frequencies from the data. For the mtREV model, all data were previously calculated, there is no free parameter. For the REVaa model, there are 189 free parameters. So the number of degrees of freedom for σ is 189. From the final result, the logarithm likelihood value is -68844.805733 and the logarithm likelihood value for REVaa model is -67728.187511, so σ is 2233.236. The 95%-confidence value for chi^2 distribution with 189 degrees of freedom is 222.0756. So, the conclusion is that we should have used the REVaa model instead of mtREV model for Arthropoda proteins.

tRNA	HKY	REV	# of freedom	95% value for χ^2	σ
	-43882.128102	-43802.215971	4	**9.487729**	**159.8243**
protein	mtREV	REVaa	# of freedom	95% value for χ^2	σ
	-68844.805733	-67728.187511	189	**222.0756**	**2233.236**

Table 5.1: Likelihood ratio tests for models of tRNA and protein

Figure 5.3 and Figure 5.4 are the results of PAML for tRNA sequences and protein sequences. Despite the same topology, the trees vary in branch lengths. The values of the distances for each species are shown in Table 5.2. The correlation coefficients for all the different distance measures are shown in Table 5.3.

It's not surprising that the correlation coefficient for breakpoint distance and inversion distance is as high as 0.99, because they are just different methods to describe the same information. The correlation coefficient for breakpoint distance and protein distance when breakpoint distance is no larger than 10 breakpoint numbers is -0.004184711 which is almost zero. The correlation coefficient when breakpoint distance is no larger than 22 breakpoint numbers is 0.3438061. The overall coefficient

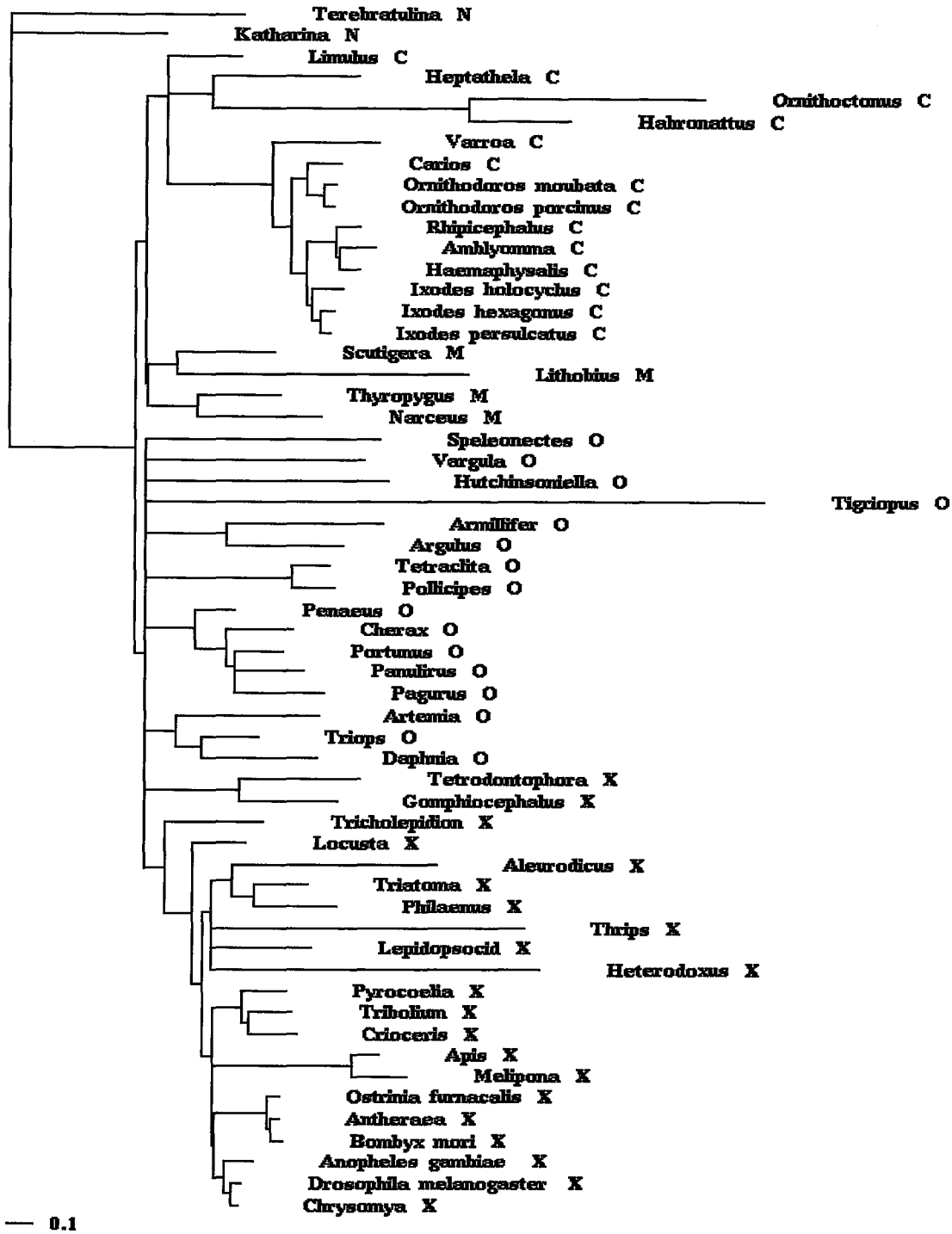

Figure 5.3: Phylogeny tree for tRNA sequence using PAML with predefined topology tree

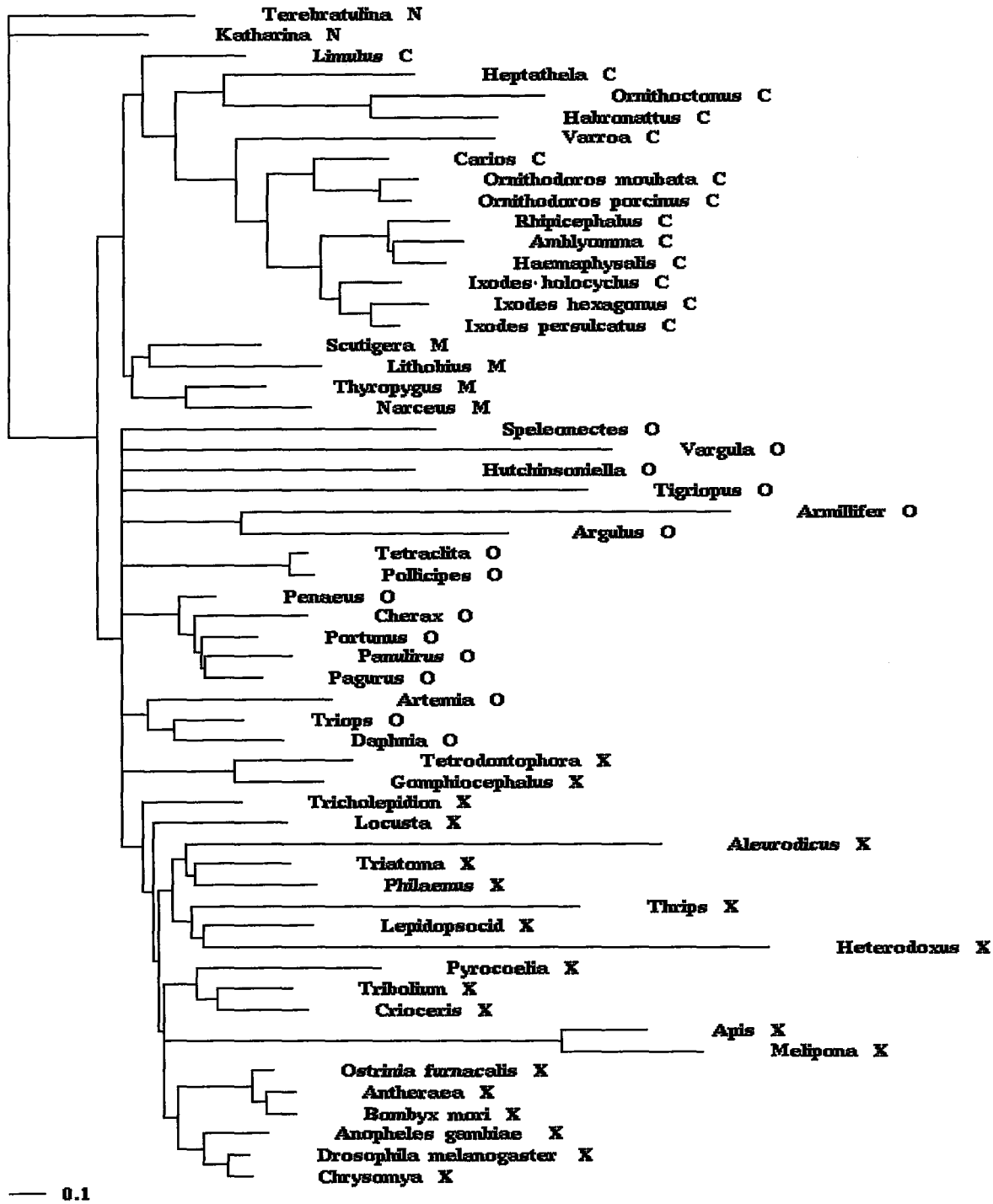

Figure 5.4: Phylogeny tree for protein sequence using PAML with predefined topology tree

species	break point	inversion	dup. and del	.trna dist	.prot dist
Carios	0	0	0	0.700731	0.792694
Heptathela	0	0	0	0.757864	0.870215
Ixodes hexagonus	0	0	0	0.736344	0.902641
Ixodes holocyclus	0	0	0	0.759042	0.827223
Ixodes persulcatus	0	0	0	0.719300	0.821178
Limulus	0	0	0	0.357568	0.401047
Ornithodoros moubata	0	0	0	0.684852	0.876607
Ornithodoros porcinus	0	0	0	0.674827	0.856152
Crioceris	3	2	0	0.554557	0.5765967
Daphnia	3	2	0	0.618057	0.507806
Drosophila melanogaster	3	2	0	0.367311	0.4165657
Gomphiocephalus	3	2	0	0.691279	0.618784
Lithobius	3	3	0	1.134174	0.611561
Panulirus	3	2	0	0.577519	0.528212
Penaeus	3	2	0	0.340374	0.320218
Philaenus	3	2	0	0.689365	0.5816597
Pyrocoelia	3	2	0	0.520205	0.7702787
Triops	3	2	0	0.419714	0.400225
Triatoma	3	2	0	0.589781	0.5025457
Tricholepidion	3	2	0	0.442811	0.393071
Chrysomya	4	2	1	0.355802	0.4214487
Antheraea	6	5	0	0.499159	0.5401667
Bombyx mori	6	5	0	0.511159	0.5435397
Locusta	6	5	0	0.384352	0.516007
Ostrinia furnacalis	6	5	0	0.494786	0.4758177
Portunus	6	5	0	0.506570	0.436482
Tribolium	6	5	0	0.545168	0.5288597
Amblyomma	7	6	0	0.878069	1.004787
Artemia	7	5	0	0.627421	0.637585
Haemaphysalis	7	6	0	0.818724	0.956307
Rhipicephalus	7	6	0	0.820515	0.964293
Aleurodicus	8	5	1	1.036355	1.5439677
Anopheles gambiae	8	6	0	0.412557	0.4694067
Tetrodontophora	8	6	0	0.770472	0.696898
Narceus	9	9	0	0.631712	0.579376
Thyropygus	9	9	0	0.493063	0.460492

species	break point	inversion	dup. and del	.trna dist	.prot dist
Armillifer	13	12	0	0.845627	1.733262
Melipona	14	8	2	0.932018	1.6587887
Varroa	14	12	0	0.827968	1.088804
Ornithoctonus	15	13	0	1.945021	1.226922
Scutigera	15	15	0	0.478776	0.441778
Habronattus	16	14	0	1.481066	1.091262
Lepidopsocid	17	16	0	0.604475	0.5873057
Vargula	17	15	0	0.786276	1.412554
Hutchinsoniella	18	16	0	0.860096	0.870521
Pagurus	18	12	0	0.645712	0.447337
Apis	19	16	0	0.839289	1.5046777
Speleonectes	19	16	1	0.834380	0.925381
Argulus	20	18	0	0.715587	1.124494
Cherax	20	16	0	0.542760	0.574595
Tetraclita	20	16	0	0.664085	0.574807
Pollicipes	22	16	2	0.685345	0.590624
Thrips	32	29	1	1.337410	1.3197147
Heterodoxus	35	32	0	1.390795	1.8337367
Tigriopus	35	32	0	2.146046	1.344014

Table 5.2: Branch lengths for 57 Arthropoda species using 4 different measures

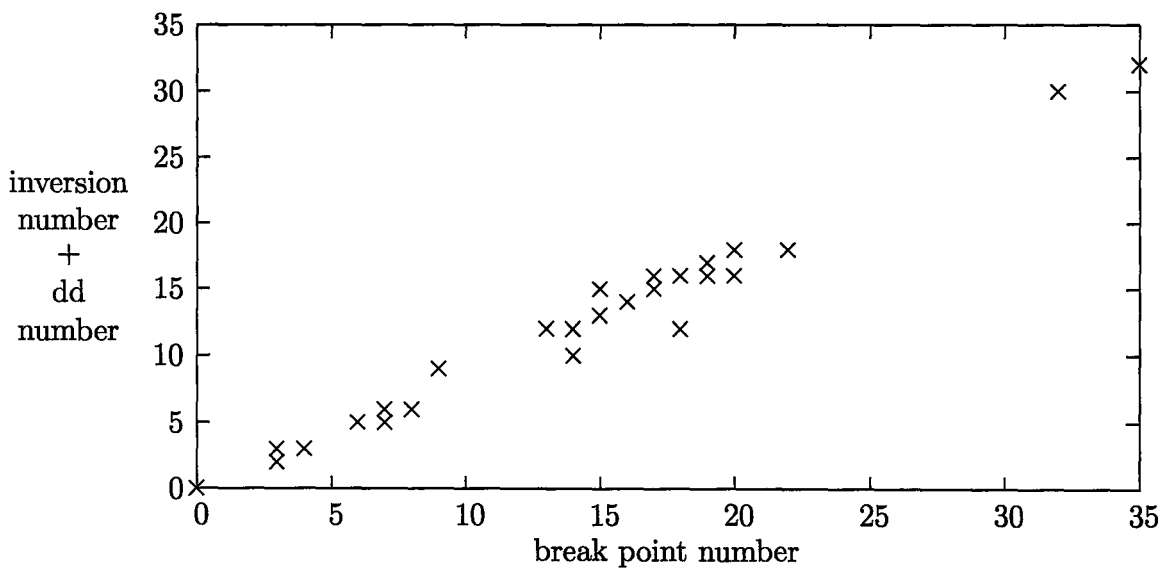

Figure 5.5: The correlation of breakpoint number and inversion number

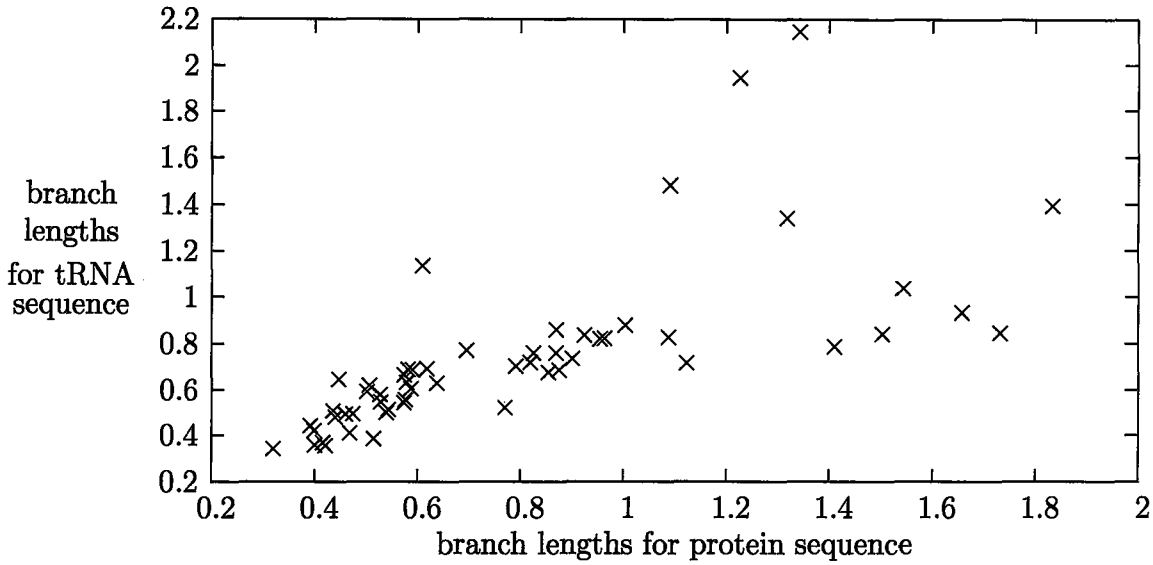

Figure 5.6: The correlation of protein distance and tRNA distance

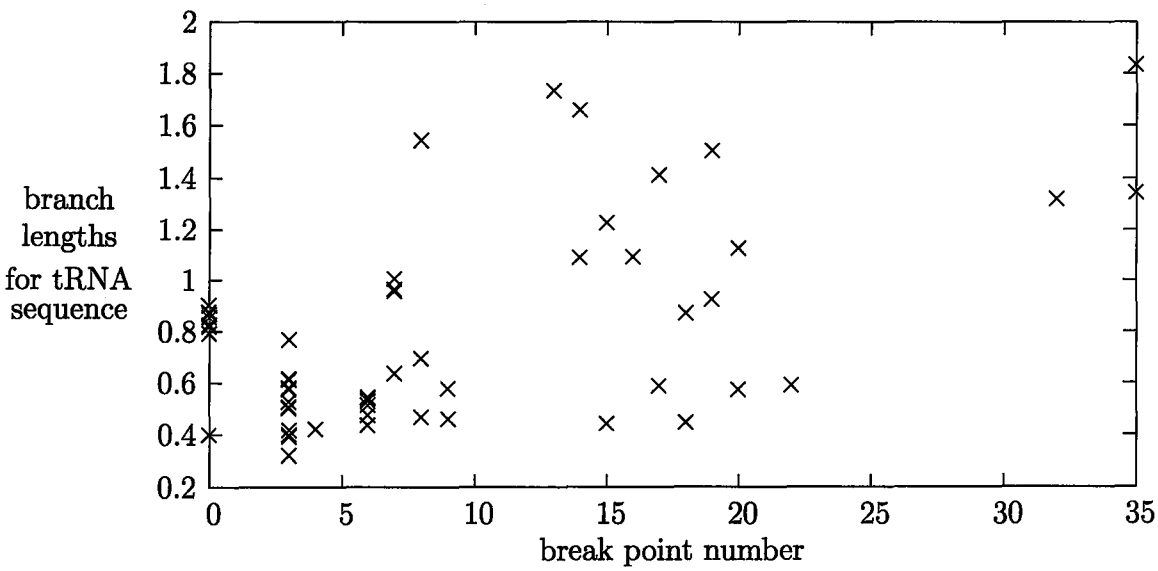

Figure 5.7: The correlation of breakpoint number and tRNA distance

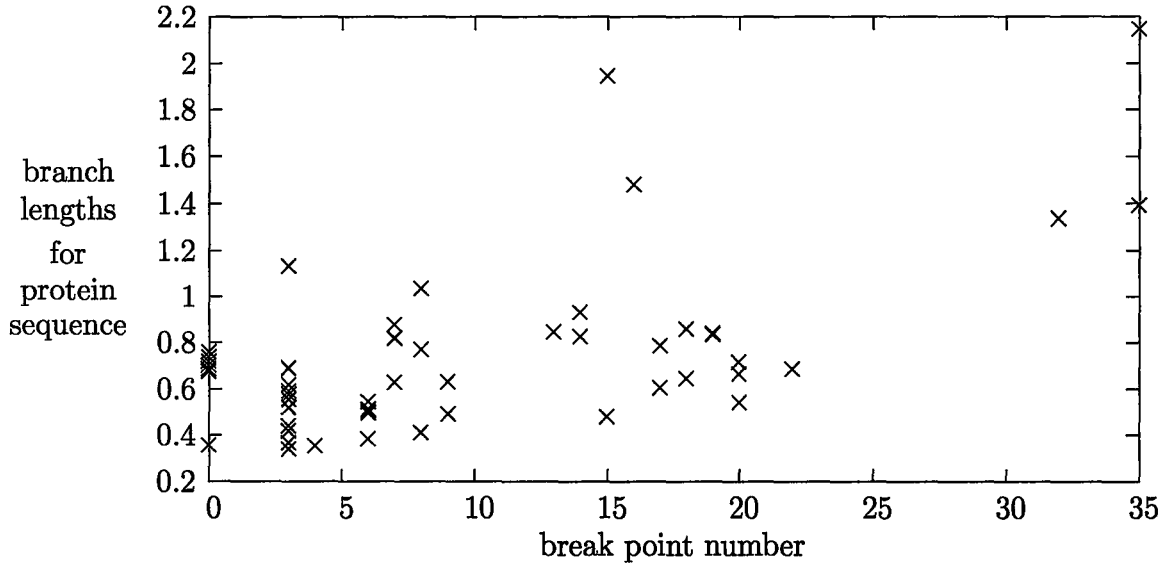

Figure 5.8: The correlation of breakpoint number and protein distance

	bp	**inv**	**trna**	**prot**
bp	1.0000000	0.9934813	0.5870733	0.5259077
inv	0.9934813	1.0000000	0.6025158	0.5352147
trna	0.5870733	0.6025158	1.0000000	0.6858336
prot	0.5259077	0.5352147	0.6858336	1.0000000

correlation	**bp<= 10**	**bp<= 22**	**overall**
bp VS protein	-0.004184711	0.3438061	0.5259077

correlation	**prot. dist.<= 1.1**	**overall**
protein VS tRNA	0.7685729	0.6858336

Table 5.3: The correlation tables

is 0.5259077. The correlation coefficient for protein distance and tRNA distance when protein distance is no larger than 1.1 is 0.7685729 and the overall coefficient is 0.6858336.

5.3 Discussion and Conclusions

The main points in this chapter are that all 4 distances are positively correlated with one another as shown in Table 5.3. The high correlation between breakpoint distance and inversion distance suggests that it doesn't matter too much which gene order rearrangement measure we use when we calculate the correlations between sequences and gene orders. Also the protein and tRNA distances are also strongly correlated which means generally that mutation influences protein and tRNA in a similar way. It won't make big difference whether we use protein or tRNA distance to discuss the correlation of gene orders with sequences.

The overall correlation of breakpoint distance with protein distance is high. In Figure 5.2, the species with long branch lengths always tend to be long in protein distances. From closer inspection of Figure 5.2, one can see that there is more randomness when mutation rate is low, and the correlation coefficient for species where the breakpoint distance is no larger than 10 is almost 0. It suggests that low mutation situations, the evolution of gene order and the evolution of sequence seem to be independent or very weakly correlated. Since gene orders and sequences evolve by different mechanisms it is not difficult to understand the very weak correlation. While on the other hand, under high mutation situations, large amounts of gene order rearrangements are always accompanied with large amounts of point mutations on sequences. This suggests that, when phylogeny analysis fails for sequences due to

long branch problems, the phylogeny analysis for gene orders also has little chance to succeed.

However, it is interesting to investigate the reason of high correlation of gene orders and protein distances under high mutation situations. Basically, gene orders and protein sequences evolve in differently. This high correlation can be the result of some factor which increases both the rate of sequence mutations and gene order rearrangements. One possible case is that when the accuracy of DNA replication enzyme reduces then the sequence mutation rate and the genome rearrangement rate increase simultaneously. The possible cause and effect relationship, between gene order and sequence evolution can cause a high correlation. Since recombination of genomes can change the gene order, if there exist large number of repeated sequences on genome, there will be a large chance for recombination to happen. Normally if the genome is at a stable state (not many genome rearrangement), there won't be many repeated sequences. If for some reason the point mutation rate on sequences becomes high, and by chance many new repeated sequences appear, the chance for genome recombination will be high, and hence there will be more gene order rearrangements. So it is possible that the unusual mutation rate on sequences can cause a large number of gene order rearrangements. Then one would like to ask, in the other direction, can the unusual change of gene order cause the high mutation rate of sequences? If one gene is changed to a different location or direction, and it becomes silent or the expression level becomes significantly lowered, it will be lethal to that species. However, if this gene belongs a gene family, the dysfunction of that gene may not influence the survivability of that species. Then this gene can evolve only under pure mutation without any selective force opposing it. Hence the substitution rate will increase because the usual rearrangements of gene orders. However, for mitochondria, this can not happen because there are no gene families in mitochondria

genomes. There is another possible way that the change of gene orders can change the subsitution rate of sequences. There are strong asymmetric mutation rates and different biases on codon usage between the two strands of DNA. If a gene changes its strand, it will find itself in a new mutation environment. Hence its substitution rate changes.

Taking a further look at Figure 5.2, we see that although generally the proteins and tRNAs are highly correlated, there are some differences for the correlations between high mutation and low mutation situations. Under low mutation situations, the correlation is stronger; while under high mutation situations the correlation is weaker. Both proteins and tRNAs genes are all sequences on a genomes. High correlation tells us that there is no obvious bias for genome evolution under low mutation rate. However, the weaker correlation under high mutation situations suggests that there should be some factors which have bias on either protein or tRNA genes during evolution. The selective pressure may influence the substitution rates of protein and tRNA genes, because selective pressure always acts on only one or several genes. However, we also note that when sequences are very divergent there is a large error in the estimate of evolutionary distances. So we would expect the scatter in the points on Figure 5.2 to be larger for high-distance points.

5.4 Future Work

Evolution and phylogeny analysis are interesting to me. From now-day information we can infer the history of life and by understanding the the history, we can know better about our human beings. What's more, this is an area that researchers from different disciplines can work through together. Personal speaking, I would like study this area from a mathematical and statistical view. This is a new, developing

and challenging area. There are many important questions have not been solved yet. About the gene orders, there is no statistical distance measure and the maximum likelihood methods haven't been applied to them. I am happy to continue the researches on this area and try to contribute as much as I can.

Bibliography

Abele, L., Kim, W., & Felgenhauer, B. 1989, Mol. Biol. Evol., 6, 685–691.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. 2002. *Molecular Biology of the Cell.* New York: Garland Science.

Altschul, S. 1991, Journal of Molecular Biology, 219, 555–565.

Altschul, S., Boguski, M., Gish, W., & Wootton, J. 1994, Nature Genetics, 6, 119–129.

Anderson, S., Bankier, A., Barrell, B., de Bruijn, M., Coulson, A., Drouin, J., Eperon, I., Nierlich, D., B.A. Roe, B. A., & Sanger, F. 1981, Nature, 290, 457–465.

Bae, J. S., Kim, I., Sohn, H. D., & Jin, B. R. 2004, Molecular Phylogenetics and Evolution, 32, 978–985.

Blanchette, M., Kunisawa, T., & Sankoff, D. 1996, Gene, 172, GC11–GC17.

Boore, J., Lavrov, D., & Brown, W. 1998, Nature, 392, 667–668.

Caprara, A. 1997, Proceedings of the First Annual International Conference on Computational Molecular Biology, RECOMB-97, 75–83.

Castro, L. & Dowton, M. 2005, Mol. Phyl. Evol., 34, 469–479.

Cavalli-Sforza, L. & Edwards, A. 1967, American Journal of Human Genetics, 19, 233–257.

Creas, T. J. 1999, Gene, 233, 89–99.

Crozier, R. H. & Crozier, Y. C. 1993, Genetic, 133, 97–117.

Dayhoff, M. & Eck, R. 1968. Atlas of protein sequence and structure.

Delsuc, F., Phillips, M., & Penny, D. 2003, Science, 301, 1482d.

Edwards, A. & Cavalli-Sforza, L. 1963, Annals of Human Genetics, 27, 105–105.

Embley, T., Horner, D., & Hirt, R. 1997, Trends in Ecology and Evolution *12*(11), 437–441.

Felsenstein, J. 1981, Journal of Molecular Evolution, 17, 368–76.

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer.

Fitch, W. & Margoliash, E. 1967, science, 155, 279–284.

Friedrich, M. & Muqim, N. 2003, Molecular Phylogenetics and Evolution, 26, 502–512.

Giribet, G., Edgecombe, G., & Wheeler, W. 2001, Nature, 413, 157–161.

Goldman, N. 1993, Feb, Journal of Molecular Evolution *36*(2), 182 – 198.

Grant, D. & Chiang, K.-S. 1980, Plasmid, 4, 82–96.

Gray, M. 1989, Annu. Rev. Cell Biol, 5, 25–50.

Halanych, K., Bacheller, J., Aguinaldo, A., Liva, S., Hillis, D., & Lake, J. 1995, Mar, Science *267*(5204), 1641–3.

Hannenhalli, S. & Pevzner, P. 1995, Proceedings of the 27th Annual ACM Symposium on the Theory of Computing, 178–189.

Hasegawa, M., Kishino, H., & Yano, T. 1985, Journal of Molecular Evolution, 22, 160–74.

Higgins, J. T. D. & Gibson, T. 1994, Nucleic Acids Res., 22, 4673–4680.

Higgs, P., Jameson, D., Jow, H., & Rattray, M. 2003, J. Mol. Evol., 57, 435–445.

Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M., & Higgs, P. 2003, Mol. Phyl. Evol., 28, 241.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. 2001, December, Science *294*(5550), 2310–2314.

Ishiwa, Y. S. H. & Chigusa, S. I. 1987, Mol. Biol. Evol, 4, 638–650.

Jameson, D. 2004. *The Comparative Analysis of Mitochondrial Genomes.* Ph. D. thesis, Manchester.

Jow, H., Hudelot, C., Rattray, M., & Higgs, P. 2002, Mol. Biol. Evol., 19, 1591–1601.

Jukes, T. & Cantor 1969, Mammalian Protein Metabolism, 3, 21–132.

Keogh, R. S., Seoighe, C., & Wolfe, K. H. 1998, Yeast *14*(5), 443 – 457.

Kovac, L., Lazowska, J., & Slonimski, P. 1984, Mol. Gen. Genet, 19, 420–4.

Lavrov, D., Boore, J., & Brown, W. 2002, Molecular Biology and Evolution, 19, 163–169.

Lavrov, D., Brown, W., & Boore, J. 2004, Proc. R. Soc. Lond. B, 271, 537–544.

Lavrov, D. V., Brown, W. M., & Boore, J. L. 2000, Proceedings of the National Academy of Sciences, 97, 13738–13742.

Lessinger, A. C., Junqueira, A. C. M., Lemos, T. A., Kemper, E. L., da Silva, F. R., Vettore, A. L., Arruda, P., , & Azeredo-Espin, A. M. L. 2000, Insect Molecular Biology, 5, 521–529.

Lowe, T. & Eddy, S. 1997, Nucl. Acids Res., 25, 955–964.

Machida, R. J., Miya, M. U., Nishida, M., & Nishida1, S. 2002, Mar. Biotechnol., 4, 406–417.

Mallatt, J., Garey, J., & Shultz, J. 2004, Mol. Phyl. Evol., 31, 178–191.

Masta, S. & Boore, J. 2004, Mol. Biol. Evol., 21, 893–902.

Masta1, S. E. & Boore, J. L. 2004, Molecular Biology and Evolution, 21, 893–902.

McClure, S. 1997, August. "object database vs. object-relational databases". http://www.ca.com/products/jasmine/analyst/idc/14821Eat.htm (Aug. 2005).

Miller, A. D., Nguyen, T. T., Burridge, C. P., & Austin, C. M. 2004, Gene, 331, 65–72.

Moret, B., Wyman, S., Bader, D., Warnow, T., & Yan, M. 2001. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing*, pp. 583. World Scientific Pub.

Morrison, C., Harvey, A., Lavery, S., Tieu, K., & Cunningham, Y. H. C. 2002, Proc. R. Soc. Lond. B, 269, 345–350.

Nardi, F., Spinsanti, G., Boore, J., Carapelli, A., Dallai, R., & Frati, F. 2003, Science, 1887-1889, 1482e.

Nardi, F., Spinsanti, G., Boore, J. L., Carapelli, A., Dallai, R., & Fratil, F. 2003, science, 299, 1887–1889.

Navajas, M., Conte, Y. L., Solignac, M., Cros-Arteil, S., & Cornuet, J.-M. 2002, Mol. Biol. Evol., 19, 2313–2317.

Notredame, C., Higgins, D., & Heringa, J. 2000, Journal of Molecular Biology, 302, 205.

O.Clary, D. & R.Wolstenholm, D. 1983, Nucleic Acids Research, 19, 6859–6872.

Ogoh, K. & Ohmiya, Y. 2004, Gene, 327, 131–139.

Pisani, D. 2004, Syst. Biol., 53, 978–989.

Regier, J. & Shultz, J. 1997, Mol. Biol. Evol., 14, 909–913.

Regier, J., Shultz, J., & Kambic, R. 2005, Proc. R. Soc. B, 272, 395–401.

Richter, S. 2002, Organisms, Diversity and Evolution, 2, 217–237.

Saitou, N. & Nei, M. 1978, Molecular Biolgy and Evolution, 4, 406–425.

Sankoff, D. & Blanchette, M. 1997, Computing and Combinatorics, Proceeedings of COCOON 97..(T. Jiang and D.T. Lee, eds) Springer Verlag, Lecture Notes in Computer Science 1276, 251.

Sankoff, D. & Blanchette, M. 1998, Proceeedings of the Second Annual International Conference on Computational Molecular Biology (RECOMB 98) (S. Istrail, P. Pevzner & M. Waterman eds) New York: ACM Press, 243.

Sankoff, D., Bryant, D., Deneault, M., Lang, B. F., & Burger, G. 2000, Aug, Journal of Computational Biology *7*(3-4), 521–535.

Serb, J. M. & Lydeard, C. 2003, Mol. Biol. Evol. *20*(11), 1854–1866.

Shultz, J. & Regier, J. 2000, Proc. R. Soc. Lond. B, 267, 1011–1019.

Spanos, L., Koutroumbas, G., Kotsyfakis, M., & Louis, C. 2000, Insect Molecular Biology, 2, 139–144.

Spears, T. & Abele, L. 2000, J. Crustacean Biol., 20, 1–24.

Stewart, J. B. & Beckenbach, A. T. 2005, Genome, 48, 46–54.

Stewarta, J. B. & Beckenbachb, A. T. 2003, Molecular Phylogenetics and Evolution, 26, 513–526.

Suyama, Y., Fukuhara, H., & Sor, F. 1985, Curr. Genet., 9, 479–493.

Tillier, E. 1994, Journal of Molecular Evolution, 39, 409–417.

Tillier, E. & Collins, R. 1998, Genetics, 148, 1993–2002.

Umetsu, K., Iwabuchi, N., Yuasa, I., Saitou, N., Clark, P. F., Boxshall, G., Osawa, M., & Igarashi, K. 2002, Electrophoresis, 23, 4080–4084.

Warrior, R. & Gall, J. 1985, Arch Sc. Genever, 38, 439–445.

Wesolowski, M. & Fukuhara, H. Mol. Cell Biol., 1981, 1, 387.

Wheeler, W., Whiting, M., Wheeler, Q., & Carpenter, J. 2001, Cladistics, 17, 113–169.

Wilson, K., Cahill, V., Ballment, E., & Benzie, J. 2000, Mol. Biol. Evol., 17, 863–874.

Yamauchi, M. M., Miya, M. U., & Nishid, M. 2002, Gene, 295, 89–96.

Yamauchi, M. M., Miya, M. U., & Nishid, M. 2004, Insect Molecular Biology, 13, 435–442.

Yang, Z. 1993, Molecular Biology and Evolution, 10, 1396–1401.

Yang, Z. 1994, Journal of Molecular Evolution, 39, 306.

Yang, Z. 2004. *User guide-PAML:Phylogenetic Analysis by Maximum Likelihood Version 3.14.*

Yang, Z. 2005, Jan. *PAML FAQ.*