TWO STUDIES ON MITOCHONDRIAL GENOME EVOLUTION

ORIGIN OF TRNA GENES IN *TRYPANOSOMA* AND *LEISHMANIA* AND COMPARISON OF EUKARYOTE PHYLOGENIES OBTAINED FROM MITOCHONDRIAL RRNA AND PROTEIN SEQUENCES

by

XIAOGUANG YANG, M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2005) McMaster University

(Theoretical Biophysics) Hamilton, Ontario

TITLE: Origin of tRNA genes in *Trypanosoma* and *Leishmania* and comparison of eukaryote phylogenies obtained from mitochondrial rRNA and protein sequences

AUTHOR: Xiaoguang Yang, M.Sc. (Tianjin University, P.R. China)

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: ix, 74

Abstract

Two studies are presented in this thesis. First part is about the origin of tRNA genes in *Trypanosoma* and *Leishmania*. These organisms have special mitochondrial DNA, termed kinetoplast DNA (kDNA), which is unique in its structure and function. kDNA is a massive network which is composed of thousands of connected DNA circles. Unlike most other mitochondrial genomes, there is no gene encoding tRNAs in their kDNAs. So all the tRNAs used in mitochondria must be encoded on nuclear genes and transported from the cytoplasm into the mitochondria. So our question of interest is where the tRNA genes in their nucleus come from. We carry out phylogenetic analysis of these genes and the corresponding ones in bacteria, mitochondria and eukaryotic nuclei. There is no evidence indicating gene transfer from mitochondria to nucleus on the basis of this analysis. These results are consistent with the simplest hypothesis, *i.e.* that all tRNA genes of *Trypanosoma* and *Leishmania* have the same origin as nuclear genes of other eukaryotes.

The second part is about the comparison of eukaryote phylogenies obtained from mitochondrial rRNA and protein sequences. We carried out phylogenetic analysis for the species which have complete mitochondrial genomes by using both concatenated mitochondrial rRNA and protein sequences. We got phylogenies for three groups, fungi/metazoan, plant/algae and stramenopile/alveolate group. The analysis is useful for the further study of position of the genetic code changes and the mechanisms involved.

Acknowledgements

I should first thank my supervisor Professor Paul Higgs for introducing me the interesting research area. I have learned a lot about evolutionary biology and many techniques to do the research from him. Thanks for his patience and clearly written book – Bioinformatics and Molecular Evolution.

I am indebted to my committee, Professor Brian Golding and Ralph Pudritz for their valuable suggestions about my thesis. I also should thank Dr. Supratim Sengupta for detailed and valuable discussion about the phylogeny. Thanks are presented to my Wenwen, my parents and my sister for their love, understanding, support and encouraging. I also thank Wei Xu for discussing not only the research but also many things about daily life.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Three domains

Life is everywhere on earth. You can find living things from backyards at your home to dry deserts, from the edge of the atmosphere to the bottom of the seas, from the freezing waters to the undersea thermal vents. The diversity of life is really amazing, but all the living creatures share some characteristics: they all can replicate; they can express the information in the DNA to produce the products that they need to live; they also can evolve, which means that they can pass on their genetic information to their offspring with some mutations.

Nowadays, life on earth is divided into three domains based on the comparative sequence analysis of 16S RNA: *Bacteria*, *Archaea*, *Eukarya*. Figure 1.1 shows the results from analysis of 16S RNA. This was first proposed by Carl Woese in 1990 (Woese *et al.* 1990). The first two groups were put into *Prokarya* domain before the analysis of 16S RNA, because they both have a relatively simple cell structure that is clearly distinct from the complex cell structure of eukaryotes. However, people found that the *Prokarya* is far more diverse than anyone had expected. *Bacteria* and *Archaea*, although both are prokaryotic cells, are different from each other as either is from *Eukarya* which are made of eukaryotic cells. Each of these three groups shares certain features with the others as well as having unique characteristics of its own: for example, it is found that *Eukarya* has some genes coming from *Bacteria* and others coming from *Archaea*. Informational genes, which

are involved in transcription, translation and other related process, are more closely

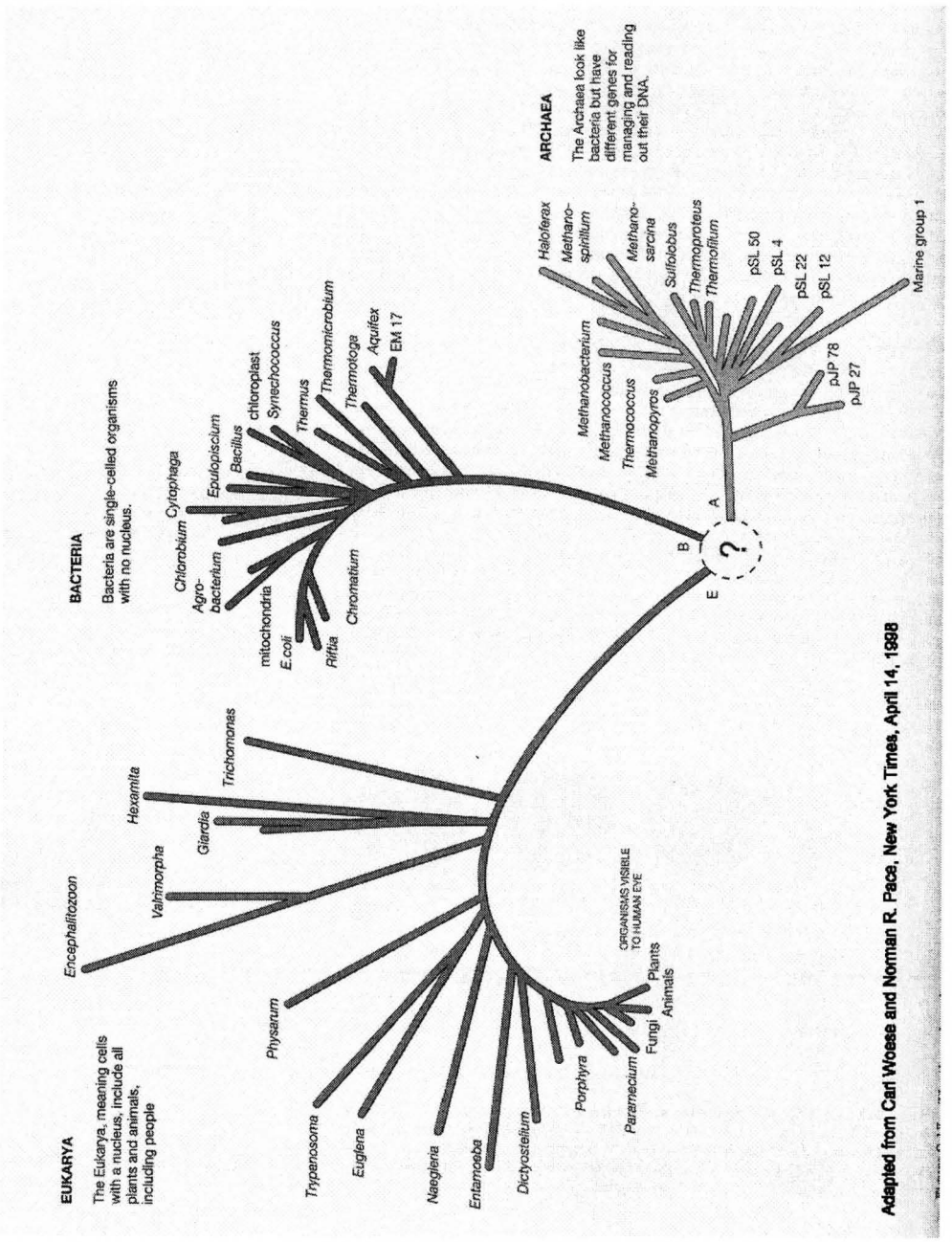

Figure 1.1 Three domains of life

http://cas.bellarmine.edu/tietjen/Ecology/Domains0.gif

related to archaeal genes; operational genes, which are involved in cellular metabolic process, are more closely related to bacterial genes (Rivera *et al.* 1998). However, *Eukarya* is much more complicated than *Bacteria* and *Archaea*. Although *Bacteria* and *Archaea* look like each other, they have different genes for managing and reading out their genetic information.

1.2 Similarities and Differences between prokaryotic and eukaryotic cells

All living organisms are made of cells which are small membrane-bounded compartments filled with solution of chemicals. Since the first observation of cell made by Robert Hooke, research led to the formation of the cell theory, which would change the basic biological theory and research forever.

All cells have three basic features. Firstly, they possess genetic material (DNA and RNA), which carry the instructions to produce proteins. Secondly, they possess a cytoplasm containing cytosol and organelles. Cytosol is a fluid, consisting mostly of water and dissolved nutrients, wastes, ions, proteins, and other molecules. Organelles are small structures suspended in the cytosol. The organelles carry out the basic functions of the cell, including reproduction, metabolism and protein synthesis. Examples of organelles are mitochondrion, chloroplast, golgi apparatus and endoplasmic reticulum, etc. Thirdly, they possess a plasma membrane consisting of a phospholipid bilayer that houses the cell. This membrane contains several structures that allow the cell to perform necessary tasks; for example, channels that allow substances to move in and out of the cell, antigens that allow the

cell to be recognized by other cells, and proteins that allow cells to attach to each other.

Apart from these three similarities, cells have diverse structures and forms and are, therefore, difficult to generalize. There are major differences in cell structure and function between different types of organisms. There are even major differences in cells within the same organism, reflecting the different functions of the cells within the organism.

One major difference among cells is the presence or absence of a nucleus, which is a sub-cellular structure that contains the genetic material. Prokaryotic cells (which include *Bacteria* and *Archaea*) lack a nucleus, while eukaryotic cells (which include single celled organisms like protists and animal and plant cells) contain a nucleus. This is in fact the characteristic used to define a eukaryote.

Figure1.2 gives us the general idea what the prokaryotic cell and eukaryotic cell look like and how different they are. In general, eukaryotic cells are much bigger and more complex than prokaryotic cells. Most bacteria are 1-10 μm in diameter, and eukaryotic cells are usually 10-100 μm in diameter, ten times bigger than bacteria. There are a variety of membrane-bounded organelles of specialized structure and function within the cytosol of a eukaryotic cell. These are absent in prokaryotic cells.

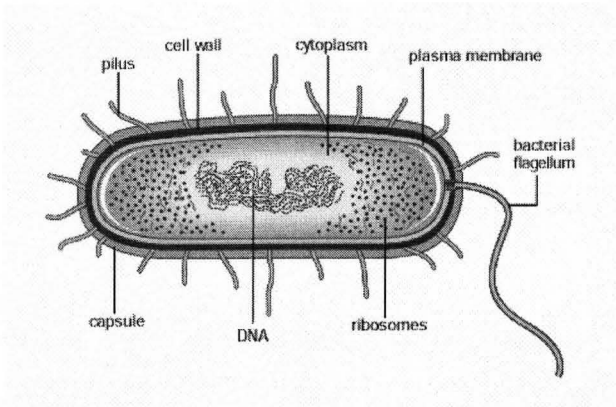

Figure 1.2 (a) Prokaryotic cell

http://www.bact.wisc.edu/Bact303/Bacterium.jpeg

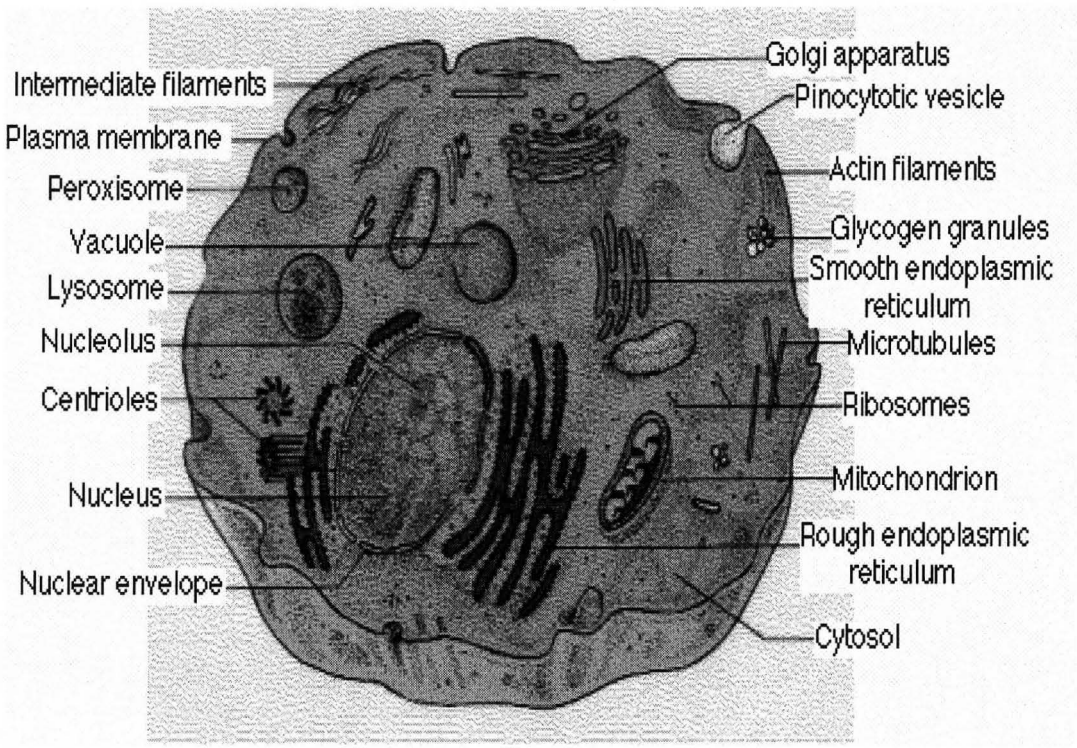

Figure 1.2 (b) Eukaryotic cell

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/A/AnimalCells.html

1.3 Mitochondria

1.3.1 Mitochondrial structure and function

Now let us focus on one kind of very important organelle, the mitochondrion. A mitochondrion is an organelle found in most eukaryotic cells, including those of animals, plants, fungi and protists. Mitochondria are about 1-10 μ m long. And a cell contains hundreds of them. Figure 1.3 shows what a mitochondrion looks like by using cartoon and transmission electron micrograph.

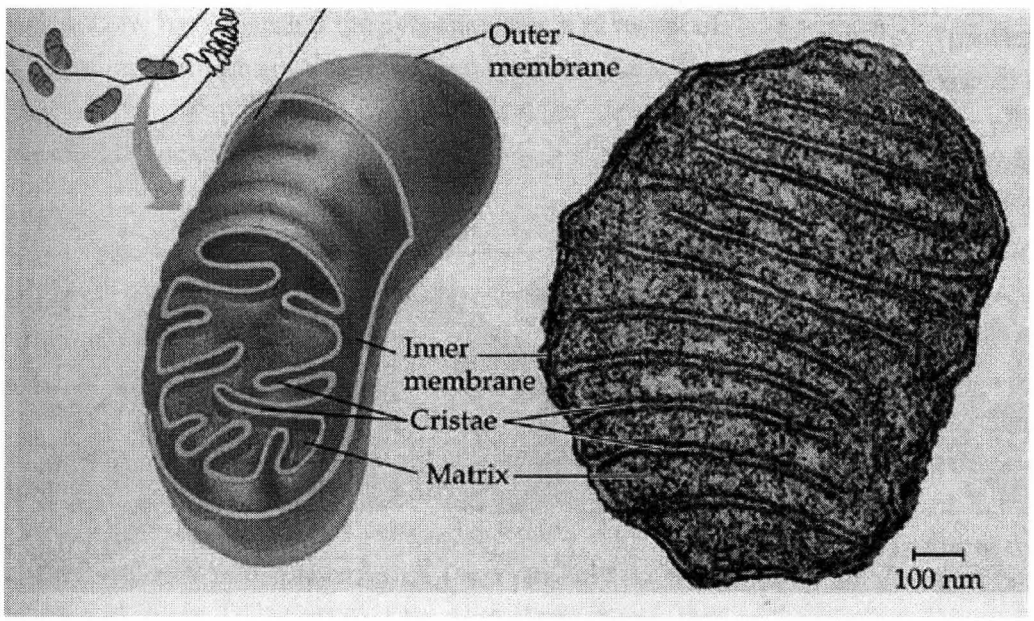

Figure 1.3 Cartoon and Transmission Electron Micrograph of a mitochondrion http://www.ualr.edu/~botany/mitochondrion.jpg

In living cells, mitochondria can move around, change their shapes and divide, unlike the static cylinders seen in electron micrograph. A mitochondrion consists of

two highly specialized membranes that play crucial roles in activities. Each membrane is a phospholipid bilayer with unique proteins embedded. The outer membrane is smooth, but the inner membrane is highly convoluted, forming infoldings called cristae. The cristae greatly increase the surface area of inner membrane. The mitochondrion is divided into two internal compartments by inner membrane. The first is the narrow region between the outer and inner membranes, called intermembrane space. The second is the mitochondrial matrix which is enclosed by inner membrane. There are mitochondrial DNAs, ribosomes and a variety of enzymes in the mitochondrial matrix. Some steps of cellular respiration are catalyzed by the enzymes in the mitochondrial matrix. Some other important steps are carried out on the cristae. It is on the cristae that sugar is combined with oxygen to produce adenosine triphosphate ATP - the primary energy source for the cell. So mitochondria are the power plant of cells.

1.3.2 Origin of mitochondria

It has been widely accepted that mitochondrion probably evolved from endosymbiotic bacteria like *Rickettsia*. The endosymbiotic hypothesis (Butow *et al.* 1988; Cavalier-Smith 1987; Gellissen and Michaelis 1987; Schwartz and Dayhoff 1978; Whatley *et al.* 1979), shown in Figure 1.4, postulates that these prokaryotic endosymbionts lived inside the ancestral eukaryotic host cells instead of being killed after they were swallowed by host cells and eventually the host and the endosymbionts developed a mutually beneficial relationship.

There are many evidences supporting the theory. For example, there are many endosymbiotic relationships in the modern world: one kind of proteobacterium called *Buchnera* can be found in the body of the pea aphid *Acyrthosiphon pisum*(Tsuchida *et al.* 2005); a rickettsia-like microorganism is found in *Culex tigripes* (Ndiaye and Mattei 1993); and other endosymbiotic virus, bacteria, archaea, algae are also found (Fernandes *et al* 1964; Hackstein and Vogels 1997; Dunn *et al.* 2004; Hoshina *et al.* 2004). Interestingly, most of the gene products in mitochondria are encoded in nucleus and transported into the mitochondria (Attardi and Schatz 1988). This fact can also support the theory.

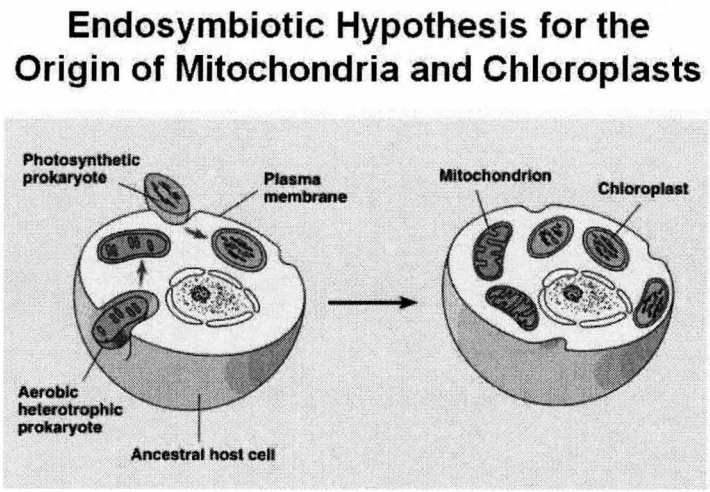

Figure 1.4 Endosymbiotic hypothesis

http://io.uwinnipeg.ca/~simmons/Chap2898/img003.jpg

Another line of evidence is the similarity between the mitochondria and bacteria. The mitochondrial genome contains a single circular DNA molecule, which is different from the linear chromosomes found in the cell nucleus, but similar to the

genomes of bacteria. Several enzymes and transport system on the inner membrane of mitochondria resemble those on the plasma memebrane of modern prokayotes (Gray and Spencer 1996). Mitochondrial ribosomes are more similar to prokaryotic ribosomes than to those in the eukaryotic cytoplasm (Olsen *et al.* 1994). There are many other similarities in terms of size and biochemical characteristics. Indeed, the closest relatives of mitochondria seem to be the *Rickettsia* (Olsen *et al.* 1994; Viale and Arakaki 1994; Gray and Spencer 1996; Sicheritz-PonteÂn *et al.* 1998). Mitohcondria and *Rickettsia* have circular DNA, similar GC content and gene order. They also share the similar reduced gene content: no genes for anaerobic glycolysis are found in either of them; the ATP production pathway is similar (Renesto *et al.* 2005); many genes involved in the biosynthesis and regulation of biosynthesis are absent from them (Andersson 1998).

1.3.3 Genomes of mitochondria

After the endocytic event, an extensive transfer of genes from mitochondrial genomes to nuclear DNA must have occurred because most of the genes encoding present-day mitochondrial proteins are in the cell nucleus (Attardi and Schatz 1988). Mitochondrial genomes of multicellular animals are mostly small, circular molecules in which 13 protein genes, two ribosomal RNA genes and 22 transfer RNA genes are closely packed (Wolstenholme 1992). Their typical size is about 16500 base-pair. Figure 1.5 shows the human mitochondrial genome.

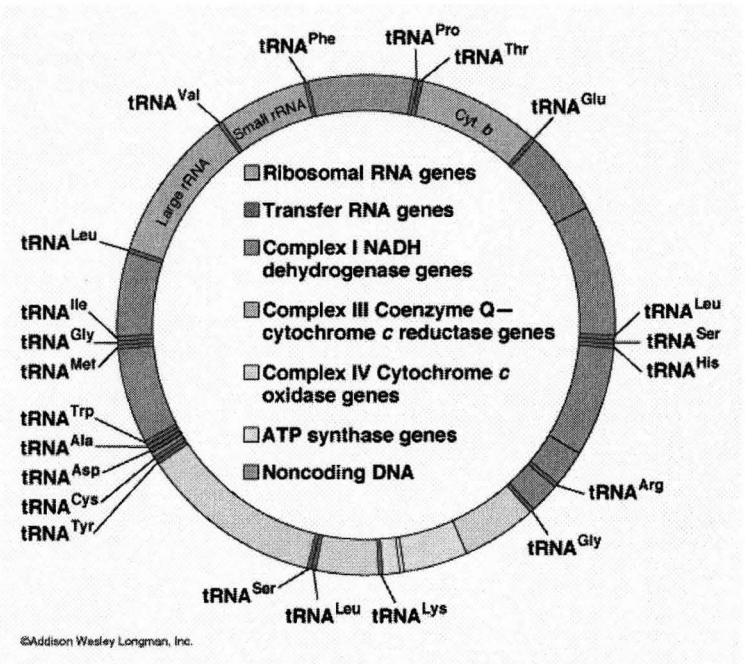

Figure 1.5 Human mitochondrial genome

http://www.mun.ca/biology/desmid/brian/BIOL2060/CellBiol16/1624.JPG

Compared to nuclear genomes, the animal mitochondrial genomes have almost no space for non-coding sequences, *i.e.*, nearly every nucleotide is part of a coding sequence. They only require 22 tRNAs for protein synthesis, compared to 30 or more in the cytosol and chloroplasts. The 'standard set' of mtDNA-encoded genes is also found in fungal mtDNA (Gray 1998). Conversely, in plants the trend has been in the opposite direction with the mtDNA tending to increase in size, 10 to 150 times larger, although the mitochondrial genomes in plants encode only a few more proteins than those of animals do. Of all the completed sequenced genomes up to date, *Zea mays* has the largest size mitochondria 569630; *Theileria parva* has the smallest size 5895; *Reclinomonas Americana* has the largest number of 63 protein genes; *Plasmodium yoelii* has the smallest number of 2 protein genes. In addition to the size and gene

content, the structure of mitochondrial DNA also varies in different species. Most of known species have circular DNA, some of them have linear DNA; the mitochondrial DNA of trypanosomatid protozoa is a massive network made of thousands of connected DNA circles which has no tRNA genes. In trypanosomatid protozoa, the tRNA genes required for protein synthesis are encoded in nucleus. The details about the origin of the tRNA genes will be discussed in the Chapter 3.

1.4 Reassignment of genetic code in mitochondria

The genes on the mitochondrial genome are transcribed and translated independently of those in the nucleus. Most mitochondria contain ribosomes with distinct rRNAs from the nuclear rRNAs and contain tRNAs that are distinct from the nuclear tRNAs. One of the most surprising features about the mitochondrial translation system is that in many species the mitochondrial genetic code is different from the universal code. The 'standard' or 'canonical' genetic code (Table 1.1) was once thought to be a 'frozen accident'. Any changes in genetic code will change the meaning of translated messages.

But people have observed many cases of codon reassignment in nuclear and mitochondrial genetic code, which differ from the 'standard' genetic code (Reviewed by Knight *et al* 2001). For example, in the mitochondria of some species UGA is translated into Trp instead of Stop; in other cases, AUA is translated into Met instead of Ile, etc. In order to explain the underlying mechanisms of these reassignments, several theories were proposed.

First Position (5' end)	Second Position				Third Position (3' end)
U	U	C	A	G	U
	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met Start	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Table 1.1 The 'standard' genetic code

Osawa and Jukes (1989) have proposed the codon capture theory. According to this hypothesis, first, the codon must disappear from the genome totally to make the mutation of tRNA neutral; then the tRNA with the corresponding anti-codon also disappear, the codon becomes unassigned; finally when the codon reappeared, it was captured by another tRNA with different specificity caused by mutation. The ambiguous intermediate theory (Schultz and Yarus 1996), unlike codon capture theory, allows intermediate stage during the reassignment process. Mutations in a tRNA can change its decoding efficiency or specificity so that one codon can have

more than one meaning. If the new meaning is advantageous under some circumstances, selection will take place the old meaning with the new one. Finally, selection leaves the new meaning unambiguous. Sengputa and Higgs (2005) proposed a unified model for the codon reassignments. This model incorporates two mechanisms mentioned above and another two possible mechanisms by changing parameter values. These parameters include the fitness factor, mutation rate and the number of the codons to be reassigned in the genome.

Genome streamlining is another possible force to drive the reassignment of the genetic code. Selection for minimization of genome length may cause the genetic code to simplify the repertoire of tRNAs and modifying enzymes. This suggests that the resident genomes, such as those of organelles, endosymbionts and obligate intracellular parasites, are more likely to be under the pressure of that force because the resident genomes intend to transfer genes to nuclear genomes of the host cells.

In chapter 4, we will present a phylogenetic analysis for the species which have complete mitochondrial genomes by using both concatenated mitochondrial rRNA and protein sequences. The obtained phylogenies will be used to determine the positions of the genetic code changes and the relevant gene losses to understand the mechanisms involved.

Chapter 2 Methods

2.1 Software packages

We use several software packages to deal with the data and do the phylogeny analysis. Perl, which stands for "Practical Extraction and Report Language", is a popular and powerful programming language that is extensively used in areas such as bioinformatics and web programming. Perl was originally invented based on UNIX system for dealing with text and table files. Now it has become popular with biologists because it is so well suited to many bioinformatics tasks. And it is available free from http://www.perl.org/ and can run on all the operating systems found in average labs.

ClustalX is a windows interface for the ClustalW multiple sequence alignment program (Thompson *et al.* 1997). It provides an integrated environment for performing multiple sequence and profile alignments and analyzing the results. The sequence alignment is displayed in a window on the screen. The pull-down menus at the top of the window allow you to select all the options required for traditional multiple sequence and profile alignment. Figure 2.1 shows that the ClustalX is in the Profile alignment mode. The latest version of the ClustalX program can be obtained by anonymous ftp to: ftp-igbmc.u-strasbg.fr or at the following WWW site: http://www-igbmc.u-strasbg.fr/BioInfo/.

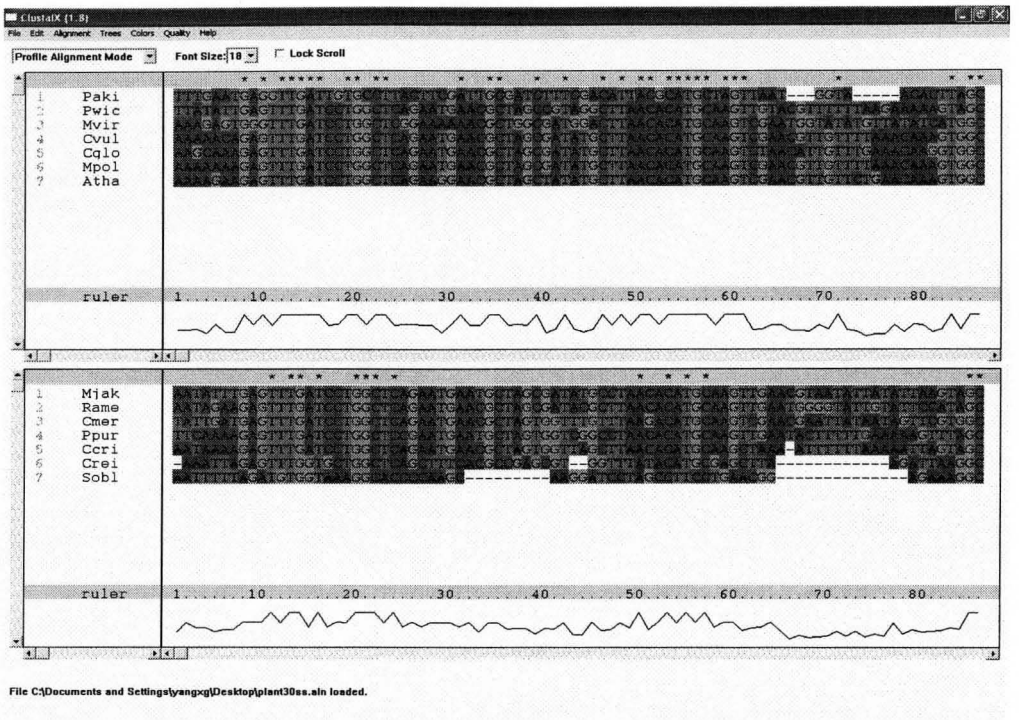

Figure 2.1 Interface of ClustalX

T-Coffee is another heuristic method for multiple alignments (Notredame *et al.* 2000). It can use the information from all the sequences in the alignment to calculate the parameters and to improve the alignments of sequences that were aligned at earlier stage.

GeneDoc provides a combination of alignment editing and alignment analysis capabilities intended to help users refine their alignments (Nicholas 1997). It can import and export files with many different formats. Users can edit the sequences with powerful functions implemented in it. It is extremely useful for the analysis of biological sequences and now has become the common tool in bioinformatics. It is also free and can be downloaded from http://www.psc.edu/biomed/genedoc. Figure 2.2 shows what the interface of GeneDoc looks like on Windows system.

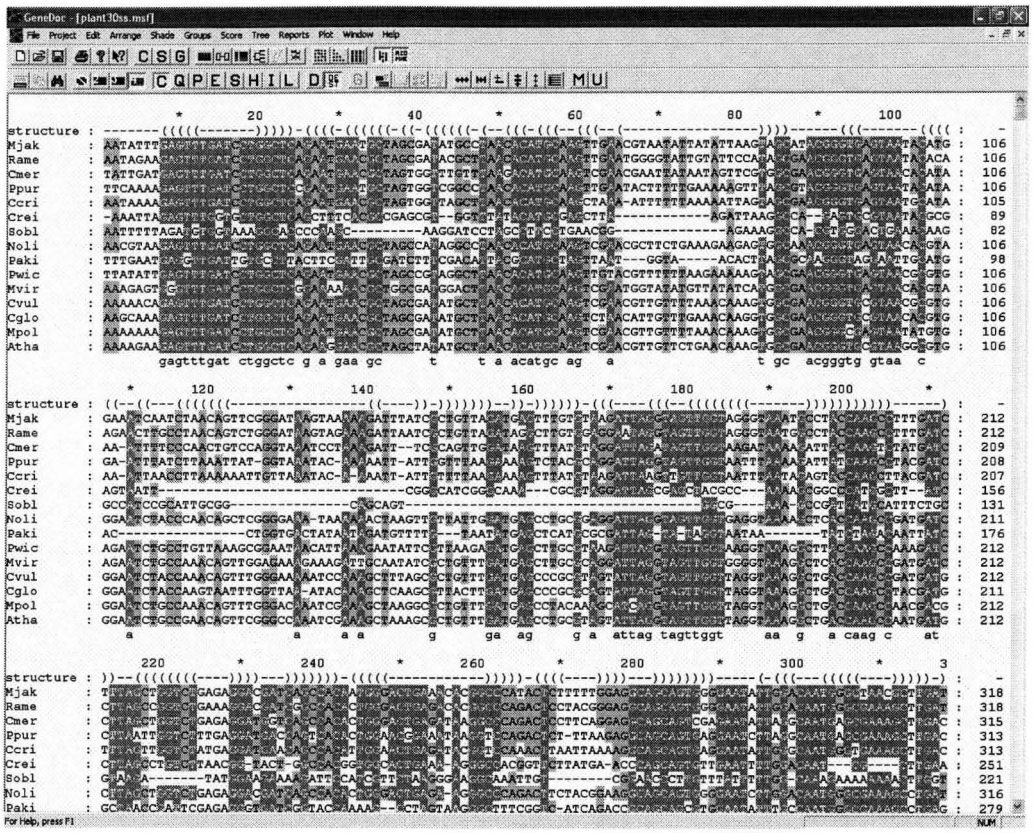

Figure 2.2 Interface of GeneDoc (GeneDoc loaded with alignment of 16S rRNA with secondary structure on the top of the sequences)

Tree-Puzzle is another package for phylogeny, which is good at maximum likelihood analysis for nucleotides and amino acids (Schmidt *et al.* 2002). It is available at http://www.nsc.liu.se/software/biology/

Phylip is one of the earliest, most famous and widely distributed packages for phylogenetics study (Felsenstein 1989). It includes programs to carry out parsimony, distance matrix methods, maximum likelihood, and other methods on a variety of types of data. It is available free from the following website http://evolution.genetics.washington.edu/phylip/software.html.

PHASE (Jow *et al.* 2002) is designed specially for the analysis of RNA sequences which have a conserved secondary structure, although it can also handle the DNA and amino acid sequences. Substitution models for pairs of sites are implemented in the package along with the standard models used for nucleotide and amino acid. Users are supposed to provide the conserved secondary structure of the RNA sequences. PHASE uses MCMC method to generate a large number of phylogenetic trees with posterior probability proportional to their likelihood. It can be downloaded from http://www.bioinf.man.ac.uk/resources/phase.

After we get the phylogenies, we use Treeview (Page 1996) to look at, edit and print the trees to do some analysis. It is available from http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

In this study, Perl, ClustalX, GeneDoc and Treeview are used on Windows system; T-Coffee, Tree-Puzzle, Phylip and PHASE are for Linux system.

2.2 Secondary structure: ribosomal RNA of Escherichia coli

The secondary structures used in this study are from the ribosomal RNA of *Escherichia coli*, which are shown in Figure 2.3, 2.4. We use them for improving the rRNA sequence alignments and doing phylogenies in the Chapter 4.

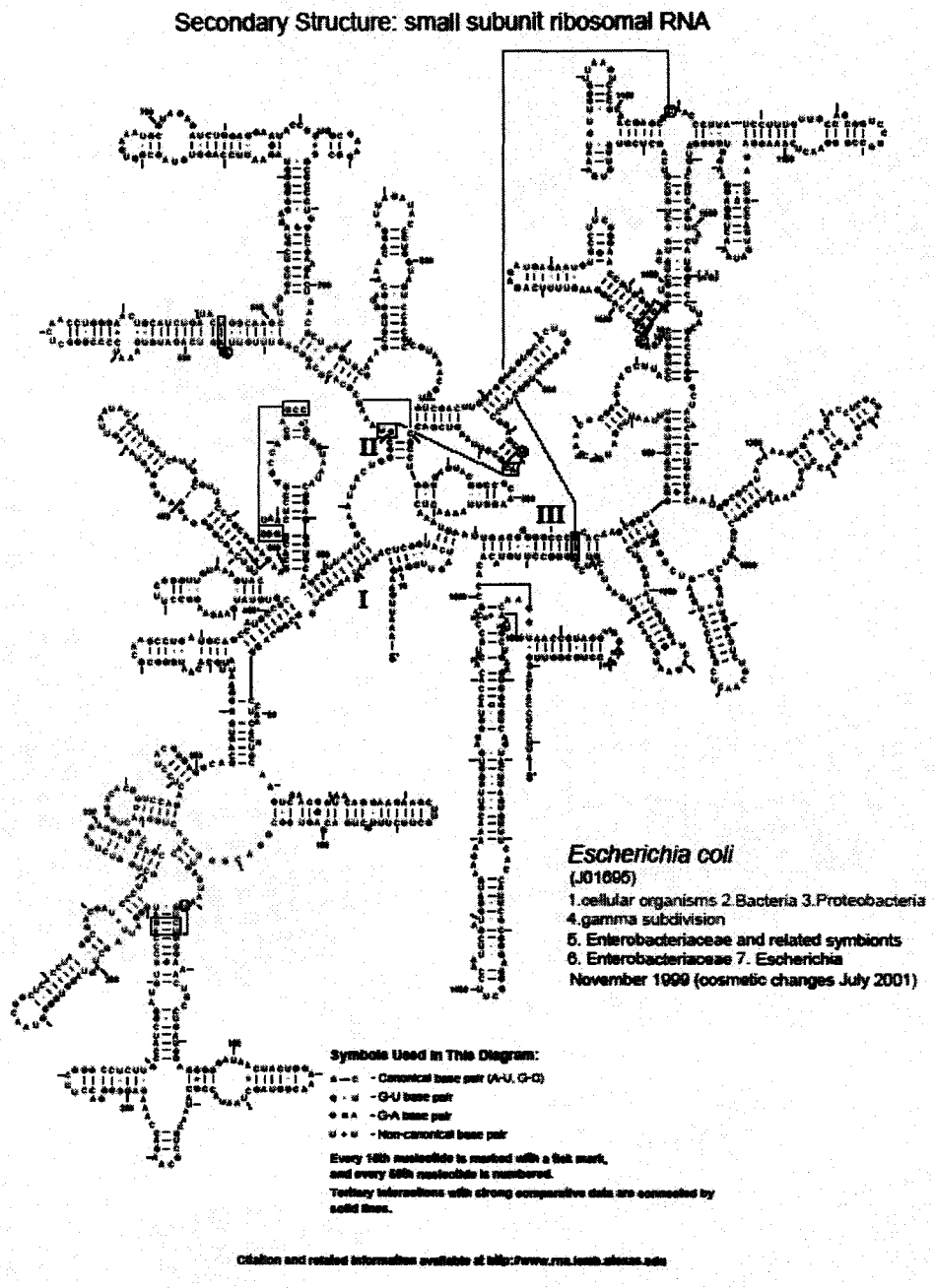

Figure 2.3 Secondary structure: small subunit ribosomal RNA of *Escherichia coli* (Cannone *et al* 2002)

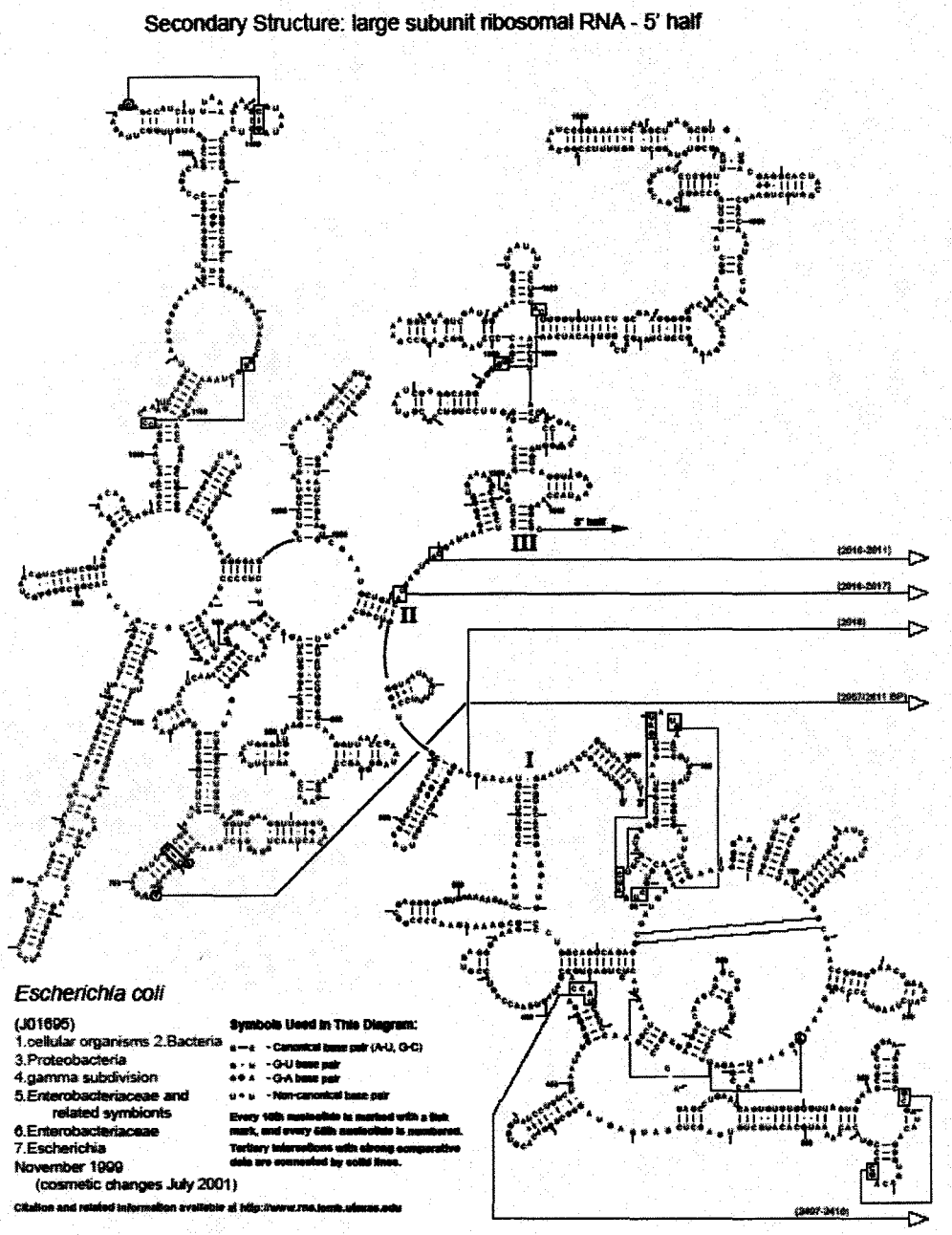

Figure 2.4(a) Secondary structure: large subunit ribosomal RNA of *Escherichia coli*-5' half (Cannone *et al* 2002)

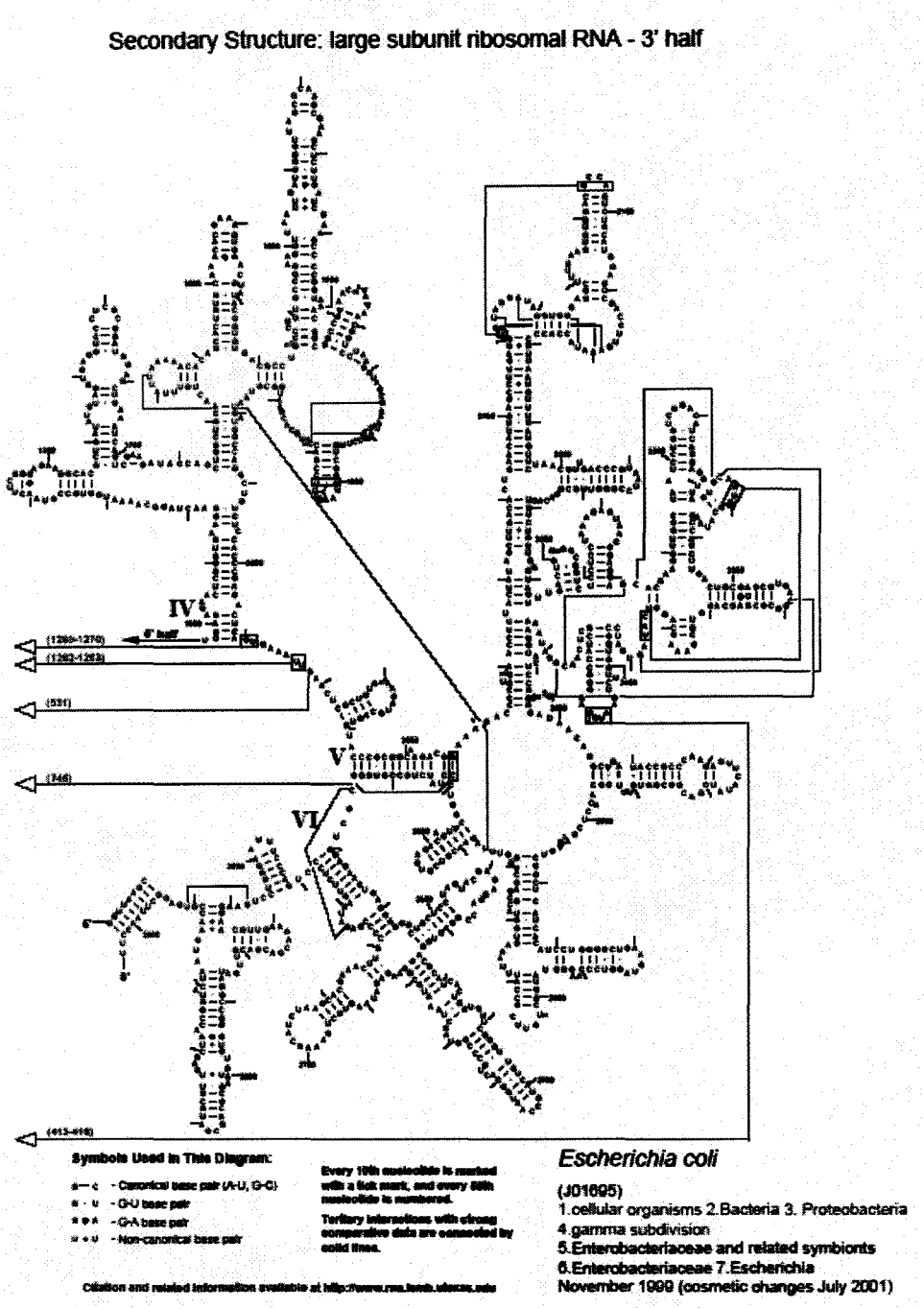

Figure 2.4(b) Secondary structure: large subunit ribosomal RNA of *Escherichia coli*-3' half (Cannone *et al* 2002)

2.3 Several nucleotide substitution rate models in PHASE

The nucleotide substitution rate models are defined by rate matrix. The off-diagonal elements $r_{ij} = \pi_i \alpha_{ij}$ define the rate of substitution from state i (row) to state j (column). In nucleotide substitution models, π_i will be π_A, π_C, π_G and π_T, which are the four equilibrium bases frequencies. In the paired-site substitution model, π_i will be the equilibrium frequencies of pairs that can be formed with four bases. α_{ij} is the rate ratio which is defined to satisfy the time-reversibility constraint, *i.e.*, $\pi_i r_{ij} = \pi_j r_{ji}$. The diagonal elements are equal to minus the sum of the other elements on the same row. The A↔G transition rate ratio is used as reference by PHASE, *i.e.*, $\alpha_{AG} = \alpha_{GA} = 1$. The AU↔GC transition rate ratio is also used as reference by PHASE. Here are three models which are used in this study:

HKY85 model (Hasegawa *et al* 1985)

$$\begin{pmatrix} & A & C & G & T \\ A & * & \pi_c\alpha_1 & \pi_G & \pi_T\alpha_1 \\ C & \pi_A\alpha_1 & * & \pi_G\alpha_1 & \pi_T \\ G & \pi_A & \pi_c\alpha_1 & * & \pi_T\alpha_1 \\ T & \pi_A\alpha_1 & \pi_c & \pi_G\alpha_1 & * \end{pmatrix}$$

The HKY85 model assumes different base frequencies and account for the difference between transition and transversions with one parameter. $\alpha_{transition}$=1.0, $\alpha_{transversion} = \alpha_1$.

REV model (Yang 1994)

$$\begin{pmatrix} & A & C & G & T \\ A & * & \pi_c\alpha_1 & \pi_G & \pi_T\alpha_2 \\ C & \pi_A\alpha_1 & * & \pi_G\alpha_3 & \pi_T\alpha_4 \\ G & \pi_A & \pi_c\alpha_3 & * & \pi_T\alpha_5 \\ T & \pi_A\alpha_2 & \pi_c\alpha_4 & \pi_G\alpha_5 & * \end{pmatrix}$$

The REV model is the most general model for nucleotide substitution subject to the time-reversibility constraint. It has four frequencies and five rate ratio parameters.

RNA7A model (Higgs 2000)

$$\begin{pmatrix} & AU & GU & GC & UA & UG & CG & MM \\ AU & * & \pi_{GU}\alpha_1 & \pi_{GC} & \pi_{UA}\alpha_2 & \pi_{UG}\alpha_3 & \pi_{CG}\alpha_4 & \pi_{MM}\alpha_5 \\ GU & \pi_{AU}\alpha_1 & * & \pi_{GC}\alpha_6 & \pi_{UA}\alpha_7 & \pi_{UG}\alpha_8 & \pi_{CG}\alpha_9 & \pi_{MM}\alpha_{10} \\ GC & \pi_{AU} & \pi_{GU}\alpha_6 & * & \pi_{UA}\alpha_{11} & \pi_{UG}\alpha_{12} & \pi_{CG}\alpha_{13} & \pi_{MM}\alpha_{14} \\ UA & \pi_{AU}\alpha_2 & \pi_{GU}\alpha_7 & \pi_{GC}\alpha_{11} & * & \pi_{UG}\alpha_{15} & \pi_{CG}\alpha_{16} & \pi_{MM}\alpha_{17} \\ UG & \pi_{AU}\alpha_3 & \pi_{GU}\alpha_8 & \pi_{GC}\alpha_{12} & \pi_{UA}\alpha_{15} & * & \pi_{CG}\alpha_{18} & \pi_{MM}\alpha_{19} \\ CG & \pi_{AU}\alpha_4 & \pi_{GU}\alpha_9 & \pi_{GC}\alpha_{13} & \pi_{UA}\alpha_{16} & \pi_{UG}\alpha_{18} & * & \pi_{MM}\alpha_{20} \\ MM & \pi_{AU}\alpha_5 & \pi_{GU}\alpha_{10} & \pi_{GC}\alpha_{14} & \pi_{UA}\alpha_{17} & \pi_{UG}\alpha_{19} & \pi_{CG}\alpha_{20} & * \end{pmatrix}$$

Specific models have been developed to deal with paired-sites in RNA helices that evolve via compensatory mutations (Tillier 1994; Tillier and Collins 1998; Schoniger and von Haeseler 1994; Higgs 2000). The RNA7A model is the most general of the seven state paired-site substitution models. It has 21 rate ratio parameters (including the reference rate ratio AU↔GC) and 7 frequencies. All mismatches are treated in a single state MM.

2.4 Algorithm

2.4.1 Maximum likelihood criterion

Given an evolutionary model and a tree topology, we can calculate the likelihood that the data evolves on this tree. By calculating each node in this tree, we can obtain the likelihood for a single given site. Then we sum over the log-likelihood of each single site to obtain the log-likelihood of the data evolving on the tree. Maximum likelihood criterion is choosing a tree with maximum likelihood (Felsenstein 1981). The different parameters, such as the tree topology, the branch length and the parameters in the model, can be optimized all together. It is also possible to optimize some parameters while others are fixed.

2.4.2 MCMC (Markov Chain Monte Carlo)

In many cases of phylogenetic analysis, there are many alternative trees that are only slightly worse than the best tree according to the maximum likelihood criterion. This indicates that the single best tree may not be correct in all aspects and we should use a method that can deal with the ensemble of possible trees, which contains more information than the single best tree.

The MCMC (Markov Chain Monte Carlo) method is introduced to generate a large set of sample trees which have the property that the probabilities of their occurring in the sample are proportional to the likelihood of the data, given the tree (which we know how to calculate), multiplied by the prior of the tree (using any expectations about the problem before knowing the information of the data). It begins

from a trial tree with certain likelihood, L_1. Then a move in the tree space is made by changing the length of branches or changing the rate parameters or changing the topology of the trial tree. So we get a new tree with likelihood, L_2 which will be slightly lower or higher than L_1. If $L_2 > L_1$, the new tree is accepted and becomes the next tree in the sample; if $L_2 < L_1$, the new tree is accepted with probability L_2 / L_1 and rejected otherwise. If it is rejected, the next tree will still be the old one. This algorithm always allows the moves that increase the likelihood, but sometimes it also allows the moves that decrease the likelihood.

The mcmcphase program in the PHASE package uses Markov Chain Monte Carlo techniques to produce large samples of trees from a given alignment. The mcmcsummarize program in the PHASE package can exploit the large samples to produce the consensus tree.

2.4.3 The Neighbor-Joining method

Neighbor-Joining (Saitou and Nei 1987) is a quick and practical method to produce trees from distance matrices. This method begins with a set of disconnected tip nodes which represent the known sequences. The two nodes with the smallest modified distance (Saitou and Nei 1987) are connected by a new internal node. Then this new internal node replaces the original two nodes in the problem. This procedure will be repeated until all the nodes are connected to one tree.

2.4.4 Bootstrapping

To assess the reliability of phylogenies under some chance fluctuations, bootstrapping method was introduced by Felsenstein (1985). It has since become a standard tool in phylogenetics. Bootstrapping method deliberately introduces slight difference into the dataset by constructing many new sequence alignments by randomly selecting columns from the original alignment. New alignments have the same length as the original one. Bootstrapping is different from shuffling the original alignment. Shuffling will not change the phylogeny because it only changes the order of these columns in the alignment. During bootstrapping, some columns in the original alignment will be used twice or more to construct randomized alignments, while some others will not be used. That means the new random sequences contain information slightly different from the original alignment. When we construct trees with the randomized sequences, we are not guaranteed to get the same phylogeny as before. If the signal in the data is strong, bootstrapping should make very little difference. However, if the signal is rather weak, the noise we introduce to the data may cause the tree-construction method to give the different results. To carry out bootstrapping analysis, many randomized datasets are constructed. The tree-construction method is then used on each dataset repeatedly to produce many trees. Some of the trees will be equivalent to the original tree and others will be different. We then look at each group of the species in the original tree and determine the percentage of the randomized trees containing the same group. This percentage tells us how it is reliable that those species really form a related group.

2.5 Programs developed by myself

To help with the sequence analysis project, I developed the following programs in Perl.

robot.pl: this program can download many separated sequence files. Robot, also called web worm, can search the links with same format and download the corresponding contents automatically. It is very convenient for downloading the separated information when the number of files is large.

name.pl: this program can classify and name the sequences from the download files automatically. Only after given names, the sequences can be used by all kinds of packages to do phylogenies.

secondarystructure.pl: after aligning sequences, we need to add the secondary structure on the top of the alignments if we want to use the information in it to infer phylogenies. This program can do this automatically.

deletemismatch.pl: this program can delete brackets which denote the mismatched pairs in the alignment of RNA sequences and delete the columns that have more gaps than some threshold value.

secondarybootstrap.pl: this program can produce many datasets of randomized sequences with secondary structure for the purpose of bootstrapping.

distphase.sh: this shell program can repeatedly call distphase program in the PHASE package to calculate the ML (Maximum Likelihood) distance matrix for each dataset of randomized sequences. This program will be used for bootstrapping.

Chapter 3

Origin of tRNA genes in *Trypanosoma* and *Leishmania*

The trypanosomatid protozoa are widespread in nature, and they are of interest and importance because they cause disease in humans, livestock, and commercially important plants (Barrett *et al.* 2003). This group includes *Trypanosoma brucei, Leishmania major* and *Leishmania infantum* which are included in our study and shown in Figure 3.1. Complete genomes of these species have recently become available (El-Sayed *et al.* 2005).

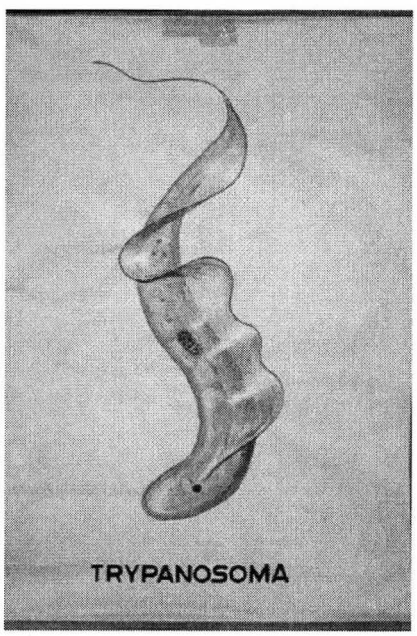

Figure 3.1(a) *Trypanosoma*

http://www.uta.edu/chagas/images/trypomast.jpg

The genus *Trypanosoma* is large and diverse. It includes several species that infect wild and domesticated animals and humans. Most of them are transmitted by insect vectors. *Trypanosoma brucei* can cause African sleeping sickness, which can be fatal, and is transmitted by tsetse fly.

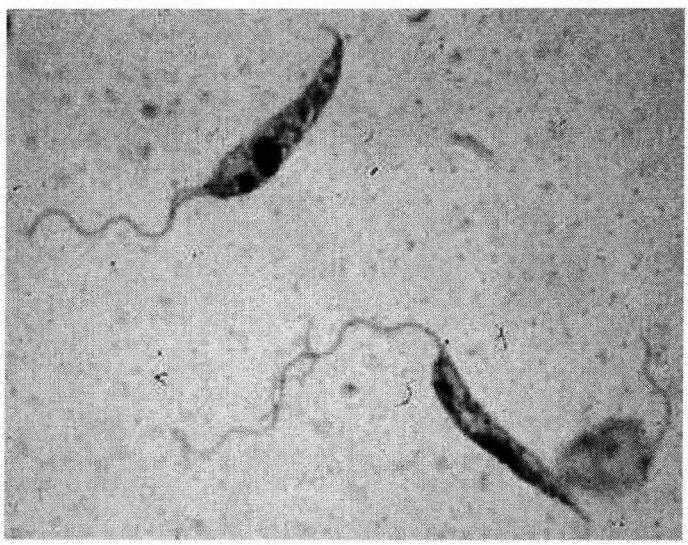

Figure 3.1 (b) *Leishmania*

http://dspace.dial.pipex.com/mark.bailey/LDC.jpg

Members of the genus *Leishmania* can infect many mammals, including humans, dogs and rodents. Their vector is sand fly. They can cause the horrible disease, Leishmaniasis, which has spread to about 100 countries around the world. The disease sometimes causes disfiguring lesions, and can also be fatal.

3.1 kDNA

The mitochondria of these ancient organisms, termed kinetoplast, have special DNA (kDNA), which is unique in its structure and function. kDNA is a massive

network which is composed of thousands of connected DNA circles. There are two types of kDNA circles, maxicircles and minicircles. The maxicircles encode ribosomal RNAs and a few mitochondrial proteins, which have similar structure and function with the mitochondrial DNA of other eukaryotes. But they do not encode tRNAs. The minicircles encode small guide RNAs that direct the specificity of RNA editing. Figure 3.2 shows us what a kDNA network looks like. More details of structure, function, replication and evolution of kDNA are presented in earlier reviews (Lukes *et al.* 2002; Morris *et al.* 2001;Shapiro and Englund 1995; Ray 1987; Stuart 1983).

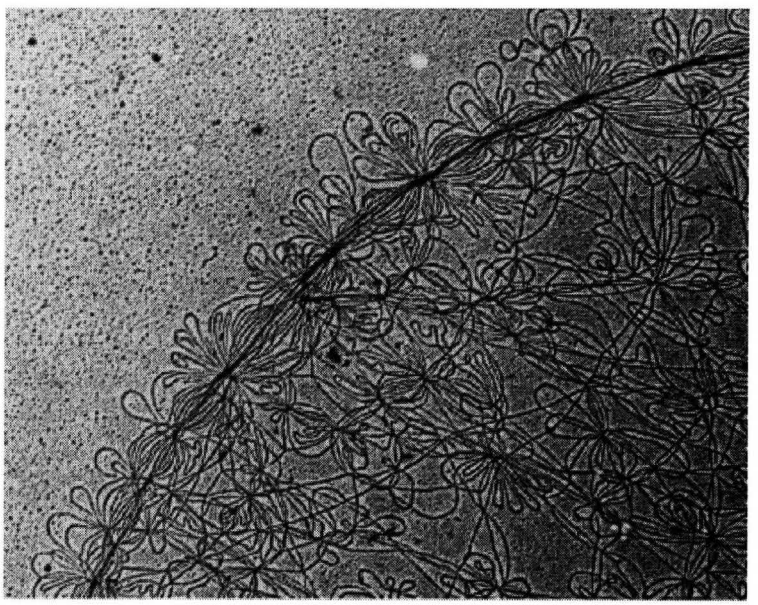

Figure 3.2 A part of a purified kDNA network from *C.fasciculata*, which also belongs to the family *Trypanosomatidae* and has similar network of *Trypanosoma* and *Leishmania*, shown by electron microscopy (EM). Small loops are the 2.5 kb minicircles, and long strands are parts of the 38 kb maxicircles (Morris *et al.* 2001).

3.2 Question of interest

The nuclear and mitochondrial genome sequences of these species, *Trypanosoma brucei, Leishmania major* and *Leishmania infantum*, are complete and therefore we have a full set of nuclear tRNA sequences available. Unlike most other mitochondrial genomes, there is no gene encoding tRNAs in their kDNAs. So all the tRNAs used in kinetoplasts must be transported from cytoplasm, and these are encoded by nuclear genes (Lukes *et al.* 2002; Morris *et al.* 2001; Shapiro and Englund 1995).

So our question of interest is where the tRNA genes in their nucleus come from. There are many examples of gene transfer from mitochondria to the nucleus in other species (Poyton and McEwen 1996; Boore 1999; Palmer 2000). Also many proteins used in mitochondria are made from genes on the nuclear genome and then transported to the mitochondrion. Therefore it is possible that the tRNAs used in the kinetoplast could have originated from kinetoplast and transferred from kinetoplast to nucleus. Then, after the transfer, the original genes were deleted in the kinetoplast. Alternatively, the tRNAs originally in the kinetoplast may have been deleted and their function was taken over by genes from the nucleus. We wish to understand the evolutionary relationship between tRNA genes of these three species and the ones of the other protists, proteobacteria and even the eukaryotic non-protists. We will ask whether any of the genes in the current nuclear genome of the trypanosomes show evidence of being related to mitochondrial genes in other species.

3.3 Dataset

We have the nuclear genes of all tRNAs of *Trypanosoma brucei*, *Leishmania major* and *Leishmania infantum* which are included in our study. Figure 1.1 shows where they are in the evolutionary tree based on the analysis of 16S rRNA.

We already have the alignments of α-proteobacterial genes (about 40 species) (Tang *et al.*, Private communication). We selected the species of interest including many protists and several eukaryotic non-protists (about 40 species). The names of these species and the accession numbers of their mitochondria are listed in Table 3.1.

The genes for tRNAs in mitochondrial and nuclear DNAs in protists and non-protists were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) (the genes in many separated files were downloaded by robot.pl). Then these genes were divided into 4 groups (mito-pro, mito-non, nucl-pro, nucl-non). 'mito-pro' means the tRNA genes in mitochondrion of protists. 'nucl-non' means the tRNA genes in nucleus of non-protists, etc. Then the sequences in each group were classified into 20 subgroups corresponding to 20 types of amino acids. Each sequence was named according to its codon, corresponding amino acid and name of species. For example, AGCa1Tbru means that a nucleus tRNA gene in *Trypanosoma brucei* has Ala as its corresponding amino acid with codon of AGC. And tgca1Hsap means a mitochondrial tRNA gene in human has Ala as its corresponding amino acid with codon of TGC.

Group	Species	Mitochodrion
Haptophyceae	Emiliania huxleyi	NC_005332
Heterolobosea	Naegleria gruberi	NC_002573
Ichthyosporea	Amoebidium parasiticum	AF538042, AF538046, AF538045
Jakobidae	Reclinomonas americana	NC_001823
Malawimonadidae	Malawimonas jakobiformis	NC_002553
Mycetozoa	Dictyostelium discoideum	NC_000895
Rhodophyta	Cyanidioschyzon merolae Porphyra purpurea Chondrus crispus	NC_000887 NC_002007 NC_001677
Acanthamoebidae	Acanthamoeba castellanii	NC_001637
Alveolata	Tetrahymena thermophila Tetrahymena pyriformis Toxoplasma gondii Paramecium aurelia Plasmodium falciparum Cryptosporidium parvum Plasmodium yoelii	NC_003029 NC_000862 NC_001324
Cryptophyta	Guillardia theta Rhodomonas salina	 NC_002572
Diplomonadida	Giardia lamblia	
Euglenozoa	Euglena gracilis Leishmania major Leishmania infantum Leptomonas collosoma Leptomonas seymouri Trypanosoma brucei Crithidia fasciculata	
Stramenopiles	Laminaria digitata Pylaiella littoralis Chrysodidymus synuroideus Cafeteria roenbergensis Ochromonas danica Phytophthora infestans	NC_004024 NC_003055 NC_002174 NC_000946 NC_002571 NC_002387
Choanoflagellida	Monosiga brevicollis	NC_004309
Plants / Chlorophyta	Chlamydomonas reinhardtii Chlamydomonas eugametos Scenedesmus obliquus Nephroselmis olivacea Pedinomonas minor Chlorella vulgaris Prototheca wickerhamii	NC_001638 NC_001872 NC_002254 AF110138 NC_000892 AY267353 NC_001613
Plants / Streptophyta	Arabidopsis thaliana Beta vulgaris Marchantia polymorpha	NC_001284 NC_002511 NC_001660

	Oryza sativa	AB076665 AB076666
	Chara vulgaris	NC_005255
	Zea mays	AY506529
	Chaetosphaeridium globosum	NC_004118
	Phaseolus vulgaris	
Fungi	Saccharomyces cerevisiae	NC_001224
	Schizosaccharomyces pombe	NC_001326
	Encephalitozoon cuniculi	
	Allomyces macrogynus	NC_001715
Animals	Caenorhabditis elegans	NC_001328
	Drosophila melanogaster	NC_001709
	Homo sapiens	NC_001807

Table 3.1 Names of species used in this study and the accession numbers of their mitochondria. The species without accession number have no tRNA genes in their mitochdria.

We aligned these sequences with the alignments of the tRNA genes of proteobacteria to get good alignments and phylogenetic trees for further study as described in the following section. Here is an example of part of an alignment file from GeneDoc (Figure 3.3).

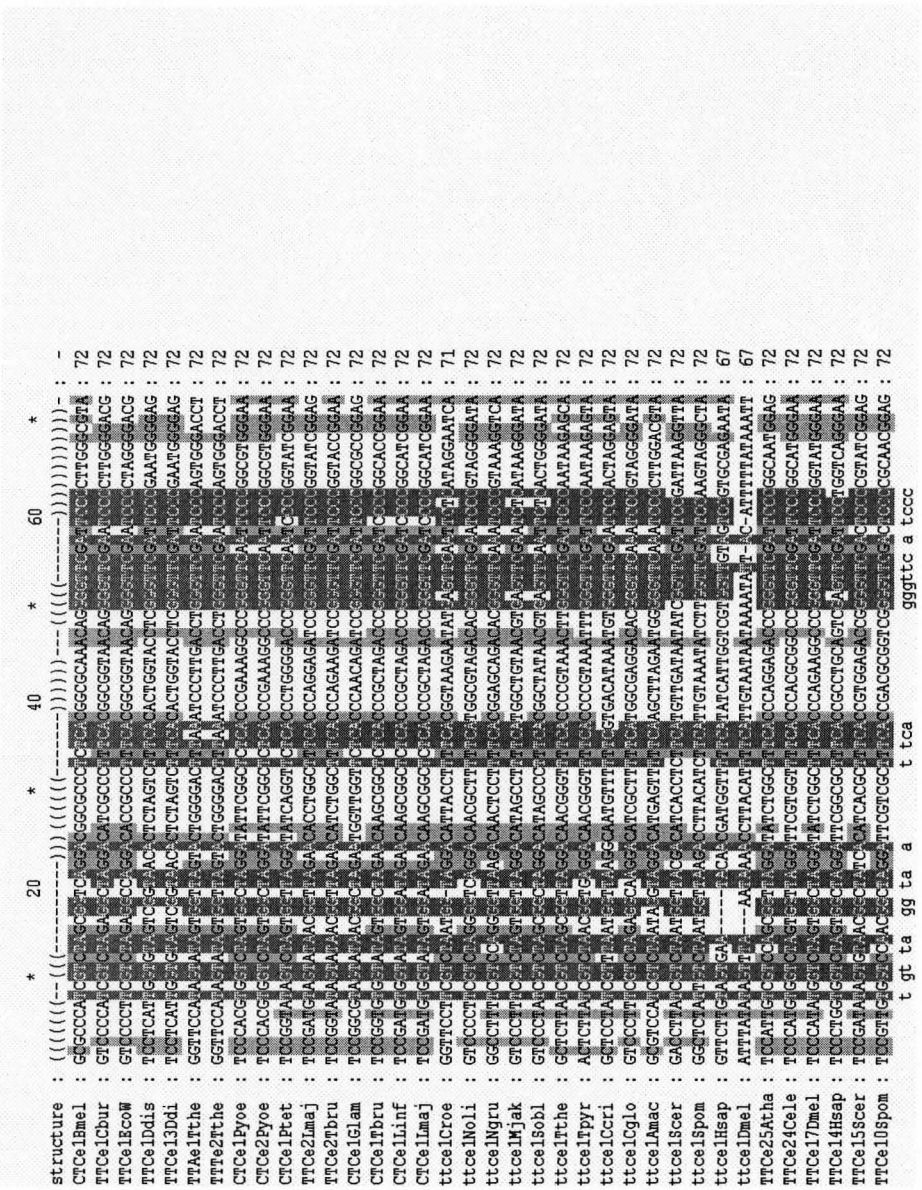

Figure 3.3 part of an alignment file from GeneDoc

3.4 Method

We use several programs to get the sequences and phylogenies and analyze them. First, we used robot.pl to download the several hundreds of separated files. Then, we used name.pl to classify the sequences into 20 subgroups and name the sequences. Some of sequences were named by hand because of their irregular format.

Using the profile alignment mode of ClustalX (Thompson *et al.* 1997), we aligned the additional sequences to the existing alignments of the tRNA genes of proteobacteria (Private communication from Tang *et al.*). The profile alignment was built up adding one group at a time until we obtained the final alignments which include the proteobateria, nucl-pro, mito-pro, mito-non and nucl-non group.

After getting the alignments, we edited them by GeneDoc (Nicholas *et al.* 1997). We want to use the information from secondary structure to improve the alignments, so we added the secondary structures from *E. coli* tRNAs to the top of the alignments and moved the residues manually according to the structures so that the alignments were more consistent with the secondary structure. Because tRNAs have very conservative secondary structure, moving the residues can improve the alignments.

We wanted a fairly rapid and simple phylogenetic method since there were a large number of short sequences. We do not expect to get fully resolved phylogenies from tRNAs. We used Tree-Puzzle to get the ML distance matrix. The model used here is the HKY-4 Γ. Then we used the Neighbor-Joining program from Phylip (Felsenstein 1989) get the tree files and 'drawtree' to draw unrooted trees.

3.5 Result

Finally, by using the alignments of tRNAs for 20 kinds of amino acids, we got the phylogenetic trees for each amino acid. Although they are not perfect, we still can get some useful information about our questions from them. Here are two examples of these trees (Figure 3.4): one is for Glu, another is for Lys. The trees in Figure 3.4 (a) and (b) have been rerooted to show a split between eukaryotic and bacterial genes. The mitochondrial genes form a subgroup within the bacteria (as we expect from the endosymbiotic theory). Importantly, the genes of *Trypanosoma* and *Leishmania* appear as a subgroup of the eukaryotic nuclear genes. That means the tRNA genes of *Trypanosoma* and *Leishmania* are similar in sequence to nuclear genes of these spieces and they are very likely to share the same ancestor with other eukaryotes.

We also did bootstrapping to assess the reliability of the trees. There are many poorly-resolved nodes on these trees because the sequence is so short, hence we have not labelled them all on the figures. There are only two important points: (i) that the eukaryotic nuclear sequences appear monophyletic, and the trypanosome sequences appear within this group, (ii) that the mitochondrial sequences appear monophyletic, and that none of the trypanosome sequences appear in this group. The bootstrap percentages of the eukaryotic nuclear clade for Glu and Lys are 53% and 42%, respectively. These figures are rather low, and do not give as strong statistical support as we would like; however, this is probably as good as can be expected for short tRNA sequences.

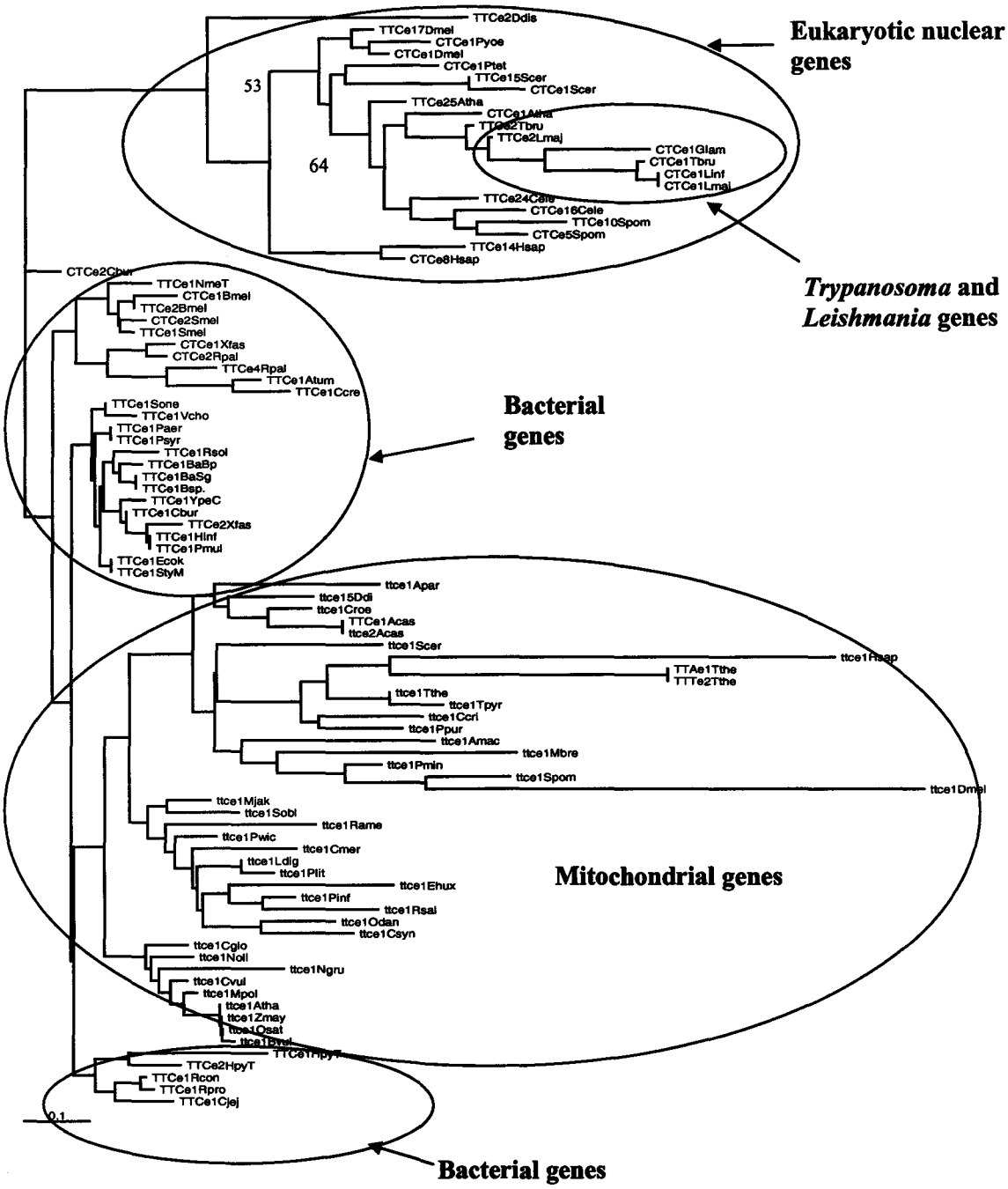

Figure 3.4 (a) Phylogeny based on tRNA genes for Glu

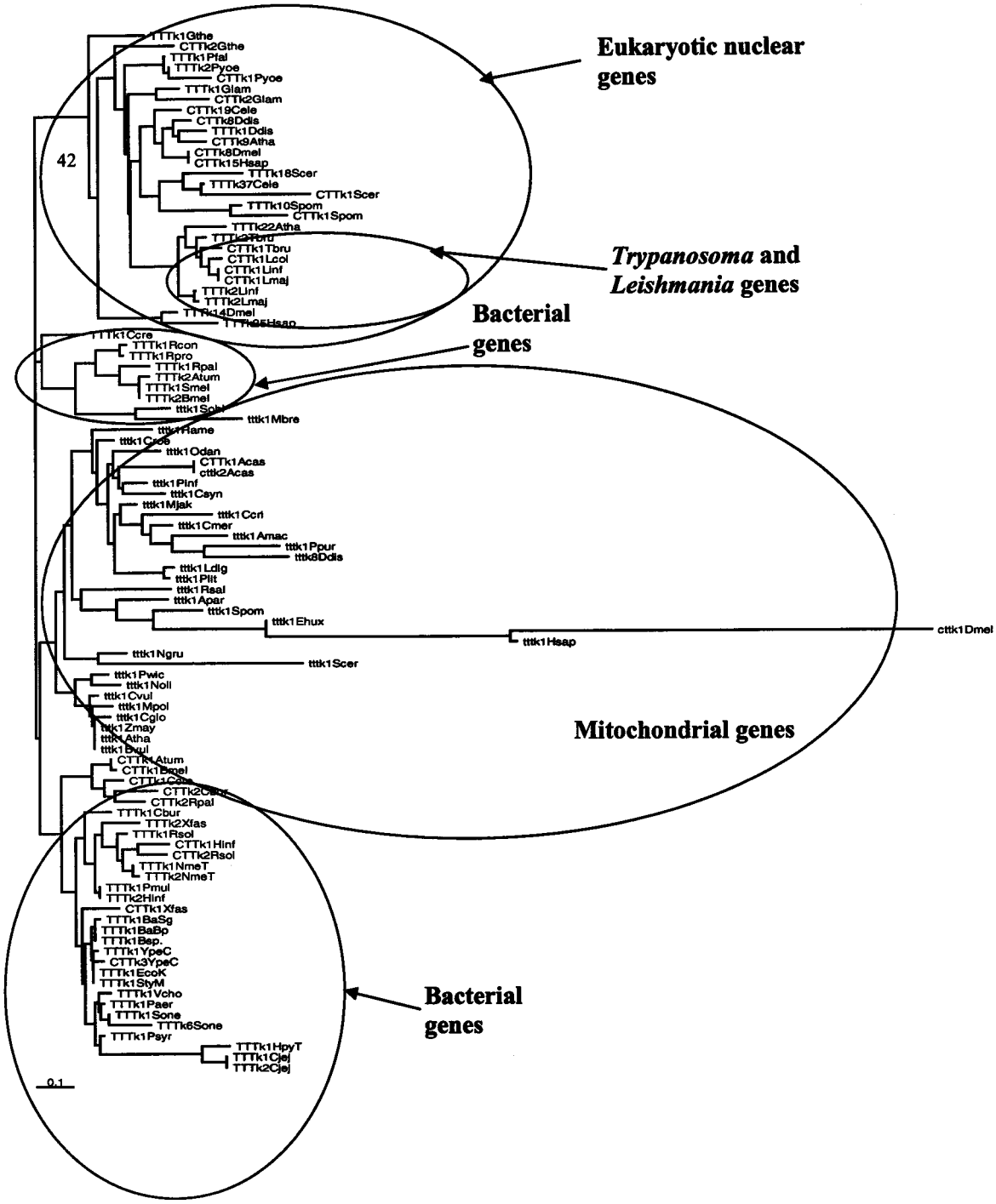

Figure 3.4 (b) Phylogeny based on tRNA genes for Lys

In the phylogenies obtained by the other tRNA genes, the same thing happened: the mitochondrial genes form a subgroup within the bacteria and the genes of *Trypanosoma* and *Leishmania* are always neighbors with the nuclear genes of other eukaryotes. In other words, all nuclear genes of the eukaryotes form a big group. The species that most frequently appear as close neighbors to the trypanosomes are *Giardia lamblia*, *Plasmodium yoelii*, *Guillardia theta*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum*.

In summary, there is no evidence indicating gene transfer from mitochondria to nucleus on the basis of these trees. These results are consistent with the simplest hypothesis, *i.e.*, that all tRNA genes of *Trypanosoma* and *Leishmania* have the same origin as nuclear genes of other eukaryotes, and that these genes (or at least some of them) have evolved to be functional inside the kinetoplast.

Chapter 4

Comparison of eukaryote phylogenies obtained from mitochondrial rRNA and protein sequences

4.1 Aim

There are several groups that have been well defined in the eukaryotic domain (Cavalier-Smith 1998). The metazoa, which means multicellular animals, is a diverse group; it is closely related to the fungi group which is also a rather diverse group; the fungi group has several important subgroups, such as Ascomycota, Basidiomycota and Chytridiomycota (Walker 1985; Prinllinger *et al.* 2002); the metazoa and fungi have been found to form a monophyletic group (Baldauf and Palmer 1993). Rhodophyta, red algae, have no flagellated stages in their life cycle; they are commonly reddish; Chlorophyta, green algae, have grass-green chloroplasts; they are closely related to land plants; Rhodophyta, Chlorophyta and plant form a monophyletic group (Cavalier-Smith 1998). Stramenopiles is a diverse protistan clade; the term stramenopiles refers to the numerous fine and hairlike projections on the flagella which are the characteristic of the organisms in this clade; Alveolata are unicellular protists which are emerging from molecular systematics (Wolters 1991; Saunders *et al.* 1995); Stramenopiles and Alveolata are thought to be monophyletic group (Wolters 1991; Saunders et al. 1995).

In this chapter, we carried out phylogenetic analysis for the species which have complete mitochondrial genomes by using both concatenated mitochondrial rRNA and protein sequences.

4.2 Dataset

The following list includes the species of interest with complete mitochondrial genomes. All RNA and protein sequences we are going to use in this chapter are from mitochondrial genomes.

Group	Species	Mitochondrion
Set 1		
Mycetozoa	Dictyostelium discoideum	NC_000895
Mycetozoa	Physarum polycephalum	NC_002508
Acanthamoebidae	Acanthamoeba castellanii	NC_001637
Fungi/insertae cedis	Hyaloraphidium curvatum	NC_003048
Fungi/Chytridiomycota	Allomyces macrogynus	NC_001715
Fungi/Chytridiomycota	Harpochytrium sp. JEL105	NC_004623
Fungi/Chytridiomycota	Harpochytrium sp. JEL94	NC_004760
Fungi/Chytridiomycota	Monoblepharella sp. JEL15	NC_004624
Fungi/Chytridiomycota	Spizellomyces punctatus	NC_003052
Fungi/Chytridiomycota	Rhizophydium sp. 136	NC_003053
Fungi/Basidiomycota	Crinipellis perniciosa	NC_005927
Fungi/Basidiomycota	Cryptococcus neoformans var. grubii	NC_004336
Fungi/Basidiomycota	Schizophyllum commune	NC_003049
Fungi/Ascomycota	Penicillium marneffei	NC_005256
Fungi/Ascomycota	Hypocrea jecorina	NC_003388
Fungi/Ascomycota	Podospora anserina	NC_001329
Fungi/Ascomycota	Pichia canadensis	NC_001762
Fungi/Ascomycota	Yarrowia lipolytica	NC_002659
Fungi/Ascomycota	Torrubiella confragosa	AF487277
Fungi/Ascomycota	Candida albicans	NC_002653
Fungi/Ascomycota	Candida glabrata	NC_004691
Fungi/Ascomycota	Candida parapsilosis	NC_005253
Fungi/Ascomycota	Candida stellata	NC_005972
Fungi/Ascomycota	Eremothecium gossypii	AF487277
Fungi/Ascomycota	Schizosaccharomyces japonicus	NC_004332

Fungi/Ascomycota	Schizosaccharomyces octosporus	NC_004312
Fungi/Ascomycota	Schizosaccharomyces pombe	NC_001326
Fungi/Ascomycota	Kluyveromyces lactis	NC_006077
Fungi/Ascomycota	Saccharomyces cerevisiae	NC_001224
Fungi/Ascomycota	Saccharomyces castellii	NC_003920
Fungi/Ascomycota	Saccharomyces servazzii	NC_004918
Ichthyosporea	Amoebidium parasiticum	AF538042, AF538046, AF538045, and other separated files
Choanoflagellida	Monosiga brevicollis	NC_004309
Metazoa/Porifera	Axinella corrugata	NC_006894
Metazoa/Cnidaria	Metridium senile	NC_000933
Metazoa/Arthropods	Drosophila melanogaster	NC_001709
Metazoa/Nematodes	Trichinella spiralis	NC_002681
Metazoa/Nematodes	Caenorhabditis elegans	NC_001328
Metazoa/Molluscs	Katharina tunicata	NC_001636
Metazoa/Platyhelminths	Fasciola hepatica	NC_002546
Metazoa/Hemichordates	Balanoglossus carnosus	NC_001887
Metazoa/Echinoderms	Paracentrotus lividus	NC_001572
Metazoa/Urochordates	Halocynthia roretzi	NC_002177
Metazoa/Vertebrates	Homo sapiens	NC_001807
Metazoa/Cephalochordates	Branchiostoma lanceolatum	NC_001912
Metazoa/Cephalochordates	Branchiostoma floridae	NC_000834
Set 2		
Malawimonadidae	Malawimonas jakobiformis	NC_002553
Jakobidae	Reclinomonas americana	NC_001823
Rhodophyta	Cyanidioschyzon merolae	NC_000887
Rhodophyta	Porphyra purpurea	NC_002007
Rhodophyta	Chondrus crispus	NC_001677
Chlorophyta	Chlamydomonas reinhardtii	NC_001638
Chlorophyta	Chlamydomonas eugametos	NC_001872
Chlorophyta	Scenedesmus obliquus	NC_002254
Chlorophyta	Nephroselmis olivacea	AF110138
Chlorophyta	Pedinomonas minor	NC_000892
Chlorophyta	Pseudendoclonium akinetum	NC_005926
Chlorophyta	Prototheca wickerhamii	NC_001613
Streptophyta	Mesostigma viride	AF353999
Streptophyta	Chara vulgaris	NC_005255
Streptophyta	Chaetosphaeridium globosum	NC_004118
Streptophyta/plants	Marchantia polymorpha	NC_001660
Streptophyta/plants	Arabidopsis thaliana	NC_001284
Streptophyta/plants	Beta vulgaris	NC_002511
Streptophyta/plants	Oryza sativa	X15901
Streptophyta/plants	Zea mays	AY506529

Set 3		
Haptophyceae	Emiliania huxleyi	NC_005332
Cryptophyta	Rhodomonas salina	NC_002572
Heterolobosea	Naegleria gruberi	NC_002573
Alveolata	Tetrahymena thermophila	NC_003029
Alveolata	Tetrahymena pyriformis	NC_000862
Alveolata	Paramecium aurelia	NC_001324
Stramenopiles	Laminaria digitata	NC_004024
Stramenopiles	Pylaiella littoralis	NC_003055
Stramenopiles	Chrysodidymus synuroideus	NC_002174
Stramenopiles	Cafeteria roenbergensis	NC_000946
Stramenopiles	Ochromonas danica	NC_002571
Stramenopiles	Phytophthora infestans	NC_002387
Stramenopiles	Saprolegnia ferax	NC_005984
Stramenopiles	Thraustochytrium aureum	AF288091

Table 4.1 species of interest with complete mitochondrial genome. These mitochondrial genomes can be downloaded from http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html and http://megasun.bch.umontreal.ca/ogmp/projects/other/all_list.html.

We have used all the species together to do phylogeny analysis using mcmcphase program in PHASE package, but the results were not good enough to do further analysis. Some species hang around on the deep branch of the trees, making it is impossible to decide which group these species belong to and the order of branching. The reason may be that the members are so diverse that the model parameters cannot be suitable for so many diverse species at the same time. When we used different random seeds, the phylogenies did not converge. That is maybe because the group is so big that the tree space is too large to explore globally and has too many local optimal point to get the converged results. So we divided them into three groups according to the widely accepted classification for the purpose of phylogenetic

analysis. Overall, the eukaryotic group contains several firmly reconstructed groups: metazoa, fungi and related species; the plantae; stramenopiles and the alveolates (Cavalier-Smith 1998).

The first data set we consider, fungi/metazoa group, contains fungi, metazoa and related species. All the species in this group are in the set 1 of Table 4.1. These species have been shown to form a monophyletic group previously (Baldauf and Palmer 1993; Drouin *et al.* 1995; Keeling and Doolittle 1996; Kuma *et al.* 1995). Fungi/metazoa group includes all the fungi in the Table 4.1. And other species included in the same group are

One Mycetozoa - *Dictiostelium discoideum* which is set as the outgroup

One Acanthamoebidae - *Acanthamoeba castellanii*

One Ichthyosporia - *Amoebidium parasiticum*

One Choanoflagellida - *Monosiga brevicollis*

Three metazoa - *Metridium senile*, *Drosophila melanogaster*, *Homo sapiens*. Although hundreds of complete mitochondrial genomes of metazoa are available, only these three representative metazoa have been included here because metazoan phylogeny has been widely studied by other research groups.

The second data set, plant/algae group, contains plants, green and red algae and related species. All the species in this group are in the set 1 of Table 4.1. This group includes three Rhodophyta, five Chlorophyta and five Streptophyta. Two other species are added as outgroups: *Malwimonas jakobiformis* and *Reclinomonas americana.* These are found to be related to the plants and algae by Ohta *et al* (1998) and Lang *et al.* (1999).

The third data set, stramenopile/alveolate group, contains the other available mitochondrial genomes from the protists. All the species in this group are in the set 3 of Table 4.1. The Stramenopiles and Alveolates are two groups that are thought to be monophyletic according to evidence from Wolters (1991) and Saunders (1995). We have included three Alveolates and seven Stramenopiles listed in Table 4.1. We have also added three other species of uncertain phylogenetic position:

One Haptophyceae - *Emiliania huxleyi* which is set as outgroup

One Cryptophyta - *Rhodomonas salina*

One Heterolobosea - *Naegleria gruberi*

For each group, we have the small subunit rRNA, large subunit rRNA sequences and the concatenated protein sequences. The phylogenetic analysis was carried out using both rRNA and protein sequences. We will focus on the case of rRNA, the latter was done by Supratim Sengupta in our group (See Appendix A for details of the latter).

For each group, we put the small and large subunit rRNA sequences into ClustalX to do alignment respectively. And then we added the rRNA secondary structures of *E. coli* on the top of the aligned sequences using GeneDoc and secondarystructure.pl.

Then we concatenated the two kinds of sequences into one file. After outputting the file from GeneDoc in Phylip format, for each group, we used the analyzer program in the PHASE package to check the mismatched pairs in the file. The analyzer gave us the positions of these mismatched pair sites above a certain cut-off threshold (between 0.0 and 1.0). The deletemismatch.pl was used to delete the

brackets which denote the mismatched pairs and count the number of gaps of each column in the alignment and delete the sites above the threshold.

For the fungi/metazoa group, the threshold is 10%. That means that the sites which have more than 10% mismatched pairs will be treated like single sites and that the sites which have more than 10% gaps are deleted. Similarly, for the plant/algae and stramenopiles/alveolate group the threshold is 30%. In order to determine the threshold values for the alignments, we have tried different percentage for each group (e.g. 10%, 20%, 30% and 40%), but it seemed that the latter two groups have more conservative insertions than the fungi/metazoa group, so the results are better when the threshold is higher. But generally speaking, the difference is small between different percentages in our study.

After the steps described above, we got the final version of the alignment file of concatenated rRNA sequences with secondary structure for each group to do phylogenetic analysis. The lengths of the alignments are 2063, 3829 and 3755 for fungi/metazoa, plant/algae and stramenopile/alveolate groups, respectively. The numbers of paired-sites are 632, 1684 and 1514, respectively.

4.3 Method

In our study, the final alignment files for each group were used by the mcmcphase program to produce large samples of trees under two different models according to whether or not the alignment had secondary structure. The model was REV+RNA7A-4Γ if the alignment had the secondary structure, otherwise the model was REV-4Γ. Then the large samples of trees were used by mcmcsummarize

program to get the consensus phylogenies of fungi/metazoa, plant/algae and stramenopile/alveolate groups, respectively, as shown in Figure 4.1, 4.2, 4.3.

To bootstrap the rRNA sequences, the secondarybootstrap.pl described in section 2.5 was used to produce datasets of randomized sequences with the secondary structure. In our study, 100 replicates were done for each group. These new sets of randomized sequences are stored in the boot file for each group. The distphase.sh was used to repeatedly call distphase program in the PHASE package to calculate the ML (Maximum Likelihood) distance matrix for each set in the boot file. The model we used to calculate the distance matrix is REV+RNA7A-4Γ. Neighbor-Joining method implemented in Phylip was used to construct trees from the distance matrix. The Neighbor-Joining method is quick and practical way of doing bootstrapping. Maximum likelihood method would be rather slow and unpractical for bootstrapping because the computer has to find the phylogeny with maximum likelihood, which would be a time-consuming procedure. It took about 22 hours for one replicate for fungi/metazoa group on SHARCNET system (http://www.sharcnet.ca). In our study, repeating the procedure 100 times will take months. After we got 100 trees, consense program in Phylip package was used to construct the consensus tree for each group. Notice that these consensus trees are different from the ones obtained from MCMC method.

When we treat the rRNA sequences as DNA sequences, *i.e.*, do not consider the secondary structure, the same method described above can be used to get the consensus tree for each group. The only difference is that the program used to produce bootstrapping sequence is seqboot program in Tree-Puzzle and that the model used to get the ML distance is REV-4Γ.

Finally, we got the bootstrapping consensus tree and output file from each group's alignments with and without secondary structure. The output files contain not only the clades that make up the consensus trees, but also the ones that do not show up in the trees. All the clades are listed in the output files according to their frequency of appearance.

We will compare our rRNA phylogenies with phylogenies derived from protein sequences from the same mitochondrial genomes. The protein sequence analysis was carried out by Supratim Sengupta in our research group (See Appendix A for details). We present some of his results here in order to discuss the similarities and differences of the rRNA and protein phyogenies.

4.4 Results and discussion

The results shown here are the consensus trees obtained from MCMC method. Two different support values are provided for each node of the trees indicating the frequency of appearance of the clade defined by the node. The first one is the Bayesian Posterior Probability (BPP) obtained from the mcmcsummarize program. The second one is the Neighbor-Joining bootstrapping support value. Almost all the nodes have very high BPP value, but the bootstrapping support values vary: some have nearly 100 percentage support, the others are very low. We should be cautious when this case occurs, and it is necessary to look at the results from other research groups, and compare with the classifications derived from the morphological methods or the analysis of other types of molecular sequences.

For the fungi/metazoa group, the topologies of the trees based on the rRNA alignments without and with secondary structure are identical and the support values are similar. We just show the phylogeny in the former case (Figure 4.1(a)). The phylogenies from rRNA and proteins are almost identical. Most of the bootstrapping support values of both trees are very high, only a few nodes have poor support values. That means there is very strong signal in both the rRNA and protein sequences.

All the fungi form a large subgroup. Several important divisions within the fungi are also seen, such as Chytridiomycota, Ascomycota and Basidiomycota. This is consistent with the results from other groups (Walker 1985; Prinllinger *et al.* 2002). The three metazoa group together as expected, and the closest relatives to the metazoa are *Amoebidium parasiticum* and *Monosiga brevicollis*.

This confirms the relationship shown by Firgure 4.1 (b) obtained from mitochondrial proteins, and shows that the mitochondrial rRNAs also support the same relationship. One difference is that *Eremothecium gossypii* was excluded from the analysis of the protein sequences because several genes are absent in this species. It therefore does not appear in Figure 4.1(b). Another difference is that *Candida stellata* and *Yarrowia lipolytica* branch together in the rRNA tree with bootstrapping support value 72%. In the protein tree, *Candida stellata* goes with other *Candida* with bootstrapping support value 2%. It is interesting to note that the genus *Candida* is not monophyletic, and is divided into three separate clades in both the rRNA and protein trees. This result is consistent with the result from Diezmann *et al.* (2004).

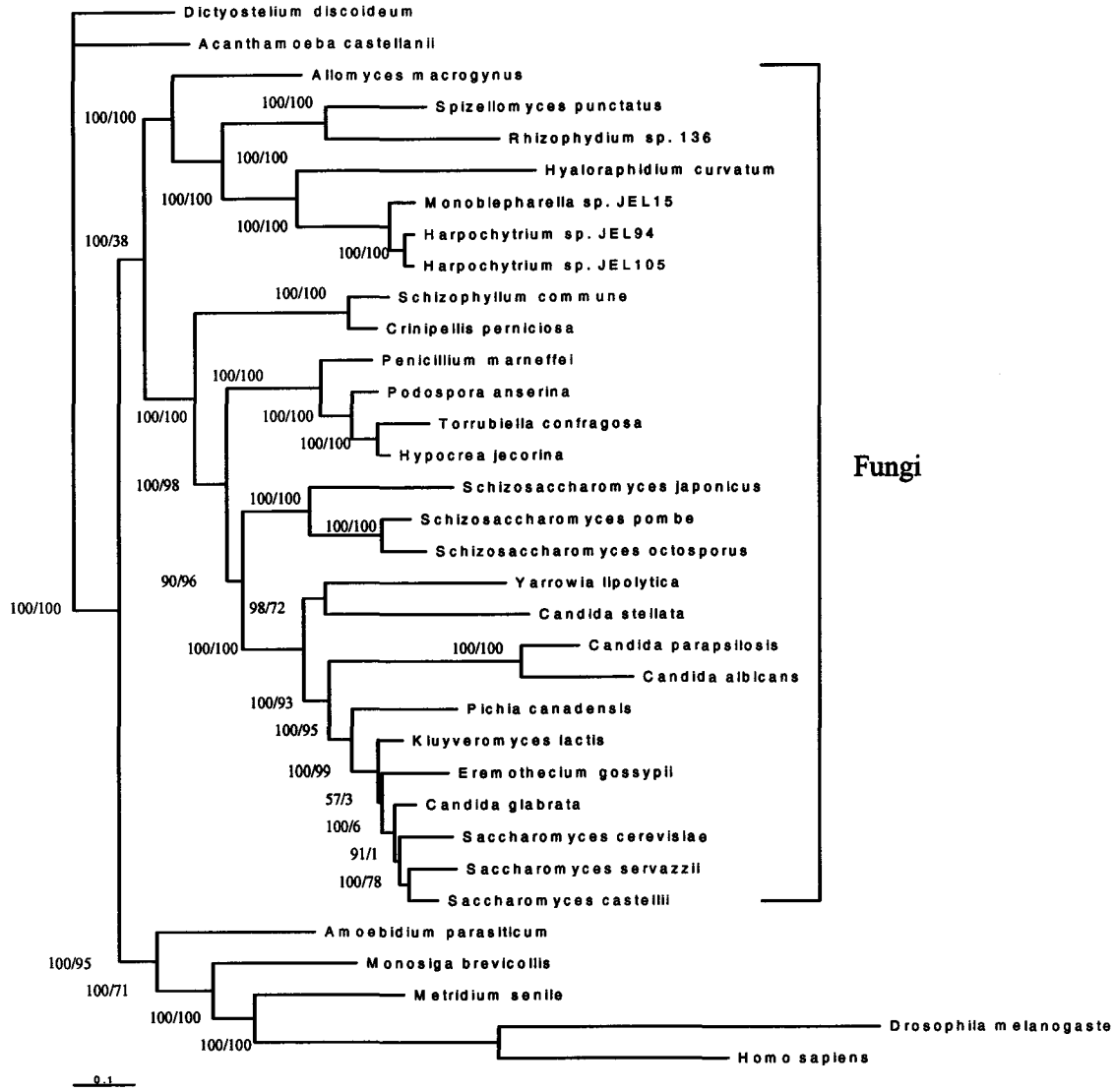

Figure 4.1 (a) Phylogeny of fungi/metazoa group based on rRNA genes (without secondary structure)

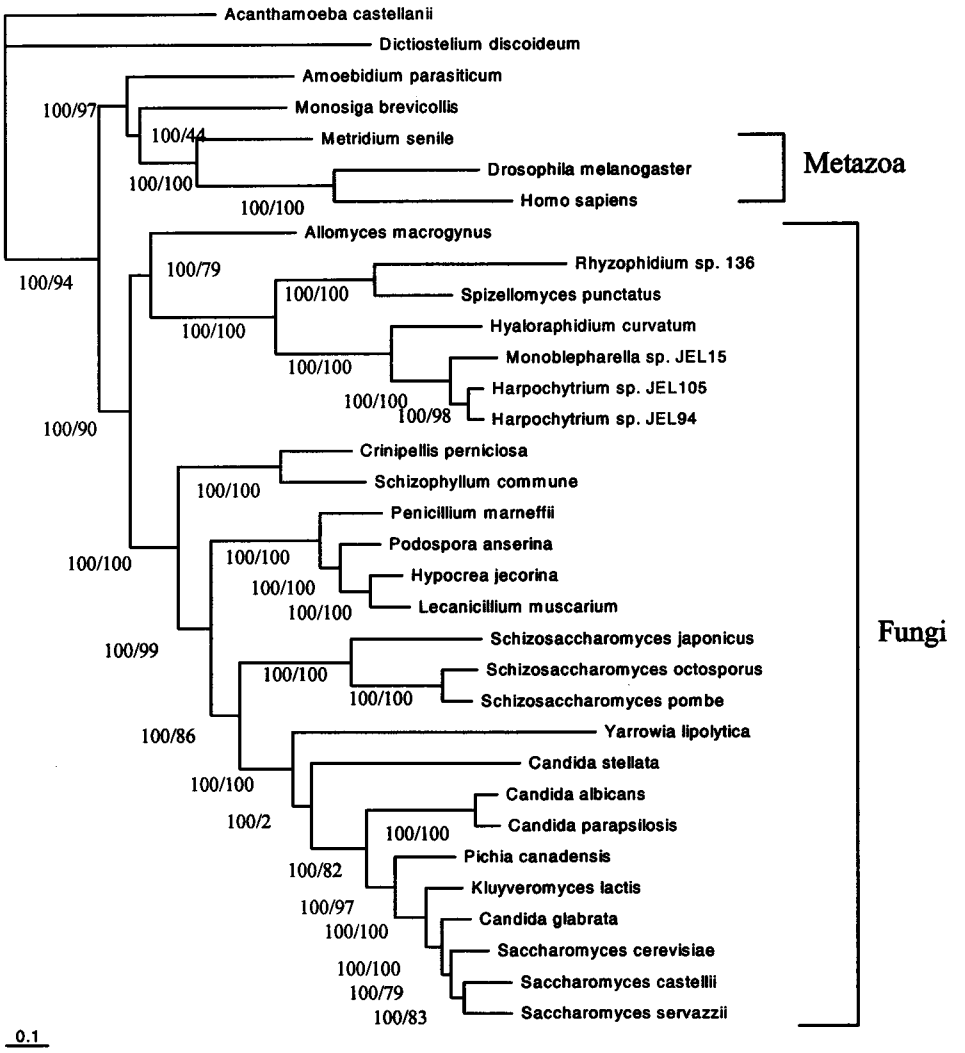

Figure 4.1 (b) Phylogeny of fungi/metazoa group based on protein sequences (courtesy of Supratim Sengupta)

For plant/algae group, we expect to find the major clades for rhodophytes (red algae), chlorophytes (green algae) and streptophytes. There is one difference between

the topologies of the trees based on the rRNA alignments without and with secondary structure (Figure 4.2(a) and (b)). In the former case, *Nephroselmis olivacea*, *Scenedesmus obliquus* and *Chlamydomonas reinhardtii*, which belong to *Chlorophyta*, go with *Streptophyta*. But in the latter case, *Scenedesmus obliquus* and *Chlamydomonas reinhardti* go with the other two species belonging to *Chlorophyta*. That means in some cases, the RNA substitution model may work better than the DNA substitution model. The result from protein sequences (Figure 4.2(c)) is almost identical to the latter tree. *Nephroselmis olivacea* form a subgroup with other *Chlorophyta* in the tree based on protein sequences. That is consistent with the morphological classification even though the bootstrapping support value on that node is zero. But in the same tree, *Reclinomonas americana*, which is supposed to be more closely related to *Malwimonas jakobiformis*, goes with *Rhodophyta* subgroup. As described above, in the tree based on rRNA alignment with secondary structure, *Nephroselmis olivacea* goes with *Streptophyta* with 100 percentage support from bootstrapping. It is hard to say which tree is right or better.

Some results from other authors (Lemieux *et al.* 2000; Turmel *et al.* 1999b) show that *Nephroselmis olivacea* will go with *Chlorophyta* based on chloroplast proteins with high support values. But when the proteins from mitochondria are used, in some cases, it will go with *Streptophyta* (Turmel *et al.* 1999a); even though in other cases it will go with *Chlorophyta*, the support values are very low (Pombert *et al.* 2004). The tree based on rRNA alignment with secondary structure may be right in some sense and indicate the different origination of its mitochondria from its chloroplast, although this would mean horizontal transfer.

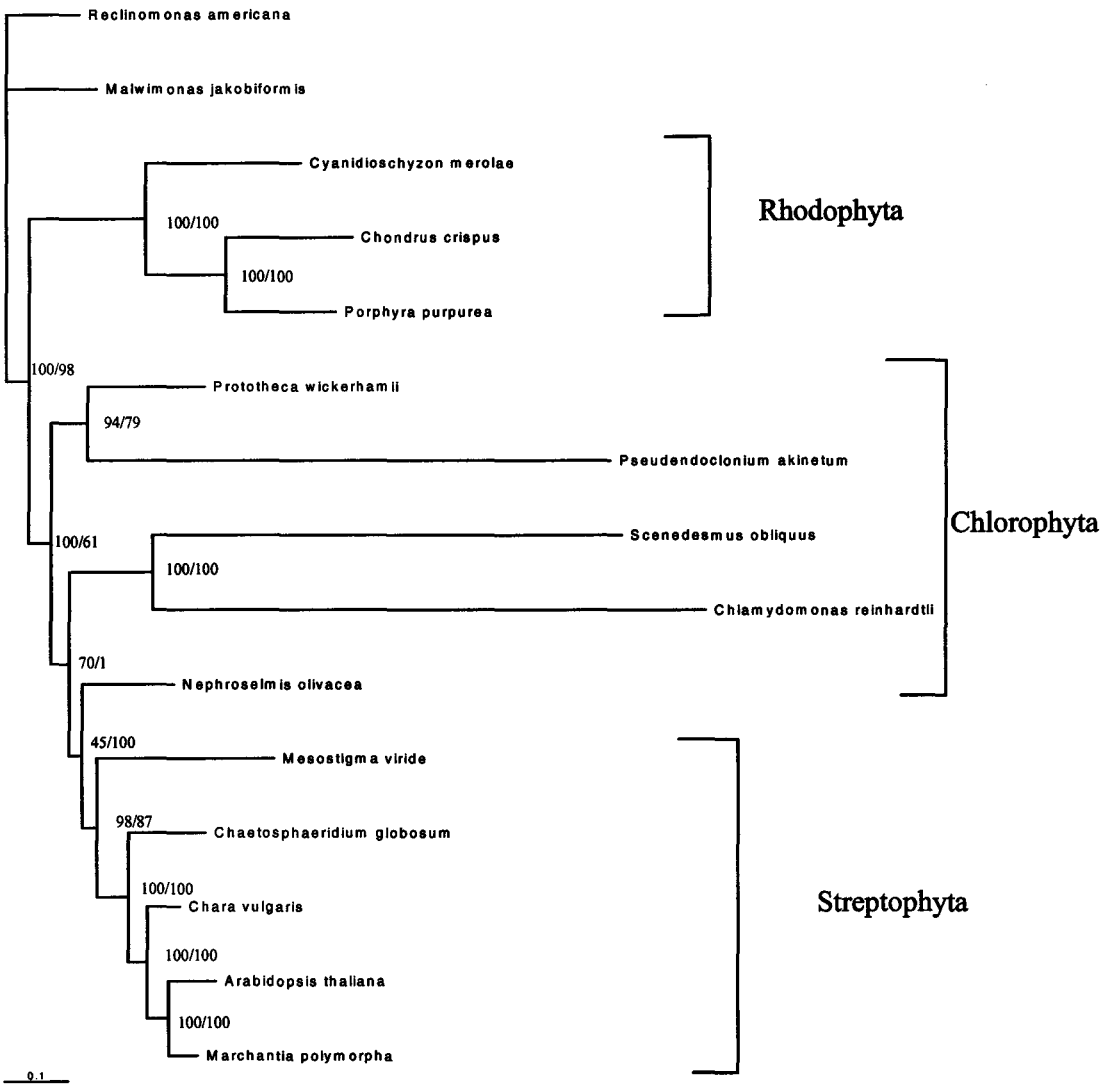

Figure 4.2 (a) Phylogeny of plant/algae group based on rRNA genes (without secondary structure)

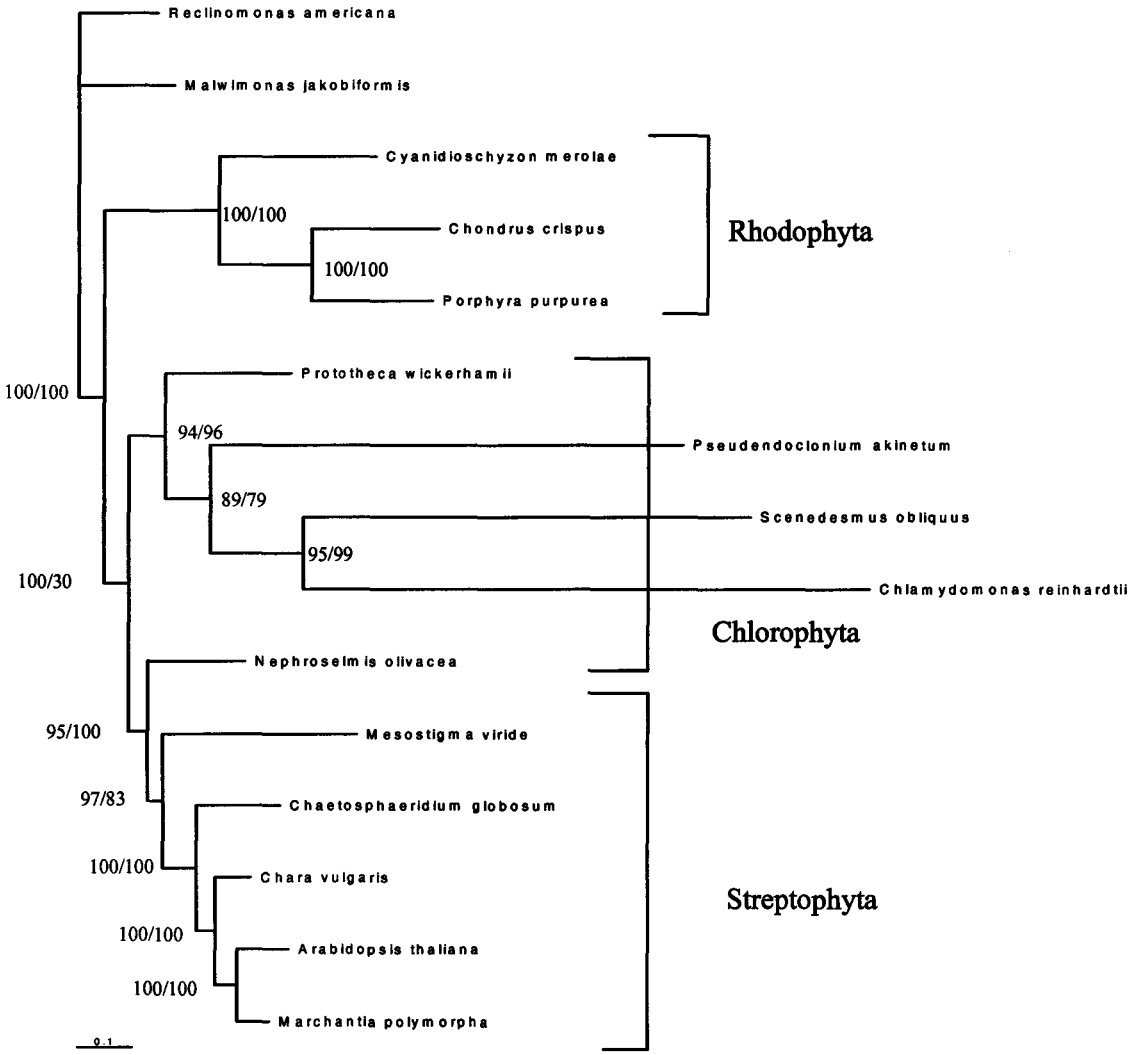

Figure 4.2 (b) Phylogeny of plant/algae group based on rRNA genes (with secondary structure)

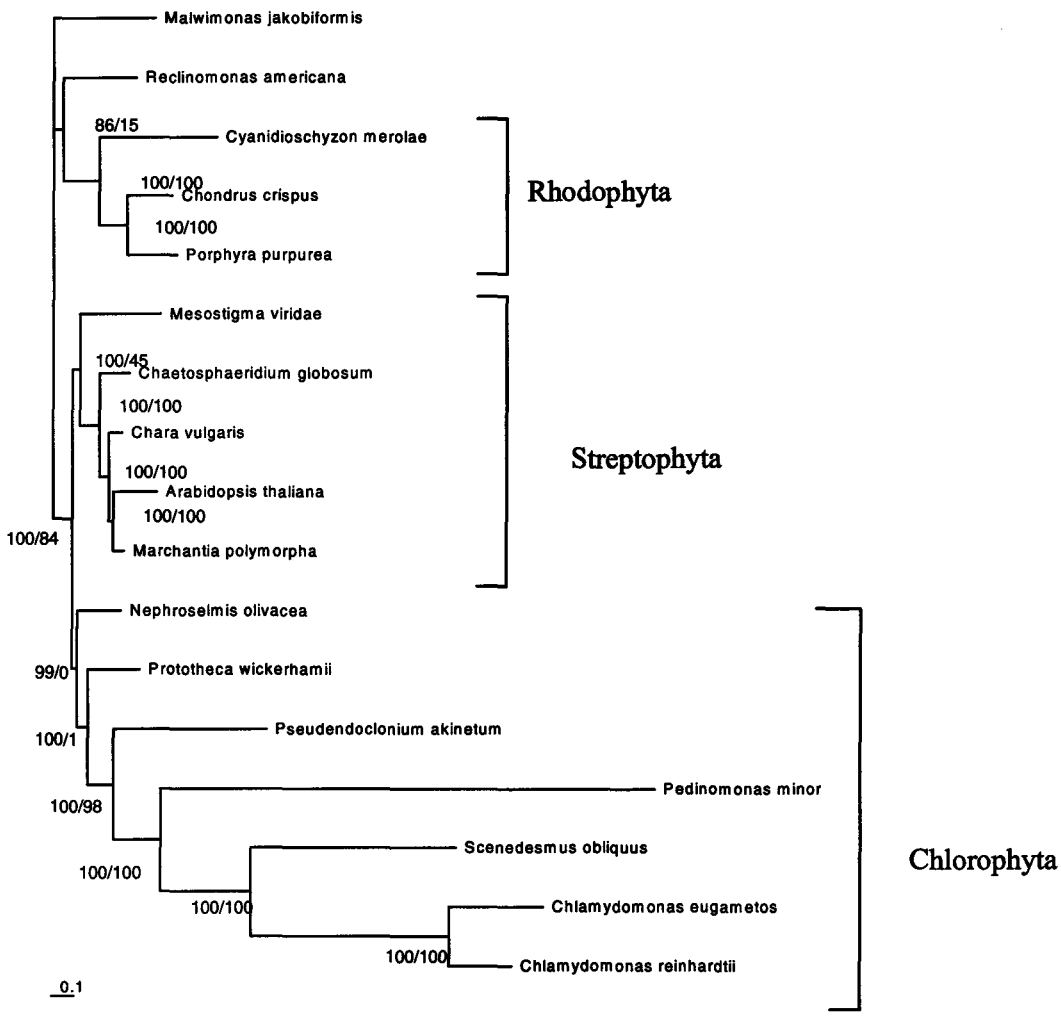

Figure 4.2 (c) Phylogeny of plant/algae group based on protein sequences (courtesy of Supratim Sengupta)

For stramenopile/alveolate group, the topologies of the trees based on the rRNA alignments without and with secondary structure are identical and the support values are similar. We show the phylogeny in the former case (Figure 4.3(a)). In this tree three species in *Alveolata* form a group with species from *Stramenopiles*, which is inconsistent with the classification previously reconstructed and supported by Wolters (1991) and Saunders *et al.* (1995). In the mean time, the deep branch nodes have very low support value. The phylogeny based on protein sequences are also shown in Figure 4.3(b). In this tree, *Stramenopiles* and *Alveolata* are separated clearly, forming two isolated groups, although the support values on deep branch nodes are not very high. This result is consistent with the previously reconstructed and supported classification (Wolters 1991; Saunders *et al.* 1995).

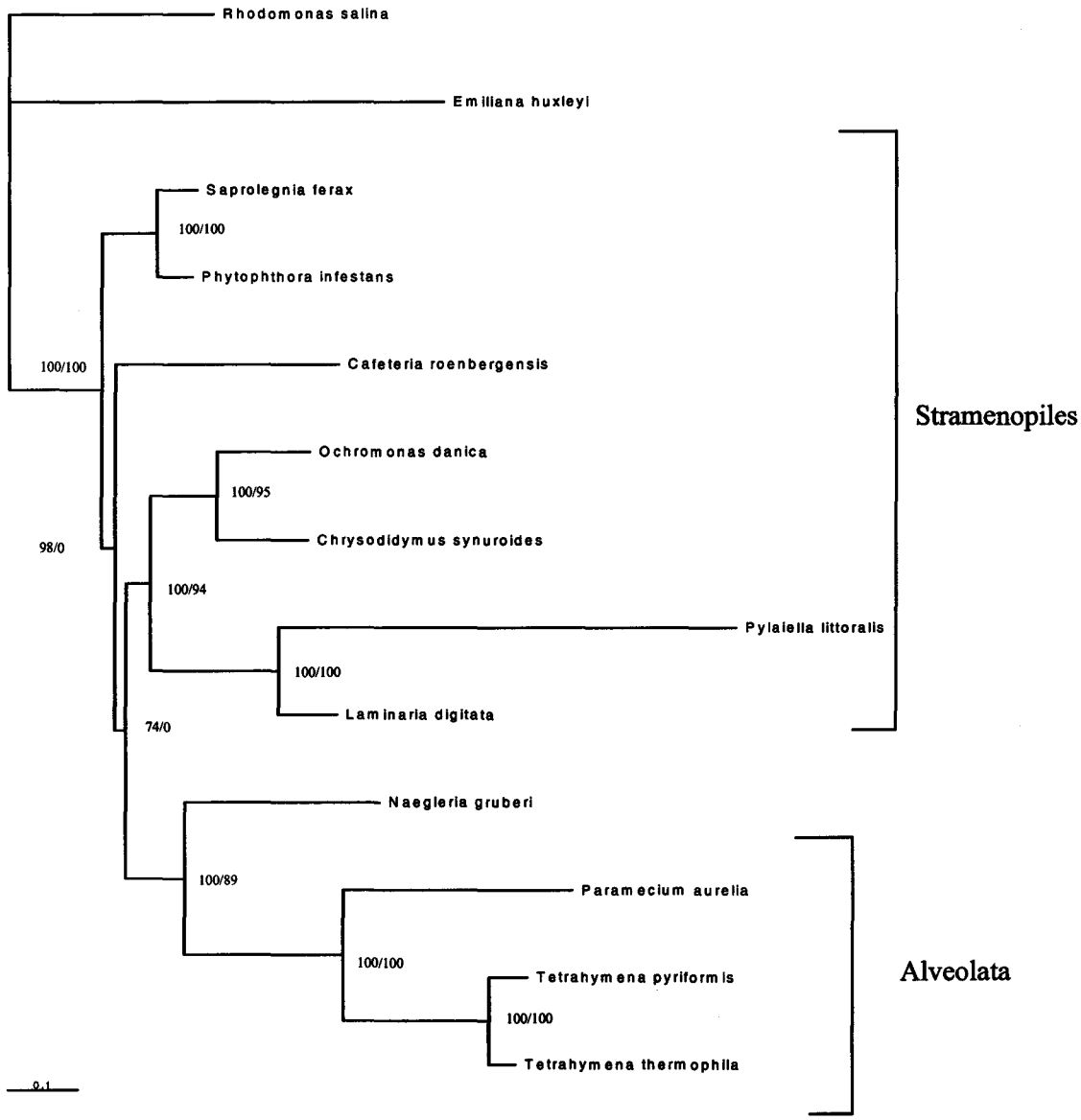

Figure 4.3 (a) Phylogeny of stramenopile/alveolate group based on rRNA genes (without secondary structure)

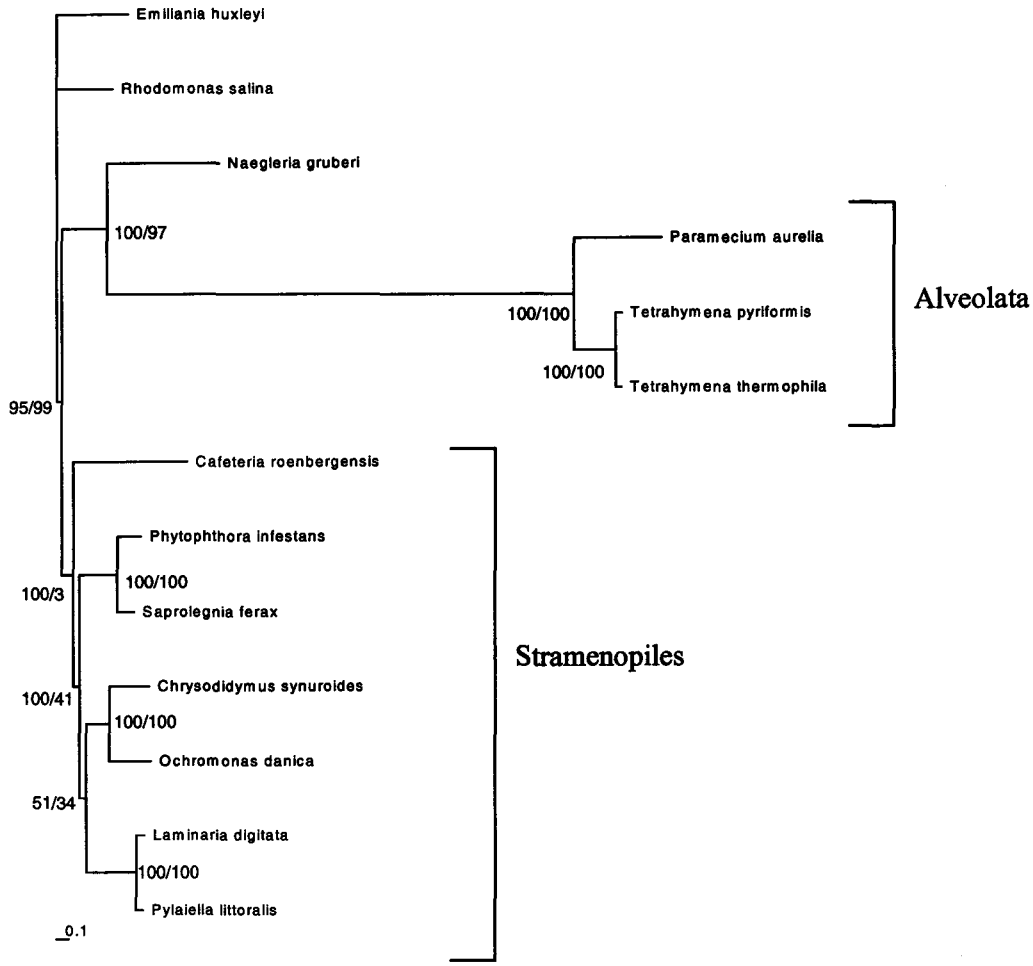

Figure 4.3 (b) Phylogeny of stramenopile/alveolate group based on protein sequences (courtesy of Supratim Sengupta)

4.5 Conclusion

The results from mitochondrial sequences are consistent with the previous classifications based on the nuclear sequences for these groups (Cavalier-Smith 1998; Drouin *et al.* 1995; Keeling and Doolittle 1996; Ohta *et al* 1998; Saunders 1995). That is to say that during and after endosymbiotic events, the resident genomes evolve within the host cells and they became dependent on each other. So the phylogenies based on the nuclear and mitochondrial sequences should be closely related to each other.

For fungi/metazoa group and plant/algae group, the phylogenies based on mitochondrial rRNA and proteins are almost identical. For stramenopile/alveolate group, the rRNA phylogeny and protein phylogeny are different. The rRNA phylogeny is not consistent with the previous results. Meanwhile the consensus tree based on rRNA from bootstrapping using Neighbor-Joining method, which is not shown here, is consistent with the protein phylogeny and previous results. MCMC method is supposed to make the most use of information in the data, so we are not sure about what happened in the rRNA phylogeny obtained by using MCMC method. Almost all the nodes in the phylogenies obtained in chapter 4 have very high BPP value for MCMC method, but some of the bootstrapping support values are very low. We should be cautious when this case occurs, and it is necessary to look at the results from other research groups, and compare with the classifications derived from the morphological methods or the analysis of other types of molecular sequences. It is also worth trying to change the model used to infer phylogenies, *e.g.*, using paired-site substitution model for secondary structure instead of nucleotide substitution model.

For fungi/metazoa and stramenopile/alveolate group, secondary structure did not change the topology of the phylogenies. It means that the information in the primary sequences is strong enough; but for plant/algae group, the secondary structure changed the position of *Scenedesmus obliquus* and *Chlamydomonas reinhardti*, which went with the other two species belonging to *Chlorophyta*, and also increase the support values. It means that the secondary structure can improve phylogenies in some cases.

Although there are still some problems with each phylogeny in each group, these results are very useful for the further analysis about the changes in gene content and the genetic code during the evolution of the mitochondrial genome.

4.6 Future work

The genetic code in mitochondrion differs from "universal code". Interestingly, many reassignments of the same codons happened independently. This indicates that there may be some similar underlying mechanisms in different species. We will use the phylogenies obtained in chapter 4 to determine the positions in the evolutionary tree of the genetic code changes and the relevant gene losses to understand the mechanisms involved.

Two interesting things have been found. One thing is that tRNA genes for Thr seem to be easy to lose during the evolution. All the species belong to *Stramenopiles* and *Alveolata* have no tRNA genes for Thr. This supports for joining these two monophyletic clades. Especially, all *Stramenopiles* only lost tRNA gene for Thr, which is also a support for this monophyletic clade. Several species in plant/algae

group have no tRNA genes for Thr, such as *Rhodophyta* clade, and *Scenedesmus obliquus*, *Chlamydomonas reinhardtii*, *Mesostigma viride*, *Arabidopsis thaliana*. And it seems that these species lost the tRNA genes for Thr independently. That means that there must be some similar mechanisms going on in there. All the species within fung/Chytridiomycota clade, except for *Allomyces macrogynus*, has no tRNA genes for Thr. *Allomyces macrogynus* branches on the base of the clade. This is another support for this clade. We have not found any papers talking about the loss of tRNA for Thr.

Another thing is that some species have retained the similar set of tRNA genes during the gene loss process. All the species within fung/Chytridiomycota clade, except for *Allomyces macrogynus*, have reduced sets of tRNA genes with similar content. *Alveolata* also have similar reduced sets of tRNA genes. *Tetrahymena pyriformis* and *Tetrahymena thermophila* have 8 tRNA genes which include the 4 genes that *Paramecium aurelia* has. The similarity supports the results obtained in chapter 4.

By the way, some positions of the genetic code changes have been determined by Dr. Supratim Sengupta based on the phylogenies obtained in chapter 4 (Supratim Sengupta, Private communication). Our future study will focus on the details of these things which may be relevant to the reassignment of the genetic code and determine the mechanisms involved.

Bibliography

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature.* 396:133-140 (1998).

Attardi G, Schatz G. Biogenesis of mitochondria. *Annu. Rev. Cell Biol.* 4:289-333 (1988).

Baldauf SL, Palmer JD. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA* 90:11558–11562 (1993).

Barrett MP, Burchmore RJ, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, Krishna S. The trypanosomiases. *Lancet.* 362:1469-1480 (2003).

Boore JL. Animal mitochondrial genomes. *Nucleic Acids Res.* 27:1767-1780 (1999).

Butow BA, Doeherty R, Parikh VS. A path from mitochondria to the yeast nucleus. Philos. *Trans. R. Soc. Lond. (Biol.)* 319:127-133 (1988).

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR. The

Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs. *BioMed Central Bioinformatics* 3:2 (2002).

Cavalier-Smith T. A revised six kingdom system of life. *Biol Rev Cambr Philos Soc* 73:203-266 (1998).

Cavalier-Smith T. The origin of eukaryotic and archaebacterial cells. *Ann. N.Y. Acad. Sci.* 503:17-54 (1987).

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ, Mitchell TG. Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis. *J Clin Microbiol.* 42:5624-5635 (2004).

Drouin G, Moniz de Sa, M, Zuker M. The Giardia lamblia actin gene and the phylogeny of eukaryotes. *J. Mol. Evol.* 41:841–849 (1995).

Dunn SR, Thomason JC, Le Tissier MD, Bythell JC. Heat stress induces different forms of cell death in sea anemones and their endosymbiotic algae depending on temperature and duration. *Cell Death Differ.* 11:1213-1222 (2004).

El-Sayed NM *et al.*. Comparative Genomics of Trypanosomatid Parasitic Protozoa. *Science* 309:404-409 (2005).

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, Martin W. A genome phylogeny for mitochondria among alpha-Proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21:1643-1660 (2004).

Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376 (1981).

Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution.* 39:773-781 (1985).

Felsenstein J. PHYLIP--Phylogeny Inference Package. *Cladistics.* 5:164-166 (1989).

Fernandes MV, Wiktor TJ, Koprowski H. Endosymbiotic relationship between animal viruses and host cells: a study of rabies virus in tissue culture. *J Exp Med.* 120:1099-1116 (1964).

Gellissen G, Michaelis G. Gene transfer: mitochondria to nucleus. *Ann. N.Y. Acad. Sci.* 503: 391-401 (1987).

Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.* 26:865-878 (1998).

Gray MW, Spencer DF. Evolution of Microbial Life (eds Roberts, D. M., Sharp, P. M., Alderson, G. & Spencer, D. F.) 109-126 (Cambridge Univ. Press, Cambridge, 1996).

Hackstein JH, Vogels GD. Endosymbiotic interactions in anaerobic protozoa. *Antonie Van Leeuwenhoek.* 71:151-158 (1997).

Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174 (1985).

Higgs PG. RNA secondary structure: physical and computational aspects. *Quart. Rev. of Bioph.* 22:199-253 (2000).

Hoshina R, Kamako SI, Imamura N. Phylogenetic position of endosymbiotic green algae in Paramecium bursaria Ehrenberg from Japan. *Plant Biol (Stuttg).* 6:447-453 (2004).

Jow H, Hudelot C, Rattray M, Higgs PG. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol Biol Evol.* 19:1591-1601 (2002).

Keeling PJ, Doolittle WF. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol. Biol. Evol.* 13:1297–1305 (1996).

Klingbeil MM, Drew ME, Liu Y, Morris JC, Motyka SA, Saxowsky TT, Wang Z, Englund PT. Kuma K, Nikoh N, Iwabe N, Miyata T. Phylogenetic position of Dictyostelium inferred from multiple protein data sets. *J. Mol. Evol.* 41:238–246 (1995).

Knight RD, Freeland SJ, Landweber LF. Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet.* 2:49-58 (2001).

Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G. A comparative genomics approach to the evolution of eukaryotes and their mitochondria. *J Eukaryot Microbiol.* 46:320-326 (1999).

Lemieux C, Otis C, Turmel M. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* 403:649-652 (2000).

Lukes J, Guilbride DL, Votypka J, Zikova A, Benne R, Englund PT. Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot Cell.* 1:495-502 (2002).

Morris JC, Drew ME, Klingbeil MM, Motyka SA, Saxowsky TT, Wang Z, Englund PT. Replication of kinetoplast DNA: an update for the new millennium. *Int J Parasitol.* 31:453-458 (2001).

Ndiaye M, Mattei X. Endosymbiotic relationship between a rickettsia-like microorganism and the male germ-cells of Culex tigripes. *J Submicrosc Cytol Pathol.* 25:71-77 (1993).

Nicholas KB. Nicholas H.B. Jr., and Deerfield, D.W. II. GeneDoc: Analysis and Visualization of Genetic Variation, *EMBNEW.NEWS* 4:14 (1997).

Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302:205-217 (2000).

Ohta N, Sato N, Kuroiwa T. Structure and organization of the mitochondrial genome of the unicellular red alga Cyanidioschyzon merolae deduced from the complete nucleotide sequence. *Nucleic Acids Res.* 26:5190-5198 (1998).

Olsen GJ, Woese CR, Overbeek R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* 176:1-6 (1994).

Osawa S, Jukes TH. Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* 28:271-278 (1989).

Page RD. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci.* 12:357-358 (1996).

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. USA*. 97:6960-6966 (2000).

Pombert JF, Otis C, Lemieux C, Turmel M. The Complete Mitochondrial DNA Sequence of the Green Alga Pseudendoclonium akinetum (Ulvophyceae) Highlights Distinctive Evolutionary Trends in the Chlorophyta and Suggests a Sister-Group Relationship Between the Ulvophyceae and Chlorophyceae. *Mol. Biol. Evol.* 21:922–935 (2004).

Poyton RO, McEwen JE. Crosstalk between nuclear and mitochondrial genomes. *Annu Rev Biochem.* 65:563-607 (1996).

Prillinger H, Lopandic K, Schweigkofler W, Deak R, Aarts HJ, Bauer R, Sterflinger K, Kraus GF, Maraz A. Phylogeny and systematics of the fungi with special reference to the Ascomycota and Basidiomycota. *Chem Immunol.* 81:207-295 (2002)

Private communication from Bin Tang, Philippe Boisvert and Paul Higgs: Department of Physics and Astronomy, McMaster University.

Private communication from Supratim Sengupta: Department of Physics and Astronomy, McMaster University.

Ray DS. Kinetoplast DNA minicircles: high-copy-number mitochondrial plasmids. *Plasmid.* 17:177-190 (1987).

Renesto P, Ogata H, Audic S, Claverie JM, Raoult D. Some lessons from *Rickettsia* genomics. *FEMS Microbiol Rev.* 29:99-117 (2005).

Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* 95:6239–6244 (1998).

Saitou N, Nei M. The neighbor-joining method: A new method of constructing phylogenetic trees. *Mol. Biol. Evol.* 4:1406-1425 (1987).

Saunders GW, Potter D, Paskind MP, Andersen RA. Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. *Proc. Natl. Acad. Sci. USA* 92:244–248 (1995).

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*. 18:502-504 (2002).

Schoniger M, von Haeseler A. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phyl. Evol.* 3:240-247 (1994).

Schultz DW, Yarus M. On malleability in the genetic code. *J. Mol. Evol.* 42:597–601 (1996).

Schwartz RM, Dayhoff MO. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* 199:395-403 (1978).

Sengputa S, Higgs P. A Unified Model of Codon Reassignment in Alternative Genetic Codes. *Genetics.* 170:831–840 (2005).

Shapiro TA, Englund PT. The structure and replication of kinetoplast DNA. *Annu Rev Microbiol.* 49:117-143 (1995).

Sicheritz-PonteÂn T, Kurland, CG, Andersson, SGEA. phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochim. Biophys. Acta* 1365:545-551 (1998).

Stuart K. Kinetoplast DNA, mitochondrial DNA with a difference. *Mol Biochem Parasitol.* 9:93-104 (1983).

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nuceicl Acids Res.* 24:4876-4882 (1997).

Tillier ERM. Maixmum likelihood with multiparameter modles of substitution. *J. Mol. Evol.* 39:409-417 (1994).

Tillier ERM, Collins R. High apparent rate of simultaneous compensatory basepair substitutions in ribosomal RNA. *Gen.* 148:1993-2002 (1998).

Tsuchida T, Koga R, Meng XY, Matsumoto T, Fukatsu T. Characterization of a facultative endosymbiotic bacterium of the pea aphid Acyrthosiphon pisum. *Microb Ecol.* 49:126-133 (2005).

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW. The Complete Mitochondrial DNA Sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: Two Radically Different Evolutionary Patterns within Green Algae. *The Plant Cell* 11:1717–1729 (1999a).

Turmel M, Otis C, Lemieux C. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast genomes. *Proc. Natl. Acad. Sci. USA* 96:10248–10253 (1999b).

Unlocking the secrets of trypanosome kinetoplast DNA network replication. *Protist* 152:255-262 (2001).

Viale A, Arakaki AK. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett.* 341:146-151 (1994).

Walker WF. 5 S and 5.8 S ribosomal RNA sequences and protist phylogenetics. Biosystems. 18:269-278 (1985).

Whatley JN, John P, Whatley FR. From extracelluar to intracellular: the establishment of mitochondria and chloroplasts. *Proc. R. Soc. Lond. (Biol.)* 204:165-187 (1979).

Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA.* 87:4576-4579 (1990).

Wolstenholme DR. Genetic novelties in mitochondrial genomes of multicellular animals. *Curr Opin Genet Dev.* 2:918-925 (1992).

Wolters J. The troublesome parasites—Molecular and morphological evidence that Apicomplexa belong to the dinoflagellate ciliate clade. *BioSystems* 25:75-83 (1991).

Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39:306-314 (1994).

Appendix A

We wished to do phylogenies using concatenated proteins. But before concatenating them, we wished to check whether there were any proteins that were significantly more divergent than the others because fast-evolving proteins are likely to be less well aligned and might add to the noise in the phylogenetic analysis.

T-COFFEE was used to carry out the multiple sequence alignments independently for each gene. The pairwise ML distance matrixes between different species for same proteins were calculated using the DISTPHASE program in the PHASE package. The mean pairwise distance for each protein was calculated by averaging over the pairwise distances between all species. The proteins whose mean pairwise distance was greater than the average of the mean pairwise distance for all proteins were identified as fast-evolving proteins and not included in our set of concatenated proteins. Based on this analysis, the atp8, nad2, nad6 were fast-evolving proteins and not included in any group. The nad proteins were not included in the fungi/metazoa group since many species of fungi do not have them.

The set of concatenated proteins for the fungi/metazoa group consisted of cox1, 2, 3, cob, atp6, 9, 1712 amino acid long. The sets for the plant/algae group and stramenopile/alveolate were made up of cox1, 2, 3, cob, atp6, 9, nad 1, 3, 4, 41, 5, having a total length of 3385 and 3360 amino acids.

The alignments of selected proteins were further edited using GeneDoc to remove the columns having more than 20% gaps. These alignment files were then concatenated to produce a single file which was used by mcmcphase program in the PHASE and the bootstrapping program. The model was mtREV24-4Γ. Neighbor-

joining Method was used with ML distance matrix for bootstrapping. Two different support values were provided for each node of the phylogenies. The first is the Bayesian posterior probability obtained by using the mcmcsummarize program in the PHASE. The second one is the percentage bootstrap value for the clade defined by that node, obtained by using the consense program in Phylip.