Title
Molecular Evolution of Mitochondrial and Nuclear Genomes in Ciliate Terahymena

Permalink
https://escholarship.org/uc/item/35x3v8pw

Author
Moradian, Mike M.

Publication Date
2007

Peer reviewed
UNIVERSITY OF CALIFORNIA

Los Angeles

Molecular Evolution of

Mitochondrial and Nuclear Genomes

in Ciliate Tetrahymena

A dissertation submitted in partial satisfaction of the
Requirements for the degree Doctor of Philosophy

In Biology

by

Mike Mkhitar Moradian

2007
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
The dissertation of Mike Mkhitar Moradian is approved.

Fred C. Fox

Robert K. Wayne

Matteo Pellegrini

Daniel Blumstein

David Jacobs, Committee Chair

University of California, Los Angeles

2007
TABLE OF CONTENTS

INTRODUCTION
  *Tetrahymena* Biology 1
  *Tetrahymena* mitochondrial genome project 10
  *Tetrahymena* Nuclear Genome Analysis 14
  References 16

CHAPTER 1
  Origin and Evolution of Mitochondria 18
  Summary and Conclusions 35
  References 36

CHAPTER 2
  Abstract 38
  Introduction 40
  Material and Methods 45
  Results and Discussion 46
  Summary and Conclusions 66
  Appendix 67
  References 77

CHAPTER 3
  Abstract 81
  Introduction 83
  Material and Methods 87
  Results and Discussion 92
  Conclusions and Future Plans 112
  References 113

CHAPTER 4
  Abstract 117
  Introduction 119
  Material and Methods 121
  Results and Discussion 123
  Summary and Conclusions 133
  References 134

CHAPTER 5
  Abstract 136
  Introduction 138
  Material and Methods 139
  Results 141
Discussion and Conclusion  148  
Appendix  151  
References  155  

CHAPTER 6  
Abstract  157  
Introduction  159  
Material and Methods  162  
Results  163  
Discussion and Conclusion  172  
References  174
LIST OF FIGURES

INTRODUCTION
  Figure 1  2
  Figure 2  4
  Figure 3  6

CHAPTER 1
  Figure 1  20
  Figure 2  21
  Figure 3  26
  Figure 4  27

CHAPTER 2
  Figure 1  47
  Figure 2  49
  Figure 3  53
  Figure 4  54
  Figure 5  57
  Figure 6  61
  Figure 7  63
  Figure 8  65

CHAPTER 3
  Figure 1  95
  Figure 2  96
  Figure 3  98
  Figure 4  101
  Figure 5  104
  Figure 6  109

CHAPTER 4
  Figure 1  124
  Figure 2  127
  Figure 3  131
  Figure 4  132

CHAPTER 5
  Figure 1  143
  Figure 2  145
  Figure 3  147

CHAPTER 6
  Figure 1  166
LIST OF TABLES

CHAPTER 2
  Table 1  47
  Table 2  48
  Table 3  55

CHAPTER 3
  Table 1  110
  Table 2  111

CHAPTER 5
  Table 1  144

CHAPTER 6
  Table 1  165
  Table 2  170
ACKNOWLEDGEMENTS

Chapter 1 is a version of:


Chapter 2 and 3 are based on:


I want to acknowledge the contribution of Clifford Brunk, Ken Jones, Jin Li, and Erik Avaniss-Aghajani for their useful comments to chapter 2 and chapter 3.

I would also want to acknowledge the contribution of Clifford Brunk to chapter 5.

Acknowledgment of Grants:
IGERT training program fellowship from NIH
California SEA grant
Mangasarian special fellowship

I would also like to acknowledge the assistance and guidance of my committee members, David Jacobs, Fred Fox, Robert Wayne, Matteo Pellegrini, and Daniel Blumstein.
VITA

June 15, 1969  Born, Tehran, Iran

1990-1992  Two years of Medical School
Tehran Medical University
Tehran, Iran

1998  B. S., Microbiology
California State University Los Angeles
Los Angeles, California

2001  California Board Certification
Clinical Laboratory Scientist
California Department of Health

2000-05  Awards and Grants
Mangasarian special fellowship, UCLA
IGERT training grant in Bioinformatics, UCLA

2001-03  Director of Information, Research, and Election
Graduate Student Government
University of California, Los Angeles

2002-05  Teaching Associate
Department of Ecology and Evolutionary Biology
University of California, Los Angeles

2005-06  Part-time Faculty, Lecturer
Department of Biology and Microbiology
California State University Los Angeles

2006-Present  R & D Scientist in Molecular Diagnostics
Medical Science Center Inc.
Glendale, California
PUBLICATION AND PRESENTATIONS


ABSTRACT OF THE DISSERTATION

Molecular Evolution of
Mitochondrial and Nuclear Genomes
in Ciliate Tetrahymena

by

Mike Mkhitar Moradian
Doctor of Philosophy in Biology
University of California, Los Angeles, 2007
Professor David Jacobs, Chair

Ciliates are single cell eukaryotic organisms with two nuclei and unusual genome biology. The purpose of this project is to study the molecular evolution of macronuclear (MAC) and mitochondrial (Mt) genomes of ciliate protist Tetrahymena (T.) to decipher and interpret their evolutionary processes. Tetrahymena species contain rapidly evolving mitochondrial (Mt) genomes, which have apparently diverged significantly from the ancestral pattern leaving half of their 44 genes without an assigned function. Thus I sequenced three Mt genomes of Tetrahymena paravorax, Tetrahymena pigmentosa, and Tetrahymena malaccensis via shotgun sequencing. I found unique features in these linear

x
~47kb long genomes including long telomeric repeats, long noncoding sequences, a pseudo-tRNA gene, unusual DNA secondary structure for *T.paravorax*, and the transcription initiation site, a GC box. More analyses were carried out using programs that calculate distance, nucleotide substitution (dn/ds), and their rate ratios (ω) to discover mutation hotspots and relaxed selective constraint in Ymf genes. Hence I concluded that the presence of mutation hot spots in *Tetrahymena* Mt genomes is a major reason and obstacle in identifying their function.

To study the MAC evolution I annotated a *Tetrahymena thermophila* chromosome, 1Mb in size, based on the translation of its possible open reading frames (ORFs) and their homologues in the GenBank. An objective of this study was to examine the TIGR automated annotation programs capability in identifying all of the biologically meaningful ORFs. Through searching the databases using BLAST I found proteins and ORFs with homologous domains that were not present in the TIGR automated annotation. Thus, I suggest phylogenetic footprinting to complement automated annotations. I also sought to reveal the extent of chromosome rearrangements between *T.thermophila* and the largest chromosome (LC) of *Paramecium (P.) tetraurelia*. Mapping all 463 putative proteins from the LC of *P.tetraurelia* to the chromosomes in *T.thermophila* genome revealed no orthologous chromosome. Interestingly, gene order between the Mt genome of these two ciliates was almost perfectly conserved, which suggests that the most plausible explanation for such an extensive gene order loss and rearrangement could be the developmentally programmed elimination of germ line sequences and randomization of MAC gene order.
INTRODUCTION

Tetrahymena Biology

Ciliates are a major evolutionary branch of eukaryotic protists composed of three primary lineages: Ciliates (e.g., Tetrahymena and Paramecium), Dinoflagellates (e.g., Symbiodinium), and the exclusively parasitic Apicomplexa (e.g., Plasmodium falciparum). Oligohymenophorea are a major class of ciliates that contain the famous families of Tetrahymenidae and Parameciidae. In Oligohymenophorea there is a ventral groove containing the mouth and distinct oral cilia, separate from those of the body (Hill, 1972). Tetrahymena (T.) cells are fairly large about 40-50 μm along the anterior-posterior axis with a considerable variety of specialized and complex cell structure.

T.thermophila (Figure 1) is a fresh water ciliate that is highly successful ecologically and inhabits streams, lakes, and ponds. T.thermophila exhibit evolutionary advances similar to features of multicellular eukaryotes such as vegetative growth in the diploid phase of their life cycle and internal fertilization through gamete nuclei exchange (Nanney, 1976). T.thermophila also demonstrates typical eukaryotic biology, its ultrastructure, cell physiology, development, biochemistry, genetics, and molecular biology have been extensively investigated. The following characteristics make T.thermophila a model eukaryotic organism to study:

*Its growth rate is <2h per generation, making it one of the fastest growing eukaryotes.
*It grows in a wide variety of media ranging from mono-bacterial cultures to defined chemical media, as well as a wide temperature range; 12°C to 41°C.

*It can grow to high densities over a wide range of scale.

*Its cells can maintain DNA constructs introduced by transformation.

*Its sexual stage, conjugation, can be induced at anytime with high efficiency.

*The availability of dozens of related species allows it to play important role in molecular evolution studies.

Figure 1 - *Tetrahymena thermophila*

The ciliates diverged earlier than plants and fungi in the evolutionary line leading to the vertebrates (Pace, 1997). *T.thermophila* exhibit evolutionary advances similar to features
of multicellular eukaryotes such as vegetative growth in the diploid phase of their life cycle and internal fertilization through gamete nuclei exchange. The possession of two related, but functionally differentiated genomes within one cell is a unique ciliate feature. The germline nucleus, micronucleus (MIC), has 5 chromosomes and is 210Mb in length. During a complex phenomenon called site specific fragmentation or chromosome breakage, about 15% of the MIC DNA is eliminated and the 5 chromosomes are broken into approximately 270 autonomously replicating pieces (ARP). The ARPs range from 50Kb to 2.2Mb. The ARPs comprise the somatic genome, which is contained in the polyploid macronucleus (Prescott, 1994).

**Nuclear dimorphism**

As is typical of ciliates, the nuclear apparatus of *Tetrahymena* is composed of two structurally and functionally differentiated types of nuclei, a phenomenon known as nuclear dimorphism. The micronucleus (MIC) is the germline, i.e. the store of genetic information for the sexual progeny. It is diploid and contains 5 pairs of chromosomes. No known genes are expressed in the MIC. Amicronucleate *Tetrahymena* cells (i.e., cells lacking a MIC) are frequently collected in nature, but in laboratory strains of *T. thermophila* the loss of the MIC leads to clone death. Only one viable laboratory-obtained amicronucleate cell line has been described. At cell division the MIC divides mitotically with kinetochores and intranuclear mitotic spindle. The macronucleus (MAC) is the somatic nucleus, i.e. the nucleus actively expressed during vegetative
multiplication. No known MAC DNA is transmitted to the sexual progeny. The MAC contains 250-300 autonomously replicating species derived from the 5 MIC chromosomes by site-specific fragmentation (Figure 2).

Figure 2. Developmentally programmed DNA rearrangement in *Tetrahymena*

Several chromosome breakage sites exist throughout the MIC chromosomes, which are essential in generation of the MAC chromosomes (Chau and Orias, 1996). During development of macronucleus DNA breaks at these sites and generates the MAC chromosomes by subsequent addition of the telomeres. Roughly 15%, mostly repetitive DNA, of the MIC DNA gets eliminated during this developmental process. The bulk of the MAC DNA pieces are present at the average level of 45 copies per MAC (Figure 2). There are no visible structures expected for mitotic distribution of MAC pieces, such as
kinetochores or mitotic spindle. The MAC is thus said to divide by amitosis.

Approximately half of the MAC DNA is distributed to each daughter MAC at cell division. Alternative allele copies of a locus segregate at random during MAC division. Physical methods are available to preparatively separate and purify MICs and MACs from one another (Yao et al., 2002).

*Life cycle*

The life cycle consists of an alternation of haploid and diploid stages (haplophase and diplophase, respectively) with reference to the germline. Cell reproduction is exclusively by binary fission; it is exclusively asexual and occurs only in the diplophase. It is remarkable that such highly differentiated cells can divide by binary fission. Cell division is accompanied by a variety of morphogenetic events that result in the development of duplicate sets of cell structures, one for each daughter. Conjugation is the sexual stage of the Tetrahymena life cycle, which includes no cell reproduction. During conjugation, two cells pair, form a temporary junction, exchange gamete nuclei and generate and differentiate the nuclear apparatus of their sexual progeny. The nuclear events of conjugation (Figure 3) normally include meiosis, gamete nucleus formation, fertilization and nuclear differentiation. It is remarkable that at the time of exconjugant separation (stage 7 in Fig. 3) there are five nuclei experiencing four extremely diverse fates, all within a common cytoplasm. Conjugation includes the only-and very brief-haploid stage of the life cycle; it follows meiosis and quickly ends at fertilization. The
haplophase is limited to a single nuclear division, without any cell division (Bettermier, 2004).

Figure 3- *Tetrahymena* life cycle.

*Why Tetrahymena thermophila?*

*T. thermophila* has a great potential for biotechnological applications and discovering unique and special features of the biology of ciliates. Here are some examples of selected contributions of this microorganism to research.

1) Research on general molecular biology problems:


- Eukaryotic telomeres: discovery of telomere molecular basis
(Blackburn and Gall 1978) and telomerase (Greider and Blackburn 1985); template role of telomerase RNA in *Tetrahymena* (Yu, *et al* 1990).

- Site specific chromosome diminution: mechanisms of diminution as well as understanding of epigenetic control and transposon relationships in *Tetrahymena* and other ciliates (Klobutcher and Herrick 1997). Thousands of site specific diminution events occur per haploid genome during MAC differentiation.

2) Applications related to nutritional, environmental, and occupational health sciences based on *Tetrahymena* as a eukaryote useful for rapid, reliable, sensitive, and inexpensive bioassays:

- Determining the protein nutritional value of human foods, based on similarity of nutritional requirements (Koehler *et al* 1987).

- Sensitive biosensor for biotoxin detection (Martin Gonzales *et al*, 1997)

- Monitoring Water Quality (Slabbert and Morgan, 1982).

3- Potential for biotechnological applications:

- Industrial production of enzymes and pharmaceuticals (Kiy *et al* 1996).

*Tetrahymena Genome Project (Initial Project)*

My initial research project was to identify the DNA sequence of ARPs that existed between 50-150 kb size range of *T. thermophila* macronucleus. ARPs in the 75kb region, assuming a practical number, would be completely sequenced in the first part of
this genome project. According to my primary assessments the region between 50-150kb should have had more than 10 ARPs, summing up to almost one million base pairs of DNA sequence. The supposition was that perhaps 2 or 3 ARPs would be in the 75kb region, which would allow me to complete my project. In the second part of this project, the completed ARPs would be annotated, and the annotations would be interpreted and confirmed. The goal was to use the complete genome sequences to find and analyze gene distribution and their corresponding protein functions. Furthermore, the intergenic regions would be searched for potential biologically significant elements. The entire sequence of the ARPs would be fully studied and analyzed in exhaustive detail to discover the organization of the *T.thermophila* macronuclear genome.

An appropriate strain of *T.thermophila* was selected, and was grown in bulk amounts. Cells were cast in agarose and run on a pulse field electrophoresis gel. The band that contained the 75kb long ARPs was excised and the DNA was extracted from the agarose gel. Since the excised band had more than three grams of agarose, I had to optimized the gel extraction protocol for the extraction of relatively minute amounts of DNA from a few grams of agarose. My developed protocols were successful and I was able to recover sufficient amount of DNA to prepare my clone libraries. A brief explanation of library preparation would be to shear the DNA and attach two sets of linkers to the polished ends, one for telomere identification and directional cloning and the other for all pieces. I was successful in preparing both clone libraries and starting the sequencing steps. After hundreds of clones were sequenced I could only identify short (3-5kb) assembled pieces, which came from *T.thermophila* genome. This affected my
decision to continue with this project since my resources were strictly limited and I concluded that there were more than 2 or 3 ARPs in the 75kb region. Later when the complete genome of *T.thermophila* was sequenced and published by The Institute for Genome Research (TIGR) my conclusion was approved. There are 133 ARPs present in 50-150kb region, from which 14 are in the 70-80kb region. Sequencing 14 ARPs 70Kb long each, excluding contamination from adjacent shorter and longer ARPs was well beyond my resources hence my decision to abandon this project was fairly justified.

Instead I switched my studies to the mitochondrial genomes of *Tetrahymena*, which were equally intriguing with several unanswered evolutionary questions. Meanwhile *T.thermophila* genome project was gaining grounds in the research community. By the time I finished sequencing the three mitochondrial genomes TIGR had generated the first draft of *T.thermophila* genome. This allowed me to work on the second part of my initial project, which was to analyze the ARP sequences using bioinformatics tools and complete a solid project on molecular evolution of *Tetrahymena* mitochondrial and nuclear genomes. The following will introduce the details of my work in the order it was performed.
Tetrahymena Mitochondrial Genome Project

Specific Goals and Objectives

Molecular Evolution

1. A primary goal of this proposal is to study the molecular evolution of divergent mitochondrial genomes of Tetrahymena through determining the complete DNA sequence for a number of Tetrahymena mitochondrial genomes, including T.paravorax, T.pigmentosa and T.malaccensis.

2. An objective of this work is to identify the function of the proteins coded in Tetrahymena mitochondrial genomes and search for sequences responsible for transcription and replication control in both previously sequenced genomes and those that will be sequenced in this project.

3. Evidence of additional examples of concerted evolution and telomeres maintenance and elongation will be sought as well as potential mechanisms responsible for these evolutionary processes.

4. The patterns of mitochondrial genome organization will be examined in detail to elucidate the mechanisms responsible for their generation and maintenance.
5. As a secondary objective, nucleotide and amino acid substitution types and patterns will be determined and quantified to be utilized to discover mechanisms responsible for failing us to identify function to many protein coding genes.

6. Also the phylogenetic relationships among the ciliates in the Tetrahymena genus and closely related taxa will be defined using complete mitochondrial genomes.

My project is comprised of a detailed and comprehensive study of molecular evolution of Tetrahymena mitochondrial and nuclear genomes. In this project I sequenced, assembled, and annotated the entire mitochondrial genomes of three Tetrahymena species, T.paravorax, T.pigmentosa, and T.malaccensis each 50kb long. During this sequencing project I got quite interested in evaluating the origin and evolution of mitochondria. This led me to examine the current hypotheses proposed for mitochondrial origin and its evolution, which presented convincing evidence for a monophyletic origin of mitochondria. Yet upon detailed review of literature it became clear that there were some genomic features in mitochondria that cannot be explained using the theory of monophyletic ancestry. Hence I wrote a review on the evolution of mitochondria where I discussed these unexplained genomic features that could not be justified with a single proteobacter ancestry. This review was the stating chapter of my dissertation and was published in the Journal of Scientific Review at UCLA in 2004.

Meanwhile I completed my sequencing projects and started the analysis of the three genomes. I found significant differences between these genomes such as presence
of a pseudo-tRNA, unusually long non-coding regions and telomeric repeats, which underscored the divergent nature of *Tetrahymena* mitochondrial genomes. But most importantly comparative genomic analysis revealed a transcription control region, a GC box in the middle of these genomes where bidirectional transcription initiates. I was also able to identify a putative sequence, which could be the origin of replication in the same region. In this project I encountered a few obstacles during cloning and completion of the gaps, yet I was able to overcome those problems by manipulating existing techniques and taking novel approaches such as direct cloning of sheared DNA. These sequencing projects, findings, and the differences are explained in chapter 2 of my dissertation.

A major question and a driving force for sequencing mitochondrial genomes was the presence of 22 genes (Ymf genes), roughly half of the genes in these mitochondrial genomes that remained without and assigned function several years after the first *Tetrahymena* mitochondrial genome was sequenced. I was eager to generate more data to study the nucleotide and amino acid substitution types and patterns and I was convinced that I would be able to demonstrate and quantify the sequence divergence among these genes. I was more compelled to conduct this analysis since I had a control group of conserved mitochondrial genes, the other 22 genes with known functions, which were under selective pressure. My analysis resulted in supporting my hypothesis that the Ymf genes contain mutation hotspots with accelerated rates of nonsynonymous substitutions. This finding could be a reason for our failure to assign functions to Ymf genes through sequence similarity and homology. This finding and its detailed analysis is presented in chapter 3, which along with chapter 2 created a comprehensive 12 pages publication in
the journal of PLoS One Evolutionary Biology earlier this year (2007),
doi:10.1371/journal.pone.0000650.

Another objective of this project was to assess the phylogenetic relationships
between *Tetrahymena* species using complete mitochondrial genomes. Ciliate nuclear
genome undergoes chromosome rearrangement and sequence elimination after
conjugation followed by rounds of asexual replication before the next conjugation. The
mitochondrial genome, on the other hand, divides asexually in every life cycle without
any complications, hence could produce accurate phylogenetic trees. Phylogenetic
relationships are usually drawn using fairly conserved sequences such as ribosomal DNA
however there is a growing interest in using mitochondrial genes and genomes to infer
relationships independent of nuclear variations. In fact when I inferred phylogenies of
*Tetrahymena* species using mitochondrial genes and genomes the generated topology
was slightly different than the only available topology in literature. I present these data
and conclusions in chapter 4, which concludes the mitochondrial genome part of my
project.


_Tetrahymena_ Nuclear Genome Analysis

Specific Goals and Objectives

1. A major goal of this project was to independently annotate a _T.thermophila_ chromosome and to evaluate the ability of automated annotation programs to identify all biologically important ORFs and control regions.

2. Also to provide analytical tools to demonstrate divergent nature of _T.thermophila_ protein coding genes and suggest alternative methods to infer homology.

3. Extending the project beyond _Tetrahymena_ genus to determine the gene order conservation between _Tetrahymena_ and _Paramecium_ macronuclear genomes for the first time since partial and complete genomic sequences for these two ciliates became available.

4. Analyzing the possibility of hybrid chromosomes in ciliate macronuclear genomes through searching for orthologous regions between _Tetrahymena_ and _Paramecium_ chromosomes.

The next two chapters of my project focus on the nuclear genome analysis. _Tetrahymena thermophila_ macronuclear genome was sequenced by TIGR and became available in 2005. Originally I wished to work on the annotation of this genome but it
was annotated using automated programs. Since ciliate genome has unique features and somewhat divergent genes evaluating its automated annotation seemed a worthy cause. I used homology to annotate the ORFs in a 1Mb chromosome of Tetrahymena thermophila genome and compare them to the automated annotation. I found 8 additional ORFs with homologous genes and 27 ORFs with homologous domains in the GenBank, which were missed by the automated annotation. Presence of ORFs with homologous domains in two non-homologous proteins suggests existence of divergent proteins in T.thermophila assuming that only the functional domain has been under sufficient selective pressure to indicate homology with other known domains. I also studied the relationship between E-value and Z-scores used in this analysis and proposed a range for these values where homology could be inferred. These analyses are presented in chapter 5 and were submitted to PLoS One Evolutionary Biology. Currently I am working on analysis of reviewer’s comments and revision request from the editors.

Finally I was able to analyze the genome rearrangement between two model ciliates Tetrahymena and Paramecium. The largest chromosome of Paramecium (P.) tetraurelia (1Mb in size) was published in 2004 and Tetrahymena thermophila genome in 2005, so I looked for gene order conservation between the P.tetraurelia chromosome and T.thermophila genome. Although these two organisms are distant relatives they are the only ciliates with complete chromosome sequences. They also contain strong gene order conservation between their mitochondrial genomes. I conducted this study to find out whether such conservation exists in macronuclear genome and whether chromosome breakage and genome rearrangement in these ciliates follow a known path. Since I found
no significant gene order conservation or shared orthologous regions between these two ciliates it became clear that they perhaps don't follow a known path during rearrangements. Hence I proposed that the reasons, which might be responsible for such reorganization, are the unique genome rearrangements and chromosome breakage sites that are not conserved in ciliates. This analysis was important since there are orthologous regions between bacterial genomes such as *E. coli* and the others, or even with mitochondrial genomes of some species, which diverged well before *Tetrahymena* and *paramecium* did. A very recent study on chromosome breakage sites provides data on chromosome breakage sites in *Tetrahymena* species, which strengthens my arguments. This analysis is the last part of my dissertation and is presented in chapter 6.

References

of the Extra-Chromosomal Ribosomal RNA Genes in *Tetrahymena*. Journal of
Molecular Biology 120:33-53.

Chau M.F., and Orias E. (1996) Developmentally Programmed DNA Rearrangement in
*Tetrahymena Thermophila*: Isolation and Sequence Characterization of Three New

Transferase Activity in *Tetrahymena* extracts. Cell 43:405-413.


Enzymes by Continuous High Cell Density Fermentation of the Ciliated Protozoon
*Tetrahymena Thermophila* in a Perfused Bioreactor. Enzyme Microbio. Technology
18:268-274.


Chapter 1

Evolution of Mitochondria

Is mitochondrial origin monophyletic?
MITOCHONDRIA: THE POWERHOUSE OF THE CELL

THE ORIGIN AND EVOLUTION OF MITOCHONDRIA

Cells in living organisms are classified into two major categories, eukaryotic cells (i.e., protist, fungi, animal, and plant) and prokaryotic cells (i.e., Bacteria and Archaea). Prokaryotic cells are primitive cells with one or two membranes, one circular chromosome, and quite simple protein synthesis machinery. Eukaryotic cells, on the other hand, contain a compact nucleus with multiple chromosomes, sophisticated cellular mechanisms, and, most importantly, membrane bound organelles such as chloroplasts and mitochondria. Mitochondria, small organelles that metabolize ingested nutrients and generate energy for the organism (Figure 1), are believed to have originated through a process called endosymbiosis. Prokaryotic cells do not have a dedicated organelle for energy production; instead they use their own cell components such as membranes to generate energy. Nevertheless, the process of energy production itself is similar to mitochondrial (Mt) function in eukaryotes. The purpose of this article is to describe and compare recent hypotheses regarding the origin of mitochondria.

What is the origin of mitochondria?

The most prominent theories explaining the origin of mitochondria are endosymbiosis, and chimeric prokaryotic ancestry. A third theory, the hydrogen hypothesis, explains the grounds and the mechanisms responsible for the endosymbiosis theory. Symbiosis is an association between two or more species living together, and occasionally this leads to endosymbiosis, the intracellular capture of the symbiont by
which two independently evolved organisms become a tightly coupled system and eventually just one organism.

Figure 1- Schematic illustration of mitochondria (picture from www.kathleensworld.com/mitochondria)

Endosymbiosis between a prokaryotic cell and a nucleated cell is a proposed hypothesis for origin of mitochondria. This hypothesis is widely accepted and there are several similarities between the mitochondria and some prokaryotes -specifically at the molecular level- that support this idea. Interestingly *Ricketssia*, a group of obligate intracellular parasites, which can cause diseases in several organisms, are among the closest known bacterial relatives of mitochondria. Endosymbiosis is also proposed to be the mechanism for creation of an early eukaryotic genome (Rivera and Lake, 2004). The eukaryotic nucleus is believed to have originated from the fusion of two diverse prokaryotic genomes, (i.e., Archaea and Bacteria). There are several molecular data that strongly suggest homology between eukaryotes and prokaryotes: they share many genes,
which contain similar DNA sequences and code for the same proteins. Researchers have suggested that the informational genes (those involved in transcription, translation, and related processes) in eukaryotic genomes are derived from Archaea and the operational genes (those involved in housekeeping) are derived from Bacteria (Rivera and Lake, 2004). Endosymbiosis is neither a trivial process nor a simple ingestion of the cells. Therefore, it is important to discuss how an important organelle such as a mitochondrion originated from a prokaryote, lost almost all of its original functions and specialized in energy generation. Besides, a thorough explanation of this theory may assist to further understand the origin of the mitochondria and the mechanisms that made it possible to create this organelle.

The endosymbiosis theory and the hydrogen hypothesis describe how a primitive prokaryote is engulfed and ingested by a primitive eukaryote to become the mitochondria in the final cell that is formed. The endosymbiosis theory further discusses whether this ingestion is either a direct or a serial endosymbiosis event; these two major mechanisms are called DET and SET respectively.

Figure 2- Direct and serial endosymbiosis theories.
In DET, which is supported by the “hydrogen hypothesis” (Martin, W., and M. Müller. 1998), the origin of mitochondria is placed very close-in the evolutionary time scale-to the origin of eukaryotic cell. The hydrogen hypothesis suggests that the driving force behind this process is metabolic syntrophy (nutritional dependency) where hydrogen producing bacterial symbiont is ingested by a hydrogen requiring Archaea (Figure 2). Therefore, a single direct endosymbiosis event may provide sufficient grounds to be considered as the mechanism for mitochondrial origin.

On the other hand, according to SET, the original prokaryotic host cell was an archaeabacterium (Margulis and Sagan 1987), which was neither photosynthetic nor capable of effectively using oxygen and obtained its energy through chemoosynthesis. This Archaea is considered to be an early primitive eukaryotic cell that played host to the mitochondrial symbiont, which already had a nucleus and the ability to ingest a smaller cell (Figure 2). Thus, this theory suggests that a primitive eukaryotic cell was created by an earlier endosymbiosis event between a bacterium and an archeum, and subsequent ingestion of another bacterium gave rise to the mitochondria.

An alternative theory to endosymbiosis is the chimeric prokaryotic ancestry of mitochondria. This theory suggests that the ancestor of animal mitochondria and the many primitive amitochondrial eukaryotes is a fusion microbe composed of a eubacterium and an archaeabacterium (Karlin et al., 1999). A Clostridium eubacterial species and a Sulfolobus archaeal species are the fusion prokaryotes in this theory. According to genome DNA data the sequences from these two genomes are moderately similar, therefore mutually compatible. These sequences are also moderately similar to
most mitochondrial sequences, whereas the genome sequence of other prokaryotes is quite distant from animal mitochondria, this may suggest that they may serve as the ancestral sequence for mitochondria. *Clostridium* species degrade the organic matter to acetate and H₂, however they lack the pathways for oxidation of acetate and energy conservation. These pathways are provided by *Sulfolobus*-like organism, where they use hydrogen and offer respiratory chain including quinones and cytochrome cofactors. *Clostridium* species also produce endospores, which may provide a means to establish nuclear and mitochondrial compartments. They also possess enzymes that are present in hydrogenosomes (a hydrogen-producing organelle with great similarity to mitochondria and with no genome). Hydrogenosomes are found in organisms that lack mitochondria such as a human genitalia parasite *Trichomonas vaginalis* (Bui et al., 1999). These organisms can be facultative anaerobes and use hydrogenosomes to break down pyruvate and produce hydrogen to generate energy. Hydrogenosomes share important similarities with mitochondria and therefore are believed to have a common evolutionary origin with mitochondria (Bui et al., 1996). When the *Clostridium* and *Sulfolobus* fuse, extensive amount of DNA is transferred between the bacterium and the archeum in this fusion cell, while forming a two-cell structure through septation. One of the septated cells engulfs the other cell in a process resembling an endospore formation, and encloses it in two membranes providing necessary structures for it to develop into mitochondria (Figure 3). This primitive mitochondriated eukaryotic cell under-different selective pressure-may have evolved into cells containing many mitochondria, cells where the mitochondria degenerates into hydrogenosome, and into amitochondriated cells (Karlin et al., 1999). In
summary, the chimeric prokaryotic ancestry suggests that the fusion of two prokaryotes
followed by exchange of their DNA gave rise to two different cells with exchanged
DNA. One of these cells was engulfed by the other one and eventually resulted in three
different cell types mentioned earlier.

These two theories are quite distinct from one another in suggesting different
mechanisms responsible for mitochondrial generation. The endosymbiosis theory
suggests a relatively more simple mechanism for the mitochondrial origin compared to a
more sophisticated process in chimeric ancestry. Although both theories are supported by
molecular data, the endosymbiosis theory seems to be less costly (e.g., time and energy)
and is favored more by researchers in this field (Burger et al., 2003).

A monophyletic origin for mitochondria?

Proponents of the endosymbiosis theory also favor a monophyletic origin for
mitochondria, meaning that all mitochondrial genomes are descendants of one bacterium
(an α-proteobacterium) via a single endosymbiosis event (Gray et al., 2001). The
evidence supporting this contention comes from molecular data. First, majority of the
genes obtained from sequencing the mitochondrial genome of Reclinomonas americana
(a Jakobid flagellate protist) could be found in bacteria, especially in α-proteobacterium.
This organism contains 97 genes the highest number of genes present in any
mitochondrial genome (Lang et al., 1997), and almost all of the mitochondrial genes with
assigned function coming from different organisms are a subset of genes in R.
americana. Second, in many instances the mitochondrial gene order is so similar to that
of their bacterial homologs that it could assist in assigning function to some divergent
genes that have retained a similar order. Third, the small subunit ribosomal RNA (SSU-
rRNA), which is largely used for phylogenetic analysis due to its conservation and
availability in almost all genomes, or a few conserved concatenated genes-tandem
attached sequences- from several mitochondrial genomes form a monophyletic clade and
support a common ancestor (Gray et al., 2001). These are compelling evidence to accept
a common ancestor for mitochondrial genomes. However, there are a few facts that may
argue against a monophyletic ancestry for the mitochondrial genomes, meaning that
mitochondrial genomes in different eukaryotic lineages may be result of more than one
independent endosymbiosis event. First, despite conserved gene order in some
mitochondrial genomes there is extensive genome reorganization in many others, which
is difficult to justify through single ancestry. Second, there is an extraordinary diversity
in size of the mitochondrial genomes, where they range from 6kb (kilo base pairs) in
protist plasmodium to several hundreds of thousands bp (base pairs) in plants, (i.e., rice
at 490 kb) (Burger et al., 2003). Third, despite the fact that some mitochondrial genomes
have retained a gene order similar to certain bacteria, there are several genomes that have
very different gene orders. Fourth, mitochondrial genomes contain genes that are longer
and proteins that possess more subunits than the proposed α-proteobacteria (Schnider
and Eebrt 2004; Marx et al., 2003). Finally, the presence of introns and the modification
of the genetic code in mitochondrial genomes require fairly sophisticated explanations
under the monophyletic origin hypothesis. Considering an extensive diversity, multiple
ancestry for mitochondrial genomes remains an open possibility. Hence, studying
mitochondrial genome diversity shall assist one to categorize them into similar groups and try to look for alternative ancestors.

**Mitochondrial genome diversity**

Mitochondrial genomes in different species vary greatly in their size and structure, gene arrangement and content, growth and shrinkage rates, and incorporation of foreign DNA (Palmer et al., 2000). Mitochondrial genomes usually consist of a single piece of DNA or chromosome, which may be circular or linear (Figure 3 and 4). In rare cases mitochondrial genome can have multiple linear (in a single cell protist *Amiboidum parasiticum*) or multiple circular (in fungus *Spizellomyces*, and animals *Globodera* and *Dicyema*) DNA coding for their genes (Burger et al., 2003). Since prokaryotic genomes are circular, a linear mitochondrial genome should have arisen during Mt genome evolution. Several organisms from various taxonomical groups such as ciliates, parasitic

![Tetrahymena mtDNA](image)

![Human mtDNA](image)

![Spizellomyces mtDNA](image)

![Amoebidium mtDNA](image)

Figure 3- Mitochondrial genome architecture (Burger et al., 2003).
apicomplexans, algae, slime molds, and fungi have linear Mt DNA. Since linear and circular Mt genomes share the same protein sets in their replication machinery, the mechanism responsible to generate a linear genome should be unrelated to the replication process. The proposed mechanism is that mobile genetic elements or plasmids (small circular DNA) present in mitochondria integrate into circular genome and result in formation of linear Mt DNA. Such sequences may very well serve as templates for telomerase and eventually become the telomeres in linear Mt genomes (Nosek and Tomaska, 2003). There are enormous differences in Mt genome sizes ranging from 6kb in *Plasmodium* (causing agent of malaria) to over 2Mb (million base pairs) in plants. Size variation occurs even in closely related species such as fission yeasts where the Mt genome ranges from 19.5 to over 80kb. Despite the fact that the Mt genomes tend to

Figure 4- A- Size and genome content for Rickettsia and mitochondrial genomes of different species, red for genes with known function, blue for unknown functions (Gray et al., 1999).
shrink their sizes over evolutionary time, which is typical in most of the eukaryotic phyla (size range 15-70kb), the plant Mt genomes seems to favor the accumulation of DNA through duplication and multiple transfers from nuclear DNA to the mitochondria (Palmer et al., 2000). Eukaryotic Mt genomes usually contain minute amount of non-coding DNA with the exception of plants, which contain more. They have extremely different gene content ranging from five genes in *Plasmodium* to almost one hundred in *Reclinomonas americana*. On average eukaryotic Mt genomes contain between 40-50 genes and the number of genes do not seem to correlate with the genome size. Duplication and reshuffling processes are responsible for radical differences in gene arrangement in Mt genomes even among closely related genera. The Mt genomes of ciliates *Paramecium* and *Tetrahymena* share over 80% of their genes, yet their gene arrangements are so distinctively dissimilar that in very few cases two or more shared genes are juxtaposed adjacent to one another due to a major inversion.

Protists are thought to be among the first organisms that acquired mitochondria, therefore information obtained from protist Mt genomes may assist in better understanding of their origin. The protist Mt DNAs are on average 40kb long and are more than 70% of its nucleotides are A+T. They have both linear and circular Mt DNA and contain the highest number of genes present in any eukaryotic Mt genome. Protists and plants share several genes and are very similar in gene content, despite the fact that plant Mt genomes are considerably larger. The transfer RNAs and the ribosomal RNAs in protist Mt genomes are quite similar to their proposed cubacterial ancestors. Although a few protists code a minimal tRNA set sufficient to recognize all the codons, majority of
them contain a few tRNA genes. This shortage is compensated by tRNAs imported from cytoplasm, which are encoded by the nucleus. Protists use either standard or the modified genetic code where UGA, which codes for a stop codon in standard code, is used to specify the amino acid tryptophan. Interestingly photosynthetic protists and plants use standard genetic code and most of the non-photosynthetic protists use the modified code. Protists rarely contain introns, however the Mt genomes of some photosynthetic green and red algae have introns, which are also prevalent in plant Mt genomes (Gray et al., 1999). Such similarities between photosynthetic protists and plants may suggest a common ancestry for their Mt genomes.

Plant Mt genomes have an unusually dynamic structure due to frequent rearrangements where recombination between repeated sequences results in accumulation of non-coding DNA. Plant Mt genomes have relatively lower rates of nucleotide substitutions than animal and fungi, yet they have the largest (up to 2.4Mb) Mt genomes. This enormous genome size is result of continuous duplication, recombination, and import of foreign DNA, which does not affect on the complexity of these genomes. Therefore the C value paradox, which suggests that the size of their genome does not correlate with the number of genes present in them, is a valid argument for plant Mt genomes. Although they have fewer genes than many protist Mt genomes, they still contain more genes than their animal counterparts. Animal Mt genomes are typically a small (15-20kb) and circular DNA, which contain 13 trans-membrane proteins, small and large ribosomal RNAs, and up to 22 tRNA genes. They have relatively higher rates of nucleotide substitutions, yet their gene orders are fairly
conserved and even identical among mammals. However, they have a few differences such as presence of introns, linear genomes, and small rearrangements, which are most likely acquired individually in separate occasions (Gray et al., 2001). The fungal Mt gnomes are more diverse than their animal counterparts in size, structure and gene content. More importantly fungal Mt genomes contain open reading frames with putative genes that have no apparent homolog in the GenBank, hence remain without an assigned function. These genomes also contain several introns, variable numbers of protein coding genes (including ribosomal proteins) and RNA genes, and translation codes. Despite such variety the common set of the fungal Mt genes are not dramatically different from that of the animals and the differences can be explained through independent mechanisms. Although, the enormous diversity among Mt genomes could be explained through often sophisticated mechanisms, it may be simply described if a prokaryotic genome is found which may provide common grounds for these differences once new genomic sequences become available.

**Differences in mitochondrial gene length and protein structure**

In symbiotic bacteria gene lengths and genome size reductions are a consequence of their intimate symbiotic lifestyle (e.g., symbiotic bacteria of legumes). Based on endosymbiosis theory mitochondria can be considered obligate intracellular symbionts, therefore reduction in their genomes may occur naturally. Mt genes such as small subunit ribosomal RNA, cytochrome c oxidase subunits I-III, and cytochrome b are used to infer phylogenetic relationships, conservation, length variation, and the rate of substitutions
among organisms. If reductions in genome size and gene length, which are prevalent in several mitochondrial genomes, are general features of obligate intracellular symbiont, then mitochondrial genes may become smaller than their putative ancestor. An advantage of this process is that shorter gene lengths may decrease the possibility of the accumulation of slightly deleterious mutations in the mitochondria. Recently, the shortening of the mitochondrial gene lengths relative to their proposed α-proteobacter ancestor is used to demonstrate a single origin for mitochondria (Schnider and Ebert, 2004). Their results suggest that conserved genes such as Cox I-III and Cob, with few exceptions, are shorter than α-proteobacteria. This is only true for mitochondrial genomes shorter than 20 kb in size, where in genomes longer than that Cox I-III genes do not support such an argument. In addition certain algae, plants, and fungal species have longer mitochondrial small and large ribosomal DNA genes than their bacterial counterparts. It has been shown that in yeast the transfer of DNA from mitochondria to the nucleus is $10^5$ times more frequent than in the opposite direction (Thorness and Fox 1990). Hence, the shortening of the genes and the genome size may be due to the transfer of mitochondrial DNA to the nucleus, and the reverse process appears to be an unlikely event. These findings argue for a single ancestry for mitochondrial genomes of certain species, and do not encompass an overwhelming majority. Unless we accept that some conserved genes in smaller mitochondrial genomes have become shorter, which might be evolutionarily advantageous, yet ribosomal DNA genes which are part of the same genome have acted contrary and added to their length for unknown reasons, we may consider an alternative unknown mitochondrial ancestor for some species.
The bc₁ complex of mitochondria is a dimeric enzyme, which catalyzes reactions to generate a proton gradient to produce ATP in mitochondria. In most bacteria (including α-proteobacteria) this complex is comprised of cytochrome b and c₁ and an iron-sulfur protein (Crofts and Berry 1998). However this complex from very different eukaryotes (i.e., potato and yeast) contains seven additional subunits (Marx et al., 2003). The monophyletic hypothesis suggests that the Jakobids (a primitive eukaryote) are evolutionary intermediates between the α-proteobacteria and the higher eukaryotes. Obviously the bc₁ complex does not support this view where this complex is almost the same in Jakobids and fungi, plants, and mammals, which includes at least five more subunits than the α-proteobacteria complex. The proposed mechanism for this discrepancy is that the migration of the mitochondrial genes to the nucleus may be the underlying cause for recruitment of additional subunits into the mitochondria (Marx et al., 2003). Although some protein and RNA molecules, which are encoded in nucleus, transfer from cytoplasm to mitochondria, the reasons for such transfers shall be explained especially if the same proteins from the ancestral genome do not require the additionally transferred subunits. A complicated mechanism like this requires further analysis and more empirical data since it argues directly against what is suggested in α-proteobacteria as the Mt origin.

*Unique mitochondrial genes with unknown origin*

In yeast mitochondrial genomes the proteins could be categorized into prokaryotic-specific (50-60%), eukaryotic-specific (20-30%), and organism-specific or unique
proteins (20%). Prokaryotic-specific proteins have counterparts in prokaryotic genomes, eukaryotic-specific proteins could be found in eukaryotic genomes but not in prokaryotic genomes, and organism-specific proteins are so far unique to *S. cerevisiae* (Gray et al., 2001). Other organisms such as protists also contain organism-specific genes. In ciliates especially *Tetrahymena*, which have divergent genomes, almost half of the genes remain without a known counterpart or homolog and also without an assigned function (Brunk et al., 2003). It has been suggested that unique genes are so diverged that based on available search tools and techniques no homolog can be found for them. Hence, sequences for all 44 genes of mitochondrial genomes from five *Tetrahymena* species were analyzed to infer relative selection on unique genes. Recently we showed that the majority of the unique genes are more diverged than the genes with assigned function (Moradian et al., 2007). However, there are a few unique genes that are very well conserved and yet still no homolog could be found for them in the GenBank. This may suggest that some of these unique genes, being in yeast, ciliates, or any other organism, which are under strong selection, might have a different, unidentified origin than a proposed common *α*-proteobacter ancestry. Horizontal gene transfer, which means genes transferred from non-ancestral organisms, is the mechanism proposed for a unique gene in any mitochondrial genome. Although horizontal gene transfer may explain the presence of unknown genes in mitochondrial genomes, the number of such transfers requires additional explanation.
**Why do mitochondria have introns?**

Several species of animals, plants, fungi and some species of protists have introns (groups I and II) in their mitochondrial genomes. The difference between groups I and II introns, is in their different secondary structure and folding process. This is particularly evident in plant mitochondrial genomes. In angiosperms intron type I appears in 48 species in the same region of CoxI gene (Palmer et al., 2000). Phylogenetic analyses of these species using the intron sequences are incongruent when compared to the same analysis using the small ribosomal RNA sequence. However, when a variable sequence such as intron is compared to a highly conserved gene such as SSU-rRNA, incongruent relationships should not be surprising. Nevertheless they propose that extensive horizontal gene transfer is responsible for the distribution of class I introns in angiosperms. This is in striking contrast to the vertical spread of type II intron in plants, which share a common ancestry gene (Palmer et al., 2000). Although, this and other proponents of a single ancestral origin for mitochondria suggest that introns are result of horizontal transfers and they have not descended vertically from a α-proteobacterium (Burger and Lang, 2003), they have not explained how feasible are so many independent horizontal transfers. The apparent path of gene transfer is from mitochondria to nucleus and the unorthodox transfer of genes from nucleus to mitochondrion is quite rare. A common progenitor with an intron, if discovered, would be a more acceptable explanation for the presence of introns in some mitochondrial genomes than the idea of several independent horizontal transfers. However, a detailed phylogenetic study of introns would reveal whether they were present in the common ancestor or they were
acquired later during evolution. The same argument holds for the changes in the genetic code (translation code). Some animal, fungi, and protist mitochondrial genomes use the standard stop codon TGA to code for amino acid tryptophan, and ATA for methionine instead of isoleucine. These reassignments would have little impact on the coding of the genes in Mt genomes since TGA is the least abundant stop codon and it can be readily be read by the tryptophan t-RNA anticodon due to the uridine in its first position (wobble position). Yet, photosynthetic algae, which are believed to be the ancestors of plant cells, and plants, use standard genetic code (Burger and Lang, 2003). It is not evident that why algae and plants have conserved the standard genetic code while many other species have departed from it.

Summary and Conclusion

In this study I present theories that explain the origin of the mitochondrial genomes in eukaryotic cells. Several studies have suggested endosymbiosis as the main mechanism that gave rise to the mitochondrial genomes and some argue that a single endosymbiosis event is responsible for their origin. Undoubtedly, the suggested α-proteobacterium could be an ancestor for the mitochondrial genome. However, there are several very important and extreme differences between mitochondrial genomes that may not be explained by available data and tools. Either mitochondrial genomes have more than one ancestor, which has not been identified yet or we do not possess the information and the techniques to decipher the reasons for such extraordinary differences. In either situation
sequencing more prokaryotic and mitochondrial genomes will provide more insights into resolving the question of the mitochondrial origin.

References


Chapter 2

Complete sequence of mitochondrial genomes of three *Tetrahymena* ciliates:

*Tetrahymena paravorax, Tetrahymena pigmentosa, and Tetrahymena malaccensis:*

Molecular evolution of *Tetrahymena* mitochondrial genome
Abstract and Summary

To help understand the molecular evolution of mitochondrial (Mt) genomes of ciliate protist *Tetrahymena (T).*, we sequenced three Mt genomes of *Tetrahymena paravorax, Tetrahymena pigmentosa,* and *Tetrahymena malaccensis.* These linear genomes are ~47kb long and have unique features that did not exist in two previously sequenced Mt genomes of *Tetrahymena thermophila* and *Tetrahymena pyriformis.* Interspecies comparisons did not reveal any different genome organization or gene order, however there were a few distinct variations, especially in *T.paravorax,* which underscored highly variable nature of *Tetrahymena* Mt DNA. These features included long telomeric repeats, long noncoding sequences, a pseudo-tRNA gene, and an unusual DNA secondary structure for *T.paravorax.* Comparative genomic analysis reveals a transcription control sequence, a GC box, located in the longest intergenic region between Cob and ymf77. In *T.paravorax* the telomeres are the same 64 bp repeats (longest known for a *Tetrahymena* Mt DNA) at either end and are multimodal 6 ± 3kb long. Also two long non-coding sequences, 495 bp at the 5′ end and 210 bp at the 3′ end, are located between the telomeres and the inverted conserved repeats. The 5′ long non-coding region has a Lysine pseudo-tRNA 118 bp away from the telomere. The longest protein ymf77 has a 1kb region in *T.paravorax* that makes strong DNA secondary structure with the highest energy level (lowest ΔG) among all genomes. *Tetrahymena Mt* genome contains 22 putative proteins of unknown function (Ymfs). Sequence similarity searches of the Ymfs reveal similar domains from different mitochondrial proteins in the database. These similarities are between ymf61 and yeast ribosomal VAR1 protein, ymf75
and atp6, and ymf77 with plasmodium and Reclinomonas RNA polymerase subunits. There were also two very short, 33 and 44 amino acid long, conserved ORFs in intergenic regions of the genes at the 5' end of the genomes, which did not show any similarity to any other sequences in the GenBank.

Introduction

Eukaryotic cells obtain their energy through a process called oxidative-phosphorylation, which is carried out in mitochondria. These organelles are apparently the product of one or a few endosymbiotic events early in the lineage leading to eukaryotic cells. Generally, the number of mitochondrial genes per genome range from 97 genes in Reclinomonas americana to 1 gene in Spizellomyces punctatus (Gray et al., 1999). Since mitochondria is specialized in oxidative-phosphorylation the vast majority of its genes are involved in electron transport pathway or in ribosomal protein synthesis. The large and small subunit ribosomal RNA genes (LSU and SSU rRNA) are almost always present. Transfer RNA genes (tRNAs) in mitochondrial genomes range from a complete set required for translation to none in apicomplexan parasites, which should be supplemented with nuclear tRNAs (Chiu et al., 1975; Suyama, 1982; Rusconi & Cech, 1996; Esseiva et al., 2004). The genes found in mitochondria can be categorized into 6 groups: 1) ribosomal RNAs, 2) tRNAs, 3) ribosomal proteins, 4) proteins involved in oxidative-phosphorylation, 5) other proteins (not ribosomal proteins and not directly involved in oxidative-phosphorylation) and 6) open reading frames, coding for putative proteins, for which homologues of known function can not be identified. Among the
genes of complete mitochondrial genomes roughly 35% are tRNA genes and about 5% are rRNA genes (Gray, et al., 1998).

Mitochondrial genomes in diverse lineages, despite their commonly accepted α-proteobacter ancestry show extensive variability in size and structural organization. Ciliate Mt DNA is no exception and is among the most rapidly evolving Mt genomes (Gray et al, 1999). The genomes of ciliates are particularly interesting since they are fairly divergent from all other genomes with rapid evolution of their DNA at the primary sequence level and genome content. A good example of rapid evolution of the ciliate Mt genomes is the structural and organizational differences between *Paramecium* and *Tetrahymena* (Burger et al, 2000). Furthermore, roughly half of the putative genes in two previously sequenced Mt genomes of *T. thermophila* and *T. pyriformis* that showed complete synteny with regards to gene arrangement still remain without an assigned function (Burger et al, 2000, Brunk et al. 2003) where by comparison, in *R. americana* also a protist, 97% of the mitochondrial genes can be assigned a function (Lang, et al., 1997). The complete mitochondrial genomes of 3 ciliates, *Paramecium aurelia*, *Tetrahymena pyriformis* and *Tetrahymena thermophila* have been determined (Prichard et al.,1990; Burger et al., 2000; Brunk et al. 2003). Several putative proteins, in these rapidly diverging Mt genomes, do not have sufficient sequence similarity to a protein of known function to insure homology. However, Brunk et al. 2003 suggest that if sequence similarity to another protein indicates homology and this second protein is homologous to a protein of known function, homology and assignment of function can be inferred though a chain of sequence similarity (Brunk et al. 2003). Thus, it is an appalling task to
try to clarify the mechanisms and the extent of this evolutionary diversification. A desirable approach to solve such evolutionary questions is comparative Mt genome analysis (Gray et al, 1998). Since evolution continually tinkers with genome sequence and tests the outcome through natural selection, this approach should provide a detailed perspective on the genome content, gene arrangement, intergenic and control regions, and other relevant information by the virtue of having conservation across related species. All of the ORFs in the two previously sequenced Tetrahymena mitochondrial genomes have sequence similarity strongly indicating homology. Yet among hypotrichous ciliates a remarkably high degree of divergence is apparent. Only half of the ORFs between Paramecium and Tetrahymena, for which function cannot be assigned, have sufficient sequence similarity to strongly indicate homology. Thus, almost a dozen of these ORFs in Paramecium and Tetrahymena are not unambiguously homologous.

One of the most interesting characteristics of the Tetrahymena Mt genomes is the sequence of their telomeres, which not only differ from the nuclear telomere sequences (i.e., C4A2) they also differ from each other (Morin & Cech 1986, Morin & Cech, 1988b, Katzen, et al., 1981). The telomeres of a number of Tetrahymena species have been characterized and shown to range in length from 31 to 53 base pair (bp) repeats, which could have multimodal distribution (Morin & Cech, 1988b and 1988c). Ciliate telomeres are dynamic short nuclear structures, which are elongated and maintained by telomerase activity (Greider & Blackburn, 1985; McEachern et al., 2000). Such mechanism may not be easily applicable to the maintenance of Mt telomeres since they contain long telomeres (31-53 bp) and consequently would require quite long internal
guide RNA sequence. This problem is also present in yeast, which has even longer
telomeres and face the same challenges for their elongation and maintenance (Nosek, et
al., 1998; McEachern & Blackburn, 1994). Morin and Cech (1986) proposed that a
possible mechanism for Mt telomere maintenance could be unequal crossing over during
recombination. In T.pigmentosa, T.hyperangularis, and T.hegewischi the telomeric
repeat sequences are different on the two ends of the mitochondrial genome (Morin &
Cech, 1988b). This additional variation among ciliate Mt genomes further underlines
their divergent nature and rapid molecular evolution.

Concerted evolution is another interesting phenomenon, which occurs in
Tetrahymena mitochondrial genomes. Concerted evolution is a mechanism that causes
different copies of the repeated genes to be homogenized within each species while
allowing them to diverge between species. Several examples of concerted evolution have
been introduced over the last three decades. In many plants and animals genes encoding
ribosomal RNAs are present in hundreds of copies and arranged end to end in long
tandem arrays in one or a few places in the genome (Brown, et al., 1972; Dover, 1993;
Elder & Turner, 1995). Terminal repeats in Tetrahymena mitochondrial genomes studied
so far include either portions or all of the large subunit ribosomal RNA genes
(LSUrRNA). Further the LSUrRNAs are separated into two units (i.e., a and b) usually
& Cech, 1988a, Brunk, et al. 2003). The a unit of LSUrRNA is 5.8S-like and about 280
bp in length, while the b unit is 21S-like and about 2315 bp in length and they are
generally separated by a lucine tRNA (tRNA^{51}). This also occurs in the terminal-repeat
regions of *T. thermophila* or *T. pyriformis*. These repeats are 2695 bp in length in both species and there are almost no differences between the 5' and 3' repeats within either species. However an 8.5% divergence exists in these sequences between the two species, which indicates a classic case of concerted evolution. A likely mechanism responsible for concerted evolution of small gene families is gene conversion, which allows a mutation that occurs in one copy of a repeated gene to spread to other copies of the gene as well (Graur D. & Li W. H., 2000).

*T. paravorax* was the first species that diverged from the *Tetrahymena* complex (Sadler & Brunk, 1992). The addition of *T.pigmentosa* and *T.malaccensis* allowed us to cover closely, intermediary, and distantly related species in this lineage. In this study we describe three linear Mt genomes of *T.paravorax*, *T.pigmentosa*, and *T.malaccensis* each about 47kb long, which contain 22 known genes, 22 putative proteins, 7 different tRNAs, and 2 psuedo-tRNAs in case of *T.paravorax*. These genomes have complete synteny with previously annotated Mt DNA of *T.pyriformis* and their genes are homologous to those of *T.pyriformis* (Burger et al., 2000). Comparative genomics of Mt DNA sequences from five species of *Tetrahymena* allowed us to find a transcription control region common to all these genomes. We also report a comprehensive study on positive selection, relative rate of evolution, and amino acid and nucleotide substitution patterns among *Tetrahymena* species in the next chapter. The purpose of this study is to show the amount of diversity in the Mt genomes of ciliated protozoa, genus *Tetrahymena*. 

44
Materials and Methods

*T. paravorax* (GenBank Acc. # DQ927304), *T. malaccensis* (GenBank Acc. # DQ927303), and *T. pigmentosa* (GenBank Acc. # DQ927305) species were obtained from ATCC and were grown in PYG medium (Orias & Bruns, 1976; Nanny & Simon, 2000). Isolation of mitochondrial DNA was conducted based on a protocol presented by Netrawali where mitochondria are isolated initially to minimize genomic DNA contamination during the DNA extraction (Netrawali et al., 1974). Extracted mt DNA was sheared to fragments of approximately 2-3kb long using a hydro-shear point device (Thorstenson et al., 1998). The ends of these fragments were filled with Taq DNA polymerase in an extension only reaction, which resulted in an adenine overhang used for direct TA cloning in pCRII vector (Hu et al., 2003; Invitrogen Inc). The clones were sequenced from both ends using M13 reverse and forward primers (~ 500 bp each) and the resulting sequences (about four fold coverage) were assembled into contigs using Phred, Phrap, and Consed (Ewing & Green, 1998; Ewing et al., 1998; Gordon et al., 1998). The gaps between the contigs were filled using PCR amplified products with primers for sequences flanking the gaps. The two long inverted repeat regions were PCR amplified using specific primers for telomeres and beginning of the Ymf57, the gene adjacent to inverted repeat region. The PCR products were cloned and sequenced using walking primers to obtain accurate consensus sequences for each repeat.

**Gene Arrangement, Homology, and Similarity Search**

Complete mt genomes of *T. paravorax*, *T. pigmentosa*, and *T. malaccensis* were aligned with the sequence of *T. pyriformis* and *T. thermophila* at both DNA and amino acid levels.
using CLUSTAL X (Thompson et al., 1997). The degree of similarity was shown using similarity metric (z score) calculations on ORFs of mt genomes of *T. paravorax*, *T. pyriformis*, and *T. thermophila* (Doolittle, 1990; Pearson, 1990; Mount, 2001). The z-scores were calculated via software that I developed in our laboratory (available upon request). In this program, the amino acid sequences of the ORFs are aligned using a global alignment (Needleman-Wunsch algorithm) to obtain an alignment score (AS). The amino acid sequences are then randomized using a shuffle algorithm and a random AS is obtained. The program repeats this process for 1000 times and calculates a mean and a standard deviation (SD) for the randomized AS distribution. This distribution approximates a Gumbel extreme value distribution (Mount, 2001). The z-score for two ORFs is the difference of the original AS and the mean of the randomized AS distribution divided by the SD for that distribution. A z-score of >6 is a strong indication of homology (Pearson, 1990; Mount 2001). The BLAST network services provided at the National Center for Biotechnology Information, tRNAscan-SE server, tools at the European Bioinformatics Institute, and 3D protein structure prediction servers were used for sequence similarity searches (Altschul et al., 1990, Lopez et al., 2003; Kelley et al., 2000).

Results and Discussion

We sequenced the entire Mt genomes of *T. paravorax*, *T. pigmentosa*, and *T. malaccensis*.

These Mt genomes, along with the two previously sequenced Mt genomes of

*T. pyriformis* and *T. thermophila*, are ~47kb long, have high A+T content, and contain 44
genes from which 22 could be assigned function (Table 1). The remaining 22 are putative proteins with either similar domains to known proteins or no similarity at all.

<table>
<thead>
<tr>
<th>Species</th>
<th>Mitochondrial genome Length(bp)</th>
<th>22 proteins with function No. of AA (*)</th>
<th>LSU/RNA Length(bp)</th>
<th>SSU/RNA Length(bp)</th>
<th>Putative control region Length(bp)</th>
<th>A+T(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. paravorax</td>
<td>47,368</td>
<td>81.5</td>
<td>5642</td>
<td>78.9</td>
<td>2572</td>
<td>74.2</td>
</tr>
<tr>
<td>T. thermophila</td>
<td>47,577</td>
<td>79.2</td>
<td>6316</td>
<td>75.1</td>
<td>2592</td>
<td>72.2</td>
</tr>
<tr>
<td>T. pyriforis</td>
<td>47,172</td>
<td>78.7</td>
<td>6487</td>
<td>76.4</td>
<td>2595</td>
<td>73.1</td>
</tr>
<tr>
<td>T. malacensis</td>
<td>47,452</td>
<td>80.1</td>
<td>6315</td>
<td>78.5</td>
<td>2595</td>
<td>72.4</td>
</tr>
<tr>
<td>T. pigmentosa</td>
<td>47,209</td>
<td>81.5</td>
<td>6204</td>
<td>78.2</td>
<td>2597</td>
<td>74.7</td>
</tr>
</tbody>
</table>

Table 1- Length and A+T content of *Tetrahymena* Mt genomes. Length is in base pair, and A+T in percentages. (*) indicates 22 coding genes with known (assigned) functions.

Genome organization and gene arrangements in all five genomes were the same with the exception of an additional putative pseudo-tRNA lysine (K) in *T. paravorax*, which was located between the 5’ telomere and the inverted repeat region (Figure 1).

![Gene map of the T.paravorax Mt genome.](image)

Figure 1. Gene map of the *T.paravorax* Mt genome.

Arrows denote direction of transcription. Intergenic region between *Cob* and *Ymf77* genes is considered to contain a putative control region. The *trnK* represents the pseudo tRNA in *T.paravorax* Mt genome.

Pseudo-tRNAs are pseudogenes that cannot form cloverleaf shape structures but can form stem loop (Figure 2). Putative Pseudo-tRNAs in *T.paravorax* Mt genome are located at both ends of the genome in the long non-coding sequences between the
telomeres and the large ribosomal subunit. These non-coding sequences are 495 bp at the 5' end and 210 bp at the 3' end (Table 2). The lysine pseudo-tRNA is 118bp away from the telomere. Presence of pseudo-tRNA genes has been reported in Mt genome of plague *Thrips imaginis* (Thao et al. 2004). The non-coding sequences are absent at the 3' end of the *T.pigmentosa* Mt genome and are only 31bp at the 5' end. None of the other three species have longer than 65bp non-coding sequences before the telomeres (Table 2). Thus the presence of these long non-coding regions with their pseudo-tRNAs in *T.paravorax* suggests that it perhaps has the most divergent Mt genome in the *Tetrahymena* complex.

<table>
<thead>
<tr>
<th></th>
<th>Telomere Repeat Length (bp)</th>
<th>Non-coding Sequence Length (bp)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5' end</td>
<td>3' end</td>
</tr>
<tr>
<td>T. paravorax</td>
<td>64</td>
<td>495</td>
<td>210</td>
</tr>
<tr>
<td>T. thermophila</td>
<td>52</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>T. pyriformis</td>
<td>31</td>
<td>49</td>
<td>34</td>
</tr>
<tr>
<td>T. malaccensis</td>
<td>52</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>T. pigmentosa</td>
<td>50(5') or 37(3')</td>
<td>31</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: The length of mitochondrial telomeric repeat sequences and the non-coding sequences between the telomeres and large ribosomal subunit.
Free Energy of Structure = -3.6 kkal/mol

Figure 2 - Secondary structure of pseudo-tRNA Lysine in *T. paravorax*.

**Putative Transcription Control Region**

*Tetrahymena* Mt DNA replication initiates bidirectionally from an origin near the center of the genome according to electron microscopy studies (Arnberg et al., 1974). Engberg and Nielsen mapped the origin of replication of rDNA of *T. thermophila* to a position
close to the middle of the molecule (Engberg and Nielsen, 1990). Similarly the promoter region of the rDNA genes was shown to be located in the same region with a few conserved repeat sequences, which bind to Topoisomerase I (Andersen et al., 1985). Presence of the origin of replication and transcription in the middle of the rDNA molecule persuaded us to search for such elements in the longest intergenic region at the middle of the Tetrahymena Mt DNA. Comparative genomic analysis of the Cob and Ynf77 intergenic region, which was suspected to contain a control region since a bidirectional transcription initiates from this region, prompted us to search for a transcription control (GC box). Sequence alignments of this region, using the five Mt genomes of Tetrahymena species showed a 94bp conserved block of sequence in this variable intergenic region (Figure 3). A supplemental figure of this region is attached to the end of this chapter. This conserved block contained a 27bp highly conserved consensus sequence (AATA\text{GCCGCACC}TAAAGAAAAAAATC). Among the 27 bases in these 5 species the consensus sequence had only 3 bases that deviated. In this region, which has 88% A+T, it is highly improbable that the putative GC control box (GCCGCACC) would occur by chance (Figure 3). When the probability of having an A or T is about 0.88 the probability of having eight nucleotides from which seven are either G or C is \((0.12)^7\times(0.88)^8=2.5\times10^{-6}\). Alignment of this intergenic region was littered with gaps, except for the conserved region containing the GC box. This conservation in a highly variable region suggests high selective pressure, which is common among functional elements in genomes. To further show the nucleotide conservation of this presumptive control region we plotted the G+C content and conservation at each
nucleotide position (Figure 4). From left to right the genes are *nad9*, *Ymf77*, the intergenic region and the *cob* gene. The *cob* gene sequence is highly conserved and has a high G+C content, which is a bit variable. The *nad9* gene is also conserved, but has a lower G+C value than even *Ymf77*. Most of *Ymf77* is less conserved, due to the highly variable nature of this gene, than either *cob* or *nad9* but has a relatively high G+C value in the carboxyl terminal region (to the left). The intergenic region is less conserved than the *cob* gene, but similar to *Ymf77*. The presumptive control region is evident by its conservation and high G+C value (Figure 4). A GC box in general contains the sequence GCCGC and is recognized by the factor SP1 (Pugh and Tjian, 1990). SP1 presumably interacts with other transcription factors (TFs) to initiate transcription. The fact that the genes flanking this sequence are transcribed in opposite orientations increases the confidence that this sequence contains a control region. This region is the site from which bi-directional transcription of most of the Mt genes is initiated. We also know that DNA replication originates at this region in a bidirectional mode and completes when the single stranded Mt DNA circularize due to presence of tandem repeats at each end. Although in bacterial chromosomes and plasmids initiation of DNA replication occurs at a single unique site (e.g., *OriC*), a consensus sequences for the origin of replication in mitochondria has not yet been established. In mammals and amphibia, some signals are located within the AT rich and variable control region for the replication initiation H-strand and for transcription of both H- and L-strands (Clayton, 1992). The origin of replication and transcription was suspected to be at the same region in Mt genome of *P.aurelia* (Burger et al., 2000). Thus it is quite possible that the conserved block of 27
nucleotides mentioned above may also have a role in replication of Tetrahymena Mt genomes. A replication origin in the middle of the Mt DNA requires a circular form to complete the process. In Tetrahymena this is solved by presence of tandem repeats at each end of the Mt DNA. These tandem repeats are complementary with a few differences in their over 2.5kb length (Table 3). Paramecium Mt DNA, on the other hand, does not contain tandem repeats hence cannot form a circular molecule. In fact there is no need for circularization in Paramecium Mt DNA since the replication originates at the end of the molecule. Origin of replication in Tetrahymena is between Ynf77 and Cob genes. A 12 gene inversion followed by the deletion of Ynf77 puts the origin of replication at the end of the Paramecium Mt DNA adjacent to Nad9. In fact a 12bp block of poly A sequences next to the GC box in Tetrahymena (Figure 3) is also present in Paramecium with 3 nucleotide differences. Hence this conserved poly A block may play a role in replication of these Mt genomes.

Differences between the two inverted terminal repeats

Tetrahymena mitochondrial genomes have inverted terminal repeats, which contain either portions or all of the large subunit ribosomal RNA genes (LSUrRNA) interrupted with a tRNA (Burger et al., 2000; Brunk et al., 2003). These two regions are clearly under concerted evolution since intra-species differences between the two ends are just a few base pairs compared to hundreds of inter-species substitutions. Table 3 illustrates the number of differences between the inverted terminal repeats of five Tetrahymena species at 5’ and 3’ ends. As is expected based on variations discussed above T.paravorax has the highest number of differences compared to rest, which exceed 300bp at either end.
Figure 3- The nucleotide alignment of the cob and ynf77 intergenic region.
Italic sequences represent the conserved region with a box around the transcription control GC sequence (CR=Control Region).
Figure 4. Nucleotide conservation in 6500bp of mitochondrial genome of *Tetrahymena species* including the control region. Arrow denotes the conservation of a putative transcription control. G+C content and nucleotide changes are calculated using windows of 100bp and a step size of 1 nucleotide and are shown with dark blue and red. Average G+C and nucleotide change are shown by light blue and magenta for the entire gene or intergenic region. A GC box, which is illustrated by a single elevated peak in control region is the location for origin of transcription. Its conservation well compares to the conservation of known genes such as cob (right) and genes with unknown function such as Ymf 77 (left).

x-axis is the gene length
y-axis-arbitrary units
Numbers of differences between the other species are less than 200, which is 30% less than what we observe in comparisons with *T. paravorax*. Not surprisingly, *T. malaccensis* and *T. thermophila* have the fewest differences since phylogenetically they are immediate descendents of a common ancestor and consequently very closely related. Concerted evolution occurs in these sequences since they need to remain complementary to one another so they can form a circular single stranded DNA during DNA replication.

<table>
<thead>
<tr>
<th>Differences</th>
<th>T. paravorax</th>
<th>T. thermophila</th>
<th>T. malaccensis</th>
<th>T. pigmentosa</th>
<th>T. pyriformis</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. paravorax</td>
<td>3</td>
<td>326(3')</td>
<td>336(3')</td>
<td>302(3')</td>
<td>321(3')</td>
</tr>
<tr>
<td>T. thermophila</td>
<td>327(5')</td>
<td>5</td>
<td>80(3')</td>
<td>169(3')</td>
<td>185(3')</td>
</tr>
<tr>
<td>T. malaccensis</td>
<td>340(5')</td>
<td>83(5')</td>
<td>5</td>
<td>158(3')</td>
<td>194(3')</td>
</tr>
<tr>
<td>T. pigmentosa</td>
<td>304(5')</td>
<td>170(5')</td>
<td>158(5')</td>
<td>4</td>
<td>134(3')</td>
</tr>
<tr>
<td>T. pyriformis</td>
<td>322(5')</td>
<td>185(5')</td>
<td>196(5')</td>
<td>134(5')</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3- The number of intra and inter-species differences between terminal inverted repeats in 5' and 3' ends of the mitochondrial genomes of five *Tetrahymena* species. Raw numbers of differences are in base pairs.

**Putative proteins of unknown function**

*Tetrahymena* Mt genome contains 22 putative proteins of unknown function called Ymf. Orthologous Ymfs from five *Tetrahymena* species compared in this study are fairly conserved and are under apparent selection. Searching through protein databases we found a few similar domains from Mt proteins for a few ymfs, (i.e., ymf61 with yeast ribosomal VAR1 protein, ymf75 with atp6, and ymf77 with RpoC). Despite the domain similarity none of the proteins passed a similarity metric test (*Z*-score < 6, appendix 1) so we could not infer common ancestry (for alignment procedure refer to material and
methods). Previously Brunk et al. reported similar sequences for ymf58 (nad4L), ymf60 (rpl6), and ymf62 (nad6) and proposed common ancestry by using chain homology (Brunk et al., 2003). The most interesting of these putative proteins is Ymf77, which is also the longest gene (4kb in length) and occupies roughly 10% of these genomes. Analysis of Ymf77 suggests that it may have a role in polymerase activity since DNA sequence and protein domain searches indicate similarity with the proten family of RpoD, RpoC, and RpoB (RNA polymerase subunits D, C, and B). We also searched through the tertiary domain structure database and came up with the same proteins from the Mt genome of a flagellate protist called R.americana (Lang et al., 1997). RpoB and RpoC are almost the same length as Ymf77, share similar domains with Ymf77, and are present in protist mitochondria. The A+T content of RpoC is over 78% whereas Ymf77 contains about 88% A and T. Although the A+T content in R.americana is 10% less than its counterpart in Ymf77 it still has a relatively high A+T ratio. The A+T content in RpoB of E.coli is slightly less than 50% and in RNA pol III of humans is about 60%. Similar domains, length, and A+T content between Ymf77 and Mt RNA polymerase subunits make them a good candidate for Ymf77 function.

Ymf77 secondary structure

Another interesting feature of Ymf77, especially in T.paravorax compared to the other genomes is the presence of a 1kb sequence in a putative gene ymf77 that constructs a strong secondary structure with the highest free energy value (Figure 5). The A+T content of this region was 96.5% resulting in secondary structure, which prohibited us
from cloning or PCR amplifying it directly from the entire Mt genome preps. Finally, a restriction digest fragment of about 3.9kb containing this region was extracted and boiled in a solution with 1% SDS to denature the secondary structure and eventually the DNA. This procedure allowed us to PCR amplify and clone this 1kb fragment with two different sets of primers and determine the sequence.

![Ymf77 Free Energy Graph](image)

Figure 5- Free energy values for 500bp intervals of Ymf77 genes in five *Tetrahymena* species. Each point denotes the free energy value of the secondary structure constructed from 500bp of Ymf77 sequences. The peak at 3kb region of Ymf77 gene represents the lowest free energy value and a strong secondary structure throughout the genome of *T.paravorax*.

**Telomeres**

The mitochondrial telomeres for *T.pigmentosa* and *T.malaccensis* have been previously identified and characterized (Morin and Ceck, 1988c). We found that *T.paravorax* has tandem 64bp repeats as its telomeres, which are the longest known telomeres in linear Mt DNA of *Tetrahymena*. Following restriction endonuclease digestion of intact *Tetrahymena* Mt DNA, the terminal restriction fragments were slightly “ smeared” due to the variable length of the telomeres. In the case of *T.paravorax*
the terminal restriction fragments could not be identified. The intact *T. paravorax* Mt DNA was digested with restriction endonuclease *Tsp 509 I*, which did not cleave within the telomere yet reduced the remainder of the Mt DNA to tiny fragments. After gel electrophoresis a prominent band centered at 6kb and ranging from 3 to 9kb was observed suggesting that the length of the two telomeric repeats comprised about 22% of the Mt genome. *T. paravorax* apparently contained on average almost 100 telomeric repeats where the other *Tetrahymena* Mt genomes appeared to have about 1 or 2 dozen repeats at each end. The sequence of this telomere is the same at both ends of the genome (5′TATCCCTATTCCCTATATTCTATCATCATTCCCATATTATCTAAACGTCTATGTACTTTGTT3′). Our suggestion is supported by the frequency of clones that contain telomeric repeat sequences after the completion of the project. From 338 clones sequenced to complete the *T. paravorax* Mt genome 30 of them, almost 9%, contained telomeric repeats. On the other hand, *T. paravorax* telomeres, with average length of 6kb, comprise more than 11% of the Mt DNA length, which is in concordance with the clone frequency. This phenomenon is not seen in the other two sequenced genomes where clones with telomeric sequences have a frequency of 4% in *T. malaccensis* and less than 2% in *T. pigmentosa*. The length of the *T. paravorax* Mt telomere and its repeats, which are quite longer than other known species put emphasis on the existence of important differences between *T. paravorax* and the other species. Multimodal distribution of telomeres in *Tetrahymena* Mt DNA is not a new discovery since Morin and Ceck have shown that it exists on the left telomere of *T. hyperangularis*.
which are different at each end, and *T. malaccensis*. They also suggest that the average length of the telomeres in *T. malaccensis* are about 2.4±1.3kb (Morin and Ceck, 1988c). Mechanisms responsible for maintenance and distribution of the telomeres in *Tetrahymena* mitochondria are not very well known, however unequal homologous recombination of telomeres is the proposed model (Morin and Cech, 1986). Single stranded Mt DNA in *Tetrahymena* is circularized due to presence of terminal tandem repeats (Goldbach et al., 1979). Figure 6 taken from Morin and Ceck (1986) also indicates recombination for the telomeres, which according to their model is responsible for maintenance and the size elongation of the Mt ends. Of course such a model should work for identical telomeric repeats at both ends of *Tetrahymena* Mt DNA, hence *T. pigmentosa* telomeres should use a different mechanism for their telomeric repeat maintenance. Another phenomenon that needs explanation is how multimodal distribution of telomeres in *T. paravorax* and *T. malaccensis* occurs. Besides suggesting a defective or altered telomere maintenance mechanism in these two genomes currently the reasons are unknown. Alternatively one may suggest that some *Tetrahymena* species have developed means for extra elongation and maintenance of their Mt telomeres. This is unlikely to be the case since telomeres in *T. thermophila*, which have very similar sequences to the telomeres of *T. malaccensis*, have unimodal distribution and are not unusually long (Morin and Ceck, 1988c). This problem becomes even more interesting when one tries to explain the size heterogeneity of the left end telomere of *T. heperangularis*. In this species, like *T. pigmentosa*, the telomere maintenance mechanism cannot be the unequal recombination of telomeric sequences since they

59
contain telomeric repeats with different lengths and different base compositions. One fact is clear that the Mt genome of *Tetrahymena* species with two different telomeric repeats at their ends are almost 5kb shorter in length than the ones with identical telomeric repeats (Morin and Cech, 1988b). We were able to prove this independently since we completely sequenced the Mt genome of *T.pigmentosa*, which was syntenic with all the other sequenced genomes. Furthermore we found very short telomeric repeat sequences at both ends of this genome suggesting that lack of unequal recombination in these genomes has resulted in shortening of their telomeric repeats and consequently the length of the Mt DNA. The telomeres on each terminus of the *T.pigmentosa*, *T.heperangularis*, and *T.hegewischi* have different sequences as previously found by Morin and Cech (1988c). In fact these species were named *T.pigmentosa* group by them. The 5' telomere was 50bp long (AGTATAGAGTAGTATACAGTATTGGACATAAATGCAGTACAT ATAAAATA) and the 3' telomere was 37bp long (TATCATATATCCATGTTAAGA ATAGGGTTAAATATAG). Nothing in the sequence of the *T.pigmentosa* Mt genome suggested how these telomeres maintained different sequences.

Table 2 shows the length of the non-coding sequences between the telomeres and the tandem repeats on each terminus of Tetrahymena Mt DNA. The tandem repeats contain almost identical sequences at either end, which is to create a circular single stranded DNA to complete Mt DNA replication. Also telomeric repeats are identical for the species that contain the same telomere at both ends. So why the joining sequences between these two are not subject to any concerted evolution remains an open question.
*T. pigmentosa*, which contains different telomeres on each terminus, has no joining sequence to the tandem repeat on the 3' end. Similarly *T. hegewischi*, contains no joining sequence at the left side (Morin and Ceck, 1988c).

![Diagram](image)

Figure 6- Model for structure of circularized single-stranded Mt DNA (taken from Morin and Ceck, 1986).

The other species of *Tetrahymena* that contain the same telomeric repeats on each terminus have joining sequences ranging from 4 in *T. malaccensis* to 495 in *T. paravorax*. Yet none of the sequences at either end are complementary to hybridize during circularization. Alignment of *T. paravorax* and *T. pyriformis* joining sequence, which are considerably longer at both ends relative to the others did not indicate absolute sequence similarity (Figure 7). Yet a small region at the 5' joining end where the pseudo-tRNA resides was similar to the 3' joining sequences. The reason for this could be the presence of another pseudo-tRNA at the 3' non-coding sequence. In fact upon detailed search of
the 3’ joining sequence I found partial sequence similarities to a Ile pseudo-tRNA.

Another explanation could be the high A+T content of these sequences that could generate higher than expected sequence similarity. This explanation was more plausible for \( Tpyriformis \) joining ends since they were shorter and did not contain any pseudo-
tRNA. The non-coding or joining sequence were too short to produce a reasonable
alignment for \( T.thermophila \) and \( T.malaccensis \) (Table 2). Here I propose that the joining
sequences do not contain any binding motif for a telomere maintaining protein nor they were complementary to assist in circularization of the single stranded Mt DNA. Hence
they do not play an important role in telomere maintenance or Mt DNA replication. It seems like their absence in \( T.pigmentosa \) group did not result in elimination or
elongation of the telomeres. The short telomere length for the species in this group is due
to lack of unequal recombination because of different telomeric sequences on each
terminus.

A) \( T.paravorax \)

<table>
<thead>
<tr>
<th>Tpar5end</th>
<th>Tpar3end</th>
</tr>
</thead>
</table>
| TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

62

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Two short conserved ORFs in intergenic regions

I found two conserved short open reading frames (ORFs) in intergenic regions between Ymf57-LSU_rnl_b and Ymf66-Ymf76. Upon translation of these ORFs it became clear that they may contain two putative proteins with reasonable conservation. The one in between Ymf57 and LSU_rnl_b is 33 amino acids long and starts with known start codons (i.e., Met, Ile, Val) in Tetrahymena mitochondrial genomes. An alignment of this sequence is shown in Figure (8A) where conservation of its sequence is apparent. This conservation becomes more distinct when T.paravorax sequence, which contains the most divergent mitochondrial genome, is removed from the alignment (Figure 8B). Similarity metric analysis of pairwise sequences indicated homology (Z-score > 6) between all species, with the exception of T.paravorax, which generated Z-scores slightly lower than 6 (Figure 8C). This phenomenon was also apparent in Ymf71, which
is the most divergent gene in *Tetrahymena* mitochondria, where a few pairwise comparisons did not pass the similarity metric test with a small margin. BLAST searches for this putative protein did not result in any similar sequences in the databases. Nevertheless the conservation of this sequence, especially on amino acid level, could mean that it has been a functional element perhaps a short gene in *Tetrahymena* mitochondria.

Similarly the intergenic region between *Ymf66* and *Ymf76* contains a conserved sequence, which upon translation generates a 44 amino acids long sequence with Met as the sole start codon. Conservation of this sequence was more than the previous one where all pairwise comparisons pass the similarity metric test with Z-scores > 6. An alignment of this sequence along with Z-scores is shown in Figure 9 for all 5 species. I was not able to find any significantly similar amino acid sequences to this putative protein in any of the databases as well. The shortest known protein in *Tetrahymena* mitochondria is almost 100 amino acids long. Presence of these much shorter sequences, which may be considered as functional elements or perhaps short genes, was an unusual phenomenon. Failure to find any similar genes or functional elements to these sequences further complicates to whether accept or reject these as putative proteins. An alternative explanation would be to think that they were part of proteins, which were eliminated from *Tetrahymena* Mt genomes.
A - Alignment With T_paravorax

Tpyr  VPRLVMQAFKFLPLPMLHEPFLPILHAYRFIIY
Tpar  MQLLILQQAKFLFFHFSFFKFLKLKKLFPP
Tpig  MPRLLMVQFVKLNLVFELIFLISLKFIY
Tmal  MPRLLMVQFVKLNLVFELIFLISLKFIY
Ttherm IPTRLMLQVFKLFNLFLFELILFILMSNFKFY

B - Alignment Without T_paravorax

Tpig  MPRLLMVQFVKLNLVFELIFLISLKFIY
Tmal  MPRLLMVQFVKLNLVFELIFLISLKFIY
Ttherm IPTRLMLQVFKLFNLFLFELILFILMSNFKFY
Tpyr  VPRLVMQAFKFLPLPMLHEPFLPILHAYRFIIY

C - Z-scores

<table>
<thead>
<tr>
<th></th>
<th>Ttherm</th>
<th>Tmal</th>
<th>Tpig</th>
<th>Tpyr</th>
<th>Tpar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ttherm</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tmal</td>
<td>11</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tpig</td>
<td>11</td>
<td>15</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tpyr</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>Tpar</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>**</td>
</tr>
</tbody>
</table>

Figure 8 - Alignment and Z-scores of conserved 33 amino acid sequences between Ymf77 and LSU_rnl_b

A - Alignment of all five species

Ttherm  MPKELHIFRTYIFTFRLKKSNINFLKNSFYTLAAYLYIINFM-
Tmal   MPKELHIFRTYIFTFRLKKSNINFLKNSFYTLAAYLYIINFM-
Tpig   MPKELHIFRTYIFTFRLKKSNINFLKNSFYTLAAYLYIINFM-
Tpyr   MPKELHIFRTYIFTFRLKKSNINFLKNSFYTLAAYLYIINFM-
Tpar   MPKELHIFRTYIFTFRLKKSNINFLKNSFYTLAAYLYIINFM-

B - Z-scores

<table>
<thead>
<tr>
<th></th>
<th>Ttherm</th>
<th>Tmal</th>
<th>Tpig</th>
<th>Tpyr</th>
<th>Tpar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ttherm</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tmal</td>
<td>17</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tpig</td>
<td>15</td>
<td>16</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tpyr</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>Tpar</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>**</td>
</tr>
</tbody>
</table>

Figure 9 - Alignment and Z-scores of conserved 44 amino acid sequences between Ymf66 and Ymf76
Summary and Conclusion

This study addresses questions regarding the molecular evolution of the ciliate Tetrahymena Mt genome. First, the differences among completely sequenced Mt genomes in this genus are reported and analyzed to indicate that Tetrahymena species have divergent Mt DNA. Second, inter-species differences are analyzed in detail and indicate that *T. paravorax* has the most divergent Mt DNA among studied genomes. Third, a putative promoter sequence that initiates the transcription of the genes in Tetrahymena Mt genomes is identified and reported. Fourth, the telomere lengths, their maintenance mechanisms, and their joining sequences are studied in detail.

In this study, we sequenced three ciliate Mt genomes from the genus *Tetrahymena*. Analyses of five available Mt genomes from this genus suggest rapid evolution in Tetrahymena Mt genomes, especially *T. paravorax*, which is an early branching species and has the most divergent Mt DNA. Despite having complete sequences of five Mt genomes, roughly half of the genes still remain without an assigned function. This and high numbers of nucleotide and amino acid substitutions in coding genes discussed in the next chapter, demonstrate divergence in Mt genomes of *Tetrahymena*. More specific findings that support this idea come from long non-coding regions in *T. paravorax*. These regions contain a pseudo-tRNA that did not exist in the other species and may indicate the loss of Mt tRNA genes in Tetrahymena Mt genome. Having tRNA degradation among species of the same genus suggests a fairly divergent Mt genome. The long non-coding regions, between the telomeres and the large ribosomal subunit tandem repeats, drift quickly relative to the genes and intergenic regions.
Moreover, the presence of different telomeres with variable sizes in *Tetrahymena* species, nad9 gene duplication in *T. thermophila* and *T. malaccensis* (Brunk et al., 2003), all indicate fast evolving Mt DNAs.

Compared to genes, intergenic regions have relatively less conservation, but enough to be able to identify a previously unknown transcription control GC box. The presence of a central bi-directional promoter has previously been suspected at the same region where we propose this GC box. Of course, having five complete Mt genomes and conducting comparative genomic analysis is the key to observing promoter sequence conservation at this region of the Mt genome where the bi-directional transcription of the genes initiate. Hence, we were able to identify a presumptive promoter sequence in an intergenic region that is believed to initiate transcription. This region also contains sequences that initiate the Mt DNA replication, yet with its extremely high A+T content, it could be in a conserved region adjacent to proposed transcription GC box.

**Appendix**

```perl
# % Score calculating program
# Mike Moradian: Nov 2003 v0.3
#

$Output = "Output.txt";
open ( OUTPUT, ">$Output" );

my %SubstitutionMatrixIndex =
  ( 'a' => 0,
    'r' => 1,
    'n' => 2,
    'd' => 3,
    'c' => 4,
);```

67
my @SubstitutionMatrix = 
{
    A R N D C Q E G H I L K M F P S T W Y V
    [5,  2,  1,  2,  1,  1,  0,  2,  1,  2,  1,  1,  0,  3,  2,  0], # A
    [2,  7,  1,  2,  4,  1,  0,  3,  0,  4,  3,  3,  2,  3,  1,  1], # R
    [1,  1,  7,  2,  2,  0,  0,  1,  3,  4,  0,  2,  4,  2,  1,  0], # N
    [2,  2,  8,  4,  0,  2,  1,  1,  4,  1,  4,  5,  1,  0,  1,  5], # D
    [2,  4,  13, 3,  3,  3,  3,  3,  2,  2,  3,  2,  2,  4,  1,  1], # C
    [1,  1,  0,  3,  7,  2,  2,  1,  3,  2,  2,  0,  4,  1,  0,  1], # Q
    [1,  1,  0,  3,  2,  6,  3,  0,  4,  3,  1,  2,  3,  1,  1,  3], # E
    [1,  0,  3,  2,  2,  8,  4,  2,  4,  3,  4,  2,  0,  2,  3,  4], # G
    [2,  0,  1,  1,  3,  1,  0,  2,  10, 4,  3,  0,  1,  1,  2,  1,  2], # H
    [-1,  4,  3,  4,  2,  3,  3,  3,  4,  4,  5,  2,  3,  2,  0,  3,  1], # I
    [-1,  3,  4,  4,  2,  3,  4,  4,  2,  2,  3,  2,  2,  3,  1,  2,  1], # L
    [-1,  1,  0,  1,  3,  2,  1,  2,  0,  3,  3,  6,  2,  3,  1,  0,  1], # K
    [-1,  2,  2,  2,  2,  3,  2,  7,  0,  3,  2,  1,  1,  0,  1], # M
    [-1,  3,  4,  5,  2,  4,  3,  4,  1,  0,  1,  4,  0,  8,  4,  2,  1], # S
    [-1,  3,  2,  1,  4,  1,  1,  2,  2,  1,  3,  3,  4,  1,  4,  3,  1], # P
    [1,  1,  0,  1,  0,  1,  0,  3,  1,  0,  2,  3,  1,  5,  2,  4,  2], # T
    [1,  0,  1,  2,  1,  1,  2,  2,  1,  1,  2,  1,  2,  5,  3,  2,  0], # W
    [2,  1,  1,  1,  1,  2,  1,  2,  1,  2,  1,  2,  1,  4,  4,  3,  15], # Y
    [2,  1,  2,  3,  1,  2,  3,  2,  1,  1,  2,  0,  4,  3,  2,  2,  8], # Y
    [2,  1,  3,  4,  1,  3,  4,  4,  1,  1,  3,  1,  1,  3,  2,  0,  3,  1], # V
};

my @FASTAMatrix;

my $SeqData1 = "";
my $SeqData2 = "";

my @Seq1;
my @Seq2;

my $AlignedSeq1 = "";
my $AlignedSeq2 = "";

68

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
sub ReadFASTAPor2Seq {
    # Read FASTA File with two sequences
    if (open(FASTAFILE, "$FASTAfilename") == 0) {
        print "Cannot open file \"$FASTAfilename\"\n\n";
        exit;
    }
    print "$FASTAfilename opened";
    print "\n";
    $FASTAinput = <FASTAFILE>;
    close FASTAFILE;

    # Extract Sequence and name
    $string = $FASTAinput[0];
    $test = substr($FASTAinput[0], 0, 1);
    unless ($test eq '>') {
        print "First character in FASTA file not >";
        exit;
    }
    $i = 0;
    $j = 0;
    $FirstSequence = -1;
    $l = scalar(@FASTAinput);
    while ($i < $l) {
        if (substr($FASTAinput[$i], 0, 1) eq '>') {
            $FirstSequence *= -1;
            if ($FirstSequence eq 1) {
                $SeqName1 = substr($FASTAinput[$i], 1, length($FASTAinput[$i])-1);
            } else {
                $SeqName2 = substr($FASTAinput[$i], 1, length($FASTAinput[$i])-1);
            }
            $j = 0;
        }
    }
}

69

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
else
{
    if ($FirstSequence eq 1)
    {
        $Seq1[$j] = $FASTAinput[$i];
    }
    else
    {
        $Seq2[$j] = $FASTAinput[$i];
    }
}
$i++;
$j++;
}

# From the lines of the sequence files, put Seq1 & Seq2 into a single string
$SeqData1 = join( ' ', @Seq1);
$SeqData2 = join( ' ', @Seq2);

# Remove whitespace
$SegName1 = s/\s\s/g;
$SegName2 = s/\s\s/g;
$SegData1 = s/\s\s/g;
$SegData2 = s/\s\s/g;

# Shift to lowercase
$SegData1 = s/[A|R|N|D|C|Q|E|G|H|I|L|K|M|F|P|S|T|W|Y|V]/\L$1/gi;
$SegData2 = s/[A|R|N|D|C|Q|E|G|H|I|L|K|M|F|P|S|T|W|Y|V]/\L$1/gi;

# Explode Sequences into arrays where each letter of the original string
# is an element in the array
@Seq1 = split( ' ', $SegData1);
@Seq2 = split( ' ', $SegData2);

$Len1 = scalar(@Seq1);
$Len2 = scalar(@Seq2);
}

sub BuildFASTA\_Matrix{
    # Build the matrix using the fasta sequences' length
    # Fill first row and column with -8 gap penalty
    $GapPenaltyConstant = -8;
    $iGapPenalty = 0;
for ($i=0; $i<$Len1+1; $i++)
{
    $FASTAMatrix[$i][0] = $iGapPenalty;
    $iGapPenalty += ($GapPenaltyConstant);
}
$iGapPenalty = 0;
for ($i=0; $i<$Len2+1; $i++)
{
    $FASTAMatrix[0][$i] = $iGapPenalty;
    $iGapPenalty += ($GapPenaltyConstant);
}

# Step 1 of 2 :Calculate the matrix values - Initial Calculation
for ($i=0; $i<$Len1; $i++)
{
    for ($j=0; $j<$Len2; $j++)
    {
        $RowIndex = $SubstitutionMatrixIndex[$Seq1[$i]];
        $ColIndex = $SubstitutionMatrixIndex[$Seq2[$j]];
        #print "\n$Seq1[$i] -> $RowIndex and $Seq2[$j] -> $ColIndex";
        $FASTAMatrix[$i+1][$j+1] = $SubstitutionMatrix[$RowIndex][$ColIndex];
    }
}

# Step 2 of 2 :Calculate the matrix values - Actual Calculation
for ($i=0; $i<$Len1; $i++)
{
    for ($j=0; $j<$Len2; $j++)
    {
        $Value1 = $FASTAMatrix[$i+1-1][$j+1-1] + $FASTAMatrix[$i+1][$j+1];
        $Value2 = $FASTAMatrix[$i+1-1][$j+1] + $GapPenaltyConstant;
        $Value3 = $FASTAMatrix[$i+1][$j+1-1] + $GapPenaltyConstant;
        $FASTAMatrix[$i+1][$j+1] = Maximum3($Value1,$Value2,$Value3);
    }
}

sub TraceBackMatrix {
    # Trace Back the elements in matrix to calculate the path
    $i = $Len1;
    $j = $Len2;
    $AlignedSeq1 = "";
    $AlignedSeq2 = "";
}

71
while( ($i ne 0) or ($j ne 0) )
{
    # Check to see if row & column are in match
    if ($Seq1[$i-1] eq $Seq2[$j-1])
    {
        if ($i eq 0)
        {
            $AlignedSeq1 = "-" . $AlignedSeq1;
        }
        else
        {
            $AlignedSeq1 = $Seq1[$i-1] . $AlignedSeq1;
            $i--;
        }
        if ($j eq 0)
        {
            $AlignedSeq2 = "-" . $AlignedSeq2;
        }
        else
        {
            $AlignedSeq2 = $Seq2[$j-1] . $AlignedSeq2;
            $j--;
        }
    }
    # print "\nSequences: $AlignedSeq1 & $AlignedSeq2"
}
else
{
    $RowValue = -999999;
    $ColValue = -999999;
    if ($i>0)
    {
        $RowValue = $FASTAMatrix[$i-1][$j];
    }
    if ($j>0)
    {
        $ColValue = $FASTAMatrix[$i][$j-1];
    }
    # Choose one of the vertical or horizontal paths
    if ($RowValue > $ColValue )
    {
        $AlignedSeq1 = $Seq1[$i-1] . $AlignedSeq1;
        $AlignedSeq2 = "-" . $AlignedSeq2;
        $i--;
    }
    else
    {
        $AlignedSeq1 = "-" . $AlignedSeq1;
        $AlignedSeq2 = $Seq2[$j-1] . $AlignedSeq2;
        $j--;
    }
}

72
# Maximum number of three entries

```perl
sub Maximum3 {
    my $myValue1 = @_[0];
    my $myValue2 = @_[1];
    my $myValue3 = @_[2];

    # Find the maximum value
    $Maximum3 = $myValue3;
    if ($myValue1 > $myValue2) {
        if ($myValue1 > $Maximum3) {
            $Maximum3 = $myValue1;
        }
    } elsif ($myValue2 > $Maximum3) {
        $Maximum3 = $myValue2;
    }

    # Print "\n $Maximum3 -> ($Value1 $Value2 $Value3)"
    return $Maximum3;
}
```

# Jumble Routine

```perl
sub Jumble {
    my $temp = 0;

    # Shuffle Sequence 1
    for ($i=0;$i<$Len1;$i++) {
        $pos1 = int((rand $Len1));
        $pos2 = int((rand $Len1));
        $temp = $Seq1[$pos1];
        $Seq1[$pos1] = $Seq1[$pos2];
        $Seq1[$pos2] = $temp;
    }

    # Shuffle Sequence 2
    for ($i=0;$i<$Len2;$i++) {
        $pos1 = int((rand $Len2));
        $pos2 = int((rand $Len2));
        $temp = $Seq2[$pos1];
        $Seq2[$pos1] = $Seq2[$pos2];
    }
}
```

73
$Seq2[$pos2] = $temp;
}

return $func_data;
}

sub PrintMatrix {
  # Just print out the matrix
  print "\n\n  ";
  for ($j=0; $j<$Len2+1; $j++) {
    if ($j>0)
      {
        print "  $Seq2[$j-1]";
      }
  }
  print "\n ";
  for ($i=0; $i<$Len1+1; $i++) {
    if ($i>0)
      {
        print "\n(Seq1[$i-1]"
      }
    for ($j=0; $j<$Len2+1; $j++)
      {
        printf "%4d",$FASTAMatrix[$i][$j];
      }
    print "\n";
  }
}

###########################################################################
##
## MAIN
###########################################################################
##
print "\n\n";
print "\n-------------------------------------";
print "\n| Z Score Calculation Program ver0.3";
print "\n| By: Mike Moradian  Date: Nov 2003";
print "\n-------------------------------------";

ReadFASTAFor2Seg;

my @JumbleScore;
my $NumberOfJumbles = 0;
my $TotalJumbleScore = 0;
my $AverageJumbleScore = 0;
my $NSDAM = 0;

my $IsThisTheOriginalOne = 1;
my $OriginalAlignmentScore = 0;
$iTestRun= $ARGV[1];
for( $iTestRun=0; $iTestRun<$ARGV[1]; $iTestRun++)
{ 
BuildFASTAMatrix;  
#PrintMatrix;  
TraceBackMatrix;  

# print "\n>>>Original Sequence.";  
# print "\n","$SeqName1"," ", @Seq1;  
# print "\n","$SeqName2"," ", @Seq2;  
# print "\n";  

# print "\n>>>Aligned Sequence.";  
# print "\n","$SeqName1"," ", $AlignedSeq1;  
# print "\n","$SeqName2"," ", $AlignedSeq2;  
# print "\n";  

if ($IsThisTheOriginalOne ne 1)  
{  
  $JumbleScore[$NumberOfJumbles] = $FASTAMatrix[$Len1][$Len2];  
  $TotalJumbleScore += $FASTAMatrix[$Len1][$Len2];  
  $NumberOfJumbles++;  
}  
else  
{  
  $OriginalAlignmentScore = $FASTAMatrix[$Len1][$Len2];  
}  

# print "\nAlignment Score: ", $FASTAMatrix[$Len1][$Len2];  

#Jumble the sequences thing  
Jumble;  

$IsThisTheOriginalOne = 0;  
}  

#Calculating the average and standard deviation  
$AverageJumbleScore = $TotalJumbleScore/$NumberOfJumbles;  

# SD Formula for reference: sd = square root [sum(( x - xbar)^2) / (N-1)]  
$TotalJumbleScore = 0;  
for($i=0;$i<$NumberOfJumbles;$i++)  
{  
  $TotalJumbleScore += (($AverageJumbleScore - $JumbleScore[$i])**2);  
  # print "\n$TotalJumbleScore";  
}  

$SD = ($TotalJumbleScore/($NumberOfJumbles-1))**(0.5);  

# print "\nOriginal Align. Score: ", $OriginalAlignmentScore;  
# print "\nAverage Jumble Score: ", $AverageJumbleScore;  
# print "\nStandard Deviation: ", $SD;  

$NSDAM = ($OriginalAlignmentScore - $AverageJumbleScore)/$SD;  
# print "\nNumber Of SDs Above the Mean: ", $NSDAM;  
print "\n";  

75
print OUTPUT "\nOpened File Name: ", $FASTAfilename, "\nOriginal Align. Score: ", $OriginalAlignmentScore,"\nAverage Jumble Score: ", $AverageJumbleScore, "\nStandard Deviation: ", $SD, "\nNumber Of SDs Above the Mean: ", $NSDAM;

close (OUTPUT);

print "\nDone.";
References


Invitrogen Corporation. 1600 Faraday Avenue PO Box 6482 Carlsbad, California 92008


80

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 3

Nucleotide and amino acid substitution types reveal mutation hotspots and accelerated nonsynonymous substitutions in Ymf genes of *Tetrahymena* mitochondrial genomes
Abstract

*Tetrahymena* species contain rapidly evolving mitochondrial (Mt) genomes, which have apparently diverged significantly from the ancestral pattern leaving half of their 44 genes without an assigned function. Genes in *Tetrahymena* mitochondria may be categorized into two major groups of known protein coding (KPC) genes and genes with unidentified function (Ymf). To gain insights into the mechanisms underlying gene divergence and molecular evolution of *Tetrahymena* Mt genomes we sequenced three Mt genomes of *T.paravorax, T.pigmentosa, and T.malaccensis*. These genomes along with two previously sequenced genomes of *T.thermophila, and T.pyriformis* were aligned and the analyses were carried out using several programs that calculate distance, nucleotide substitution (dn/ds), and their rate ratios (ω) on individual codon sites and via a sliding window approach. A significant finding was the presence of at least one or more highly variable regions in Ymf genes where majority of substitutions were concentrated. These regions were mutation hotspots where elevated distances and the dn/ds ratios were primarily due to an increase in the number of nonsynonymous substitutions suggesting relaxed selective constraint. However, in a few Ymf genes, accelerated rates of nonsynonymous substitutions may be due to positive selection. Similarly, on protein level the majority of amino acid replacements occurred in these regions. Analysis of more than 1300 pairwise comparisons of the Ymf and KPC genes revealed that there were almost no nonsynonymous mutation hot spots in KPC genes compared to at least one or more highly divergent regions in the Ymfs.
Therefore I hypothesize that the reason why no homologues could be found for half of the genes in *Tetrahymena* Mt genome (i.e., *Ymf* genes) is that they are highly divergent in at least one or more regions (mutation hotspots), accept more nonsynonymous substitutions, and evolve more rapidly than KPC genes. *Ymf* genes comprise half of the genes in *Tetrahymena* Mt genomes so understanding why they have not been assigned definitive functions is an important aspect of molecular evolution. Importantly nucleotide substitution types and rates suggest possible reasons for not being able to find homologues for *Ymf* genes. Hence I suggested that the *Ymf* genes in *Tetrahymena* Mt genomes contain mutation hot spots, which could be a major obstacle in identifying their function.

**Introduction**

Despite their commonly accepted α-proteobacter ancestry, mitochondrial genomes in diverse lineages demonstrate extensive variability in size and structural organization. Ciliate Mt DNA is no exception, and is among the most rapidly evolving Mt genomes (Gray et al., 1999). The genomes of ciliates are particularly interesting as their genome content and rapid evolution of DNA at the primary sequence level makes them fairly divergent from all other genomes. The structural and organizational differences between *Paramecium* and *Tetrahymena* exemplify the rapid evolution of the ciliate Mt genomes (Burger et al., 2000). Although the previously sequenced Mt genomes of *T. thermophila* and *T. pyriformis* show complete synteny with regards to gene arrangement, they can be considered divergent due to considerable differences in nucleotide and amino acid
substitutions. This view is supported by the fact that *Tetrahymena* Mt genomes, in addition to 22 known KPC (known protein coding) genes, 7 different tRNAs, and the small and large subunits of ribosomal RNA, contain 22 Ymf genes with unidentified functions (Burger et al., 2000). The Ymfs, which are fairly conserved between *Tetrahymena* species, could be classified as transmembrane or ribosomal proteins by similarity of their hydrophobicity plots, physico-chemical properties, nucleotide composition, and codon usage biases (Brunk et al., 2003). Yet, none of the available similarity based searches could definitively find a homologue in GenBank for almost 50% of *Tetrahymena* Mt genes. This problem becomes particularly challenging when it is compared to the Mt genome of the jacobid protozoan *Reclinomonas Americana*, which at 100kb is more than twice as large as *Tetrahymena* Mt genome and carries 97 genes where about 96% of them can be assigned function (Lang et al. 1997).

It is an important task to try to clarify the mechanisms and the extent of this evolutionary diversification. Comparisons between DNA regions in mitochondria can be useful for differentiating between genome-specific factors such as DNA replication rate, and gene-specific factors such as substitution and selection (Kimura, 1980). Sequences from Mt genomes of *Tetrahymena* are found to have sufficient similarity to each other to show a reliable alignment of orthologous genes and control regions, but they are sufficiently diverged to obtain an estimate of the rate of evolutionary change, positive selection, and amino acid and nucleotide substitution patterns. Of course, using more than a few genes allows more accurate quantification of these evolutionary processes with a smaller error margin.

84
Comparisons between DNA regions in mitochondria can be useful for differentiating between genome-specific factors such as DNA substitution rates and gene-specific factors such as selection (Grey et al., 1998). Sequences from Mt genomes from *Tetrahymena* are sufficiently similar allowing reliable alignment of orthologous genes and control regions, but they are sufficiently diverged to allow an estimate of the rate of evolutionary change, positive selection, and amino acid replacement and nucleotide substitution patterns. Investigating variations in Mt genes enables understanding of evolutionary forces acting at individual loci and whole genomes (Ballard and William, 2000). The dn/ds ratio is an important tool for determining the levels of selective pressure. If the dn/ds > 1, then one may suspect positive selection; the lower the dn/ds ratio the more selection is acting on the protein (Li, 1997). However excess of nonsynonymous substitutions is not sufficient to demonstrate positive selection since such an increase may be due to relaxed selective constraints along certain lineages (Krandal and Hillis, 1997). The dn/ds rate ratio (\( \omega \)) produced by codon based models, which allow for variable selection intensity among sites, could detect sites under positive selection when \( \omega > 1 \) (Yang and Nielsen, 2000). To determine \( \omega \) these methods, in addition to previously used variables, consider two major features of DNA evolution: the transition/transversion bias and the base or codon frequency bias. Therefore \( \omega \) can be estimated at each codon site using aligned sequences from several species. Alternatively, Tajima's \( D \) statistical analysis (Tajima, 1989) is performed to detect selection and deviation from neutrality (Chiang et al., 2003). Significantly negative Tajima's \( D \) values are consistent with positive selective pressure (Carlson et al., 2005; Yu et al., 2005).
Methods for estimating dn/ds ratios usually consider whole genes. However during the course of evolution some sites are strictly conserved (for correct folding) whereas others could be subject to positive selection (Siltberg and Liberles, 2002). In several proteins that have been shown to be under positive selection only a few amino acid sites were found to be responsible for adaptive evolution (Yokoyama and Yokoyama, 1996). To detect regions of protein that are under positive selection where the whole protein may not be, one may use a sliding window program, which analyses nucleotide substitution and dn/ds rate ratios for any desired length. Mutation hotspots with accelerated nonsynonymous substitutions could occur in regions that are under positive selection however these do not present positive selection unless they contain sites where $\omega > 1$. An explanation for the presence of these regions is that they code for parts of the protein, which do not play essential roles in its function allowing more nonsynonymous nucleotide substitutions or radical amino acid replacements. Amino acid composition and replacement frequencies and patterns are essential to identifying how forces such as selection are acting upon proteins. The evolution of Mt proteins has been studied using different classifications of amino acid replacements based on charge, polarity, and size (Blouin et al., 1998). These classifications categorize the replacements into radical (non-similar) and conservative (similar), amino acid replacements. It is generally understood that proteins with different functions or from different genomes have different amino acid replacement patterns resulting in variable radical to conservative replacement (Rad/Cons) ratios (Zhang, 2000). Therefore, it is valuable to demonstrate whether such
replacement patterns are the product of functional constraints or relaxed selective constraints along certain lineages.

In this study I report a comprehensive study and quantification of nucleotide substitution and amino acid replacement types and patterns in Mt genomes of Tetrahymena for 22 KPC and 22 Ymf genes. I investigate the relationships between the distance, \( dn/ds \), and \( \omega \) ratios as well as nucleotide and amino acid substitutions. I utilize such relationships to quantify selection and explain the reasons for failing to assign function for the Ymf genes. I also present which types of substitution analyses may be beneficial in explaining the evolutionary forces operating on Tetrahymena Mt genomes.

Materials and Methods

Gene arrangement, homology, nucleotide substitution and amino acid replacement

*T.*paravorax (GenBank Acc. \# DQ927304), *T.*malaccensis (GenBank Acc. \# DQ927303), and *T.*pigmentosa (GenBank Acc. \# DQ927305) species were obtained from ATCC and their Mt genome was sequenced as was explained in chapter 2 of this dissertation. Complete Mt genomes of *T.*paravorax, *T.*pigmentosa, and *T.*malaccensis were aligned with the genomes of previously sequenced *T.*pyriformis (GenBank Acc. \# AF160864), and *T.*thermophila (GenBank Acc. \# AF396436) at both DNA and amino acid levels using CLUSTAL X (Thompson et al., 2997; Higgins and Sharp 1998, 1999). The degree of similarity was shown using similarity metric (Z score) calculations on ORFs from Mt genomes sequenced in this project with that of previously sequenced *T.*pyriformis, and *T.*thermophila. The Z-score is obtained by comparing the original

87
alignment score of two sequences with an average score obtained form 1000 alignments of randomized sequence of the two original sequences. The randomized alignment scores are not normally distributed and follow an extreme value distribution, yet after 1000 permutations the distribution approximates to a normal distribution (Mount, 2001). The difference between the alignment scores is divided by the SD of the randomized alignment score distribution where the scores greater than 6 are indicative of homology between two sequences (Doolittle, 1990; Pearson, 1990; Mount, 2001). The Z-scores were calculated via software developed in our laboratory (for more explanation of the models see Brunk et al., 2003). The BLAST network services provided at the National Center for Biotechnology Information, tools at the European Bioinformatics Institute, tRNAscan-SE server, and 3D protein structure prediction servers were used for sequence similarity searches to identify genes, proteins, and tRNAs ((Altschul et al., 1990; Lowe and Eddie, 1997; Kelly et al., 2000; Lopez et al., 2003). DNA from Mt genes from T.paravorax, T.pigmentosa, and T.malaccensis were aligned with sequences from T.pyriformis and T.thermophila using CLUSTAL X. The CLUSTAL X gene alignments were checked by eye to obtain correct outputs in PHYLIP format. There were 10 different alignment combinations for each gene, however an average value is reported in results. The DNA and protein distances were calculated using the DNADIST and PROTDIST applications from the PHYLIP package (Felsenstein 1989, 1993). The distances were calculated based on Kimura’s two-parameter model with correction of multiple substitutions (Kimura, 1980). For comparison, the DNA distances were alternatively obtained from LogDet, and the protein distances from JTT programs
available in the PHYLIP software package ((Felsenstein 1993). The relative-rate of
substitutions were calculated using pairwise distances obtained for genes from five
Tetrahymena Mt genomes using *T. paravorax* as the outgroup (Yang and Nielsen, 2000).
The transitional and transversional differences, the Ts/Tv ratios were calculated by
software developed based on the “K80” Kimura model (Kimura, 1980). The dn/ds ratios
of nucleotide substitutions were calculated using a program from the Los Alamos
National Security (Nei and Gojobori, 1986; Korber, 2001). This program was based on
Nei and Gojobori methods for estimating the numbers of synonymous and
nonsynonymous nucleotide substitutions. Tajima’s *D* statistical analysis (Tajima, 1989)
was carried out using DnaSP (Rozas et al., 2003). Tajima’s test is based on the fact that
under the neutral model estimates of the number of segregating/polymorphic sites and of
the average number of nucleotide differences are correlated. If the value of *D* is too large
or too small, the neutral ‘null’ hypothesis is rejected. Hence it can be used to detect rare
alleles, positive selection and population bottlenecks. Positive selection is implied by
negative values (Yu et al., 2005). Above explained models and programs used here to
quantify the nucleotide substitutions correct for multiple substitutions according to their
authors. The amino acid replacement and conservation patterns were studied where two
aligned amino acid sequences in PHYLIP format were compared and the amino acid
replacements were quantified. The replacement patterns were analyzed under two
different classifications based on charge, and volume and polarity. Classification 1 (by
charge) divided the amino acids into three categories: positive (R,H,K), negative (D,E),
polarity) divided the amino acids into six categories: special (C), neutral and small 
(A,G,P,S,T), polar and relatively small (N,D,Q,E), polar and relatively large (R,H,K),
nonpolar and relatively small (I,L,M,V), and nonpolar and relatively large (F,W,Y)
(Yang et al., 1998). Intra-group replacements were considered as conservative and inter-
group ones as radical replacements. The Rad/Cons ratios were calculated by using the
amino acid alignments and the conversion categories in classification 2 and were
normalized by the length of the protein. Assuming random nucleotide substitution in all
possible codons using classification 2 the ratio of radical to conservative replacements
for the first, second, and third codon position changes are 17:19, 25:3, and 5:1
accordingly.

**PAML and Sliding window programs**

The codonml software from PMAL package was used to determine the dn/ds rate ratios
(\(\omega\)) for individual codon site (Yang, 1997). The minimum number of sequences
recommended for PAML is 4 or 5 if the divergence is optimal. *Tetrahymena*
mitochondrial genes are fairly divergent and they generated acceptable results. For
sliding window analysis, I first developed a software program, which uses a sliding
window approach to simultaneously calculate distance, dn/ds, Ts/Tv and Rad/Cons
ratios. This program can accept variable window and step sizes depending on the length
of the analyzed gene and return the desired ratios for that specific length. To optimize the
window and step sizes each gene was analyzed using several different combinations of
window length and step sizes starting with the smallest step size of three. The window
length and step size that generated the least noise and revealed the substitution hotspots was chosen as the most optimum. Calculated values were on average based on a window of 180 and a step size of 30 nucleotides yet in a few occasions they varied depending on the gene length. For example, the window size used for Ymf77 (more than 4000bp in length) was 270 with a step size of 90, which considerably reduced the graph noise. On the contrary for Rps 19 (less than 300bp) the window size of 90 and step size of 21 was used to obtain optimum results. With the exception of dn/ds ratio all other variables were calculated by this software using models explained above. The dn/ds ratio was calculated by calling the program developed at LANS (referenced above). The minimum number of sequences for this program was 2 yet more sequences would increase the reliability of the results. All software developed for this study was coded in perl script and is available upon request. To confirm our results from sliding window analysis we used a software package called SWAPSC, which similarly uses a sliding window analysis procedure to calculate selective constraint (Fares, 2004). SWAPSC automatically optimizes the window size based on a randomized sequence of the input genes with a maximum window size of 20 codons per permutation. This test detects significant selective constraints at specific codon regions of a protein alignment in single branches of a protein. The program detects positive selection, mutation hot spots, saturation of synonymous substitution, negative selection, accelerated rates of nonsynonymous substitution by calculating dn, ds, and \( \omega \). A detailed table, which could explain different mutational dynamics based on the values of dn, ds, and \( \omega \) was used to determine which one of the mechanisms mentioned above explains the substitution patterns. This program
has no limitations for the number of sequences in the alignment yet for the sequences shorter than 200 amino acids it is recommended to use more than the minimum two sequences (Fares, 2004).

**Results and Discussion**

**Nucleotide substitutions in KPC and Ymf genes**

*Ymf* genes in *Tetrahymena* Mt genome are in conserved ORFs, and are homologous among all five *Tetrahymena* species, based on a similarity metric test (for review see Brunk et al., 2003). Therefore they are certainly under selection. An incomplete cDNA library of *Tetrahymena* Mt genome, which contained 16 mRNA sequences, indicated that at least four of these *Ymf* genes were transcribed in *T.pyriformis* (Edqvist, 2000). The codon usage between the KPC and *Ymf* genes in *Tetrahymena* Mt genomes was extremely similar with very little variation (ANOVA pvalue=0.9988), however there were three exceptions in *Ymf* 69, 74, and 71 genes. These genes were diverging rapidly and had elevated dn/ds ratios throughout their sequence such that a few pairwise comparisons failed the Z-score test (Z-scores of <6) by a small margin. These *Ymf* genes were relatively short genes (average length of 110 amino acids) where the entire sequences have diverged so rapidly that they failed to indicate homology. This was also the case for most of the mutation hot spots in *Ymf* genes (described in the next section) where they failed to indicate strong homology, when considered separately, after a Z-score analysis.
To study the nucleotide substitution patterns we determined the relative rate of nonsynonymous substitutions for all Ymf and KPC genes using T.paravorax sequences as outgroup. The substitution rates should be similar for any mutation in these genes, the difference is in the rate at which these mutations are fixed. We found no clear division line or conclusive difference between the Ymf and KPC gene groups regarding their nonsynonymous nucleotide substitutions rates. For example Nad10, which is a very conserved KPC gene, contained extremely similar rates of substitutions close to 1, and so did the less conserved Ymf61 and Ymf68.

\textit{Analysis of dn/ds and their rate ratio (\omega)}

To investigate how nucleotide and amino acid replacement types correlated during the evolution of the Tetrahymena Mt genes we compared the dn/ds ratios using the entire sequence for each gene. Pairwise comparisons indicated that Ymf genes on average accumulated more nonsynonymous substitutions resulting in almost three times higher dn/ds ratios than KPC genes (0.48 vs. 0.17). None of the Ymf\textit{s} had an average dn/ds > 1, therefore we suspended the possibility of positive selection pending further analysis.

Thus to analyze the main reasons for higher dn/ds values in Ymf genes we introduced a sliding window software program, which calculated nucleotide substitutions as distances and dn/ds ratios for each window. We confirmed our analysis with an alternative program called SWAPSC (see material and methods). These programs completed the tasks by showing the relationships between dn, ds, and their ratio. This relationship was quite clear for KPC genes where they had very low dn and much higher ds values.
resulting in low dn/ds ratios (Figure 1). Conversely the Ymf genes contained regions with elevated numbers of nonsynonymous substitutions resulting in higher dn/ds ratios. These regions, which were present in almost all of the Ymf genes and in small regions of a few KPC genes such as Nad5, Cox2, and Rps3, could be considered as mutation hotspots. An illustration of these hotspots in Ymf genes is shown in figure 2, where they were compared to other regions of the Ymf genes to show that dn/ds ratios in these genes ranged from almost zero to over 1.5. On the other hand, the conserved regions of Ymf genes had dn/ds ratios comparable to that of KPC genes, which could potentially represent the functional domains of their protein products.

A previous study suggested that in Mt genomes of C. elegans the dn/ds ratio increased by more than five fold when the effects of natural selection were minimized (Denver et al., 2000). Our analysis, which showed on average an almost three fold increase in the Ymf dn/ds ratios, seems to support such a conclusion. However increased dn/ds ratios in C.elegans Mt genomes occurred throughout the entire Mt genome with little spatial preference. If minimized natural selection was the reason for increased dn/ds ratios in Tetrahymena Mt genomes then these ratios should have increased in most if not all of the KPC and Ymf genes. Yet, our analysis of dn/ds ratios using a sliding window program, which revealed substitution variations in these genes in detail, did not quite support minimized natural selection throughout the Tetrahymena Mt genome. The rapid divergence of Ymf genes and elevated dn/ds ratios were primarily due to presence of regions with accelerated rates of nonsynonymous mutations (AdN), which were detected by SWAPSC.
Figure 1 - Relationship between dn/ds ratios, dn, and ds values for KPC genes. Values for each point are average comparison of *T. thermophila* and *T. pyriformis*. They are from average window of size 180 sliding 30 nucleotides per permutation.
Figure 2: Relationship between \( \text{dn/ds} \) ratios, \( \text{dn} \), and \( \text{ds} \) values for *Ymf* genes. Values for each point are average comparison of *T.thermophilia* and *T.pyriformis*. They are from average window of size 180 sliding 30 nucleotides per permutation.
In addition to identifying regions with AdN we were able to locate regions under positive selection, mutation hot spots, regions with saturated synonymous substitution, and negative selection at specific codon regions. The significant results of SWAPSC output, which identified the variable regions in some $Ymf$ genes that seemed to contain mutation hotspots with accelerated nonsynonymous substitutions are shown in Figure 3. SWAPSC also uses a sliding window approach yet the program itself determines the most appropriate window size with a maximum of 20 amino acids per window. Hence there are some $d_{n/ds}$ value differences in figure 1 vs figure 3, which are due to usage of smaller window sizes and $\omega$ in SWAPSC. There were a few small regions in some $Ymf$ genes that suggested positive selection yet they could not be considered as a major cause for such an extensive variation. Hence we reconsidered the possibility that these variable regions of $Ymf$ genes, could be under positive selection. To confirm our argument we determined the $d_{n/ds}$ rate ratios in genes with accelerated nonsynonymous substitutions for each codon site using software from PAML package (see material and method). The codonml software in this package revealed likelihood ratios of positive selection or relaxed selective constraints along lineages based on the $d_{n/ds}$ rate ratios ($\omega$) per individual codon for $Ymf$ genes from all five genomes. The significant increases in $\omega$ were observed in some $Ymf$ genes and Nad5 (Figure 4). In sum, results from four different software packages, which determined $d_{n/ds}$, $\omega$, mutation hotspots, accelerated rates of nonsynonymous mutations, and positive selection, indicated that the primary reason for presence of variable regions in $Ymf$ genes were accelerated rates of nonsynonymous mutations.
Figure 3 - dn/ds ratio Variation in Ymf genes

A - Ymf64 (5' end 327bp)

B - Ymf68 (5' end 348bp)

C - Ymf69 (204bp)

D - Ymf71 (303bp)
Figure 3- dn/ds ratio Variation in Ymf genes

Plots A, B, and F are the 5' avriable region of Ymf 64, 64, and 76 genes.
Plots C, D, E, G, and H present variable regions throughout Ymf 69, 71, 74, 67, and 77 genes.
AHS: mutation hot spots, S: saturation of synonymous substitutions, NS: negative selection,
AdN: accelerated rate of nonsynonymous substitutions, PS: positive selection.
Y axis represents values for dn/ds ratios. X axis indicates types of variation for each codon.
Figure 4: dN/dS ratio for each codon site

The non-synonymous substitution rate ratios for all Ymr genes. For comparison the same is shown for Nad5 (with a highly variable region) and Nad10 (most conserved KPC gene). Rate ratios > 1 represent regions which could be under positive selection.
However cases of small sites under positive selection were present in parts of the Ymf 57, 60, 61, 64, 67, 68, 71, 74, 76, and 77 where the dn/ds were elevated and \( \omega > 1 \). We also calculated Tajima’s \( D \) values for KPC and Ymf genes to detect selection. We found negative \( D \) values in all KPC and Ymf genes. But the significantly negative \( D \) values in variable regions of Ymf 57, 60, 61, 64, 67, 68, 71, 74, 76, and 77 stood out (Figure 5). These results were consistent with the results from regions presumably under positive selective pressure based on dn/ds and \( \omega \). We also found positive selection in the 5′ region of the Nad5 gene with significant dn/ds, \( \omega \), and Tajima’s \( D \) values. Such variable regions with AdN along with more substitutions in Ymf genes could be the cause for our inability to find homologues for them. Thus we conclude that substitution types, numbers, patterns, and fixation rates support the idea that variable regions with AdN in Ymf genes in Tetrahymena Mt cause them to evolve so rapidly that they could not be assigned definitive functions based on sequence similarity or homology. Also presence of sites under positive selection in some Ymf genes contributed to such rapid evolution. The presence of regions with AdN in a few KPC genes (e.g., Nad5) did not weaken our argument since, unlike in Ymf genes, the remaining regions of the aforementioned KPC genes were highly conserved and had preserved their ancestral sequence (Figure 3 and 4).
Figure 5- Tajima's D values for all Ymf and Nad5 genes

*** denotes significant Tajima's negative D values which represent regions under positive selection.
Amino acid composition and replacement patterns in Tetrahymena Mt genomes

On the amino acid level pairwise comparisons of Rad/Cons ratios for all 44 genes in Tetrahymena Mt genomes suggested that there was on average an almost two-fold increase in Rad/Cons ratios for Ymf s (1.20) compared to KPC genes (0.67). Since radical amino acid replacements are more likely to alter protein structure their presence would mean more structural variation in any protein. An independent sliding window analysis for Rad/Cons amino acid replacement patterns indicated that radical substitutions occurred with higher frequency in parts of the protein where the dn/ds ratios were elevated (data not shown). If elevated Rad/Cons ratios in Ymf genes were the result of accelerated rates of nonsynonymous substitutions in mutation hotspots then there would be a positive correlation between these ratios. Analysis of the relationship between dn/ds and Rad/Cons ratios among the Ymf genes indicated that there was a significant positive correlation (r=0.93, p-value < 0.05) between these two ratios (Figure 6). Alternatively, when such correlation was analyzed for KPC genes the correlation coefficient (r) value dropped to 0.73 (r=0.73, p < 0.05), which suggested a weaker relationship between dn/ds and Rad/Cons ratios among KPC genes. This weaker correlation was due to the absence of regions with AdN in KPC genes. Relaxed selective constraint or nonfunctional regions in these proteins could explain this strong correlation where more nonsynonymous substitutions resulted in more radical replacements and consequently more diverged protein structure. The overall amino acid conservation in KPC genes was over 80% compared to significantly lower conservation of 57% in Ymf genes (Table 1). This difference came from the greater number of amino acid replacements in regions...
with AdN, hence a more detailed study of amino acid replacement types and patterns between KPC and Ynf genes was conducted. According to different amino acid replacement classifications (see materials and methods) there were far fewer radical replacements in KPC genes relative to Ynfs when orthologous proteins were compared. In classification 1 (based on charge) the KPC genes were quite conserved where 11% of accrued replacements were radical in transmembrane and 18% in ribosomal proteins. Conservation was less noticeable in classification 2 (based on polarity and size) where about 40% of the amino acid replacements in KPC genes were radical (Table 1). Insertions and deletions (indels) did not play a major role and comprised approximately 0.5% of the replacements. In Ynf genes, besides a significant increase in the number of amino acid replacements, there was a significant increase in the percentage of radical replacements. The radical amino acid replacements in Ynf genes for classification 1 and 2 were 80 and 40 percent higher respectively, relative to KPC genes. Radical replacements are assumed to be more likely than conservative replacements to alter the structure and function of a protein based on compositional factors and may not be used to infer positive selection (Dagan et al., 2002).
Figure 6- Correlation graph for dn/ds vs. Rad/Cons ratios for Ymf genes.

Each point is based on average pairwise comparison of Ymf genes. Total of 10 comparisons per orthologous gene using five mt genomes.
Table 1. Total amino acid conservation and replacement percentages

<table>
<thead>
<tr>
<th>Tetrahymena Mitochondria</th>
<th>KPC Genes</th>
<th>Ymf Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Change Classifications/Protein Class</td>
<td>Tm</td>
<td>Rp</td>
</tr>
<tr>
<td>Length of AA sequences Compared</td>
<td>44050</td>
<td>12838</td>
</tr>
<tr>
<td>Total AA Conservation (%)</td>
<td>84.5</td>
<td>77.9</td>
</tr>
<tr>
<td>Total Conservative Changes Classification 1(%)</td>
<td>89</td>
<td>82</td>
</tr>
<tr>
<td>Total Radical Changes Classification 1(%)</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Total Conservative Changes Classification 2(%)</td>
<td>60</td>
<td>56</td>
</tr>
<tr>
<td>Total Radical Changes Classification 2(%)</td>
<td>40</td>
<td>44</td>
</tr>
<tr>
<td>Positions Containing Gaps(%)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Classification 1 Conservative Changes

<table>
<thead>
<tr>
<th></th>
<th>KPC</th>
<th>Ymf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive &lt;-&gt; Positive changes (RHK)</td>
<td>1.5</td>
<td>5.9</td>
</tr>
<tr>
<td>Negative &lt;-&gt; Negative changes (DE)</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Uncharged &lt;-&gt; Uncharged changes (ANCQGILMFPTWYY)</td>
<td>97.3</td>
<td>92.6</td>
</tr>
</tbody>
</table>

Classification 1 Radical Changes

<table>
<thead>
<tr>
<th></th>
<th>KPC</th>
<th>Ymf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive &lt;-&gt; Negative changes</td>
<td>4.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Positive &lt;-&gt; Uncharged changes</td>
<td>51.6</td>
<td>76.7</td>
</tr>
<tr>
<td>Negative &lt;-&gt; Uncharged changes</td>
<td>44.2</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Classification 2 Conservative Changes

<table>
<thead>
<tr>
<th></th>
<th>KPC</th>
<th>Ymf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special &lt;-&gt; Special changes (C)</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Neutral Small &lt;-&gt; Neutral Small changes (AGPST)</td>
<td>12.8</td>
<td>16.7</td>
</tr>
<tr>
<td>Polar Small &lt;-&gt; Polar Small changes (NDQE)</td>
<td>7.2</td>
<td>6.7</td>
</tr>
<tr>
<td>Polar Large &lt;-&gt; Polar Large changes (RHK)</td>
<td>2.2</td>
<td>8.8</td>
</tr>
<tr>
<td>Non-Polar Small &lt;-&gt; Non-Polar Small changes (ILMV)</td>
<td>69.9</td>
<td>58.6</td>
</tr>
<tr>
<td>Non-Polar Large &lt;-&gt; Non-Polar Large changes (FWY)</td>
<td>7.9</td>
<td>9.2</td>
</tr>
</tbody>
</table>

Classification 2 Radical Changes

<table>
<thead>
<tr>
<th></th>
<th>KPC</th>
<th>Ymf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special &lt;-&gt; Neutral Small changes</td>
<td>1.4</td>
<td>2.4</td>
</tr>
<tr>
<td>Special &lt;-&gt; Polar Small changes</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>Special &lt;-&gt; Polar Large changes</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Special &lt;-&gt; Non-Polar Small changes</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Special &lt;-&gt; Non-Polar Large changes</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>Neutral Small &lt;-&gt; Polar Small changes</td>
<td>19.2</td>
<td>19.4</td>
</tr>
<tr>
<td>Neutral Small &lt;-&gt; Polar Large changes</td>
<td>2.2</td>
<td>7.7</td>
</tr>
<tr>
<td>Neutral Small &lt;-&gt; Non-Polar Small changes</td>
<td>19.7</td>
<td>16.5</td>
</tr>
<tr>
<td>Neutral Small &lt;-&gt; Non-Polar Large changes</td>
<td>5.9</td>
<td>4.8</td>
</tr>
<tr>
<td>Polar Small &lt;-&gt; Polar Large changes</td>
<td>8.2</td>
<td>14.7</td>
</tr>
<tr>
<td>Polar Small &lt;-&gt; Non-Polar Small changes</td>
<td>6.9</td>
<td>7.3</td>
</tr>
<tr>
<td>Polar Small &lt;-&gt; Non-Polar Large changes</td>
<td>4.8</td>
<td>4.1</td>
</tr>
<tr>
<td>Polar Large &lt;-&gt; Non-Polar Small changes</td>
<td>2.7</td>
<td>5.3</td>
</tr>
<tr>
<td>Polar Large &lt;-&gt; Non-Polar Large changes</td>
<td>2.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Non-Polar Small &lt;-&gt; Non-Polar Large changes</td>
<td>26.1</td>
<td>11.8</td>
</tr>
<tr>
<td>Non-Polar Large &lt;-&gt; Non-Polar Large changes</td>
<td>33.3</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Transmembrane (Tm) and ribosomal (Rp) proteins in Tetrahymena Mt genomes.
The amino acid replacement frequencies and patterns in two classifications based on charge (classification 1), and polarity and size (classification 2).
For more detail see materials and methods.
Nonsynonymous substitutions and radical to conservative amino acid replacements

To analyze the nonsynonymous substitution patterns in each codon position in *Tetrahymena* Mt genomes, we compared the average frequency of radical and conservative amino acid replacements in *Ymf* and KPC genes in each nonsynonymous codon position (Table 2). Results for KPC genes showed that more than half (54%) of the amino acid replacements caused by a substitution at the second codon position were radical (non-similar). The third codon position had fewer, while the first had the fewest radical replacements. A similar pattern was seen for the *Ymf* genes however with higher (69%) radical replacement percentages (Table 2). When amino acid replacements caused by two or three nucleotide substitutions were added to the analysis, we observed an increase in radical replacement since almost all amino acid replacements caused by these codons were radical. Although the second codon position had the fewest number of nonsynonymous transversions (data not shown), the frequency of radical replacements was higher than the other codon positions suggesting that transversions did not associate with radical replacements in *Tetrahymena* Mt genomes. More radical replacements

| Table 2. Average frequencies for radical and conservative amino acid replacements |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                 | KPC genes       |                 | Ymf genes       |                 |
| Replacement                     | Position1 position2 position3 | All | Position1 position2 position3 | All |
| Radical                         | 80              | 99              | 69              | 247             | 284 | 353 | 231 | 868 |
| Conservative                    | 254             | 84              | 90              | 428             | 453 | 154 | 219 | 826 |
| Sum                             | 333             | 183             | 159             | 675             | 737 | 507 | 450 | 1694|
| % Radical                       | 24              | 54              | 43              | 37              | 39  | 69  | 51  | 51  |
| % Conservative                  | 76              | 46              | 57              | 63              | 61  | 31  | 49  | 49  |

Adding codons with two or three nucleotide substitutions:

| % Radical | 48 | 78 | 52 | 58 | 46 | 75 | 55 | 60 |
| % Conservative | 52 | 22 | 48 | 42 | 54 | 25 | 45 | 40 |

Data from pairwise comparison of *Ymf* and KPC genes in *Tetrahymena* Mt genomes; codons with one nucleotide substitution (top); codons with two and three nucleotide substitutions added (bottom).
occurred in $Ynf$ genes, 51% on average, than in KPC genes (37%), which was the case for all three codon positions (Table 2). If nucleotide substitutions were to occur randomly (equally for all bases) the ratio of radical to conservative amino acid replacements for first, second, and third codon positions would be 1:1, 8:1, and 5:1 respectively. Although none of the ratios for KPC and $Ynf$ genes were close to that of random patterns, increases in radical replacements in all codon positions suggested that the majority of $Ynf$ genes could be under relaxed selection where they accrued more radical replacements.

**Conclusion and Future Plans**

In sequencing and assembling of *Tetrahymena* Mt genomes I faced a few minor obstacles of strong secondary structures and nuclear contamination, however the overall final sequences are quite accurate and reliable. The sequence data from this study are an invaluable addition to previously available *Tetrahymena* Mt genomes to study evolutionary and functional genomics elements. One of our main goals was to gain insights regarding the reasons for failing to find definitive homologues for $Ynf$ genes. Our analysis suggested that $Ynf$ genes contain mutation hotspots and regions with accelerated rates of nonsynonymous substitutions. We also found relatively shorter regions under positive selection in some $Ynf$ genes. Thus we concluded that a major obstacle in identifying function for $Ynf$ genes was the presence of regions in these genes, which evolved more rapidly relative to the rest of the gene. In a few shorter $Ynf$ genes
the entire gene was evolving rapidly. A plausible explanation could be relaxed selective constraint in these variable regions portraying them as nonessential to the function of the protein. Although presence of sites under positive selection could also account for failing to identify function for \textit{Ymf} genes, their relatively smaller numbers and lengths did not lay sufficient grounds to be considered as a major reason. Also, having five complete Mt genomes allowed us to find a transcription control region in \textit{Tetrahymena} mitochondria through comparative genome analysis. This project allowed us to develop reliable picture for \textit{Tetrahymena} Mt genome organization and conduct molecular evolution analysis. Addition of more complete Mt genomes of \textit{Tetrahymena} and other closely related species genera such as \textit{Glaucoma} will certainly enhance these evolutionary studies and provide more information for studying ciliate Mt genomics.

**References**


115


Chapter 4

Phylogenetic analysis using complete mitochondrial sequences of five *Tetrahymena*

species suggest a slightly different topology
Abstract

The phylogenetic relationships between *Tetrahymena* species have been previously determined using intergenic regions from histone H3 and H4 genes. This topology suggested that *T.pyriformis* forms a natural clade with *T.thermophila* and *T.malaccensis*, and *T.pigmentosa* falls right outside that clade. In this study the addition of three complete mitochondrial genomes from *Tetrahymena* genus allowed me to analyze the phylogenetic relationships among five *Tetrahymena* species (i.e., *T.thermophila* and *T.malaccensis*, *T.pyriformis*, *T.pigmentosa*, and *T.paravorax*) with complete mitochondrial genome sequence. Genes from *P.aurelia* mitochondrial genome were used as outgroup when single genes or concatenation of common genes were used to infer phylogenetic relationships. Complete mitochondrial genome data supported a topology using maximum likelihood (ML) and parsimony methods, which was slightly different than the previous analysis. In this tree *T.pyriformis* and *T.pigmentosa* construct a sister group to *T.thermophila* and *T.malaccensis* clade suggesting that *T.pyriformis* is closer to *T.pigmentosa* than to *T.thermophila* and *T.malaccensis*. When all 44 genes from these genomes were used individually to construct trees a vast majority supported the new topology, however high variability among generated trees suggested that many of the mitochondrial genes were not robust enough to obtain correct phylogeny. Hence I enhanced this analysis by concatenating these genes based on their similarities and common features to seven groups. All concatenated groups supported the new topology using ML method and six groups using the Parsimony method. The final analysis using the entire mitochondrial genome sequence data (with and without intergenic sequences)
also supported the new topology with almost perfect bootstrap values. Therefore I concluded that based on mitochondrial genome data *T. pyriformis* is closer to *T. pigmentosa*, which is slightly in odds with the previous analysis.

**Introduction**

Mitochondrial sequence data have recently been commonly used to infer or resolve phylogenetic relationships among species, genera, families, and orders (Friedrich and Muqim 2003; Sasaki et al., 2005; Steinauer et al., 2005). Mitochondrial genes are inherited as a single unit thus they have a single evolutionary history (Corneli and Ward, 2000), which makes them suitable for phylogenetic analysis. Additionally mitochondrial genes are an attractive marker for inferring the phylogeny of closely related species because of the rapid evolution of the mitochondrial genome and slight chance of recombination (Biswas et al., 2005; Munemasa et al., 2006). Phylogenies based on single short genes (< 2kb) may lack reliability and resolution since such genes may be under different selection and produce different trees (Waits 1996; Zardoya and Meyer 1996). This phenomenon could particularly apply to *Tetrahymena* mitochondrial gene since more than half of them contain highly variable regions, which may affect the construction of correct and reliable phylogenetic trees. However, the source of error is not always easily identified. The most accurate trees are constructed when large amounts of molecular data are used, which is invariably more reliable than when limited data is available. Thus the best way to minimize random error is to use large amounts of data.
Phylogenetic methods will not produce dependable results when sequences fail to supply sufficient phylogenetic information either, because they are short and lack variation or they are hyper-variable. Appropriate rate of evolution is essential for sequences to be considered for phylogenetic analysis. Fast evolving sequences could be used to distinguish close relationships and sequences that evolve slower could answer questions about distantly related species. One such example is the mitochondrial gene coding for cytochrome b, which is extensively used to infer phylogenetic relationships (Wang et al., 1998; Yokoyama et al., 2001). If substitution rates vary considerably among sequences then a model that reflects the actual substitution probabilities should be used to obtain correct trees (Takezaki and Gojobori, 1999).

One way to produce large amounts of data is to combine sequences from different data sets. These sequences should have similar rates of nucleotide substitution or belong to groups of genes that conduct similar functions. This approach could be used when genes from Tetrahymena mitochondrial genome are considered for phylogenetic analysis. There are five completely sequenced Tetrahymena mitochondrial genomes (three produced in this project), which could be used for phylogeny. These genomes contain four major gene groups coding transmembrane proteins, ribosomal proteins, small and large ribosomal RNA, and the Ymf genes that are putative proteins without an assigned function. Thus combined sequences in each category should relatively produce a more reliable phylogeny compared to when just individual genes from Tetrahymena mitochondria are used. However as shown in previous chapter, the rates of nucleotide substitutions among Tetrahymena genes are not significantly different, the differences
come from rates at which these substitutions were fixed. Therefore all gene groups could be combined or simply the entire mitochondrial genomes could be used.

In this chapter I have analyzed the phylogenetic relationship between five *Tetrahymena* species using sequences from their complete mitochondrial DNA. I have determined topologies produced by individual genes, several appropriate combinations, and the entire mitochondrial genome sequences.

**Materials and Methods**

*Sequence data and alignment*

The entire mitochondrial genomes of five *Tetrahymena* species and one *Paramecium* were used to determine their phylogenetic relationship. The *Tetrahymena* species are *T. paravorax* (GenBank Acc. # DQ927304), *T. malaccensis* (GenBank Acc. # DQ927303), and *T. pigmentosa* (GenBank Acc. # DQ927305) (obtained from this project), *T. pyriformis* (GenBank Acc. # AF160864), and *T. thermophila* (GenBank Acc. # AF396436) were obtained from GenBank. The outgroup *P. aurelia* was also obtained from GenBank (Acc. # X15917). Mitochondrial DNA sequences were aligned using CLUSTAL X (Higgins et al., 1998) and the output alignments in phylip format were used for phylogenetic analysis.

*Data analysis*

Maximum Likelihood and Parsimony methods from Phylip package version 3.65 (Felsenstein 1989 and 2005) were used to conduct the phylogenetic analysis. ML uses a
Hidden Markov Model to infer evolutionary rates at different sites so it can compute the likelihood by summing them over all possible assignments of rates to sites, which are weighted by their prior probabilities of occurrence (Felsenstein and Churchill, 1996). I used transition/ tranversion ratios calculated for each gene from the previous chapter so the substitution pattern will be more likely to represent the actual ratios. Parsimony method uses slightly different assumptions and rate probabilities to construct the most parsimonious tree. It assumes that the probability of a base substitution at a given site is small over the time period involved in that branch of phylogeny. It also assumes that variation of expected changes among sites are small so it can find the informative sites and use them to draw the tree with least number of changes or the most parsimonious phylogeny (Kluge and Farris, 1969). This method can handle both bifurcating and multifurcating trees by adding branches in the middle of forks or at the ends of new branches. Therefore in resulted tree a branch exists if only some character has the most parsimonious reconstruction that would involve change in that branch (Kluge and Farris, 1969). ML trees were constructed from each of the 22 known protein coding (KPC) genes, 22 Ymf genes, and from eight different combinations. Several criteria were used to concatenate the genes into these different combinations, which were primarily based on their nucleotide substitution rates and functional similarities. These functional groups were the Nad group genes, electron transport protein group, ribosomal protein group, all KPC gene group. Also the Ymfs were divided into two major groups depending on if they had orthologous genes in *P. aurelia* or not, as well as all Ymfs genes concatenated. The ML method accommodates unequal (empirical) base frequencies and a transition
transversion ratio of 2:1, which assumes rate homogeneity (Ward and Corneli, 2000). Parsimony method assumptions are based on identification of informative sites and their patterns. This program searches for the most parsimonious trees while adding species by creating new forks in the middle of existing branches and also putting them at the end of new branches, which are added to the existing forks (Felsenstein, 1988). Trees were constructed using a randomly different order of species input. The degree to which the data support the formation of each clade was estimated by a majority consensus program using trees built from 100 bootstrap replicates (Felsenstein, 1985). The outgroup was chosen based on their availability from *G. chattoni* and *P. aurelia*. If no orthologous sequence was available *T. paravorax* sequences were assigned as the outgroup.

**Results and Discussion**

*The expected tree*

Phylogenetic relationship among *Tetrahymena* species has been determined using intergenic sequences from Histon H3 and H4 genes (Sadler and Brunk, 1992). Assuming Sadler and Brunk’s tree topology the five species that were used in this project with complete mitochondrial DNA sequence should have had the following phylogenetic relationship (Figure 1). *T. thermophila* and *T. malaccensis*, which are closely related species, group together and make a clade with the next closest species, which is *T. pyriformis*. Hence based on intergenic sequences from Histon H3 and H4 genes *T. pyriformis* is more closely related to *T. thermophila* and *T. malaccensis* group than to *T. pigmentosa*, which appears to be outside this clade. Such a relationship is shown in
figure 1 for the five *Tetrahymena* species with *G.chattoni* as an outgroup. Although this tree is based on intergenic sequences between Histone H3 and H4 it still could be used as a reference and a start point for comparison.

```
+---------------------T.pyriformis
|                     |
|                     | +-----T.thermophila
+---------------------+-----T.malaccensis

+---------------------T.pigmentosa
|                     |
+---------------------T.paravorax

+---------------------G.chattoni
```

Figure 1- Phylogenetic tree based on Histone H3 and H4 intergenic sequences (taken from Sadler and Brunk, 1992). For simplicity only five *Tetrahymena* species with complete mitochondrial genomes are shown. The outgroup is *Glaucoma chattoni*.

**Tree topologies**

**Single gene data**

*Tetrahymena* mitochondria contain 22 genes, which are known protein coding and 22 genes without an assigned function called *Ymfs*. Each one of these genes has been used individually to assess the phylogenetic relationship among five *Tetrahymena* species with complete mitochondrial sequence. The results suggest that only 2 genes using maximum likelihood (ML) method and 4 genes using Parsimony method support a topology proposed by Sadler and Brunk in figure 1. On the other hand 26 genes (ML method) and 35 genes (Parsimony method) generate a slightly different tree. These genes are *cob, cox1, nad1, nad2, nad4, nad5, nad9, Rps14, Rps13, Rpl2, LSUrRNA, SSUrRNA,*
Ymf59, Ymf60, Ymf61, Ymf63, Ymf65, Ymf66, Ymf67, Ymf68, Ymf69, Ymf72, Ymf74, Ymf75, Ymf76, and Ymf77. They support a tree (called the alternative tree hereafter) where *T. pyriformis* and *T. pigmentosa* construct a sister group to *T. thermophila* and *T. malaccensis* clade suggesting that *T. pyriformis* is closer to *T. pigmentosa* than to *T. thermophila* and *T. malaccensis*. In this alternative tree *T. paravorax* is placed outside the clade with the remainder of the species (Figure 2A). There were genes that constructed 3 unusual topologies (Figure 2B-2D) due to their highly divergent or highly conserved sequences. They also failed to generate high bootstrap values, which further emphasized that they were unlikely to represent the correct phylogenetic relationship. The first group of these genes, which were *atp9, cox2, nad3, nad7, Rps3, Ymf58, Ymf64*, grouped *T. paravorax* with *T. pyriformis* and *T. pigmentosa* with an average bootstrap value of 45, which is not a significant value and decreases the reliability of this topology (Figure 2B). There were also 3 genes (ML method) that group *T. paravorax* with *T. thermophila* and *T. malaccensis* (Figure 2C). These genes, which also fail to produce a significant bootstrap value (i.e. 42), were *Rpl14, Rpl16, and nad10*. The next generated tree, which is constructed by 3 genes (i.e., Ymf56, Ymf57, and Ymf62) groups *T. pigmentosa* with *T. thermophila* and *T. malaccensis* clade (Figure2D). This tree topology is also less probable since it is quite different than the reference tree and only supported by 3 Ymf genes, which are variable genes with no definitive function. The remaining gene (i.e, *Ymf71*) generated an outrageous tree by separating *T. thermophila* and *T. malaccensis* and grouping them with other species. This tree is highly improbable and not reliable since mitochondrial genomes of these two species contain extremely
similar sequences and are almost identical. *Ymf71* is the most divergent gene in *Tetrahymena* mitochondria and substitutions are prevalent across its 303bp length (for reference see chapter 3). On the other hand *nad10*, which is the most conserved gene and its nucleotide substitution rates are quite constant with very few substitutions, generates a highly improbable tree as well. These two genes are good examples of genes that are not suitable for phylogenetic analysis on individual basis. Despite the fact that majority of topologies generated based on individual gene data supported the alternative tree it was not sufficient to draw conclusions and explain the reasons. This also suggests that phylogenetic analysis using individual genes from *Tetrahymena* mitochondrial genome may not generate reliable and accurate topologies due to variable patterns of divergence and conservation across the genome. Therefore concatenation of genes may produce less variable and more reliable topologies.

![Phylogenetic tree]

2A- Phylogenetic relationship constructed by *cob*, *cox1*, *nad1*, *nad2*, *nad4*, *nad5*, *nad9*, *Rps14*, *Rps13*, *Rpl2*, *LSUrRNA*, *SSUrRNA*, *Ymf59*, *Ymf60*, *Ymf61*, *Ymf63*, *Ymf65*, *Ymf66*, *Ymf67*, *Ymf68*, *Ymf69*, *Ymf72*, *Ymf74*, *Ymf75*, *Ymf76*, *Ymf77*.
2B- Phylogenetic relationship constructed by *atp9, cox2, nad3, nad7, Rps3, Ymf58, Ymf64.*

2C- Phylogenetic relationship constructed by *Rpl14, Rpl16, nad10.*

2D- Phylogenetic relationship constructed by *Ymf56, Ymf57, Ymf62.*

Figure 2- Phylogenetic trees constructed using 44 genes from *Tetrahymena* mitochondrial genome and orthologous gene as outgroup from *P. aurelia* if available. Average bootstrap values shown are shown for each topology.
Concatenated gene data

Mitochondrial genes could be concatenated based on their function, variability, and conservation across species (Takezaki and Gojobori, 1999). I divided *Tetrahymena* mitochondrial genes into two major categories, which are known protein coding (KPC) and *Ymf* genes based on their conservation and function. The KPC genes were divided into ribosomal proteins, transmembrane proteins (Tm), and small and large ribosomal RNA units. The Tm group was further divided into Nad genes group and Atp9, Cox1, Cox2 group. Orthologous genes from *P. aurelia* mitochondrial genome were used as outgroup for trees constructed using KPC genes. On the other hand, the *Ymf* genes were not divided into any groups since some of the *Ymf* genes shared orthologues with *P. aurelia* mitochondrial genes. Therefore I was able to study the phylogenetic relationships among five *Tetrahymena* species by constructing 7 trees from aforementioned concatenated data. As for single genes ML and Parsimony were the two methods used to construct these trees.

As it appears in figures 3A-F sequence data from all different concatenations from 22 KPC genes described above support two distinct sister clades, one with *T. thermophila* and *T. malaccensis* and the other with *T. pyriformis* and *T. pigmentosa*. These sister groups together form a natural clade with *T. paravorax* suggesting that the alternative tree was the tree of choice for concatenated mitochondrial data (Figures 3A-E). Furthermore the concatenation of 22 *Ymf* genes also generated the alternative tree topology (Figure 3F). Although they all generated the same topology the bootstrap values are different hence I prefer to show all generated trees. This consensual topology with highly reliable
bootstrap values was obtained from both ML and Parsimony methods, which is slightly in odds with the reference tree (Figure 1) mentioned above. However there was an exception when analyzing the small and large ribosomal RNA genes with Parsimony method. This group generated a tree similar to the reference topology (figure 1) where *T. pyriformis* does not form a clade with *T. pigmentosa* and appears to be closer to *T. thermophila* and *T. malaccensis* (Figure 3G). Since only concatenated small and large ribosomal RNAs generated this topology then one could suggest that selection plays a decisive role on trees generated from mitochondrial data. In fact ribosomal RNA genes have been used extensively in phylogenetic analysis since they were thought to evolve in appropriate rates and are also universally present in almost all organisms. Yet an issue with them could be that they might be under more selection than the rest of the genome and generate inaccurate topologies, which would only reflect their evolution.

Nevertheless I showed that use of concatenated data from mitochondrial genome almost eliminated the false positive topologies and resulted in two acceptable phylogenetic trees.

```
+-------------------T. malaccensis
|                  +100.0-|
|                  +------T. thermophila
|                       +100.0-|
|                       +-----T. pigmentosa
|                               +87.0-|
|                               +------T. pyriformis
|                                           +----------T. paravorax
|                                           +-----------------P. aurelia
```

3A- Phylogenetic relationship constructed by concatenated 22 KPC genes

129
3B- Phylogenetic relationship constructed by concatenated \textit{Atp9, Cob, Cox1, Cox2} genes

3C- Phylogenetic relationship constructed by concatenated \textit{Nad1, Nad2, Nad3, Nad4, Nad5, Nad7, Nad9, Nad10} genes

3D- Phylogenetic relationship constructed by concatenated ribosomal proteins
3E- Phylogenetic relationship constructed by concatenated transmembrane proteins

3F- Phylogenetic relationship constructed by concatenated 22 Ymf genes

3G- Phylogenetic relationship constructed by small and large ribosomal RNA genes

Figure 3- Phylogenetic trees constructed using concatenated genes from Tetrahymena mitochondrial genome and orthologous gene as outgroup from P.aurelia if available. Average bootstrap values shown are shown for each topology.
**Complete mitochondrial genome data**

As final analysis the entire mitochondrial genome sequence from five *Tetrahymena* species were aligned and used to study phylogenetic relationships. Both ML and Parsimony methods support, the most popular, alternative tree where *T.pyriformis* and *T.pigmentosa* form a separate clade (figure 4A). Although the intergenic sequences were fairly short in *Tetrahymena* mitochondrial genome they were eliminated from the genomes leaving only concatenated genes, which also generated the alternative tree with almost identical bootstrap values (Figure 4B). Interestingly these trees had the highest bootstrap values among trees generated by all analyzed sequence data, which allows me to draw a reliable conclusion.

![Phylogenetic tree](image)

**Figure 4A- Phylogenetic relationship constructed by entire mitochondrial genome sequence data**

![Phylogenetic tree](image)

**Figure 4B- Phylogenetic relationship constructed by entire mitochondrial genome sequence data without the intergenic sequences**
Summary and Conclusion

In this chapter I present the phylogenetic analysis of *Tetrahymena* species based on their complete mitochondrial genome. Using several combinations of concatenated data as well as the entire mitochondrial genome data I showed that species used in this project form two distinct clades of *T. thermophila*, *T. malaccensis* and *T. pyriformis*, *T. pigmentosa*. These sister groups together form a natural clade with *T. paravorax* suggesting that this topology was the tree of choice for concatenated mitochondrial data. Therefore I conclude that phylogenetic analysis using complete mitochondrial genomes from 5 *Tetrahymena* species generates a topology, which is slightly different than the only available reference tree. The reference tree (Sadler and Brunk, 1992) was constructed using 375bp of intergenic regions between Histone H3 and H4 genes from 29 *Tetrahymena* species. Histone genes are known to be under strong selective pressure and since their intergenic region is quite short, it could also be under strong selection. On the other hand *Tetrahymena* mitochondrial genomes are fairly divergent with half of their genes remaining without and assigned function. Also there are only 5 *Tetrahymena* species with complete mitochondrial genomes. Hence I conclude that both reference topology based on Histone intergenic sequences and the alternative tree generated by mitochondrial genomes could be authentic. The difference may be due to the number of species used and the intensity of selection on the sequences. After all the Histone intergenic and the ribosomal RNA genes in mitochondria, also under strong selection, generated the reference tree. One way to resolve this phylogenetic paradox and to obtain
correct topology is to sequence more *Tetrahymena* mitochondrial genomes, preferably all 29 of them.

References


Chapter 5

Homology based annotation of a Tetrahymena thermophila chromosome:
Correlation between the BLAST E-values and the similarity metric Z-scores
Abstract

The function of a newly sequenced gene may be predicted on the basis of its homology with a known gene in a database. In this study we have annotated a Tetrahymena thermophila chromosome, 1Mb in size, based on the translation of its possible open reading frames (ORFs) and their homologues in the GenBank. We identified 5697 ORFs of length greater than 100 amino acids and searched for their homologues in a non-redundant protein database using BLAST. This resulted in the identification of 1011 putative protein-coding ORFs with E-values less than 10. These ORFs are almost equally transcribed from each of the opposite strands of this chromosome. A similarity metric (Z-score) was used to identify potential homologues between T. thermophila ORFs and the proteins in GenBank. We also identified homologous for predicted proteins from The Institute for Genomic Research’s (TIGR’s) automated annotation (308 total). 228 of the TIGR proteins (>100AA in length) have putative homologues in GenBank, of which 56 were known proteins and 113 have only homologous domains found in GenBank. We found 35 ORFs with homologous domains from which eight had homologous proteins in GenBank that were not present in the TIGR automated annotation.

We have also studied the relationships between the BLAST E-values and the Z-scores and found a significant correlation in proteins with Z-score > 6, which comprise about 25% of the total. Thus, we suggest that in many cases E-values as well as Z-score and actual alignment can be used in inferring homology. An objective of this study was to
examine the TIGR automated annotation programs capability in identifying all of the biologically meaningful ORFs.

Introduction

The Basic Local Alignment Search Tool (BLAST) is a similarity search program that can be used, in conjunction with a protein database, to identify potentially homologous protein among newly sequenced putative proteins. When homologues are identified they may provide clues regarding the function of the putative proteins (1). BLAST utilizes a dynamic programming based local alignment program to identify sections with sequence similarity between the query and database sequences. The results may be continuous or gapped alignments. A raw score for local alignment, S, obtained using an alignment matrix is also provided (2). An expectation value (E-value) is calculated to estimate the significance of the match. The E-value is the number of alignments that is expected to generate an S-value equal to or greater than the observed S-value simply by chance alone (1). The E-values depend on the lengths and composition of the query and subject sequences, as well as the size of the database (3).

A conventional method by which to infer homology (common ancestry) is to use similarity metric statistics, a Z-score. Empirically it has been observed that if the Z-score is greater than 6 there is a very high probability that the query and subject sequences are homologues and not simply similar by chance (4,5). We wished to examine whether one may infer homology between the query and the subject sequences based on the E-value. For example if the E-value is greater than some cutoff value (e.g. 0.01) is there a high probability that the subject sequence is not homologous to the query sequence.
Alternatively, there may be a lower E-value threshold below which there is a reasonable probability that the query and subject sequences are homologous.

Protein-coding regions (or more specifically exons) must lie between two stop codons, an ORF. It is observed that homologous ORFs are generally conserved between closely related species. TIGR has recently produced an automated annotation of the recently sequenced T.thermophila genome. This provides a basis to begin the identification of the proteins coded within the genome (6). To produce an automated annotation one requires a set of known proteins, provided by the user to train the algorithm for gene prediction. There are less than 200 T.thermophila genes in GenBank, which is a meager set for training a gene-finding program. This could potentially pose a serious challenge for automated identification of exons. Unfortunately, identifying the genes in the T.thermophila genome is particularly challenging because T.thermophila has only a single stop codon. Thus, ORFs are relatively common and even ORFs of 150 amino acids may appear by chance (7). It is also worth mentioning that usually frame shifts are difficult for automated annotation systems to detect. In this study, we use homology with proteins in a database as a means of identifying putative proteins. We use this approach to evaluate the accuracy of the automated annotation. We also explore the potential of using BLAST E-values as a measure of homology.

Materials and Methods

The ORFs longer than 300 bp (100 AA) were identified and translated using the ciliate genetic code (21). The codes for ORF finding program could be found in
Appendix 1 in this chapter. These amino acid sequences were subjected to a protein-protein Batch BLAST similarity search against the non-redundant protein database using the search engine from The FIRC Institute of Molecular Oncology (http://bio.ifom-firc.it/BLAST/index.shtml). Significant results were confirmed with NCBI BLASTP (8). Potential homologues from the protein database were aligned with the T. thermophila ORF amino acid sequences using a global alignment (Needleman-Wunsch algorithm) implemented by CLUSTAL W (22) to obtain an alignment score (AS) and trimmings of unaligned end sequences were performed (9). I used global alignment in cases where the entire ORF was included in the GenBank protein but the BLAST result returned partial matches (the local alignment). However when only a segment of the ORF was similar to a segment of the GenBank protein according to the BLAST results, I used the local alignment to obtain Z-scores for putative homologous domains identified by BLAST. Basically when the entire ORF was subject to homology testing a global alignment was used and when just a segment of the ORF was subject to the test a local alignment was used. An example of homologous domain is shown in result section. The sequence similarity between the ORF amino acid sequences and the putative homolog was determined by computing a Z-score as follows. The amino acid sequences are then randomized using a shuffle algorithm and a random AS is obtained. This process was repeated 1000 times and a mean and standard deviation (SD) for the randomized AS distribution was calculated (4,5). The Z-score for two amino acid sequences is the difference of the original AS and the mean of the randomized AS distribution divided by the SD for that distribution. A Z-score of >6 is strongly indicative of homology (10).
The Z-score codes could be found in appendix 1 of chapter 2 in this dissertation. For homologous domains we performed this analysis on sequences obtained from a local alignment from BLAST results.

Tetrahymena proteins predicted by TIGR’s automated programs were downloaded from Eukaryotic Genome Projects website (http://www.tigr.org/tdb/e2k1/ttg/). In this study I only included proteins with greater than 100 amino acids. The Tetrahymena 1Mb scaffold can be found in TIGR scaffolds with accession number 8254638 or 1173196. Make reference to TGD website.

Results

We analyzed 5697 ORFs from a 1-Mb long T.thermophila chromosome. BLAST results with E-values of 10 and higher were eliminated from our analysis resulting in 1011 ORFs. We also analyzed 228 proteins predicted by the TIGR annotation for this 1Mb chromosome using the same criteria as for the ORFs. We compared these two analyses by converting the annotated proteins into ORFs.

Analysis of TIGR predicted proteins and ORFs

In an analysis of 228 predicted proteins 56 appear to have homologues in the available databases. However, 113 proteins, with relatively low BLAST E-values, contain only domains with homologues in the GenBank. In these cases homology could be inferred (z-score) only if the sequences obtained from a local alignment are considered, a global alignment of these proteins does not have sufficient similarity to a database protein.
Thus, we just report them as proteins with homologous domains (see discussion). The remaining 59 proteins have no significant homologues in the GenBank. The 228 TIGR predicted proteins mapped to 558 ORFs, which accounts for slightly more than half of the total analyzed (i.e., 1011). An independent analysis of the remaining 453 ORFs indicated that they were not present among TIGR’s predicted proteins. We identified 8 ORFs, not predicted by the TIGR annotation, that appear to be homologous to 8 proteins in the database (Table 1). Three of these ORFs (ID: 102A-3, 339-1, 452-5) are homologous to single exon proteins while the remaining 5 ORFs (ID: 545-4, 286-1, 308-1, 545-1,302-1) are homologous to a segment of multi-exon proteins. The Z-scores obtained for these ORFs were estimates after sequence trimming was performed. The GenBank protein was aligned with the complete ORF sequence, which might have contained non-coding sequences. Since presence of non-coding sequences in ORFs will decrease the Z-score by lowering the original alignment score, unaligned ends were trimmed. We also found 27 additional ORFs, which have homologous domains in the GenBank.
Figure 1- (Flow chart) Schematic presentation of the processes used to analyze and compare the two annotation methods.
Table 1- List of 8 proteins not present in the TIGR automated annotation and the Z-scores, the number of exons in their homolog from GenBank, and their percent amino acid identity to the GenBank homolog.

<table>
<thead>
<tr>
<th>ORF ID-Frame</th>
<th>GenBank Homolog</th>
<th>Accession #</th>
<th>Z-score</th>
<th>Exons in Homolog</th>
<th>%AA identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>102A-3</td>
<td>hypothetical protein</td>
<td>APE1620-G725</td>
<td>9.5</td>
<td>Single</td>
<td>44</td>
</tr>
<tr>
<td>286-1</td>
<td>putative transcriotional regulatory protein</td>
<td>NP_593431.1</td>
<td>6.2</td>
<td>Multiple</td>
<td>29</td>
</tr>
<tr>
<td>302-1</td>
<td>guanylyl cyclase</td>
<td>CAB52247.1</td>
<td>8.6</td>
<td>Multiple</td>
<td>31</td>
</tr>
<tr>
<td>308-1</td>
<td>hypothetical coiledcoil protein</td>
<td>NP_588370.1</td>
<td>6.1</td>
<td>Multiple</td>
<td>27</td>
</tr>
<tr>
<td>339-1</td>
<td>metG</td>
<td>NP_871539.1</td>
<td>6.1</td>
<td>Single</td>
<td>26</td>
</tr>
<tr>
<td>452-5</td>
<td>unnamed protein product [Mus musculus]</td>
<td>BAC33023</td>
<td>11.0</td>
<td>Single</td>
<td>29</td>
</tr>
<tr>
<td>545-1</td>
<td>PERAXIN</td>
<td>NP_870998</td>
<td>7.4</td>
<td>Multiple</td>
<td>28</td>
</tr>
<tr>
<td>545-4</td>
<td>hybrid sensory kinase [Synechocystis sp.]</td>
<td>NP_442152</td>
<td>6.4</td>
<td>Multiple</td>
<td>26</td>
</tr>
</tbody>
</table>

It is probable that many authentic proteins will not have homologues in the database, because no homolog has been characterized or because the T. thermophila proteins have diverged to an extent that the homology is no longer recognizable. Less than a quarter of the putative proteins identified by the TIGR automated annotation appear to have homologues in the database, yet almost half of them contain homologous domains. It appears that 8 potential T. thermophila ORFs with homologous proteins in the database were not identified by TIGR automated annotation. Since we have found definitive homologues for 56 of predicted proteins then it might be expected that as many as 10 to 15% of the potential proteins were missed by the TIGR automated annotation, which actually is very good for an automated annotations system using a limited training set.

**The E-value vs. Z-score**

We have examined the relationship between the E-values and the Z-scores for the 228 predicted proteins identified by TIGR and 8 additional putative proteins that we have
identified. This T.thermophilia chromosome contains five proteins previously identified proteins (Histone H3, Histone H4, Actin, P type ATPase, and an anlagen stage inducer protein). These proteins were excluded from our analysis because their E-values = 0. The objective of this analysis is to examine whether there is a reliable E-value that could be used to infer homology. Currently we infer homology if two proteins have a Z-score > 6.

E-value vs. Z-score

![Graph showing the correlation between E-values and Z-scores](image-url)

Figure 2- Correlation between the BLAST E-values and the similarity metric Z-scores

To examine these relationships we plotted E-values vs. Z-scores for 231 predicted proteins on the 1 Mb T.thermophilia chromosome (Figure 2). There is a significant correlation between E-values and the Z-scores when we examine proteins with Z-scores >6, our criteria for homology. The genes in this group generally have E-values < 0.01 with regression statistics of R=0.93, R-squared=0.87, (p-value < 0.01). Proteins in this
group comprise about 1/4 of the total proteins analyzed. The remaining proteins do not have homologues in GenBank by our criteria, but many of them do have relatively low E-values. Upon examination it became clear that these proteins with low E-values have domains that are highly similar to domains of the GenBank proteins. An example of such a case is shown in figure 3, where the Z-score is 2.7 and the –Log(E-value) is 81.

The predicted T. thermophila protein is 942 amino acids long, but has a 186 amino acid domain toward the carboxy terminus that aligns very well with a putative ABC transporter protein from the GenBank. The Z-score for the comparison of these domains is 32.0. Clearly these domains are homologous by our criteria, but it is also clear that the entire proteins do not have sufficient sequence similarity to indicate homology. A significant fraction of the T. thermophila predicted proteins, possibly as many as half of them, fall into this class. Figure 3 shows some proteins with Z-scores less than 6.0 that have relatively low E-values. These proteins constitute about 1/4 of the presumptive T. thermophila proteins.
Figure 3 - Alignment of gene TIGR_8594 from T. thermophila 1Mb chromosome with a BLAST protein putative ABC TRANSPORTER from GenBank. The domain identified by BLAST is in bold and italic.
Discussion

We have examined two questions. First, does the TIGR automated gene-finding programs yield the majority of proteins in a genome? Second, do BLAST E-values provide a reliable basis for which to infer homology?

In order to investigate the efficiency and accuracy of the gene-finding program used by TIGR to identify the T. thermophila proteins we require an independent estimate of the proteins present in T. thermophila. We used a 1 Mb chromosome of the T. thermophila genome and compared its ORFs with the proteins currently in GenBank for sequence similarity using BLAST. Assuming that a Z-score of greater than 6 is indicative of homology, we determined the number of homologous proteins in this T. thermophila chromosome and compared this determination with the proteins predicted for this region by the TIGR analysis. Among these ORFs 8 are from proteins with apparent homologues in GenBank that were missing from TIGR predicted proteins. This indicates that about 85 to 90% of the proteins on this chromosome are probably identified by TIGR gene-finding program. This is a remarkable efficiency, particularly given the challenges presented by the small number of known T. thermophila proteins available as a training set and the general characteristics of the T. thermophila genome. On the other hand the set of T. thermophila proteins that have homologues in GenBank are among the most conserved proteins in T. thermophila, thus the ability of an automated annotation program to identify highly conserved proteins may not be unexpected. In any case our analysis certainly indicates that most of T. thermophila proteins that we could identify independently were also identified by the TIGR analysis.
A comparison of closely related genomes using sequence similarity to identify genes and regulatory regions, "phylogenetic footprinting", appears to have great potential (20). Several studies have done whole genome comparisons among closely related species: bacteria (14), yeast (15,16), and Caenorhabditis (17). Our comparison of the T. thermophila ORFs with the proteins in GenBank utilizes sequences from relatively distant species for comparison and is thus very limited. The availability of genome sequences from a species closely related to T. thermophila would facilitate an extremely accurate and efficient identification of the genes and control regions in T. thermophila. A BLAST analysis employs a local alignment algorithm while the alignment used to calculate Z-scores is a global alignment. Thus, these two different approaches to alignment present different pictures of sequence similarity. A limited region of significant sequence similarity occurring in a conserved protein domain will yield a low E-value, but may not be sufficient to produce a high Z-score for the whole protein. Almost half of the T. thermophila predicted proteins we examined fall into this category. The criteria we accept for homology is based on a common evolutionary origin for the whole proteins being compared. However, the domains within proteins may be homologous without the entire protein having a common ancestral origin. It is probable that such a domain may have been introduced into the T. thermophila protein by an exchange process so that the protein is actually a chimera. In this case the proteins as a whole would not be homologous, although it may contain domains that are in fact homologous to other domains. Alternatively, portions of the T. thermophila protein may have drifted at such a high rate that only a highly conserved functional region retains
sufficient sequence similarity to meet the criteria for homology. In this case the proteins would in fact be homologues even though the whole protein no longer has sufficient sequence similarity to meet the accepted criteria. A related situation has been noted among T. thermophila mitochondrial proteins where there are terminal extensions of the proteins (11). If these extensions are sufficiently extensive the whole protein may fail to meet the homology criteria, unless the extensions are "trimmed" away. Thus, the alignment of sequences being compared must be examined carefully. In the case of T. thermophila proteins we have examined here, roughly half of the proteins fall into this category. It does not appear that a simple set of criteria base on E-value can easily be put forward that would be equivalent to the current Z-score based homology criteria. We find that the TIGR gene-finding program does a remarkable job in identifying the proteins located on T. thermophila chromosome we analyzed. However, we remain convinced that a "phylogenetic-footprinting" approach will yield the most efficient and accurate identification of the T. thermophila genes and regulatory regions. Due to its local alignment approach BLAST E-values cannot easily be correlated with Z-scores and thus a criteria for protein homology based on E-values is not readily available.
print "\n\n";
print "**********************************************\n";
print "* ORF Finder Program ver:1.0\n";
print "* By: Mike Moradian Date: MARCH 2003\n";
print "**********************************************\n";

my $lineno=1;

#
# Defining the sequence data file names
#
#$fileName1 = "test seq for ORF";
$fileName1 = "Ttherm_rpl2.txt";

#
# read the first sequence data file
#
print "-> Reading first sequence data file: ["$fileName1"]... ";
open FILE, $fileName1 or die $!

my $data1 = "";
while(<FILE>)
{
    if(m//)
    {
        print "Data File ID: $_"
    }
    else {
        s/\s+/\g; # Remove the white spaces
        s/(ATCG)/\L/$i; # Shift to lowercase
        $data1 .= 
    }
}
chomp($data1);

print "done.\n";

#
# Start to translate the sequence
#

my $code1;

my $ResultString = "";

# Defining simple state machine
# 0=initial state
# 1=Started
# 2=Stopped
my $state = 0;

print "nData Length: ", length($data1);
for ($i=0;$i<length($data1);$i+=3)
{
    $code1 = substr($data1,$i,3);
    if (IsStop($code1) eq 1 && $state ne 1) {  # isstart is a subroutine returns 1 if the
code is start codon and we are not in the state of started
        $ResultString .= "\n[[";
        $state = 1;  # Stated
    }

    $ResultString .= $code1;  # continues appending code after [[[

    if (IsStop($code1) eq 1 && $state ne 2) {  # if we are not in stopped state find the
        stop codon and put ]]] change the state
        $ResultString .= "]]\n";
        $state = 2;  # Stopped
    }
}

print "nORF Strings : ", $ResultString;

print "n*******************************\n";

$state = 0;
$ResultString = "";

for ($i=1;$i<length($data1);$i+=3)
{
    $code1 = substr($data1,$i,3);
    if (IsStop($code1) eq 1 && $state ne 1) {

152
$ResultString .= "n[[]";
    $state = 1; #Started

$ResultString .= $code1;

    if (IsStop($code1) eq 1 && $state ne 2) {
        $ResultString .= "]]\n";
        $state = 2; #Stopped
    }

print "\nORF Strings : ", $ResultString;

print "\n******************************************************************************\n";

    $state = 0;
    $ResultString = "";

    for ($i=2;$i<length($data1);$i+=3)
    {
        $code1 = substr($data1,$i,3);
        if (IsStop($code1) eq 1 && $state ne 1) {
            $ResultString .= "n[[";
            $state = 1; #Stated
        }

        $ResultString .= $code1;

        if (IsStop($code1) eq 1 && $state ne 2) {
            $ResultString .= "]]\n";
            $state = 2; #Stopped
        }
    }

print "\nORF Strings : ", $ResultString;

print "\n\nEnd.\n\n";

#
# Subroutine for find if this is a start code
sub IsStop
{
    my $arg = @_[0];
    if ($arg eq "taa") {
        $IsStop = 1;
    }
    else {
        $IsStop = 0;
    }

    return $IsStop;
}

# Subroutine for find if this is a stop code
sub IsStop
{
    my $arg = @_[0];
    if ($arg eq "taa") {
        $IsStop = 1;
    }
    else {
        $IsStop = 0;
    }

    return $IsStop;
}
References


Chapter 6

Evidence for Extensive Gene Order Rearrangement Between *Tetrahymena thermophila* and *Paramecium tetraurelia* macronuclear Chromosomes
Abstract

Ciliates are single cell eukaryotic organisms with two nuclei and an unusual genome biology where a few germ line micronuclear (MIC) chromosomes are fragmented into hundreds of transcriptionally active macronuclear (MAC) chromosomes. Such developmental rearrangements could potentially result in extensive gene order loss among these organisms. TIGR has recently sequenced the complete macronuclear genomes of *Tetrahymena thermophila* hence I sought to reveal the extent of chromosome rearrangements between *T. thermophila* and the largest chromosome (LC) of *Paramecium tetraurelia*, which is 1Mb long and contains 463 putative proteins. Initially I sought to identify an orthologous chromosome for the LC of *P. tetraurelia*. To achieve this goal I tried to map all 463 putative proteins from the LC of *P. tetraurelia* to one of the chromosomes in *T. thermophila* genome using BLAST. This analysis revealed that there was no orthologous chromosome for the LC of *P. tetraurelia* in *T. thermophila*, however I found 29 chromosomes with at least 25 homologous proteins in both organisms. Each gene was part of an orthologous pair and could be part of one or more paralogous pairs. I determined the gene order conservation (GOC) index between the orthologous genes and found only one pair of orthologues that were contiguous in both genomes suggesting an extensive chromosome rearrangement and gene order loss between the LC of *P. tetraurelia* and *T. thermophila* genome. I also showed that MAC chromosomes in these two ciliates could be hybrid and composed of more than one MIC DNA fragment. Interestingly, gene order between the mitochondrial genome of these two ciliates was almost perfectly conserved and generated a high GOC index. Since
these mitochondrial genomes do not go through a rearrangement process, I suggest that the most plausible explanation for such an extensive gene order loss and rearrangement could be the developmentally programmed elimination of germ line sequences and randomization of MAC gene order.

Introduction

*Paramecium* and *Tetrahymena* are two model eukaryotic organisms, which have been frequently used in several unicellular biology studies. These two genera are part of the phylum Ciliophora and class Oligohymenophorea, which contain two distinct nuclei in a single cell, transcriptionally silent diploid micronucleus (MIC) and somatic macronucleus (MAC) with actively transcribed acentric polygenomic chromosomes [1]. The MIC, which is responsible for transmission of genetic information via conjugation, generates a new somatic MAC at each sexual cycle. Differentiation of MAC requires processes called developmentally programmed DNA rearrangements, called anlagen development, during which the MIC is fragmented into smaller DNA pieces [2,3]. One such process is the deletion of segments of the MIC called internally eliminated sequences (IES), which range from a few hundred to a few thousand bp in *Tetrahymena* [4]. The IES are relatively smaller in *Paramecium* MIC ranging from 26 to 883bp [5]. There are other processes such as elimination of chromosome breakage sites (Cbs) in *Tetrahymena* [4] and imprecise mechanisms that remove transposons or minisatellites in *Paramecium* [6], which contribute to elimination of approximately 10-15% of the mostly repetitive DNA from the MIC of both *Paramecium* and *Tetrahymena* [4]. It has been suggested that
during the elimination of the MIC DNA in *Tetrahymena* some of the breakage sites
rejoin [7] and create hybrid MAC chromosomes, which may map to different locations
on any of the 5 chromosomes in MIC. Similarly imprecise deletions in *Paramecium* MIC
lead to the rejoining of flanking sequences before they are capped by telomere addition
[6].

Gene order in eukaryotic genomes, which appear to evolve by macro and micro
rearrangements, is not random [8]. Inversion, insertion and deletion (indels) of a single
or a few genes are the major causes of micro rearrangements where macro
rearrangements occur mostly in telomeres and centromeres [9]. Gene order conservation
(GOC) index is frequently used to measure the genome stability among related
organisms. The starting point to study chromosome rearrangement is to find orthologous
chromosomes between two or more genomes. Once the orthologous chromosomes are
identified then the gene map is searched in order to find orthologous genes pairs, which
are contiguous between these chromosomes. The values of GOC range from 0 to 1 such
that a high level of synteny between two chromosomes will result in a GOC of close to 1
[10]. A recent study suggests low rates of rearrangements and more conserved pairs of
genes among several bacterial genomes, where after 500 Myr of evolution over 50% of
originally contiguous orthologous genes remain in their original conserved order [10].
On the contrary rates of genome rearrangements could be highly variable depending on
the studied lineages. An example of such variability was shown in yeast lineages, which
generated a wide range of GOC values from 0.11 to 0.99 using comparisons from
orthologous chromosomes [11]. Nevertheless to determine the GOC between any two
genomes one needs to initially identify the orthologous chromosomes and subsequently the orthologous gene pairs that are contiguous on these chromosomes.

*T. thermophila* was the first ciliate with a completely sequenced macronuclear genome. The Institute for Genomic research (TIGR) has produced an automated annotation of the recently sequenced *T. thermophila* genome, where over 100 Mb of nucleotides comprise about 225 macronuclear chromosomes [12]. Zagulski *et al.*, in a relatively smaller study sequenced and annotated the largest chromosome (LC) of *P. tetraurelia*, almost 1Mb in size with 463 putative genes products [13]. Availability of DNA sequences from these two ciliate genomes provides us with a unique opportunity to study the evolution of genome organization in these unusual single cell eukaryotes. The LC of *P. tetraurelia* may or may not have an orthologue in *T. thermophila* genome since larger chromosomes in ciliates are possibly a combination of rejoined different DNA pieces resulting from breakage of micronuclear chromosomes [6,7]. However there is a possibility that there could be orthologous regions between the LC of *P. tetraurelia* and *T. thermophila* genome, which would allow us to determine the GOC index between these two ciliates. Hence I identified the homologues for all the 463 genes from the LC of *P. tetraurelia* in *T. thermophila* genome and determined the GOC index. I also showed that the putative genes in *P. tetraurelia* had homologues with more similar sequences to genes from organism other than *T. thermophila*. The main purpose of this study was to show extensive differences in genome organization between *P. tetraurelia* and *T. thermophila* macronuclear chromosomes.
Materials and Methods

Genomic sequences for the LC of \textit{P. tetraurelia} and \textit{T. thermophila} were downloaded at http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview &list_uids=440 and http://www.tigr.org/tdb/e2k1/ttg/ sites, respectively. BLAST [17] at the NCBI and the WU-BLAST version 2.0 [18] at the TIGR sites were used to conduct sequence similarity searches. Orthologous pairs were assigned based on an additional Z-score statistics analysis. A BLAST result with an E-value of 0.01 or smaller was subjected to a Z-score analysis, where any Z-score $>6$ was assumed to strongly suggest homology. The $z$-score is obtained by comparing the original alignment score of two sequences with an average score obtained from 1000 alignments of randomized sequence of the two original sequences. The difference between the alignment scores is divided by the SD of the randomized alignment score distribution where the scores greater than 6 are indicative of homology between two sequences [19-21]. The same analysis was conducted on similar domains of greater than 50 amino acids. I used both orthologous genes and domains, which most likely come from paralogous proteins, to study chromosome rearrangements in these genomes since homologous domains, in divergent ciliate genomes, are likely to present orthologous genes with common ancestry. Although paralogous genes are not appropriate to determine the GOC index, I was forced to include them in this analysis to compare and find the actual orthologous pairs. The GOC index was calculated using the following equation [10]: (number of contiguous orthologous pairs /total number of orthologues). Gene maps were generated on scatter plots based on orthologous gene location on either chromosome.
Results

We searched to identify similar sequences for predicted genes from *P. tetraurelia* largest somatic chromosome in TIGR’s *T. thermophila* genome database using BLAST. Each one of the 463 predicted genes from the LC of *P. tetraurelia* generated results, which were subsequently examined for homology. Significant BLAST results were subjected to a Z-score analysis to confirm their homology and to be considered as homologous genes. *T. thermophila* macronuclear genome contains about 225 somatic chromosomes and several of them contained genes homologous to the genes on the LC of *P. tetraurelia*. One of these homologous pairs was the real orthologues, where the other pairs were probably paralogues or gene duplication of the original orthologues. Table 1 lists 29 of such chromosomes that contain 25 or more orthologous and paralogous gene between the LC of *P. tetraurelia* and *T. thermophila*. The largest chromosome in *T. thermophila* is over 2.2Mb along with quite a few chromosomes over 1Mb in size. The 2.2Mb chromosome contains the highest number of homologous genes followed by relatively smaller ones, yet none of the chromosomes presented in table 1 were smaller than 500kb in size. Since in several cases it was difficult to determine the actual orthologous pair, I presumed all the homologous pairs as orthologues so I can complete my analysis. The percentage of orthologous genes relative to the total number of genes in each *T. thermophila* chromosome is between 7-29% where the two smallest chromosomes get the highest percentage of orthologous genes (Table 1). Here one could suggest that the largest chromosome from *P. tetraurelia* does not have any orthologous counterpart in
*T. thermophila* macronuclear genome. However presence of dozens of orthologous genes between these two ciliate genomes could be used to identify the GOC index.

**Gene order conservation**

I determined the GOC for genes from mitochondrial genomes of *P. tetraurelia* and *T. thermophila* to have a comparison point since mitochondrial DNA replicate independently. There were 32 common orthologous genes between the two mitochondria from which 30 pairs were contiguous in both genomes resulting in a GOC of 0.93, which suggest strong stability between them. However there is a major inversion of 18 genes starting from *Nad4* to *Cob* (Figure 1A). I conducted the same analysis using pairs of orthologous genes that were present in both the LC of *P. tetraurelia* and 29 different macronuclear chromosomes from *T. thermophila* genome (a list of which presented in table 1) to determine the GOC index between them. Interestingly there were only one pair of genes that were contiguous in analyzed sequences resulting in a GOC value of 0.04 (1/27) for assembly 4886 of *T. thermophila* genome. These genes were putative retinol dehydrogenase and microtubule binding protein, which were located at positions 232905 and 239509 on *P. tetraurelia* chromosome. Contiguous orthologues to this pair were located at positions 207000 and 209866 on assembly 4886 of *T. thermophila* genome. The remaining 25 orthologues appear to have no significant pattern of GOC (Figure 1B) suggesting extensive rearrangements between these two ciliate genomes. Despite the fact that assembly 4886 contained the only pair of contiguous orthologues between the LC of *P. tetraurelia* and *T. thermophila* genome I observed interesting facts in other studied assemblies.
<table>
<thead>
<tr>
<th>Assembly number</th>
<th>Number of Orthologues</th>
<th>Assembly Length (bp)</th>
<th>Assembly GenBank Acc.</th>
<th>Number of genes</th>
<th>Percentage of Orthologues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79</td>
<td>2216158</td>
<td>CH445775</td>
<td>573</td>
<td>13.8</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>1557438</td>
<td>CH445663</td>
<td>408</td>
<td>15.4</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>1807540</td>
<td>CH445789</td>
<td>438</td>
<td>12.1</td>
</tr>
<tr>
<td>16</td>
<td>46</td>
<td>935000</td>
<td>CH445632</td>
<td>270</td>
<td>17.0</td>
</tr>
<tr>
<td>4877</td>
<td>43</td>
<td>979536</td>
<td>CH445644</td>
<td>296</td>
<td>14.0</td>
</tr>
<tr>
<td>5026</td>
<td>41</td>
<td>542724</td>
<td>CH445543</td>
<td>140</td>
<td>28.9</td>
</tr>
<tr>
<td>4003</td>
<td>40</td>
<td>589220</td>
<td>CH670347</td>
<td>154</td>
<td>25.5</td>
</tr>
<tr>
<td>4824</td>
<td>40</td>
<td>715637</td>
<td>CH445577</td>
<td>194</td>
<td>20.0</td>
</tr>
<tr>
<td>4915</td>
<td>38</td>
<td>891270</td>
<td>CH445588</td>
<td>239</td>
<td>15.4</td>
</tr>
<tr>
<td>13</td>
<td>36</td>
<td>1004159</td>
<td>CH445754</td>
<td>298</td>
<td>12.1</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>1231906</td>
<td>CH445655</td>
<td>314</td>
<td>11.1</td>
</tr>
<tr>
<td>4847</td>
<td>35</td>
<td>819503</td>
<td>CH445654</td>
<td>180</td>
<td>18.7</td>
</tr>
<tr>
<td>14</td>
<td>34</td>
<td>983259</td>
<td>CH445770</td>
<td>267</td>
<td>12.7</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>1171415</td>
<td>CH445735</td>
<td>304</td>
<td>10.9</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>1191929</td>
<td>CH445709</td>
<td>294</td>
<td>10.9</td>
</tr>
<tr>
<td>4995</td>
<td>32</td>
<td>1407426</td>
<td>CH445782</td>
<td>333</td>
<td>9.3</td>
</tr>
<tr>
<td>4860</td>
<td>31</td>
<td>758301</td>
<td>CH445689</td>
<td>189</td>
<td>15.9</td>
</tr>
<tr>
<td>4912</td>
<td>31</td>
<td>1049106</td>
<td>CH445575</td>
<td>270</td>
<td>11.1</td>
</tr>
<tr>
<td>41</td>
<td>30</td>
<td>605964</td>
<td>CH445606</td>
<td>168</td>
<td>17.9</td>
</tr>
<tr>
<td>4844</td>
<td>30</td>
<td>663878</td>
<td>CH445625</td>
<td>189</td>
<td>15.5</td>
</tr>
<tr>
<td>4886</td>
<td>27</td>
<td>1073828</td>
<td>CH445601</td>
<td>269</td>
<td>10.0</td>
</tr>
<tr>
<td>87</td>
<td>27</td>
<td>384242</td>
<td>CH445561</td>
<td>133</td>
<td>20.3</td>
</tr>
<tr>
<td>4084</td>
<td>26</td>
<td>556539</td>
<td>CH670348</td>
<td>144</td>
<td>17.6</td>
</tr>
<tr>
<td>4963</td>
<td>26</td>
<td>638766</td>
<td>CH445555</td>
<td>164</td>
<td>15.0</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>1227316</td>
<td>CH445695</td>
<td>343</td>
<td>7.3</td>
</tr>
<tr>
<td>4871</td>
<td>25</td>
<td>816060</td>
<td>CH445618</td>
<td>204</td>
<td>11.7</td>
</tr>
<tr>
<td>4873</td>
<td>25</td>
<td>714276</td>
<td>CH445580</td>
<td>162</td>
<td>14.4</td>
</tr>
<tr>
<td>4902</td>
<td>25</td>
<td>1078769</td>
<td>CH445650</td>
<td>279</td>
<td>8.4</td>
</tr>
<tr>
<td>4940</td>
<td>25</td>
<td>1053479</td>
<td>CH445731</td>
<td>266</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Table1. List of *Tetrahymena* chromosomes with homologous (orthologous and paralogous) genes in Paramecium.
Figure 1. Chromosome mapping between *T. thermophila* and *L.C* of *P. tetraurelia*

A- Mapping of mitochondrial genomes  
B- The only orthologous pair contiguous in both genomes shown in red (assembly 4886)  
C- Absence of orthologues between two chromosomes in 500kb (assembly 4902)  
D- Orthologous domains mapping to large proteins (assembly 5026)  
E- Contiguous orthologous genes groups yet interrupted with unrelated genes (not shown in graph) (assembly 13)  
F- The largest chromosome of *T. thermophila* with the most orthologues and a 200kb unrelated region (assembly 1)
I found examples of rejoining of broken fragments from MIC chromosomes, which may produce hybrid MAC chromosomes in these ciliates. I showed that some *T. thermophila* chromosomes contained parts that shared no orthologous genes with the LC of *P. tetraurelia*. Assembly 4902 is a good example of such arrangement where despite its almost equal size (1Mb) to the LC of *P. tetraurelia* it shares only 25 orthologous genes. There is a considerably large, approximately 500Kb, region towards the end of the assembly that shares no orthologues with *P. tetraurelia* (Figure 1C). Many hypothetical proteins on the LC of *P. tetraurelia* shared homologous domains, which most probably come from paralogous genes with proteins in *T. thermophila* genome generating specific patterns in gene mapping charts (Figure 1C and 1D). Assembly 5026 (600kb long) shares 39 orthologues with the LC of *P. tetraurelia*, 20 of which share only homologous domains with two long proteins on this assembly. These two hypothetical proteins were 2200 and 3712 amino acids long, shared 11 and 9 orthologous domains, and were located on positions 64945 and 528961 of the assembly 5026, respectively. Despite their unknown function, further analysis revealed that they contained domains similar to coiled-coil, zinc finger, and microtubule domains, as well as transmembrane helices. These domains are quite prevalent and widely used in different gene groups and could be part of a paralogous gene family.

Although there was only a single pair of contiguous orthologues between the LC of *P. tetraurelia* and *T. thermophila* genome I found several groups of genes, which appeared to partially conserve their orders. These genes could be considered contiguous
orthologues if one or more insertions are allowed between them. Figure 1E shows three
of such gene groups, which share orthologues with assembly 13 and are represented by
triangles, crosses, and squares. The triangle group is comprised of 4 genes, which start at
position 728278 and end at position 891781 on the LC of *P. tetraurelia*. The sequences
between these four genes contain 36, 8, and 16 genes from which 4, 0, and 1 share
orthologues with genes on assembly 13, respectively. The next group represented by
crosses contains five genes, 1 of which is a duplication. The remaining four genes are
located between positions 125856 and 428620 on the LC of *P. tetraurelia*.* 44, 61, and 31
genes separate them, from which 3, 3, and 0 share orthologues with genes on assembly
13, respectively. This group could serve as a clear example of extensive gene
rearrangements between chromosomes of these two ciliates. The last and smallest group
represented by squares in figure 1E contains just two genes on positions 739275 and
759147 separated by 9 genes on the LC of *P. tetraurelia*. These 9 genes share only one
orthologue with genes on assembly 13. Presence of such groups was observed in other
assemblies as well. Assembly 1 in *T. thermophila* genome is the largest chromosome,
over 2.2Mb in size, and shares the greatest number of orthologues (i.e., 79) with the LC
of *P. tetraurelia*. Unlike the situation in figure 1C where a considerable segment of the
assembly 4902 did not share any orthologues with the genes on the LC of *P. tetraurelia*,
in assembly 1 the last 300kb of the LC of *P. tetraurelia* did not contain any orthologues
(Figure 1F). A plausible explanation could be that this segment of the chromosome
comes from a MIC piece that was not incorporated in assembly 1 in *T. thermophila*. 169
<table>
<thead>
<tr>
<th>Complete Protein Sequence</th>
<th>BLAST Result</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine carboxypeptidase II, putative</td>
<td>refNP_956844.11</td>
<td>Vertebrata*</td>
</tr>
<tr>
<td>AAA ATPase, cell division control protein</td>
<td>refNP_998649.11</td>
<td>Vertebrata*</td>
</tr>
<tr>
<td>60S ribosomal L18A, putative</td>
<td>gbIAAS2055.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>CTD-like phosphatase, putative</td>
<td>refXP_954746.1</td>
<td>Apicomplexa</td>
</tr>
<tr>
<td>GTP-binding protein RAB2 homolog</td>
<td>embCAI39281.1</td>
<td>Chlorophyta</td>
</tr>
<tr>
<td>ER retention receptor, putative</td>
<td>embCA00841.1</td>
<td>Vertebrata*</td>
</tr>
<tr>
<td>DHHC-type Zn-finger containing protein, puto</td>
<td>refXP_966647.1</td>
<td>Insecta*</td>
</tr>
<tr>
<td>Protein kinase, putative</td>
<td>gbAAC04932.1</td>
<td>Apicomplexa</td>
</tr>
<tr>
<td>Succinate dehydrogenase, putative</td>
<td>gbIAAW50854</td>
<td>Ciliophora</td>
</tr>
<tr>
<td>Protein kinase, putative</td>
<td>refXP_572873.1</td>
<td>Ascomycota*</td>
</tr>
<tr>
<td>Prenylcysteine lyase, putative</td>
<td>refXP_546312.1</td>
<td>Mammalia*</td>
</tr>
<tr>
<td>40S ribosomal protein S9, putative</td>
<td>refXP_730059.1</td>
<td>Apicomplexa</td>
</tr>
<tr>
<td>Insulin degrading enzyme</td>
<td>embCAC67408.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Conserved hypothetical protein</td>
<td>refXP_001222013.1</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>60S ribosomal protein L37a, putative</td>
<td>gbEAU90177.1</td>
<td>Basidiomycota</td>
</tr>
<tr>
<td>40S ribosomal protein S8, putative</td>
<td>refXP_627792.1</td>
<td>Apicomplexa</td>
</tr>
<tr>
<td>Ubiquitin-specific protease, putative</td>
<td>embCAL51083.1</td>
<td>Chlorophyta</td>
</tr>
<tr>
<td>40S ribosomal protein S15, putative</td>
<td>gbIAAX62477.1</td>
<td>Insecta*</td>
</tr>
</tbody>
</table>

**Domain Sequences**

<table>
<thead>
<tr>
<th>Domain Sequences</th>
<th>BLAST Result</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical protein PTMB.01c</td>
<td>refNP_207763.1</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>Unc51-like kinase, putative</td>
<td>gbIAAZ9490.1</td>
<td>Ciliophora</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.08c</td>
<td>embCAF92169.1</td>
<td>Vertebrata</td>
</tr>
<tr>
<td>ZZ domain protein, putative</td>
<td>gbEAS29272.1</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>Hypothetical protein with leucine-rich repeats</td>
<td>refXP_658029.1</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.44c</td>
<td>refYP_723847.1</td>
<td>Cyanobacteria</td>
</tr>
<tr>
<td>Zn-finger protein, putative</td>
<td>refXP_387220.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Hypothetical protein, transmembrane helices</td>
<td>refNP_610351.1</td>
<td>Insecta</td>
</tr>
<tr>
<td>Microtubule-binding protein, putative</td>
<td>refXP_001088113.1</td>
<td>Primates</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.102c</td>
<td>refXP_770485.1</td>
<td>Giardinaceae</td>
</tr>
<tr>
<td>Ubiquitin protein ligase, putative</td>
<td>refNP_190877.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.188-bc</td>
<td>gbEAS03089.1</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.152c</td>
<td>refZP_0414925.1</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.155c</td>
<td>refXP_638372.1</td>
<td>Mycetozoa</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.169c</td>
<td>refYP_126799.1</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>Mevalonate kinase, putative</td>
<td>refNP_819639.1</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>Casein kinase II catalytic subunit, putative</td>
<td>dbjIAA01091.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Proteosome subunit, putative</td>
<td>gbCA73621.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Hypothetical protein PTMB.219</td>
<td>refXP_00105397.1</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Hypothetical protein, RING Zn-finger</td>
<td>embCA0202.1</td>
<td>Mycetozoa</td>
</tr>
<tr>
<td>Hypothetical protein, C48 peptidase motif</td>
<td>embCAL56509.1</td>
<td>Chlorophyta</td>
</tr>
<tr>
<td>Lipid-binding protein, putative</td>
<td>refNP_001048139.1</td>
<td>Embryophyta</td>
</tr>
<tr>
<td>Hypothetical transmembrane protein</td>
<td>sp</td>
<td>OS3W1ERF8</td>
</tr>
<tr>
<td>Acetyl-Coenzyme A transporter, putative</td>
<td>refXP_001209110.1</td>
<td>Ascomycota</td>
</tr>
<tr>
<td>Acid phosphatase, putative</td>
<td>gbEA132463.1</td>
<td>Insecta</td>
</tr>
<tr>
<td>Ca2+ channel, putative</td>
<td>refXP_634007.1</td>
<td>Mycetozoa</td>
</tr>
<tr>
<td>K+ channel, putative</td>
<td>gbIAAF70087.1</td>
<td>Vertebrata</td>
</tr>
</tbody>
</table>

* represents genes absent in *Tetrahymena*
In summary after determining the orthologous genes between the LC of *P.tetraurelia* and *T.thermophila* genome I discovered low levels of GOC between these two ciliates. In addition I showed that there were gene groups that appeared to contain contiguous orthologues yet they were interrupted with many non-related genes. Both the LC of *P.tetraurelia* and *T.thermophila* assemblies contained segments that did not share any orthologues. And lastly there were large putative proteins on different *T.thermophila* assemblies that shared promiscuous domains with several putative proteins, perhaps paralogous proteins, from the LC of *P.tetraurelia*.

**Highly divergent genes**

*Paramecium* and *Tetrahymena* are the only species in the order Oligohymenophorea with partial or complete genome sequences in the databases. Therefore one would expect to find the most similar sequences for predicted proteins from the LC of *P.tetraurelia* to come from *T.thermophila* genome. Hence I searched the GenBank to identify similar sequences for all 463 predicted proteins from the LC of *P.tetraurelia*. As expected, in the vast majority of the BLAST results *T.thermophila* genes were the best matches however I were able to identify 18 complete proteins and 27 protein domains with homologues in the GenBank that had sequences more similar to organisms other than *T.thermophila*. Majority of these sequences come from fungi and algae yet more distant organisms such as vertebrates, insects, and bacteria were also present (Table 2). Surprisingly six of these proteins were not even present in *T.thermophila* genome. A complete list of their names, positions, and the type of organism that they come from is presented in table 2. In sum
4% (18/463) of analyzed genes from the LC of *P. tetraurelia* were relatively more conserved in *P. tetraurelia* than in *T. thermophila*, similarly 6% (27/463) of them contained more conserved domains suggesting more rapidly evolving genes and divergent nature in the *T. thermophila* genome.

**Discussion**

Chromosome rearrangements and gene order randomization could result from micro- and macro rearrangements in a genome. Possible reasons for genome restructuring are variable and could be in response to environmental changes, reduced selection and accumulation of indels, retrotransposons, etc... Each one of these mechanisms could potentially be responsible for randomization of gene order between *P. tetraurelia* and *T. thermophila* chromosomes. However a more plausible explanation for such an extensive gene order loss between these two ciliates could be the mechanisms through which the MIC chromosomes are generated. Chromosome breakages in MIC genome of ciliates produce subchromosomal DNA pieces, which subsequently create the MAC chromosomes. There are roughly 6,000 deletion sites per haploid genome in *Tetrahymena* [4], and assuming an over 100 Mb genome [12] the average size of these DNA fragments could be between 15-20 kb. In *Paramecium* there are 60,000 deletion sites, yet its genome is amplified 500 folds before being fragmented into smaller DNA pieces [14], which complicates speculating their sizes. Nevertheless these DNA fragments or blocks could be used to investigate chromosome rearrangements on a much smaller scale since they could potentially contain conserved syntenic gene blocks. This
analysis suggests extensive chromosome rearrangement and gene order loss between the LC of *P. tetraurelia* and *T. thermophila* genome, which is particularly large when it is compared to mitochondrial GOC between these ciliates. I also observed that even the conserved gene blocks were interrupted by many genes. Our criteria to obtain the GOC index may be too stringent since I look for two consecutive uninterrupted genes, however I showed that even the closest consecutive orthologous genes pairs were separated by at least 8 and up to 60 unrelated genes. Hence one may propose that the most plausible mechanism for such a gene order loss could be the programmed genome rearrangement in ciliates where thousands of internally eliminated sequences (IESs) are dispersed throughout the MIC. I presumed that the presence of an extremely efficient mechanism to conserve the location and sequence of the IESs is unlikely. Hence the randomization of the gene order occurs during fragmentation of the MIC chromosomes. Interestingly Huvos in a recent study confirms that the IESs are not conserved between *T. thermophila* and *T. malaccensis* [15]. In fact she shows that from three studied loci two are missing in *T. malaccensis* and the third one is at a different location with different sequence. Furthermore, Huvos reports that there are indel polymorphisms in IESs in *T. thermophila* MIC [16], which could even add more to the differences between these developmentally important sequences and consequently accelerate randomization of the gene order. When more closely related ciliate genomes become available, one may also use them to understand the patterns of the genome rearrangements and to what extent is gene order under selection in these genomes.
References


