Copyright
by
Chaitanya Muralidhara
2010

The Dissertation Committee for Chaitanya Muralidhara
certifies that this is the approved version of the following dissertation:

# Matrix and tensor decomposition methods as tools to understanding sequence-structure relationships in sequence alignments

Committee:

Orly Alter, Supervisor

Robin R. Gutell, Supervisor

Ron Elber

Lauren Ancel Meyers

Claus O. Wilke

# Matrix and tensor decomposition methods as tools to understanding sequence-structure relationships in sequence alignments

**by**

## Chaitanya Muralidhara, B.E.

**DISSERTATION**

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN
December 2010

To Ajit, for always encouraging me to be the best I can be.

# Acknowledgments

Let me start by thanking my PhD advisor Dr. Orly Alter. Her incisive observations and questions kept me constantly questioning my premises and thus honing my knowledge. I would like to thank her for the academic opportunities she has provided me as well as her support during difficult times.

I would like to thank my co-PI, Dr. Robin Gutell who has provided me significant insight into the mechanisms of rRNA structure. Dr. Alter and Dr. Gutell are true pioneers in their chosen fields, and I am fortunate to have been mentored by them.

I would also like to thank Dr.Lauren Myers, Dr. Claus Wilke and Dr.Ron Elber for taking an interest in my work and serving on my committee.

This work was sponsored by the National Human Genome Research Institute R01 Grant HG-004302 and National Science Foundation CAREER Award DMS-0847173 to Dr. Alter. My travel to several international conferences was made possible by a travel award from the National Science Foundation (for the SIAM Annual Meeting, 2008), and two professional development awards from the UT Graduate School (for the BMES Fall Meeting, 2008 and the CR Rao Conference, 2009).

Many thanks are due to the administrative staff at the Institute for Cellular and Molecular Biology as well as Biomedical Engineering for all their help and

co-operation during my years at UT.

I would like to thank the past members of the Alter Lab for lightening the long hours spent in the lab. Andy Gross helped convert parts of the code to Mathematica. Many thanks also to Jamie Cannone from the Gutell Lab for help with navigating the CRW.

Looking back over the years spent in school, I realize the importance of many subtle lessons taught by my parents and grandparents, and the stimulating environment they provided for me. Words are inadequate to express gratitude to my family for their love and support.

# Matrix and tensor decomposition methods as tools to understanding sequence-structure relationships in sequence alignments

Publication No. ____________________

Chaitanya Muralidhara, Ph.D.
The University of Texas at Austin, 2010

Supervisors: Orly Alter
Robin R. Gutell

We describe the use of a tensor mode-1 higher-order singular value decomposition (HOSVD) in the analyses of alignments of 16S and 23S ribosomal RNA (rRNA) sequences, each encoded in a cuboid of frequencies of nucleotides across positions and organisms. This mode-1 HOSVD separates the data cuboids into combinations of patterns of nucleotide frequency variation across the positions and organisms, i.e., “eigenorganisms” and corresponding nucleotide-specific segments of “eigenpositions,” respectively, independent of a-priori knowledge of the taxonomic groups and their relationships, or the rRNA structures. We show that this mode-1 HOSVD provides a mathematical framework for modeling the sequence alignments where the mathematical variables, i.e., the significant eigenpositions and eigenorganisms, are consistent with current biological understanding of the 16S and 23S rRNAs.

First, the significant eigenpositions identify multiple relations of similarity and dissimilarity among the taxonomic groups, some known and some previously unknown.

Second, the corresponding eigenorganisms identify positions of nucleotides exclusively conserved within the corresponding taxonomic groups, but not among them, that map out entire substructures inserted or deleted within one taxonomic group relative to another. These positions are also enriched in adenosines that are unpaired in the rRNA secondary structure, the majority of which participate in tertiary structure interactions, and some also map to the same substructures. This demonstrates that an organism's evolutionary pathway is correlated and possibly also causally coordinated with insertions or deletions of entire rRNA substructures and unpaired adenosines, i.e., structural motifs which are involved in rRNA folding and function.

Third, this mode-1 HOSVD reveals two previously unknown subgenic relationships of convergence and divergence between the Archaea and Microsporidia, that might correspond to two evolutionary pathways, in both the 16S and 23S rRNA alignments. This demonstrates that even on the level of a single rRNA molecule, an organism's evolutionary pathway is composed of different types of changes in structure in reaction to multiple concurrent evolutionary forces.

# Table of Contents

**Acknowledgments** **v**

**Abstract** **vii**

**List of Tables** **xii**

**List of Figures** **xiii**

**Chapter 1. Introduction** **1**

1.1 Motivation . . . . . 1
1.2 Ribosomal RNA Sequence Alignments . . . . . 2
  1.2.1 16S Ribosomal RNA . . . . . 3
  1.2.2 23S Ribosomal RNA . . . . . 5
  1.2.3 Evolution of ribosomal RNAs . . . . . 5
  1.2.4 Comparative Analysis of RNA sequences . . . . . 8
1.3 Genomic Signal Processing . . . . . 9
1.4 Mathematical Framework . . . . . 11
  1.4.1 Singular Value Decomposition . . . . . 11
  1.4.2 Tensors . . . . . 12
  1.4.3 HOSVD . . . . . 12
  1.4.4 Matrix Decompositions and Sequence Analysis . . . . . 13
1.5 Our Aims . . . . . 16
1.6 Organization . . . . . 17

**Chapter 2. Materials and Methods** **18**

2.1 Data . . . . . 18
  2.1.1 Alignment . . . . . 18
  2.1.2 Structure . . . . . 20

2.1.3 Taxonomy . . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.2 Mathematical Framework . . . . . . . . . . . . . . . . . . . . 21
2.2.1 Encoding . . . . . . . . . . . . . . . . . . . . . . . . . . 21
2.2.2 Mode-1 HOSVD . . . . . . . . . . . . . . . . . . . . . . . 22
2.2.3 Interpretation . . . . . . . . . . . . . . . . . . . . . . . . 23
2.3 Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . 25
2.3.1 Organisms . . . . . . . . . . . . . . . . . . . . . . . . . . 25
2.3.2 Positions . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
2.3.2.1 Conservation . . . . . . . . . . . . . . . . . . . . . 27
2.3.2.2 Enrichment . . . . . . . . . . . . . . . . . . . . . . 28

**Chapter 3. Results** **32**
3.1 Most significant eigenposition is invariant . . . . . . . . . . . . . 32
3.2 Eigenpositions correspond to phylogenetic groups . . . . . . . . . . 33
3.2.1 Eigenpositions in the 16S rRNA . . . . . . . . . . . . . . . . 33
3.2.2 Eigenpositions in the 23S rRNA . . . . . . . . . . . . . . . . 37
3.3 Eigenorganisms identify positions uniquely conserved within phylogenetic groups . . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.3.1 Naming conventions for figures . . . . . . . . . . . . . . . . 41
3.3.2 Eigenposition 2 separates the Bacteria and Eukarya . . . . . . 42
3.3.3 Other significant eigenpositions . . . . . . . . . . . . . . . . 45
3.4 Eigenorganisms identify characteristic sites . . . . . . . . . . . . . 45
3.4.1 Sites are insertions/deletions of structure motifs . . . . . . . . 46
3.4.2 Sites are structure motifs: Unpaired adenosines . . . . . . . . 49
3.5 Eigenpositions identify multiple pathways of evolution . . . . . . . 59
3.5.1 Eigenposition 3 shows that Archaea are similar to Microsporidia . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.5.2 Eigenposition 5 shows that Archaea are dissimilar to Microsporidia . . . . . . . . . . . . . . . . . . . . . . . . . . 60

Chapter 4. Discussion 74
4.1 Most significant eigenposition is invariant 74
4.2 Eigenpositions correspond to phylogenetic groups 75
4.3 Eigenorganisms identify positions uniquely conserved within phylogenetic groups 75
4.4 Unpaired adenosines are significant in distinguishing structure motifs 76
4.5 Eigenpositions identify multiple pathways of evolution 79
4.5.1 The Archaea 79
4.5.2 The Microsporidia 80
4.5.2.1 The Archezoa Hypothesis 81
4.5.2.2 Conflicting phylogenies from other genes 82
4.5.3 Archaea/Microsporidia relationship in the rRNAs 87
4.5.3.1 Similarity between the Archaea and the Microsporidia 87
4.5.3.2 Dissimilarity between the Archaea and the Microsporidia 89
4.5.4 Genome compaction and evolution 90
4.6 Robustness 91
4.6.1 Domain relationships 91
4.6.2 Archaea and Microsporidia relationship 91
4.7 Conclusions 95
4.8 Implications for Future Research 96

Appendices 98

Appendix A. Mode-1 HOSVD analysis of 5S rRNA 99
A.1 Introduction 99
A.2 Data 100
A.3 Results 100

Appendix B. Taxonomy of sequences in the rRNA Datasets 103

Bibliography 128

Vita 146

# List of Tables

2.1 **Composition of rRNA alignments** . . . . . . . . . . . . . . . . . 20

3.1 **Association of phylogenetic groups with the most dominant eigenpositions in the 16S and 23S alignments.** . . . . . . . . . . . 36
3.2 **Enrichment of structure motifs in the dominant 16S eigenorganisms** . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
3.3 **Enrichment of structure motifs in the dominant 23S eigenorganisms** . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
3.4 **Enrichment of tertiary interactions in the 100 nucleotides in the A segment most negatively correlated with the second eigenorganism.** . . . . . . . . . . . . . . . . . . . . . . . . . 56
3.5 **Unpaired Adenosines exclusively conserved in the Bacterial 16S rRNA, and their tertiary interactions** . . . . . . . . . . . . . . 73

4.1 **Summary of results from Mode-1 HOSVD analysis of various rRNA datasets, showing that the phylogenetic relationships are always revealed in the most dominant eigenpositions.** . . . . . . 93

A.1 **Probabilistic significance of the enrichment of the $k$=50 organisms in the 5S rRNA** . . . . . . . . . . . . . . . . . . . . . . . 101

B.1 **Organisms in the 339-sequence 16S rRNA dataset, and their phylogeny** . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
B.2 **Organisms in the 75-sequence 23S rRNA dataset, and their phylogeny** . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
B.3 **Organisms in the 242-sequence 5S rRNA dataset, and their phylogeny** . . . . . . . . . . . . . . . . . . . . . . . . . . . 119

# List of Figures

1.1 The conserved secondary structure of the 16S ribosomal RNA (Reproduced from CRW). . . . . . . . . . . . . . . . . . . . . . . . 4

1.2 The conserved secondary structure of the 23S ribosomal RNA (3' end) (Reproduced from CRW). . . . . . . . . . . . . . . . . . . 6

1.3 The conserved secondary structure of the 23S ribosomal RNA (5' end) (Reproduced from CRW). . . . . . . . . . . . . . . . . . . 7

1.4 Mathematical models for DNA microarray data derived from genomic signal processing techniques (reproduced from Alter, 2007). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.1 Flowchart showing the steps involved in the analysis of ribosomal RNA alignments using Mode-1 HOSVD . . . . . . . . . . . 19

2.2 Structure categories in the ribosomal RNAs (reproduced from Smit, *et. al.*, 2006) . . . . . . . . . . . . . . . . . . . . . . . . 21

2.3 Mode-1 higher-order singular value decomposition (HOSVD) of the 16S rRNA alignment. . . . . . . . . . . . . . . . . . . . 26

2.4 Significant 16S eigenpositions and their correlation with the NCBI Taxonomy Browser taxonomic groups . . . . . . . . . . . 30

2.5 Significant 23S eigenpositions and their correlation with the NCBI Taxonomy Browser taxonomic groups . . . . . . . . . . . 31

3.1 Significant 16S eigenpositions. . . . . . . . . . . . . . . . . . . . 35

3.2 Significant 23S eigenpositions. . . . . . . . . . . . . . . . . . . . 38

3.3 Sequence gaps exclusive to Eukarya or Bacteria 16S rRNAs. . . 43

3.4 Non-homologous features that distinguish the three domains, represented on the secondary structure of *E. coli* (Reproduced from Winker & Woese,1991). . . . . . . . . . . . . . . . . . . . 44

3.5 Sequence gaps exclusive to Eukarya or Bacteria 16S rRNAs . . . 48

3.6 Sequence gaps and unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs. . . . . . . . . . 50

3.7 Sequence gaps exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . 51

3.8 Unpaired adenosines exclusive to Bacteria 16S rRNAs. . . . . . . 52
3.9 Unpaired adenosines exclusive to Eukarya excluding Microsporidia 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
3.10 Unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . 55
3.11 Unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . 57
3.12 Sequence gaps exclusive to both Archaea and Microsporidia 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
3.13 Other nucleotides exclusive to Archaea and Microsporidia 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
3.14 Sequence gaps exclusive to Bacteria or Archaea and Microsporidia 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
3.15 Unpaired adenosines exclusive to Bacteria or Archaea and Microsporidia 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . 64
3.16 Nucleotides exclusive to Microsporidia 16S rRNAs. . . . . . . . 66
3.17 Adenosine and Cytosine nucleotides exclusive to Archaea 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
3.18 Guanosine and Uracil nucleotides exclusive to Archaea 16S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
3.19 Sequence gaps and unpaired adenosines exclusive to Microsporidia or Archaea 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . 70
3.20 Sequence gaps exclusive to Archaea or Microsporidia 23S rRNAs. 71
3.21 Unpaired adenosines exclusive to Archaea or Microsporidia 23S rRNAs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72

4.1 Relative percentages of paired and unpaired nucleotides in Bacterial rRNAs (data from CRW) . . . . . . . . . . . . . . . . . 77
4.2 A eukaryotic tree from early SSU rRNA analysis (reproduced from Embley, 2006). . . . . . . . . . . . . . . . . . . . . . . . 83
4.3 Consensus tree of the eukaryotes, representing current hypotheses about early eukaryotic evolution (reproduced from Embley, 2006). . . . . . . . . . . . . . . . . . . . . . . . . . 88
4.4 Eigenpositions from two reduced 16S rRNA datasets showing the Archaea/Microsporidia relationships. . . . . . . . . . . . . 94

A.1 The conserved secondary structure of the 5S ribosomal RNA (Reproduced from CRW). . . . . . . . . . . . . . . . . . . . . . . . 100

A.2 **Significant 5S eigenpositions.** . . . . . . . . . . . . . . . . . . . . . . 102

# Chapter 1

# Introduction

## 1.1 Motivation

Rapid advances in high-throughput sequencing technologies have created an abundance of DNA and RNA sequence data. To make sense of this data is a challenge that requires, in addition to increased understanding of the biology of cells and organisms, methods to organize and classify the data. The comparative analysis and mathematical modeling of these data holds the key to fundamental understanding of biological processes, evolution, and human diseases.

The International HapMap project [25] catalogs human single nucleotide polymorphisms (SNPs), with the aim of associating allelic variation with disease phenotypes. The analysis of these SNPs is already providing causative linkages for common human diseases, while also uncovering therapeutic targets [53]. It is now becoming increasingly clear that future predictive power, discovery, and control in biology and medicine will result from the ability to accurately analyse and model these large-scale sequence data.

## 1.2 Ribosomal RNA Sequence Alignments

The ribosomal RNA (rRNA) is an essential component of the ribosome, the cellular organelle that associates the cell's genotype with its phenotype by catalyzing protein synthesis in all known organisms, and therefore also underlies cellular evolution [113]. RNAs are thought to be among the most primordial macromolecules. This is because an RNA template, similar to a DNA template, can be used to synthesize DNA and RNA, while RNA, similar to proteins, can form three-dimensional structures and catalyze reactions. It was suggested, therefore, that rRNA sequences and structures, that are similar or dissimilar among groups of organisms, are indicative of the relative evolutionary pathways of these organisms [26, 82, 111].

Advances in sequencing technologies have resulted in an abundance of rRNA sequences from organisms spanning all taxonomic groups. Today, the small subunit ribosomal RNA (16S rRNA) is the gene with the largest number of determined sequences. Comparative analyses of these rRNA sequences promise to give insights into the universality and specialization of evolutionary, genetic and biochemical pathways. These analyses may also prove useful in drug design, since most natural as well as synthetic antibiotics target the ribosome.

The analysis of RNA sequences is made complicated by the fact that most functional RNAs conserve structure more than they conserve sequence. Ribosomal RNAs also exhibit a significant degree of non-canonical base pairing, and, as observed in the crystal structure of the 30S ribosomal subunit, short single-stranded RNA segments make idiosyncratic long-range interactions to stabilize the packing of helical elements [109]. These interactions determine not only folding pathways, but in many cases, rRNA function as well [41]. Therefore, methods used to analyse RNA alignments should be, in principle, able to capture this complexity of the data, by integrating diverse sources of biological information.

### 1.2.1 16S Ribosomal RNA

The 16S rRNA is part of the small subunit (SSU) of the ribosome, which functions in protein translation by providing the mRNA-binding machinery and most components that control translation fidelity [11]. Prokaryotic 16S rRNAs consist of $\sim$1500 nucleotides, while their eukaryotic counterparts, the 16S rRNAs, are $\sim$1900 nucleotides long. In prokaryotes, the 3' end of the 16S rRNA consists of a pyrimidine-rich region ('anti Shine-Dalgarno'), which assists in mRNA placement.

The 16S ribosomal RNA shows a high degree of sequence conservation across all organisms [114] (Figure 1.1). Its universal distribution, high conservation, some moderate variability, and minimal lateral genetic transfer have made the 16S rRNA a good candidate for use in phylogenetic analyses of widely varying species [112]. The discovery of the microbial kingdom Archaea [40, 113], and later, the reorganization of life into the three domains, Archaea, Bacteria, and Eukarya [115], can be attributed to comparative studies of the 16S rRNA.

There are now 73640 16S rRNA sequences and 662 comparative secondary structure models available on the CRW [16]. Like many functional RNAs, the 16S rRNA structure is highly conserved across all organisms (Figure 1.1).

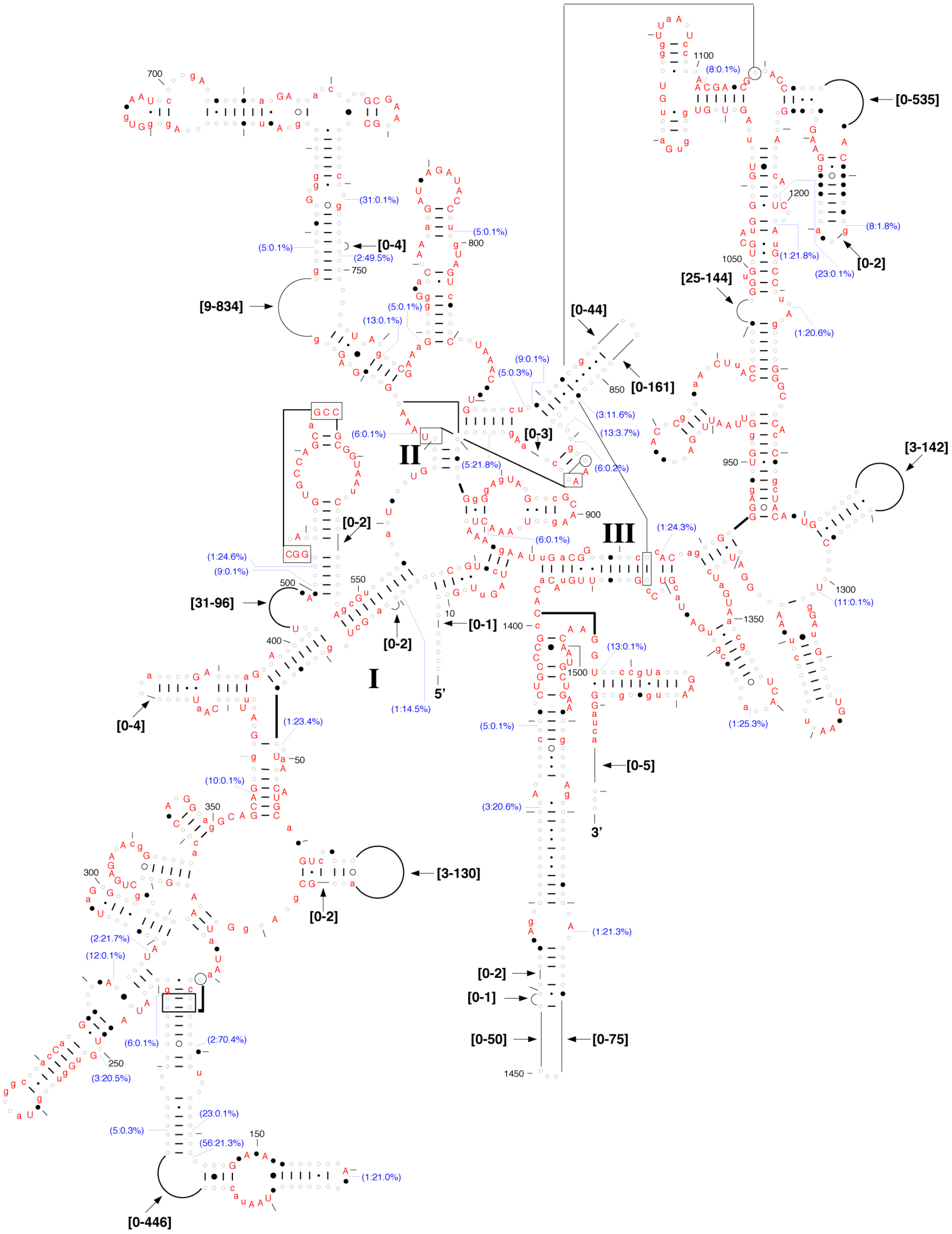

**Fig.** 1.1: **The conserved secondary structure of the 16S ribosomal RNA.** Positions in the 16S ribosomal RNA with a nucleotide in more than 95% of the sequences are shown superimposed onto the *E. coli* secondary structure. Phylogenetic conservation is derived from the comparative analysis of 6326 sequences (Reproduced from CRW).

### 1.2.2 23S Ribosomal RNA

The 23S rRNA is part of the large subunit (LSU) of the ribosome, which is the center of amino acid polymerization, the main catalytic function during protein translation [11]. The 23S rRNAs are not as highly conserved as the 16S, varying in length from $\sim$2900 nucleotides (prokaryotic 23S) to $\sim$4700 nucleotides (eukaryotic 28S).

The 23S rRNA sequence was first reported in 1980 [12], and was soon followed by a comparative secondary structure model [78]. The CRW now lists a total of 11610 23S rRNA sequences, and 86 comparative secondary structure models [16].

### 1.2.3 Evolution of ribosomal RNAs

Sites in the rRNAs do not evolve independently, but are constrained by selection to maintain base complementarity in the paired regions. Both paired and unpaired regions have been shown to contain phylogenetic signal [29].

Because stems in rRNAs are assumed to be largely structural, any substitution of one base pair for another should typically be acceptable, borne out by the extensive presence of non-canonical base pairs. In contrast, unpaired regions are thought to depend more specifically on their sequence. It has been observed that most of the highly conserved regions in 16S rRNAs, with little to no variability at the sequence level, were unpaired. Base pairing, therefore, appears to be a weak constraint on sequence compared to other influences on the sequence near the active site of the ribosome.

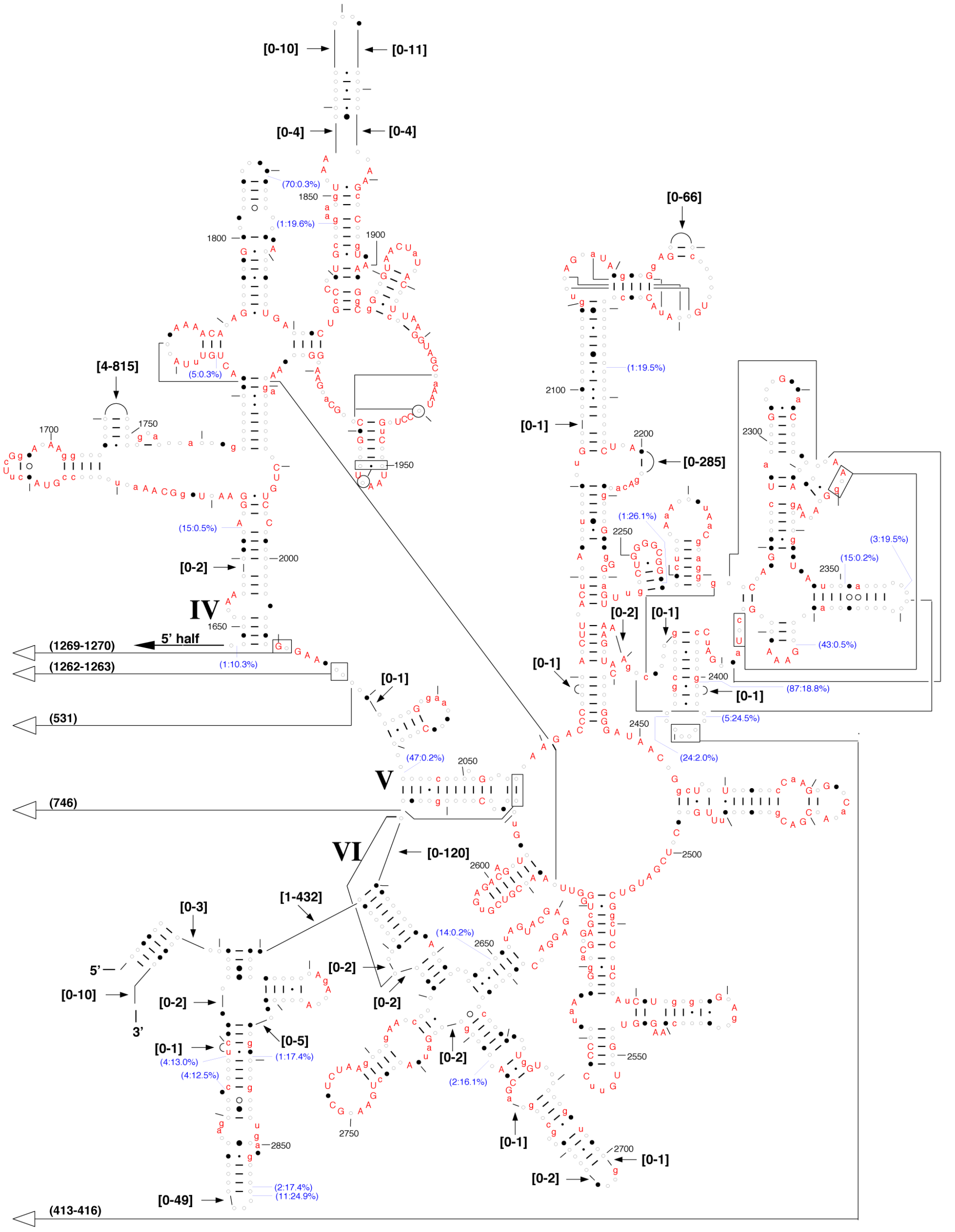

**Fig.** 1.2: **The conserved secondary structure of the 23S ribosomal RNA (3' end) (Reproduced from CRW).** See Figure 1.3 for 5' end.

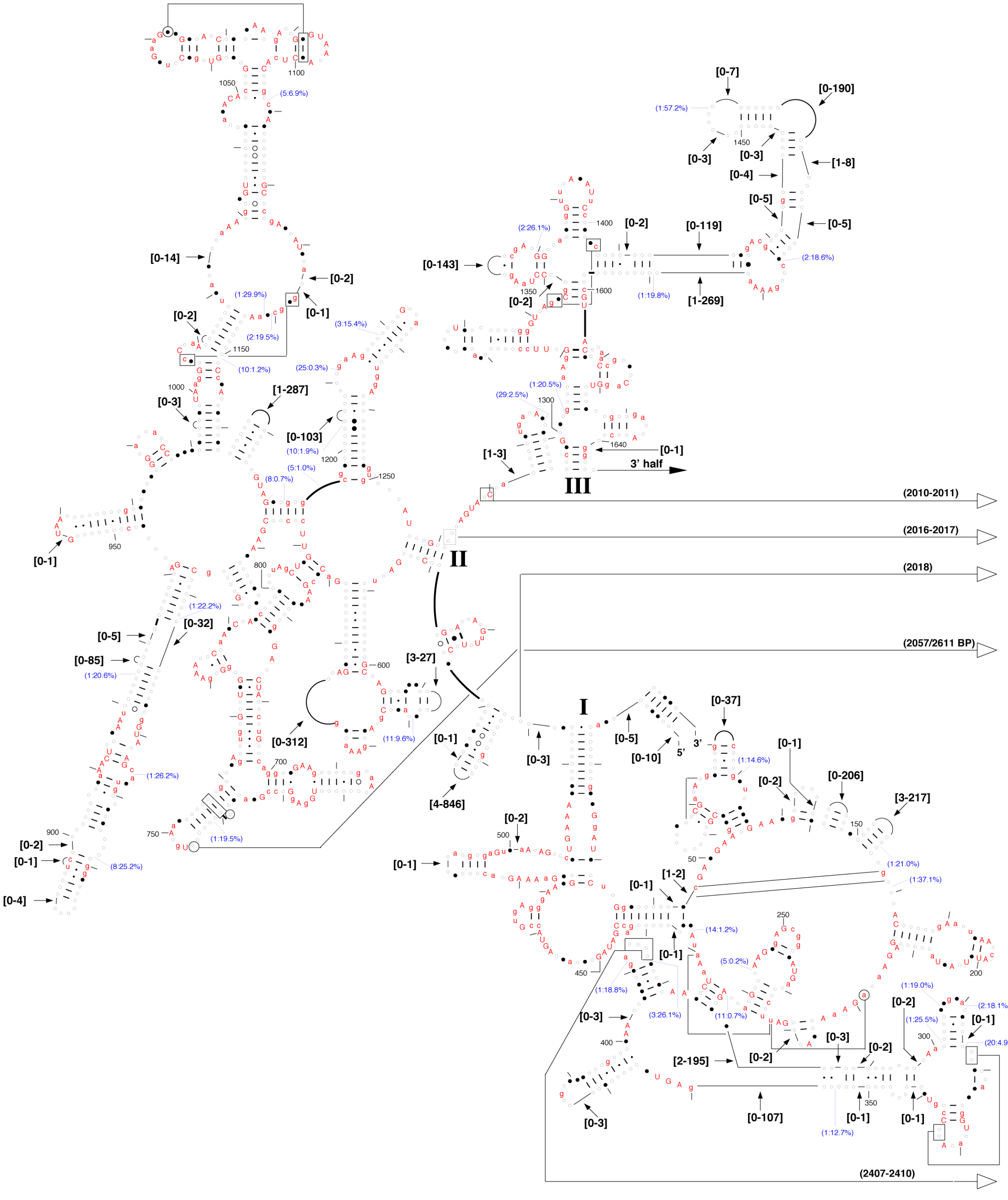

**Fig.** 1.3: **The conserved secondary structure of the 23S ribosomal RNA (5' end).** Positions in the 23S ribosomal RNA with a nucleotide in more than 95% of the sequences are shown superimposed onto the *E. coli* secondary structure. Phylogenetic conservation is derived from the comparative analysis of 592 sequences (Reproduced from CRW).

Different categories of rRNA secondary structure show distinct, characteristic base compositions. However, these patterns of variation are similar among sequences from 16S and 23S rRNAs, and across all domains of life [95]. Structural categories in the ribosomal RNA have been found to evolve at different rates, with the rates varying across phylogenetic domains: in the bacteria and the archaea, stems evolve faster, while in the eukarya, loops evolve faster. While highly conserved regions tend to be unpaired, the converse is not always true [94].

### 1.2.4 Comparative Analysis of RNA sequences

Zuckerkandl and Pauling observed in 1962, from a comparison of the amino acid sequence of haemoglobin from various species, that the more varied two species are, their haemoglobin sequences differ by a greater number of amino acids [117].

All the above studies use folding free energies to quantify the potential of a sequence to form secondary structure. While this approach has been successful in predicting the structures of small RNAs, it may be undesirable for the analysis of longer RNAs for several reasons. First, the minimum free energy structure may not be the structure that is formed *in vivo*, due to the effect of several factors like the directionality and velocity of transcription, binding of ribosomes, RNA chaperones and other RNA binding proteins, presence of metal ions and small noncoding RNAs, etc. [91]. Second, it has been shown that as the length of RNA increases, fold prediction methods that rely on free energy criteria perform less accurately [57]. Finally, it is now recognized that RNA sequence evolution is constrained by structure. It is therefore desirable to infer RNA structure and function using comparative methods.

Comparative analyses of rRNA sequences are already being used to determine the two-dimensional, i.e., secondary structure of rRNAs and enhance fundamental understanding of the rRNAs three-dimensional, i.e., tertiary structure. The underlying assumption of these comparative analyses is that sequence positions with similar patterns of variation across multiple organisms are base-paired in the rRNA structure [33, 48, 49]. The determination of the high resolution crystal structures of the ribosome [10, 90, 109] substantiate these secondary and tertiary structure models, with approximately 97% of the proposed base pairs present in the crystal structures.

The comparative analyses of sequence alignments require mathematical tools that are able to simultaneously identify relations of similarity and dissimilarity among the organisms, as well as the corresponding sequence positions and nucleotides that underlie these relations. These tools should provide mathematical frameworks for the modeling of these data, where the mathematical variables, i.e., significant patterns, that are uncovered in the data, of nucleotide-specific frequency variation across the organisms and sequence positions, represent biological reality.

## 1.3 Genomic Signal Processing

Tools from matrix algebra have been used, with great success, for the integrative analysis and modeling of large-scale biological data. Studies on genome-wide microarray expression data have shown that singular value decomposition describes the overall observed signal as the outcome of a simple network, where a few independent sources of variation affect the genes and samples in the dataset [3]. This model has been successfully extended to the comparative analysis of mRNA expression from two organisms using the generalized singular value decomposition [4], and in the integrative analysis of mRNA expression as well as DNA copy number data using pseudo-inverse projection [5] (Figure 1.4).

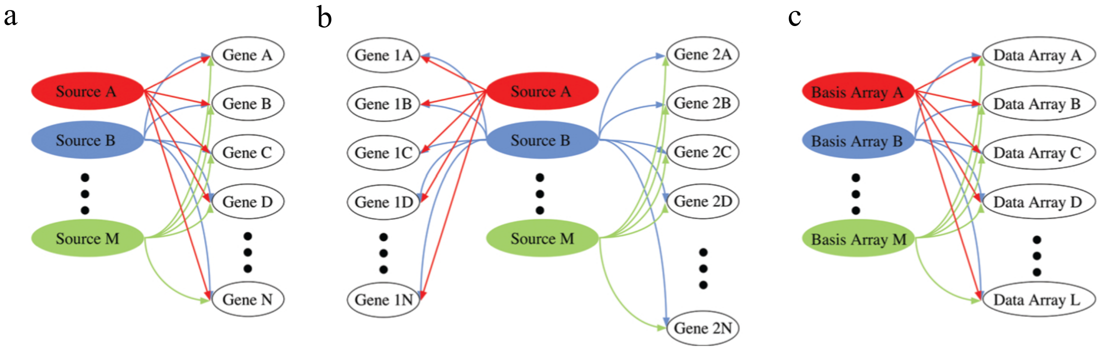

**Fig.** 1.4: **Mathematical models for DNA microarray data derived from genomic signal processing techniques (reproduced from Alter, 2007 [2]).**
(a) The SVD model describes the data as the outcome of a simple linear network, with a few independent sources (experimental or biological) affecting all the genes and arrays in the dataset. (b) The GSVD model describes the two datasets as the outcome of a simple linear comparative network, with a few independent sources, some common to both datasets whereas some are exclusive to one dataset or the other, affect all the genes in both datasets. (c) The pseudoinverse projection integrative model approximates any number of datasets as the outcome of a simple linear integrative network, where the cellular states, which correspond to one chosen basis set of observed samples, affect all the samples, or arrays, in each dataset.

Recently, the application of tensor decomposition methods has resulted in the prediction as well as experimental verification of genome-scale correlations between DNA replication and mRNA transcription in *S. cerevisiae* [80, 81]. The application of these signal processing methods has now created a framework where biological data may be analysed and modeled the way physical systems are today.

## 1.4 Mathematical Framework

### 1.4.1 Singular Value Decomposition

The Singular Value Decomposition (SVD) [45] is also known as Karhunen–Loève expansion in pattern recognition, and is similar to Principal Component Analysis (PCA) in statistics. SVD finds applications in signal processing, image compression, solutions to inverse problems, etc. Singular value decomposition is closely related to the eigenvalue decomposition, and in the case of Hermitian positive semi-definite matrices, the SVD is the same as the EVD.

If $D$ is an $m{\times}n$ matrix with $m{>}n$ then the SVD of $D$ is the linear transformation is given by:

$$D = U\Sigma V^T \tag{1.1}$$

$U_{m\times n}$ and $V^T_{n\times n}$ are orthogonal matrices, and $\Sigma_{n\times n}$ is a diagonal matrix whose elements are the ordered singular values of $D$. Each column of $U$ is associated with only the row of $V^T$, with the corresponding $\sigma$ indicating their relative significance.

### 1.4.2 Tensors

A tensor is a multidimensional or $N$-way array [67]. Tensors have been recognized as a logical way to model multidimensional biological data, and have been used successfully in chemometrics [93], psychometrics [51], and more recently, in genomic signal processing [80].

Of the several tensor decompositions, CANDECOMP/PARAFAC and Tucker decomposition (N-mode SVD) can be considered higher-order generalizations of the matrix SVD. The PARAFAC (Parallel Factorization) or CANDECOMP (Canonical Decomposition), variously attributed to Hitchcock [54, 55], Cattell [20, 21], Carroll and Chang [17], and Harshman [51], is a rank-k approximation that preserves the diagonality of the core tensor. The Tucker decomposition [106] or HOSVD [28], on the other hand, is an exact decomposition that preserves the orthogonality of the singular vectors. This is the decomposition that will be discussed for our application.

### 1.4.3 HOSVD

The N = 3-mode SVD, a Higher-Order SVD (HOSVD) [28] of the third-order data tensor, is a multilinear transformation of the data tensor $T_{K\times L\times M}$ given by:

$$T = R \times_a U \times_b V_x \times_c V_y \tag{1.2}$$

where $\times_a U$, $\times_b V_x$, and $\times_c V_y$ denote multiplications of the tensor and the matrices $U$, $V_x$, and $V_y$, which contract the first, second, and third indices of with the second indices of $U$, $V_x$, and $V_y$, or, equivalently, the first indices of $U^T$, $V_x^T$, and $V_y^T$, respectively.

To ensure ease of interpretation, the decomposition in (Eq 2.2) may be reformulated such that it decomposes $T$ into a linear superposition of rank-1 subtensors, with the superposition coefficients tabulated in the core tensor $R$ [66]:

$$T = \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} R_{abc} U_a \otimes V_{x,b:}^T \otimes V_{y,c:}^T = \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} R_{abc} S(a,b,c) \tag{1.3}$$

where the subtensor $S(a,b,c)$ is the outer product of the eigenvectors $U_{:,a}$, $V_{x,b:}^T$, and $V_{y,c:}^T$.

In the integrative analysis of DNA microarray data from different studies, HOSVD has been shown to identify the effects of different drugs on cell cycle progression, and the genes associated with these effects [80].

### 1.4.4 Matrix Decompositions and Sequence Analysis

Several matrix-based methods have found application in the analysis of sequences of proteins, RNAs, and even whole genomes.

Fogolari *et al.* used the SVD as a dimensionality-reduction tool, to analyze a matrix of pairwise similarity scores of proteins in the calycin superfamily [39].

Lee and Seung pioneered the use of the Non-negative Matrix Factorization (NMF) for image analysis, with the aim of obtaining basis vectors that are non-subtractive linear combinations of the data [69]. Heger and Holm applied this principle to a hierarchical clustering of distantly related proteins (40% overall sequence homology) from the urease superfamily, in order to obtain 'fuzzy' alignments [52].

Stuart and Berry developed an SVD-based method for reconstructing phylogeny from whole genome sequences of bacteria, encoded using a correlated peptide score [98]. Kitazoe *et al.* reformulated the phylogeny reconstruction problem as the successive splitting of branch vectors in a multidimensional vector space (MVS) [64, 65].

Pazos *et al.* used the Multiple Correspondence Analysis (MCA), a multivariate extension of the PCA, on protein sequence similarity scores, to detect functionally significant residues in SH3 domains and TIM-barrel hydrolases [84]. Their method claims to be independent of phylogeny, in that it uses an *a priori* definition of functional classes that is independent of the phylogeny implicit in the sequence alignment.

Paschou *et al.* used a PCA-based algorithm to detect population structure in SNPs derived from admixed human populations, without prior knowledge of ancestry [83]. Building upon this idea, Mahoney and Drineas proposed the CUR decomposition, a low-rank matrix decomposition, as a more biologically relevant representation of the SNP data [72].

Casari *et al.* first proposed the PCA as a tool to analyze protein sequence similarity in the Ras-Rab-Rho superfamily [18]. They showed that the principal components of the protein similarity matrix identify the directions in protein sequence space most strongly populated by members of the three protein families. Although they used a 20-bit vector representation for each protein sequence, they did not explore this dimension of the data in their analysis. Building upon this idea, Sagara *et al.* used PCA recursively on an alignment of tRNA sequences, to detect amino acid-specific clusters in sequence space [88]. They then used multi-dimensional scaling (MDS) to trace the principal components back to individual bases and positions that characterize individual groups. Suh *et al.* found from the PCA analysis of a matrix of Group 1 intron sequence distances that several previously unclassified sequences clustered together, separately from the recognized structural classes, leading them to propose a new class of Group 1 introns [99].

While the above studies illustrate the widespread applicability of matrix decompositions in the analysis of sequence data to derive biologically meaningful results, they suffer from two major limitations. First, the studies that use the SVD fail to fully exploit its ability to simultaneously classify sequences along not one, but both dimensions of the matrix. Second, they flatten inherently cuboidal data into a matrix, thus losing information along the third dimension of nucleotides.

## 1.5 Our Aims

The evolutionary forces that act on genomes are essentially stochastic. Detecting significant similarities between anciently diverged sequences in the background of random mutation, natural selection, and genetic drift may therefore be viewed as a signal to noise problem.

We therefore propose matrix decomposition-based algorithms for the comparative analysis of sequences, as a method to simultaneously classify sequences in an alignment into clusters and identify the signatures defining these clusters, compare these patterns among different datasets, and integrate data from various sources, with the ultimate goal of being able to create predictive models. By using a tensor HOSVD, we ensure that the information contained in the nucleotide dimension is not lost.

Our method will be data-driven, and allow for the simultaneous classification of the sequences in the alignment, and identification of positions in the alignment that contribute to the classes, without requiring *a priori* definitions of the classes, to enable the discovery of known as well as new relationships between sequences in the data.

We use the rRNA as our model so that the relationships we discover among the sequences may be verified against known phylogenetic relationships. We use only sequences with known structure models, so that the positions we identify may be correlated with structure elements, and lead to hypotheses about RNA folding and function.

## 1.6 Organization

This dissertation is organized as follows. Chapter 2 describes the data used in our HOSVD analyses, along with the mathematical and computational methods developed for this purpose. Chapter 3 lists the results we obtained from the analysis of 16S and 23S rRNA alignments. A discussion of these results in an evolutionary and structural context is presented in Chapter 4, followed by conclusions and proposed future research.

We also analysed an alignment of 5S rRNA using the mode-1 HOSVD: a discussion of these results appears in Appendix 1. Finally, supplementary tables are presented in Appendix 2.

# Chapter 2

# Materials and Methods

This chapter describes the mathematical and computational methods we developed for the mode-1 HOSVD analysis of rRNA sequence alignments. Figure 2.1 gives an overview of the steps involved in the analysis. These steps are described in detail in the sections to follow.

## 2.1 Data

### 2.1.1 Alignment

We describe results from the analysis of 16S and 23S rRNA sequence alignments. The sequences, obtained from the Comparative RNA Website (CRW) [16], represent all 16S, 23S, and 5S sequences for which a secondary structure model is available. The organisms in these alignments are from different National Center for Biotechnology Information (NCBI) Taxonomy Browser groups [89].

The compositions of the three ribosomal RNA alignments we analyzed are shown in Table 2.1.

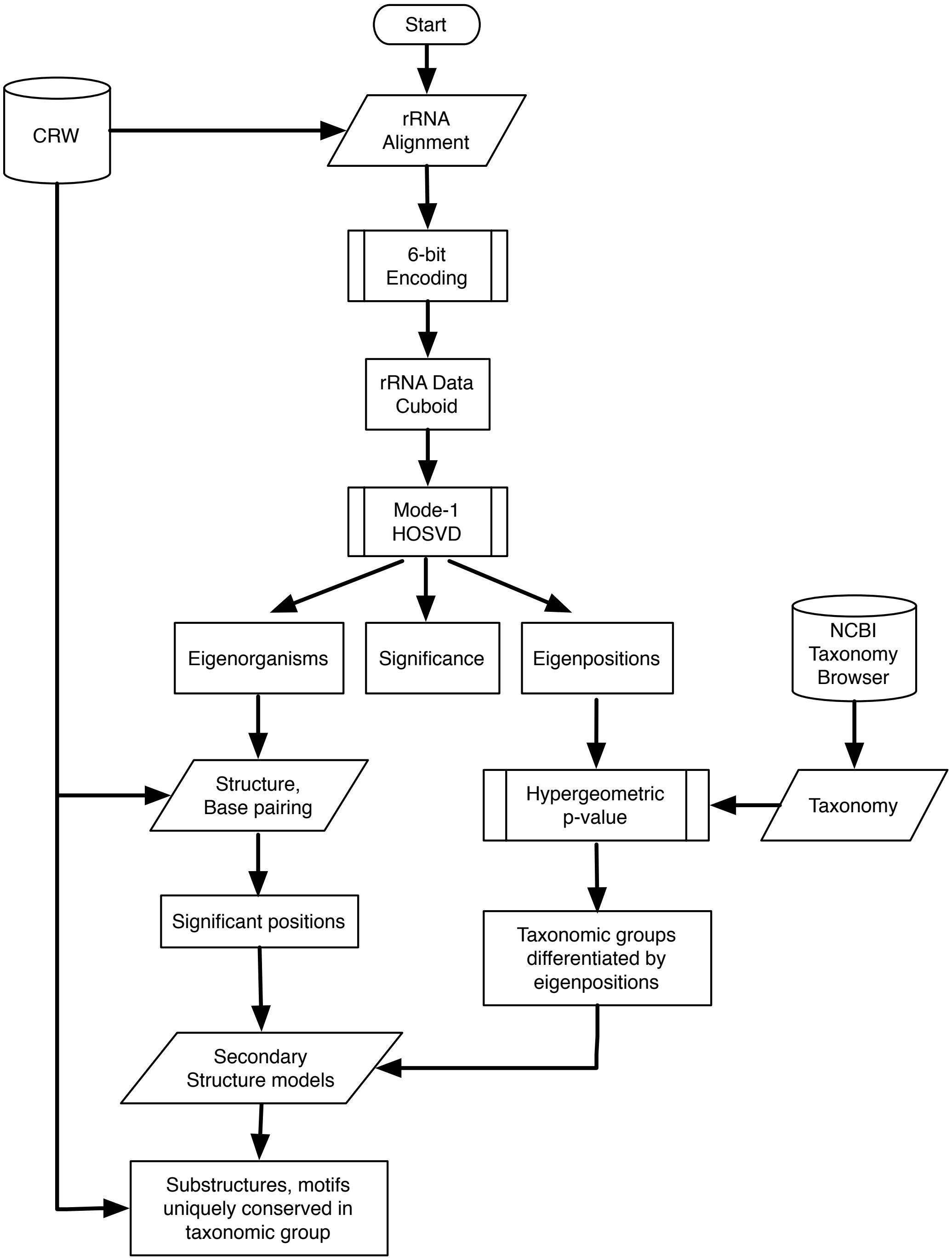

**Fig.** 2.1: **Flowchart showing the steps involved in the analysis of ribosomal RNA alignments using Mode-1 HOSVD.**

| **rRNA Alignment** | **Positions** | **Organisms** | | | |
|---|---|---|---|---|---|
| | | **Total** | **Archaea** | **Bacteria** | **Eukarya** |
| 16S | 3249 | 339 | 21 | 175 | 143 |
| 23S | 6636 | 75 | 6 | 57 | 12 |
| 5S | 152 | 242 | 28 | 83 | 131 |

Table 2.1: **Composition of rRNA alignments**

### 2.1.2 Structure

For each sequence in the 16S and 23S rRNA alignments, we obtain base pairing (‘*.bpseq’) files from the CRW [16]. These files tabulate each nucleotide’s base-pairing status, and where relevant, list the position that the nucleotide is base-paired to. We also obtain for each sequence, the structure (‘*.alden’) files from the CRW, which classify each nucleotide into one of six structural categories, following Smit, *et. al.* (Figure 2.2, [95]).

These base-pairing and structure files are then used to annotate each position in the alignment, as described in Section 2.3.2.1.

### 2.1.3 Taxonomy

For each sequence in the alignment, we retrieve its organismal taxonomy from the NCBI Taxonomy Browser [89]. We then assign to each sequence five annotations, based on the five topmost hierarchical levels defined in the Taxonomy Browser for that organism. The sequences in the 16S, 23S, and 5S alignments are

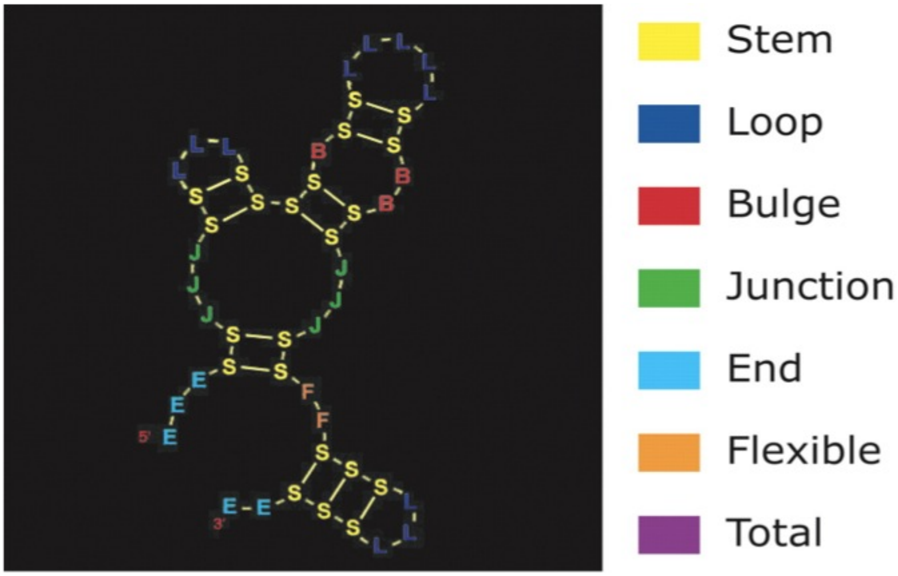

**Fig.** 2.2: **Structure categories in the ribosomal RNAs (reproduced from Smit,** *et. al.*, **2006 [95])**

listed in Appendix 2 along with their NCBI Taxonomy annotations.

## 2.2 Mathematical Framework

### 2.2.1 Encoding

The 16S alignment matrix we analyze tabulates six sequence elements or "nucleotides," i.e., A, C, G and U nucleotides, unknown ("N") and gap ("–"), across the 339 organisms and the 3249 sequence positions with A, C, G or U nucleotides in at least 1% of the 339 organisms. Similarly, the 23S alignment matrix tabulates six sequence elements across the 75 organisms and the 6636 sequence positions with A, C, G or U nucleotides in at least 1% of the 75 organisms.

A six-bit binary encoding [88],

$$\begin{aligned}A &= (1,0,0,0,0,0)\\ C &= (0,1,0,0,0,0)\\ G &= (0,0,1,0,0,0)\\ U &= (0,0,0,1,0,0)\\ N &= (0,0,0,0,1,0)\\ - &= (0,0,0,0,0,1),\end{aligned} \tag{2.1}$$

transforms each alignment matrix into a third-order tensor, i.e., a cuboid, of six "slices," one slice for each nucleotide, tabulating the frequency of this nucleotide across the organisms and positions (Figure 2.3).

### 2.2.2 Mode-1 HOSVD

The mode-1 HOSVD transforms each $K$-organisms × $L$=6-nucleotides × $M$-positions data tensor $\mathcal{D}$ into the reduced and diagonalized $K$-"eigenpositions" × $K$-"eigenorganisms" matrix $\Sigma$, by using the $K$-eigenorganisms × $L$=6-nucleotides × $M$-positions transformation tensor $\mathcal{U}$ and the $K$-organisms × $K$-eigenpositions transformation matrix $V^T$,

$$\mathcal{D} \quad = \quad \mathcal{U} \quad \Sigma V^T. \tag{2.2}$$

This mode-1 HOSVD is computed from the singular value decomposition (SVD) [3, 45, 80, 81] of each data tensor unfolded along the $K$-organisms axis such that its nucleotide-specific slices $D_i$ are appended along the organisms axis,

$$\begin{pmatrix} D_A \\ D_C \\ D_G \\ D_U \\ D_N \\ D_- \end{pmatrix} = \begin{pmatrix} U_A \\ U_C \\ U_G \\ U_U \\ U_N \\ U_- \end{pmatrix} \Sigma V^T. \tag{2.3}$$

The transformation tensor $\mathcal{U}$ is obtained by stacking the nucleotide-specific slices $U_i$ along the organisms axis.

### 2.2.3 Interpretation

The “eigenpositions”, or $V_i^T$, are the patterns of variation among the organisms. We show in the following sections that these eigenpositions correspond to phylogenetic variation among the organisms examined. The “eigenorganisms”, or nucleotide-specific $U_i$, are patterns of variation among the positions in the alignment. They represent the relative nucleotide frequency of positions in the alignment, and identify positions that uniquely characterize taxonomic groups uncovered by the corresponding eigenposition.

**Eigenposition:** An eigenposition is a *position-like* vector that describes the variation in the data across the *organisms*. The eigenpositions are orthogonal to one another, i.e., the patterns of variation that they describe are uncorrelated. There are as many eigenpositions as there are organisms.

The significance of each eigenposition and the corresponding eigenorganism, is defined in terms of the fraction of the overall information that these orthogonal patterns of nucleotide frequency variation across the $K$-organisms and $L$=6-nucleotides × $M$-positions, respectively, capture in the data tensor and is proportional to the corresponding singular value that is listed in $\Sigma$, that is, $\sigma_i$. These singular values are ordered in decreasing order, such that the patterns are ordered in decreasing order of their relative significance.

This fraction $p_i$ is calculated as:

$$p_i = \frac{\sigma_i^2}{\sum_{k=1}^{L} \sigma_k^2} \tag{2.4}$$

The normalized Shannon entropy of the dataset:

$$0 \leq d = -\frac{1}{L} \sum_{k=1}^{L} p_k \log p_k \leq 1 \tag{2.5}$$

measures the complexity of the data from the distribution of the overall nucleotide frequency variation between the different eigenpositions and corresponding eigenorganisms, where d = 0 corresponds to an ordered and redundant dataset in which all nucleotide frequency variation is captured by one eigenposition and the corresponding eigenarray, and d = 1 corresponds to a disordered and random dataset where all eigenpositions and eigenorganisms are equally significant.

**Eigenorganism:** An eigenorganism is an *organism-like* vector that describes the variation in the data across the *nucleotides*×*positions*. The eigenorganisms, like the eigenpositions, are orthogonal to one another, i.e., they describe uncorrelated patterns of variation. There are as many eigenorganisms as there are organisms, and each eigenorganism is associated with one eigenposition.

Figure 2.3 shows the Mode-1 higher-order singular value decomposition (HOSVD) of the 16S rRNA alignment. The structure of the alignment is of an order higher than that of a matrix. The organisms, the positions, as well as the "nucleotides," i.e., sequence elements (Equation 2.3), each represent a degree of freedom in a cuboid, i.e., a third-order tensor. We compare these data by using a tensor mode-1 HOSVD, which uncovers in the data tensor "eigenpositions" and nulceotide-specific segments of "eigenorganisms," i.e., patterns of nucleotide frequency variation across the organisms and positions, respectively (Equation 2.2). This is depicted in a raster display with increased nucleotide frequency (red), no change in frequency (black) and decreased frequency (green) relative to the average frequency variation across the organisms and positions, which is captured by the most significant eigenposition and eigenorganism, respectively.

## 2.3 Data Analysis

### 2.3.1 Organisms

Eigenpositions are correlated and anticorrelated with taxonomic groups in a data-driven manner as follows. For each eigenposition, we calculate the probabilistic enrichment of all five taxonomic levels among the organisms most correlated and anticorrelated with the eigenposition, under the assumption of the hypergeometric distribution, as described by Tavazoie *et al* [102]. The $P$-value of a given association is the hypergeometric probability of the $J$ annotations among the $K$ organisms, and of the subset of $j \subseteq J$ annotations among the subset of $k$ organisms:

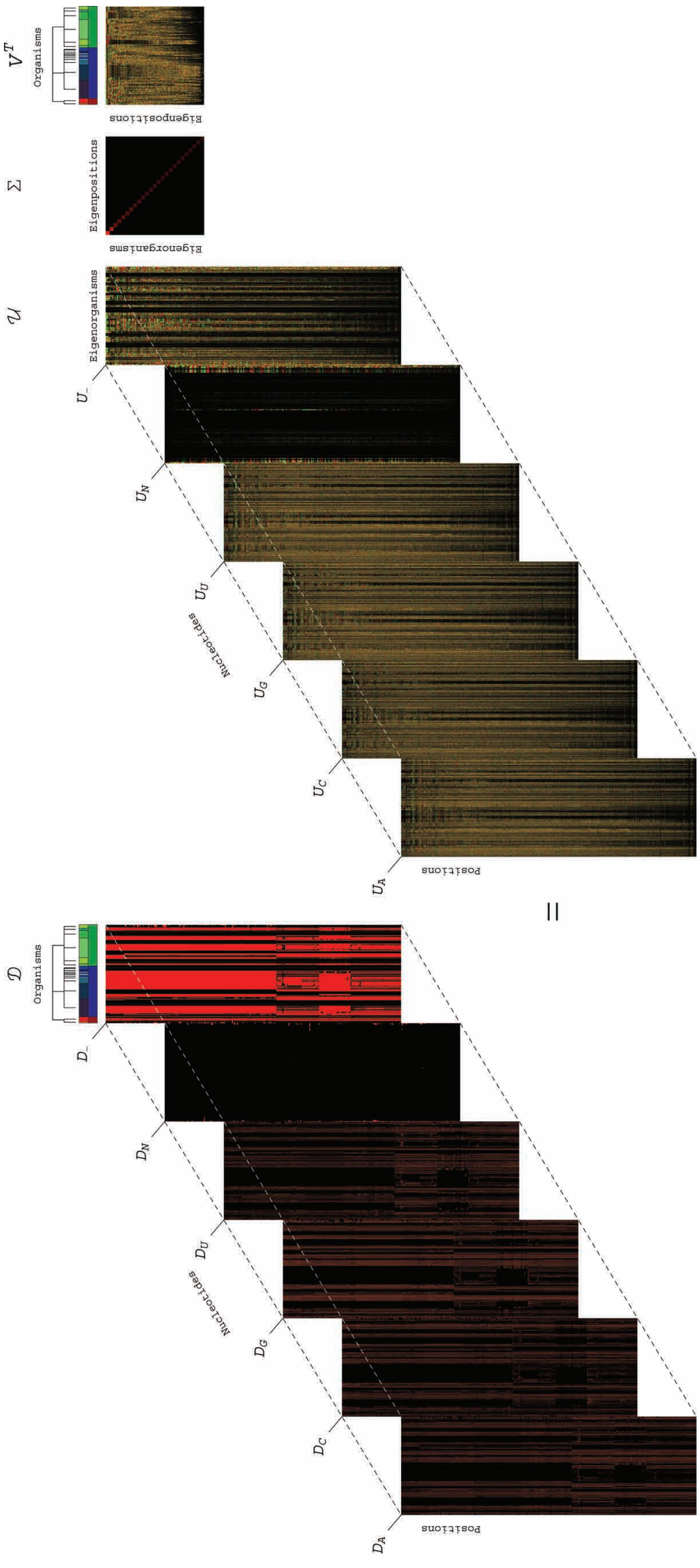

**Fig.** 2.3: **Mode-1 higher-order singular value decomposition (HOSVD) of the 16S rRNA alignment.**

$$P(j;k,K,J) = \binom{K}{k}^{-1} \sum_{i=j}^{k} \binom{J}{i} \binom{K-J}{k-i} \tag{2.6}$$

where $\binom{N}{n}$ is the Newton binomial coefficient, given by:

$$\binom{N}{n} = N!n!^{-1}(N-n)!^{-1} \tag{2.7}$$

The taxonomic groups with the most significant enrichments in the subsets of $k$ most correlated and anticorrelated organisms are then used for the analysis of the corresponding eigenorganisms.

### 2.3.2 Positions

We annotate positions based on structure information obtained from the CRW. For the analysis of each eigenorganism, we annotate all positions according to the phylogenetic groups separated by the corresponding eigenposition, as described below.

#### 2.3.2.1 Conservation

We define exclusive nucleotide or gap conservation as conservation of the nucleotide or gap within at least 80% of the organisms of the corresponding taxonomic group but in less than 20% of the remaining organisms.

Similarly, we define exclusive paired (or unpaired) nucleotide conservation as conservation of the nucleotide within at least 80% of the organisms of the group but in less than 20% of the remaining organisms, together with greater frequency of paired (or unpaired) nucleotides within the group rather than among the remaining organisms.

We define structure motifs as conservation of the motif within at least 60% of the organisms of the group.

#### 2.3.2.2 Enrichment

We calculate the enrichment of each structural attribute (nucleotides, paired or unpaired nucleotides, structure motifs) in the positions most positively and negatively correlated with each eigenorganism. The $P$-value of a given association is calculated assuming hypergeometric probability distribution of the $J$ annotations among the $K$ total positions in the alignment, and of the subset of $j \subseteq J$ annotations among the subset of $k$ positions most positively and negatively correlated with each eigenorganism, as described by Tavazoie *et al* [102] (Equation 2.6).

Figure 2.4 shows the significant eigenpositions uncovered by the mode-1 HOSVD in the 16S rRNA alignment, and their correlation with taxonomic groups from the NCBI Taxonomy Browser [89]. The classification of the organisms in the alignment into taxonomic groups according to the top six hierarchical levels of the NCBI Taxonomy Browser is shown in (*a*). The 25 most significant eigenpositions are displayed in raster form in (*b*), with increased frequency (red), no change in frequency (black) and decreased frequency (green) relative to the average frequency variation across the organisms, captured by the most significant eigenposition. The fractions of nucleotide frequency variation that the 25 most significant eigenpositions capture in the 16S alignment is displayed as a bar chart in (*c*).

The corresponding results for the 23S rRNA alignment are displayed in Figure 2.5 (*a*), (*b*), and (*c*) respectively.

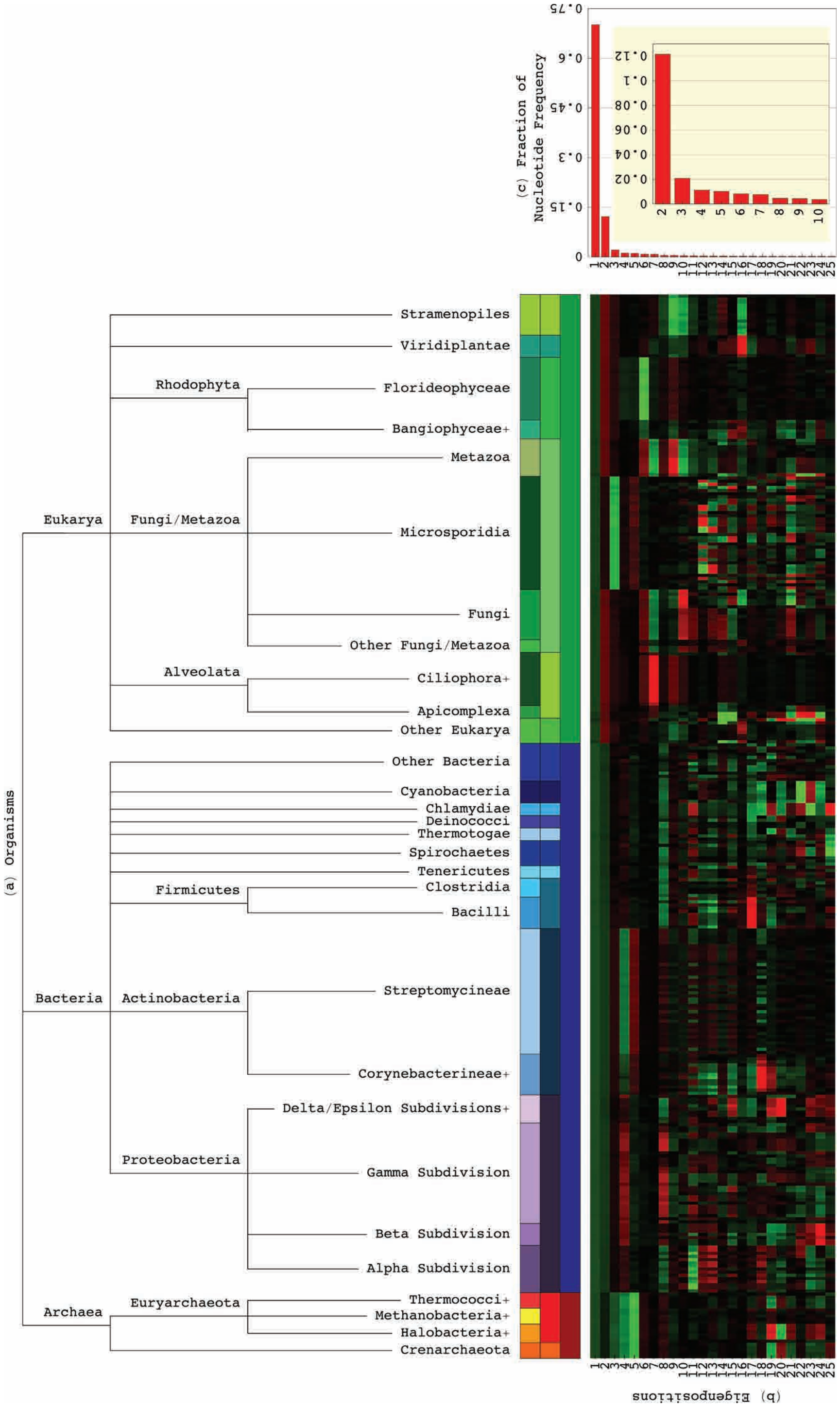

Fig. 2.4: **Significant 16S eigenpositions and their correlation with the NCBI Taxonomy Browser taxonomic groups.**

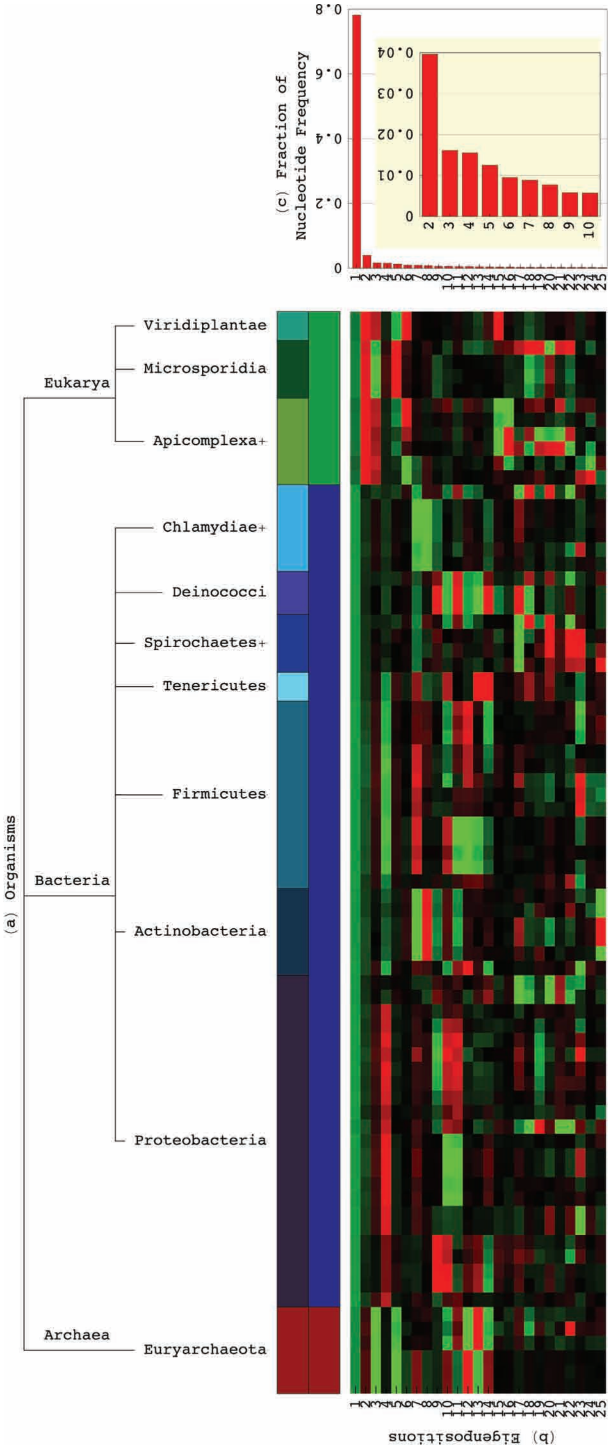

Fig. 2.5: **Significant 23S eigenpositions and their correlation with the NCBI Taxonomy Browser taxonomic groups.**

# Chapter 3

# Results

In this chapter, the results from the Mode-1 HOSVD on the 339-sequence 16S rRNA alignment and the 75-sequence 23S rRNA alignment are presented [76].

We correlate and anticorrelate an eigenposition with increased relative nucleotide frequency across a taxonomic group according to the NCBI Taxonomy Browser annotations [89] (Figures 2.4 and 2.5) of the two groups of $k$ organisms each, with largest and smallest levels of nucleotide frequency in this eigenposition among all $K$ organisms, respectively. The $P$-value of a given association is calculated assuming hypergeometric probability distribution of the $J$ annotations among the $K$ organisms, and of the subset of $j \subseteq J$ annotations among the subset of $k$ organisms (Equation 2.6) [102].

## 3.1 Most significant eigenposition is invariant

In the 16S alignment, the seven most significant eigenpositions and corresponding eigenorganisms uncovered capture $\sim$88$\%$ of the nucleotide frequency information in the alignment (Figures 2.4). Similarly, in the 23S alignment, the five most significant eigenpositions and corresponding eigenorganisms capture 87$\%$ of the information (Figure 2.5). In both alignments, the most significant eigenposition is approximately invariant across the organisms, and correlates with the average frequency of all nucleotides across the positions with the correlation $> 0.995$ [15]. The correlation of each nucleotide-specific segment of the most significant eigenorganism with the average frequency of this nucleotide across the positions is $> 0.999$.

We interpret the remaining eigenpositions and the nucleotide-specific segments of the corresponding eigenorganisms as patterns of nucleotide frequency variation relative to these averages. We find that the patterns uncovered in the 16S and 23S are qualitatively similar.

## 3.2 Eigenpositions correspond to phylogenetic groups

The remaining significant eigenpositions uncovered in both the 16S (Figure 3.1) and 23S (Figure 3.2) data cuboids reveal the dominant taxonomic groups among the organisms and their relations of similarity and dissimilarity.

### 3.2.1 Eigenpositions in the 16S rRNA

Among the 16S rRNAs, the second through seventh most significant eigenpositions (Figure 3.1) describe relationships among the taxonomic groups as follows. The second most significant eigenposition ((*a*), red) differentiates the Eukarya excluding the Microsporidia from the Bacteria, as indicated by the color bar (Table 3.1(a)). The fourth ((*a*), blue) distinguishes between the Gamma Proteobacteria and the Actinobacteria and Archaea. The third ((*b*), red) and fifth ((*b*), blue) eigenpositions describe the similar and dissimilar among the Archaea and Microsporidia, respectively. The sixth ((*c*), red) and seventh ((*c*), blue) eigenpositions differentiate the Fungi/Metazoa excluding the Microsporidia from the Rhodophyta and the Alveolata, respectively.

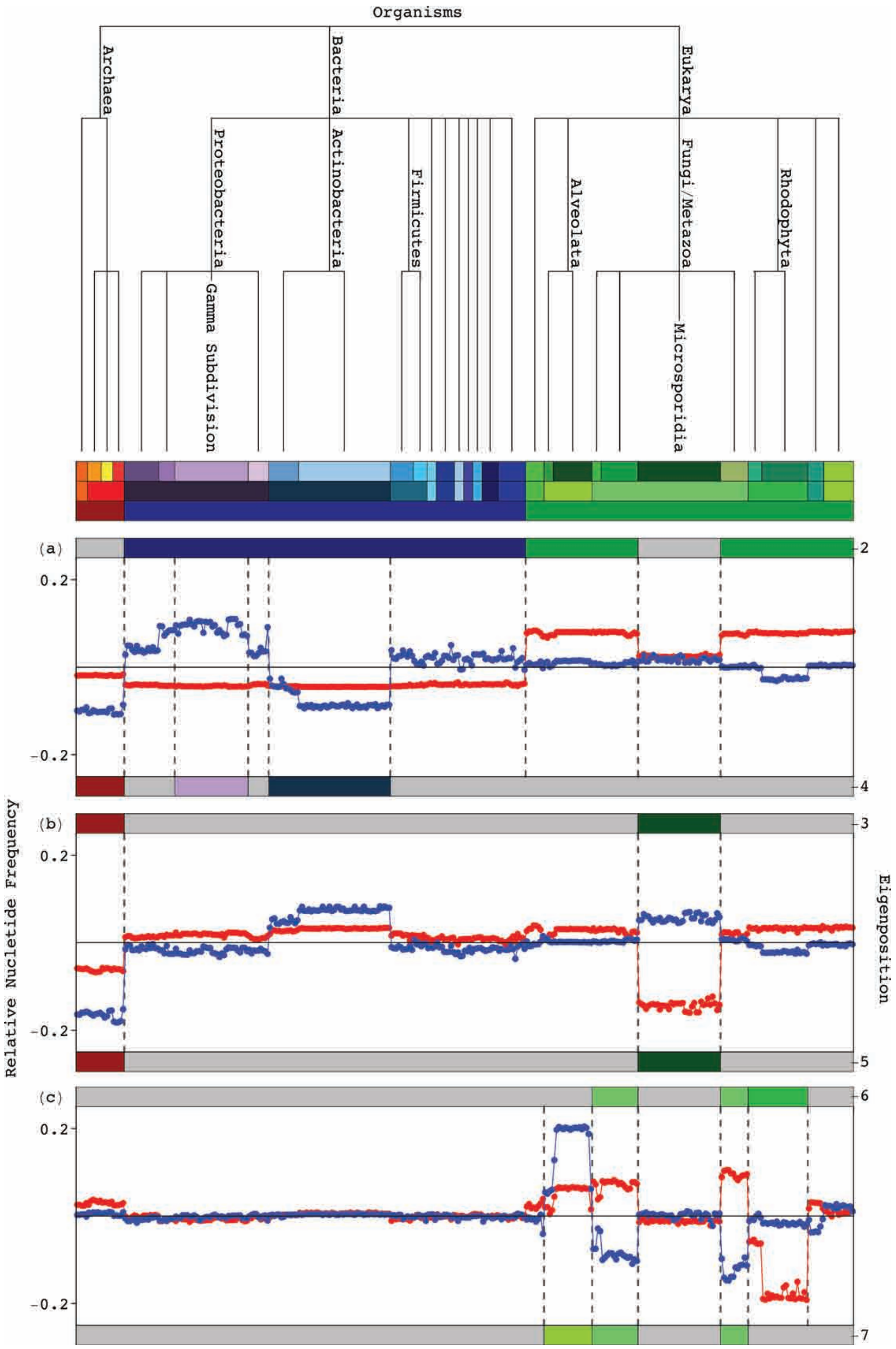

**Fig.** 3.1: **Significant 16S eigenpositions.** Line-joined graphs of the second through seventh 16S eigenpositions, i.e., patterns of nucleotide frequency across the organisms, and their correlation with the taxonomic groups in the 16S alignment, classified according to the top six hierarchical levels of the NCBI Taxonomy Browser [89] (Figure 2.4).

(a) Probabilistic significance of the enrichment of the $k$=75 organisms in the 16S rRNA

| **16S Eigenposition** | **Correlated** | | | | **Anticorrelated** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Group** | $n$ | $N$ | **p-value** | **Group** | $n$ | $N$ | **p-value** |
| 2 | Eukarya-Microsporidia | 75 | 107 | $5.7 \times 10^{-50}$ | Bacteria | 75 | 175 | $1.5 \times 10^{-26}$ |
| 3 | | | | | Archaea+Microsporidia | 57 | 57 | $3.4 \times 10^{-49}$ |
| 4 | Gamma Proteobacteria | 32 | 32 | $2.0 \times 10^{-24}$ | Actinobacteria+Archaea | 72 | 74 | $2.5 \times 10^{-67}$ |
| 5 | Microsporidia | 29 | 36 | $1.8 \times 10^{-15}$ | Archaea | 21 | 21 | $1.5 \times 10^{-15}$ |
| 6 | Fungi/Metazoa-Microsporidia | 32 | 32 | $2.0 \times 10^{-24}$ | Rhodophyta | 26 | 26 | $1.8 \times 10^{-19}$ |
| 7 | Alveolata | 21 | 21 | $1.5 \times 10^{-15}$ | Fungi/Metazoa-Microsporidia | 32 | 32 | $2.0 \times 10^{-24}$ |

(b) Probabilistic significance of the enrichment of the $k$=15 organisms in the 23S rRNA

| **23S Eigenposition** | **Correlated** | | | | **Anticorrelated** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Group** | $n$ | $N$ | **p-value** | **Group** | $n$ | $N$ | **p-value** |
| 2 | Eukarya-Microsporidia | 8 | 8 | $3.8 \times 10^{-7}$ | Bacteria | 15 | 57 | $9.7 \times 10^{-3}$ |
| 3 | | | | | Archaea+Microsporidia | 10 | 10 | $3.6 \times 10^{-9}$ |
| 4 | Proteobacteria | 15 | 23 | $2.2 \times 10^{-10}$ | Firmicutes | 12 | 13 | $2.2 \times 10^{-10}$ |
| 5 | Microsporidia | 4 | 4 | $1.1 \times 10^{-3}$ | Archaea | 6 | 6 | $2.5 \times 10^{-5}$ |

Table 3.1: **Association of phylogenetic groups with the most dominant eigenpositions in the 16S and 23S alignments.**

### 3.2.2 Eigenpositions in the 23S rRNA

In the 23S rRNA alignment, the second through fifth most significant eigenpositions in the 23S rRNAs describe relationships among the organisms similar to those in the 16S rRNA, as observed in Figure 3.2. The second most significant eigenposition ((*a*), red) differentiates the Eukarya excluding the Microsporidia from the Bacteria, as indicated by the color bar (Table 3.1(b)). The fourth ((*a*), blue) distinguishes between the Proteobacteria and the Firmicutes. The third ((*b*), red) and fifth ((*b*), blue) eigenpositions describe the similar and the dissimilar among the Archaea and Microsporidia, respectively.

## 3.3 Eigenorganisms identify positions uniquely conserved within phylogenetic groups

We correlate and anticorrelate an eigenposition with increased relative nucleotide frequency across a taxonomic group according to the NCBI Taxonomy Browser annotations [89] (Figures 2.4 and 2.5) of the two groups of $k$ organisms each, with largest and smallest levels of nucleotide frequency in this eigenposition among all $K$ organisms, respectively. The $P$-value of a given association is calculated assuming hypergeometric probability distribution of the $J$ annotations among the $K$ organisms, and of the subset of $j \subseteq J$ annotations among the subset of $k$ organisms, as described by Tavazoie *et al* [102] (Equation 2.6).

The significant enrichments are listed in Tables 3.2 and 3.3, and are discussed in detail, in the context of phylogenetic relationships, in the following sections.

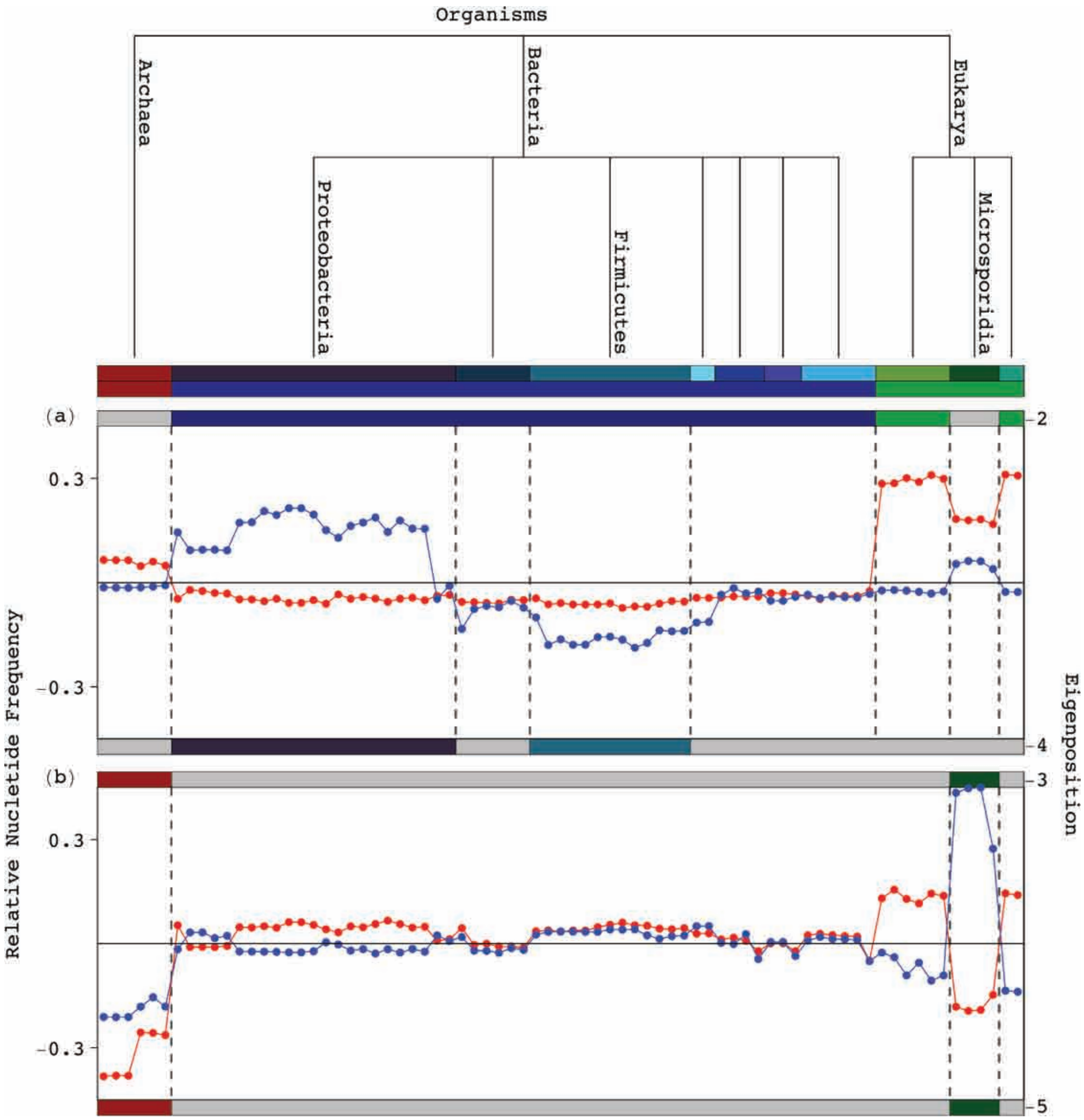

**Fig.** 3.2: **Significant 23S eigenpositions.** Line-joined graphs of the second through fifth 23S eigenpositions, i.e., patterns of nucleotide frequency across the organisms, and their correlation with the taxonomic groups in the 23S alignment, classified according to the top six hierarchical levels of the NCBI Taxonomy Browser [89] (Figure 2.5).

| **Eigenorganism** | **Correlation** | **Nucleotide Segment** | **Structure Motif** | **Conserved in** | $n$ | $N$ | **p-value** |
|---|---|---|---|---|---|---|---|
| 2 | Correlated | Gap | Gap | Eukarya-Microsporidia | 124* | 211 | $4.0 \times 10^{-167}$ |
| | | A | Unpaired A | Eukarya-Microsporidia | 48 | 66 | $2.3 \times 10^{-63}$ |
| | | Gap | Unpaired A | Bacteria | 13* | 50 | $2.1 \times 10^{-8}$ |
| | Anticorrelated | Gap | Gap | Bacteria | 57 | 58 | $9.8 \times 10^{-94}$ |
| | | A | Unpaired A | Bacteria | 50 | 50 | $1.2 \times 10^{-82}$ |
| 3 | Anticorrelated | C | Helix | Archaea+Microsporidia | 76 | 1148 | $4.3 \times 10^{-17}$ |
| | | G | Helix | Archaea+Microsporidia | 68 | 1148 | $1.5 \times 10^{-11}$ |
| | | U | Helix | Archaea+Microsporidia | 65 | 1148 | $8.5 \times 10^{-10}$ |
| | | Gap | Gap | Archaea+Microsporidia | 6 | 6 | $7.3 \times 10^{-10}$ |
| 4 | Correlated | A | Unpaired A | Gamma Proteobacteria | 11 | 11 | $1.4 \times 10^{-17}$ |
| | | Gap | Helix | Actinobacteria+Archaea | 34 | 153 | $8.6 \times 10^{-22}$ |
| 5 | Correlated | C | Helix | Microsporidia | 58 | 947 | $9.6 \times 10^{-10}$ |
| | | U | Helix | Microsporidia | 55 | 947 | $3.6 \times 10^{-8}$ |
| | | Gap | Unpaired A | Archaea | 7 | 14 | $6.1 \times 10^{-8}$ |
| | Anticorrelated | A | Unpaired A | Archaea | 14 | 14 | $2.7 \times 10^{-22}$ |
| | | G | Helix | Archaea | 84 | 933 | $1.8 \times 10^{-31}$ |
| | | C | Helix | Archaea | 85 | 933 | $1.3 \times 10^{-32}$ |
| | | U | Helix | Archaea | 57 | 933 | $1.8 \times 10^{-9}$ |
| 6 | Correlated | A | Unpaired A | Fungi/Metazoa-Microsporidia | 9 | 16 | $1.7 \times 10^{-10}$ |
| | Anticorrelated | A | Unpaired A | Rhodophyta | 25 | 27 | $2.2 \times 10^{-37}$ |
| 7 | Correlated | A | Unpaired A | Alveolata | 25 | 31 | $4.3 \times 10^{-34}$ |
| | Anticorrelated | A | Unpaired A | Fungi/Metazoa-Microsporidia | 10 | 16 | $3.4 \times 10^{-12}$ |

Table 3.2: **Enrichment of structure motifs in the dominant 16S eigenorganisms.** P-values were calculated under the assumption of the hypergeometric distribution, with $k$=100. The variables $k$, $n$, and $N$ are as described in Equation 2.6.

| Eigenorganism | Correlation | Nucleotide Segment | Structure Motif | Conserved in | $n$ | $N$ | p-value |
|---|---|---|---|---|---|---|---|
| 2 | Correlated | Gap | Gap | Eukarya-Microsporidia | 136 | 145 | $2.3 \times 10^{-220}$ |
| | | A | Unpaired A | Eukarya-Microsporidia | 59 | 59 | $1.7 \times 10^{-94}$ |
| | | Gap | Unpaired A | Bacteria | 15 | 41 | $2.9 \times 10^{-13}$ |
| | Anticorrelated | Gap | Gap | Bacteria | 14* | 14 | $9.8 \times 10^{-27}$ |
| | | A | Unpaired A | Bacteria | 41 | 41 | $6.1 \times 10^{-65}$ |
| | | Gap | Unpaired A | Eukarya-Microsporidia | 8* | 59 | $1.1 \times 10^{-6}$ |
| 3 | Correlated | Gap | Gap | Bacteria | 12 | 14 | $3.5 \times 10^{-17}$ |
| | | A | Unpaired A | Bacteria | 28 | 41 | $4.8 \times 10^{-34}$ |
| | Anticorrelated | Gap | Gap | Archaea+Microsporidia | 41* | 45 | $2.2 \times 10^{-74}$ |
| | | A | Unpaired A | Archaea+Microsporidia | 11 | 11 | $1.4 \times 10^{-17}$ |
| | | Gap | Unpaired A | Bacteria | 8* | 41 | $1.3 \times 10^{-7}$ |
| 4 | Correlated | A | Unpaired A | Proteobacteria | 8 | 8 | $5.9 \times 10^{-13}$ |
| | Anticorrelated | A | Unpaired A | Firmicutes | 5 | 5 | $2.4 \times 10^{-8}$ |
| 5 | Correlated | Gap | Gap | Microsporidia | 191 | 387 | $3.3 \times 10^{-245}$ |
| | | A | Unpaired A | Microsporidia | 16 | 31 | $5.1 \times 10^{-17}$ |
| | Anticorrelated | Gap | Gap | Archaea | 15* | 59 | $1.1 \times 10^{-10}$ |
| | | A | Unpaired A | Archaea | 39 | 49 | $6.6 \times 10^{-52}$ |
| | | Gap | Unpaired A | Microsporidia | 9* | 31 | $1.9 \times 10^{-7}$ |

Table 3.3: **Enrichment of structure motifs in the dominant 23S eigenorganisms** P-values were calculated under the assumption of the hypergeometric distribution, with $k$=200 (except for the gap segments of the second, third and fifth eigenorganisms, where the largest nucleotide frequency decrease is shared by $^*m$=91, 100 and 199 positions, respectively). The variables $k$, $n$, and $N$ are as described in Equation 2.6.

The $P$-value of each enrichment is calculated as described [102] assuming, for each nucleotide, hypergeometric distribution of the motifs among the positions.

Exclusive sequence gap conservation is defined as conservation of gaps within at least 80% of the organisms of the corresponding taxonomic group but in less than 20% of the remaining organisms. Exclusive unpaired A nucleotide conservation is defined as conservation of an adenosine within at least 80% of the organisms of the group but in less than 20% of the remaining organisms, together with greater frequency of unpaired nucleotides within the group rather than among the remaining organisms.

### 3.3.1 Naming conventions for figures

In this section, significant positions identified by each eigenorganism are displayed in two ways. First, they are mapped on the secondary structure of the corresponding taxonomic group. The secondary structures shown here were modified from the Comparative RNA Website (www.rna.ccbb.utexas.edu) [16].

Second, the significant positions are displayed as rasters, to visualize the nucleotide variation at these positions across the alignment. The nucleotides are color-coded A (red), C (green), G (blue), U (yellow), unknown (gray) and gap (black). The color bars above the rasters highlight the taxonomic groups that are differentiated by the second 23S eigenposition and eigenorganism, i.e., the Eukarya excluding the Microsporidia and the Bacteria, and correspond to the trees in Figures **??** and 2.5

### 3.3.2 Eigenposition 2 separates the Bacteria and Eukarya

In both alignments, the second most significant eigenposition captures the dissimilarities between the Eukarya excluding the Microsporidia, and the Bacteria. These patterns of relative nucleotide frequency across the organisms correlate with increased frequency across the Eukarya excluding the Microsporidia, and decreased frequency across the Bacteria, with both $P$-values $< 10^{-25}$ and $< 10^{-2}$ in the 16S and 23S alignments, respectively.

In Figure 3.3, the sequence gaps conserved exclusively in the Eukarya and Bacteria, identified by second most significant eigenorganism, are mapped on the secondary structure diagrams of *E. coli* and *S. cerevisiae* respectively. These positions identify gaps exclusively conserved in either the Eukarya excluding the Microsporidia, or the Bacteria (Table 3.2), that map out known as well as previously unrecognized entire substructures deleted or inserted, respectively, in the Eukarya relative to the Bacteria.

The 124 positions with largest increase in relative nucleotide frequency in the gap segment of the second eigenorganism, i.e., the 124 positions of gap variation across the organisms most correlated with the second eigenposition, map out the exclusively conserved substructures in the secondary structure model of the bacterium *E. coli* [16] (Figure 3.3(a)). The substructures I and II were identified by Winker & Woese [110] (Figure 3.4), and the substructures III and IV were previously unrecognized.

Of the 100 positions of gap variation across the organisms most anticor-

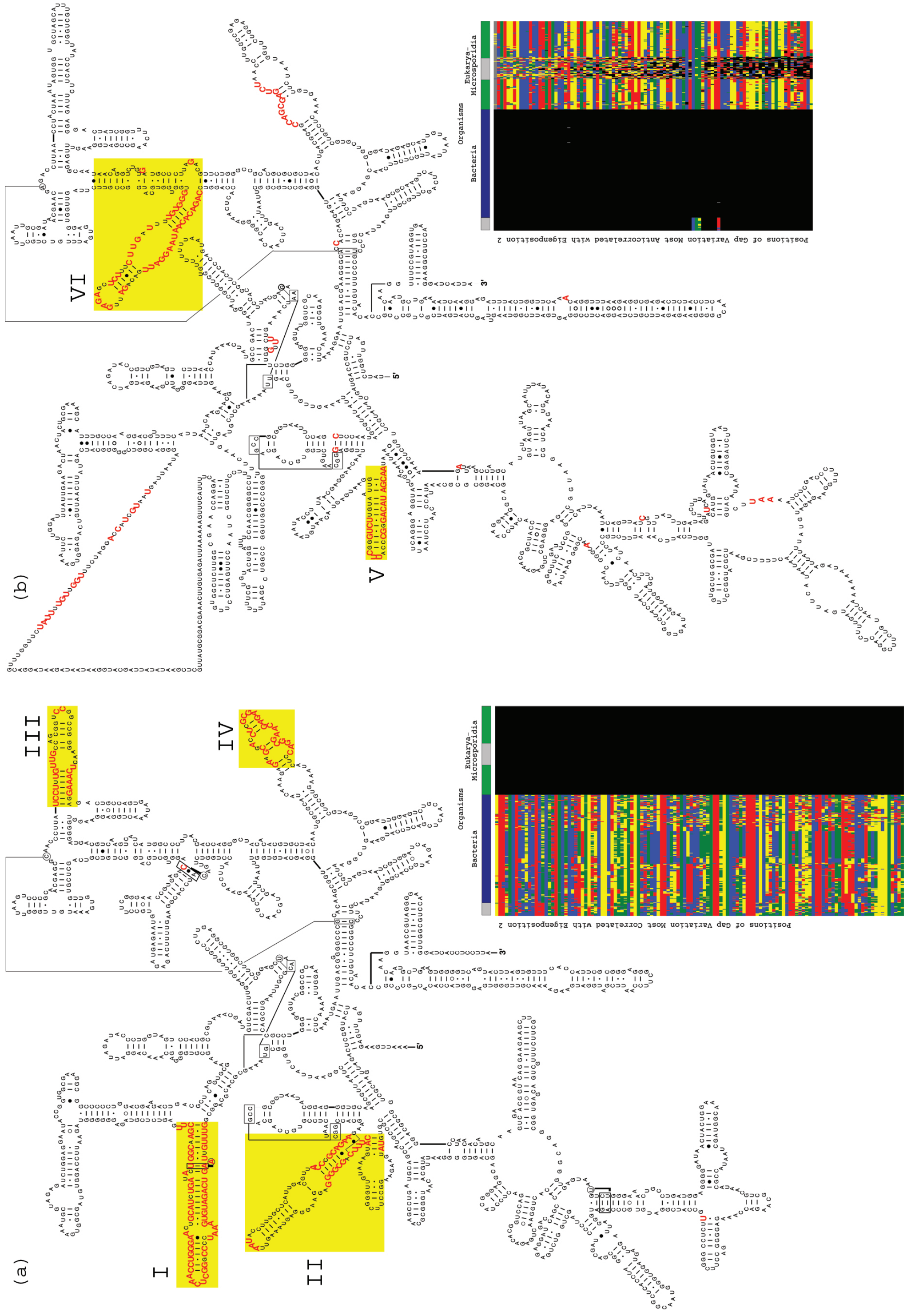
(b)
VI
V
Positions of Gap Variation Most Anticorrelated with Eigenposition 2
Organisms
Bacteria
Eukarya
Microsporidia
3'
5'
(a)
I
II
III
IV
Positions of Gap Variation Most Correlated with Eigenposition 2
Organisms
Bacteria
Eukarya
Microsporidia
3'
5'

**Fig.** 3.3: **Sequence gaps exclusive to Eukarya or Bacteria 16S rRNAs.** (*a*) The 124 positions of gap variation across the organisms most correlated with the second eigenposition, shown on the secondary structure model of *E. coli*, and in raster (inset). (*b*) The 100 positions of gap variation across the organisms most anticorrelated with the second eigenposition, shown on secondary structure model of *S. cerevisiae*, and in raster (inset).

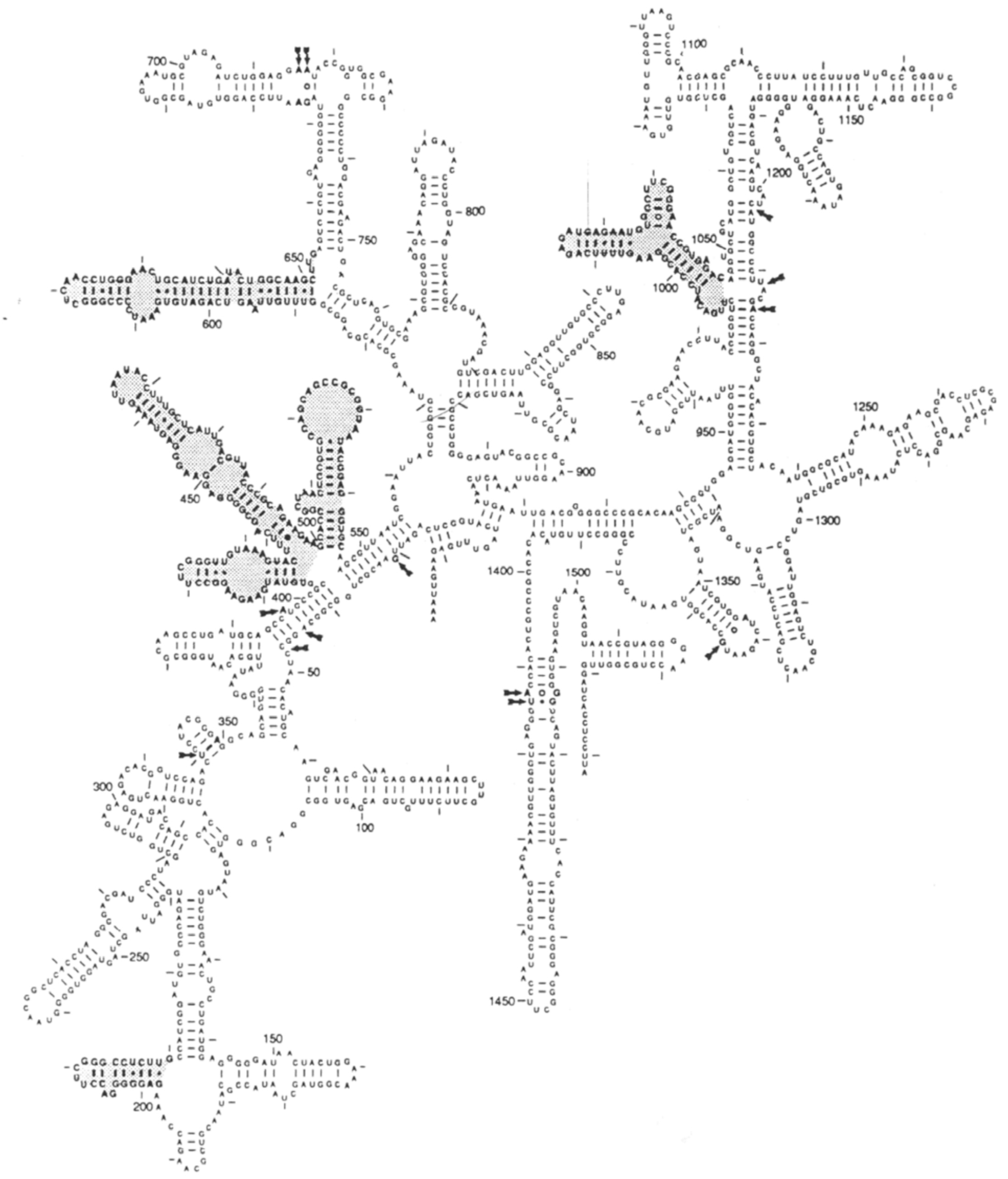

**Fig.** 3.4: **Non-homologous features that distinguish the three domains, represented on the secondary structure of** *E. coli* **(Reproduced from Winker and Woese [110]).** The regions characteristic to the Bacteria are shaded.

related with the second eigenposition, 99 map out the substructures V and VI in the secondary structure model of the eukaryote *S. cerevisiae* (Figure 3.3(b)). The 100th position is an unknown nucleotide at the 3'-end of the molecule, which is not displayed. These 100 positions are also displayed in the inset raster.

### 3.3.3 Other significant eigenpositions

The fourth 16S eigenposition correlates with increased nucleotide frequency across the Gamma Proteobacteria and decreased frequency across the Actinobacteria and Archaea, with both $P$-values $< 10^{-23}$. The Gamma Proteobacteria and the Actinobacteria are the two largest bacterial groups in this alignment. The fourth 23S eigenposition captures the dissimilar between the Proteobacteria and the Firmicutes, the two largest bacterial groups in this alignment.

In both alignments, the third and fifth eigenpositions capture the similarities and dissimilarities between the Archaea and Microsporidia, respectively. In the 16S alignment, the sixth and seventh eigenpositions identify dissimilarities among the Fungi/Metazoa excluding the Microsporidia and the Rhodophyta and separately the Alveolata, respectively.

## 3.4 Eigenorganisms identify characteristic sites

Consistent with the eigenpositions, the corresponding 16S and 23S eigenorganisms identify positions of nucleotides that are approximately conserved within the respective taxonomic groups, but not among them. These positions are significantly enriched in conserved sequence gaps which map out entire substructures inserted or deleted in the 16S and 23S rRNAs of one taxonomic group relative to another as well as adenosines that are unpaired in the rRNA secondary structure and are conserved exclusively in the respective taxonomic groups. The majority of these adenosines participate in tertiary structure interactions, and some also map to the same substructures. We consider the $m$ positions with largest increase or decrease in the relative nucleotide frequency in each nucleotide-specific segment of each eigenorganism (Table 3.2 and 3.3 ).

These positions exhibit the frequency variations across the organisms that are most correlated or anticorrelated, respectively, with the corresponding eigenposition. We calculate the $P$-value of the enrichment of these positions in sequence and structure motifs conserved across the corresponding taxonomic groups by assuming hypergeometric probability distribution of the $N$ conserved motifs among the $M$ positions, and of the subset of $n \subseteq N$ motifs among the subset of $m$ positions, as described [102], $P(n;m,M,N) = \binom{M}{m}^{-1} \sum_{i=n}^{m} \binom{N}{i}\binom{M-N}{m-i}$.

### 3.4.1 Sites are insertions/deletions of structure motifs

The positions identified by the eigenorganisms include entire substructures inserted or deleted in the structure of one taxonomic group relative to another. Consider for example the 124 positions with largest nucleotide frequency increase in the gap segment of the second most significant 16S eigenorganism, i.e., the positions for which the frequency of gaps across the organisms is most correlated with the second eigenposition. These positions are enriched in sequence gaps conserved in the Eukarya excluding the Microsporidia (Figure 3.5(a)). These include 13 of the 50 positions with unpaired A nucleotides exclusively conserved in the Bacteria (Figure 3.8). The 100 positions with largest frequency decrease are enriched in gaps conserved in the Bacteria (Figure 3.5(b)). These include 8 of the 66 positions with unpaired A nucleotides exclusively conserved in the Eukarya (Figure 3.9). Both $P$-values $< 10^{-93}$.

Mapped onto the secondary structure models of the bacterium *E. coli* and the eukaryote *S. cerevisiae* [16], these positions map out known as well as previously unrecognized insertions and deletions of not only isolated nucleotides but entire substructures in the Eukarya with respect to the Bacteria [110] (Figure 3.3).

Similarly, the positions identified by the gap segment of the second 23S eigenorganism map out entire substructures inserted and deleted in 23S rRNAs of the Eukarya excluding the Microsporidia relative to the Bacteria.

In Figure 3.6, the positions of gap variation most correlated and anticorrelated with the second eigenposition are marked on the secondary structure models of the bacterium *E. coli* and the eukaryote *S. cerevisiae* respectively. The 200 positions of gap variation across the organisms most correlated with the second eigenposition (green), map out entire substructures in the secondary structure model of the bacterium *E. coli* (Figure 3.6(a), yellow). The 200 positions with largest frequency decrease in the A nucleotide segment of the same eigenorganism, identify all 41 unpaired A nucleotides that are exclusively conserved in the Bacteria (red). Of these, 15 correspond to gaps conserved in the Eukarya excluding the Microsporidia. The 91 positions of gap variation across the organisms most anticorrelated with the second eigenposition (green) map out entire substructures in the secondary structure

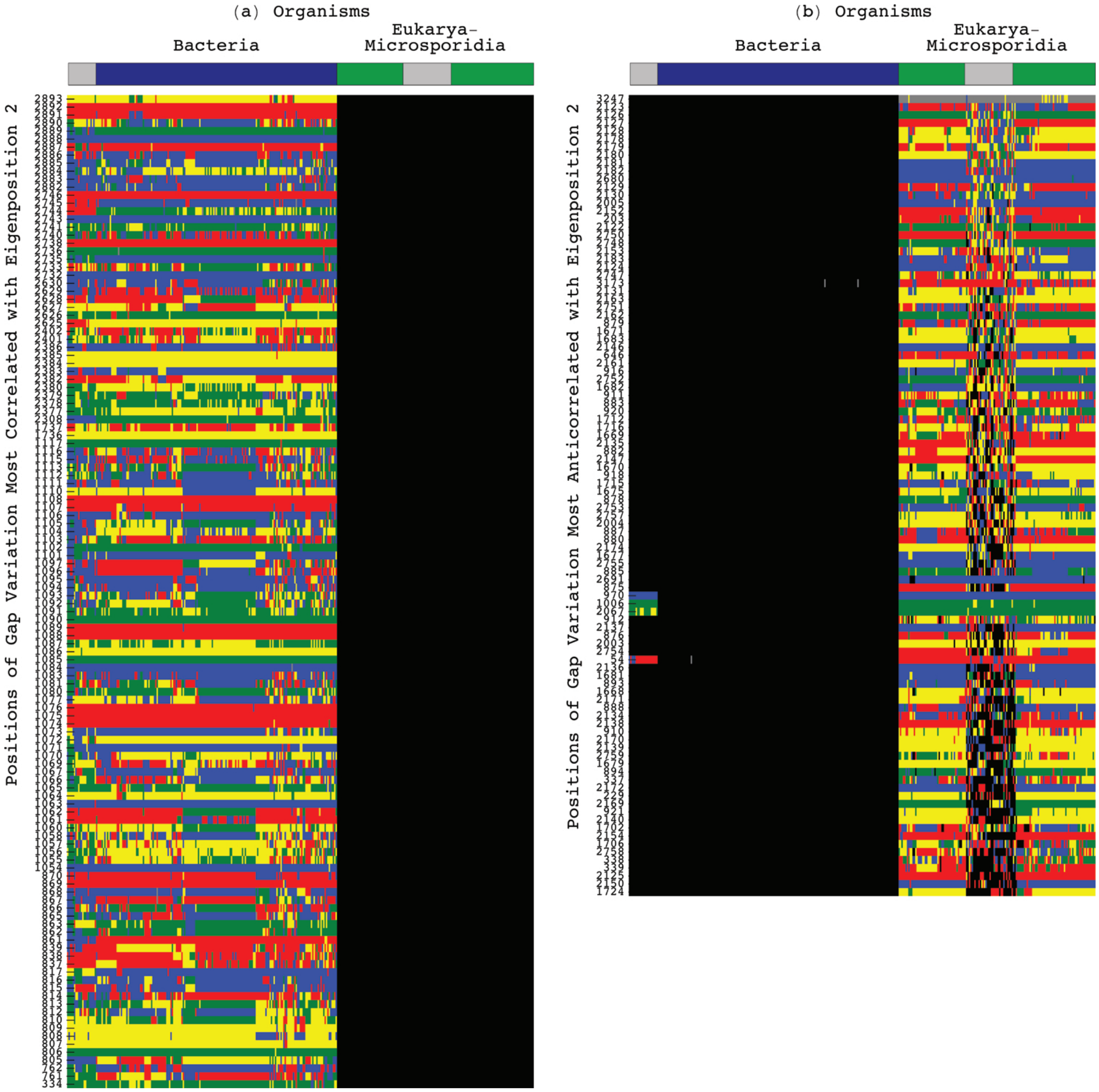

**Fig.** 3.5: **Sequence gaps exclusive to Eukarya or Bacteria 16S rRNAs** Raster displays of the positions in the alignment for which the gap frequency variation is most correlated or anticorrelated with the second eigenposition (Figure 3.3), as identified by the gap segment of the second eigenorganism. (*a*) The 124 correlated positions display gaps exclusively conserved in the Eukarya. (*b*) The 100 anticorrelated positions display gaps exclusively conserved in the Bacteria.

model of the eukaryote *S. cerevisiae* (Figure 3.6(b), yellow). The 200 positions with largest frequency increase in the A nucleotide segment of the same eigenorganism, identify all 59 unpaired A nucleotides that are exclusively conserved in the Eukarya excluding the Microsporidia (red). Of these, eight correspond to gaps conserved in the Bacteria.

Figure 3.7 shows the raster displays of these 200 and 91 positions in the 23S alignment for which the gap frequency variation is most correlated or anticorrelated, respectively, with the second 23S eigenposition (Figure 3.2), as identified by the gap segment of the second eigenorganism (Table 3.3). The 200 correlated positions display gaps exclusively conserved in the Eukarya, plotted on the secondary structure model of the eukaryote *S. cerevisiae*.

### 3.4.2 Sites are structure motifs: Unpaired adenosines

The eigenorganisms identify adenosines, unpaired in the rRNA secondary structure, which are conserved exclusively in the respective taxonomic groups, most of which participate in tertiary structure interactions and map to the substructures inserted or deleted within taxonomic groups.

We find the positions with largest nucleotide frequency increase in the A segment of the second 16S eigenorganism to be enriched in unpaired adenosines, which are exclusively conserved in the Eukarya excluding the Microsporidia (Figures 3.2 and 3.5). The positions with largest decrease in relative nucleotide frequency include all 50 unpaired adenosines exclusively conserved in the Bacteria ($P$-values $< 10^{-62}$, Table 3.1).

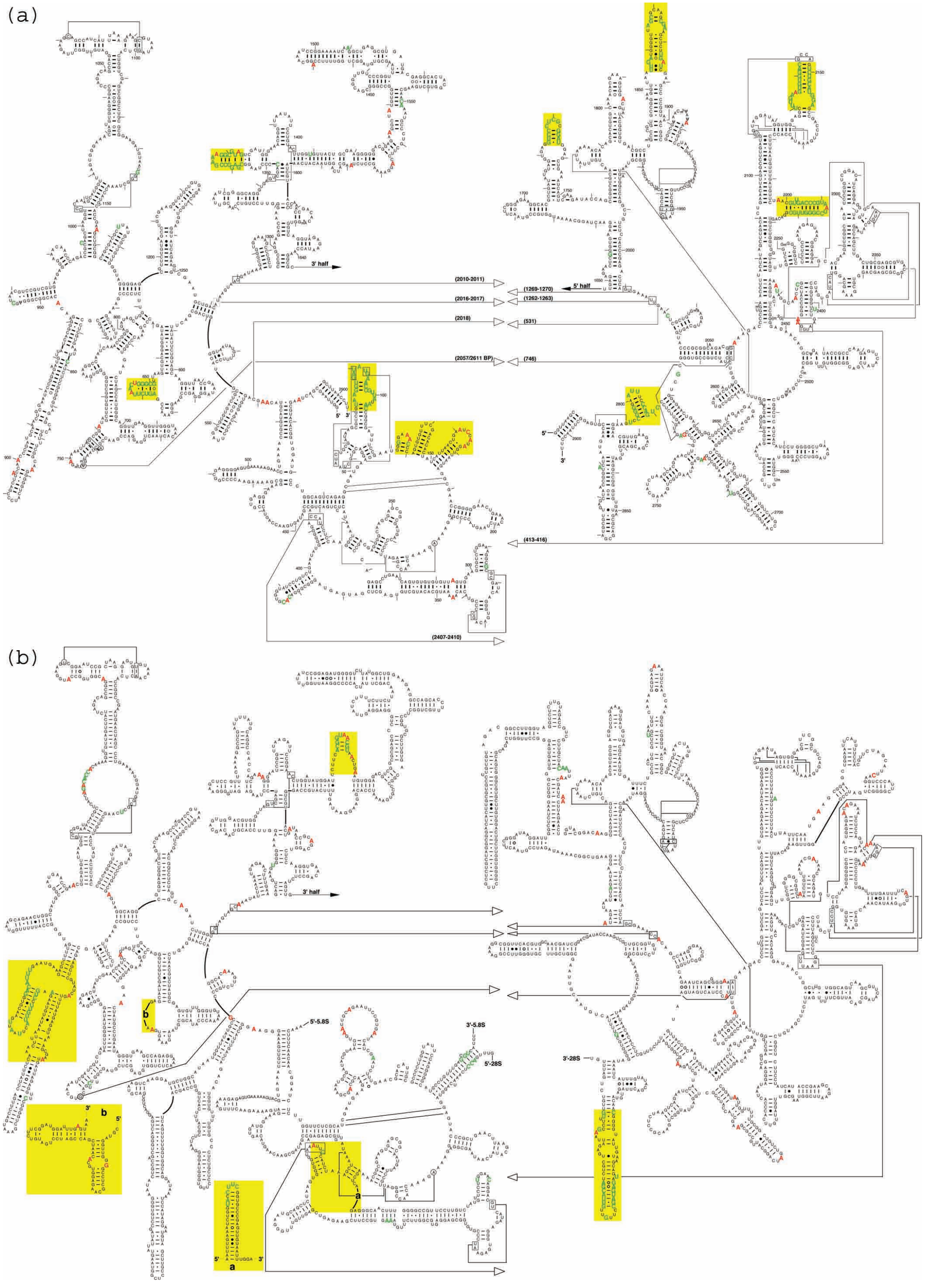

**Fig.** 3.6: **Sequence gaps and unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs.**

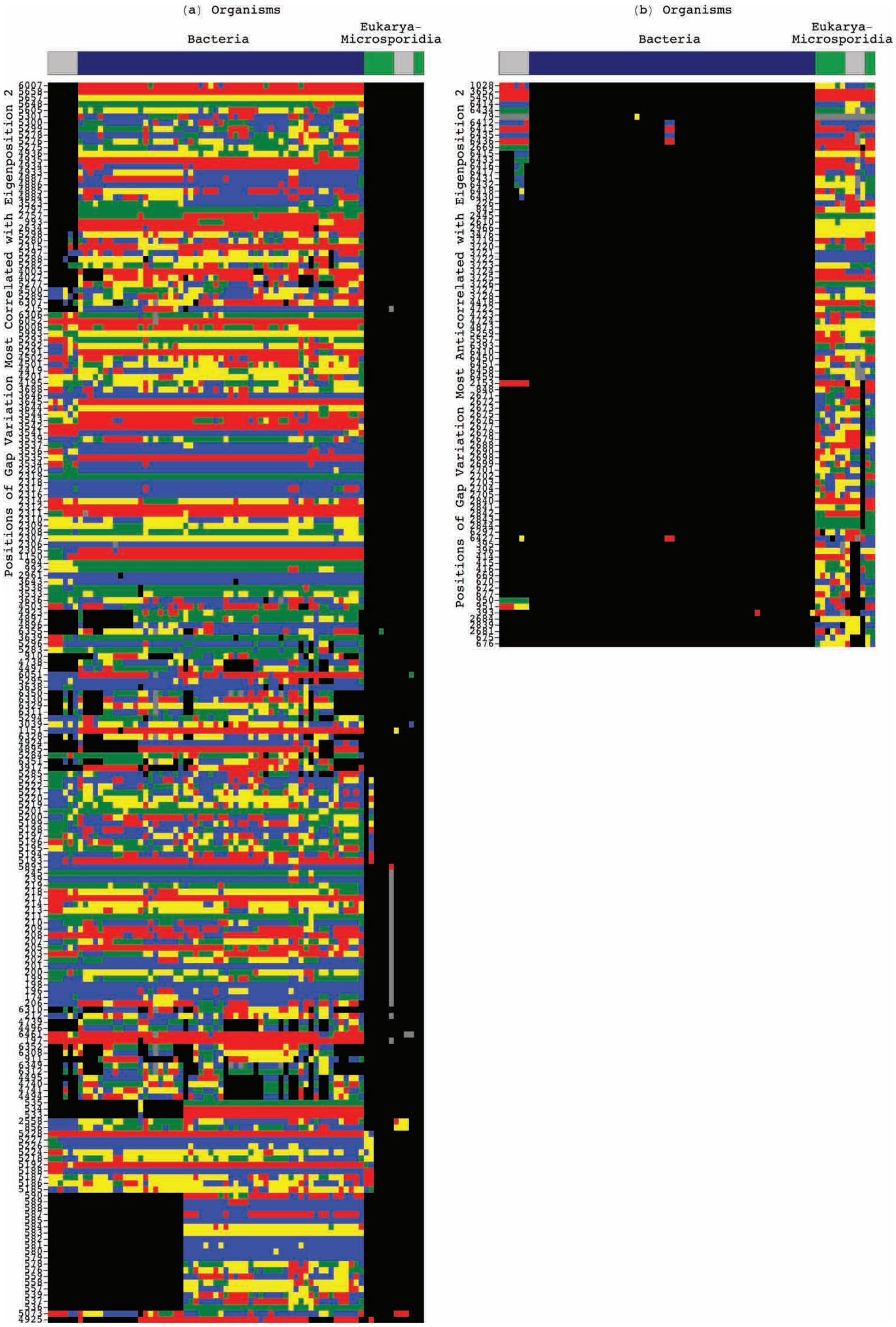

**Fig.** 3.7: **Sequence gaps exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs.** (*a*) The 200 correlated positions display gaps exclusively conserved in the Eukarya. (*b*) The 91 anticorrelated positions display gaps exclusively conserved in the Bacteria.

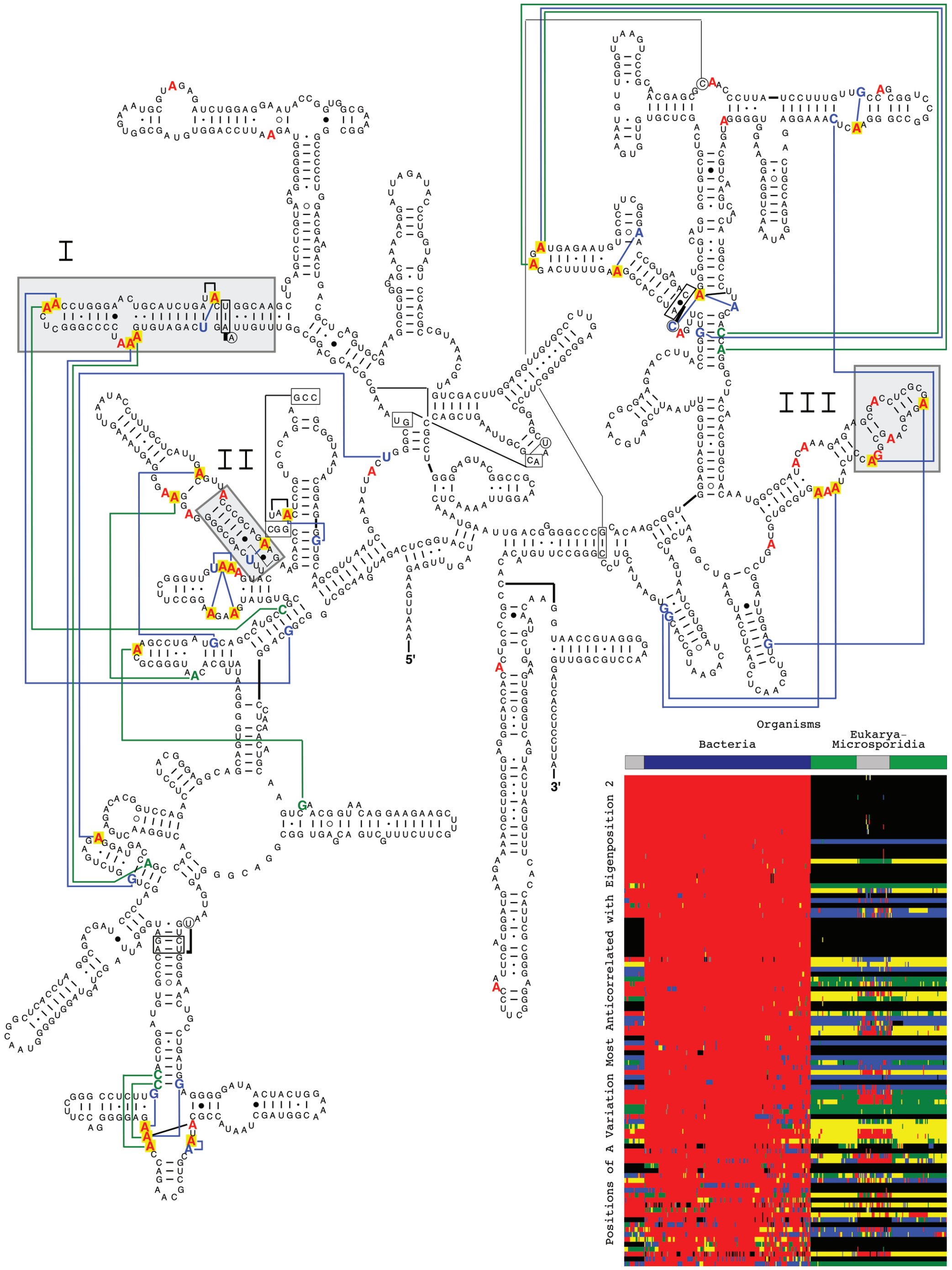
I
II
III
5'
3'
Organisms
Bacteria
Eukarya-Microsporidia
Positions of A Variation Most Anticorrelated with Eigenposition 2

**Fig.** 3.8: **Unpaired adenosines exclusive to Bacteria 16S rRNAs.** The 100 positions with largest decrease in relative A nucleotide frequency in the second eigenorganism are mapped on the *E. coli* secondary structure [70], and displayed in raster (inset). The blue and green lines indicate known tertiary base-base and base-backbone interactions respectively, from the crystal structure of *T. thermophilus*.

In Figure 3.8, the 100 positions identified in the A nucleotide segment of the second eigenorganism with the largest decrease in relative nucleotide frequency include all 50 positions (red) in the alignment with unpaired A nucleotides exclusively conserved in the Bacteria. Of these 50 positions, 28 (yellow) map to known tertiary interactions in the crystal structure of the bacterium *T. thermophilus*, plotted on the secondary structure model of the bacterium *E. coli* [16]. These include 22 base-base interactions (blue) and eight base-backbone interactions (green). These interactions represent a significant enrichment among all tertiary interactions in the 16S rRNA crystal structure of the bacterium *T. thermophilus* (Table 3.4).

Of the 50 positions of unpaired A nucleotides exclusively conserved in the Bacteria, 13 correspond to gaps conserved exclusively in the Eukarya excluding the Microsporidia ($P$-value $< 10^{-7}$). These 13 positions map to the entire 16S rRNA substructures that are deleted in the Eukarya with respect to the Bacteria (gray), identified by the gap segment of the second eigenorganism.

The 100 most anticorrelated A positions are also displayed in raster form in Figure 3.10(b). The color bars highlight the Bacteria.

Similarly, in Figure 3.9, the 100 positions identified in the A nucleotide segment of the second eigenorganism with the largest increase in relative nucleotide frequency include 48 of the 66 positions (red) in the alignment with unpaired A nucleotides conserved exclusively in the Eukarya. Eight of these 48 positions correspond to gaps conserved exclusively in the Bacteria, and map to the entire 16S rRNA substructures that are inserted in the Eukarya with respect to the Bacteria,

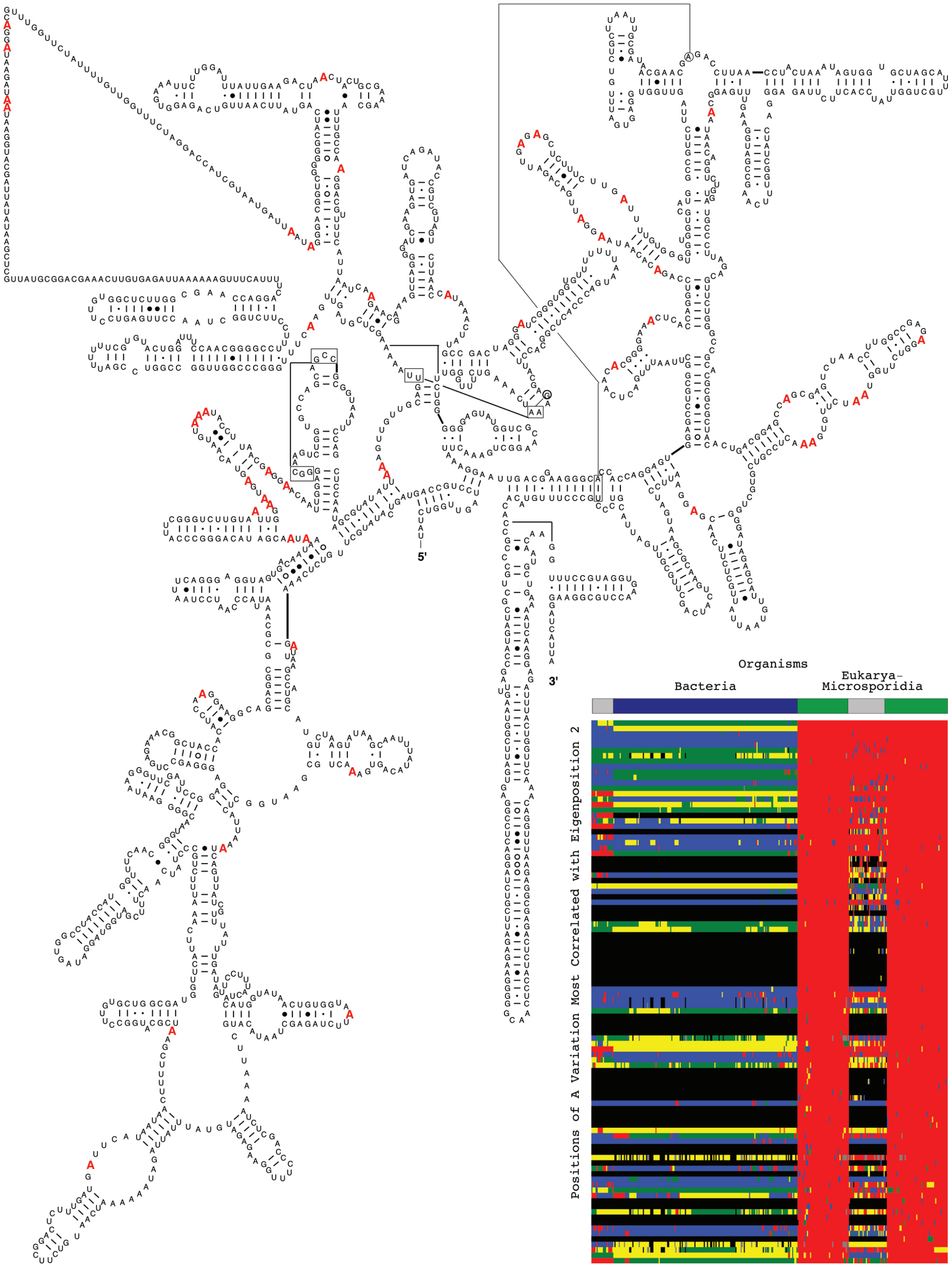
5'
3'
Organisms
Bacteria
Eukarya-
Microsporidia
Positions of A Variation Most Correlated with Eigenposition 2

**Fig.** 3.9: **Unpaired adenosines exclusive to Eukarya excluding Microsporidia 16S rRNAs.** The 100 positions identified in the A nucleotide segment of the second eigenorganism with the largest increase in relative nucleotide frequency plotted on the secondary structure model of *S. cerevisiae* and displayed in raster (inset).

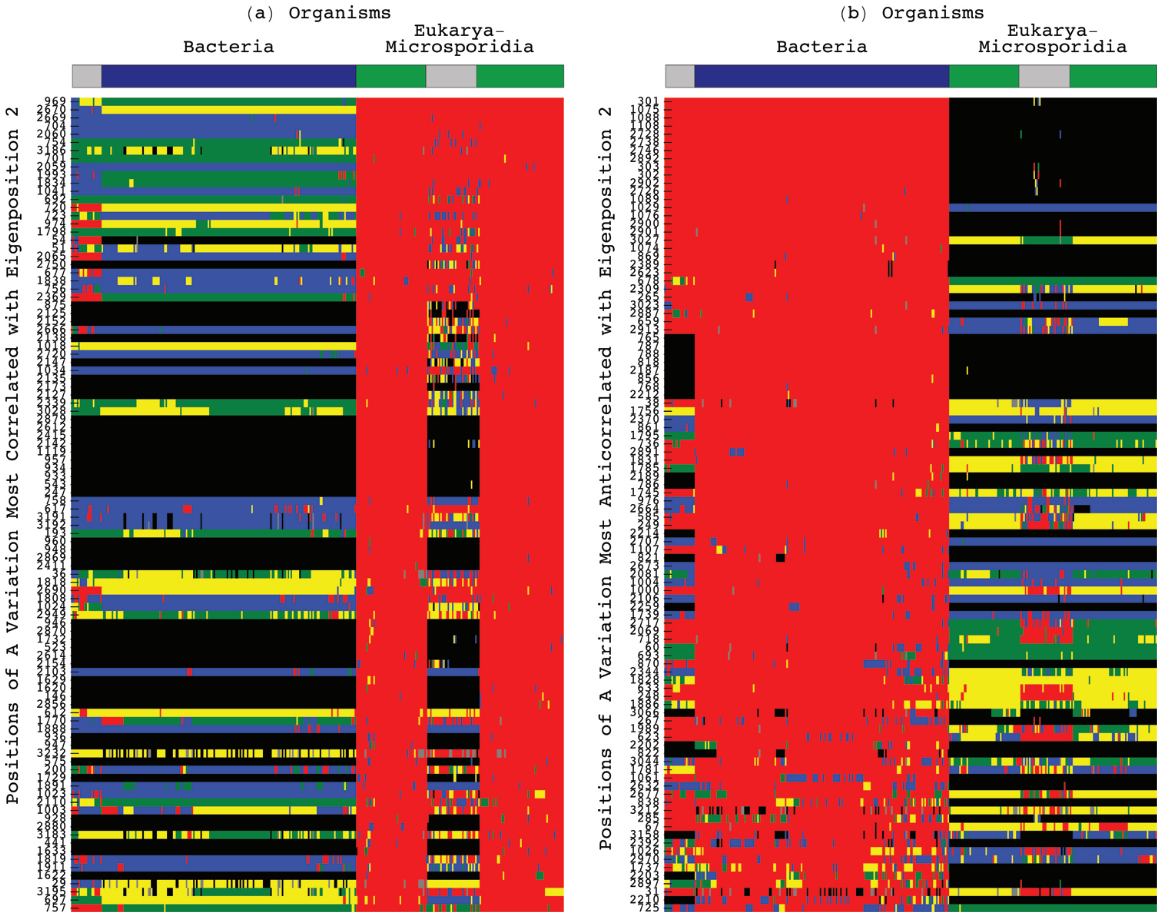

**Fig.** 3.10: **Unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 16S rRNAs.** Raster displays of the 100 positions in the alignment for which the A nucleotide frequency variation is most correlated or anticorrelated with the second eigenposition, as identified by the A segment of the second eigenorganism. (*a*) The 100 correlated positions include 48 of the 66 unpaired A nucleotides exclusively conserved in the Eukarya excluding the Microsporidia (Figure 3.9). (*b*) The 100 anticorrelated positions include all 50 unpaired A nucleotides exclusively conserved in the Bacteria (Figure 3.8).

| **Tertiary Interaction** | $N$ | $n$ | **p-value** |
|---|---|---|---|
| Unpaired A Backbone | 25 | 9 | $2.3 \times 10^{-8}$ |
| Unpaired A Base-base | 41 | 14 | $4.8 \times 10^{-12}$ |
| Paired A Backbone | 28 | 7 | $1.5 \times 10^{-5}$ |
| Paired A Base-base | 48 | 18 | $4.4 \times 10^{-16}$ |
| Nucleotides involved in at least one tertiary interaction | 303 | 44 | $1.1 \times 10^{-20}$ |

Table 3.4: **Enrichment of tertiary interactions in the 100 nucleotides in the A segment most negatively correlated with the second eigenorganism.**
These include the 50 unpaired A's annotated as conserved exclusively in the Bacteria. P-values were calculated under the assumption of the hypergeometric distribution, with $k$=100. The variables $k$, $n$, and $N$ are as described in Equation 2.6

identified by the gap segment of the second eigenorganism (Figure 3.3). A raster of these same 100 positions can be seen in Figure 3.10($a$). The color bars highlight the Eukarya excluding the Microsporidia.

In the 23S rRNAs, the second most significant eigenorganism identifies gaps exclusively conserved in either the Eukarya excluding the Microsporidia or the Bacteria (Table 3.3) that map out entire substructures deleted or inserted, respectively, in the Bacteria relative to the Eukarya. The same eigenorganism also identifies unpaired adenosines, exclusively conserved in either the Eukarya excluding the Microsporidia or the Bacteria, some of which map to the same substructures. The 200 positions with largest frequency decrease in the A nucleotide segment of the same eigenorganism identify all 41 unpaired A nucleotides that are exclusively conserved in the Bacteria (Figure 3.11($b$)). Of these, 15 correspond to gaps conserved in the Eukarya excluding the Microsporidia. The 200 positions with largest frequency increase in the A nucleotide segment of the same eigenorganism,

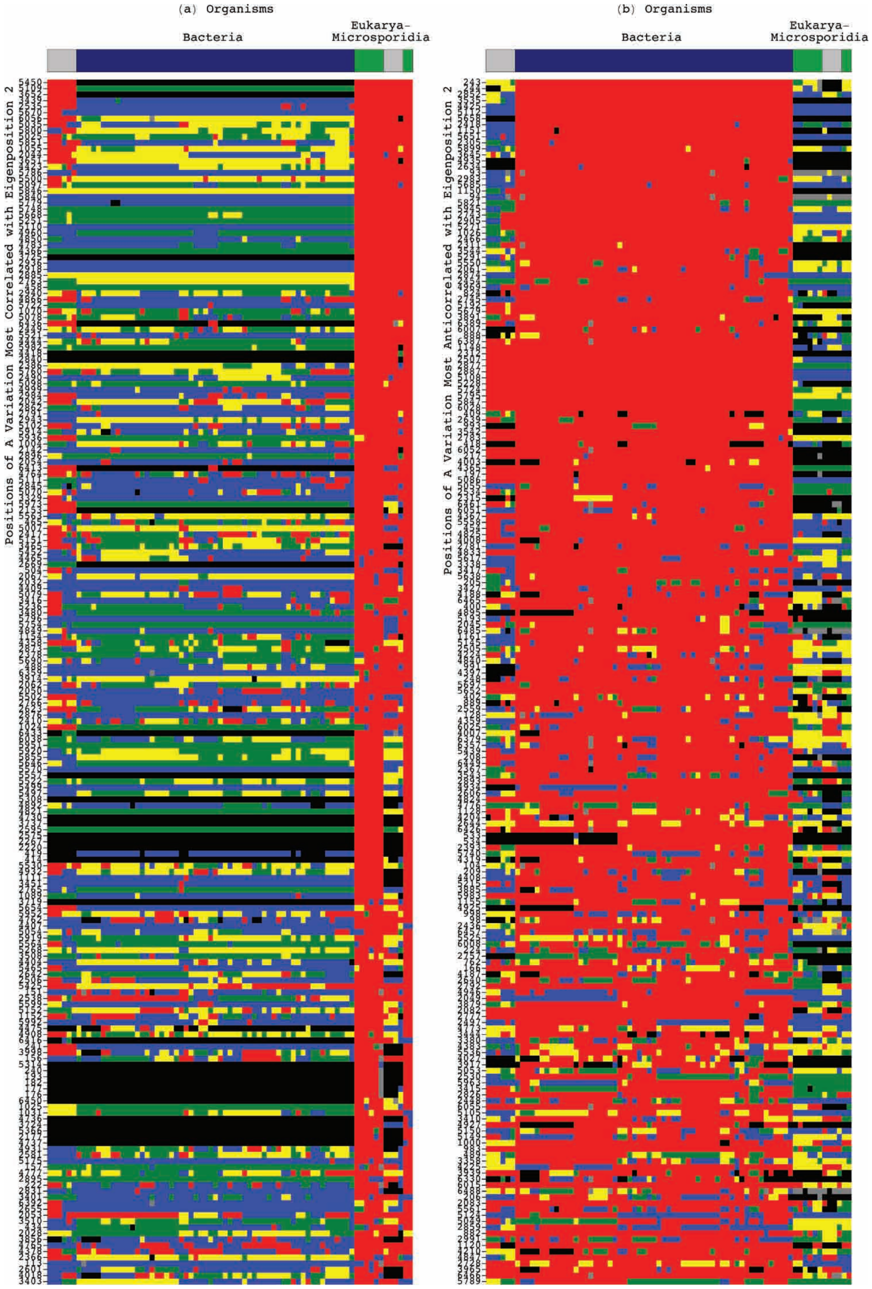

**Fig.** 3.11: **Unpaired adenosines exclusive to Eukarya excluding Microsporidia or Bacteria 23S rRNAs.** Raster displays of the 200 positions in the 23S alignment for which the A nucleotide frequency variation is most (*a*) correlated or (*b*) anticorrelated with the second eigenposition, as identified by the A segment of the second eigenorganism (Figure 3.6).

identify all 59 unpaired A nucleotides that are exclusively conserved in the Eukarya excluding the Microsporidia (Figure 3.11($a$))). Of these, eight correspond to gaps conserved in the Bacteria.

In addition, the 200 correlated gap positions exclusively conserved in the Eukarya include 15 of the 41 positions with unpaired A nucleotides exclusively conserved in the Bacteria (Figure 3.7(a)) . The 91 anticorrelated gap positions exclusively conserved in the Bacteria include eight of the 59 positions with unpaired A nucleotides exclusively conserved in the Eukarya excluding the Microsporidia (Figure 3.7(b)).

We find a similar enrichment of unpaired A nucleotides exclusively conserved in the taxonomic groups identified by the fourth through seventh 16S eigenpositions and by the third through fifth 23S eigenpositions. In the 16S, the 100 positions with largest frequency increase or decrease in the A nucleotide segment of the fourth, fifth, sixth, and seventh eigenorganism, i.e., the positions for which the A nucleotide frequency across the organisms is most correlated or anticorrelated, respectively, with the fourth, fifth, sixth or seventh eigenposition, include all or most of the unpaired A nucleotides exclusively conserved in either the Gamma Proteobacteria, Archaea, Rhodophyta, Alveolata or Fungi/Metazoa excluding the Microsporidia, with all $P$-values $< 10^{-9}$ (Table 3.2). In the 12S, the 200 positions with largest frequency increase or decrease in the A nucleotide segment of the third, fourth or fifth eigenorganism include all or most of the unpaired A nucleotides exclusively conserved in either the Proteobacteria, Firmicutes, Archaea or Microsporidia, with all $P$-values $< 10^{-8}$ (Table 3.3).

## 3.5 Eigenpositions identify multiple pathways of evolution

We find twopreviously unknown relationships between the Archaea and Microsporidia: the third eigenposition captures the similarities between these groups, while the fifth eigenposition captures the dissimilarities.

### 3.5.1 Eigenposition 3 shows that Archaea are similar to Microsporidia

In both 16S and 23S alignments, the third most significant eigenposition captures the similarities among the Archaea and the Microsporidia, and correlates with decreased nucleotide frequency across both the Archaea and Microsporidia relative to all other organisms with the $P$-values $< 10^{-23}$ and $10^{-9}$, respectively. The 100 positions with largest nucleotide frequency decrease in the gap segment of the third 16S eigenorganism identify all six gaps exclusively conserved in both the Archaea and Microsporidia with the corresponding $P$-value $< 10^{-9}$. Mapped onto the secondary structure model of the bacterium *E. coli*, these 100 positions identify deletions of not only isolated nucleotides but entire substructures in the Archaea and Microsporidia with respect to the Bacteria (Figure 3.12(a), substructures I–III).

In Figure 3.12(c), the same 100 positions from (b) are displayed across an alignment of 858 mitochondrial 16S rRNA sequences. These positions show that the gaps are conserved in most Metazoan mitochondria. The other groups of Eukarya represented in the mitochondrial alignment are Alveolata (1), Euglenozoa (2), Fungi (3) and Rhodophyta and Viridiplantae (4).

The 100 positions with the largest nucleotide frequency decrease in the C, G, and U nucelotide segments (Figure 3.13) are enriched in helices, i.e., base-paired nucleotides, exclusively conserved in both the Archaea and Microsporidia, with the $P$-values $< 10^{-9}$.

Similar to the results observed in the 16S rRNA, the 100 positions with largest nucleotide frequency decrease in the gap segment of the third 23S eigenorganism identify 41 of the 45 gaps that are exclusively conserved in both the Archaea and Microsporidia (Figure 3.14). The 200 correlated positions identified in the A segment include 28 of 41 unpaired A nucleotides exclusively conserved in the Bacteria, while the 200 anticorrelated positions include all 11 unpaired A nucleotides exclusively conserved in the Archaea and Microsporidia (Figure 3.15). All three $P$-values $< 10^{-16}$.

### 3.5.2 Eigenposition 5 shows that Archaea are dissimilar to Microsporidia

The fifth 16S and 23S eigenpositions both capture the dissimilarities between Archaea and Microsporidia and correlate with increased and decreased frequency across the Microsporidia and the Archaea, with the $P$-values $< 10^{-14}$ and $10^{-2}$, respectively.

In the gap segment of the 16S fifth eigenorganism, the 100 positions with largest nucleotide frequency increase include seven of the 14 unpaired A nucleotides exclusively conserved in the Archaea, implying that these seven unpaired adenosines are exclusively missing in the Microsporidia (Figure 3.16(c)). The 100 positions with largest nucleotide frequency increase in the C and U segments of the fifth eigenorganism are enriched in helices exclusively conserved in the Microsporidia (Figure 3.16 (a,b)).

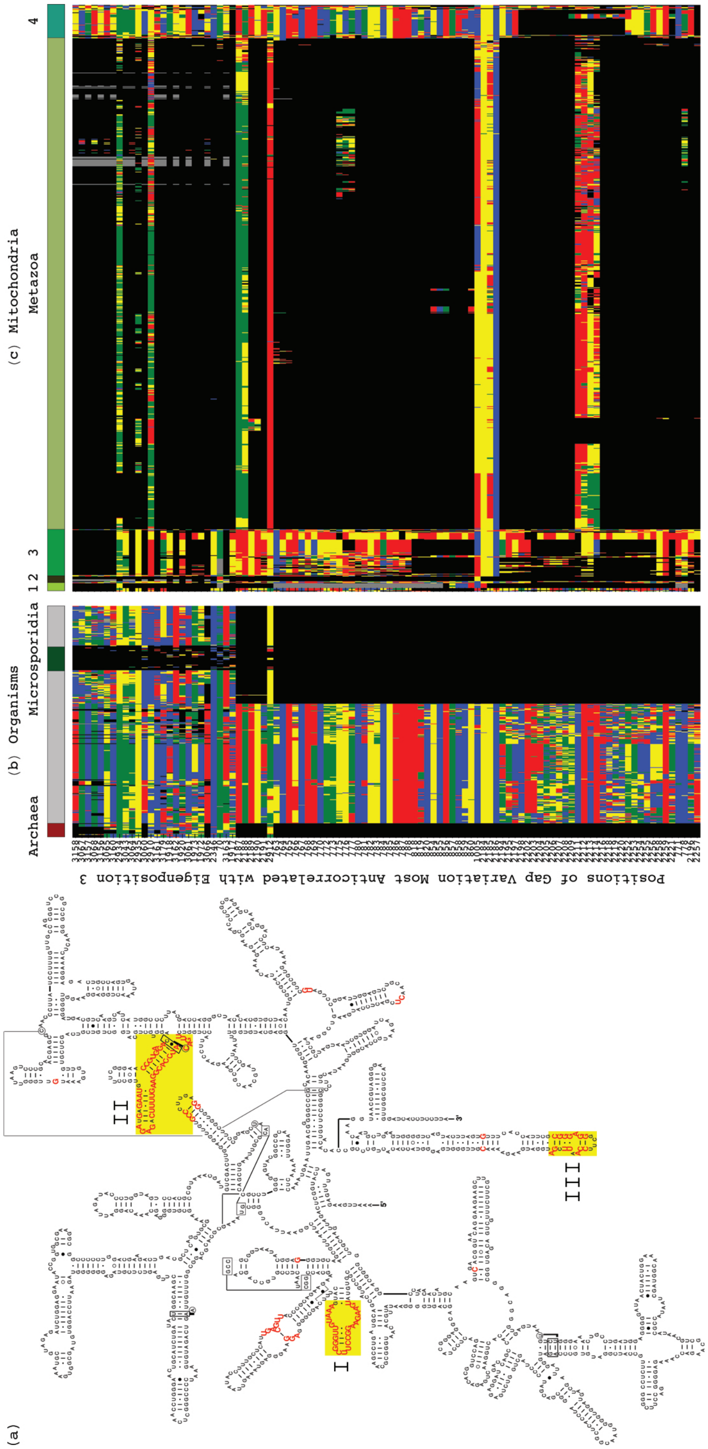

**Fig.** 3.12: **Sequence gaps exclusive to both Archaea and Microsporidia 16S rRNAs.**

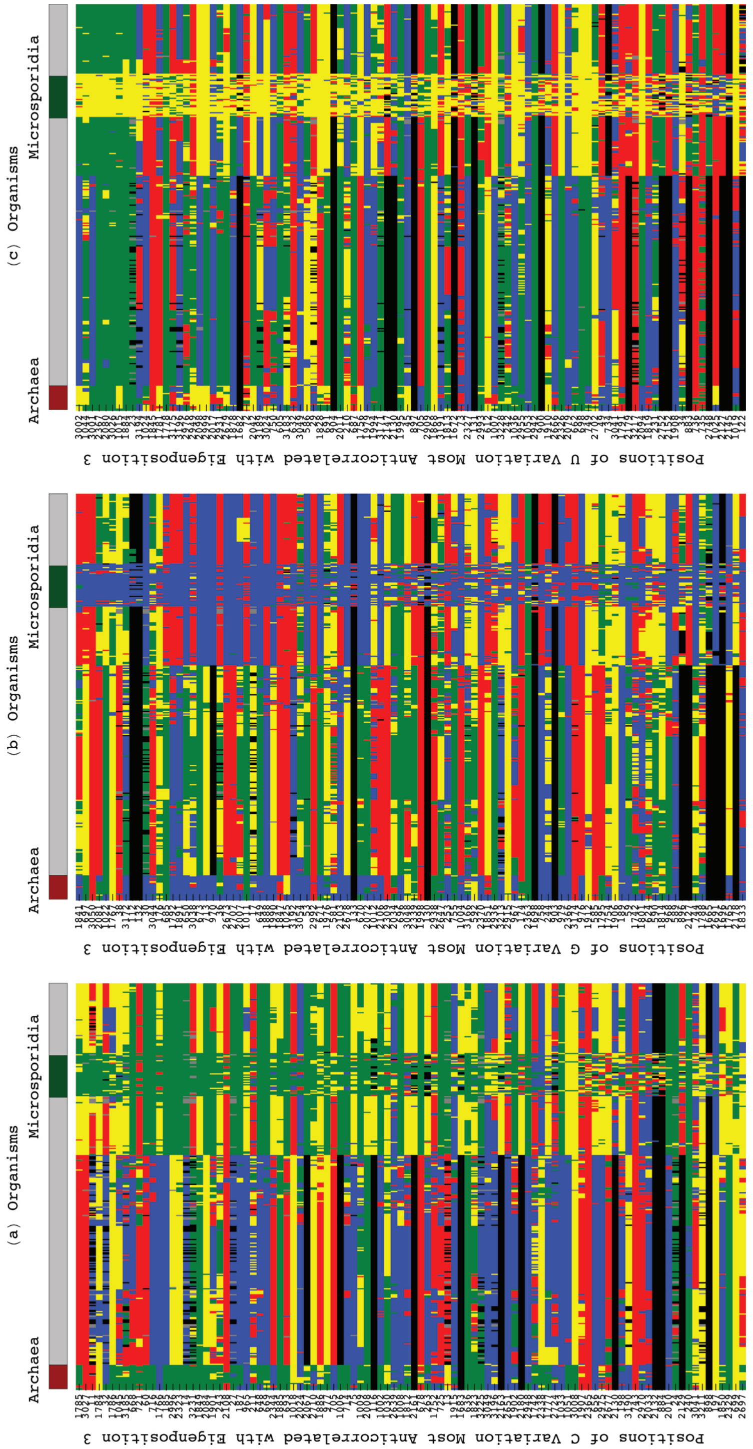

**Fig.** 3.13: **Other nucleotides exclusive to Archaea and Microsporidia 16S rRNAs.** Raster displays of the 100 positions each, identified in the (*a*) C, (*b*) G and (*c*) U nucelotide segments of the third eigenorganism with the largest decrease in relative nucleotide frequency.

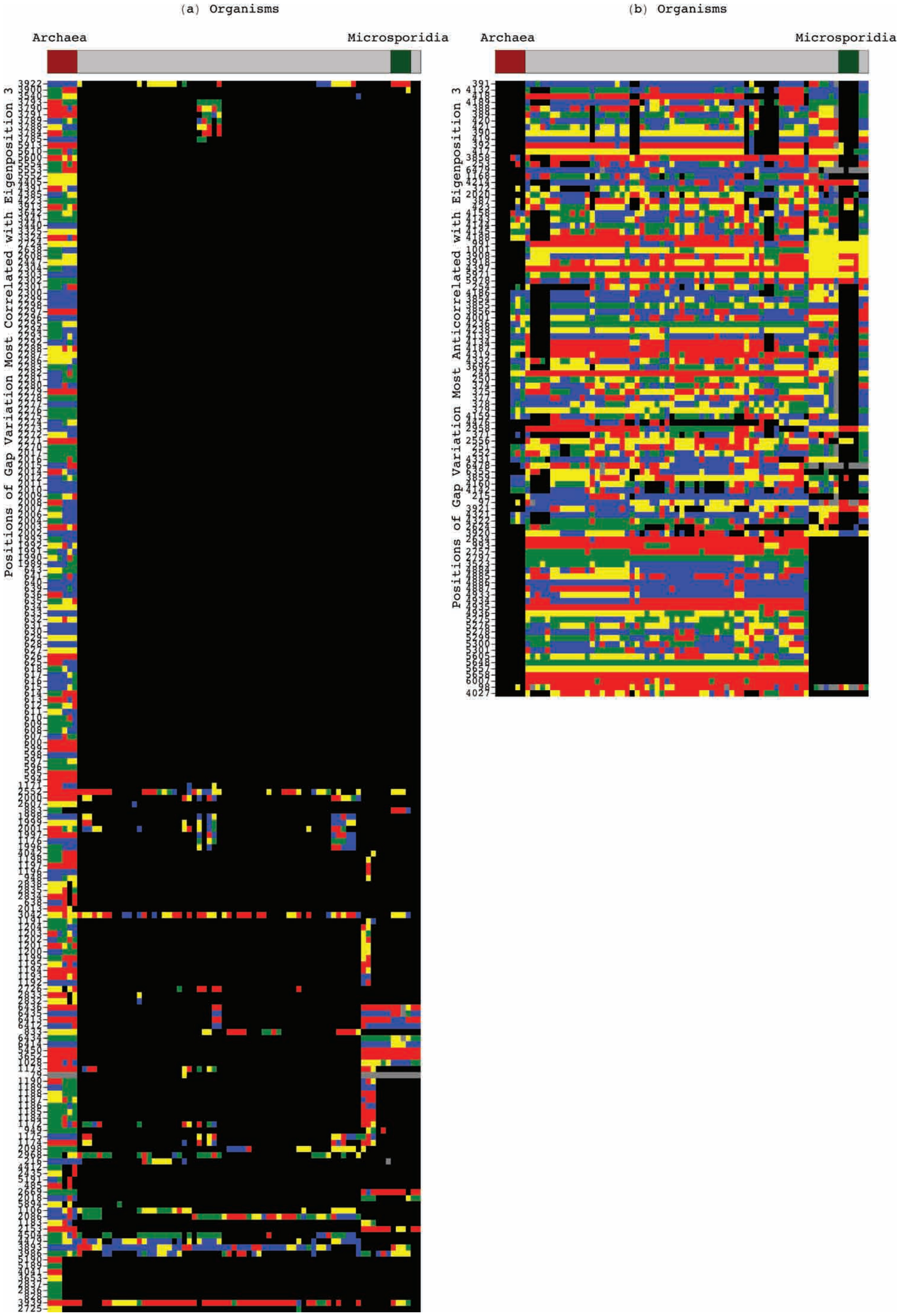

**Fig.** 3.14: **Sequence gaps exclusive to Bacteria or Archaea and Microsporidia 23S rRNAs.** Raster displays of the 200 and 100 positions in the 23S alignment for which the gap frequency variation is most (*a*) correlated or (*b*) anticorrelated, respectively, with the third 23S eigenposition, as identified by the gap segment of the third eigenorganism.

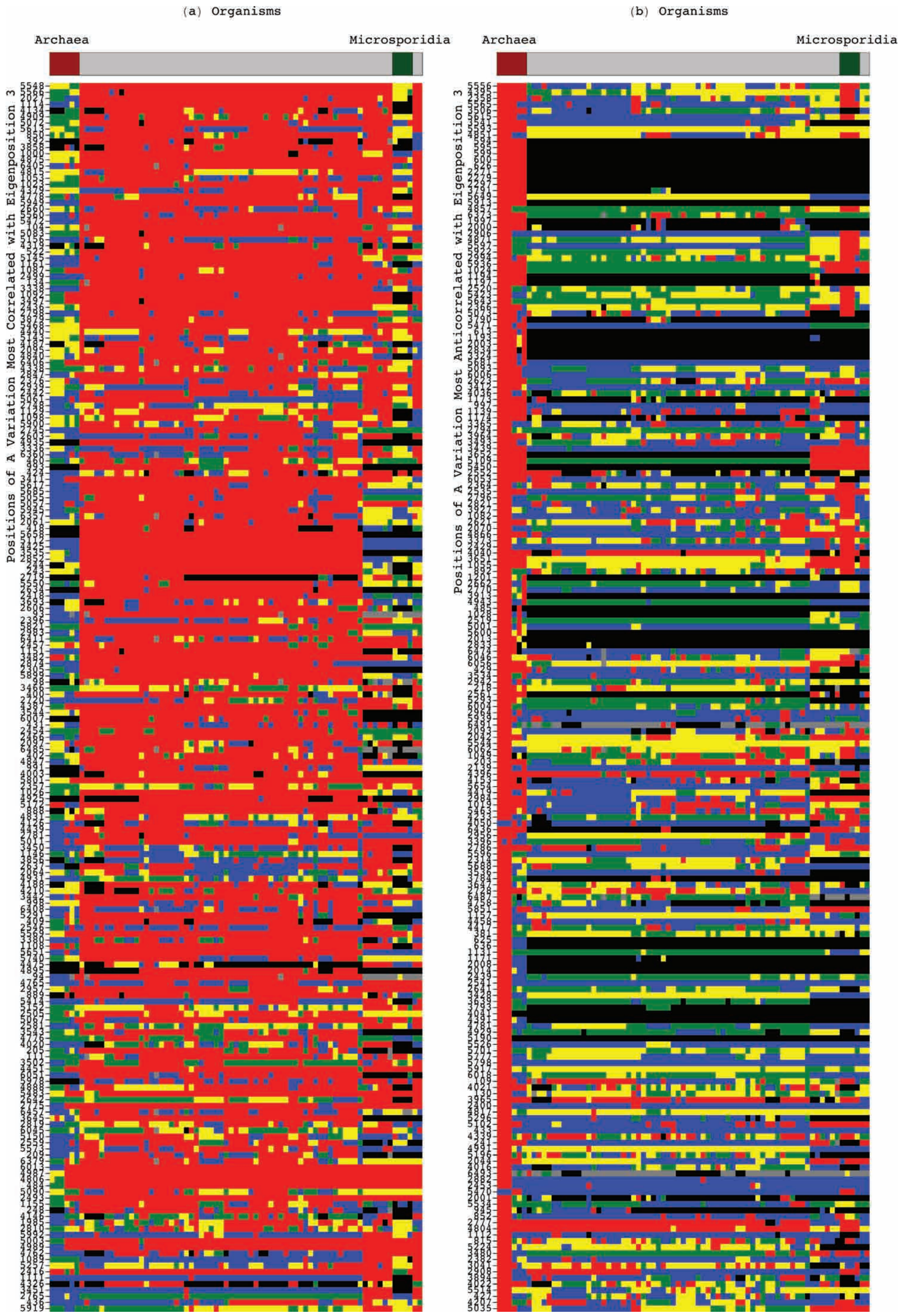

**Fig.** 3.15: **Unpaired adenosines exclusive to Bacteria or Archaea and Microsporidia 23S rRNAs.** Raster displays of the 200 positions in the 23S alignment for which the A nucleotide frequency variation is most (*a*) correlated or (*b*) anticorrelated with the third eigenposition, as identified by the A segment of the third eigenorganism.

The 100 positions with largest nucleotide frequency decrease in the A nucleotide segment of this eigenorganism include all 14 unpaired A nucleotides exclusively conserved in the Archaea (Figure 3.17(a)), implying that these seven unpaired adenosines are exclusively missing in the Microsporidia. In the C, G and U segments, the 100 positions with largest nucleotide frequency decrease are enriched in helices exclusively conserved in the Archaea (Figure 3.17(b) and Figure 3.18), with the $P$-values $< 10^{-8}$. These same positions in the mitochodrial 16S rRNA do not follow a trend similar to either the Archaea or the Microsporidia.

The fifth most significant 23S eigenorganism identifies gaps exclusively conserved in either the Microsporidia or the Archaea (Table 3.3) that map out entire substructures (yellow) deleted or inserted, respectively, in the Microsporidia relative to the Archaea, and vice versa (Figure 3.19). The gap segment of the 23S fifth eigenorganism identifies 191 of the 387 and 15 of the 59 sequence gaps exclusive to the Microsporidia and the Archaea, respectively, with both $P$-values $< 10^{-9}$.

The same eigenorganism also identifies unpaired adenosines, exclusively conserved in either the Microsporidia or the Archaea, some of which map to the same substructures. The 200 positions with largest frequency decrease in the A nucleotide segment of the same eigenorganism identify 39 of the 49 unpaired A nucleotides that are exclusively conserved in the Archaea (Figure 3.19(a), red). The 200 positions with largest frequency increase in the A nucleotide segment of the same eigenorganism identify 16 of the 31 unpaired A nucleotides that are exclusively conserved in the Microsporidia (Figure 3.19(b)). Of these, nine correspond to gaps conserved in the Archaea.

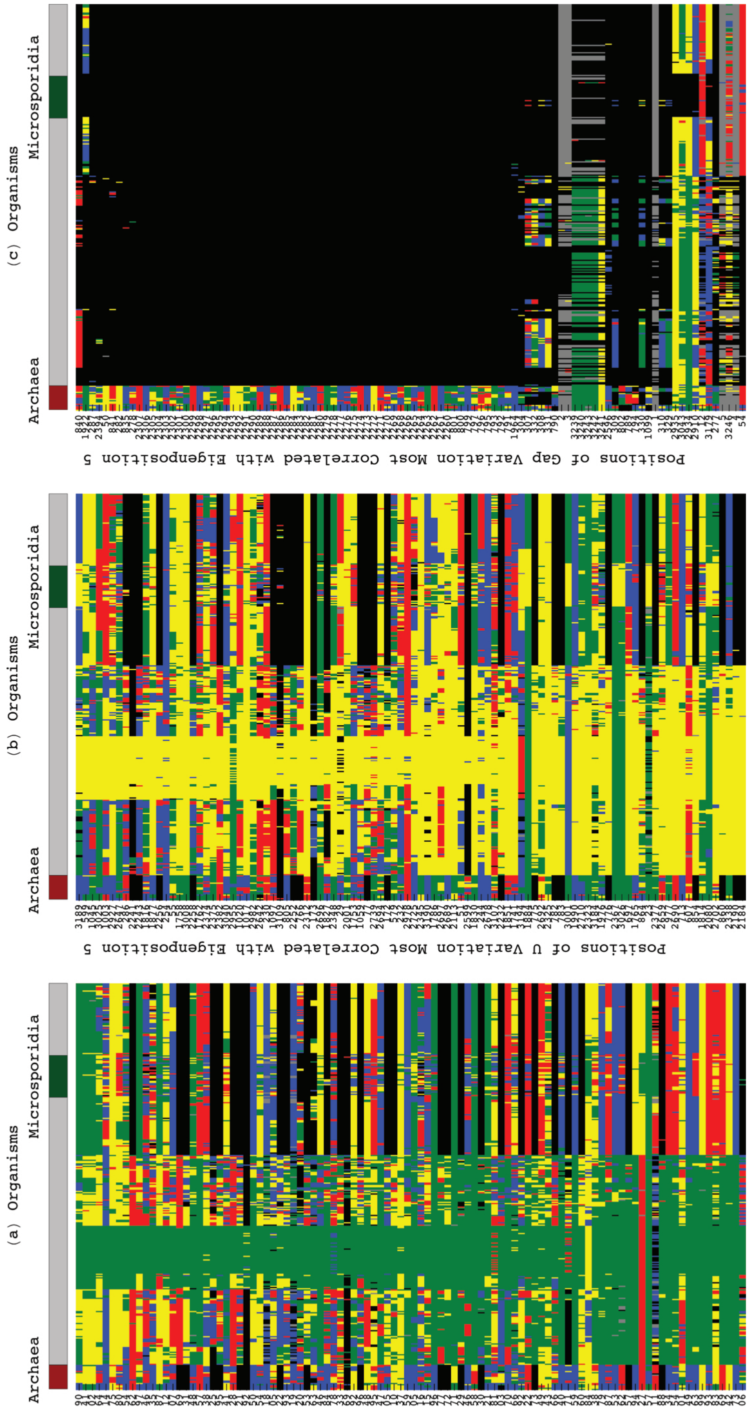

**Fig.** 3.16: **Nucleotides exclusive to Microsporidia 16S rRNAs.** Raster displays of the 100 positions each identified in the (*a*) C and (*b*) U nucleotide and (*c*) gap segments of the fifth eigenorganism with the largest increase in relative nucleotide frequency.

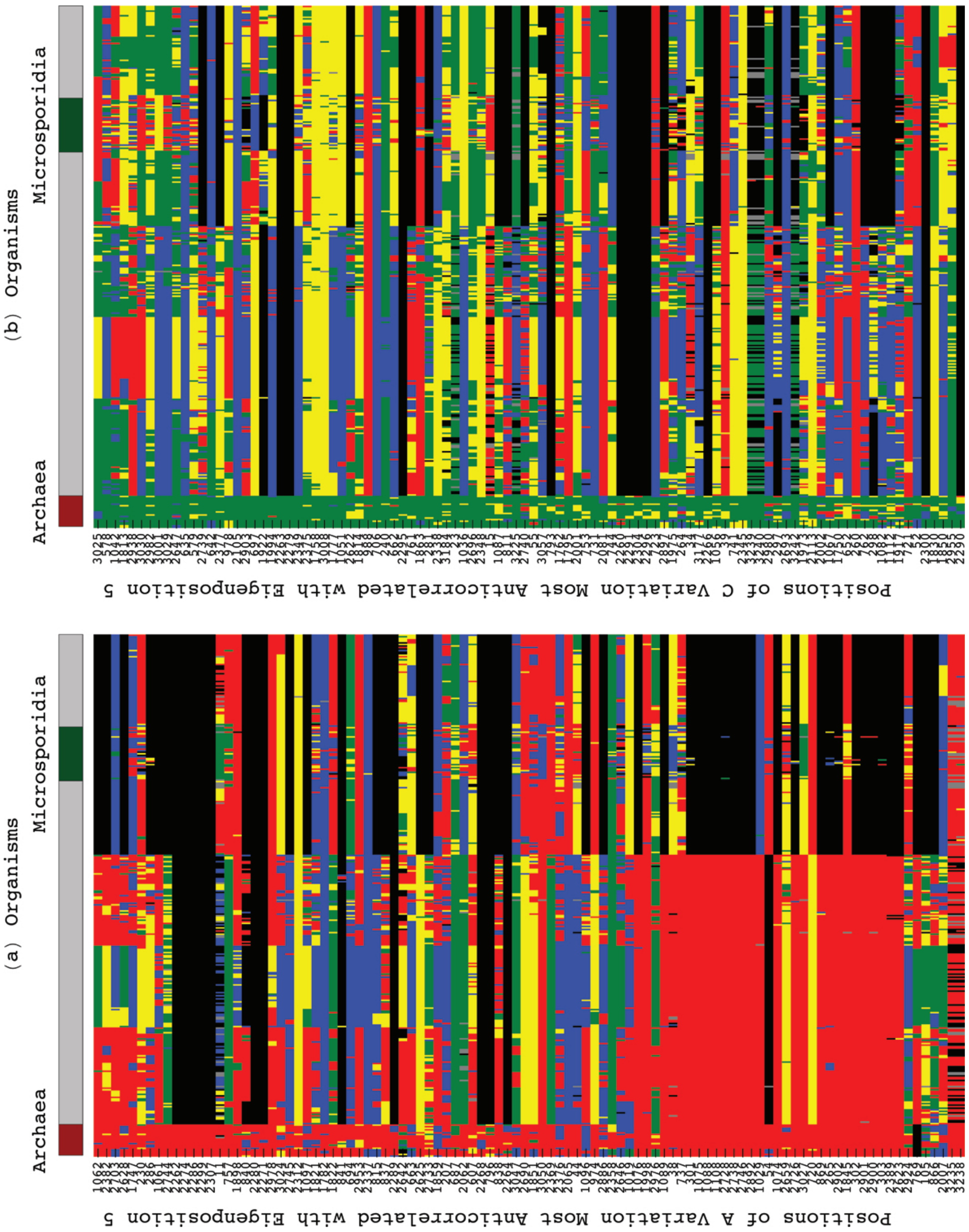

**Fig.** 3.17: **Adenosine and Cytosine nucleotides exclusive to Archaea 16S rRNAs.** Raster displays of the 100 positions each identified in the (*a*) A and (*b*) C nucleotide segments of the fifth eigenorganism with the largest decrease in relative nucleotide frequency.

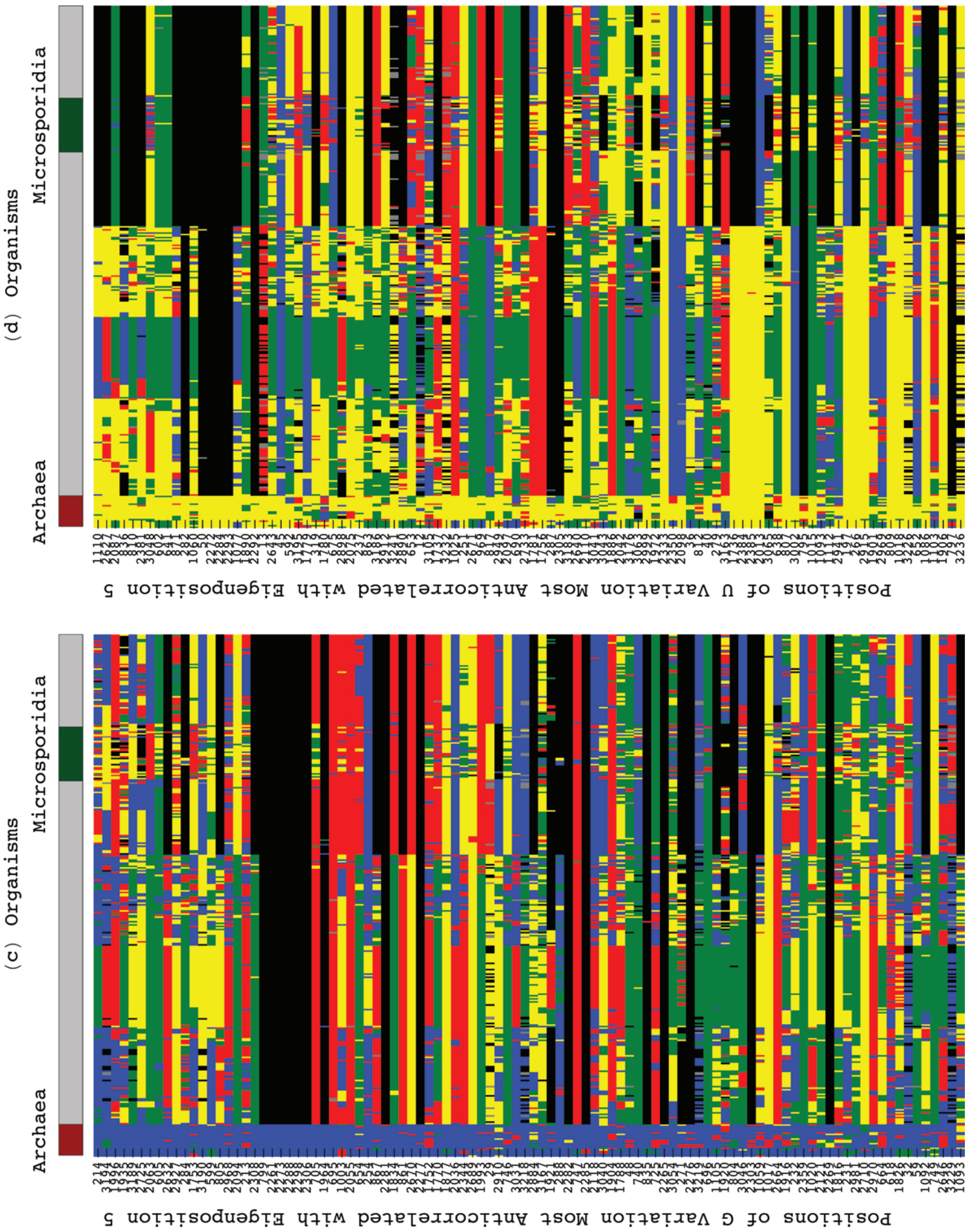

**Fig.** 3.18: **Guanosine and Uracil nucleotides exclusive to Archaea 16S rRNAs.** Raster displays of the 100 positions each identified in the (*c*) G and (*d*) U nucleotide segments of the fifth eigenorganism with the largest decrease in relative nucleotide frequency.

The A nucleotide segment of this eigenorganism identifies 16 of the 31 adenosines exclusively conserved in the Microsporidia, and 39 of the 49 unpaired adenosines exclusively conserved in the Archaea, respectively, with both $P$-values $< 10^{-16}$ (Figure 3.21).

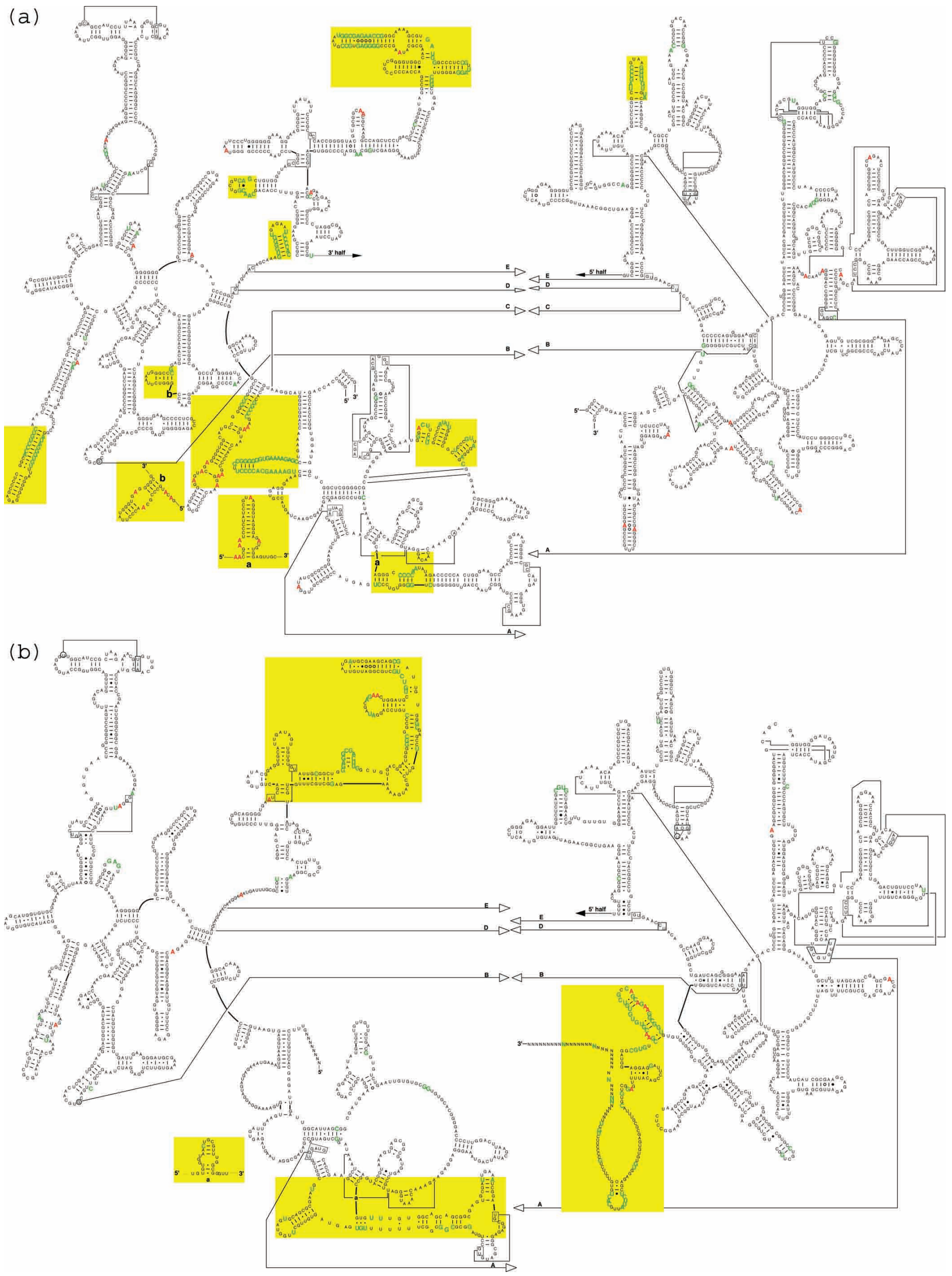

**Fig.** 3.19: **Sequence gaps and unpaired adenosines exclusive to Microsporidia or Archaea 23S rRNAs.** (*a*) The 200 positions of gap variation (green) across the organisms most correlated with the fifth eigenposition plotted on the secondary structure model of *M. jannaschii* . (*b*) The 199 positions of gap variation (green) across the organisms most anticorrelated with the fifth eigenposition plotted on the secondary structure model *E. cuniculi*.

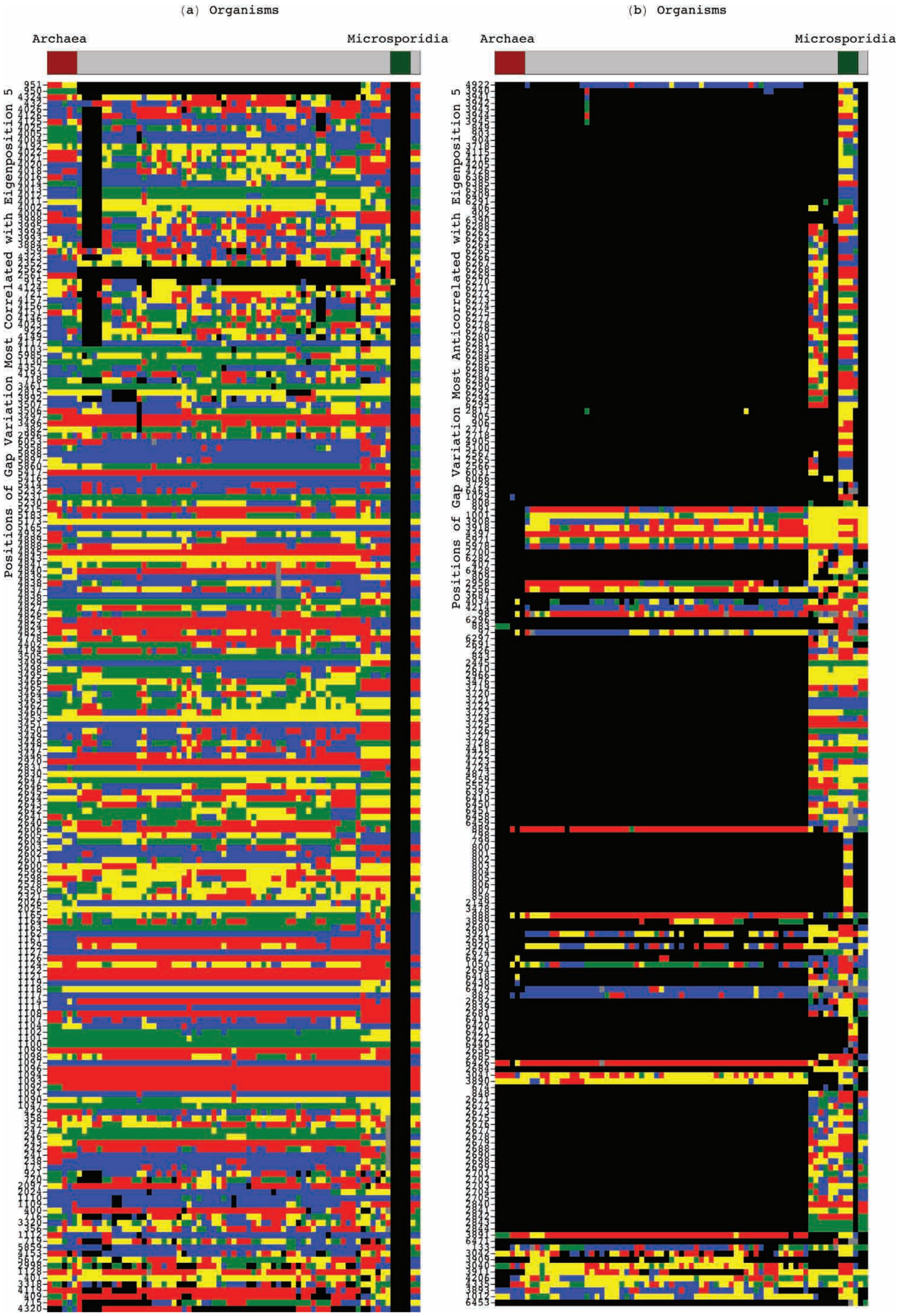

**Fig.** 3.20: **Sequence gaps exclusive to Archaea or Microsporidia 23S rRNAs.** Raster displays of the 200 and 199 positions in the 23S alignment for which the A nucleotide frequency variation is most (*a*) correlated or (*b*) anticorrelated with the fifth eigenposition, as identified by the gap segment of the fifth eigenorganism.

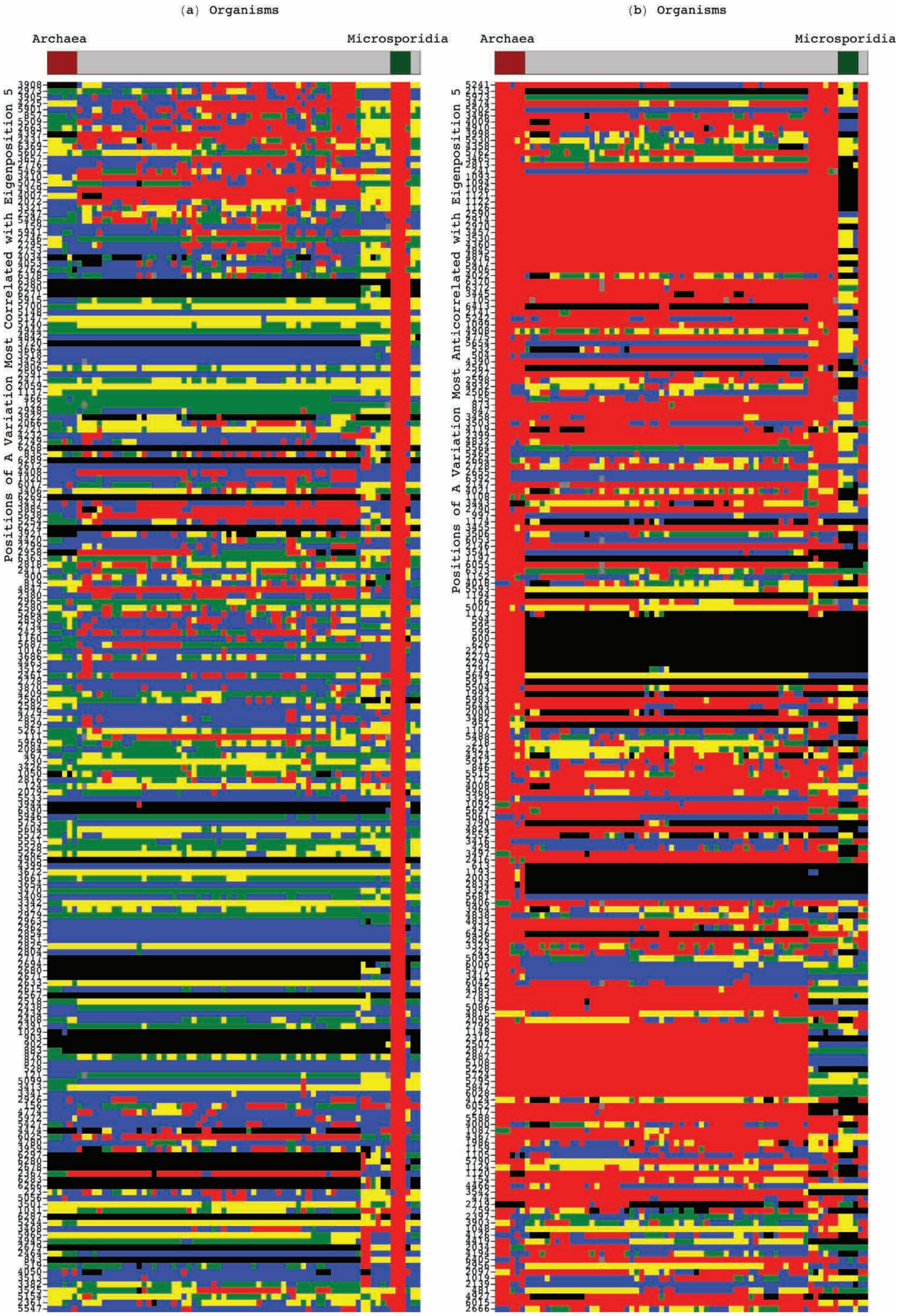

**Fig.** 3.21: **Unpaired adenosines exclusive to Archaea or Microsporidia 23S rRNAs.** Raster displays of the 200 positions in the 23S alignment for which the A nucleotide frequency variation is most (*a*) correlated or (*b*) anticorrelated with the second eigenposition, as identified by the A nucleotide segment of the second eigenorganism.

| *E. coli* position number | Secondary interaction | Secondary Motif | Tertiary interaction | Interaction type |
|---|---|---|---|---|
| 179 | A179:A196 | | | |
| 181 | G181:U182 | | | |
| 195 | A195:U180 | | U222:A141 | Backbone (U:A).(A.U) |
| 196 | A196:A179 | | C221:G142 | Backbone (A:A).(C:G) |
| | | | G142:C221 | Base-base (A:A)(G:C) |
| 197 | | | G220:A143 | Base-base (G:A)A |
| 300 | A300:G297 | AA.AG@helix.ends | U565 | Base-base (G:A)U |
| 382 | | Tetraloop | G64 | Backbone G.A |
| 411 | | | A430 | Base-base AA |
| 414 | | | A430 | Base-base AA |
| 430 | | | A411 | Base-base AA |
| | | | A414 | Base-base AA |
| 431 | | | | |
| 432 | A432:G410 | AA.AG@helix.ends | | |
| 448 | A448:U486 | E loop, Tandem GA | | |
| 451 | | | A373 | Backbone A.A |
| 452 | A452:U480 | | | |
| 482 | | | G391:C370 | Base-base (G:C)A |
| 487 | A487:G447 | AA.AG@helix.ends, E loop | | |
| 495 | A495:U438 | LUA@helix.ends | | |
| 510 | A510:C508 | | G542:C503 | Base-base (C:A)(C:G) |
| 563 | A563:U884 | LUA@helix.ends | | |
| 607 | | | G309:C291 | Backbone (G:C).A |
| 608 | | | G292:C308 | Base-base (G:C)A |
| 609 | | | | |
| 621 | | | C401:G41 | Backbone (C:G).A |
| 622 | A622:C618 | | G42:C400 | Base-base (C:A)(G:C) |
| 642 | A642: U641 | | U598:A640 | Base-base (U:A)(U:A) |
| 675 | A675:A715 | | | |
| 702 | | K-turn | | |
| 994 | | | | |
| 1004 | | | A1035:G1026 | Base-base (G:A)A |
| 1014 | | Tetraloop | U1219:A986 | Backbone (U:A).A |
| 1016 | A1016:G1013 | Tetraloop, AA.AG@helix.ends | G988:C1217 | Base-base (C:G)(A:G) |
| | | | C1217:G988 | Backbone (G:C).(A:G) |
| 1046 | A1046:U1211 | | A1213:U991 | Base-base (U:A)(A:U) |
| | | | C995 | Base-base (U:A)C |
| 1110 | | | | |
| 1130 | | | | |
| 1146 | | | G1127:C1145 | Base-base (C:G)A |
| 1188 | | | | |
| 1248 | A1248:A1289 | AA.AG@helix.ends | | |
| 1250 | | | | |
| 1261 | A1261:G1274 | AA.AG@helix.ends, GGA/GAA | | |
| 1269 | A1269:G1266 | Tetraloops, AA.AG@helix.ends | G1312:C1325 | Base-base (G:A)(G:C) |
| 1275 | A1275:C1260 | GGA/GAA | | |
| 1279 | | | | |
| 1280 | | | C1149:G1124 | Base-base (G:C)A |
| 1287 | | | G1370:C1352 | Base-base (C:G)A |
| 1288 | A1288:C1249 | | | |
| 1289 | A1289:A1248 | AA.AG@helix.ends | G1371:U1351 | Base-base (A:A)(G:U) |
| 1299 | A1299:A1239 | | | |
| 1408 | A1408:A1493 | AA.AG@helix.ends | | |
| 1447 | A1447:G1459 | | | |

Table 3.5: **Unpaired Adenosines exclusively conserved in the Bacterial 16S rRNA, and their tertiary interactions**. The 50 unpaired A's in Bacteria (*E. coli*) that are significantly anticorrelated with the second eigenorganism are listed here, along with their secondary and tertiary interactions, derived from the *T. thermophilus* crystal structure (compiled from RNA2DMap v2 [16]).

# Chapter 4

# Discussion

We describe here a novel application of the matrix decomposition techniques in the analysis of RNA sequence alignments. A six-bit binary code is used to convert the alphanumeric alignments to numeric tensors. The tensors are flattened back into matrices, and SVD or mode-1 HOSVD is applied. Results are presented from the analysis of alignments of 16S and 23S ribosomal RNA sequences. In each case, the decompositions simultaneously uncover uncorrelated patterns of variation across both dimensions of the alignments.

## 4.1 Most significant eigenposition is invariant

We find that the most significant eigenposition in our rRNA datasets, which captures $\sim$70% of the variation in the data, is approximately invariant across the organisms. This eigenposition correlates with the average frequency of all nucleotides across the positions, consistent with the most significant principal component in the PCA of an uncentered matrix [15].

We interpret the remaining eigenpositions and the nucleotide-specific segments of the corresponding eigenorganisms as patterns of nucleotide frequency variation relative to these averages.

## 4.2 Eigenpositions correspond to phylogenetic groups

The remaining significant eigenpositions uncovered in both the 16S and 23S data cuboids identify the dominant taxonomic groups among the organisms and their relations of similarity and dissimilarity. Further, the taxonomic groups identified in the various rRNA datasets examined are qualitatively similar (Table 4.1).

In more general terms, the eigenpositions can be understood as a clustering of sequences in the alignment (See Section 2.2.3 on page 23). In the case of the rRNA sequences, this similarity is correlated with taxonomic groups, owing to the high degree of sequence conservation of the rRNAs. Our preliminary studies of Group I introns from various structural classes suggest that the principal components uncovered in the data are indeed correlated with the structural classes (data not shown). In preliminary analyses of mouse SNPs, we found that eigenpositions are correlated with the mouse strains from which the SNPs are derived (data not shown).

## 4.3 Eigenorganisms identify positions uniquely conserved within phylogenetic groups

The eigenorganisms in our 16S and 23S data cuboids identify positions in the rRNA structure that are uniquely conserved in the taxonomic groups separated by the corresponding eigenpositions. The eigenorganisms indicate the degree of correlation of positions in the alignment with the eigenpositions. They may be thought of as identifying positions that confer similarity or dissimilarity upon the organisms (See Section 2.2.3 on page 24).

We are able to identify, from the second eigenorganism, previously known as well as new structure motifs that uniquely define the Bacterial 16S rRNA structure (Figure 3.3, compare with Figure 3.4). Similarly, the second eigenorganism in the 23S data identifies structure motifs uniquely conserved in the Bacterial and Eukaryotic 23S (Figure 3.6, regions marked in yellow).

The positions of nucleotide variation that are most correlated and anticorrelated with each eigenorganism map out not only isolated nucleotides, but also entire substructures deleted or inserted in one taxonomic group with respect to another. This suggests that entire structure motifs are involved in rRNA function and folding, and that mutational changes in isolated nucleotides often result in compensatory changes, or insertions and deletions, in interacting nucleotides.

## 4.4 Unpaired adenosines are significant in distinguishing structure motifs

The eigenorganisms identify adenosines, unpaired in the rRNA secondary structure, which are conserved exclusively in the respective taxonomic groups, most of which participate in tertiary structure interactions and map to the substructures inserted or deleted within taxonomic groups.

Previous comparative studies observed that nearly 66% of all A nucleotides in Bacterial rRNAs (from an analysis of 66017 sequences) are unpaired, as compared to 24%, 30%, and 40% of C's, G's, and U's respectively [47, 49]. Both the 16S and 23S rRNAs show this marked bias towards unpaired A's (Figure 4.1).

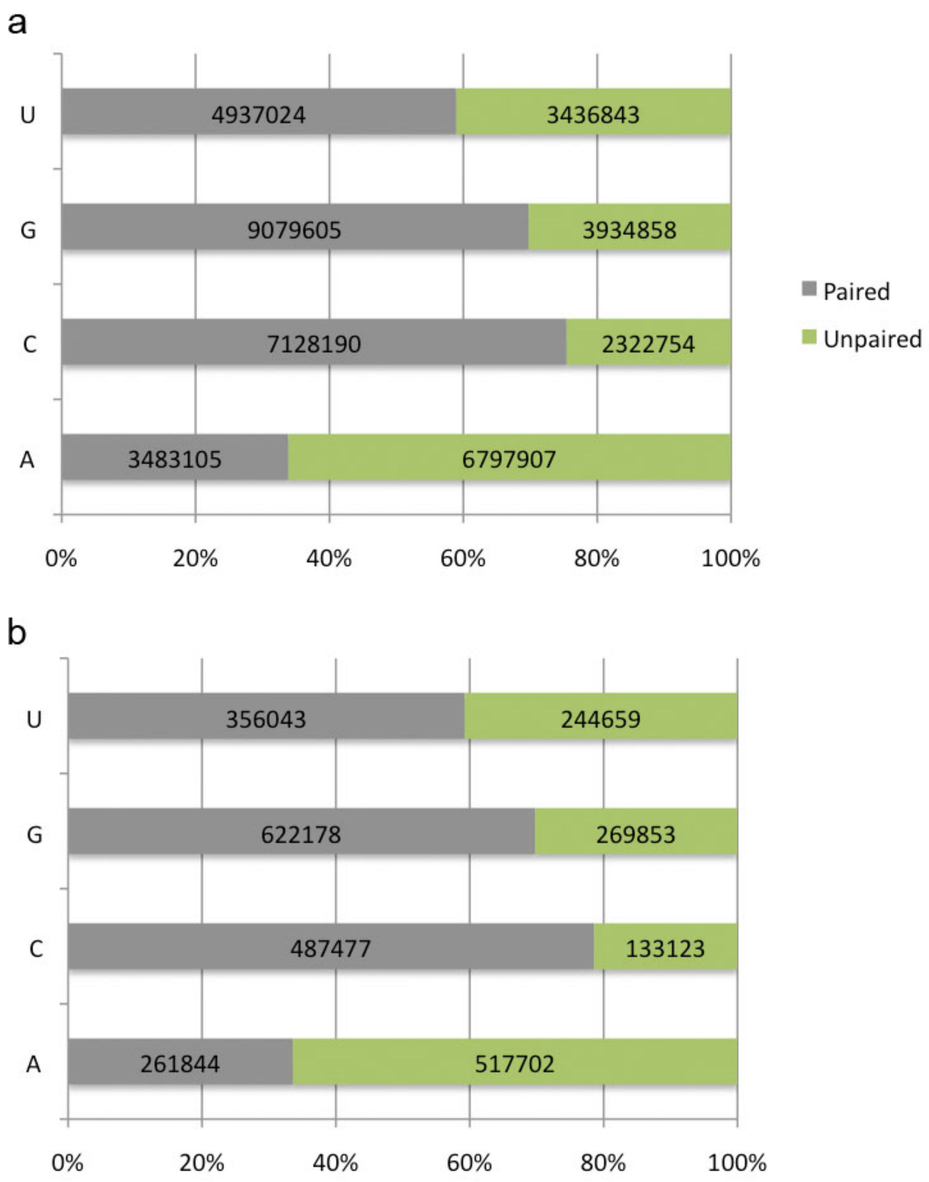

**Fig.** 4.1: **Relative percentages of paired and unpaired nucleotides in Bacterial rRNAs (data from CRW [16]).**
(a) Nucleotide statistics from 59711 16S rRNA sequences, showing that 66.1% of A's are unpaired, in comparison with 24.6% C's, 30.2% G's, and 41% U's.
(b) Nucleotide statistics from 66017 23S rRNA sequences, showing that 66.4% of A's are unpaired, in comparison with 21.5% C's, 30.2% G's, and 40.7% U's.

It was also noted that these unpaired adenosines are especially abundant in tertiary structure motifs such as tetraloops [116], E-loops [42], adenosine platforms [19], and AA side-step [24].

**Tetraloops:** Tetraloops are four-base hairpin loops that cap many double helices in rRNAs. It was observed from comparative analysis of 16S rRNAs that tetraloop sequences are highly constrained, independent of the location of the loops in the secondary structure [116]. Of the 256 possible tetraloop sequence configurations, only 16 occur in nature, and the majority of tetraloops fit the sequence pattern GNRA [74].

Experimental studies of naturally occurring tetraloop sequences show a positive selection for thermodynamic stability [8], suggesting a significant role in RNA folding [105]. The recognition of both the 16S and 23S rRNAs by the cytotoxic protein ricin is mediated by GNRA tetraloops [44]. Experimental observations of intra- and intermolecular interactions involving these loops and other motifs rich in unpaired adenosines [19] suggested a role for these unpaired nucleotides in a universal mode of RNA helical packing [30, 77] as well as in the accuracy and specificity of the translational function of the rRNA protein synthesis [58, 68, 71, 79].

Our results show an enrichment of unpaired A's in positions that distinguish taxonomic groups. A significant number of these unpaired A's we identify as distinguishing the Bacteria from the Eukarya are involved in tertiary base-base and base-backbone interactions in the bacterial 16S rRNA crystal structure (Tables 3.4, 3.5) [22]. In the absence of 16S crystal structures from other domains, we cannot rule out the possibility of compensatory tertiary interactions that result in similar thermodynamic and folding profiles.

However, the exclusive conservation of different sets of unpaired A's in several phylogenetic groups in both the 16S and 23S rRNAs, combined with the preponderence of unpaired A's in structure motifs experimentally verified to be significant in RNA function and folding, leads us to believe that the unpaired A's are indeed determinants of folding pathways that are unique to the phylogenetic groups.

We hypothesize, therefore, that though the 16S and 23S rRNA possess a high degree of sequence similarity across the tree of life, these differences in motifs involved in rRNA folding and function could result in significant differences in transcriptional regulation and efficiency.

## 4.5 Eigenpositions identify multiple pathways of evolution

The third and fifth eigenpositions and eigenorganisms in the 16S and 23S data reveal two orthogonal, i.e., uncorrelated, evolutionary pathways relating the Archaea and Microsporidia, demonstrating the ability of this mode-1 HOSVD to uncover multiple subgenic patterns of evolution in an aligment of sequences of a single rRNA molecule.

### 4.5.1 The Archaea

The Archaea are single cell prokaryotes of extremely small genomes. Archaeal rRNAs are more similar to bacterial rather than eukaryotic rRNAs.

Archaeal ribosomal proteins, however, are more similar to eukaryotic rather than bacterial ribosomal proteins [115].

### 4.5.2 The Microsporidia

The Microsporidia are a diverse, species-rich group of unicellular eukaryotes. They are obligate intracellular parasites which infect a wide variety of animals, as well as certain ciliates and gregarine apicomplexa [38]. They have also been used as agents for biological control of insect pests (e.g., *Nosema locustae* against tropical grasshoppers) [43]. There has been a renewed interest in the study of microsporidia since their discovery as major opportunistic pathogens in immunocompromised HIV patients [73].

Outside their host cells, microsporidia exist as hardy spores protected by protein and chitin walls. Infection occurs by the piercing of the host cell by a tightly-bound organelle called the polar tubule [38]. Apart from their curious infection mechanism, the microsporidia have been an interesting group in systematics owing to their largely simplified genomes [38]. They are not only one of three major amitochondriate eukaryotic lineages, but also lack most other membrane-bound organelles like the Golgi complex, peroxisomes, etc. This led to considerable ambiguity in their phylogenetic position by traditional morphology-based systematics. For a long time, they were grouped along with such diverse organisms as the archamoebae and the parabasala [108]. A chief concern in assigning phylogenetic positions to such organisms is the fact that, unlike plants and animals, they share no true synapomorphies, that is, there is no trait that unifies them to the exclusion of other groups [96]. With the advent of molecular systematics, there has been a re-classification of these amitochondriate eukaryotes, leading to a re-thinking of hypotheses about events in early eukaryotic evolution.

#### 4.5.2.1 The Archezoa Hypothesis

Early eukaryotic evolution has been hypothesized as a period of anaerobic evolution producing a nucleated phagocytic cell which engulfed a mitochondrial endosymbiont, thought to be an $\alpha$-proteobacterium [14, 36]. The acquisition of this endosymbiont was thought to confer an evolutionary advantage to the host cell, allowing it to colonize emerging aerobic environments. The existence of anaerobic, amitochondriate eukaryotes lent credence to this theory, since they were thought to be examples of primitive organisms which were hosts to the endosymbiont [36]. Building on this hypothesis, in 1983, Cavalier-Smith proposed a eukaryotic subkingdom, the Archezoa, which included eukaryotes that predated the mitochondrial acquisition. Historically, this group has included four phyla: the Archamoebae (e.g., Entamoeba), the Metamonads (e.g., Giardia), the Parabasala (e.g., Trichomonas), and the Microsporidia [87].

Shortly after, this hypothesis gained support from Vossbrinck *et. al.* [108], who first included a microsporidian, *Vairimorpha necatrix,* in their phylogenetic analysis of 18S small sub unit (SSU) rRNA sequences from 10 eukaryotes. Using a distance-based approach as well as maximum parsimony on this data, they inferred a phylogeny in which the microsporidia were at the base of the eukaryotic tree (Figure 4.3). Based on this tree, they hypothesized that the microsporidia-eukaryote divergence must have occurred very early in time, possibly 2.9–2.7 BYA, when the earth's atmosphere lacked free oxygen. The authors however caution that the *V. necatrix* SSU rRNA molecule "lacks various regions of the molecule considered to be 'eukaryotic' ", and this deviation from the mean SSU rRNA length may have had an effect on their analysis.

Brown and Doolittle [13] attempted to reconstruct a rooted universal tree using aminoacyl-tRNA synthetases, which are thought to have diverged prior to the emergence of prokaryotic and eukaryotic lineages. The phylogenies reported in this study, derived from a consensus parsimony analysis as well as neighbor joining analysis, support the basal positioning of the microsporidia. Kamaishi and coworkers [59, 60] found further support for the early divergence of the microsporidia, from the analysis of elongation factors EF-1$\alpha$ and EF-2 from the microsporidian *Glugea plecoglossi*. Thus, early molecular data apparently confirmed the Archezoa, or the "Microsporidia-early" hypothesis.

#### 4.5.2.2 Conflicting phylogenies from other genes

The first evidence contradicting the basal position of the microsporidia came from $\alpha$- and $\beta$-tubulins, whose phylogenetic analysis showed that the microsporidia surprisingly emerged within the fungi, with strong bootstrap support [61]. This result was further supported by a congruent $\beta$-tubulin tree, leading the authors to

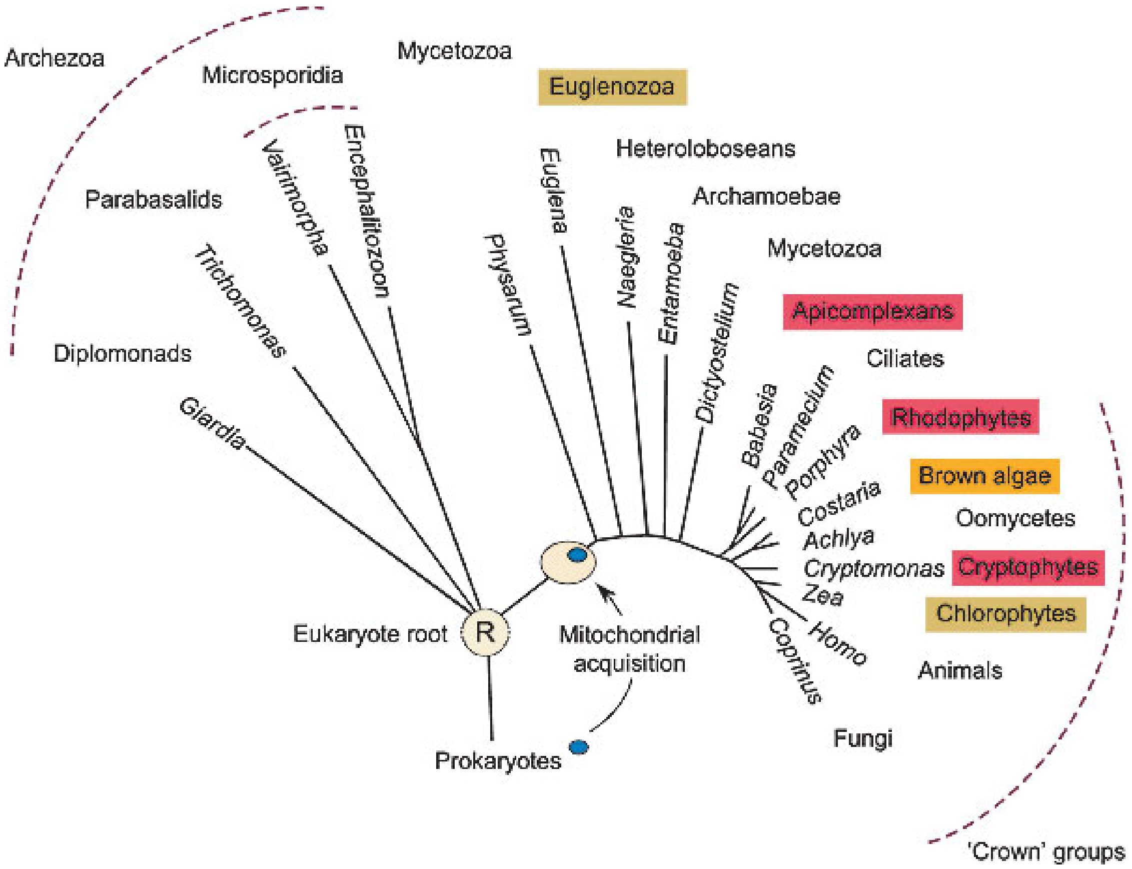

**Fig.** 4.2: **A eukaryotic tree from early SSU rRNA analysis (reproduced from Embley, 2006 [35]).** The tree supports the Archezoa hypothesis, which classifies the microsporidia as basal Eukarya. Analysis of elongation factors EF-1$\alpha$ and EF-2 also supported similar tree topologies.

consider the possibility that earlier support for the archezoa hypothesis may have been an artifact of long-branch attraction. Independent of these results, Edlind *et. al.* [34] found that an analysis of $\beta$-tubulin sequences from a set of eukaryotes including four microsporidia, using both distance-based methods and parsimony, grouped the microsporidia as a sister group of the fungi. Again, this led them to hypothesize that the microsporidia are not primitive at all, but rather, have evolved degeneratively from higher, free-living eukaryotes.

In the analysis of sequences of the TATA box-Binding Protein (TBP), a universal transcription factor, from the microsporidian *Nosema locustae*, maximum likelihood (ML) and neighbour-joining analysis showed a weak but consistent fungal affinity for the microsporidian [37]. Stiller and Hall carried out an investigation to see if the earlier results from SSU rRNA that supported the basal positioning of microsporidia [108] were indeed artifacts of systematic phylogenetic reconstruction error [97]. They note that the sequences that cluster near the base of the eukaryotic tree (Microsporidia and Diplomonads) tend to be more similar in length to one another than to the "crown taxa". Also, when the archaeal outgroups in the analysis were replaced with randomly generated sequences with base composition similar to that of eukaryotes, they clustered along with the Microsporidia. Although there was no observable correlation between variation from the "standard" 1.8 kb sequence length and position on the tree, a study of previously published eukaryotic trees from rRNA sequences showed that in all cases, the variation in sequence lengths of basal taxa is highly significant. This led the authors to hypothesize that "large insertions or deletions in rRNA genes could be either the cause or a consequence of an increased rate of sequence evolution". Together, their analyses indicate that the "crown taxa" are merely a group of eukaryotes that have undergone a more normal mode of evolution, while the "basal eukaryotes" represent an artificial clustering of more rapidly evolving sequences.

Van de Peer and coworkers [107] constructed a large subunit (LSU) rRNA phylogenetic tree based on 42 sequences of representatives of the different eukaryotic crown taxa plus the sequences of the microsporidia *Nosema* and *Encephalitozoon*. They found that the microsporidia diverged from within the fungal cluster with a relatively low bootstrap support (62%), which they explain as caused by the long branch.

This conflict in gene trees can be explained in two different ways: the phylogenies of one or the other of these genes is reconstructed incorrectly owing to an artifact in the reconstruction method, or that microsporidia may have acquired a subset of genes, such as the tubulins, by lateral transfer from their hosts [87]. Uneven taxonomic sampling, wide disparities in evolutionary rates among lineages, and/or inadequate characterization have been suggested as causes for artifacts in phylogenetic reconstruction. Baldauf *et al.* [9] created a phylogeny comparable to that of SSU rRNA by combining the deduced amino acid sequences of four protein-encoding genes. The encoded proteins $\alpha$-tubulin, $\beta$-tubulin, actin, and elongation factor 1alpha (EF-1$\alpha$) were analysed using the phylogeny inference package PAUP [100]. Their analyses places the Microsporidia along with Fungi, with a strong bootstrap support of 95%, suggesting that the early branching of Microsporidia is an artifact of their accelerated evolutionary rates for these genes.

More evidence for the long-branch attraction artifact came from the work of Philippe and Germot [85], who performed a combined analysis of SSU and LSU rRNA from around 136 eukaryotes and archaea. They found that the ML tree inferred assuming the model of equal evolutionary rates across sites (E) was similar to the one obtained from the analysis of SSU rRNA alone, while the tree inferred assuming a gamma distribution of rates ($\Gamma$) placed the Archezoa no longer basal to the other species. Conclusive evidence for the long-branch attraction artifact came much later, when in 1995, Fischer and Palmer showed, from an analysis of SSU rRNA from 83 available eukaryotic species, that “a fungal origin of Microsporidia is not statistically distinguishable from an ancient origin”, and that “a basal position of Microsporidia is no better than a position within the eukaryotic crown”.

Thus, most evidence subsequent to the first SSU rRNA analysis seems to indicate that the basal positioning of the microsporidia in that analysis was due to the fast-evolving long branches of the microsporidia being falsely attracted towards the long branch of the archaeal outgroup [9, 27, 85, 97], which can be attributed to differing G+C contents, rate heterogeneity, and an increased proportion of variable positions in these sequences. Improved phylogenetic reconstruction methods, accounting for among-sites rate variation, and additional taxon sampling have been suggested as solutions to overcome this problem. Noting that sequence data is prone to systematic errors due to homoplasy, Baldauf [9] suggested that, in addition to increased taxon sampling, it may be necessary to sample more than one molecule to be able to reconstruct higher-order taxonomy. Another alternative would be to use phylogenetic markers such as insertions and deletions, which although not free from homoplasy, make it easier to detect. Indeed, Baldauf's analysis of the 12-aa insertion in the EF-1$\alpha$ molecule, shared by all major animal and plant lineages and the microsporidia, but not by ciliates and other protists [9], strongly supports the fungal placement of the microsporidia. The disadvantage of such an approach, especially true in the case of organisms with highly reduced genomes, is that it is not trivial to find such conserved markers for analysis.

### 4.5.3 Archaea/Microsporidia relationship in the rRNAs

We uncover, from the analysis of a single alignment, two orthogonal relations between the Archaea and Microsporidia - one showing similarity, and one showing dissimilarity between the two groups.

#### 4.5.3.1 Similarity between the Archaea and the Microsporidia

In both 16S and 23S alignments, the third most significant eigenposition captures the similarities among the Archaea and the Microsporidia, and correlates with decreased nucleotide frequency across both the Archaea and Microsporidia relative to all other organisms.

The 100 positions with largest nucleotide frequency decrease in the gap segment of the third 16S eigenorganism identify all six gaps exclusively conserved in both the Archaea and Microsporidia. Mapped onto the secondary structure model of

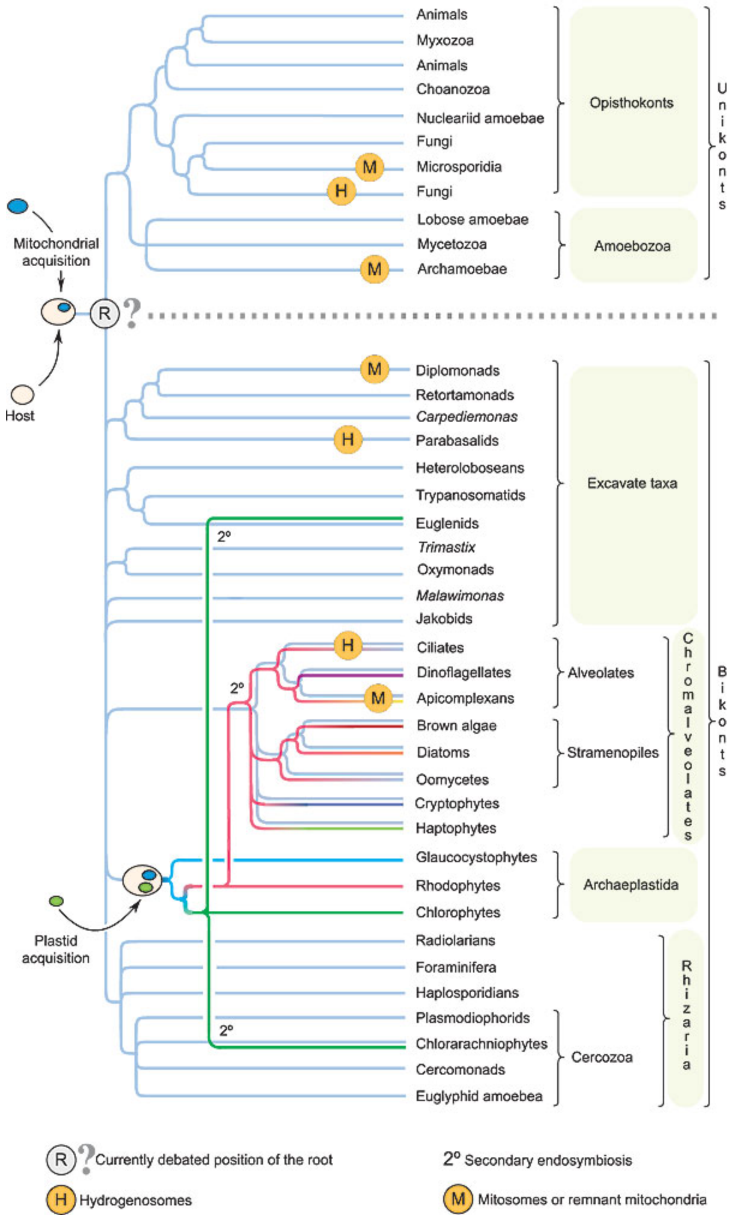

**Fig.** 4.3: **Consensus tree of the eukaryotes, representing current hypotheses about early eukaryotic evolution (reproduced from Embley, 2006 [35]).** It is now almost universally accepted that the common ancestor of all eukaryotes contained mitochondria, which then underwent reduction independently in several lineages, but was never completely lost.

*E. coli*, these 100 positions identify deletions of entire substructures in the Archaea and Microsporidia with respect to the Bacteria (Figure 3.12(a), substructures I–III), indicating a convergent loss in both the Archaea and Microsporidia with respect to the Bacteria as well as the Eukarya.

#### 4.5.3.2 Dissimilarity between the Archaea and the Microsporidia

The fifth 16S and 23S eigenpositions both capture the dissimilarities between Archaea and Microsporidia and correlate with increased frequency across the Microsporidia, and decreased frequency across the Archaea.

The positions that are exclusively conserved in the Microsporidia include C and U nucleotides in helix regions, in addition to unpaired A's (Figure 3.16). The positions exclusively conserved in the Archaea include C, G, and U nucleotides in helix regions, as well as unpaired A's (Figure 3.17). These same positions in the mitochodrial 16S rRNA do not follow a trend similar to either the Archaea or the Microsporidia.

We observe these similarities and differences in the 23S rRNA as well, which follow the same trends as the 16S rRNA.

Together, the third and fifth eigenpositions and eigenorganisms reveal two orthogonal, i.e., uncorrelated, evolutionary pathways relating the Archaea and Microsporidia, demonstrating the ability of this mode-1 HOSVD to uncover multiple subgenic patterns of evolution in an aligment of sequences of a single rRNA molecule.

### 4.5.4 Genome compaction and evolution

The loss of identical structures in the Archaea and the Microsporidia could be either due to independent convergent events or the result of a single evolutionary pressure. However, given that the Mitochondria also show a loss of the same structures, the first scenario seems unlikely.

It has been noted that the microsporidian genomes are very small in size, ranging from 19.5 Mbp in *Glugea atherinae*, to only 2.3 Mbp in *Encephalitozoon cuniculi* [62]. The *E. cuniculi* genome codes for only about 2000 proteins, indicating genome compaction by substantial gene loss. It was observed that gene loss is not random: although genes for certain metabolic pathways are completely absent, genes related to basic cellular processes like DNA replication and transcription are conserved [63]. This loss has been attributed to the parasitic lifestyle of the organism. Mitochondria are believed have undergone a similar compaction in their genomes [23]. Although the Archaea do not share this characteristic, their 16S rRNAs are comparable in size to those of the Microsporidia and the mitochondria.

We examined the positions that indicate similarity between Archaea and Microsporidia 16S, in a mitochondrial 16S rRNA alignment (Figure 3.12(c)), and found that the gaps conserved among the Archaea and Microsporidia are also conserved across the Metazoan mitochondrial 16S rRNA, but not among the other eukaryotic mitochondrial rRNA. Together, these results suggest that the similarity between the Archaea and the Microsporidia could be explained best by losses due to evolutionary forces driving genome compaction, and particularly, compaction of the 16S rRNA.

## 4.6 Robustness

Our analysis is data-driven; therefore, the relationships between organisms that are retrieved by the eigenpositions are dictated by the composition of the alignment. However, we find that the phylogenetic relationships retrieved by the most dominant eigenpositions are fairly robust to perturbations in the rRNA alignments.

### 4.6.1 Domain relationships

We performed the mode-1 HOSVD analysis on several 16S rRNA alignments with different taxonomic compositions, all derived from the same super alignment in the CRW, and also 23S and 5S rRNA alignments. In each case, the most significant eigenpositions differentiate the three domains, Archaea, Bacteria, and Eukarya (Table 4.1(a)). The enrichments of the taxonomic groups and of the structure motifs conserved within these groups, was, as expected, dependent on the number of organisms as well as positions in the alignment.

### 4.6.2 Archaea and Microsporidia relationship

The multiple evolutionary pathways that connect the Archaea and the Microsporidia are also robust to changes in the composition of the datasets. In the 339-organism 16S alignment, the two relationships are revealed even upon the removal of the Bacteria or the Eukarya (excluding the Microsporidia) (Figure 4.4

and Table 4.1(b)). In the 23S dataset, we see these two relationships despite the reduced number of Archaea and Microsporidia in the alignment (Table 4.1(b)).

(a) **Domain relationships**

| rRNA | Organisms | | | | Eigenpositions | | | |
|---|---|---|---|---|---|---|---|---|
| | **Total** | **Archaea** | **Bacteria** | **Eukarya** | **2** | | **3** | |
| 16S | 30 | 10 | 10 | 10 | Bacteria | Eukarya | Archaea | Bacteria |
| 16S | 62 | 13 | 28 | 21 | Archaea+Bacteria | Eukarya | Archaea | Bacteria+Eukarya |
| 16S | 220 | 13 | 117 | 90 | Bacteria | Eukarya | Archaea | Bacteria |
| 16S | 339 | 21 | 175 | 143 | Bacteria | Eukarya | Archaea+Micro | Bacteria |
| 23S | 75 | 6 | 57 | 12 | Bacteria | Eukarya | Archaea+Micro | Bacteria |
| 5S | 242 | 28 | 83 | 131 | Bacteria | Eukarya | Archaea | Actinobacteria |

(b) **Archaea and Microsporidia relationships**

| Dataset | Organisms | | | | | Eigenposition showing | |
|---|---|---|---|---|---|---|---|
| | **Total** | **Archaea** | **Bacteria** | **Eukarya** | **Microsporidia** | **Similarity** $(p_i)$ | **Difference** $(p_i)$ |
| 16S | 339 | 21 | 175 | 143 | 36 | 3 (0.02) | 5 (0.01) |
| 16S_ABM | 232 | 21 | 175 | 36 | 36 | 2 (0.047) | 4 (0.016) |
| 16S_AEM | 164 | 21 | 0 | 143 | 36 | 2 (0.08) | 3 (0.027) |
| 23S | 75 | 6 | 57 | 12 | 4 | 3 (0.016) | 5 (0.013) |

Table 4.1: **Summary of results from Mode-1 HOSVD analysis of various rRNA datasets, showing that the phylogenetic relationships are always revealed in the most dominant eigenpositions.** The amount of nucleotide frequency variation captured by each pattern, $p_i$, is calculated according to Equation 2.4.

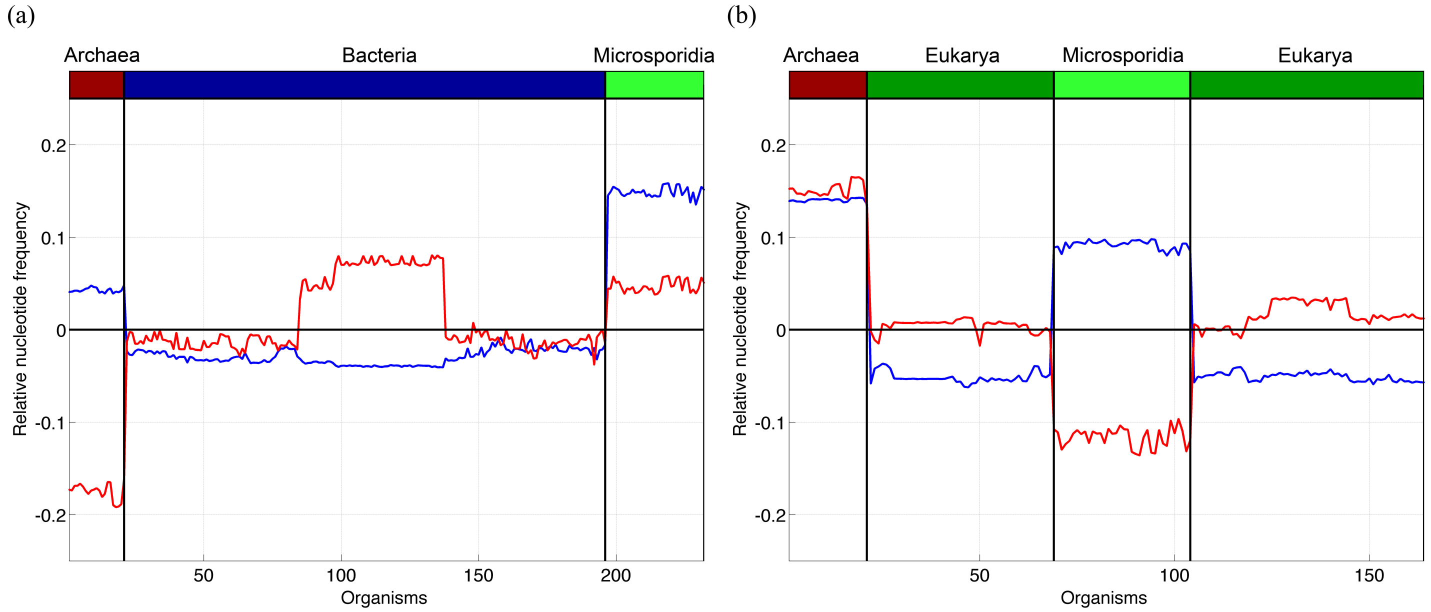

**Fig.** 4.4: **Eigenpositions from two reduced 16S rRNA datasets showing the Archaea/Microsporidia relationships.**
(a) In a 16S alignment with 21 Archaea, 175 Bacteria, and 36 Microsporidia, the second (blue, $p_i$=0.047) and fourth (red, $p_i$=0.016) most significant eigenpositions show a similarities and differences respectively between the Archaea and Microsporidia. (b) In a 16S alignment with 21 Archaea and 143 Eukarya including 36 Microsporidia, the second (blue, $p_i$=0.08) and third (red, $p_i$=0.027) most significant eigenpositions show a similarities and differences respectively between the Archaea and Microsporidia. The amount of nucleotide frequency variation captured by each pattern, $p_i$, is calculated according to Equation 2.4.

## 4.7 Conclusions

It was shown that the singular value decomposition (SVD) provides a mathematical framework for the modeling of DNA microarray data, where the mathematical variables and operations represent biological reality [1]. The variables, significant patterns uncovered in the data, correlate with activities of cellular elements, such as regulators or transcription factors. The operations, such as classification, rotation or reconstruction in subspaces of these patterns, were shown to simulate experimental observation of the correlations and possibly even the causal coordination of these activities. Recent experimental results [81] demonstrate that SVD modeling of DNA microarray data can be used to predict previously unknown cellular mechanisms [5, 80].

We now show that the mode-1 HOSVD, which is computed by using the SVD, provides a mathematical framework for the modeling of rRNA sequence alignments, independent of a-priori knowledge of the taxonomic groups and their relationships, or the rRNA structures, where the mathematical variables, significant eigenpositions and corresponding nucleotide-specific segments of eigenorganisms, represent multiple subgenic patterns of evolution.

The eigenpositions identify multiple orthogonal i.e., uncorrelated, relations of similarity and dissimilarity among the taxonomic groups of organisms, that might result from convergent as well as divergent evolutionary pathways. The corresponding eigenorganisms identify positions of nucleotides exclusively conserved within the taxonomic groups, but not among them, which map out entire substructures inserted or deleted in one taxonomic group relative to another, and are enriched in unpaired adenosines. These results suggest that insertions or deletions of entire substructures and unpaired adenosines, motifs which are known to be involved in rRNA folding and function, are correlated and possibly also causally coordinated with an organism's evolutionary pathway.

We also find in our analysis two orthogonal, i.e., uncorrelated, evolutionary pathways relating the Archaea and Microsporidia, demonstrating the ability of this mode-1 HOSVD to uncover multiple subgenic patterns of evolution in an aligment of sequences of a single rRNA molecule.

## 4.8 Implications for Future Research

We have created, in this work, a novel framework for the analysis of sequence alignments. Our methods provide a way for the data-driven classification of a set of aligned sequences, based on some metric of similarity, without *a priori* knowledge of the classes. We envision this property to be useful in the analysis of protein sequence alignments, to detect residues that confer binding specificities or functional diversity among proteins with sequence homology.

While it is common practice to infer phylogenetic trees from sequence alignments, it is now recognized that the true phylogeny of a set of organisms is more likely a network, with more than one line of descent [46, 75]. The phylogenetic tree of a group of organisms may be viewed as resulting from the superposition of multiple evolutionary pressures. Our methods enable us not only to detect these evolutionary forces, and the groups of organisms they act upon, but also identify sites in the alignment that are mutated as a result of these forces.

In recent years, genome-wide association studies [53] have identified multiple loci contributing to several human diseases involving complex traits, most notably type-2 diabetes [92], Crohn's disease [50], breast cancer [32], prostate cancer [104], lung cancer [7], and colorectal cancer [103]. Our HOSVD framework may be adapted to the analysis of SNPs, to simultaneously associate SNPs with disease phenotypes, and also to detect and remove systematic biases arising from population stratification in allele frequency data [86].

# Appendices

# Appendix A

# Mode-1 HOSVD analysis of 5S rRNA

## A.1 Introduction

The 5S ribosomal RNA is the smallest component of the large subunit, and is present in almost all organisms. It was found to be absent from the mitochondrial ribosomes of some fungi, vertebrates and most protists. It is approximate;y 120 nucleotides long, and like other rRNAs, has a strongly conserved secondary structure (Figure A.1). It has been observed that a small number of nucleotides in the internal loop E of the 5S rRNA are notable in distinguishing the bacterial 5S from its eukaryotic and archaeal counterparts [101].

The precise role of 5S rRNA in ribosome function is not fully understood. It has been suggested to play a role as a signal transducer between the peptidyltransferase centre and domain II responsible for translocation [31], or as a determinant of large-subunit stability [56]. It is, however, essential for protein biosynthesis: in *E. coli*, the deletion of more than one copy of the 5S rRNA is shown to impair growth rate [6].

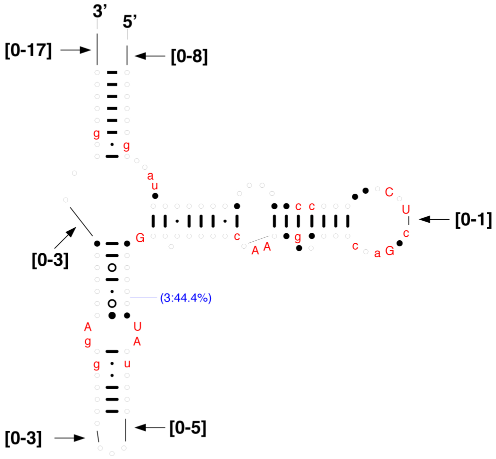

**Fig.** A.1: **The conserved secondary structure of the 5S ribosomal RNA.** Positions in the 5S ribosomal RNA with a nucleotide in more than 95% of the sequences are shown superimposed onto the *E. coli* secondary structure. Phylogenetic conservation is derived from the comparative analysis of 682 sequences (Reproduced from CRW).

## A.2 Data

We performed our Mode-1 HOSVD analysis described previously on an alignment of 242 5S rRNA sequences from the CRW (Table **??**) [16]. The taxonomy of the sequences in this alignment is shown in Table B.3.

## A.3 Results

The four most significant eigenpositions and corresponding eigenorganisms capture ∼79% of the nucleotide frequency information in the alignment . The most significant eigenposition, which captures ∼63% of the nucleotide frequency informationis approximately invariant across the organisms. The remaining significant eigenpositions uncovered identify the dominant taxonomic groups among the organisms and their relations of similarity and dissimilarity.

The second most significant eigenposition (Figure A.2 (*a*)) differentiates the Bacteria from the Eukarya, as indicated by the color bar (Table A.1). The third (*b*) distinguishes between the Archaea and the Actinobacteria, the largest Bacterial subgroup in the alignment. The fourth (*c*) distinguishes the Fungi/Metazoa and the Viridiplantae, the two largest Eukaryotic subgroups in this alignment.

The results described here are qualitatively similar to those obtained from the analysis of the 16S and 23S rRNA sequence alignments. We did not detect significant enrichments of structure motifs among the most correlated and anticorrelated positions in the corresponding eigenorganisms, conceivably due to the small number of positions in the alignment.

| **5S Eigenposition** | **Correlated** | | | | **Anticorrelated** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Group** | $n$ | $N$ | **p-value** | **Group** | $n$ | $N$ | **p-value** |
| 2 | Bacteria | 50 | 83 | $6.8 \times 10^{-30}$ | Eukarya | 50 | 131 | $2.2 \times 10^{-16}$ |
| 3 | Archaea | 28 | 28 | $1.5 \times 10^{-26}$ | Proteobacteria | 47 | 56 | $3.6 \times 10^{-37}$ |
| 4 | Fungi/Metazoa | 48 | 83 | $1.8 \times 10^{-25}$ | Viridiplantae | 24 | 24 | $1.5 \times 10^{-19}$ |

Table A.1: **Probabilistic significance of the enrichment of the $k$=50 organisms in the 5S rRNA.**

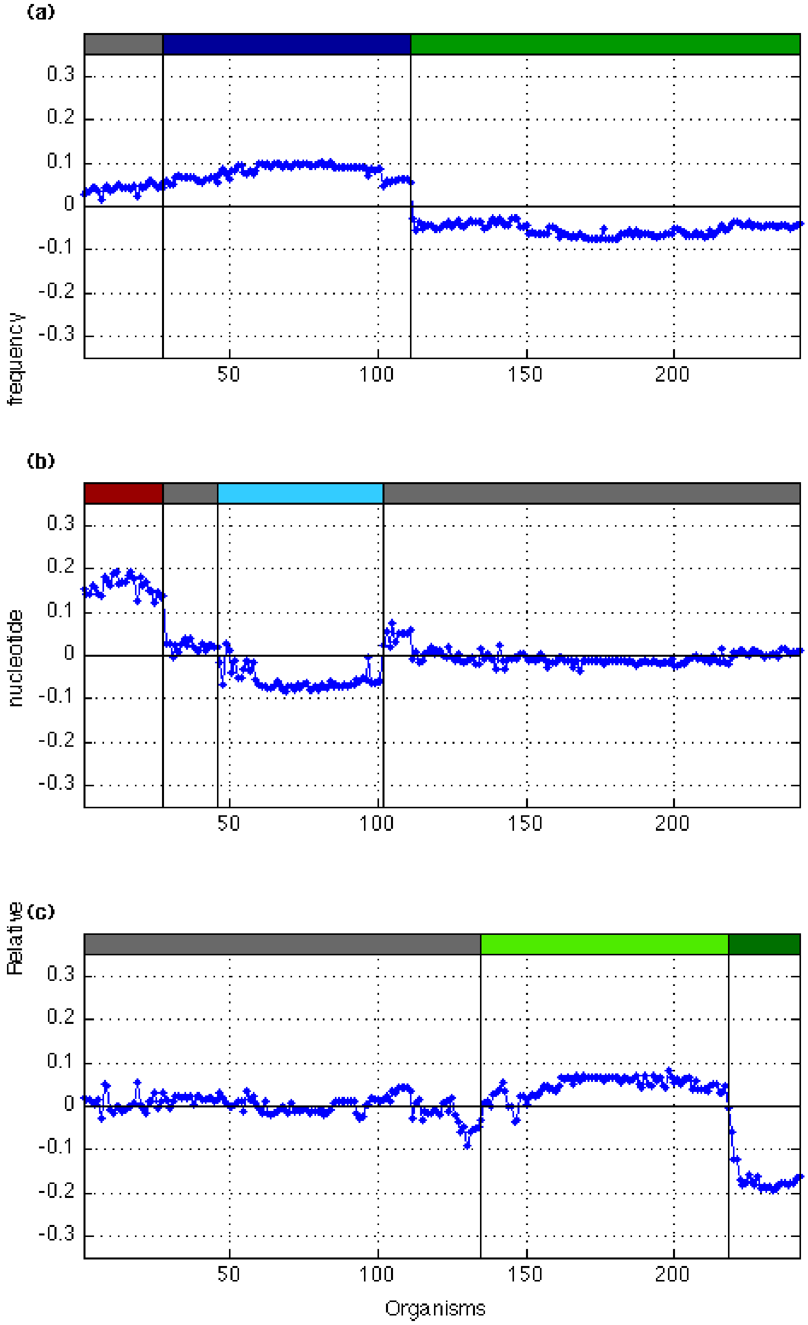

**Fig.** A.2: **Significant 5S eigenpositions.** Line-joined graphs of the (a) second, (b) third, and (c) fourth 5S eigenpositions, i.e., patterns of nucleotide frequency across the organisms, and their correlation with the taxonomic groups in the 5S alignment, classified according to the top six hierarchical levels of the NCBI Taxonomy Browser [89].

# Appendix B

# Taxonomy of sequences in the rRNA Datasets

Tables B.1, B.2, and B.3 list the organisms in the 16S, 23S, and 5S datasets respectively, along with their taxonomic groups. This data was retrieved from the NCBI Taxonomy Browser [89]. Although only the taxonomic groups from the three top hierarchical levels are shown, six levels were used for the calculation of enrichment of taxonomic groups among eigenpositions.

Table B.1: **Organisms in the 339-sequence 16S rRNA dataset, and their their associated taxonomic groups.**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 1 | Aeropyrum pernix | Archaea | Crenarchaeota | Thermoprotei |
| 2 | Pyrodictium occultum | Archaea | Crenarchaeota | Thermoprotei |
| 3 | Sulfolobus acidocaldarius | Archaea | Crenarchaeota | Thermoprotei |
| 4 | Sulfolobus solfataricus | Archaea | Crenarchaeota | Thermoprotei |
| 5 | Thermoproteus tenax. | Archaea | Crenarchaeota | Thermoprotei |
| 6 | Archaeoglobus fulgidus | Archaea | Euryarchaeota | Archaeoglobi |
| 7 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 8 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 9 | Haloferax volcanii | Archaea | Euryarchaeota | Halobacteria |
| 10 | Natronobacterium innermongoliae. | Archaea | Euryarchaeota | Halobacteria |
| 11 | Natronobacterium bangense | Archaea | Euryarchaeota | Halobacteria |
| 12 | Methanobacterium formicicum | Archaea | Euryarchaeota | Methanobacteria |
| 13 | Methanobacterium thermoautotrophicum | Archaea | Euryarchaeota | Methanobacteria |
| 14 | Methanobacterium thermoautotrophicum | Archaea | Euryarchaeota | Methanobacteria |
| 15 | Methanococcus vannielii | Archaea | Euryarchaeota | Methanococci |
| 16 | Methanospirillum hungatei | Archaea | Euryarchaeota | Methanomicrobia |
| 17 | Pyrococcus abyssi. | Archaea | Euryarchaeota | Thermococci |
| 18 | Pyrococcus furiosus | Archaea | Euryarchaeota | Thermococci |
| 19 | Pyrococcus horikoshii | Archaea | Euryarchaeota | Thermococci |
| 20 | Thermococcus celer | Archaea | Euryarchaeota | Thermococci |
| 21 | Thermoplasma acidophilum | Archaea | Euryarchaeota | Thermoplasmata |
| 22 | Gluconacetobacter liquefaciens | Bacteria | Proteobacteria | Alphaproteobacteria |
| 23 | Bartonella vinsonii | Bacteria | Proteobacteria | Alphaproteobacteria |
| 24 | Bartonella henselae | Bacteria | Proteobacteria | Alphaproteobacteria |
| 25 | Bartonella quintana | Bacteria | Proteobacteria | Alphaproteobacteria |
| 26 | Bradyrhizobium japonicum | Bacteria | Proteobacteria | Alphaproteobacteria |
| 27 | Brucella melitensis | Bacteria | Proteobacteria | Alphaproteobacteria |
| 28 | Azorhizobium caulinodans | Bacteria | Proteobacteria | Alphaproteobacteria |
| 29 | Blastobacter sp. | Bacteria | Proteobacteria | Alphaproteobacteria |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 30 | Mesorhizobium loti | Bacteria | Proteobacteria | Alphaproteobacteria |
| 31 | Agrobacterium tumefaciens | Bacteria | Proteobacteria | Alphaproteobacteria |
| 32 | Rhodobium orientis | Bacteria | Proteobacteria | Alphaproteobacteria |
| 33 | Rickettsia prowazekii | Bacteria | Proteobacteria | Alphaproteobacteria |
| 34 | Rickettsia rickettsii | Bacteria | Proteobacteria | Alphaproteobacteria |
| 35 | Rickettsia prowazekii | Bacteria | Proteobacteria | Alphaproteobacteria |
| 36 | Rickettsia bellii | Bacteria | Proteobacteria | Alphaproteobacteria |
| 37 | Bordetella parapertussis | Bacteria | Proteobacteria | Betaproteobacteria |
| 38 | Bordetella pertussis | Bacteria | Proteobacteria | Betaproteobacteria |
| 39 | Comamonas testosteroni | Bacteria | Proteobacteria | Betaproteobacteria |
| 40 | Lautropia mirabilis | Bacteria | Proteobacteria | Betaproteobacteria |
| 41 | Neisseria gonorrhoeae | Bacteria | Proteobacteria | Betaproteobacteria |
| 42 | Neisseria meningitidis | Bacteria | Proteobacteria | Betaproteobacteria |
| 43 | Neisseria meningitidis | Bacteria | Proteobacteria | Betaproteobacteria |
| 44 | Aeromonas salmonicida | Bacteria | Proteobacteria | Gammaproteobacteria |
| 45 | Dichelobacter nodosus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 46 | Edwardsiella tarda | Bacteria | Proteobacteria | Gammaproteobacteria |
| 47 | Escherichia coli | Bacteria | Proteobacteria | Gammaproteobacteria |
| 48 | Escherichia coli O157 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 49 | Escherichia coli O157 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 50 | Plesiomonas shigelloides | Bacteria | Proteobacteria | Gammaproteobacteria |
| 51 | Proteus vulgaris | Bacteria | Proteobacteria | Gammaproteobacteria |
| 52 | Salmonella typhimurium | Bacteria | Proteobacteria | Gammaproteobacteria |
| 53 | Shigella dysenteriae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 54 | Yersinia pestis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 55 | Yersinia pseudotuberculosis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 56 | Chromohalobacter marismortui | Bacteria | Proteobacteria | Gammaproteobacteria |
| 57 | Haemophilus influenzae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 58 | Haemophilus influenzae | Bacteria | Proteobacteria | Gammaproteobacteria |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 59 | Haemophilus influenzae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 60 | Haemophilus influenzae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 61 | Pasteurella multocida. | Bacteria | Proteobacteria | Gammaproteobacteria |
| 62 | Psychrobacter pacificensis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 63 | Pseudomonas aeruginosa | Bacteria | Proteobacteria | Gammaproteobacteria |
| 64 | Pseudomonas putida | Bacteria | Proteobacteria | Gammaproteobacteria |
| 65 | Francisella tularensis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 66 | Beggiatoa sp. | Bacteria | Proteobacteria | Gammaproteobacteria |
| 67 | Vibrio cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 68 | Vibrio cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 69 | Vibrio Cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 70 | Vibrio cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 71 | Vibrio cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 72 | Vibrio cholerae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 73 | Xanthomonas albilineans | Bacteria | Proteobacteria | Gammaproteobacteria |
| 74 | Xanthomonas campestris | Bacteria | Proteobacteria | Gammaproteobacteria |
| 75 | Xylella fastidiosa | Bacteria | Proteobacteria | Gammaproteobacteria |
| 76 | Desulfovibrio desulfuricans | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 77 | Myxococcus xanthus | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 78 | Campylobacter jejuni | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 79 | Campylobacter jejuni. | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 80 | Campylobacter sputorum | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 81 | Helicobacter pylori | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 82 | Helicobacter pylori 26695 | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 83 | Helicobacter pylori J99 | Bacteria | Proteobacteria | delta/epsilon subdivisions |
| 84 | Pseudomonas sp. | Bacteria | Proteobacteria | unclassified Proteobacteria |
| 85 | Actinomyces israelii | Bacteria | Actinobacteria | Actinobacteria (class) |
| 86 | Corynebacterium diphtheriae | Bacteria | Actinobacteria | Actinobacteria (class) |
| 87 | Mycobacterium avium | Bacteria | Actinobacteria | Actinobacteria (class) |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 88 | Mycobacterium leprae | Bacteria | Actinobacteria | Actinobacteria (class) |
| 89 | Mycobacterium leprae | Bacteria | Actinobacteria | Actinobacteria (class) |
| 90 | Mycobacterium tuberculosis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 91 | Mycobacterium tuberculosis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 92 | Mycobacterium tuberculosis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 93 | Mycobacterium tuberculosis CDC1551 | Bacteria | Actinobacteria | Actinobacteria (class) |
| 94 | Nocardia asteroides | Bacteria | Actinobacteria | Actinobacteria (class) |
| 95 | Rhodococcus erythropolis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 96 | Frankia sp. | Bacteria | Actinobacteria | Actinobacteria (class) |
| 97 | Arthrobacter globiformis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 98 | Streptomyces acidiscabies | Bacteria | Actinobacteria | Actinobacteria (class) |
| 99 | Streptomyces albidoflavus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 100 | Streptomyces ambofaciens | Bacteria | Actinobacteria | Actinobacteria (class) |
| 101 | Streptomyces bikiniensis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 102 | Streptomyces bluensis. | Bacteria | Actinobacteria | Actinobacteria (class) |
| 103 | Streptomyces bottropensis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 104 | Streptomyces caelestis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 105 | Streptomyces diastatochromogenes | Bacteria | Actinobacteria | Actinobacteria (class) |
| 106 | Streptomyces espinosus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 107 | Streptomyces eurythermus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 108 | Streptomyces felleus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 109 | Streptomyces galbus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 110 | Streptomyces glaucescens. | Bacteria | Actinobacteria | Actinobacteria (class) |
| 111 | Streptomyces gougerotii | Bacteria | Actinobacteria | Actinobacteria (class) |
| 112 | Streptomyces griseus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 113 | Streptomyces hygroscopicus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 114 | Streptomyces intermedius | Bacteria | Actinobacteria | Actinobacteria (class) |
| 115 | Streptomyces limosus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 116 | Streptomyces lincolnensis | Bacteria | Actinobacteria | Actinobacteria (class) |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 117 | Streptomyces macrosporus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 118 | Streptomyces mashuensis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 119 | Streptomyces megasporus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 120 | Streptomyces neyagawaensis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 121 | Streptomyces nodosus. | Bacteria | Actinobacteria | Actinobacteria (class) |
| 122 | Streptomyces odorifer | Bacteria | Actinobacteria | Actinobacteria (class) |
| 123 | Streptomyces ornatus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 124 | Streptomyces pseudogriseolus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 125 | Streptomyces rimosus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 126 | Streptomyces rutgersensis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 127 | Streptomyces sampsonii | Bacteria | Actinobacteria | Actinobacteria (class) |
| 128 | Streptomyces scabies | Bacteria | Actinobacteria | Actinobacteria (class) |
| 129 | Streptomyces setonii | Bacteria | Actinobacteria | Actinobacteria (class) |
| 130 | Streptomyces sp. | Bacteria | Actinobacteria | Actinobacteria (class) |
| 131 | Streptomyces subrutilus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 132 | Streptomyces tendae | Bacteria | Actinobacteria | Actinobacteria (class) |
| 133 | Streptomyces thermodiastaticus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 134 | Streptomyces thermolineatus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 135 | Streptomyces thermoviolaceus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 136 | Streptomyces thermonitrificans | Bacteria | Actinobacteria | Actinobacteria (class) |
| 137 | Streptomyces thermovulgaris | Bacteria | Actinobacteria | Actinobacteria (class) |
| 138 | Bacillus cereus | Bacteria | Firmicutes | Bacilli |
| 139 | Bacillus halodurans | Bacteria | Firmicutes | Bacilli |
| 140 | Bacillus subtilis | Bacteria | Firmicutes | Bacilli |
| 141 | Staphylococcus aureus | Bacteria | Firmicutes | Bacilli |
| 142 | Staphylococcus aureus | Bacteria | Firmicutes | Bacilli |
| 143 | Enterococcus faecalis | Bacteria | Firmicutes | Bacilli |
| 144 | Enterococcus faecium | Bacteria | Firmicutes | Bacilli |
| 145 | Lactococcus lactis subsp. lactis | Bacteria | Firmicutes | Bacilli |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 146 | Streptococcus pneumoniae | Bacteria | Firmicutes | Bacilli |
| 147 | Streptococcus pyogenes | Bacteria | Firmicutes | Bacilli |
| 148 | Clostridium botulinum | Bacteria | Firmicutes | Clostridia |
| 149 | Clostridium perfringens | Bacteria | Firmicutes | Clostridia |
| 150 | Clostridium tetani | Bacteria | Firmicutes | Clostridia |
| 151 | Eubacterium brachy | Bacteria | Firmicutes | Clostridia |
| 152 | Heliobacterium chlorum | Bacteria | Firmicutes | Clostridia |
| 153 | Epulopiscium sp. | Bacteria | Firmicutes | Clostridia |
| 154 | Mycoplasma capricolum | Bacteria | Firmicutes | Mollicutes |
| 155 | Mycoplasma gallisepticum | Bacteria | Firmicutes | Mollicutes |
| 156 | Mycoplasma hyopneumoniae | Bacteria | Firmicutes | Mollicutes |
| 157 | Ureaplasma urealyticum | Bacteria | Firmicutes | Mollicutes |
| 158 | Gemmata obscuriglobus | Bacteria | Planctomycetes | Planctomycetacia |
| 159 | Planctomyces sp. | Bacteria | Planctomycetes | Planctomycetacia |
| 160 | Brachyspira hyodysenteriae | Bacteria | Spirochaetes | Spirochaetes (class) |
| 161 | Leptonema illini | Bacteria | Spirochaetes | Spirochaetes (class) |
| 162 | Leptospira borgpetersenii | Bacteria | Spirochaetes | Spirochaetes (class) |
| 163 | Borrelia burgdorferi. | Bacteria | Spirochaetes | Spirochaetes (class) |
| 164 | Borrelia burgdorferi | Bacteria | Spirochaetes | Spirochaetes (class) |
| 165 | Borrelia hermsii | Bacteria | Spirochaetes | Spirochaetes (class) |
| 166 | Brevinema andersonii | Bacteria | Spirochaetes | Spirochaetes (class) |
| 167 | Treponema pallidum | Bacteria | Spirochaetes | Spirochaetes (class) |
| 168 | Geotoga subterranea | Bacteria | Thermotogae | Thermotogae (class) |
| 169 | Petrotoga miotherma | Bacteria | Thermotogae | Thermotogae (class) |
| 170 | Thermotoga maritima | Bacteria | Thermotogae | Thermotogae (class) |
| 171 | Thermotoga maritima | Bacteria | Thermotogae | Thermotogae (class) |
| 172 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 173 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 174 | Thermus aquaticus | Bacteria | Deinococcus-Thermus | Deinococci |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 175 | Thermus thermophilus | Bacteria | Deinococcus-Thermus | Deinococci |
| 176 | Bacteroides fragilis | Bacteria | Bacteroidetes/Chlorobi group | Bacteroidetes |
| 177 | Porphyromonas gingivalis | Bacteria | Bacteroidetes/Chlorobi group | Bacteroidetes |
| 178 | Chlorobium vibrioforme | Bacteria | Bacteroidetes/Chlorobi group | Chlorobi |
| 179 | Chlamydia trachomatis | Bacteria | Chlamydiae/Verrucomicrobia group | Chlamydiae |
| 180 | Chlamydia trachomatis | Bacteria | Chlamydiae/Verrucomicrobia group | Chlamydiae |
| 181 | Chlamydophila pneumoniae | Bacteria | Chlamydiae/Verrucomicrobia group | Chlamydiae |
| 182 | Chlamydophila pneumoniae J138 | Bacteria | Chlamydiae/Verrucomicrobia group | Chlamydiae |
| 183 | Microcystis aeruginosa | Bacteria | Cyanobacteria | Chroococcales |
| 184 | Synechococcus sp. | Bacteria | Cyanobacteria | Chroococcales |
| 185 | Synechocystis PCC6803 | Bacteria | Cyanobacteria | Chroococcales |
| 186 | Nostoc muscorum | Bacteria | Cyanobacteria | Nostocales |
| 187 | Oscillatoria agardhii | Bacteria | Cyanobacteria | Oscillatoriales |
| 188 | Pleurocapsa sp. | Bacteria | Cyanobacteria | Pleurocapsales |
| 189 | Chlorogloeopsis sp. | Bacteria | Cyanobacteria | Stigonematales |
| 190 | Acidobacterium capsulatum | Bacteria | Fibrobacteres/Acidobacteria group | Acidobacteria |
| 191 | Holophaga foetida | Bacteria | Fibrobacteres/Acidobacteria group | Acidobacteria |
| 192 | Aquifex aeolicus. | Bacteria | Aquificae | Aquificae (class) |
| 193 | Deferribacter thermophilus | Bacteria | Deferribacteres | Deferribacteres (class) |
| 194 | Fusobacterium necrophorum | Bacteria | Fusobacteria | Fusobacteria (class) |
| 195 | Streptobacillus moniliformis | Bacteria | Fusobacteria | Fusobacteria (class) |
| 196 | Thermomicrobium roseum | Bacteria | Chloroflexi | Thermomicrobia (class) |
| 197 | Acanthamoeba castellanii | Eukaryota | Acanthamoebidae | Acanthamoeba |
| 198 | Plasmodium falciparum (A stage) | Eukaryota | Alveolata | Apicomplexa |
| 199 | Plasmodium vivax | Eukaryota | Alveolata | Apicomplexa |
| 200 | Babesia bigemina | Eukaryota | Alveolata | Apicomplexa |
| 201 | Babesia canis | Eukaryota | Alveolata | Apicomplexa |
| 202 | Euplotes aediculatus | Eukaryota | Alveolata | Ciliophora |
| 203 | Onychodromus quadricornutus | Eukaryota | Alveolata | Ciliophora |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 204 | Paraurostyla weissei | Eukaryota | Alveolata | Ciliophora |
| 205 | Engelmanniella mobilis | Eukaryota | Alveolata | Ciliophora |
| 206 | Cyrtohymena citrina | Eukaryota | Alveolata | Ciliophora |
| 207 | Gastrostyla steinei | Eukaryota | Alveolata | Ciliophora |
| 208 | Oxytricha granulifera | Eukaryota | Alveolata | Ciliophora |
| 209 | Oxytricha granulifera | Eukaryota | Alveolata | Ciliophora |
| 210 | Oxytricha longa. | Eukaryota | Alveolata | Ciliophora |
| 211 | Pleurotricha lanceolota. | Eukaryota | Alveolata | Ciliophora |
| 212 | Stylonychia lemnae | Eukaryota | Alveolata | Ciliophora |
| 213 | Stylonychia mytilus | Eukaryota | Alveolata | Ciliophora |
| 214 | Paruroleptus lepisma | Eukaryota | Alveolata | Ciliophora |
| 215 | Uroleptus gallina. | Eukaryota | Alveolata | Ciliophora |
| 216 | Uroleptus pisces. | Eukaryota | Alveolata | Ciliophora |
| 217 | Urostyla grandis. | Eukaryota | Alveolata | Ciliophora |
| 218 | Alexandrium fundyense | Eukaryota | Alveolata | Dinophyceae |
| 219 | Euglypha rotunda | Eukaryota | Cercozoa | Euglyphida |
| 220 | Paulinella chromatophora | Eukaryota | Cercozoa | Euglyphida |
| 221 | Cyanophora paradoxa | Eukaryota | Glaucocystophyceae | Cyanophoraceae |
| 222 | Glaucocystis nostochinearum | Eukaryota | Glaucocystophyceae | Glaucocystales |
| 223 | Gloeochaete wittrockiana. | Eukaryota | Glaucocystophyceae | Gloeochaetales |
| 224 | Balamuthia mandrillaris. | Eukaryota | Lobosea | Leptomyxida |
| 225 | Phreatamoeba balamuthi | Eukaryota | Pelobiontida | Mastigamoebidae |
| 226 | Acanthocoepsis unguiculata | Eukaryota | Choanoflagellida | Acanthoecidae |
| 227 | Diaphanoeca grandis | Eukaryota | Choanoflagellida | Acanthoecidae |
| 228 | Aspergillus flavus | Eukaryota | Fungi | Dikarya |
| 229 | Coccidiodes immitis | Eukaryota | Fungi | Dikarya |
| 230 | Neurospora crassa | Eukaryota | Fungi | Dikarya |
| 231 | Saccharomyces cerevisiae | Eukaryota | Fungi | Dikarya |
| 232 | Candida albicans | Eukaryota | Fungi | Dikarya |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 233 | Pneumocystis carinii | Eukaryota | Fungi | Dikarya |
| 234 | Filobasidiella neoformans serotype D | Eukaryota | Fungi | Dikarya |
| 235 | Ustilago maydis | Eukaryota | Fungi | Dikarya |
| 236 | Allomyces macrogynus | Eukaryota | Fungi | Blastocladiomycota |
| 237 | Blastocladiella emersonii | Eukaryota | Fungi | Blastocladiomycota |
| 238 | Smittium culisetae | Eukaryota | Fungi | Fungi incertae sedis |
| 239 | Absidia corymbifera | Eukaryota | Fungi | Fungi incertae sedis |
| 240 | Absidia corymbifera | Eukaryota | Fungi | Fungi incertae sedis |
| 241 | Mucor circinelloides f. lusitanicus | Eukaryota | Fungi | Fungi incertae sedis |
| 242 | Mucor racemosus | Eukaryota | Fungi | Fungi incertae sedis |
| 243 | Rhizopus arrhizus | Eukaryota | Fungi | Fungi incertae sedis |
| 244 | Culicosporella lunata | Eukaryota | Fungi | Microsporidia |
| 245 | Enterocytozoon bieneusi. | Eukaryota | Fungi | Microsporidia |
| 246 | Bacillidium sp. | Eukaryota | Fungi | Microsporidia |
| 247 | Ichthyosporidium sp | Eukaryota | Fungi | Microsporidia |
| 248 | Nosema algerae. | Eukaryota | Fungi | Microsporidia |
| 249 | Nosema apis. | Eukaryota | Fungi | Microsporidia |
| 250 | Nosema bombycis | Eukaryota | Fungi | Microsporidia |
| 251 | Nosema necatrix | Eukaryota | Fungi | Microsporidia |
| 252 | Vittaforma corneae | Eukaryota | Fungi | Microsporidia |
| 253 | Spraguea lophii. | Eukaryota | Fungi | Microsporidia |
| 254 | Encephalitozoon cuniculi | Eukaryota | Fungi | Microsporidia |
| 255 | Encephalitozoon hellem | Eukaryota | Fungi | Microsporidia |
| 256 | Encephalitozoon sp | Eukaryota | Fungi | Microsporidia |
| 257 | Microgemma sp. | Eukaryota | Fungi | Microsporidia |
| 258 | Endoreticulatus schubergi | Eukaryota | Fungi | Microsporidia |
| 259 | Amblyospora connecticus | Eukaryota | Fungi | Microsporidia |
| 260 | Amblyospora sp. | Eukaryota | Fungi | Microsporidia |
| 261 | Parathelohania anophelis | Eukaryota | Fungi | Microsporidia |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 262 | Vairimorpha imperfecta | Eukaryota | Fungi | Microsporidia |
| 263 | Vairimorpha necatrix | Eukaryota | Fungi | Microsporidia |
| 264 | Loma acerinae | Eukaryota | Fungi | Microsporidia |
| 265 | Pleistophora hippoglossoideos. | Eukaryota | Fungi | Microsporidia |
| 266 | Pleistophora mirandellae. | Eukaryota | Fungi | Microsporidia |
| 267 | Pleistophora sp. LS. | Eukaryota | Fungi | Microsporidia |
| 268 | Pleistophora sp. ATCC 50040 | Eukaryota | Fungi | Microsporidia |
| 269 | Trachipleistophora hominis | Eukaryota | Fungi | Microsporidia |
| 270 | Vavraia culicis. | Eukaryota | Fungi | Microsporidia |
| 271 | Polydispyrenia simulii | Eukaryota | Fungi | Microsporidia |
| 272 | Thelohania solenopsae | Eukaryota | Fungi | Microsporidia |
| 273 | Janacekia debaisieuxi | Eukaryota | Fungi | Microsporidia |
| 274 | Ameson michaelis | Eukaryota | Fungi | Microsporidia |
| 275 | Antonospora scoticae | Eukaryota | Fungi | Microsporidia |
| 276 | Edhazardia aedis | Eukaryota | Fungi | Microsporidia |
| 277 | Microsporidium 5786. | Eukaryota | Fungi | Microsporidia |
| 278 | Microsporidium prosopium | Eukaryota | Fungi | Microsporidia |
| 279 | Visvesvaria acridophagus. | Eukaryota | Fungi | Microsporidia |
| 280 | Dermocystidium sp. | Eukaryota | Fungi/Metazoa incertae sedis | Ichthyosporea |
| 281 | Psorospermium haeckelii | Eukaryota | Fungi/Metazoa incertae sedis | Ichthyosporea |
| 282 | Echinococcus granulosus | Eukaryota | Metazoa | Eumetazoa |
| 283 | Oryctolagus cuniculus | Eukaryota | Metazoa | Eumetazoa |
| 284 | Homo sapiens | Eukaryota | Metazoa | Eumetazoa |
| 285 | Mus musculus | Eukaryota | Metazoa | Eumetazoa |
| 286 | Xenopus borealis | Eukaryota | Metazoa | Eumetazoa |
| 287 | Xenopus laevis | Eukaryota | Metazoa | Eumetazoa |
| 288 | Mytilus edulis | Eukaryota | Metazoa | Eumetazoa |
| 289 | Placopecten magellanicus | Eukaryota | Metazoa | Eumetazoa |
| 290 | Androctonus australis | Eukaryota | Metazoa | Eumetazoa |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 291 | Artemia salina | Eukaryota | Metazoa | Eumetazoa |
| 292 | Drosophila melanogaster | Eukaryota | Metazoa | Eumetazoa |
| 293 | Okanagana utahensis | Eukaryota | Metazoa | Eumetazoa |
| 294 | Bangia sp. (Alaska/AK) | Eukaryota | Rhodophyta | Bangiophyceae |
| 295 | Bangia sp. (Virgin Islands/VIS7) | Eukaryota | Rhodophyta | Bangiophyceae |
| 296 | Compsopogon coeruleus. | Eukaryota | Rhodophyta | Bangiophyceae |
| 297 | Erythrotrichia carnea | Eukaryota | Rhodophyta | Bangiophyceae |
| 298 | Porphyridum aerugineum | Eukaryota | Rhodophyta | Bangiophyceae |
| 299 | Rhodella maculata | Eukaryota | Rhodophyta | Bangiophyceae |
| 300 | Audouinella dasyae | Eukaryota | Rhodophyta | Florideophyceae |
| 301 | Audouinella hermannii | Eukaryota | Rhodophyta | Florideophyceae |
| 302 | Ahnfeltia plicata | Eukaryota | Rhodophyta | Florideophyceae |
| 303 | Batrachospermum gelatinosum | Eukaryota | Rhodophyta | Florideophyceae |
| 304 | Batrachospermum macrosporum | Eukaryota | Rhodophyta | Florideophyceae |
| 305 | Nemalionopsis tortuosa | Eukaryota | Rhodophyta | Florideophyceae |
| 306 | Thorea violacea. | Eukaryota | Rhodophyta | Florideophyceae |
| 307 | Bonnemaisonia hamifera | Eukaryota | Rhodophyta | Florideophyceae |
| 308 | Ceramium rubrum | Eukaryota | Rhodophyta | Florideophyceae |
| 309 | Bostrychia moritziana. | Eukaryota | Rhodophyta | Florideophyceae |
| 310 | Corallina officinalis | Eukaryota | Rhodophyta | Florideophyceae |
| 311 | Gelidium vagum | Eukaryota | Rhodophyta | Florideophyceae |
| 312 | Chondrus crispus | Eukaryota | Rhodophyta | Florideophyceae |
| 313 | Gracilariopsis sp. England-1 | Eukaryota | Rhodophyta | Florideophyceae |
| 314 | Halymenia plana. | Eukaryota | Rhodophyta | Florideophyceae |
| 315 | Hildenbrandia rubra | Eukaryota | Rhodophyta | Florideophyceae |
| 316 | Nemalion helminthoides | Eukaryota | Rhodophyta | Florideophyceae |
| 317 | Plocamiocolax pulvinata | Eukaryota | Rhodophyta | Florideophyceae |
| 318 | Rhodogorgon carriebowensis | Eukaryota | Rhodophyta | Florideophyceae |
| 319 | Rhodymenia leptophylla | Eukaryota | Rhodophyta | Florideophyceae |

Continued on next page

**Table B.1 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 320 | Chlorella luteoviridis (B) | Eukaryota | Viridiplantae | Chlorophyta |
| 321 | Oryza sativa. | Eukaryota | Viridiplantae | Streptophyta |
| 322 | Solanum tuberosum | Eukaryota | Viridiplantae | Streptophyta |
| 323 | Fragaria x ananassa | Eukaryota | Viridiplantae | Streptophyta |
| 324 | Sinapis alba | Eukaryota | Viridiplantae | Streptophyta |
| 325 | Arceuthobium verticilliflorum | Eukaryota | Viridiplantae | Streptophyta |
| 326 | Genicularia spirotaenia | Eukaryota | Viridiplantae | Streptophyta |
| 327 | Aulacoseira ambigua. | Eukaryota | stramenopiles | Bacillariophyta |
| 328 | Corethron criophilum. | Eukaryota | stramenopiles | Bacillariophyta |
| 329 | Coscinodiscus radiatus | Eukaryota | stramenopiles | Bacillariophyta |
| 330 | Melosira varians | Eukaryota | stramenopiles | Bacillariophyta |
| 331 | Stephanopyxis cf. broschii | Eukaryota | stramenopiles | Bacillariophyta |
| 332 | Cymatosira belgica | Eukaryota | stramenopiles | Bacillariophyta |
| 333 | Lauderia borealis. | Eukaryota | stramenopiles | Bacillariophyta |
| 334 | Ditylum brightwelli. | Eukaryota | stramenopiles | Bacillariophyta |
| 335 | Skeletonema costatum. | Eukaryota | stramenopiles | Bacillariophyta |
| 336 | Thalassiosira eccentrica | Eukaryota | stramenopiles | Bacillariophyta |
| 337 | Fragilaria striatula | Eukaryota | stramenopiles | Bacillariophyta |
| 338 | Rhaphoneis belgicae | Eukaryota | stramenopiles | Bacillariophyta |
| 339 | Labyrinthuloides minuta | Eukaryota | stramenopiles | Labyrinthulida |

Table B.2: **Organisms in the 75-sequence 23S rRNA dataset, and their associated taxonomic groups.**

| **No.** | **Organism name** | **Taxonomy** | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 1 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 2 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 3 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 4 | Methanothermobacter thermoautotrophicus | Archaea | Euryarchaeota | Methanobacteria |
| 5 | Methanococcus jannaschii | Archaea | Euryarchaeota | Methanococci |
| 6 | Thermococcus celer | Archaea | Euryarchaeota | Thermococci |
| 7 | Micrococcus luteus | Bacteria | Actinobacteria | Actinobacteria(class) |
| 8 | Mycoplasma leprae | Bacteria | Actinobacteria | Actinobacteria(class) |
| 9 | Mycobacterium tuberculosis | Bacteria | Actinobacteria | Actinobacteria (class) |
| 10 | Streptomyces ambofaciens | Bacteria | Actinobacteria | Actinobacteria (class) |
| 11 | Streptomyces carnosus | Bacteria | Actinobacteria | Actinobacteria (class) |
| 12 | Tropheryma whippelii | Bacteria | Aquificae | Aquificae (class) |
| 13 | Aquifex aeolicus | Bacteria | Chlamydiae/Verrucomicrobiagroup | Chlamydiae |
| 14 | Chlamydophila psittaci | Bacteria | Chlamydiae/Verrucomicrobiagroup | Chlamydiae |
| 15 | Chlamydia suis | Bacteria | Chlamydiae/Verrucomicrobiagroup | Chlamydiae |
| 16 | Chlamydia trachomatis | Bacteria | Chlamydiae/Verrucomicrobiagroup | Chlamydiae |
| 17 | Parachlamydia acanthamoebae | Bacteria | Chlamydiae/Verrucomicrobiagroup | Chlamydiae |
| 18 | Simkania negevensis | Bacteria | Deinococcus-Thermus | Deinococci |
| 19 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 20 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 21 | Thermus thermophilus | Bacteria | Firmicutes | Bacilli |
| 22 | Bacillus anthracis | Bacteria | Firmicutes | Bacilli |
| 23 | Bacillus subtilis | Bacteria | Firmicutes | Bacilli |
| 24 | Enterococcus faecium | Bacteria | Firmicutes | Bacilli |
| 25 | Lactobacillus delbrueckii | Bacteria | Firmicutes | Bacilli |
| 26 | Lactococcus lactis | Bacteria | Firmicutes | Bacilli |
| 27 | Listeria monocytogenes | Bacteria | Firmicutes | Bacilli |
| 28 | Listeria monocytogenes | Bacteria | Firmicutes | Bacilli |
| 29 | Staphylococcus aureus | Bacteria | Firmicutes | Clostridia |
| 30 | Clostridium botulinum | Bacteria | Firmicutes | Clostridia |

Continued on next page

**Table B.2 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 31 | Clostridium botulinum | Bacteria | Firmicutes | Clostridia |
| 32 | Clostridium botulinum | Bacteria | Firmicutes | Clostridia |
| 33 | Clostridium botulinum | Bacteria | Firmicutes | Erysipelotrichi |
| 34 | Erysipelothrix rhusiopathiae | Bacteria | Proteobacteria | Alphaproteobacteria |
| 35 | Acetobacter calcoaceticus | Bacteria | Proteobacteria | Alphaproteobacteria |
| 36 | Bartonella bacilliformis | Bacteria | Proteobacteria | Alphaproteobacteria |
| 37 | Rhodopseudomonas palustris | Bacteria | Proteobacteria | Alphaproteobacteria |
| 38 | Rickettsia prowazekii | Bacteria | Proteobacteria | Betaproteobacteria |
| 39 | Rickettsia rickettsii | Bacteria | Proteobacteria | Betaproteobacteria |
| 40 | Bordetella bronchiseptica | Bacteria | Proteobacteria | Betaproteobacteria |
| 41 | Burkholderia mallei | Bacteria | Proteobacteria | Betaproteobacteria |
| 42 | Bordetella pertussis | Bacteria | Proteobacteria | Betaproteobacteria |
| 43 | Burkholderia pseudomallei | Bacteria | Proteobacteria | Betaproteobacteria |
| 44 | Burkholderia cepacia | Bacteria | Proteobacteria | Betaproteobacteria |
| 45 | Nesseria gonorrhoeae | Bacteria | Proteobacteria | delta/epsilonsubdivisions |
| 46 | Neisseria meningitidis | Bacteria | Proteobacteria | delta/epsilonsubdivisions |
| 47 | Campylobacter jejuni | Bacteria | Proteobacteria | Gammaproteobacteria |
| 48 | Helicobacter pylori | Bacteria | Proteobacteria | Gammaproteobacteria |
| 49 | Aeromonas hydrophila | Bacteria | Proteobacteria | Gammaproteobacteria |
| 50 | Coxiella burnetii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 51 | Citrobacter freundii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 52 | Escherichia coli | Bacteria | Proteobacteria | Gammaproteobacteria |
| 53 | Haemophilus influenzae Rd | Bacteria | Proteobacteria | Gammaproteobacteria |
| 54 | Klebsiella pneumoniae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 55 | Plesiomonas shigelloides | Bacteria | Proteobacteria | Gammaproteobacteria |
| 56 | Ruminobacter amylophilus | Bacteria | Spirochaetes | Spirochaetes (class) |
| 57 | Pseudomonas aeruginosa | Bacteria | Spirochaetes | Spirochaetes (class) |
| 58 | Borrelia burgdorferi | Bacteria | Spirochaetes | Spirochaetes(class) |
| 59 | Leptospira interrogans | Bacteria | Tenericutes | Mollicutes |

Continued on next page

**Table B.2 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 60 | Treponema pallidum | Bacteria | Tenericutes | Mollicutes |
| 61 | Mycoplasma genitalium | Bacteria | Thermotogae | Thermotogae (class) |
| 62 | Mycoplasma pneumoniae | Eukaryota | Alveolata | Apicomplexa |
| 63 | Thermotoga maritima | Eukaryota | Alveolata | Apicomplexa |
| 64 | Plasmodium falciparum | Eukaryota | Alveolata | Apicomplexa |
| 65 | Plasmodium falciparum | Eukaryota | Alveolata | Ciliophora |
| 66 | Toxoplasma gondii | Eukaryota | Fungi | Microsporidia |
| 67 | Tetrahymena thermophila | Eukaryota | Fungi | Fungi incertae sedis |
| 68 | Encephalitozoon cuniculi | Eukaryota | Fungi | Microsporidia |
| 69 | Microsporidium 57864 | Eukaryota | Fungi | Fungiincertaesedis |
| 70 | Mucor racemosus | Eukaryota | Fungi | Microsporidia |
| 71 | Nosema apis | Eukaryota | Fungi | Microsporidia |
| 72 | Nosema apis | Eukaryota | Fungi | Dikarya |
| 73 | Saccharomyces cerevisiae | Eukaryota | Viridiplantae | Streptophyta |
| 74 | Arbisopsis thaliana | Eukaryota | Viridiplantae | Streptophyta |
| 75 | Oryza sativa | Eukaryota | Viridiplantae | Streptophyta |

Table B.3: **Organisms in the 242-sequence 5S rRNA dataset, and their associated taxonomic groups.**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 1 | Pyrobaculum aerophilum | Archaea | Crenarchaeota | Thermoprotei |
| 2 | Sulfolobus solfataricus | Archaea | Crenarchaeota | Thermoprotei |
| 3 | Sulfolobus acidocaldarius | Archaea | Crenarchaeota | Thermoprotei |
| 4 | Pyrodictium occultum | Archaea | Crenarchaeota | Thermoprotei |
| 5 | Aeropyrum pernix | Archaea | Crenarchaeota | Thermoprotei |
| 6 | Desulfurococcus mobilis | Archaea | Crenarchaeota | Thermoprotei |
| 7 | Thermoplasma acidophilum | Archaea | Euryarchaeota | Thermoplasmata |
| 8 | Thermococcus celer | Archaea | Euryarchaeota | Thermococci |
| 9 | Pyrococcus woesei | Archaea | Euryarchaeota | Thermococci |
| 10 | Natrialba magadii | Archaea | Euryarchaeota | Halobacteria |
| 11 | Halococcus morrhuae | Archaea | Euryarchaeota | Halobacteria |
| 12 | Halococcus morrhuae | Archaea | Euryarchaeota | Halobacteria |
| 13 | Halorubrum saccharovorum | Archaea | Euryarchaeota | Halobacteria |
| 14 | Haloferax mediterranei | Archaea | Euryarchaeota | Halobacteria |
| 15 | Haloferax volcanii | Archaea | Euryarchaeota | Halobacteria |
| 16 | Halobacterium salinarum | Archaea | Euryarchaeota | Halobacteria |
| 17 | Halobacterium salinarum | Archaea | Euryarchaeota | Halobacteria |
| 18 | Haloarcula marismortui | Archaea | Euryarchaeota | Halobacteria |
| 19 | Archaeoglobus fulgidus | Archaea | Euryarchaeota | Archaeoglobi |
| 20 | Methanolobus tindarius | Archaea | Euryarchaeota | Methanomicrobia |
| 21 | Methanosarcina vacuolata | Archaea | Euryarchaeota | Methanomicrobia |
| 22 | Methanosarcina barkeri | Archaea | Euryarchaeota | Methanomicrobia |
| 23 | Methanocaldococcus jannaschii | Archaea | Euryarchaeota | Methanococci |
| 24 | Methanocaldococcus jannaschii | Archaea | Euryarchaeota | Methanococci |
| 25 | Methanothermobacter thermautotrophicus | Archaea | Euryarchaeota | Methanobacteria |
| 26 | Methanothermus fervidus | Archaea | Euryarchaeota | Methanobacteria |
| 27 | Methanothermococcus thermolithotrophicus | Archaea | Euryarchaeota | Methanococci |
| 28 | Methanobacterium formicicum | Archaea | Euryarchaeota | Methanobacteria |
| 29 | Spiroplasma melliferum | Bacteria | Tenericutes | Mollicutes |
| 30 | Mycoplasma capricolum | Bacteria | Tenericutes | Mollicutes |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 31 | Mycoplasma pneumoniae M129 | Bacteria | Tenericutes | Mollicutes |
| 32 | Bacillus pasteurii | Bacteria | Firmicutes | Bacilli |
| 33 | Bacillus subtilis | Bacteria | Firmicutes | Bacilli |
| 34 | Geobacillus stearothermophilus | Bacteria | Firmicutes | Bacilli |
| 35 | Geobacillus stearothermophilus | Bacteria | Firmicutes | Bacilli |
| 36 | Geobacillus stearothermophilus | Bacteria | Firmicutes | Bacilli |
| 37 | Geobacillus stearothermophilus | Bacteria | Firmicutes | Bacilli |
| 38 | Staphylococcus aureus | Bacteria | Firmicutes | Bacilli |
| 39 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 40 | Deinococcus radiodurans | Bacteria | Deinococcus-Thermus | Deinococci |
| 41 | Thermus sp. | Bacteria | Deinococcus-Thermus | Deinococci |
| 42 | Thermus thermophilus | Bacteria | Deinococcus-Thermus | Deinococci |
| 43 | Thermus thermophilus | Bacteria | Deinococcus-Thermus | Deinococci |
| 44 | Thermus thermophilus | Bacteria | Deinococcus-Thermus | Deinococci |
| 45 | Thermus aquaticus | Bacteria | Deinococcus-Thermus | Deinococci |
| 46 | Planctomyces brasiliensis | Bacteria | Planctomycetes | Planctomycetacia |
| 47 | Comamonas acidovorans | Bacteria | Proteobacteria | Betaproteobacteria |
| 48 | Alcaligenes faecalis | Bacteria | Proteobacteria | Betaproteobacteria |
| 49 | Rhodobacter capsulatus | Bacteria | Proteobacteria | Alphaproteobacteria |
| 50 | Agrobacterium tumefaciens | Bacteria | Proteobacteria | Alphaproteobacteria |
| 51 | Ectothiorhodospira shaposhnikovii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 52 | Halorhodospira halophila | Bacteria | Proteobacteria | Gammaproteobacteria |
| 53 | Thiothrix sp. | Bacteria | Proteobacteria | Gammaproteobacteria |
| 54 | Thiothrix nivea | Bacteria | Proteobacteria | Gammaproteobacteria |
| 55 | Beggiatoa alba | Bacteria | Proteobacteria | Gammaproteobacteria |
| 56 | Acidithiobacillus thiooxidans | Bacteria | Proteobacteria | Gammaproteobacteria |
| 57 | Acidithiobacillus ferrooxidans | Bacteria | Proteobacteria | Gammaproteobacteria |
| 58 | Acidithiobacillus ferrooxidans | Bacteria | Proteobacteria | Gammaproteobacteria |
| 59 | Haemophilus influenzae | Bacteria | Proteobacteria | Gammaproteobacteria |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 60 | Listonella anguillarum | Bacteria | Proteobacteria | Gammaproteobacteria |
| 61 | Listonella pelagia | Bacteria | Proteobacteria | Gammaproteobacteria |
| 62 | Grimontia hollisae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 63 | Vibrio logei | Bacteria | Proteobacteria | Gammaproteobacteria |
| 64 | Vibrio fischeri | Bacteria | Proteobacteria | Gammaproteobacteria |
| 65 | Vibrio tubiashii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 66 | Vibrio ordalii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 67 | Vibrio metschnikovii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 68 | Vibrio nereis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 69 | Vibrio natriegens | Bacteria | Proteobacteria | Gammaproteobacteria |
| 70 | Vibrio mediterranei | Bacteria | Proteobacteria | Gammaproteobacteria |
| 71 | Vibrio gazogenes | Bacteria | Proteobacteria | Gammaproteobacteria |
| 72 | Vibrio diazotrophicus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 73 | Vibrio fluvialis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 74 | Vibrio harveyi | Bacteria | Proteobacteria | Gammaproteobacteria |
| 75 | Vibrio mimicus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 76 | Vibrio vulnificus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 77 | Vibrio proteolyticus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 78 | Vibrio parahaemolyticus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 79 | Vibrio harveyi | Bacteria | Proteobacteria | Gammaproteobacteria |
| 80 | Vibrio alginolyticus | Bacteria | Proteobacteria | Gammaproteobacteria |
| 81 | Vibrio cincinnatiensis | Bacteria | Proteobacteria | Gammaproteobacteria |
| 82 | Photobacterium angustum | Bacteria | Proteobacteria | Gammaproteobacteria |
| 83 | Photobacterium sp. | Bacteria | Proteobacteria | Gammaproteobacteria |
| 84 | Photobacterium damselae subsp. damselae | Bacteria | Proteobacteria | Gammaproteobacteria |
| 85 | Plesiomonas shigelloides | Bacteria | Proteobacteria | Gammaproteobacteria |
| 86 | Escherichia coli | Bacteria | Proteobacteria | Gammaproteobacteria |
| 87 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 88 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 89 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 90 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 91 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 92 | Salmonella typhimurium LT2 | Bacteria | Proteobacteria | Gammaproteobacteria |
| 93 | Shewanella hanedai | Bacteria | Proteobacteria | Gammaproteobacteria |
| 94 | Shewanella putrefaciens | Bacteria | Proteobacteria | Gammaproteobacteria |
| 95 | Shewanella colwelliana | Bacteria | Proteobacteria | Gammaproteobacteria |
| 96 | Azotobacter vinelandii | Bacteria | Proteobacteria | Gammaproteobacteria |
| 97 | Pseudomonas stutzeri | Bacteria | Proteobacteria | Gammaproteobacteria |
| 98 | Pseudomonas stutzeri | Bacteria | Proteobacteria | Gammaproteobacteria |
| 99 | Pseudomonas fluorescens | Bacteria | Proteobacteria | Gammaproteobacteria |
| 100 | Pseudomonas aeruginosa | Bacteria | Proteobacteria | Gammaproteobacteria |
| 101 | Pseudomonas aeruginosa | Bacteria | Proteobacteria | Gammaproteobacteria |
| 102 | Campylobacter jejuni | Bacteria | Proteobacteria | delta/epsilonsubdivisions |
| 103 | Empedobacter brevis | Bacteria | Bacteroidetes/Chlorobigroup | Bacteroidetes |
| 104 | Chlorobium limicola | Bacteria | Bacteroidetes/Chlorobigroup | Chlorobi |
| 105 | Pseudonocardia hydrocarbonoxydans | Bacteria | Actinobacteria | Actinobacteria(class) |
| 106 | Actinomadura madurae | Bacteria | Actinobacteria | Actinobacteria(class) |
| 107 | Arthrobacter oxydans | Bacteria | Actinobacteria | Actinobacteria(class) |
| 108 | Arthrobacter globiformis | Bacteria | Actinobacteria | Actinobacteria(class) |
| 109 | Arthrobacter globiformis | Bacteria | Actinobacteria | Actinobacteria(class) |
| 110 | Micrococcus luteus | Bacteria | Actinobacteria | Actinobacteria(class) |
| 111 | Mycobacterium bovis | Bacteria | Actinobacteria | Actinobacteria(class) |
| 112 | Cryptomonas paramecium | Eukaryota | Cryptophyta | Cryptomonadaceae |
| 113 | Cyanophora paradoxa | Eukaryota | Glaucocystophyceae | Cyanophoraceae |
| 114 | Dictyostelium discoideum | Eukaryota | Amoebozoa | Mycetozoa |
| 115 | Physarum polycephalum | Eukaryota | Amoebozoa | Mycetozoa |
| 116 | Acanthamoeba castellanii | Eukaryota | Amoebozoa | Centramoebida |
| 117 | Trypanoplasma borreli | Eukaryota | Euglenozoa | Kinetoplastida |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 118 | Crithidia fasciculata | Eukaryota | Euglenozoa | Kinetoplastida |
| 119 | Phytomonas sp. Isolate Alp1 | Eukaryota | Euglenozoa | Kinetoplastida |
| 120 | Trypanosoma cruzi | Eukaryota | Euglenozoa | Kinetoplastida |
| 121 | Trypanosoma brucei | Eukaryota | Euglenozoa | Kinetoplastida |
| 122 | Euglena gracilis | Eukaryota | Euglenozoa | Euglenida |
| 123 | Euglena gracilis | Eukaryota | Euglenozoa | Euglenida |
| 124 | Schizochytrium aggregatum | Eukaryota | stramenopiles | Labyrinthulida |
| 125 | Diatoma tenue | Eukaryota | stramenopiles | Bacillariophyta |
| 126 | Crypthecodinium cohnii | Eukaryota | Alveolata | Dinophyceae |
| 127 | Plasmodium falciparum | Eukaryota | Alveolata | Apicomplexa |
| 128 | Blepharisma japonicum | Eukaryota | Alveolata | Ciliophora |
| 129 | Bresslaua vorax | Eukaryota | Alveolata | Ciliophora |
| 130 | Paramecium tetraurelia | Eukaryota | Alveolata | Ciliophora |
| 131 | Tetrahymena thermophila | Eukaryota | Alveolata | Ciliophora |
| 132 | Tetrahymena thermophila | Eukaryota | Alveolata | Ciliophora |
| 133 | Euplotes woodruffi | Eukaryota | Alveolata | Ciliophora |
| 134 | Euplotes eurystomus | Eukaryota | Alveolata | Ciliophora |
| 135 | Gracilaria compressa | Eukaryota | Rhodophyta | Florideophyceae |
| 136 | Amoebidium parasiticum | Eukaryota | Fungi/Metazoa group | Fungi/Metazoa incertae sedis |
| 137 | Blastocladiella simplex | Eukaryota | Fungi/Metazoa group | Fungi |
| 138 | Mortierella formosensis | Eukaryota | Fungi/Metazoa group | Fungi |
| 139 | Exobasidium vaccinii | Eukaryota | Fungi/Metazoa group | Fungi |
| 140 | Christiansenia pallida | Eukaryota | Fungi/Metazoa group | Fungi |
| 141 | Filobasidiella neoformans | Eukaryota | Fungi/Metazoa group | Fungi |
| 142 | Hyphodontia paradoxa | Eukaryota | Fungi/Metazoa group | Fungi |
| 143 | Lentinula edodes | Eukaryota | Fungi/Metazoa group | Fungi |
| 144 | Kabatiella microsticta | Eukaryota | Fungi/Metazoa group | Fungi |
| 145 | Ascobolus immersus | Eukaryota | Fungi/Metazoa group | Fungi |
| 146 | Candida albicans | Eukaryota | Fungi/Metazoa group | Fungi |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 147 | Saccharomyces cerevisiae | Eukaryota | Fungi/Metazoa group | Fungi |
| 148 | Schizosaccharomyces pombe | Eukaryota | Fungi/Metazoa group | Fungi |
| 149 | Schizosaccharomyces pombe | Eukaryota | Fungi/Metazoa group | Fungi |
| 150 | Pneumocystis carinii | Eukaryota | Fungi/Metazoa group | Fungi |
| 151 | Chrysaora quinquecirrha | Eukaryota | Fungi/Metazoa group | Metazoa |
| 152 | Aurelia aurita | Eukaryota | Fungi/Metazoa group | Metazoa |
| 153 | Aurelia aurita | Eukaryota | Fungi/Metazoa group | Metazoa |
| 154 | Nemopsis dofleini | Eukaryota | Fungi/Metazoa group | Metazoa |
| 155 | Actinia equina | Eukaryota | Fungi/Metazoa group | Metazoa |
| 156 | Brachionus plicatilis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 157 | Onchocerca cervicalis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 158 | Caenorhabditis elegans | Eukaryota | Fungi/Metazoa group | Metazoa |
| 159 | Caenorhabditis elegans | Eukaryota | Fungi/Metazoa group | Metazoa |
| 160 | Globodera pallida | Eukaryota | Fungi/Metazoa group | Metazoa |
| 161 | Saccoglossus kowalevskii | Eukaryota | Fungi/Metazoa group | Metazoa |
| 162 | Branchiostoma belcheri | Eukaryota | Fungi/Metazoa group | Metazoa |
| 163 | Lethenteron japonicum | Eukaryota | Fungi/Metazoa group | Metazoa |
| 164 | Scyliorhinus canicula | Eukaryota | Fungi/Metazoa group | Metazoa |
| 165 | Pleurodeles waltl | Eukaryota | Fungi/Metazoa group | Metazoa |
| 166 | Notophthalmus viridescens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 167 | Gastrotheca riobambae | Eukaryota | Fungi/Metazoa group | Metazoa |
| 168 | Xenopus laevis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 169 | Iguana iguana | Eukaryota | Fungi/Metazoa group | Metazoa |
| 170 | Bos taurus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 171 | Rattus norvegicus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 172 | Rattus norvegicus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 173 | Rattus norvegicus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 174 | Mus musculus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 175 | Mesocricetus auratus | Eukaryota | Fungi/Metazoa group | Metazoa |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 176 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 177 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 178 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 179 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 180 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 181 | Homo sapiens | Eukaryota | Fungi/Metazoa group | Metazoa |
| 182 | Oncorhynchus mykiss | Eukaryota | Fungi/Metazoa group | Metazoa |
| 183 | Micropterus salmoides | Eukaryota | Fungi/Metazoa group | Metazoa |
| 184 | Misgurnus fossilis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 185 | Acheilognathus tabira | Eukaryota | Fungi/Metazoa group | Metazoa |
| 186 | Cyprinus carpio | Eukaryota | Fungi/Metazoa group | Metazoa |
| 187 | Stichopus oshimae | Eukaryota | Fungi/Metazoa group | Metazoa |
| 188 | Pseudocentrotus depressus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 189 | Hemicentrotus sp. | Eukaryota | Fungi/Metazoa group | Metazoa |
| 190 | Asterias vulgaris | Eukaryota | Fungi/Metazoa group | Metazoa |
| 191 | Asterina pectinifera | Eukaryota | Fungi/Metazoa group | Metazoa |
| 192 | Phascolopsis gouldii | Eukaryota | Fungi/Metazoa group | Metazoa |
| 193 | Urechis unicinctus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 194 | Enchytraeus albidus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 195 | Perinereis brevicirris | Eukaryota | Fungi/Metazoa group | Metazoa |
| 196 | Lineus geniculatus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 197 | Emplectonema gracile | Eukaryota | Fungi/Metazoa group | Metazoa |
| 198 | Bugula neritina | Eukaryota | Fungi/Metazoa group | Metazoa |
| 199 | Cerastoderma edule | Eukaryota | Fungi/Metazoa group | Metazoa |
| 200 | Octopus vulgaris | Eukaryota | Fungi/Metazoa group | Metazoa |
| 201 | Illex illecebrosus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 202 | Sepia officinalis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 203 | Artemia salina | Eukaryota | Fungi/Metazoa group | Metazoa |
| 204 | Asellus aquaticus | Eukaryota | Fungi/Metazoa group | Metazoa |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | **Level 1** | **Level 2** | **Level 3** |
| 205 | Proasellus coxalis | Eukaryota | Fungi/Metazoa group | Metazoa |
| 206 | Acheta domesticus | Eukaryota | Fungi/Metazoa group | Metazoa |
| 207 | Harpalus rufipes | Eukaryota | Fungi/Metazoa group | Metazoa |
| 208 | Calliphora vicina | Eukaryota | Fungi/Metazoa group | Metazoa |
| 209 | Drosophila melanogaster | Eukaryota | Fungi/Metazoa group | Metazoa |
| 210 | Drosophila melanogaster | Eukaryota | Fungi/Metazoa group | Metazoa |
| 211 | Drosophila mauritiana | Eukaryota | Fungi/Metazoa group | Metazoa |
| 212 | Samia cynthia | Eukaryota | Fungi/Metazoa group | Metazoa |
| 213 | Antheraea pernyi | Eukaryota | Fungi/Metazoa group | Metazoa |
| 214 | Acyrthosiphon magnoliae | Eukaryota | Fungi/Metazoa group | Metazoa |
| 215 | Planocera reticulata | Eukaryota | Fungi/Metazoa group | Metazoa |
| 216 | Dugesia japonica | Eukaryota | Fungi/Metazoa group | Metazoa |
| 217 | Hymeniacidon sanguinea | Eukaryota | Fungi/Metazoa group | Metazoa |
| 218 | Haliclona oculata | Eukaryota | Fungi/Metazoa group | Metazoa |
| 219 | Spirogyra sp. | Eukaryota | Viridiplantae | Streptophyta |
| 220 | Funaria hygrometrica | Eukaryota | Viridiplantae | Streptophyta |
| 221 | Plagiomnium trichomanes | Eukaryota | Viridiplantae | Streptophyta |
| 222 | Cycas revoluta | Eukaryota | Viridiplantae | Streptophyta |
| 223 | Ephedra kokanica | Eukaryota | Viridiplantae | Streptophyta |
| 224 | Gnetum gnemon | Eukaryota | Viridiplantae | Streptophyta |
| 225 | Metasequoia glyptostroboides | Eukaryota | Viridiplantae | Streptophyta |
| 226 | Larix decidua | Eukaryota | Viridiplantae | Streptophyta |
| 227 | Pinus radiata | Eukaryota | Viridiplantae | Streptophyta |
| 228 | Ginkgo biloba | Eukaryota | Viridiplantae | Streptophyta |
| 229 | Beta vulgaris | Eukaryota | Viridiplantae | Streptophyta |
| 230 | Quercus petraea | Eukaryota | Viridiplantae | Streptophyta |
| 231 | Linum usitatissimum | Eukaryota | Viridiplantae | Streptophyta |
| 232 | Phaseolus vulgaris | Eukaryota | Viridiplantae | Streptophyta |
| 233 | Lupinus luteus | Eukaryota | Viridiplantae | Streptophyta |

Continued on next page

**Table B.3 – continued from previous page**

| No. | Organism name | Taxonomy | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| 234 | Brassica napus | Eukaryota | Viridiplantae | Streptophyta |
| 235 | Gossypium arboreum | Eukaryota | Viridiplantae | Streptophyta |
| 236 | Petunia x hybrida | Eukaryota | Viridiplantae | Streptophyta |
| 237 | Petunia x hybrida | Eukaryota | Viridiplantae | Streptophyta |
| 238 | Triticum monococcum | Eukaryota | Viridiplantae | Streptophyta |
| 239 | Triticum aestivum | Eukaryota | Viridiplantae | Streptophyta |
| 240 | Oryza sativa | Eukaryota | Viridiplantae | Streptophyta |
| 241 | Oryza sativa | Eukaryota | Viridiplantae | Streptophyta |
| 242 | Equisetum arvense | Eukaryota | Viridiplantae | Streptophyta |

# Bibliography

[1] O. Alter. Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci U S A*, 103(44):16063–16064, 2006.

[2] O. Alter. Genomic signal processing: from matrix algebra to genetic networks. *Methods Mol Biol*, 377:17–60, 2007.

[3] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–10106, 2000.

[4] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100(6):3351–3356, 2003.

[5] O. Alter and G. H. Golub. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc Natl Acad Sci U S A*, 101(47):16577–16582, 2004.

[6] D. Ammons, J. Rampersad, and G. E. Fox. 5S rRNA gene deletions cause an unexpectedly high fitness loss in Escherichia coli. *Nucleic Acids Res.*,

27:637–642, 1999.

[7] C. I. Amos, X. Wu, P. Broderick, I. P. Gorlov, J. Gu, T. Eisen, Q. Dong, Q. Zhang, X. Gu, J. Vijayakrishnan, K. Sullivan, A. Matakidou, Y. Wang, G. Mills, K. Doheny, Y. Y. Tsai, W. V. Chen, S. Shete, M. R. Spitz, and R. S. Houlston. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.*, 40:616–622, 2008.

[8] V. P. Antao and I. Tinoco. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.*, 20(4):819–824, 1992.

[9] S. L. Baldauf, A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290(5493):972–977, 2000.

[10] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4Å resolution. *Science*, 289:905–920, 2000.

[11] A. Bashan and A. Yonath. Correlating ribosome function with high-resolution structures. *Trends Microbiol.*, 16:326–335, 2008.

[12] J. Brosius, T. J. Dull, and H. F. Noller. Complete nucleotide sequence of a 23S ribosomal RNA gene from Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 77:201–204, 1980.

[13] J. R. Brown and W. F. Doolittle. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci U S A*, 92(7):2441–2445, 1995.

[14] C. E. Bullerwell and B. F. Lang. Fungal evolution: the case of the vanishing mitochondrion. *Curr Opin Microbiol*, 8(4):362–369, 2005.

[15] J. Cadima and I. Jolliffe. On relationships between uncentered and column-centered principal component analysis. *Pak J Statist*, 25:473–503, 2009.

[16] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3:2, 2002.

[17] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young decomposition. *Psychometrika*, 35:283–319, 1970.

[18] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2:171–178, 1995.

[19] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, A. A. Szewczak, C. E. Kundrot, T. R. Cech, and J. A. Doudna. RNA tertiary structure mediation by adenosine platforms. *Science*, 273:1696–1699, 1996.

[20] R. B. Cattell. Parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika*, 9:267–283, 1944.

[21] R. B. Cattell. The three basic factor-analytic research designs – their interrelations and derivatives. *Psychological Bulletin*, 49:449–452, 1952.

[22] M. Chastain and I. Tinoco. Structural Elements in RNA. volume 41 of *Progress in Nucleic Acid Research and Molecular Biology*, pages 131 – 177. Academic Press, 1991.

[23] X. J. Chen and R. A. Butow. The organization and inheritance of the mitochondrial genome. *Nat. Rev. Genet.*, 6:815–825, 2005.

[24] G. L. Conn, R. R. Gutell, and D. E. Draper. A functional ribosomal RNA tertiary structure involves a base triple interaction. *Biochemistry*, 37(34):11980–11988, 1998.

[25] The International HapMap Consortium*. The International HapMap Project. *Nature*, 426:789–796, 2003.

[26] F. H. Crick. The origin of the genetic code. *J Mol Biol*, 38:367–379, 1968.

[27] J. B. Dacks and W. F. Doolittle. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell*, 107(4):419–425, 2001.

[28] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[29] M. T. Dixon and D. M. Hillis. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.*, 10:256–267, 1993.

[30] E. A. Doherty, R. T. Batey, B. Masquida, and J. A. Doudna. A universal mode of helix packing in RNA. *Nat Struct Biol*, 8:339–343, 2001.

[31] S. Dokudovskaya, O. Dontsova, O. Shpanchenko, A. Bogdanov, and R. Brimacombe. Loop IV of 5S ribosomal RNA has contacts both to domain II and to domain V of the 23S RNA. *RNA*, 2:146–152, 1996.

[32] D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C. S. Healey, R. Bowman, K. B. Meyer, C. A. Haiman, L. K. Kolonel, B. E. Henderson, L. Le Marchand, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C. Y. Shen, P. E. Wu, H. C. Wang, D. Eccles, D. G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M. R. Stratton, N. Rahman, G. Chenevix-Trench, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, D. J. Hunter, S. E. Hankinson, D. G. Cox, P. Hall, S. Wedren, J. Liu, Y. L. Low, N. Bogdanova, P. Schurmann, T. Dork, R. A. Tollenaar, C. E. Jacobi, P. Devilee, J. G. Klijn, A. J. Sigurdson, M. M. Doody, B. H.

Alexander, J. Zhang, A. Cox, I. W. Brock, G. MacPherson, M. W. Reed, F. J. Couch, E. L. Goode, J. E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J. L. Hopper, M. McCredie, M. Southey, G. G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y. D. Ko, A. B. Spurdle, J. Beesley, X. Chen, A. Mannermaa, V. M. Kosma, V. Kataja, J. Hartikainen, N. E. Day, D. R. Cox, and B. A. Ponder. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093, 2007.

[33] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res*, 22:2079–2088, 1994.

[34] T. D. Edlind, J. Li, G. S. Visvesvara, M. H. Vodkin, G. L. McLaughlin, and S. K. Katiyar. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. *Mol Phylogenet Evol*, 5(2):359–367, 1996.

[35] T. M. Embley. Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philos Trans R Soc Lond B Biol Sci*, 361(1470):1055–1067, 2006.

[36] T. M. Embley and W. Martin. Eukaryotic evolution, changes and challenges. *Nature*, 440(7084):623–630, 2006.

[37] N. M. Fast, J. S. Law, B. A. P. Williams, and P. J. Keeling. Bacterial catalase in the microsporidian Nosema locustae: implications for microsporidian metabolism and genome evolution. *Eukaryot Cell*, 2(5):1069–1075, 2003.

[38] N. M. Fast, J. M. Jr Logsdon, and W. F. Doolittle. Phylogenetic analysis of the TATA box binding protein (TBP) gene from Nosema locustae: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Mol Biol Evol*, 16(10):1415–1419, 1999.

[39] F. Fogolari, S. Tessari, and H. Molinari. Singular value decomposition analysis of protein sequence alignment score data. *Proteins*, 46:161–170, 2002.

[40] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci U S A*, 74(10):4537–4541, 1977.

[41] M. G. Gagnon, A. Mukhopadhyay, and S. V. Steinberg. Close packing of helices 3 and 12 of 16 S rRNA is required for the normal ribosome function. *J. Biol. Chem.*, 281:39349–39357, 2006.

[42] D. Gautheret, D. Konings, and R. R. Gutell. A major family of motifs involving G.A mismatches in ribosomal RNA. *J Mol Biol*, 242(1):1–8, 1994.

[43] A. Germot, H. Philippe, and H. Le Guyader. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae. *Mol Biochem Parasitol*, 87(2):159–168, 1997.

[44] Glück, A. and Endo, Y. and Wool, I. G. Ribosomal rna identity elements for ricin a-chain recognition and catalysis : Analysis with tetraloop mutants. *Journal of Molecular Biology*, 226(2):411 – 424, 1992.

[45] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[46] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J Computational Biology*, 3:479–502, 1996.

[47] R. R. Gutell, J. J. Cannone, Z. Shang, Y. Du, and M. J. Serra. A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol*, 304(3):335–354, 2000.

[48] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*, 20(21):5785–5795, 1992.

[49] R. R. Gutell, B. Weiser, C. R. Woese, and H. F. Noller. Comparative anatomy of 16-S-like ribosomal RNA. *Prog Nucleic Acid Res Mol Biol*, 32:155–216, 1985.

[50] J. Hampe, A. Franke, P. Rosenstiel, A. Till, M. Teuber, K. Huse, M. Albrecht, G. Mayr, F. M. De La Vega, J. Briggs, S. Gunther, N. J. Prescott, C. M. Onnie, R. Hasler, B. Sipos, U. R. Folsch, T. Lengauer, M. Platzer, C. G. Mathew, M. Krawczak, and S. Schreiber. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.*, 39:207–211, 2007.

[51] R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16:1–84, 1970.

[52] A. Heger and L. Holm. Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. *Bioinformatics*, 19 Suppl 1:130–137, 2003.

[53] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6:95–108, 2005.

[54] H. L. Hitchcock. Multilple invariants and generalized rank of a p-way matrix or tensor. *Mathematics and Physics*, 7:39–79, 1927.

[55] H. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Mathematics and Physics*, 6:164–189, 1927.

[56] L. Holmberg and O. Nygård. Release of ribosome-bound 5S rRNA upon cleavage of the phosphodiester bond between nucleotides A54 and A55 in 5S rRNA. *Biol. Chem.*, 381:1041–1046, 2000.

[57] M. Huynen, R. R. Gutell, and D. Konings. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol*, 267(5):1104–1112, 1997.

[58] J. Isaksson, S. Acharya, J. Barman, P. Cheruku, and J. Chattopadhyaya. Single-stranded adenine-rich dna and rna retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry*, 43:15996–16010, 2004.

[59] T. Kamaishi, T. Hashimoto, Y. Nakamura, Y. Masuda, F. Nakamura, K. Okamoto, M. Shimizu, and M. Hasegawa. Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, Glugea plecoglossi: implications for the deepest branching of eukaryotes. *J Biochem (Tokyo)*, 120(6):1095–1103, 1996.

[60] T. Kamaishi, T. Hashimoto, Y. Nakamura, F. Nakamura, S. Murata, N. Okada, K. Okamoto, M. Shimizu, and M. Hasegawa. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. *J Mol Evol*, 42(2):257–263, 1996.

[61] P. J. Keeling and W. F. Doolittle. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol Biol Evol*, 13(10):1297–1305, 1996.

[62] P. J. Keeling and C. H. Slamovits. Simplicity and complexity of microsporidian genomes. *Eukaryot Cell*, 3(6):1363–1369, 2004.

[63] P. J. Keeling and C. H. Slamovits. Causes and effects of nuclear genome reduction. *Curr Opin Genet Dev*, 15(6):601–608, 2005.

[64] Y. Kitazoe, H. Kishino, T. Okabayashi, T. Watabe, N. Nakajima, Y. Okuhara, and Y. Kurihara. Multidimensional vector space representation for convergent evolution and molecular phylogeny. *Mol. Biol. Evol.*, 22:704–715, 2005.

[65] Y. Kitazoe, Y. Kurihara, Y. Narita, Y. Okuhara, A. Tominaga, and T. Suzuki. A new theory of phylogeny inference through construction of multidimensional vector space. *Mol. Biol. Evol.*, 18:812–828, 2001.

[66] T. G. Kolda. Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.*, 23(1):243–255, 2001.

[67] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 2008.

[68] L Lancaster and H F Noller. Involvement of 16s rrna nucleotides g1338 and a1339 in discrimination of initiator trna. *Mol Cell*, 20:623–32, 2005.

[69] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[70] J. C. Lee, J. J. Cannone, A. Wongsa, S. Ozer, D. P. Gardner, and R. R. Gutell. Comparative RNA Web (CRW): Structure. *(in preparation)*.

[71] G. Lentzen, R. Klinck, N. Matassova, F. Aboul-ela, and A. Murchie. Structural basis for contrasting activities of ribosome binding thiazole antibiotics. *Chem Biol*, 10(8):769–78, 2003.

[72] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 106:697–702, 2009.

[73] A. Mathis. Microsporidia: emerging advances in understanding the basic biology of these unique organisms. *Int J Parasitol*, 30(7):795–804, 2000.

[74] P. B. Moore. Structural motifs in RNA. *Annu. Rev. Biochem.*, 68:287–300, 1999.

[75] B. M. Moret, L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Trans Comput Biol Bioinform*, 1:13–23, 2004.

[76] C. Muralidhara, A. M. Gross, R. R. Gutell, and O. Alter. Tensor Mode-1 Higher-Order Singular Value Decomposition Reveals Subgenic Evolutionary Relationships of Convergence and Divergence and Correlations with Structural Motifs in Ribosomal RNA. *(submitted)*.

[77] P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci USA*, 98:4899–4903, 2001.

[78] H. F. Noller, J. Kop, V. Wheaton, J. Brosius, R. R. Gutell, A. M. Kopylov, F. Dohme, W. Herr, D. A. Stahl, R. Gupta, and C. R. Waese. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.*, 9:6167–6189, 1981.

[79] J. M. Ogle, D. E. Brodersen, W. M. Jr Clemons, M. J. Tarry, A. P. Carter, and V. Ramakrishnan. Recognition of cognate transfer rna by the 30s ribosomal subunit. *Science*, 292:897–902, 2001.

[80] L. Omberg, G. H. Golub, and O. Alter. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from

different studies. *Proc. Natl. Acad. Sci. U.S.A.*, 104:18371–18376, 2007.

[81] L. Omberg, J. R. Meyerson, K. Kobayashi, L. S. Drury, J. F. Diffley, and O. Alter. Global effects of dna replication and dna replication origin activity on eukaryotic gene expression. *Mol Syst Biol*, 5:312, 2009.

[82] L. E. Orgel. Evolution of the genetic apparatus. *J Mol Biol*, 38:381–393, 1968.

[83] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.*, 3:1672–1686, 2007.

[84] F. Pazos, A. Rausell, and A. Valencia. Phylogeny-independent detection of functional residues. *Bioinformatics*, 22:1440–1448, 2006.

[85] H. Philippe and A. Germot. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol Biol Evol*, 17(5):830–834, 2000.

[86] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38:904–909, 2006.

[87] A. J. Roger. Reconstructing Early Events in Eukaryotic Evolution. *Am Nat*, 154(S4):S146–S163, 1999.

[88] J. I. Sagara, S. Shimizu, T. Kawabata, S. Nakamura, M. Ikeguchi, and K. Shimizu. The use of sequence comparison to detect 'identities' in tRNA genes. *Nucleic Acids Res*, 26(8):1974–1979, 1998.

[89] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. John Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 38:D5–16, 2010.

[90] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, 102:615–623, 2000.

[91] R. Schroeder, R. Grossberger, A. Pichler, and C. Waldsich. RNA folding *in vivo*. *Curr Opin Struct Biol*, 12(3):296–300, 2002.

[92] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner,

D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[93] A. Smilde, R. Bro, and P. Geladi. *Multi-Way Analysis: Applications in the Chemical Sciences*. Wiley, West Sussex, England, 2004.

[94] S. Smit, J. Widmann, and R. Knight. Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res.*, 35:3339–3354, 2007.

[95] S. Smit, M. Yarus, and R. Knight. Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, 12:1–14, 2006.

[96] M. L. Sogin and J. D. Silberman. Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int J Parasitol*, 28(1):11–20, 1998.

[97] J. W. Stiller and B. D. Hall. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol*, 16(9):1270–1279, 1999.

[98] G. W. Stuart and M. W. Berry. A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. *J Bioinform Comput Biol*, 1:475–493, 2003.

[99] S. O. Suh, K. G. Jones, and M. Blackwell. A Group I intron in the nuclear small subunit rRNA gene of Cryptendoxyla hypophloia, an ascomycetous

fungus: evidence for a new major class of Group I introns. *J Mol Evol*, 48(5):493–500, 1999.

[100] D. L. Swofford. Paup: phylogenetic analysis using parsimony, version 3.1. 1. *Illinois Natural History Survey*, 1993.

[101] M. Szymański, M. Z. Barciszewska, V. A. Erdmann, and J. Barciszewski. 5 S rRNA: structure and interactions. *Biochem. J.*, 371:641–651, 2003.

[102] S Tavazoie, J D Hughes, M J Campbell, R J Cho, and G M Church. Systematic determination of genetic network architecture. *Nat Genet*, 22:281–285, 1999.

[103] A. Tenesa, S. M. Farrington, J. G. Prendergast, M. E. Porteous, M. Walker, N. Haq, R. A. Barnetson, E. Theodoratou, R. Cetnarskyj, N. Cartwright, C. Semple, A. J. Clark, F. J. Reid, L. A. Smith, K. Kavoussanakis, T. Koessler, P. D. Pharoah, S. Buch, C. Schafmayer, J. Tepel, S. Schreiber, H. Volzke, C. O. Schmidt, J. Hampe, J. Chang-Claude, M. Hoffmeister, H. Brenner, S. Wilkening, F. Canzian, G. Capella, V. Moreno, I. J. Deary, J. M. Starr, I. P. Tomlinson, Z. Kemp, K. Howarth, L. Carvajal-Carmona, E. Webb, P. Broderick, J. Vijayakrishnan, R. S. Houlston, G. Rennert, D. Ballinger, L. Rozek, S. B. Gruber, K. Matsuda, T. Kidokoro, Y. Nakamura, B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Montpetit, T. J. Hudson, S. Gallinger, H. Campbell, and M. G. Dunlop. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.*, 40:631–637, 2008.

[104] G. Thomas, K. B. Jacobs, M. Yeager, P. Kraft, S. Wacholder, N. Orr, K. Yu, N. Chatterjee, R. Welch, A. Hutchinson, A. Crenshaw, G. Cancel-Tassin, B. J. Staats, Z. Wang, J. Gonzalez-Bosquet, J. Fang, X. Deng, S. I. Berndt, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, O. Cussenot, A. Valeri, G. L. Andriole, E. D. Crawford, M. Tucker, D. S. Gerhard, J. F. Fraumeni, R. Hoover, R. B. Hayes, D. J. Hunter, and S. J. Chanock. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, 40:310–315, 2008.

[105] I. Tinoco and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, 1999.

[106] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[107] Y. Van de Peer, A. Ben Ali, and A. Meyer. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene*, 246(1-2):1–8, 2000.

[108] C. R. Vossbrinck, J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*, 326(6111):411–414, 1987.

[109] B. T. Wimberly, D. E. Brodersen, W. M. Jr Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30s ribosomal subunit. *Nature*, 407:327–339, 2000.

[110] S. Winker and C. R. Woese. A definition of the domains *Archaea*, *Bacteria* and *Eucarya* in terms of small subunit ribosomal RNA characteristics. *Syst Appl Microbiol*, 14:305–310, 1991.

[111] C. R. Woese. *The genetic code: The molecular basis for genetic expression.* Harper & Row, Publishers, 1967.

[112] C. R. Woese. Bacterial Evolution. *Microbiol Rev*, 51:221–271, 1987.

[113] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–5090, 1977.

[114] C. R. Woese, G. E. Fox, L. Zablen, T. Uchida, L. Bonen, K. Pechman, B. J. Lewis, and D. Stahl. Conservation of primary structure in 16S ribosomal RNA. *Nature*, 254(11):83–86, 1975.

[115] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci U S A*, 87:4576–4569, 1990.

[116] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A*, 87(21):8467–8471, 1990.

[117] E. Zuckerkandl and L. Pauling. *Horizons in Biochemistry*, pages 189–225. Academic Press, New York, 1962.

# Vita

Chaitanya Muralidhara was born in Mysore, India in 1982. She received the Bachelor of Engineering (Honors) degree in Computer Science from the Birla Institute of Technology and Science, Pilani, India. She was admitted into the PhD program in Cellular and Molecular Biology at the University of Texas at Austin in 2004, and has been working in the Alter lab since May 2005.

Chaitanya has presented the research described in this dissertation at several international conferences as invited and contributed talks, most notably, at the C. R. Rao Conference for the Interface between Statistics and the Sciences (Hyderabad, India, December 2009), the BMES Annual Fall Meeting (St. Louis, MO, October 2008), and the SIAM Annual Meeting (San Diego, CA, July 2008).

Permanent address: 411 E Buckingham Road
Apt 1114
Richardson, TX 75081

This dissertation was typeset with LaTeX[†] by the author.

---

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.