CHAOS GAME REPRESENTATION FOR GENOME SEQUENCE ANALYSIS

Thesis Submitted to

Cochin University of Science and Technology

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Under The Faculty of Technology

By

JIJOY JOSEPH

Under the Supervision of

Dr. ROSCHEN SASIKUMAR

Computational Modelling and Simulation

Process Engineering and Environmental Technology Division

National Institute for Interdisciplinary Science and Technology (CSIR)

Thiruvananthapuram, Kerala, India

May 2010

DECLARATION

I hereby declare that the matter embodied in the thesis entitled "**Chaos Game Representation for genome sequence analysis**" is the result of the investigations carried out by me at the Computational Modelling and Simulation Section of National Institute for Interdisciplinary Science and Technology (CSIR), Thiruvananthapuram, under the supervision of Dr**. Roschen Sasikumar** and the same has not been submitted elsewhere for a degree.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

Thiruvananthapuram **(Jijoy Joseph)**

May 2010

Dr. Roschen Sasikumar
Scientist G
Computational Modelling and Simulation Section
National Institute for Interdisciplinary Science and Technology (CSIR)
Thiruvananthapuram
roschen.csir@gmail.com

May 26, 2010

CERTIFICATE

This is to certify that the work embodied in the thesis entitled "**Chaos Game Representation for genome sequence analysis**" has been carried out by **Mr. Jijoy Joseph** under my supervision at the Computational Modelling and Simulation Section of National Institute for Interdisciplinary Science and Technology (CSIR), Thiruvananthapuram and the same has not been submitted elsewhere for a degree.

Thiruvananthapuram **Roschen Sasikumar**

May, 2010 **(Thesis Supervisor)**

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

It is difficult to overstate my gratitude to my Ph.D. supervisor, ***Dr. Roschen Sasikumar****. Her enthusiasm for research and her ability and willingness to explain things in a simple form helped me to be comfortable with the previously unacquainted subject. I wish to thank her, for her patient listening and tolerant supervising, without which I could not have finished my dissertation successfully.*

I wish to thank Prof. T. K. Chandrashekar and Dr. Suresh Das, Directors of the National Institute for Interdisciplinary Science and Technology (NIIST) for providing me the necessary facilities for carrying out the work.

My sincere thanks are also due to

- *Dr. Elizabeth Jacob, Dr. Savithri S., Dr. Suresh C.H., and Dr. Vijayalakshmi K.P., scientists of the Computational Modelling and Simulation group, for all their support and advice during my tenure at the laboratory.*
- *Dr. C.S. Bhat, for his help on CGR, during the initial stages in particular*
- *Mr. K. Rajeev, Govt. Sanskrit College, for introducing us to endosymbiosis*
- *Prof. Guenter E. Peschek, University of Vienna, for his collaboration on work regarding endosymbiosis*
- *Dr. Achuthsankar S. Nair, University of Kerala and Dr. Sumam Mary Idiculla, CUSAT, for their valuable suggestions*
- *Dr. U. Syamaprasad, Convenor, NIIST-CUSAT Research Committee*
- *CSIR, New Delhi for financial assistance*
- *The student family at CMS, for their companionship and emotional support. A lengthy list which I care for in my heart, but won't be able to put it in here*
- *Friends and well-wishers at NIIST*
- *My family, for their prayers, love and tolerance which helped me endure the period*

"Not on our merits, but by His grace"

Jijoy Joseph

CONTENTS

PREFACE

Computational biology is an area of study that applies the techniques of computer science, applied mathematics and statistics to address biological problems. Sequence analysis, which forms an integral part of this highly interdisciplinary field, deals with the computational examination of nucleotide or aminoacid sequences, considering them plainly as strings of characters. Various algorithms and statistical techniques are employed to interpret and understand these huge rapidly increasing datasets of biological sequences. The aims of these techniques are multifaceted and include gene finding, gene structure prediction, functional annotation of genes, identifying Single Nucleotide Polymorphisms, reconstructing evolutionary relationships, determination of functional regions in sequences, prediction of gene expression, etc. Chaos Game Representation (CGR) is one such algorithm, originally proposed by Jeffrey (1990) as a technique for studying the "non-randomness" of genomic sequences. This iterated function system method has remarkable potential and this thesis entitled: “Chaos Game Representation for genome sequence analysis”, reports our efforts to exploit this technique further, particularly to analyze large nucleotide sequences.

The Chaos Game is an algorithm which produces pictures of fractal structures for fairly large nucleotide sequences. CGR is mathematically an iterative mapping technique that processes sequences such as nucleotides, in order to find the coordinates for their position in a continuous space. This distribution of positions has two properties: it is unique, and the source sequence can be recovered from the coordinates such that distance between positions measures similarity between the corresponding sequences. These properties, suggest the possibility to employ CGR as a sequence alignment and comparison tool.

Frequency Chaos Game Representation (FCGR) is a special form of CGR in which the CGR plot is discretized using a grid. FCGR has the property that the numbers of different oligonucleotides in the sequence can be quickly counted. This straight away makes FCGR, a handy tool to evaluate "genomic signature", which is essentially the oligonucleotide frequency profile of the sequence. In this work, techniques based on CGR for nucleotide sequence analysis are conceived, which include a fast algorithm for identifying all local alignments between two large DNA sequences and a tool for phylogenetic analysis based on genomic signature.

The first chapter of the thesis gives an overview of CGR. In a CGR for a particular oligonucleotide of length k (k-mer), its CGR co-ordinates will always be contained in a specific square with side length 2^{-k}. Thus counting the points in the square give the frequency of that k-mer and that gives rise to the matrix representation called Frequency Chaos Game Representation. The construction of FCGR of a given sequence is also described in detail. The concept of genomic signature, which is based on the observance that subsequences of a genome exhibit an oligonucleotide frequency profile that is characteristic of the whole genome, is explained. Previous works which deals with various applications of CGR and genome signature are reviewed in this chapter.

The potential of CGR for making alignment-based comparisons of whole genome sequences is being exploited in the second chapter. An algorithm for identifying all local alignments between two long DNA sequences using the sequence information contained in CGR points is presented. Since determination of distance between all pairs of CGR points, is costly in time (complexity $O(N \times M)$, N and M being the length of the two sequences) , we speed up the program by using an anchored alignment approach similar to that used in other programs such as FASTA.

The local alignments thus obtained are chained together at the same time allowing mismatches. The method is demonstrated through comparison of whole genomes of several microbial species.

Chapter three describes the exploration of the phylogenetic signal contained in genomic signature represented by the Frequency Chaos Game Representation. A statistical measure of dissimilarity between two sequences is discussed which ultimately leads to the construction of phylogenetic trees based on FCGR. The phylogenetic signal in FCGR is validated by the evolutionary trees thus obtained. Further, the effect of varying the order of the FCGR on the reconstructed trees is discussed. In addition to it, this alignment free technique also gives rise to a chromosome wise comparison, which cannot be done using a traditional alignment based method, and the results are discussed in this chapter.

The fourth chapter deals with the application of FCGR as a tool to investigate a specific problem namely, the evolutionary origin of the eukaryotic organelles, mitochondria and chloroplasts. These organelles are believed to have originated from two bacterial ancestors, an alpha proteobacterium being the ancestor of mitochondria and a cyanobacterium being the ancestor of chloroplasts.

Phylogenetic analysis based on genome signature tree however shows that cyanobacteria and chloroplasts are closer to mitochondria than most alpha-proteobacteria. An alternate hypothesis is proposed which says that, a single endosymbiotic uptake of a cyanobacterium led to the birth of both the organelles, mitochondria and chloroplasts. Arguments from different viewpoints that support this new hypothesis are discussed in this chapter. This chapter underlines the necessity to take a re-look at established phylogenetic relationships based solely on amino acid sequence similarities.

It may be mentioned that each chapter of the thesis is presented as an independent unit and therefore the tables and figures are numbered chapter wise. Relevant references are given at the end of each chapter. A summary of the work and future directions are given towards the end of the thesis.

Chapter 1

Introduction to Chaos Game Representation

1.1 Introduction

Computational biology deals with the use of mathematical tools to extract useful information from biological data. Representative problems in computational biology range from the assembly of high-quality DNA sequences from fragmentary 'shotgun' DNA sequencing to the prediction of gene regulation with data from mRNA micro-arrays and protein chips. Although efforts are continuously being made towards understanding the characteristics of genomes, any particular genome is too long and too complex for a person to directly comprehend its characteristics. This chapter gives an introduction to one such mathematical technique called Chaos Game Representation (CGR). CGR was originally proposed as a scale-independent representation for genomic sequences by Jeffrey in 1990 (Jeffrey, 1990). The technique, formally an iterative function system, can be traced further back to the foundations of statistical mechanics, in particular to Chaos theory (Bar-Yam, 1997).

1.2 Introduction to Chaos Game

The Chaos Game is an algorithm which produces pictures of fractal structures. In mathematics, the term chaos game, as coined by Michael Barnsley (1988), originally refers to a method of creating a fractal, using a polygon and a random point inside it. In a simple form, it proceeds as follows.

1. Plot three non-collinear points on a paper. Label the points as A, B and C.
2. Plot another point anywhere on the plane. This is the current point.
3. Now take a six sided die and roll it. If the number which appears on top is 1 or 2, then plot a point mid-way of the current point and A. If the number is 3 or 4, then plot the same towards B and if the number is 5 or 6 then plot the same towards C. The point which you have last plotted is the current point.
4. Again roll the die and repeat step 3, where the current point is the point is the point you have plotted last.

If these steps are repeated many times, one might expect a paper covered with random dots or perhaps a triangle filled with random dots. Such is not the case. What we obtain is seen in the figure, a triangle filled with a sequence of smaller and further smaller triangles. This figure is called the *'Sierpinski Gasket'* after the mathematician who first defined it.

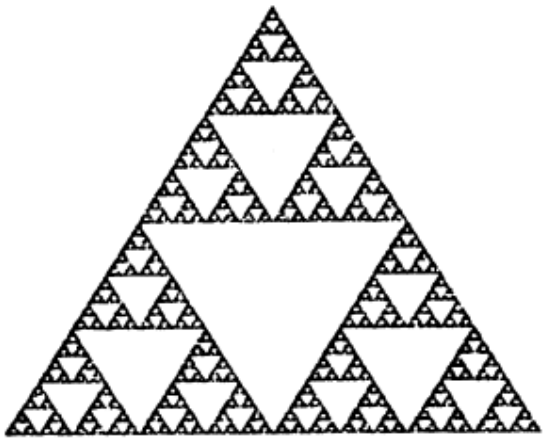


Figure 1.1 - Sierpinski's gasket

With four initial points the result is different. We will not obtain squares inside squares. What we obtain is a square uniformly filled with points.

Mathematically the chaos game is represented by an Iterated Function System (IFS). IFS is a finite collection of mappings F_i: $X \to X$ defined on a metric space X, with

$$x_i(n) = \sum_i F(x_i(n-1))$$

.

Each equation gives the formula for computing the new values of x_i.

1.3 Chaos Game Representation of DNA sequences

The DNA sequence is composed of four nucleotides Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). For our use, it is treated formally as a simple string comprising of the characters A, T, G and C. Suppose that each nucleotide is assigned a point as follows; A is (0, 0), T is (1, 0), G is (1, 1) and C is (0, 1). Given a DNA sequence, it can be visually represented in a CGR as follows. Plot the initial point in the centre of the square (0.5, 0.5) formed by the four points A, T, G and C as

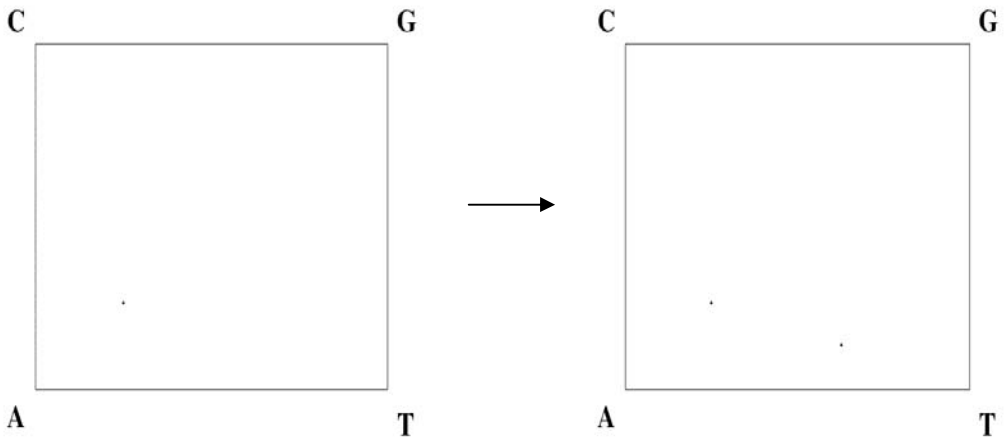
its vertices. The first nucleotide in the sequence is considered. Plot a point exactly midway between the current point and the vertex corresponding to the nucleotide. For the next nucleotide, take this second point as the current point and repeat the same procedure. i.e. The CGR of a nucleotide at position i of a sequence is exactly halfway between the previous point and the vertex corresponding to the present nucleotide.

Mathematically it is represented by an Iterated Function System, here a pair of linear equations defined by,

$$\left.\begin{aligned} x_i &= 0.5\left(x_{i-1} + g_x(i)\right) \\ y_i &= 0.5\left(y_{i-1} + g_y(i)\right) \end{aligned}\right\} \quad (1)$$

where $g_x(i)$ is the x coordinate of the vertex corresponding to the nucleotide at position i and $g_y(i)$ is the y coordinate of that vertex.

As an illustration consider the sequence ATGCGAGTGT



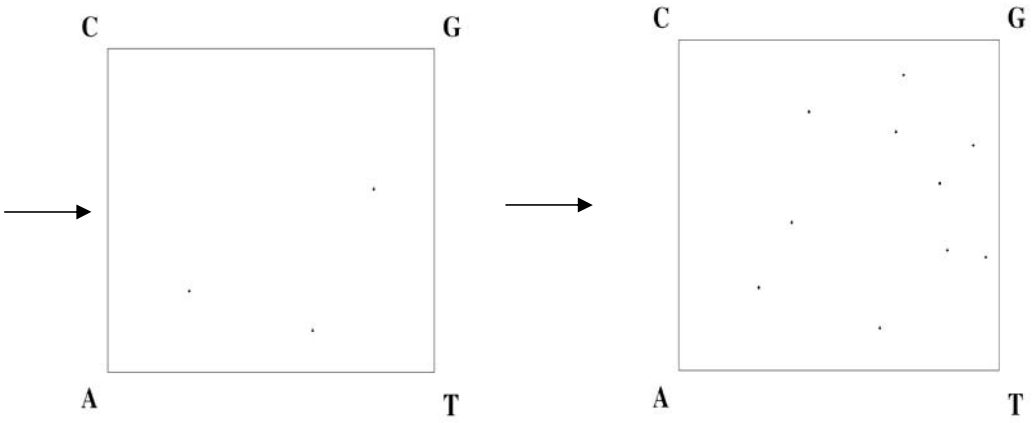


Figure 1.2 - CGR of ATGCGAGTGT

But if we continue plotting the same way for a genome region, the resulting figure is not a square filled with random dots. The CGRs for the complete genome of the bacterium Clostridium Tetani E88 and the mitochondrial genome of the plant Arabidopsis Thaliana is given as illustrations. Observe that the uniformly filled square for random probabilities strongly contrasts with the apparent structure displayed by the CGR for the DNA sequences. Also notice the difference between the patterns formed.

Figure 1.3 - Clostridium Tetani E88 - Complete Genome

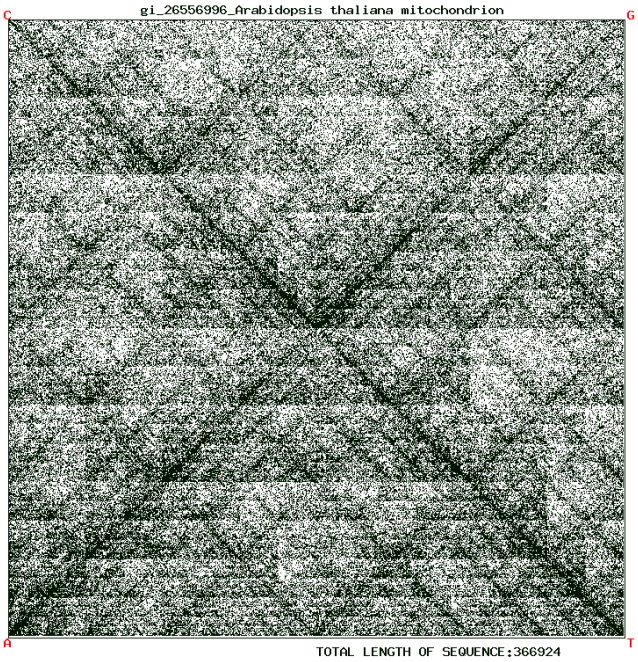


Figure 1.4 - Arabidopsis Thaliana Mitochondrion

Earlier we have shown that plotting points randomly in a square using Chaos Game, give a square randomly filled with dots, that is without any particular patterns. But plotting DNA sequences using CGR show visible patterns in the picture. What we see is the attractor formed by the iterated function system. The pictures have a complex structure which varies depending on the input sequence. H.J. Jeffrey (1990) proposed this method and visualised the patterns of different sequences. Intuitively, non-randomness in the picture corresponds to non-randomness in the sequence. It implies that the nucleotide sequences are following some kind of rule. Jeffrey noted that a pattern in one part of the picture was repeated in many places, but in varying magnitudes. The CGR thus exhibits the property of self-similarity which is very important in the study of fractals and chaotic dynamics. He noticed that there is a one to one correspondence between the sequence and the points in the CGR. Hence any visible pattern in the CGR corresponds to some pattern in the sequence of bases. It is

to be noted that adjacent nucleotides in the sequence may not be plotted adjacent to each other. He observed that the visible patterns represent global as well as local patterns in the sequence.

1.4 Frequency Chaos Game Representation

Jeffrey (1990) has observed that each point in the CGR corresponds to exactly one subsequence (starting from the first base). Though he mentioned the one to one correspondence, he did not give a method to reconstruct the sequence from the CGR. The point in a CGR corresponding to one base of a sequence is plotted in the quadrant of the square labelled with that base. This is because each quadrant comprises all points that are halfway between one corner and any other point within the square. Conversely, all points plotted within a quadrant must correspond to subsequences of the DNA sequence that end with the base labelling the corner of that quadrant. For example, any base G gives rise to a point in the G (upper-right) quadrant of the square; and every point in that quadrant corresponds to a base G in the DNA sequence. This association between points and subsequences continues recursively to sub-quadrants, sub-sub-quadrants etc.

A correspondence between the subsequence and the CGR points is described as follows. In a CGR whose side is of length 1, two sequences with suffix of length 'k', are contained within the square with side length 2^{-k}. i.e. For a particular k-mer, its CGR co-ordinates will always be contained in a specific square with side length 2^{-k}.

In the figure shown below, the CGR point which appears bold in the lower right quadrant is used to trace the sequence backwards. Since the point lies in the lower right quadrant the last nucleotide in the sequence is T. Subdivide this quadrant

into four and we can see that our point lies in the upper right quadrant which stands for G. i.e. our sequence ends with GT. Further subdividing the current quadrant into four will make the point fall into the lower right quadrant which stands for A. Here we have divided the CGR square four times to obtain the last four nucleotides which is GAGT. This can be extended to any further resolution as desired. In theory, it is possible to reconstruct the entire sequence from the first base.

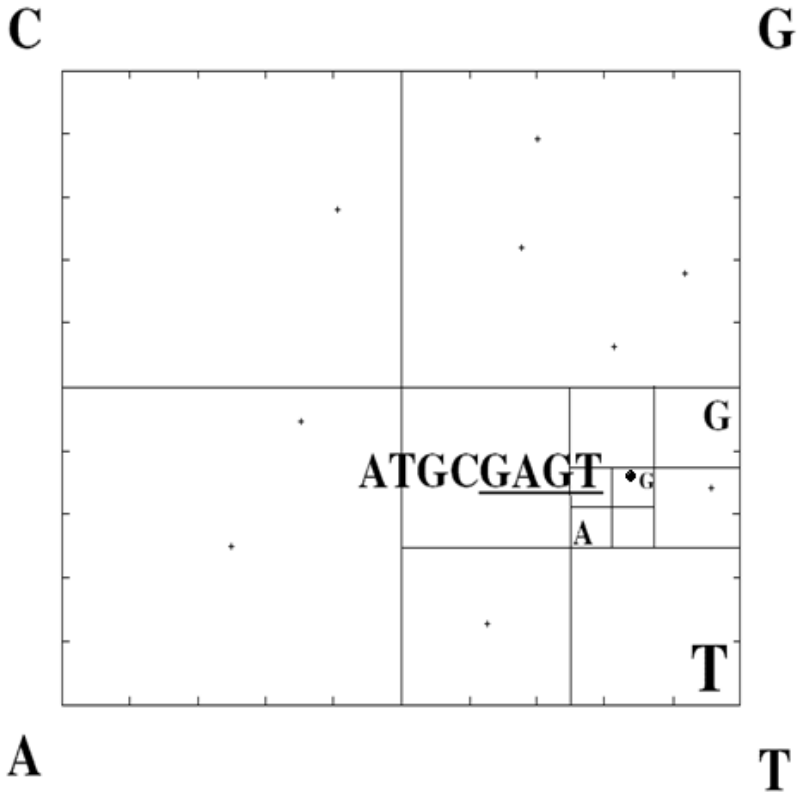
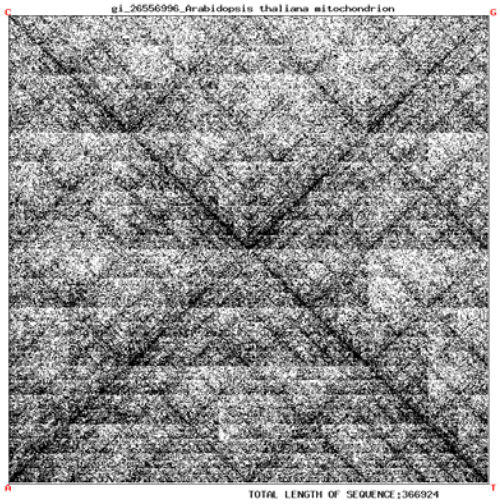


Figure 1.5 - Resolving a CGR

Another way to look at the above picture is that, whenever the pattern 'GAGT' appears in the sequence, a dot will be plotted somewhere in the corresponding square. That is, counting the number of points in the corresponding square will give the frequency of appearance of 'GAGT' in the sequence. Thus counting the CGR points in the squares of a 2^k x 2^k grid gives the number of occurrences of all possible k-mers

in the sequence. This type of representation is called a Frequency Chaos Game Representation (FCGR). The structure of FCGR was introduced by Deschavanne et al. (1999) and the name FCGR was proposed by Almeida et al. (2001). Note that those points on the grid square lines are not to be counted because they represent the length *k-1* oligonucleotide at the beginning of the DNA sequence. These *k-1* points can be omitted as long as the DNA sequence is much longer than *k*.



53	44	29	31
72	28	44	46
60	59	74	83
91	90	92	101

x 1000 bp

Figure 1.6 - CGR and FCGR of order 2

Figure 1.6 shows a CGR and its corresponding FCGR of order 2. Here the CGR is divided into 2^2 x 2^2 squares and the number of points in each square is written in the corresponding FCGR matrix. The values in this FCGR gives various dimer frequencies (multiplied by 1000) of the given sequence.

It is also possible to calculate oligonucleotide frequencies of non-integer lengths by resolving the CGR using grids of sizes other than powers of two. (Almeida et al, 2001)

Thus CGR, which was primarily meant only to be a visualization technique of nucleotide sequences, was shown to give rise to a fast algorithm for computing oligonucleotide frequencies of any length. Instead of being a graphical representation like CGR, an FCGR is a numerical matrix. The method, thus provides a graphical representation as well as a storage tool.

1.5 Review of CGR in sequence analysis

After its introduction in 1990, the potential of CGR to analyze sequences generated much interest among researchers. The observation, that the visible patterns in CGR represent global as well as local patterns in the sequence, was relevant to the DNA sequence organization. This attracted immediate further research (Basu et al, 1992; Hill et al, 1992 and Oliver et al, 1993). Hill et al. examined the CGRs of coding sequences of 7 human globin genes and 29 relatively conserved alcohol dehydrogenase genes from phylogenetically divergent species. The results showed that, CGRs of human globin cDNAs were similar to one another and to the entire human globin gene complex. Moreover, Adh CGRs were similar for genes of the same or closely related species but were different for relatively conserved Adh genes from distantly related species. The paper suggested that dinucleotide frequencies may account for the self-similar pattern that is characteristic of vertebrate CGRs and the genome-specific features of CGR patterns. Three years after the original proposition, Goldman (Goldman, 1993) interpreted that the frequency of dots in the CGR quadrants was nothing more than the oligonucleotide frequencies. CGR research received a setback when he asserted that simple Markov Chain models based solely on di-nucleotide and tri-nucleotide frequencies can completely account for the complex patterns exhibited in CGRs of DNA sequences. He concluded that the CGR

gives no further insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies and unless more complex patterns are found in CGRs, there is no justification for ascribing their patterns to anything other than the oligonucleotide frequencies. Jeffrey (1990) had earlier plotted CGR of Human Beta Globin Region on Human Chromosome 11 and the most noticeable feature in it was the repeated (self-similar) pattern of sparse 'double scoop' shaped regions, the largest of which is at the top of the G quadrant. Goldman pointed out that the double scoop is nothing more than the relative rarity of CG dinucleotides. He claimed that a four state discrete time Markov Model could easily simulate the "double scoop" pattern and other features obtained in the CGR. According to this conclusion, CGR should be relegated to the status of a pictorial representation of nucleotide, dinucleotide and trinucleotide frequencies. These sobering conclusions had the effect that CGRs have subsequently been much less studied from this perspective.

The use of CGR for the study of the entropy of genomic sequences was noted by Roman-Ronald et al (1994) and Oliver et al (1993). Oliver et al divided the square into 4^n smaller squares as in the case of an FCGR and counted the point density in each square. A histogram of the densities was prepared after determining appropriate intervals. Shannon's formula was applied to the probability distribution histogram, thus obtaining an entropic estimate of the DNA sequence. The entropic profile of the sequence was drawn by considering entropies at various resolution levels. Oliver et al. showed that the entropic profiles clearly discriminate between random and natural DNA sequences. The paper also illustrates that the entropic profile show a different degree of variability within the genome and between

genomes. The paper observes that vertebrate nuclear genomes show more variable entropic profiles than bacterial and mitochondrial ones.

The original proposition of CGR was meant for genomic sequences only. In later works it was more generalized and was shown to represent other biological sequences such as proteins (Basu et al., 1997; Pleißner et al., 1997) and also sequences of arbitrary finite number of symbols (Tino, 1999). Basu et al. used concatenated amino acid sequences of proteins belonging to a particular family. A new method of CGR was used with a 12 sided polygon in place of the CGR square. Each vertex of the polygon represented a group of amino acid residues leading to conservative substitutions. The CGR was partitioned into grids, and an estimation of the percentages of points plotted in the different segments allowed quantification of the nonrandomness of the CGR patterns generated. The CGRs of different protein families exhibited distinct visually identifiable patterns.

Deshavanne et al (1999) showed that subsequences of a genome exhibit the main characteristics of the whole genome, attesting to the validity of a genomic signature concept. The short oligonucleotide composition of a particular genome is more or less same throughout the entire genome. This property of the nucleotide sequence of an organism is known as the genome signature of that particular organism. His experiments showed that variation between CGR images along a genome was smaller than variation among genomes. He claimed that these facts strongly support the concept of genomic signature and qualify the CGR as a powerful tool to unveil it.

The measure generated on the attractor of the CGR (which is an Iterated Function System) provides more information on the sequence (Guiterrez et al, 1998;

Hao, 2000). Guitierrez et al. (Guitierrez et al, 2001) mapped a DNA symbolic sequence onto a singular measure on the attractor of a particular Iterated Function System model. A multifractal analysis of this measure is performed and singularities were interpreted in terms of mutual information and statistical dependency among subsequence symbols.

It was Almeida et al. (Almeida et al, 2001) who demonstrated that CGR may be upgraded from a mere representation technique to a sequence modeling tool. He showed that the distribution of points in the CGR has two properties: it is unique, and the source sequence can be recovered from the coordinates such that distance between positions measures similarity between the corresponding sequences. The frequency of various oligonucleotide combinations, the 'genomic signature', can be determined by dividing the CGR space with a grid of appropriate size and counting occurrence in each quadrant. In order to obtain the frequency matrix of oligonucleotide length *n*, a $2^n \times 2^n$ grid must be used. Almeida et al showed that Markov chain models are in fact particular cases of CGRs contrary to the claim by Goldman (Goldman, 1993). The frequency matrices extracted from CGR is called a Frequency Chaos Game Representation (FCGR) and can now be reordered in the more useful Markov Chain model (MCM) format (Goldman, 1993; Almagor, 1983; Avery, 1987). Almeida showed that the conversion from FCGR to MCM is straight forward only if the number of quadrants k satisfy the condition, $k = 2^{2n}$, where $n \geq 1$ is an integer. i.e. The FCGR represents an MCM only when, $k = 2^{2n}$ is satisfied. In other words, they showed that the distribution of points in CGR is a generalization of Markov chain probability tables that accommodates non-integer orders. Unlike MCM, FCGR is not constrained to represent sequences with an integer number of bases. This fundamental

characteristic of CGR is illustrated by Almeida et al. for *E.coli thrA* where the frequency of oligonucleotides with a fractionary length has been computed by dividing the CGR plane with a 10 × 10 grid (k = 100 violates condition in the above equation). Almeida et al. also suggested a global distance measure to measure the dissimilarity between the sequences. The measure was based on a weighted Pearson correlation coefficient r_w between the FCGRs. Let the two sets of FCGR quadrants be x and y with x_i and y_i representing the frequency in the i^{th} quadrant. The weighted Pearson correlation coefficient is calculated as follows:

$$nw = \sum_{i=1}^{N} x_i y_i$$

$$xw = \frac{\sum_{i=1}^{N} x_i^2 y_i}{nw} \qquad yw = \frac{\sum_{i=1}^{N} y_i^2 x_i}{nw}$$

$$sx = \frac{\sum_{i=1}^{N} (x_i - \bar{x}w)^2 x_i y_i}{nw} \qquad sy = \frac{\sum_{i=1}^{N} (y_i - \bar{y}w)^2 x_i y_i}{nw}$$

$$rw_{x,y} = \frac{\sum_{i=1}^{N} \frac{x_i - \bar{x}w}{\sqrt{sx}} \frac{y_i - \bar{y}w}{\sqrt{sy}} x_i y_i}{nw}$$

The advantage of using weighted correlation coefficient is that, the importance of each quadrant is made proportional to its magnitude. Hence a quadrant with a significantly high occurrence of a particular oligonucleotide is given more importance while determining similarity. The distance between the sequences is defined to be $d = 1 - r_w$ and this value ranges between 0 and 2. Note that the distance 0 corresponds to perfect correlation between the sequences, i.e. the sequences are similar. Almeida

et al. also recognized the property of CGR in finding out local similarity. The paper notes that two sequences with the same last nucleotide cannot be further than 0.5 distance apart. Also, two sequences with the same last two nucleotides cannot be further than 0.25 distance apart. The presence of similar nucleotides upstream will further shorten this distance. Note that each similar pair of nucleotides halves the distance between the sequences prior to it. This method of finding regions of local similarity was not further explored by the scientific community. The work of Almeida et al thus positioned CGR as a powerful sequence modelling tool that has the advantages of computational efficiency and scale independence.

Anh et al. (2002) considered the problem of matching a DNA fragment to an organism using its entire genome. The authors used Recurrent Iterative Function System (RIFS) another iterative function system which has resemblance to CGR. Their hypothesis was that the multifractal characteristic of the probability measure of a complete genome, as captured by the RIFS, is preserved in its reasonably long fragments. The RIFS of the fragments of various lengths were compared with that of the original sequences using Euclidean distance as a distance measure. The hypothesis is supported by results obtained on five randomly selected genomes.

Wang et al. (Wang et al, 2005) made a detailed and comprehensive study on various genomic signatures. The papers first concern was to prove that while nucleotide, di-nucleotide and tri-nucleotide frequencies are able to influence the patterns in CGRs these frequencies cannot solely determine the patterns in CGRs. Their work generated a new sequence which simulated the dinucleotide frequency of another sequence. The CGR of the original sequence and the simulated sequence were seen not to be same. The same procedure was repeated for trinucleotide

frequencies and then also the CGRs of the original sequence and the simulated sequence did not match. These were counter examples to the result claimed by Goldman (1993). It was shown that the CGR of a sequence was not solely dependent on oligonucleotide frequencies. They showed that frequencies of oligonucleotides of all lengths are needed to determine the CGR absolutely. The second part of this paper by Wang et al. concerns various genomic signatures. In parallel to CGR research, Karlin and Burge proposed the concept of genomic signature (Karlin and Burge, 1995) which says that Dinucleotide Relative Abundance Profiles (DRAPs) of different DNA sequence samples from the same organism are generally much more similar to each other than to those of sequences from other organisms. In addition, closely related organisms generally have more similar DRAPs than distantly related organisms. Wang et al. (2005) demonstrated that DRAP is one particular genomic signature contained within a broader spectrum of signatures. He claimed that CGR, which provides a unique visualization of patterns in sequence organization, is another alternative genomic signature within this spectrum. In his opinion, DRAP can be considered as a second order FCGR, where the relative frequency of nucleotides is plotted instead of the usual frequency. Note that relative frequency is defined as the original frequency divided by the product of frequencies of the component monomers. Expanding this, he generalized DRAP by defining trinucleotide relative FCGR. The trinucleotide relative frequency is defined as trinucleotide frequency divided by the product of frequencies of the component monomers. Based on these the paper proposes that various kinds of genomic signatures exist, and they can be considered as members of a spectrum of genomic signatures. The paper notices that, before computing the FCGR the sequence has to be concatenated with the reverse complement strand to nullify strand bias. Another thing is that, different organisms

will be having sequences of varying length and hence FCGRs have to be standardized by nullifying the effect of sequence length, in order to effectively compare between two of them. Thirdly, the paper also proposes some distance measures between genomic signatures of two DNA sequences. Two geometric distances which he proposes are the usual Euclidean distance and Hamming distance between two standardized FCGRs. Another geometric distance the paper proposes is the Image distance, an innovation in this paper, which is computed using two concepts neighbourhood of an integer and density in that neighbourhood. Yet another distance he mentions is a statistical one called the Pearson distance based on weighted correlation coefficient introduced by Almeida et al (2001), which we have already mentioned. He further evaluated the phylogenetic tree produced by these various distances by comparing it with the phylogenetic tree obtained using CLUSTALW.

Dufraigne et al. (2005) used the property of genome signature to detect horizontal transfer of genes between various organisms. Since DNA transfers originate from species with a signature different from those of the recipient species, the analysis of local variations of signature along recipient genome may allow for detecting exogenous DNA. First the entire genome is scanned with a sliding window while calculating the corresponding local signature. Then, the signature of each window is evaluated by measuring its deviation from the signature of the whole genome. If the signature of a window is markedly different from that of the whole genome similar signature is searched for in a database of genomic signatures to find the putative origin of that particular fragment. Deschavenne et al. analyzed a total of 22 prokaryote genomes in this way. It has been observed that atypical regions make up ~ 6% of each genome on the average. Most of the claimed Horizontal Transfers as

well as new ones were detected using this method. The origin of putative DNA transfers is looked for among ~12000 species. Donor species are proposed and sometimes strongly suggested, considering similarity of signatures.

Cenac et al (2006) considered a possible representation of a DNA sequence in a quaternary tree, based on CGR, in which one can visualize repetitions of subwords. A CGR-tree was created, which turns a sequence of letters into a Digital Search Tree (DST), obtained from the suffixes of the reversed sequence.

Fertil et al (2005) created a workspace, named GENSTYLE, for nucleotide sequence analysis based on CGR. In addition to visualization of genomic signature, the toolbox provides for comparing different signatures for the purpose of building phylogenetic tree. The origin of short DNA fragments can be searched for using this tool. The homogeneity of the signature along an entire genome could be studied which can lead to detecting Horizontal Transfers as mentioned by Dufraigne et al (2005). The software further provides for measuring similarity and differences among sequences using statistical methods such as Principal Component Analysis.

1.6 Central idea of the thesis

This thesis is an attempt to explore and enhance the potential of Chaos Game Representation as a tool for Genome sequence analysis and comparison. We demonstrate for the first time the potential of CGR for making alignment-based comparisons of whole genome sequences. A fast algorithm for identifying all local alignments between two long DNA sequences using the sequence information contained in CGR points is developed and demonstrated. Another focus of the thesis is the use of CGR as a tool to explore the concept of genomic signature and use it for

deducing phylogenetic relationships. A number of studies have demonstrated that genome signature is a phylogenetic signal which means that genome signatures of evolutionarily related organisms tend to resemble each other. In this thesis, using the different oligonucleotide frequency profiles obtained by FCGR as different representations of the genome signature, we classify different groups of organisms based on similarity of the genome signature. We find that different representations of the genome signature lead to different resolutions of the levels of classification. We apply the tool to investigate the bacterial origin of the eukaryotic organelles- mitochondria and chloroplast- by comparing the genome signatures of the organelles with those of bacteria. This leads us to formulating an alternate hypothesis for the origin of mitochondria.

This work adds to the repertoire of sequence analysis applications of Chaos Game and positions CGR as a powerful too for genome sequence analysis and comparison.

1.7 Organization of the rest of the thesis

The potential of CGR in making alignment-based comparisons of whole genome sequences is explored in the next chapter. In this chapter local alignments between two long DNA sequences are identified using the sequence information contained in CGR points. An algorithm is developed so as to compute the length of aligned sequence from the distance between corresponding CGR points of the pair of sequences. The algorithm is made faster by reducing the complexity from $O(n \times m)$ to $O(n)$. This is done by anchoring the alignment using the FCGR matrix.

The third chapter describes the investigations of the phylogenetic signal contained in genome signatures using FCGR. The chapter begins with a description

of the traditional methods of molecular phylogeny and outlining their limitations. Phylogenetic relationships based on similarity of genome signature are determined for different groups of organisms and different representations of the genome signature.

The fourth chapter deals with the application of FCGR to investigate a specific problem namely, the evolutionary origin of the eukaryotic organelles, mitochondria and chloroplasts The genome signature tree shows a major discrepancy from the established hypothesis that the bacterial ancestor of mitochondria is a member of the group alpha proteobacteria. We find that the genome signatures of mitochondria are closer to cyanobacteria than to most alpha proteobacteria. The unique capability of cyanobacteria to perform both oxygenic photosynthesis and aerobic respiration prompts a more parsimonious hypothesis that a single endosymbiotic uptake of a cyanobacterium could have led to the birth of both the organelles. Other arguments such as timing of evolutionary and geological events, selectional advantages conferred by combined photosynthesis and aerobic respiration and structural and functional similarity of cyanobacterial membranes to both the organellar membranes are brought together so as to demonstrate the plausibility of this alternate hypothesis. This chapter underlines the necessity to take a re-look at established phylogenetic relationships based solely on amino acid sequence similarities.

Summary and future directions are given towards the end of the thesis.

1.8 References

1. Almagor (1983) A Markov analysis of DNA sequences, *J. Theor. Biol.* 104: 633--645

2. Almeida JS, Carrico JA, Maretzek A, Noble PA and Fletcher M (2001) Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics,* 17(5): 429—437
3. Anh VV, Lau KS and Yu ZG (2002) Recognition of an organism from fragments of its complete genome, *Phys. Rev.* E66, 031910
4. Avery PJ (1987) The analysis of intron data and their use in the detection of short signals, *J. Mol. Evol.* 26: 335--340
5. Hao BL (2000) Fractals from genome--exact solutions of a biology-inspired problem *Physica A* 282: 225--246
6. Bar-Yam (1997) Dynamics of complex systems, Perseus Books, Cambridge, USA
7. Barnsley M (1988) *Fractals Everywhere*, Academic Press, New York
8. Basu S, Pan A, Dutta C, Das J (1992) Mathematical characterization of chaos game representation. New algorithms for nucleotide sequence analysis. *J. Mol. Biol* 228:715–719
9. Basu S, Pam A, Dutta C and Das J (1997) Chaos Game Representation of proteins, *J. Mol. Graph. Model.*,15: 279–289
10. Cenac P, Chauvin B, Ginouillac S and Pouyanne N (2006) Digital search trees and chaos game representation,
11. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.* 16(10):1391—1399
12. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* 33: e6

13. Fertil B, Massin M, Lespinats S, Devic C, Dumee P, Giron A (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature, *Nucl Acids Res,* 33:W512-W515

14. Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res*. 21: 2487–2491

15. Gutiérrez JM, Iglesias A, Rodriguez MA, Burgos JD and Moreno P (1998) Analyzing the multifractal structure of DNA nucleotide sequences, In: *Chaos and Noise in Biology and Medecine*, Barbi M and Chillemi S (eds) World Scientific, Series on Biophysics and Biocybernetics - 7, Singapore, pp 315-319

16. Gutiérrez JM, Rodríguez MA and Abramson G (2001) Multifractal analysis of DNA sequences using a novel chaos-game representation, *Physica A: Statistical Mechanics and its Applications* 300(1-2): 271--284

17. Hill KA, Schisler NJ and Singh SM (1992) Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J. Mol. Evol*. 35:261–269.

18. Jeffrey H.J (1990) Chaos game representation of gene structure, *Nucleic Acids Res*. 18: 2163—2170

19. Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature, *Trends Genet.* 11: 283--290

20. Oliver JL, Bernaola-Galvan P, Guerrero G, and Roman-Roldan R (1993) Entropic profiles of DNA sequences through chaos-game-derived images. *J. Theor. Biol.* 160(4): 457–470

21. Pleißner KP, Wernisch L, Osvald H and Fleck E (1997) *Electrophoresis* 18: 1709–2713

22. Tino P (1999) Spatial representation of symbolic sequences through iterative function systems, *IEEE Transaction on Signal Processing* 386--393

23. Wang Y, Hill K, Singh S and Kari L (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation, *Gene* 346: 173—185

Chapter 2

Whole genome sequence alignment using Chaos Game Representation

2.1 Introduction

Chaos Game Representation of genome sequences has been initially a tool for visual representation of genome sequence patterns. Later on it was employed for alignment-free comparisons of genome sequences based on oligonucleotide frequencies. However the potential of this representation for making alignment-based comparisons of whole genome sequences has not been exploited. In this chapter, a fast algorithm for identifying all local alignments between two long DNA sequences using the sequence information contained in CGR points is developed. The local alignments can be depicted graphically in a dot-matrix plot or in text form, and the significant similarities and differences between the two sequences can be identified. The method is demonstrated through comparison of whole genomes of several microbial species. Given two closely related genomes, information on mismatches, insertions, deletions and shuffles that differentiate the two genomes is found out using this algorithm.

2.2 Overview of sequence alignment

Over the past years, sequence comparison has evolved from an obscure pursuit of a few evolutionary biologists to a routine event that is performed 100,000's times a day. This is because sequence comparison is the simplest, quickest and most inexpensive way of determining whether a newly sequenced gene or protein is in fact "new" and whether this new gene might do something interesting. By comparing a sequence to others that have already been painstakingly characterized, it is possible to infer not only functional and structural similarity, but also detailed phylogenetic relationships -- simply on the basis of sequence similarity alone. In many respects, sequence searching and the assessment of sequence similarity lie at the heart of bioinformatics.

Sequence alignment is an arrangement of two or more sequences, highlighting their regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Pair wise sequence alignment methods are concerned with finding the best-matching piecewise or global alignments of protein or DNA sequences. It is typical to assume, while aligning two sequence segments that both the sequences have evolved from a common ancestor. In that case, the mismatches in the alignment can be considered as those which correspond to mutations and gaps correspond to insertions or deletions in one of the sequences.

One of the uses of sequence alignment is to find homologues of a gene or gene-product in a database of known examples. This information is useful for answering a variety of biological questions. Another very important application of pair wise alignment is identification of sequences of unknown structure or function. From the similarity of the sequences one can deduce the structure or function of the

unannotated sequence. Further, conserved regions in both the sequences may imply the structural or functional significance of the motif. A most widely used application of sequence alignment is in the study of molecular evolution. DNA carries over genetic material from generation to generation, by virtue of its semi-conservative duplication mechanism. Changes in the material are introduced by occasional errors and mutations in the duplication, and by viruses and other mechanisms which sometimes move sub-sequences within the chromosome and between individuals. Consequently, an alignment between sequences indicates that the sequences evolved from a common ancestor which contained the matching subsequences. Using assumptions about the probabilities of these change events, we can estimate the time when sequences diverged from a common ancestor or the time required for changing one sequence into another. The actual biological meaning of any alignment can never be absolutely guaranteed. However, statistical methods can be used to assess the likelihood of finding an alignment between two regions (or sequences) by chance, given the size of the database and its composition.

2.2.1 Algorithms for sequence alignment

String representation allows researchers to apply various string comparison techniques available in computer science. As a result, various applications have been developed that facilitate the task of sequence alignment. Computational approaches to sequence alignment generally fall into two categories: global alignments and local alignments. Finding a global alignment is a form of global optimization that makes the alignment to span the entire length of all query sequences. Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. A general global

alignment technique is the Needleman-Wunsch algorithm, which is based on dynamic programming. However, local alignments identify regions of similarity within long sequences that are often widely divergent overall. They are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

A variety of computational algorithms have been applied to the sequence alignment problem, including slow but formally optimizing methods like dynamic programming, and efficient, but not as thorough heuristic algorithms or probabilistic methods designed for fast large-scale database search. The dot-matrix approach, which implicitly produces a family of alignments for individual sequence regions, is qualitative and simple, though time-consuming to analyze on a large scale. To construct a dot-matrix plot, the two sequences are written along the top row and leftmost column of a two-dimensional matrix and a dot is placed at any point where the characters in the appropriate columns match. Word methods identify a series of short, nonoverlapping subsequences in the query sequence that are then matched to candidate database sequences. The relative positions of the word in the two sequences being compared are subtracted to obtain an offset; this will indicate a region of alignment if multiple distinct words produce the same offset. Only if this region is detected do these methods apply more sensitive alignment criteria; thus, many unnecessary comparisons with sequences of no appreciable similarity are eliminated. FASTA is a dynamic programming algorithm that compares two sequences to find the

best alignment. It finds regions of exact local matches between two sequences and then tries to connect them to get a global alignment. BLAST stands for "Basic Local Alignment Search Tool." BLAST searches for common words or k-tuples in the selected sequence and each database sequence and then tries to extend them beyond a selected threshold. Both FASTA and BLAST are heuristics of Smith-Waterman algorithm. Further, word methods are best known for their implementation in the database search tools FASTA and the BLAST family. Both the above algorithms are extremely efficient for aligning 'gene sized' sequences but are not suitable for aligning large sequences like whole genomes.

However, as more and more genomes are being sequenced it has important to develop efficient programs for detecting and aligning matching segments in pairs of megabase scale sequences for comparing whole genomes and determining evolutionary relationships. Several programs for large-scale genome comparison have been developed in recent years, for example, MUMmer (Kurtz et al, 2004), SSAHA (Ning et al, 2001), AVID (Bray et al, 2003), BLASTZ (Schwartz et al, 2003), LAGAN (Brudno et al, 2003), DIALIGN (Morgenstern, 2004), WABA (Baillie et al., 200) and GLASS (Leplae, 1998). All these programs follow an anchor-based approach like FASTA in which all matching *n*-mers for a fixed *n* are initially identified as potential anchors and later the anchors are extended into longer alignments. SSAHA stands for Sequence Search and Alignment by Hashing Algorithm. Its uses a hashing function with a 'k-mer' seed. Its fast and needs less memory than the traditional suffix-tree method. However the length of k is limited since if k=15 then it uses around 4 GB of memory. MUMmer which stands for Maximum Unique Matching (mer) is one of the extremely popular programs used for

megabase scale sequence alignment. It uses a suffix tree method for alignment. A Maximal Unique Match (MUM) for x and y is a pair of subsequences (x^1,y^1) that exactly match and there is no other subsequence pair that contain x^1 and y^1 simultaneously. MUMmer first constructs a suffix tree for x and later the suffixes of y are inserted into the same tree. All the MUMs are detected by traversing this suffix tree. The gaps between consecutive MUMs are aligned with the help of Smith-Waterman algorithm. MUMmer is linear in time consumption and memory usage. However, the program works best when the similarity of the input sequences are very high.

Our proposed tool uses the sequence information contained in the CGR points for detecting local alignments between large genome sequences.

2.3 Methods - CGR for alignment based comparison

2.3.1 Overview of the method

Here we develop a fast algorithm for pair wise local alignment of long sequences, of the order of megabases, using the information contained in CGR points. It is to be noted that in the CGR, a point corresponding to a sequence of length '*n*' is contained within a square with side of length 2^{-n}. This holds true for any positive integer value 'n'. In other words, a point in the CGR can be used to trace back its corresponding original sequence upto *n* nucleotides backwards, where this *n* can be arbitrarily large. This tremendous information content in the CGR has been left relatively unexplored. Most applications of CGR have been based merely on point counts calculated at various grid resolutions (FCGR). However, Almeida et al. (2001), mentions that, regions of local similarity between two sequences is reflected in the distance between CGR points. CGR points come closer together as sequence

similarity increases. They defined a measure of local similarity as length of similar sequence n_H calculated as a function of the maximum absolute difference between either CGR coordinate. Nevertheless, no attempt was made to use the information for developing an algorithm for aligning and comparing whole genomes.

We first show how the similar segments of two sequences can be identified based on the distance between the CGR points of the two sequences. Since determination of distance between all pairs of CGR points, is costly in time (complexity *O(N x M)*, *N* and *M* being the length of the two sequences) , we speed up the program by using an anchored alignment approach similar to that used in other sequence alignment programs. We use FCGR resolved by a 2^k x 2^k grid for the initial location of the matching k-mers which form the anchors. The distance between CGR points corresponding to each pair of matching k-mers, is then used to see if the matching k-mers can be extended into longer local alignments. We allow for mismatches by chaining together close local alignments. The program finds multiple local alignments between two sequences, allowing the detection of homologous segments, internal sequence duplications and shuffling of segments.

2.3.2 Using CGR points for finding identical segments in two sequences

Now, we show how the distance between CGR points can be used to identify sequence identities, and thereby a local alignment, without having to match the sequences nucleotide by nucleotide. Consider the i^{th} nucleotide of one sequence and the j^{th} nucleotide of the other. Co-ordinates of the CGR points corresponding to these positions on the two sequences are given by:

$$\left.\begin{aligned} X_i &= 0.5\ (X_{i-1} + g_{ix}) \\ Y_i &= 0.5\ (Y_{i-1} + g_{iy}) \\ X_j &= 0.5\ (X_{j-1} + g_{jx}) \\ Y_j &= 0.5\ (Y_{j-1} + g_{jy}) \end{aligned}\right\} \quad (2)$$

Define a distance between the two CGR points by

$$d(i,j) = \max\left(abs(X_i - X_j), abs(Y_i - Y_j)\right) \quad (3)$$

If the nucleotides at positions 'i' and 'j' of the first and the second sequence respectively are equal then $g_{ix} = g_{jx}$ and $g_{iy} = g_{jy}$ Then from equations (2) and (3) we get,

$$d(i,j) = 0.5d(i-1, j-1) \quad (4)$$

i.e. A pair of similar nucleotides makes the distance between the corresponding CGR points, half the distance between the previous pair of points. Extending this argument, we can say that if k consecutive nucleotides previous to positions i and j on the two sequences are identical, the distance between the CGR points corresponding to i and j is given by

$$d(i,j) = (0.5)^k d(i-k, j-k) \quad (5)$$

As *k* increases *d(i, j)* becomes smaller, i.e. as the length of identical sequence increases, the CGR points come closer together.

It must be noted that the closeness of two CGR points is not a sufficient condition to conclude that there is a length of similar sequence behind them. *d(i,j)* can become very low even when the sequences are very different. Such cases correspond to points on either side of, but close to, the borders of the quadrants corresponding to the four nucleotides. However if eqn. (5) is satisfied it can be inferred that the sequence

segment (i-k to i) in one sequence is identical to the segment (j-k to j) in the other sequence.

Taking log on both sides of eqn. (5), we get,

$$k = \frac{\log(d(i,j)) - \log(d(i-k, j-k))}{\log(0.5)} \quad (6)$$

We can get an upper bound for k by putting *d(i-k, j-k)* = 1 in eqn.(6):

$$k_{\max} = -\log_2(d(i,j)) \quad (7)$$

This can be seen to be the same as the length of similar sequence proposed by Almeida et al. as a measure for assessing local similarity in two sequences.

Equations 5 and 7 can be used to develop an algorithm for detection of all identical segments in two sequences based on the distance between CGR points.

Calculating k_{max} for a pair of positions (i,j) on the two sequences we can estimate that , at the most, the sequence segment from i to i- k_{max} in one sequence could be identical to the segment from j to j- k_{max} in the other sequence . We then check whether eqn. (5) is satisfied for k= k_{max} to see if these segments are truly identical. If not, we substitute k-1 for k and check again if eqn. (5) is satisfied. If not, then the procedure is repeated till the condition is satisfied. Thus starting from (i- k_{max}, j- k_{max}), the first position (i-k, j-k) that satisfies eqn. (5) is determined. This gives the length k up to which segments prefixed to positions i and j in the two sequences, are identical. The flow chart of this procedure is shown in the following figure:

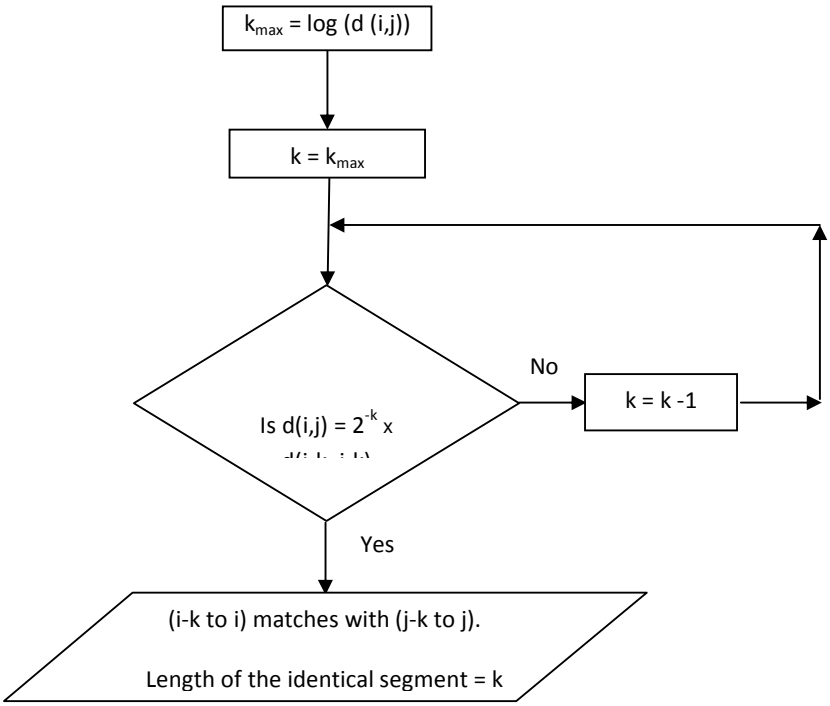


Figure 2.1 - Flow chart of the procedure for identifying matching segments

The method thus identifies identical segments without having to match the whole segment nucleotide by nucleotide. Search can be completely avoided if k_{max} is found to be less than a threshold and long homologous segments can be identified by checking only a few points from (i- k_{max}, j- k_{max}) instead of matching the whole length of the segment.

2.3.3 Speeding up the algorithm

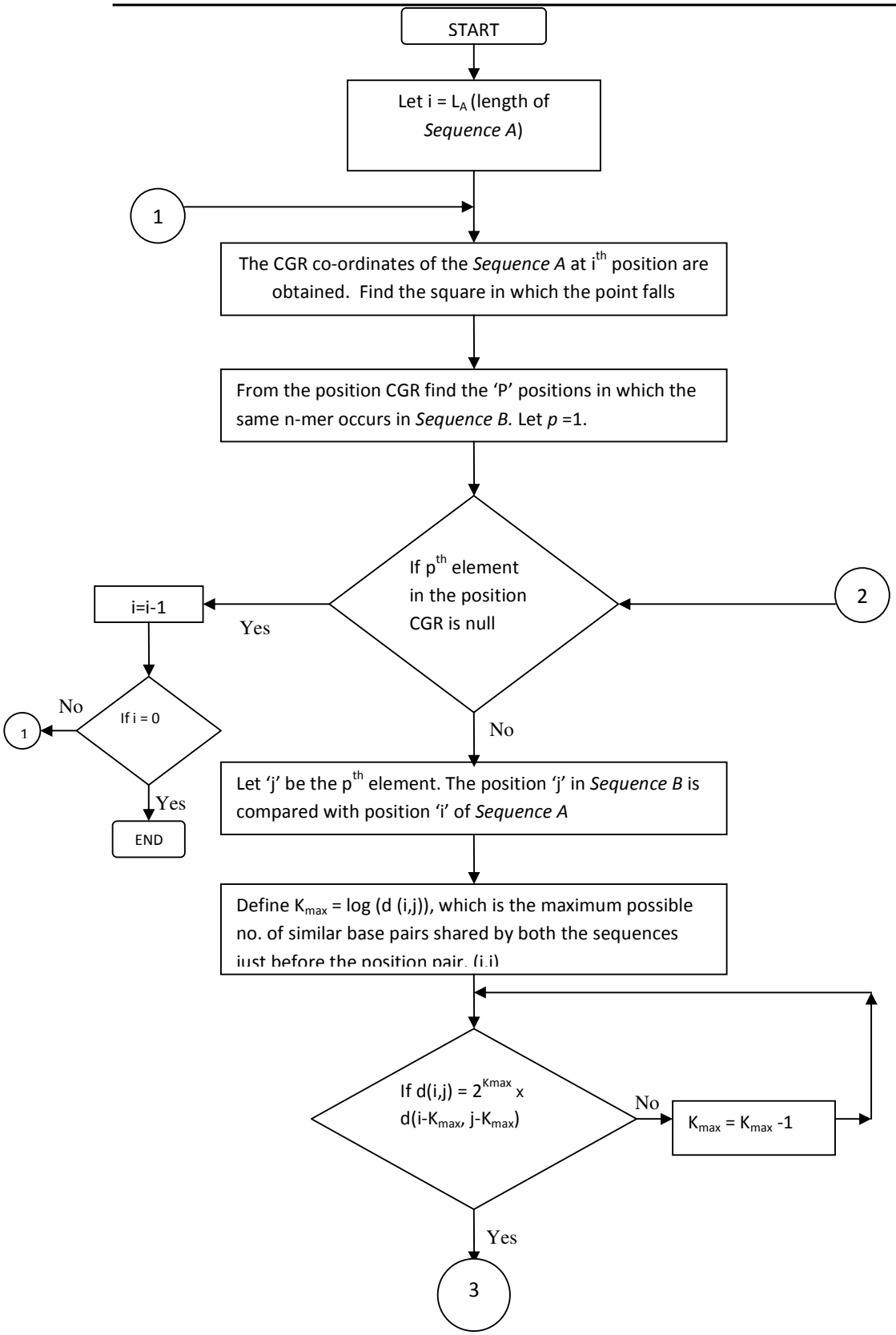
A shortcoming of the above method is that the computational cost of obtaining the pair-wise alignment is high because *d(i,j)* has to be determined for all pairs of CGR points of the two sequences and therefore the cost is of the order of the product of the length of the two sequences. In order to speed up the program, we find a way to avoid computing *d(i,j)* for all pairs of CGR points of the two sequences. For this we use information from a resolved CGR in which the CGR square is divided into grid of

size 2^n x 2^n. All CGR points falling in a square denotes the existence of a particular n-mer prefixed to that position. The 'k_{max}' is found only for those nucleotides in the other sequence which have their corresponding CGR point in the above mentioned square. It is obvious that the increase in size of 'n' will increase the speed of the algorithm. However, mismatches within the length 'n' cannot be found by this method.

The algorithm for comparing two sequences A and B is described below:

1. The CGR co-ordinates for both the sequences are calculated
2. The CGR is resolved using a 2^n x 2^n grid and the CGR points of sequence B that fall in each square are noted and stored
3. Starting from the last nucleotide of sequence A, we identify the square in which the corresponding CGR point *i* falls.
3. The CGR points of B that fall in the same square, correspond to the n-mers in B that match the n-mer which is prefixed to the position i in A
4. We calculate *d(i,j)* and k_{max} for those CGR points j of the sequence B, which fall in the same square as the CGR point i of sequence A.
5. Using *d(i,j)* and k_{max} ,we determine the length of matching segments, as described in the flowchart.
6. The longest matching segment is taken as the best local alignment at position *i*
7. The procedure is repeated next for the point i-k in A, k being the length of the longest matching segment.

The flowchart of the entire algorithm is as follows:


START
Let i = L_A (length of *Sequence A*)
1
The CGR co-ordinates of the *Sequence A* at i^{th} position are obtained. Find the square in which the point falls
From the position CGR find the 'P' positions in which the same n-mer occurs in *Sequence B.* Let *p* =1.
If p^{th} element in the position CGR is null
2
Yes
i=i-1
No
If i = 0
1
Yes
END
No
Let 'j' be the p^{th} element. The position 'j' in *Sequence B* is compared with position 'i' of *Sequence A*
Define K_{max} = log (d (i,j)), which is the maximum possible no. of similar base pairs shared by both the sequences just before the position pair. (i,j)
If $d(i,j) = 2^{Kmax} \times d(i-K_{max}, j-K_{max})$
No
$K_{max} = K_{max} - 1$
Yes
3

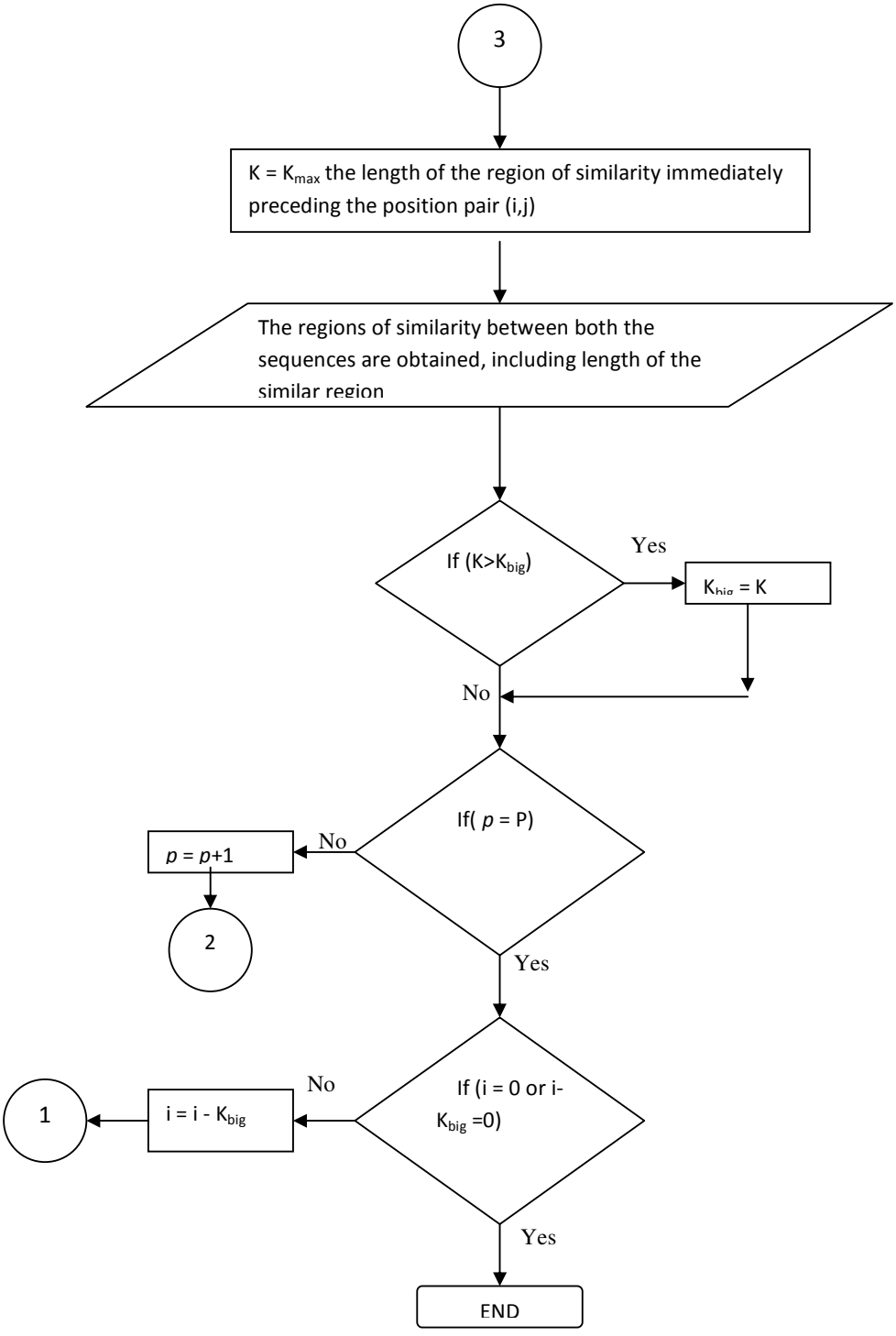


Figure 2.2 – Flowchart of the whole algorithm

We can thus find all the non-overlapping local alignments between the two sequences. Using this approach, the all-to-all comparisons of the previous section is reduced to some-to-some comparisons, which speeds up the algorithm considerably. This technique is similar to the anchored alignment method used in other alignment programs. It is to be noticed that here we use information from CGR, both for finding the anchors as well as for extending them.

The program yields the list of all local alignments between the two sequences in the order of their position in the sequence A. A sample list of alignments can be seen in Table 2.3

.2.3.4 Floating point error

For long identical sequence segments, the distance value may go below the minimum value possible for a floating point variable. The distance defined in double precision variable becomes zero when the length of identical segment is greater than 64. . Therefore in our implementation, when we encounter zero value for the distance we jump back by sixty positions and check distance again; if the distance is again zero, we jump back another sixty positions and so on until the distance becomes non-zero. We add all the skipped positions to the k that we finally calculate with the non-zero distance value.

2.3.5 Analysing the local alignments for shuffles, mismatches and insertion/deletions

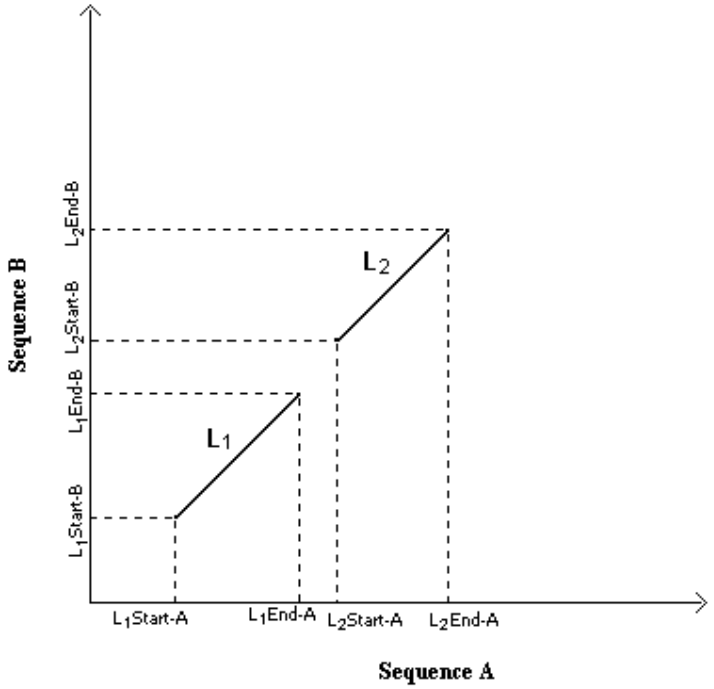


Figure 2.3 – Local Alignments

Many a time, a quantitative measure of the degree of similarity/dissimilarity between two genomes is needed in a pairwise sequence alignment. The number of (or length of) similar regions, shuffled regions, mismatches, insertions/deletions, duplications, inversions etc. help to quantify this concept. These measures are defined below. A local alignment can be defined by the start and end positions of identical segments in the two sequences. A pair of local alignments is given in the figure 2.3.

Consider two local alignments L_1 and L_2 defined by

(L_1.StartA, L_1.EndA, L_1.StartB, L_1.EndB) and

(L_2.StartA, L_2.EndA, L_2.StartB, L_2.EndB)

(a) Shuffles/Rearrangements

Consider the list of local alignments that are ordered in increasing order of L.EndA. This list may not be in increasing order of L.EndB and any disruption of order in the list with respect to position in Sequence B is indicative of shuffling. By examining the disruption of order in L.EndB we can estimate the number of shuffles that have taken place in Sequence B with respect to Sequence A.

(b) Mismatches

Let, ΔA = abs (L_2.StartA – L_1.EndA) and ΔB = abs (L_2.StartB – L_1.EndB) where L_1 and L_2 are two consecutive alignments in the ordered list.

Mismatch length between the alignments can be calculated as:

Mismatch length = min (ΔA, ΔB)

L_1EndA	*L_2StartA*	*L_1EndB*	*L_2StartB*	*Mismatch*
56	64	56	64	8
408	410	391	393	2
636	637	619	620	1
684	706	667	689	22
843	857	826	840	14
882	886	865	869	4

Table 2.1 – Mismatches in the alignment

A few mismatches between the forward strands of the genomes of E.Coli K12 and E.Coli O157:H7 is illustrated in the above table.

(c) Insertions/Deletions

Diagonal off-set between two consecutive local alignments that are consecutive in Sequence B also, indicate deletions and insertions and can be calculated as

IN/DEL length= max (ΔA, ΔB) - min(ΔA, ΔB)

In/Del	*L_1EndA*	*L_2StartA*	*L_1EndB*	*L_2StartB*	*In/Del Length*
Del	228	242	228	224	14
Ins	317	318	299	301	1
Ins	15394	15395	15377	16727	1349
Ins	38484	38485	34110	34112	1
Del	46517	46584	42111	42171	7
Del	54724	55372	50310	50397	561

Table 2.2 - In/Dels in the alignment

A sample of Insertions/Deletions between the forward strands of the genomes of E.Coli K12 and E.Coli O157:H7 is given in the table 2.2. Here the In/Dels in Sequence B is given with respect to Sequence A.

(d) Duplications

Duplications in B can be identified wherever L_1.StartA = L_2.StartA and L_1.EndA = L_2.EndA

(e) Inversions

Inversion of segments is detected by finding local alignments between Sequence A and the reverse complement of Sequence B

2.3.6 Chaining local alignments and filtering background noise

Short spurious alignments or background noise can be removed by filtering out all alignments below a certain threshold length. However this carries with it the danger of filtering out many "true" alignments that are separated by small mismatches. Therefore before filtering it is better to chain together 'nearby' perfect local alignments by allowing a certain amount of mismatches. We allow for short mismatches by chaining together local alignments that have no diagonal off-set and differ only by mismatches of a few nucleotides. We specify the maximum allowable mismatches per length of the chained alignment. If there is no diagonal off-set between them i.e. $\Delta A = \Delta B$, and the mismatch falls below the threshold value, the two alignments are chained together into a single alignment. Chained alignments having length below a threshold are discarded to filter out the background noise.

Order in A	*StartA*	*EndA*	*Order in B*	*StartB*	*EndB*	*Length*
0	0	228	0	0	228	228
1	242	317	1	224	299	75
2	318	15394	2	301	15377	15076
3	15395	18462	3	16727	19794	3067
4	24968	25707	8	20563	21302	739
5	25715	30126	9	21304	25715	4411

Table 2.3 – Sample list of alignments

The above table shows some of the matching regions between the forward strands of the genomes of E.Coli K12 and E.Coli O157:H7, after filtering background noise.

Inversions in any one of the genomes is taken care of by aligning the forward strand of one with the reverse complementary strand of the other.

Order in A	*StartA*	*EndA*	*Order in B*	*StartB*	*EndB*	*Length*
18	226955	228101	301	2729325	2728179	1146
19	227193	229051	369	3426692	3424834	1858
20	227285	227349	306	2728995	2728931	64
21	229018	229051	293	2727268	2727235	33
22	229068	230258	284	2727227	2726037	1190

Table 2.4 - Alignments between forward and reverse complementary strand

The above table shows local alignments when the forward strand of E.coli K12 is compared with the reverse complementary strand of E.coli O157:H7. Note that in Sequence B, the alignments are obtained in the reverse direction.

2.4 Results and Discussion

The following figures show dot matrix plots showing the local alignments obtained using the method. Line segments from the bottom-left to top-right direction shows regions of similarity. Line segments from the top-left to bottom-right direction shows regions of inversion. Parallel diagonal lines corresponding to the same region in the x-axis (or y-axis) show duplication in the sequence in the y-axis (correspondingly x-axis)

The pair wise alignment was performed between Human Immunodefeciency Virus and Chimpanzee Immunodeficiency Virus, Pyrococcus Abyssi and Pyrococcus Horikoshii, E. Coli OH157:H7 and E. Coli K12, Mycobacterium leprae TN and Mycobacterium tuberculosis H37Rv respectively. In all these computations the following values were used:

Length of the k-mer which is used as the anchor – 9

Threshold length for accepting as local alignment (before chaining) – 20

Maximum no. of mismatches allowed for chaining - 0.25 per matched length

Threshold length for filtering background noise, after chaining the local alignments – 30

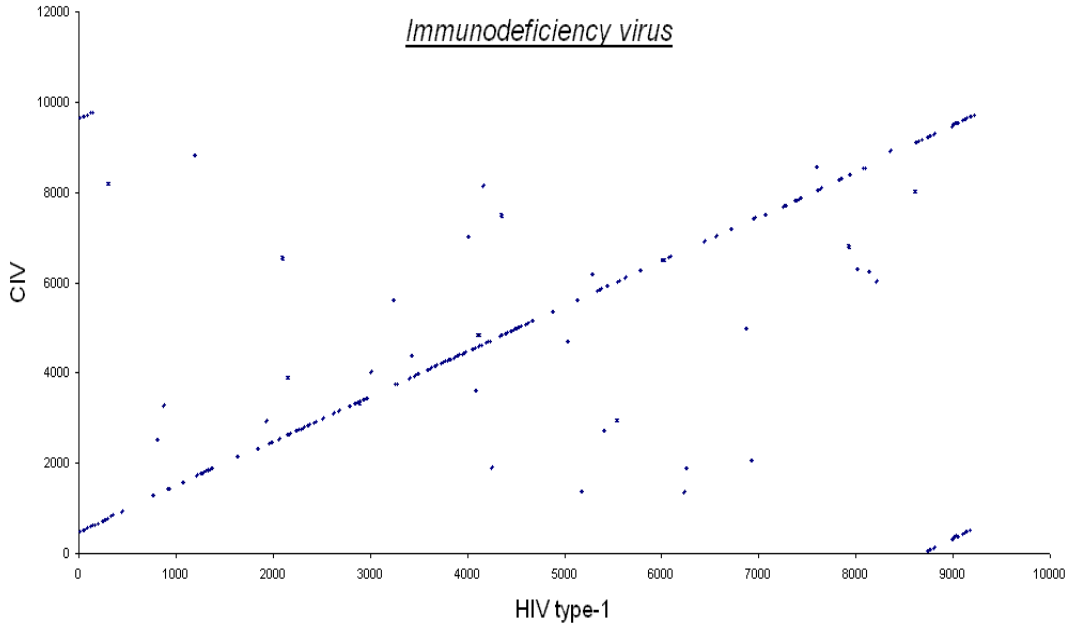


Figure 2.4 Human immunodeficiency virus and Chimpanzee immunodeficiency virus

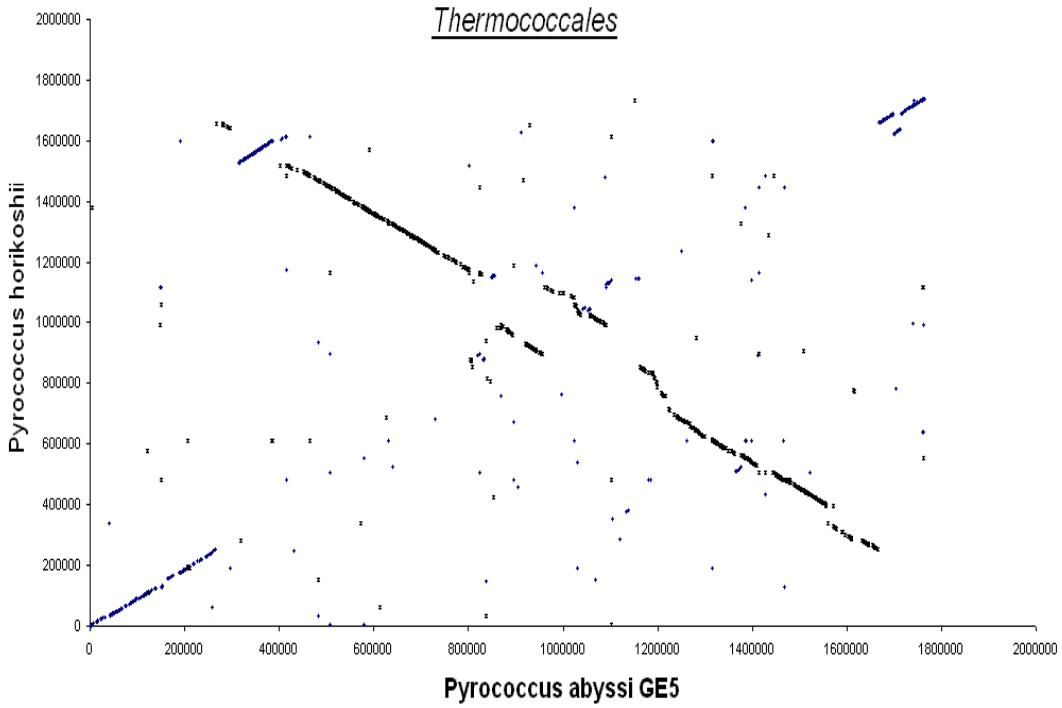


Figure 2.5 - Pyrococcus Abyssi and Pyrococcus Horikoshii. Large scale inversions can be observed in the genomes

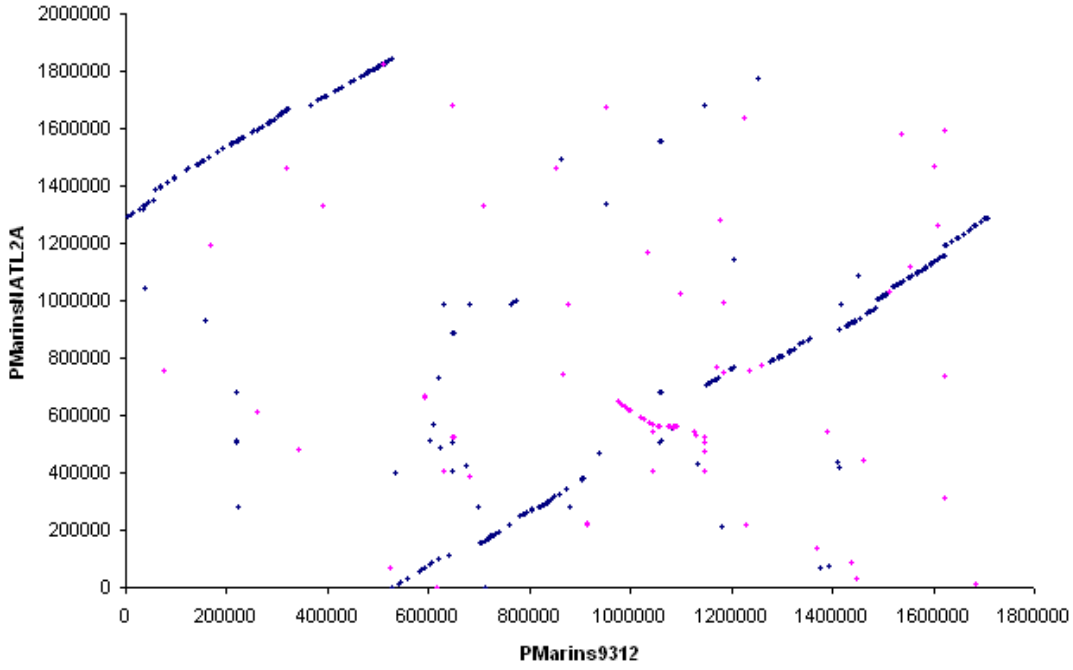


Figure 2.6 – Prochlorococcus Marinus 9312 and Prochlorococcus Marinus NATL2A. Shuffling of a large segment (~500000bp) can be found visibly in the figure. Here the alignment in reverse complementary direction is marked by a different colour.

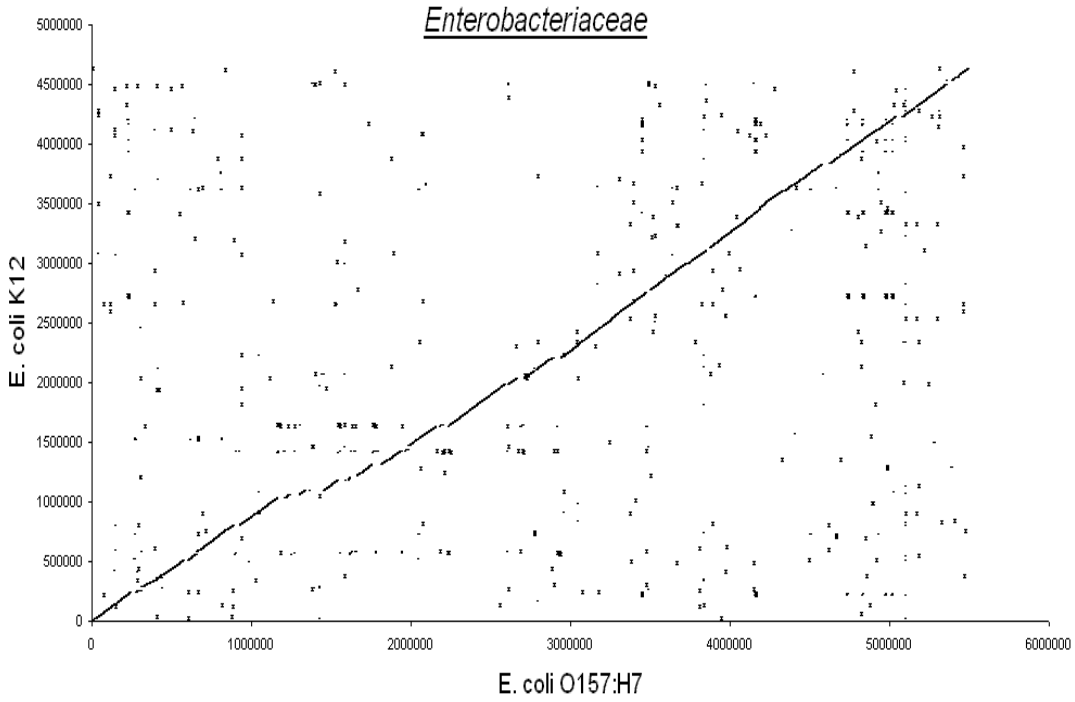


Figure 2.7 - E.coli K12 and E.coli O157:H7

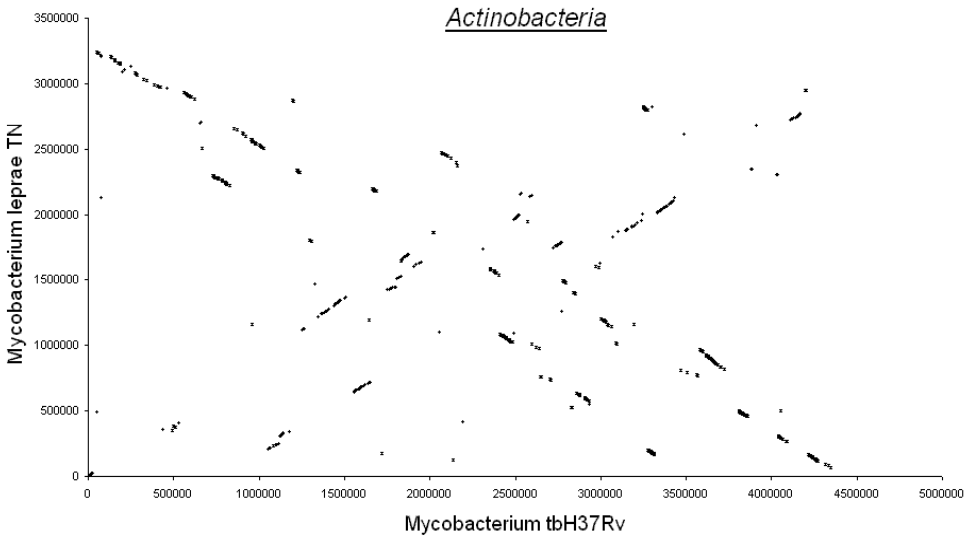


Figure 2.8 - Mycobacterim Tuberculosis H37Rv and Mycobacterium Leprae TN.
Inversions can be noticed in the figure in the alignment of these genomes

2.4.1 Computational Time

The following table (Table 2.5) shows the computation time taken for finding all local alignments between pairs of sequences of different sizes with the program running on a Pentium IV 2.5GHz machine.

It can be seen that the factors affecting the time of execution of the program are not only on the length of the sequences, but also the degree of similarity between them and the amount of internal duplications. For example, the time taken for comparing M.leprae and M.tuberculosis is much greater than the time taken for comparing M. bovis and M.tuberculosis even though the sizes of the genomes are similar. However, the time taken by this program for comparing the two E.Coli genomes is 68 seconds while MUMmer an extremely popular large-scale sequence alignment tool available takes only 17 seconds. The emphasis of this paper is on the theoretical development of the method rather than on software development and it is possible that with better programming inputs the implementation can be made more efficient and faster. The focal characteristic of this method comes from the fact that CGR simultaneously facilitates other types of sequence comparisons ranging from visual comparisons of patterns to oligonucleotide frequency spectrums and genome signatures.

Organisms	**Length A** (In base pairs)	**Length B** (In base pairs)	**Time** (forward strand) **in seconds**	**Time** (reverse complement) **in seconds**
HIV *vs.* CIV	9229	9811	<1	1
P. Abyssi *vs.* P. Horikoshii	1765118	1738505	24	27
E. coli O157:H7 *vs.* E. coli K12	5498450	4639675	68	156
R. Madrid E *vs.* R. Malish 7	1111523	1268755	18	24
M. tuberculosis H37Rv *vs.* M. leprae TN	4411532	3268203	119	120
M. bovis AF2122 *vs.* M.tuberculosis H37Rv	4345492	4411532	10	230
M. tuberculosis H37Rv *vs.* M. tuberculosis CDC1551	4411532	4403662	6	232

Table 2.5 - Computational Time Taken

2.5 Conclusion

A novel algorithm, which uses information from chaos game representation of genome sequences, for finding all local alignments between the sequences has been developed. The CGR of a nucleotide at a particular position is affected, though partially, by every nucleotide preceding it. This property of CGR has tremendous potential and this hitherto completely unutilized potential of CGR is employed in the pair wise sequence alignment algorithm. The algorithm is made faster by anchored-

alignment method. Anchoring is done using detection of n-mers through CGR. Thus, both the anchoring and the progression of the alignment are computed using CGR. The method takes maximal advantage of the CGR, regarding computational speed and its 'more-than-oligomer-representation' property. Fast comparisons can be made between sequences of mega-base size using a Pentium IV machine. As far as the speed of alignment is concerned, the program, in its present state does not offer any major improvements over MUMmer. However, it is possible that the method can be implemented more efficiently through better programming inputs. Addition of the possibility of large scale sequence alignment to the existing repertoire of alignment-free sequence analysis possibilities from CGR, positions CGR as a powerful quantitative sequence analysis tool.

2.6 References

1. Almeida JS, Carrico JA, Maretzek A, Noble PA and Fletcher M (2001) Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics,* 17(5): 429—437
2. Baillie DL, Rose AM (2000) WABA success: a tool for sequence comparison between large genomes, *Genome Res.* 10(8): 1071--1073
3. Bray N, Dubchak I and Pachter L (2003) AVID: A global alignment program, *Genome Res.* 13(1):7-102
4. Brudno M, Do C, Cooper G, Kim MF, Davydov E, Green ED, Sidow A and Batzoglou S (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA, *Genome Res.* 13(4): 721--731

5. Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res*. 21: 2487–-2491

6. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL (2004) Versatile and open software for comparing large genomes, *Genome Biology* 5: R12

7. Leplae R, Hubbard T , Tramontano A (1998) GLASS: A tool to visualize protein structure prediction data in three dimensions and evaluate their consistency, *Proteins: Structure, Function and Bioinformatics* 30(4): 339--351

8. Morgenstern B (2004) DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ, *Nucleic Acids Research* 32: W33-W36

9. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases, *Genome Res.* 11: 1725--1729

10. Schwartz S, Kent J W, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D and Miller W (2003) Human-Mouse Alignments with BLASTZ, *Genome Res.* 13:103-107

11. Wang Y, Hill K, Singh S and Kari L (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation, *Gene* 346: 173—185

Chapter 3

Phylogenetic Analysis using Frequency Chaos Game Representation

3.1 Introduction

Traditional phylogenetic methods based on molecular data regard the alignment of aminoacid sequences of homologous genes as the primary source of deducing evolutionary relationships. This chapter first discusses the inherent difficulties in this method that include, the selection of orthologous genes, likelihood of horizontal transfer of genes, the fact that mutation takes place at the nucleotide level, the information contained in the so called 'junk' DNA, ambiguities of the alignments, to name a few. The property of uniformity in oligomer profile throughout the length of an entire genome has given rise to the concept of 'genomic signature' for each genome. This genomic signature is further found to carry a distinct phylogenetic signal. i.e. Closely related species are found to have similar genomic signature. This makes genomic signature a potential tool for computing evolutionary relationships.

Moreover, this method is an alignment free sequence comparison technique and thus free of many of the errors that plague traditional alignment based methods. We use Frequency Chaos Game Representation (FCGR) of nucleotide sequences as a tool for computing genome signature. We compute pairwise distances between sequences using a statistical distance measure proposed by Almeida et al.(2001), based on weighted correlation on the FCGR of two sequences. This distance is incorporated to a distance matrix to form an evolutionary tree. The validity of the method is explored by examining whether and to what extent this method is capable of recreating established phylogenetic relationships and where this method shows unique capabilities with respect to traditional phylogenetic methods. We also explore the effect of using different representations of the genome signature in this chapter.

3.2 Computational Phylogeny - Traditional Methods and their limitations

Phylogenetics is the study of evolutionary relationship among various groups of organisms. The relatedness between the biological species is studied through morphological data matrices and molecular sequencing data. Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. The objective of the researcher is to reconstruct evolutionary trees and branches by computational methods. A wide variety of information, including molecular data analysis, biochemical analysis, and analysis of morphology of the organisms are made use of to show different reconstruction possibilities. A dilemma phylogenetics faces is that molecular data are available only for the present, while fossil records are sporadic and

unreliable. It is a complex and delicate task to fabricate trees that perfectly reproduce the evolutionary tree that represents the historical relationships between the species. Our knowledge of how evolution operates comes to help in assembling the whole tree.

3.2.1 Phylogenetic tree based on morphology

A phylogenetic tree, also called an evolutionary tree, is a graph showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. It is a form of a cladogram. In this graph, each node with descendants represents the most recent common ancestor of the descendants, and edge lengths correspond to time estimates. The earliest phylogenetic trees were simply based on the morphology and physiology of the species. Traditional phylogenetic methods rely on morphological data obtained by measuring and quantifying the phenotypic properties of representative organisms. The basic difficulty in morphological phylogenetics is the assembly of a matrix which quantifies the phenotypic characteristics being used as a classifier. Every possible phenotypic characteristic could be measured and encoded for analysis. But, the selection of which features are to be used as a basis for the matrix is a major inherent obstacle to this method. The answer lies in, which traits of a species are evolutionarily relevant. Further, morphological studies can be confounded by examples of convergent evolution of phenotypes. A high likelihood of inter-taxon overlap also present a hurdle. The inclusion of extinct taxa in morphological analysis is often difficult due to absence of or incomplete fossil records. Extinct taxa have been demonstrated to have significant effect on the phylogenetic tree produced. In one study only the inclusion of extinct species of apes produced a morphologically derived tree that was

consistent with that produced from molecular data (Strait & Grine, 2004). Phenotypic classifications, particularly those used when analyzing very diverse groups of taxa, are discrete and unambiguous. For example, classifying organisms as possessing or lacking a tail is straightforward in the majority of cases, as is counting features such as vertebrae. However, the appropriate representation of continuously varying phenotypic measurements is a contentious problem without a general solution. A frequently used method is simply to sort the measurements of interest into two or more classes, rendering continuous observed variation as discretely classifiable (e.g., all members with humerus bones longer than a given cutoff are scored as members of one state, and the others as members of a second state). This results in an easily obtained data set but has been criticized for poor reporting of the basis for the class definitions and for sacrificing information compared to methods that use a continuous weighted distribution of measurements (Wiens, 2001). The labour intensiveness of collecting morphological data prompts researchers to reuse previously compiled data matrices. This results in the propagation of flaws in the original matrix into multiple derivative analyses.

3.2.2 Phylogenetic tree based on molecular data

Molecular sequences of macromolecules, such as genes and proteins, have surpassed morphological and other organismal characters as the most popular forms of data for phylogenetic analyses. Trees based on biomolecular sequences are much better since the characters (i.e. the amino acids or the nucleotides) are more closely equal in weight. Molecular phylogenetics, also known as molecular systematics, is the use of the structure of molecules to gain information on an organism's

evolutionary relationships. Every living organism contains DNA, RNA, and proteins. Closely related organisms generally have a high degree of agreement in the molecular structure of these substances. Conserved sequences are expected to accumulate mutations over time. Assuming a constant rate of mutation provides a molecular clock for dating divergence. Molecular phylogeny uses such data to build a relationship tree that shows the probable evolution of various organisms. Not until recent decades, however, has it been possible to isolate and identify these molecular structures. As DNA sequencing has become cheaper and easier, molecular systematics has become a popular way to reconstruct phylogenies.

Many of the disadvantages of the traditional phylogeny construction based on morphology are overcome by the introduction of phylogeny based on similarity between the molecules. Early works in molecular phylogenetics made use of proteins, enzymes, carbohydrates and other molecules which were separated and characterized using techniques such as chromatography. Another early approach was to determine the divergences between the genotypes of individuals by DNA-DNA hybridisation. The advantage for using hybridisation rather than gene sequencing was that it was based on the entire genotype, rather than on particular sections of DNA. Modern sequence comparison techniques try to overcome this objection by the use of multiple sequences. Currently the exact sequences of DNA, RNA or protein segments are extracted using different techniques.

3.2.3 Multiple Sequence Alignment for building Phylogenetic tree

Currently, a large majority of methods to compute phylogenetic trees based on biomolecular sequences are based on Multiple Sequence Alignment (MSA).

Sequence alignment is the method of arranging similar regions of a pair of sequences alongside so that the differences between them can be found out. MSA refers to the sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. When the alignment is done for the purpose of evolution, the input set of query sequences is assumed to have an evolutionary relationship. i.e. It is supposed to have diverged from a common ancestor sequence. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins.

The similarity of the alignment is quantified and is made use to produce evolutionary trees. However, defining homology can be challenging due to the inherent difficulties of multiple sequence alignment. It is rarely possible to determine the evolutionary alignment of two divergent sequences with confidence because this would require knowledge of the precise history of substitutions, insertions and deletions that have led to the creation of the present day sequences from their common ancestor. For a given gapped MSA, several rooted phylogenetic trees can be constructed that vary in their interpretations of which changes are mutations, and which events are insertion-mutations or deletion-mutations. For example, given only a pairwise alignment with a gap region, it is impossible to determine whether one sequence bears an insertion mutation or the other carries a deletion. The problem is magnified in MSAs with unaligned and nonoverlapping gaps. In practice, sizable regions of a calculated alignment may be discounted in phylogenetic tree construction to avoid integrating noisy data into the tree calculation.

One way to minimize the errors caused by an errant interpretation of mutation is by using a suitable substitution model. Molecular phylogenetics methods rely on a substitution model that encodes a hypothesis about the relative rates of mutation at various sites. The substitution models aim to correct for differences in the rates of mutations in nucleotide sequences. The use of substitution models is necessitated by the fact that the genetic distance between two sequences increases linearly only for a short time after the two sequences diverge from each other. The longer the amount of time after divergence, the more likely it becomes that two mutations occur at the same nucleotide site. Simple genetic distance calculations will thus undercount the number of mutation events that have occurred in evolutionary history. The extent of this undercount increases with increasing time since divergence, which can lead to the phenomenon of long branch attraction, that is wrongly marking two distantly related but convergently evolving sequences as closely related.

Once the sequences are aligned, there are various methods for phylogenetic analyses, which can be performed (Holder & Lewis, 2003). The methods for calculating phylogenetic trees can be broadly divided into two categories. The first category consists of distance matrix based methods, also known as clustering or algorithmic methods (e.g. Neighbor-joining, Fitch-Margoliash, UPGMA). The latter category includes discrete data (or character) based methods, also known as tree searching methods (maximum parsimony, maximum likelihood and Bayesian methods).

3.2.4 Distance Matrix Methods

Distance between two sequences is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Once the distances between all pairs of samples have been determined, the resulting triangular matrix of differences is submitted to some form of statistical analysis. Distance methods attempt to construct an all-to-all distance matrix from the sequence query set describing the distance between each sequence pair. The resulting dendogram created from this distance matrix is examined in order to see whether the samples cluster in the way that would be expected from current ideas about the taxonomy of the group. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as the basis for progressive and iterative types of multiple sequence alignments. There are different tree building methods from a distance matrix. Neighbor-joining methods, for tree building, apply general data clustering techniques using genetic distance as a clustering metric. The simple neighbor-joining method produces unrooted trees. It does not assume a constant rate of evolution (i.e., a molecular clock) across lineages. Its relative, UPGMA (Unweighted Pair Group Method with Arithmetic mean) produces rooted trees and requires a constant-rate assumption. That is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal. Another tree building algorithm, the Fitch-Margoliash method uses a weighted least squares method for clustering based on genetic distance. Here, closely related sequences are given more weight to correct for

the increased inaccuracy in measuring distances between distantly related sequences. The distances calculated by this method must be linear; the linearity criterion for distances requires that the expected values of the branch lengths for two individual branches must equal the expected value of the sum of the two branch distances. This property applies to biological sequences only when they have been corrected for the possibility of back mutations at individual sites. This correction is done through the use of a substitution matrix such as that derived from the Jukes-Cantor model of DNA evolution. The distance correction is only necessary in practice when the evolution rates differ among branches. In order to root unrooted trees, a standard practice in distance matrix methods is, to include an outgroup sequence. The outgroup should appear near the root of the tree. For that this sequence should be moderately related to the sequences in the query set. An outgroup sequence which is too closely related to the query set defeats the purpose and one which is too distantly related only adds to the noise.

3.2.5 Character based methods

Character-based methods are based on the idea of generating all the possible tree topologies with the input sequences and then searching among these trees for the tree that best matches the data given some criteria. Maximum parsimony (MP) method is based on shared and derived characters. It does not reduce sequence information to a single number. It works with original data (alignment) and tries to provide the information about the ancestral sequences. The principle of this method is to find a tree with the smallest number of evolutionary changes. It is based on the hypothesis that evolution prefers the smallest number of mutations. MP is a method

of identifying the potential phylogenetic tree that requires the smallest total number of evolutionary events to explain the observed sequence data. Some ways of scoring trees also include a "cost" associated with particular types of evolutionary events and attempt to locate the tree with the smallest total cost. The most naive way of identifying the most parsimonious tree is simple enumeration - considering each possible tree in succession and searching for the tree with the smallest score. However, this is only possible for a relatively small number of sequences or species because the problem of identifying the most parsimonious tree is known to be NP-hard. The MP method is particularly susceptible to long branch attraction due to its explicit search for a tree representing a minimum number of distinct evolutionary events. The Sankoff-Morel-Cedergren algorithm was one of the early methods to simultaneously produce an MSA and a phylogenetic tree for nucleotide sequences (Sankoff, 1973). The method uses a maximum parsimony calculation in combination with a scoring function that penalizes gaps and mismatches, thus favoring the tree that introduces a minimal number of such events. The sequences at the interior nodes of the tree are scored and summed over all the nodes in each possible tree. The lowest-scoring tree sum provides both an optimal tree and an optimal MSA given the scoring function. One shortcoming of the method is that it is highly computationally intensive. Hence an approximate method is first performed in which initial guesses for the interior alignments are refined one node at a time. Both the full and the approximate methods are in practice done by dynamic programming. More recent phylogenetic tree - MSA methods use heuristics to isolate high scoring, but not necessarily optimal, trees. The MALIGN and POY is such a technique. The MALIGN method uses a maximum-parsimony technique to compute a multiple alignment by maximizing a

cladogram score, and its companion POY uses an iterative method that couples the optimization of the phylogenetic tree with improvements in the corresponding MSA. However, the use of these methods in constructing evolutionary hypotheses has been criticized as biased due to the deliberate construction of trees reflecting minimal evolutionary events (Simmons, 2004). Maximum Likelihood method relies on the fact that, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences. However, it formally requires search of all possible combinations of tree topology and branch length and hence it is computationally expensive to perform on more than a few sequences.

3.2.6 Disadvantages of MSA in rebuilding phylogeny

One of the main applications of a successful Multiple Sequence Alignment is in constructing phylogenetic trees. Here we consider the disadvantages of using an MSA to construct a phylogenetic tree.

A major difficulty that arises is in the selection of conserved sequences for the purpose of alignment. For obtaining a meaning phylogeny from a specific gene, the genes considered in all the organisms should be orthologous genes. Note that one must distinguish between gene trees and species trees due to the presence of orthologous and paralogous genes. Orthologous genes are homologous genes that are

truly the same and are separated by speciation alone. Paralogous genes are homologous genes that resulted from a gene duplication in the parent organism. The genes used in analysis can have more than one copy in a particular genome (i.e. paralogous genes). This provides a cause for uncertainty such as which copy to select for the construction of the tree. Yet another difficulty arises when different copies are found in different species. The resulting phylogenetic tree will obviously not provide the correct relationships between the species In addition to these, there are xenologous genes, those that are transferred from one organism to the other horizontally. Taking a single gene sequence into consideration does not provide the best resolution in any phylogenetic tree. For constructing a reliable tree multiple genes are required (Sandersson and Driskell 2003). Note that, in case of less conserved sequences, it can be difficult to find orthologs from all the species under study.

The next problem is the alignment itself. Automatic alignments may fail to correctly identify conserved regions, whereas manual alignments allow this, but they are much more laborious. Many a times, automated alignment methods encounter the local minimum crisis. An alignment with minimum number of penalties is the optimum one. However, often automated alignment methods do not reach this global optimum and are trapped at local optima. The alignments have to be then viewed by the user and should be manipulated by him manually to get the optimal alignment.

In a gapped pairwise alignment, the gap can be attributed to insertion in one sequence or deletion in the other. The problem augments geometrically in an MSA with more than two sequences. Similarly, when there is a mismatch between two

characters at a particular position, the event may be a single mutation or more than one mutation at the same site. Even when two aligned characters are similar, in reality, there might have been a mutation at that point in one sequence, and later another reverse mutation for the same character. All these are to be taken into account for assembling an exact phylogenetic tree. Only an ideal evolutionary model can tackle this issue perfectly. However, this is near to impossible.

Alignments often become meaningless and are difficult to perform when the percentage of similarity between the sequences is very low. A minimum threshold is needed to produce a significant evolutionary alignment.

Another crisis with the automated alignment programs is the time required for their execution. Dynamic programming methods for alignment are computationally expensive. The problem of aligning two sequences with '*n*' and '*m*' character positions respectively is of the order of *n* x *m.* The computational time rises geometrically as the product of sequence length as more and more sequences are added for an MSA.

In order to handle this computational expense, a heuristic method called Progressive alignment deviced by Feng and Doolittle (1987) is employed. The method clearly does not guarantee the optimal alignment for a set of sequences. The most successful implementation of this progressive alignment method is used in the CLUSTALX (Des Higgins) programme. Nevertheless, CLUSTALX has no way of quantifying whether or not the alignment is good. Moreover the user, without using manual visualization, does not even know whether the alignment is correct. Further this method has no objective function. CLUSTALX initially computes pairwise

distance between sequences. Sequences which have very less distance between them are placed as neighbours in the guide tree. These neighbours are aligned in the early stages. This aligned pair is treated as a single sequence for latest comparison. The programme does not allow for subsequent corrections to be made later in the alignment process. Suppose if there is a gap in an initial alignment, it is not possible to correct it, even if a correction may be needed subsequently. Hence there is a possibility of a suboptimal alignment when the program reaches a local minimum. Thus the tree created can be of sub-optimal tree topology. The technique in overall can be a very good estimate or an equally poor one. Many a times the result has to be improved manually by visualizing the alignment. The user has to make a judgment whether the regions aligned are reliable or not.

Once the data are aligned we use distance matrix based methods or character based methods for phylogenetic analysis. The distance matrix format after computing an MSA enables rapid tree building compared to character based methods. However, there are two major shortcomings. Firstly, there is a loss of information when going from molecular data to distance data. Secondly, it is not trivial to choose a good distance measure among the many distance measures that exist. The distance measures must among other things compensate for the possibility that there may be multiple substitutions at a particular site of the alignment which, if not corrected for, would result in an underestimation of the evolutionary distance between two sequences.

One drawback of character based methods, such as Maximum Likelihood (ML) and Maximum Parsimony (MP) method, is that they are very time consuming

since the number of tree topologies grows very rapidly with the number of sequences. There are fortunately heuristic search methods that avoid having to evaluate all the trees but, nevertheless, the need to evaluate many different trees makes the optimality methods considerably more time consuming than the distance based clustering methods. Considerably more time consuming means that for clustering methods the result is ready within seconds whereas for optimality methods the result may easily take several minutes and maybe hours (depending on the number of sequences and their length). A specific problem of MP is long-branch attraction. It is caused by the fact that rapidly evolving lineages are considered closely related, regardless of their true evolutionary relationships (Bergsten 2005). Differential rates of substitution among lineages or breaking up long branches by adding taxa that are related to those with the long branches have to be employed to minimize this miscalculation. Maximum likelihood methods require a model of evolution. It has the disadvantage that the use of inadequate likelihood models can lead to faulty interpretation in real data sets. The method is computationally expensive and is not suited for large number of sequences since it has to search all possible combinations and tree topology.

3.2.7 Phylogeny based on nucleotide sequences

Molecular sequencing of macromolecules, such as genes and proteins, have surpassed morphological and other organismal characters as the most popular forms of data for phylogenetic analyses. The literature contains different opinions about whether nucleotide sequences or amino acid sequences should be used to decipher ancient phylogenetic relationships. Slowly evolving sequence characters have

generally been favored over faster evolving characters and this has led to the preferential use of amino acid characters over nucleotide characters.

However, Simmons et al. (2004) talks of taking into consideration at least six factors before deciding whether to use nucleotide sequences or amino acid sequences for computing phylogeny. The amino acids are composite characters formed by combining three separate nucleotide characters. This causes loss of hierarchical information while using amino acid characters. The second factor is that those amino acids that are specified by more than one codon can appear to be convergently derived on a tree when the underlying nucleotides are not. The silent substitutions, which generally occur more frequently than replacement substitutions, are ignored if we only consider amino acid characters. Studies have reported greater phylogenetic signal at the third codon position than the first or second codon position combined. The above three factors highlight the advantages of selecting nucleotides for molecular phylogeny. The fourth factor depends on the property that, amino acid characters have an increased potential character-state space relative to nucleotide characters, which makes convergence less likely to occur (i.e., 20 vs. four possible character states). However, there are functional constraints on the protein that effectively limit the character-state space for amino acid characters (Dayhoff et al., 1972; Miyamoto & Fitch, 1995; Naylor and Gerstein, 2000). Convergent changes in base composition can cause the same problem as long-branch attraction (Felsenstein, 1978), wherein unrelated organisms that derive similar compositional biases can be resolved as sister groups (Lockhart et al., 1992). The fifth factor is that amino acid characters are less sensitive to such changes in base composition. This is because shifts in nucleotide composition are concentrated at third codon positions at which most substitutions are

silent. However, amino acid composition is also affected by changes in base composition. Moreover, simulation studies have indicated that the compositional heterogeneity among sequences must be extreme in order for phylogenetic reconstruction methods to fail, suggesting that this is less of a problem than previously believed (Conant and Lewis, 2001; Rosenberg and Kumar, 2003), except in cases of short internal branches (Jermiin et al, 2004). The sixth factor is that amino acid characters are not as subject to saturation as faster evolving silent substitutions because amino acid character-state changes are only caused by replacement substitutions. Simmons et al. (2004) conclude, by mentioning the advantage of greater potential phylogenetic signal for nucleotide characters, and the greater observed character-state space and lower heterogeneity of amino acid characters, which was confirmed based on a broad selection of protein-coding loci. Although the greater potential phylogenetic signal for nucleotide characters was found to be enormous, the greater observed character-state space for amino acid characters was less impressive. Agosti et al. (1996) consider that the most judicious approach may be to incorporate phylogenetic signal from both nucleotide and amino acid characters in a simultaneous analysis.

3.2.8 Significance of silent mutations

A basic assumption underlying the use of the amino acid sequence variation in deducing evolutionary relationships is that selection takes place predominantly at the level of protein function which is, of course, decided by the amino acid sequence. This in turn implies that nucleotide character changes that lead to synonymous codons can take place without selectional constraints. However, recent studies contradict this argument by bringing out evidence of functional changes produced by synonymous

single-nucleotide polymorphisms (SNPs). A Synonymous SNP is a mutation that changes the triplet, but leaves the amino acid unchanged. Hence they do not produce altered coding sequences, and therefore they are not expected to change the function of the protein in which they occur. Sarfaty et al. (Sarfaty et al., 2007) in a recent work studied two synonymous SNPs that crop up frequently in a human protein that pumps toxins out of cells. They found that one of the SNPs affects the timing of co-translational folding and insertion of a protein into the membrane. This in turn alters the structure of substrate and inhibitor interaction sites. Similar mRNA and protein levels, but altered conformations, were found while comparing the wild-type and polymorphic one. This study implies that naturally occurring silent mutations may also have such an effect and hence cannot be written off. Silent mutations are known to have other effects too. They can change the way that RNA, the molecule that bridges DNA to protein production, is cut and spliced together. Pagani F. et al. (2005) showed that many silent mutations in the gene responsible for the lung disease cystic fibrosis can cause splicing changes that inactivate the protein. These works point out that the silent mutations are not so silent after all. It reveals that selection can constrain changes at the nucleotide sequence level itself so that the so-called neutral changes at the third codon position cannot be really considered as neutral. Variations like this in the DNA sequence can be taken into account while constructing phylogeny only if the nucleotide sequence itself is considered. It is important to keep in mind that it is the changes in the nucleotide sequence that ultimately bring about evolutionary changes.

3.2.9 Importance of considering the whole genome

Any alteration in the nucleotide sequence in the non-coding region cannot be taken into account if only amino acid sequences are taken into consideration. This 'junk' DNA, as it was previously considered, is now proved not be junk at all (Makalowski, 2003). Another important factor to consider is that it is not only individual protein function but also the regulation of their expression that contributes to the making of an organism which is more than just the sum of several (be it even millions of) amino acid sequences. Ideas about the evolutionary significance of noncoding mutations are nearly as old as the discovery of regulatory sequences themselves. Soon after publishing their ground-breaking paper describing the *lac* operon in 1961 (Jacob & Monod, 1961). In another paper, Monod and Jacob (1961) speculated about the unique role that mutations in *cis*-regulatory regions might have during the course of evolution. Therefore "tinkering" with gene regulation has been recognized as a particularly powerful mode of evolution (Jacob, 1977), and gene regulation involves non-coding nucleotide sequences as well. G.A Wray in his review work (Wray, 2007) even suggests that some phenotypic changes are more likely to result from cis-regulatory mutations than from coding mutations. Further he suggests that mutations in *cis*-regulatory regions are often co-dominant in contrast to, many or most coding mutations which are recessive. Consequently it is obvious that we lose important evolutionary information by looking only at amino acid sequence changes without considering the nucleotide sequence changes.

Phylogenetic trees based on amino acid sequences make the trees gene (protein) specific. Basing the analysis on a single feature, such as a single gene or protein, is often unreliable since such trees constructed from another unrelated data

source such as another gene often differ from the first. Hence great care is to be taken while inferring phylogenetic relationships among species. This is most true of genetic material that is subject to horizontal gene transfer and recombination, where different haplotype blocks can have different histories. In general, the output tree of such a phylogenetic analysis is an estimate of the feature's phylogeny (i.e. a gene tree) and not the phylogeny of the taxa (i.e. species tree) from which these features were sampled, though ideally, both should be very close. Further, in many organisms, endosymbionts have an independent genetic history from the host and mistakenly selecting a protein from an endosymbiont would give fictitious results. Hence, serious phylogenetic studies have to take the pain of using a combination of genes that come from different genomic sources (e.g., from mitochondrial or plastid vs. nuclear genomes), or genes that would be expected to evolve under different selective regimes, so that result would not show a false homology.

3.3 The concept of Genome Signature

Over the last decade considerable evidence has come up that DNA sequences evolve under the constraint of a "genome signature" which is related to the frequency of occurrence of short oligonucleotides in the DNA sequence (Karlin and Ladunga, 1994; Edwards et al., 2002; Qi et al., 2004; Dehnert et al., 2005; Wang et al., 2005; Chapus et al., 2005). Oligonucleotide frequency differences between species are seen to change too slowly to be purely the result of random mutational drift. This slow pattern of change reflects the direct or indirect action of purifying selection and the presence of functional constraints. In fact, the constraints that slow down the divergence of genome signature could play an important role in determining evolutionary distances. The concept of a genomic signature was introduced with the

observation of species-type specific Dinucleotide Relative Abundance Profiles (DRAPs) by Karlin and Burge in 1995 (Karlin & Burge, 1995). Dinucleotides were identified as the subsequences with the greatest bias in representation in a majority of genomes. As the name suggests, the relative abundance of dinucleotides in a genome was measured. The computation was done by dividing the frequency of each dinucleotide by the frequency of its component monomers. This was performed so as to nullify the effect caused by monomer frequency, such as being GC rich, since they are hypothesized to be not playing a part in evolution. The key observation behind the genomic signature concept is that DRAP of different DNA sequence samples from the same organism are generally much more similar to each other than to those of sequences from other organisms. In addition, closely related organisms generally have more similar DRAPs than distantly related organisms. It was concluded from these observations that the DRAP values constitute a genomic signature of an organism. Later it came to be known that DRAP is one particular genomic signature contained within a broader spectrum of signatures (Wang et al., 2005). The invariance of the genome signature over different parts of the same genome indicates that there must be genome-wide processes that impose limits upon oligonucleotide composition in any particular genome. The mechanisms by which the genome conserves its signature are still in the realm of speculations. If these indications are put on stronger scientific foundations, our current evolutionary models will have to be modified considerably.

Much work is being done in the field of genome signatures. Campbell et al. (1999) compared genomic signatures of prokaryote, plasmid, and mitochondrial DNA. Deschavanne et al. (1999) observed that subsequences of a genome exhibit the main characteristics of the whole genome. Further, Deschavanne et al. (2000) showed

that word usage in short fragments of genomic DNA (as short as1 kb) is similar to that of the whole genome, thus providing a strong support to the concept of genomic signature. Gentles and Karlin (2001) looked at the genome signature of various eukaryotes. Sandberg et al. (2001) proposed a method to classify sequence segments using genomic signatures. They used a Bayesian classifier to analyse bacterial and archaeal genomes and investigated the possibility of predicting the genome of origin from a genome fragement. Genome signatures were used in phylogenetic analysis by Edwards et al. (2002). In fact, the remarkable agreement between phylogenetic relationships based on genomic signature distances and some of the well-established phylogenetic relationships is a strong indication that sequence diversification does take place under a strong constraint to conserve the signature. Bush and Lahn (2006) in a recent paper show that the distances based on di-nucleotide and tetra-nucleotide frequencies from any part of the genome, and octa-nucleotide frequencies from the promoter regions are correlated to evolutionary time. They suggest that, while genome-wide processes like DNA replication and repair constrain the divergence of the shorter DNA words, the octa-nucleotide frequencies may be getting constrained by the slow evolution of the gene regulatory machinery. PhyloPythia, a phylogenetic classifier using multi-class Support Vector Machine classifier with the oligonucleotide composition of genome fragments as input was built by McHardy et al. (2007). The method allows for classification of genomic fragments for different taxa and more importantly for previously unseen fragments which originate from novel organisms. The remarkable agreement between phylogenetic relationships based on genomic signature distances and some of the well-established phylogenetic relationships (Qi et al., 2004, Dehnert et al., 2005, Wang et al., 2005) is a strong

indication that sequence diversification does take place under a strong constraint to conserve the genomic signature. It is to be remembered that this signature is present in the entire genome and not restricted to coding regions alone. The below graph (Figure 3.1) is drawn based on a paper by Dufraigne et. al (2005), where the *y*-axis is formed by all possible tetranucleotides and the *x*-axis runs through the genome. The colour gradient represents the frequency of a particular tetramer (*y*) at a particular region (*x*) in the genome. The figure shows the tetranucleotide patterns in the 1million nucleotide long genome of *Rickettsia Prowazekii.* Each horizontal line is an indication of the uniformity of a particular tetramer pattern throughout the genome. This visual representation shows that the genome signature is conserved in all over the entire genome.

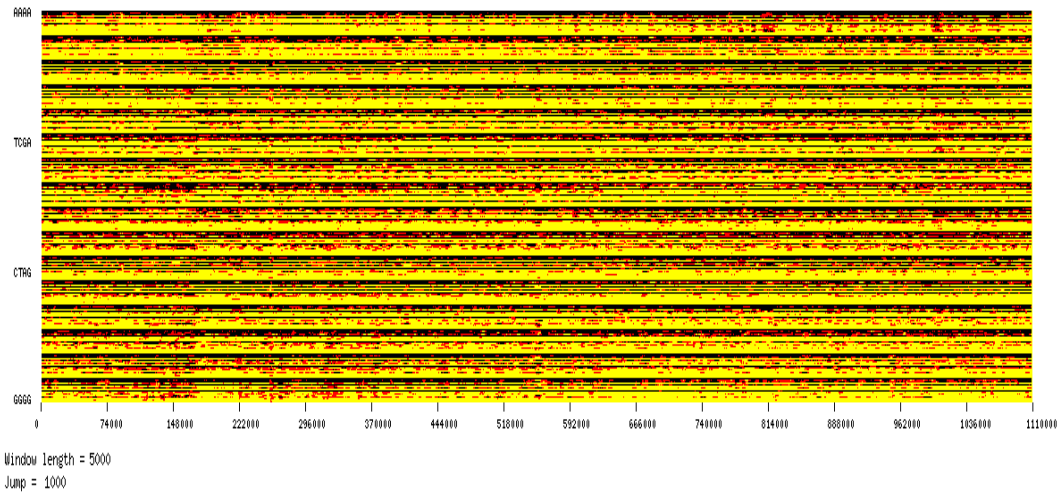


Figure 3.1 - Tetramer frequency pattern in Rickettsia Prowazekii genome - 1,111,523 nt

These observations open to question the tacit assumption in most phylogenetic analyses that selection acts at the level of protein function only. The amino acid sequence variations cannot be regarded as the sole foundation that holds the key to

evolutionary relationships. Unless the phylogenies based on amino acid sequence variations are supported by those based on nucleotide sequences, there is room for legitimate doubt. This is especially true when chances of convergent evolution are high. In summary, the phylogeny reconstructed using various methods morphological, molecular or otherwise, should support each other. If the trees obtained are not alike, then the conclusion obtained from any one of them can be inaccurate. In the following we explore whether and to what extent the phylogenetic relationships based on genome signature as represented by the Frequency Chaos Game Representation of the sequence have the capability to reproduce established relationships. We also investigate how the estimated trees differ when they are based on different representations of the genome signature.

3.4 Building phylogenetic trees using FCGR

In order to build a phylogenetic tree based on genome signature, two quantitative representations are required:

1. A representation of the genome signature
2. A representation of the distance (dissimilarity) between two signatures

Genome signature is represented as the frequency profile of oligonucleotides of a particular length i.e. frequency profile of n-mers in the genome where n =1,2,3,4 etc. As described in Chapter 1 the Frequency Chaos Game Representation (FCGR) of order 'n' is the frequency profile of all oligomers of length 'n' in the genome. Depending on the value of n chosen we can make different representations of the genome signature. Another variation in the representation of genome signature is

brought by whether the frequency profile is corrected for the biases due to the base composition of the genome. For this purpose the frequency of a particular oligomer is divided by the frequency of its component monomers. This corrected profile is called the relative frequency profile. In this work we have used FCGRs of order 2 to 10 as genome signature representations and observed the differences in the phylogenetic classifications caused by the difference in the order of the FCGR.. We have also investigated whether correction for the base composition (i.e. relative frequency profile) makes a difference to the classification.

A number of distance measures are possible between two FCGRs (mentioned in chapter 1). Among these, we use the statistical distance based on weighted correlation coefficient mainly because it is more suited to comparing genomes of very different lengths. This measure proposed by Almeida et al. (2001) is determined as follows. Let the two sets of FCGR quadrants be x and y with x_i and y_i representing the frequency in the i^{th} quadrant. The weighted Pearson correlation coefficient is calculated as follows:

$$nw = \sum_{i=1}^{N} x_i y_i$$

$$xw = \frac{\sum_{i=1}^{N} x_i^2 y_i}{nw} \qquad yw = \frac{\sum_{i=1}^{N} y_i^2 x_i}{nw}$$

$$sx = \frac{\sum_{i=1}^{N} (x_i - \bar{x}w)^2 x_i y_i}{nw} \qquad sy = \frac{\sum_{i=1}^{N} (y_i - \bar{y}w)^2 x_i y_i}{nw}$$

$$rw_{x,y} = \frac{\sum_{i=1}^{N} \frac{x_i - \bar{x}w}{\sqrt{sx}} \frac{y_i - \bar{y}w}{\sqrt{sy}} x_i y_i}{nw}$$

The distance between the sequences is defined to be $d = 1 - rw_{x,y}$ and this value ranges between 0 and 2 since correlation ranges between -1 and 1. Note that the distance 0 corresponds to perfect correlation between the sequences, i.e. the signatures are identical. An advantage of using such a statistical distance based on correlation coefficient is that sequences of highly varying length can be compared without standardizing the CGRs initially. Another benefit in using weighted correlation coefficient is that, the importance of each quadrant is made proportional to its magnitude. Hence a quadrant with a significantly high occurrence of a particular oligonucleotide is given more importance while determining similarity. Thus the overrepresentation of any particular oligomer gets more highlighted which in turn serves to increase the correlation value for the sequence pair. However, this method has the disadvantage that under-representation of a particular oligomer does not get its deserved priority. Studies by Wang et al. (2005) show that this method produces fairly reliable phylogenies and hence we select it for computing phylogeny based on genomic signature.

3.5 Results and Discussion

3.5.1 Exploration of the potential of genome signature as a phylogenetic signal

We present first of all some of the results obtained from phylogenetic classification using genome signature, represented by 3rd order FCGR, as the phylogenetic signal. We present here phylogenetic classifications obtained from

different sources of DNA sequences viz. whole genomes and chromosomes, mitochondrial genomes and the ubiquitous 16srRNA gene.

3.5.1.1 Phylogenetic Classification using Genome Signature of whole genomes and chromosomes

It is clear from Section 3.2 that phylogenetic signal is contained not only in proteins and protein-coding genes, but also in the whole genome sequence. In an alignment based method comparison of whole genomes or chromosomes would not make any sense since there would not be any significantly similar regions on these sequences to be aligned. At the most, there would be similar genes in the two organisms, but that too would not yield a meaningful alignments since the genes are distributed throughout the genome or chromosome. Alignment-free methods like genome signature comparisons are the only ways by which the phylogenetic information contained in the whole genome or chromosomes can be considered for deducing phylogenetic relationships.

The chromosomes of nine belonging to the class "Metazoa" were compared to each other. The first five chromosomes were taken from each organism, with each chromosome taken as if it were a separate organism. The organisms are H.Sapiens, P.Troglodytes, R.Norvegicus, M.Musculus, B.Taurus, G.Gallus, A.Gambiae, D.Melanogaster and C.Elegans. The pair-wise Pearson distances mentioned earlier were computed between the organisms. The resulting distance matrix was supplied as input to the tree drawing software, 'PHYLIP'. The KITSCH algorithm (Fitch-Margoliash and Least Squares Methods with evolutionary clock) was used and the resulting unrooted tree is drawn**.** The unrooted distance tree obtained using 3^{rd} order FCGR show interesting results as observed in Figure 3.2.

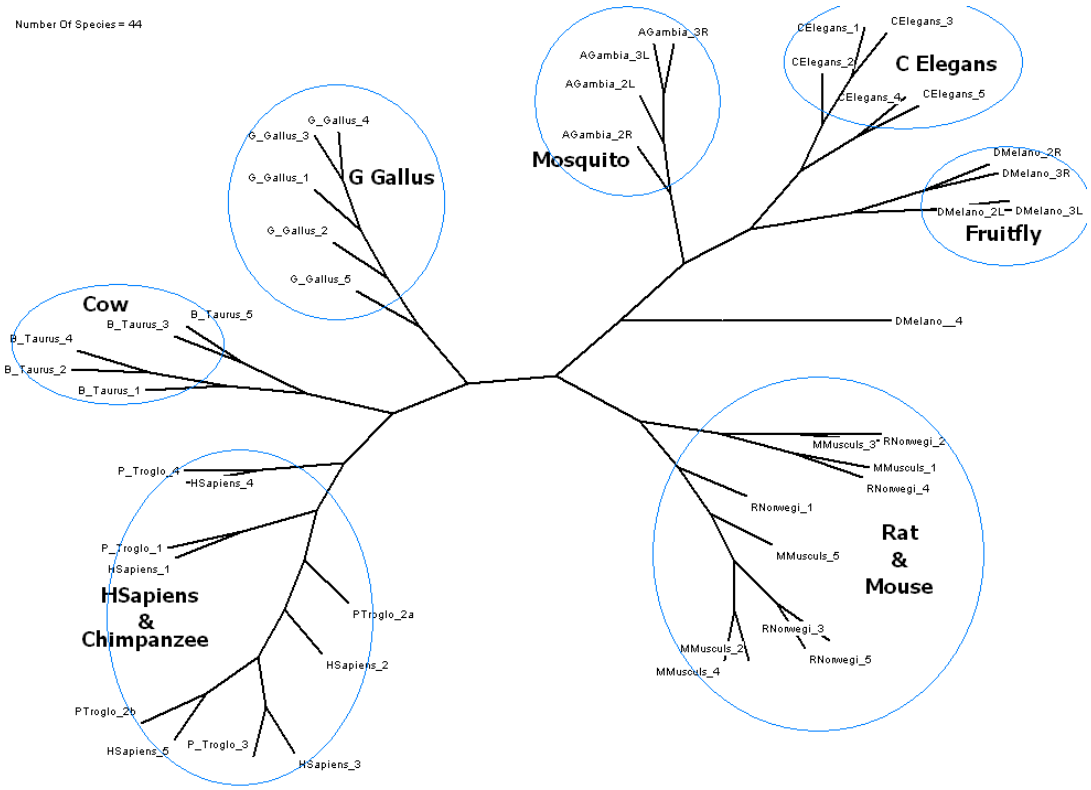


Figure 3.2 - Metazoan tree based on 3rd order FCGR

It can be noticed that the vertebrates are clearly separated from non-vertebrates. Similar organisms (Rat and Mouse, Human and Chimpanzee) cluster together. Chromosomes of each organism are seen to form separate clusters except for the chromosomes of close organisms like Human and Chimpanzee or Rat and Mouse which may be too close to segregate into separate clusters. This result shows a distinctive property of genomic signature where all parts of a genome of an organism show similar signature, irrespective of the chromosome from where the sequence is taken. The similarity in the genomic signature is the only reason that chromosomes of the same organism cluster together. Similar organisms are seen to be close to each

other on the tree. This distinctive property of genomic signature makes it a valuable tool for looking into phylogenetic relationships from an entirely new perspective.

Going to lower organisms whole genomes of sixty six species of bacteria were classified based on genome signature represented by the 3rd order FCGR (Figure 3.3). There is a great deal of clustering of similar species of bacteria showing the presence of phylogenetic signal in the genome signature. However the clustering is not faultless. The inter-species relationships do not reflect established relationships – for example ε-proteobacteria and the Rickettsiales are seen to cluster separately from the rest of the proteobacteria. However, bacterial phylogeny is notoriously ambiguous and different methods and different genes are known to give very different results. It is still unclear how the different main groups are related to each other and how they branched off from a common ancestor (Gupta et al., 2002).

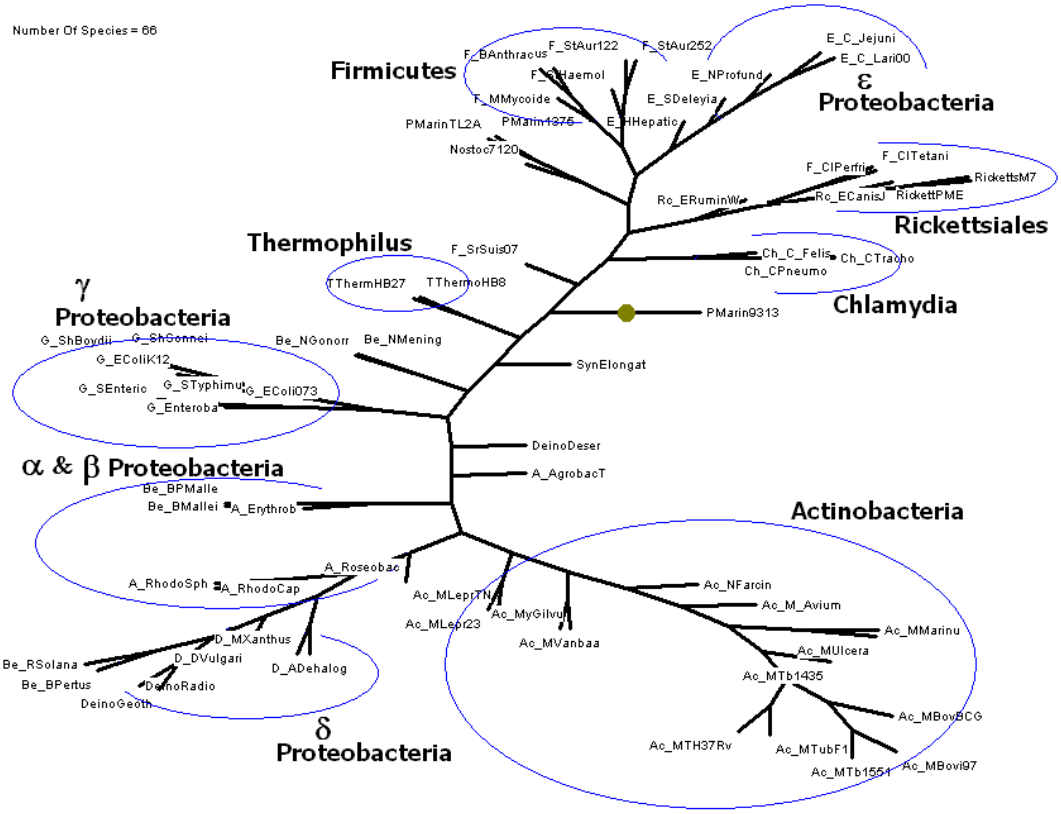


Figure 3.3 - Bacterial phylogenetic tree based on 3rd order FCGR

3.5.1.2 Phylogenetic Classification using Genome Signature of mitochondrial genomes

A phylogenetic tree is created for 28 organisms belonging to the class "Eutheria" using the signature of the mitochondrial genome of all the organisms. The genome signature is represented by the 3rd order FCGR. One can clearly notice the clustering of the primates as a group and it can see that many closely related organisms are paired together (Figure 3.4). However there are errors like incorrect grouping of rodents. The fact that phylogenetic signal is present in the mitochondrial sequence signature is clearly indicated by the clustering of closely related organisms.

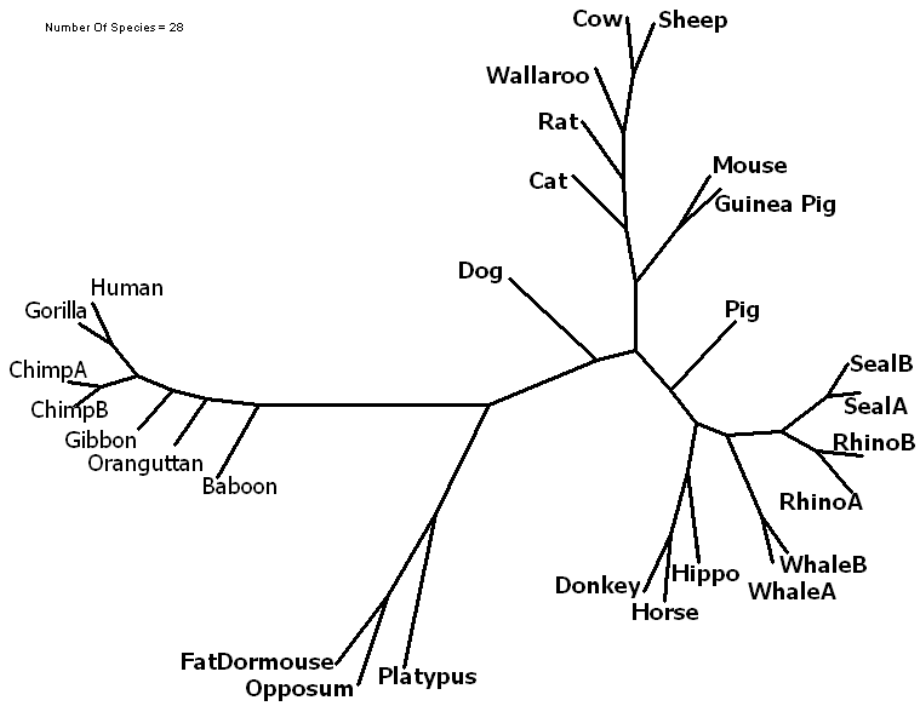


Figure 3.4 - Eutherian tree based on 3rd order FCGR on mitochondrial genomes

Further, mitochondrial genomes of a larger class of organisms, "Eukaryota" are examined for their phylogenetic signal. The tree based on 3rd order FCGR was constructed for 50 eukaryotic mitochondrial genomes. Various groups such as Vertebrates, Lower animals, Fungi, Lower plants, Higher plants can be seen *(*Figure 3.5*)* to form separate clusters in the resulting phylogenetic tree.

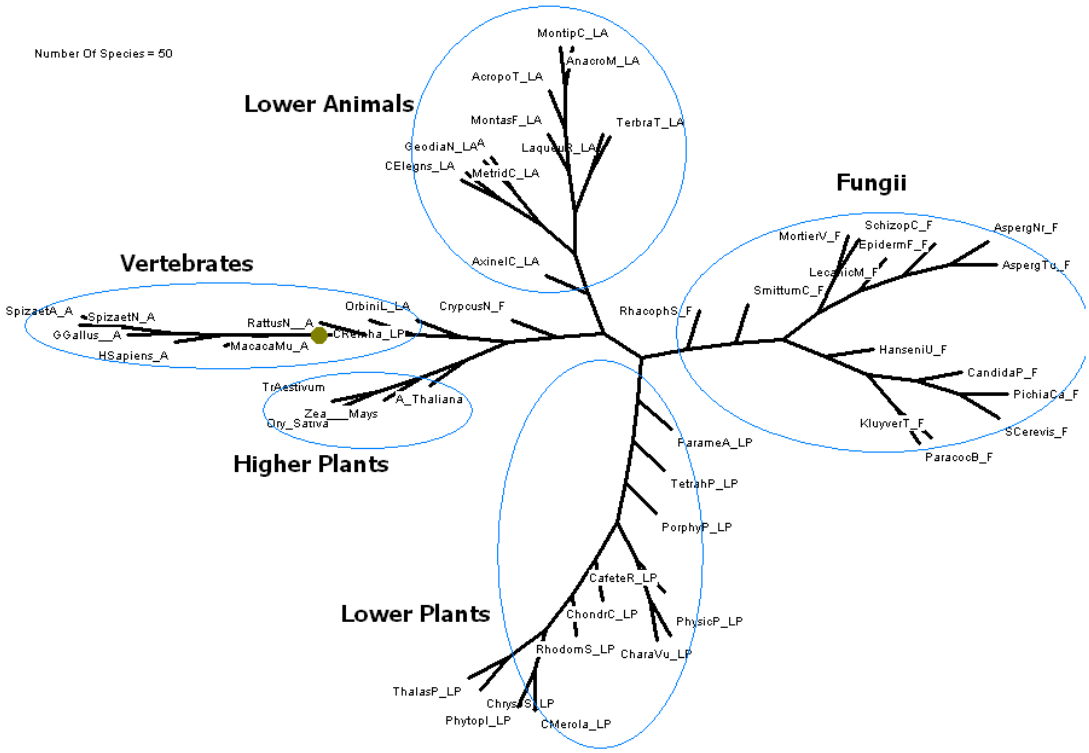


Figure 3.5 - Tree based on 3rd order FCGR of mitochondrial genomes of a larger class, Eukaryota

The above results indicate that that there is a strong phylogenetic signal in the genome signature of mitochondrial genomes. Since mitochondrial genomes of different organisms contain very different sets of genes it would be very difficult to compare them by alignment based methods. Therefore the genome signature based method has a clear advantage over traditional methods in this respect.

3.5.1.3 Phylogenetic Classification of 16SrRNA genes based on relative FCGR

The mitochondrial 16S rRNA gene fulfils the requirements for a universal DNA bar-coding marker. The gene conveys sufficient phylogenetic information to assign species to major taxa. We wanted to find out if the phylogenetic information is carried not only by the sequence itself but also by relative abundance profile of the

oligonucleotides in the gene sequence. We took the 16S rRNA gene of 66 different organisms from eukaryotes, bacteria and archaea. The genome signature was computed by dividing the 3rd order FCGR of the gene sequence by the monomer frequency. The Neighbour-Joining algorithm was used and the resulting unrooted tree can be observed in Figure 3.6. It can be seen that the three major kingdoms (Eukaryotes, Bacteria and Archaea) are clearly demarcated in the picture in accordance with established phylogeny.

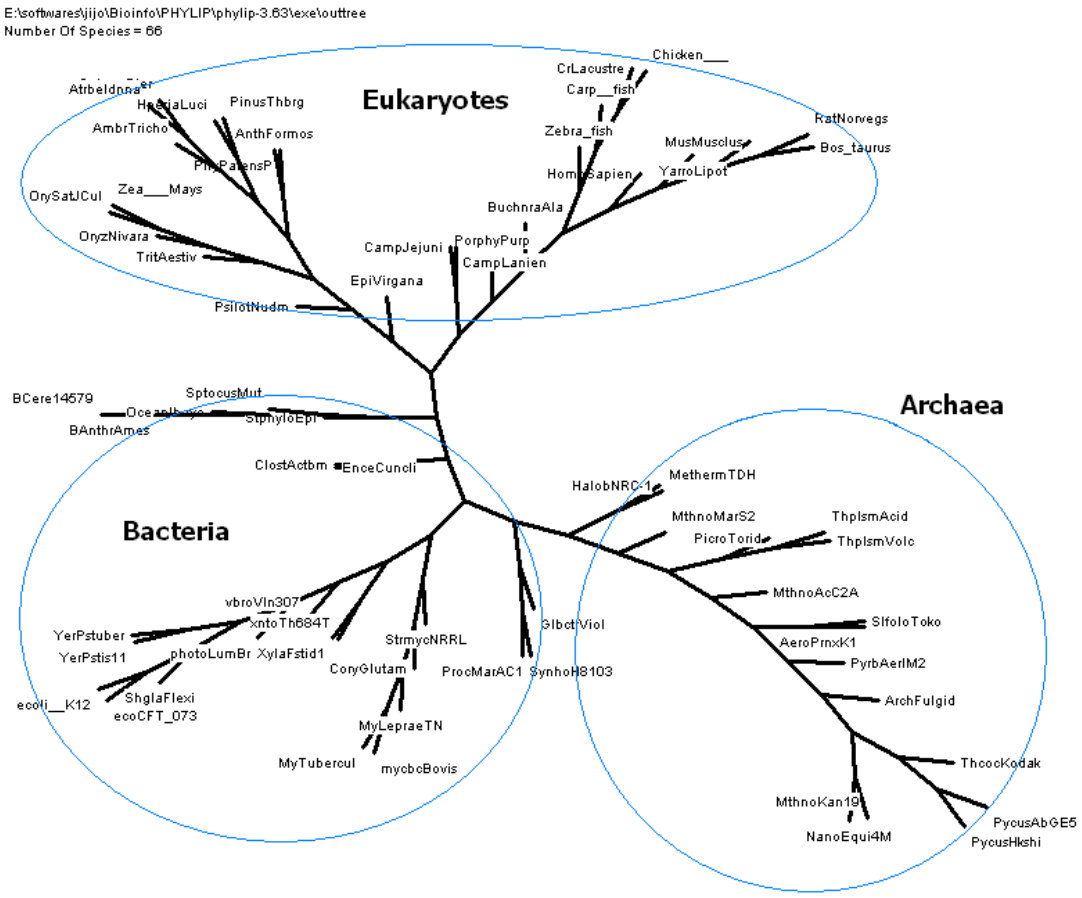


Figure 3.6 - 16SrRNA tree based on trimer-relative frequency

3.5.2 Effect of the order of FCGR on phylogenetic classification

All the above classifications were based on a particular representation of the genome signature namely the 3rd order FCGR. Further, we investigate whether changing the order of the FCGR makes a difference to the classification. We also see whether compensating for biases created by base composition makes a difference to the classification by dividing the FCGR by the frequency of the component monomers.

3.5.2.1 FCGR order variation on the 'Metazoan' tree

The same set of chromosomes of nine Metazoa, as taken earlier, was taken as test sequence. FCGRs of order 3 to 10 and the bias compensated 3rd order FCGR were taken for phylogenetic tree construction. The resulting trees are shown in the figures 3.7, 3.8 and 3.9.

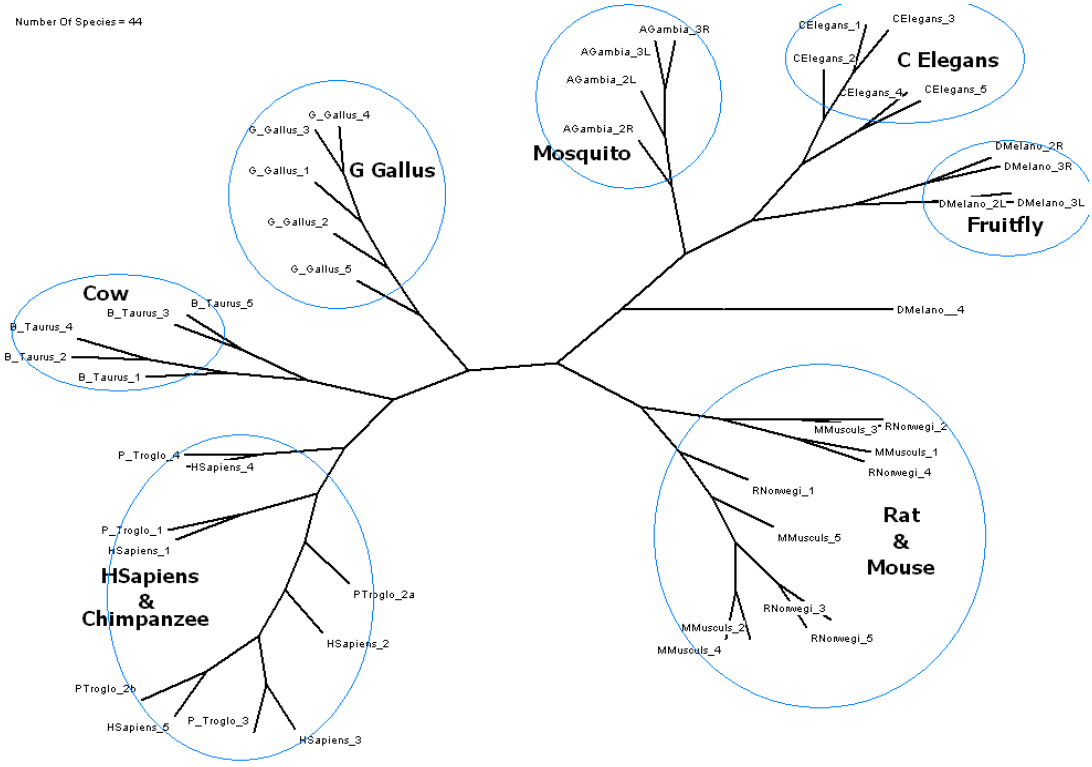


Figure 3.7 - Tree based on 3rd order FCGR of Metazoan chromosomes

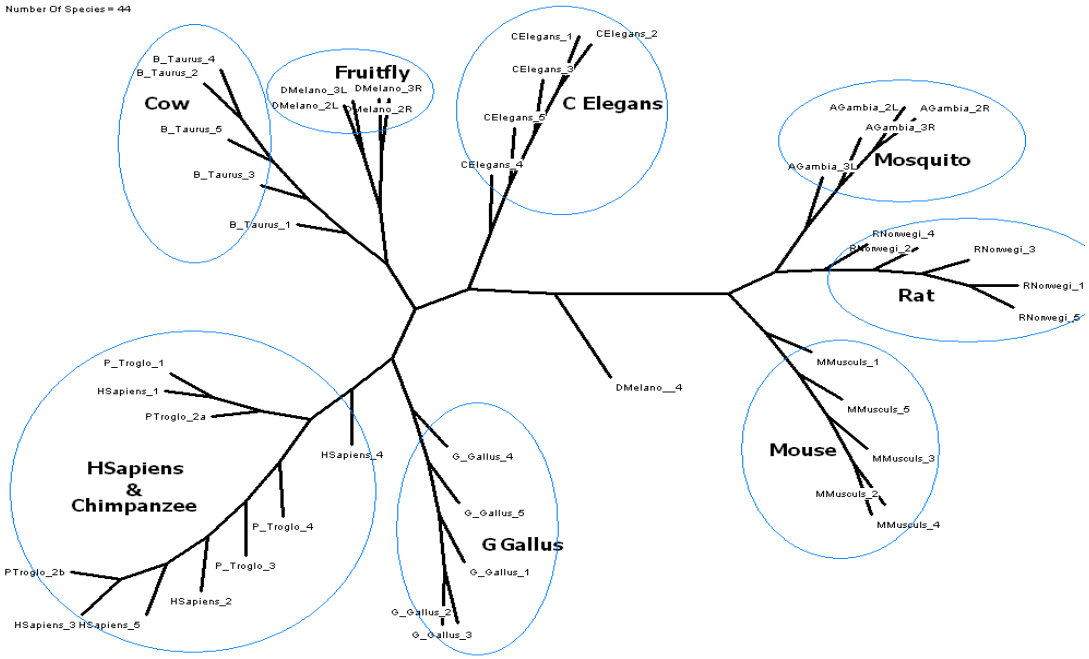


Figure 3.8 - Tree based on 10th order FCGR of Metazoan chromosomes

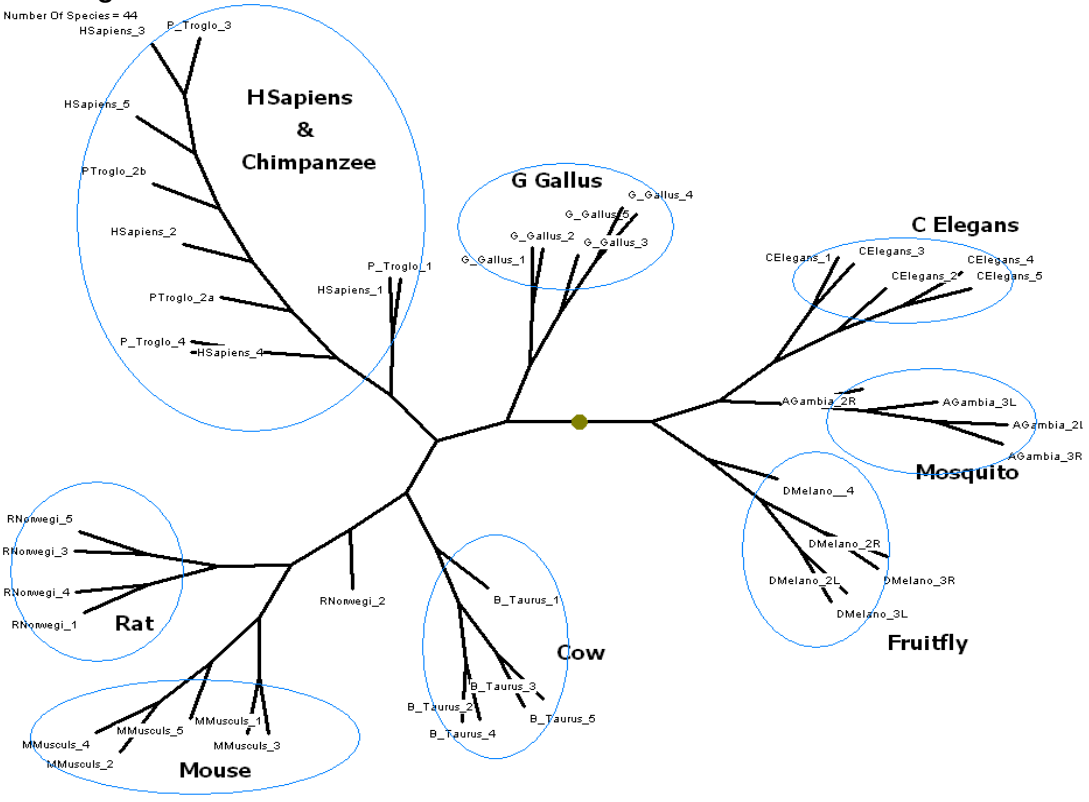


Figure 3.9 - Tree based on on bias-compensated 3rd order FCGR of Metazoan chromosomes

Figure 3.7 shows the unrooted tree obtained by considering the 3rd order FCGR. It can be observed in the 3rd order FCGR tree that the vertebrates and invertebrates cluster separately. Further, closely related organisms such as Rat - Mouse and Human – Chimpanzee are clustered together. However, this happens at the cost of intermingling of chromosomes of these closely related organisms. One can notice that some of the Rat chromosomes are dispersed among the Mouse chromosomes and similar intermingling has occurred between the Chimpanzee and Human chromosomes. In the 10th order FCGR tree the *Rat* and *Mouse* chromosomes are seen to form exclusive organism-wise clusters and there is no intermingling of their chromosomes. However, the inter-species relationships are not depicted well here (for example the vertebrates and invertebrates do not form separate groups in this tree) The bias compensated 3rd order FCGR tree gives the best result among all the trees constructed. Here the organisms including the Rat and Mouse chromosomes form separate groups. The interspecies evolutionary relationships also come out good, showing the invertebrates as a separate group and Rat and Mouse chromosomes in nearby branches. Further, chromosome number 4 of D.Melanogaster, which did not cluster with any group and was an exception in both the 3rd order FCGR tree and the 10th order FCGR tree, groups with the other D.Melanogaster chromosomes. However, in all the trees constructed the chromosomes of Human and Chimpanzee were always intermingled. This phenomenon may be attributed to the high degree of similarity between the genome signatures of these closely related species.

3.5.2.2 FCGR order variation in the Human – Rhesus Monkey chromosome tree

The effect of varying the order of FCGR on the intermingling of chromosomes of two moderately related organisms, Human and Rhesus Monkey (Macaca Mulatta), was further examined separately. All the chromosomes including the mitochondrial genomes were taken for building a distance tree. It is observed that in the 3rd order FCGR tree (Figure 3.10) the chromosomes appear intermingled. This shows that the trinucleotide frequency profile similarity between the chromosomes is greater than the similarity between chromosomes of the same organism. However the 10th order FCGR tree (Figure 3.11) classifies the chromosomes organism-wise into separate clusters, with the exception that chromosome numbers 10, 16 and 19 of M.Mulatta cluster with human chromosomes.

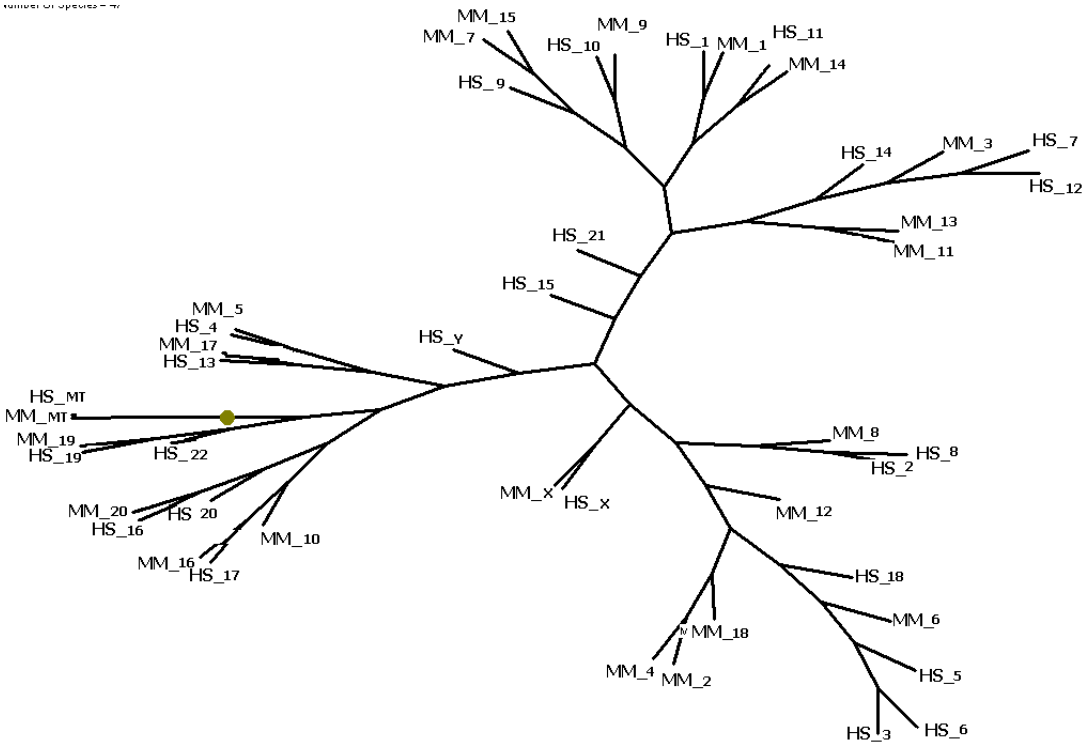


Figure 3.10 - 3rd order FCGR distance tree of Human (HS) - Rhesus Monkey (MM) chromosomes

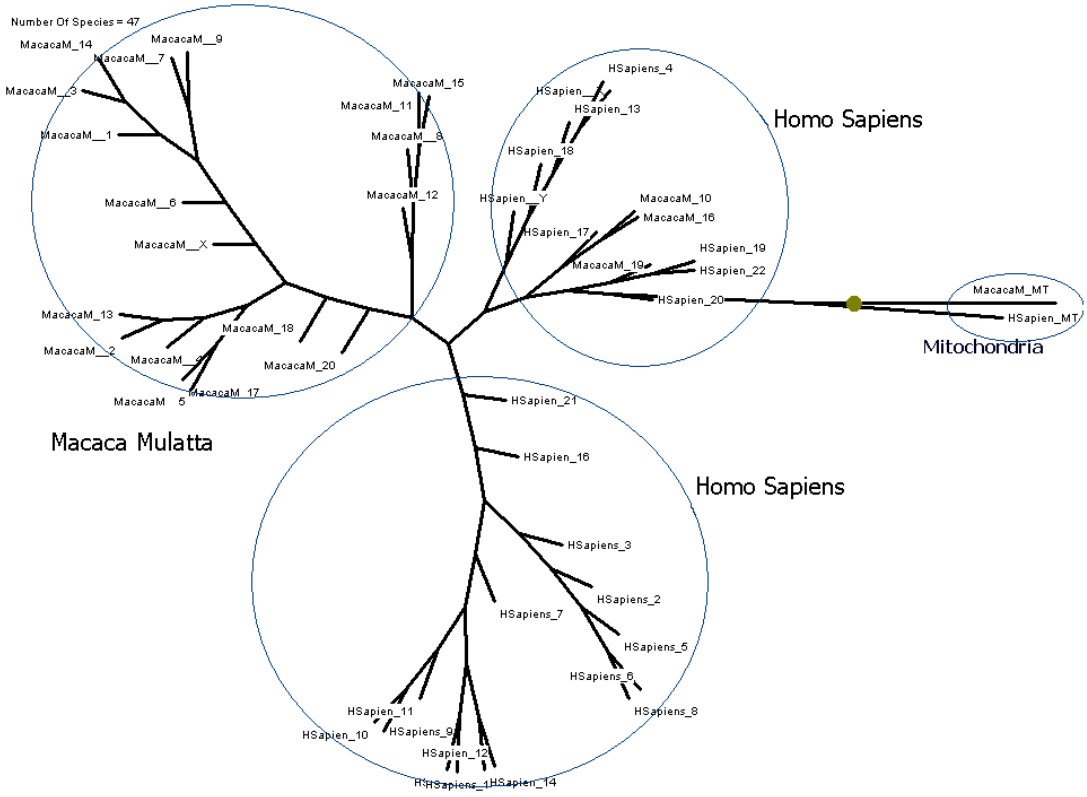


Figure 3.11 – 10th order FCGR distance tree of Human - Rhesus Monkey chromosomes

3.5.2.3 FCGR order variation in Human – Common Chimpanzee chromosome tree

The 3rd order FCGR based genome signature distance tree (Figure 3.12) between Human and Common Chimpanzee chromosomes shows non-random intermingling of chromosomes. It can be observed that almost all chromosome 'counterparts' in both the organisms have paired together. However this similarity between chromosomal counterparts is not observed for higher orders (e.g. 10-mer, Figure 3.13). The pairing of counterpart chromosomes as seen in the 3rd order FCGR tree is not observed here, except for chromosomes 7, 17 and 20. Such 'counterpart' chromosomes are present only in highly similar species pairs and that may be the

reason why the mingling of chromosomes in lower orders of 'not so highly similar' pair of species appears to be random.

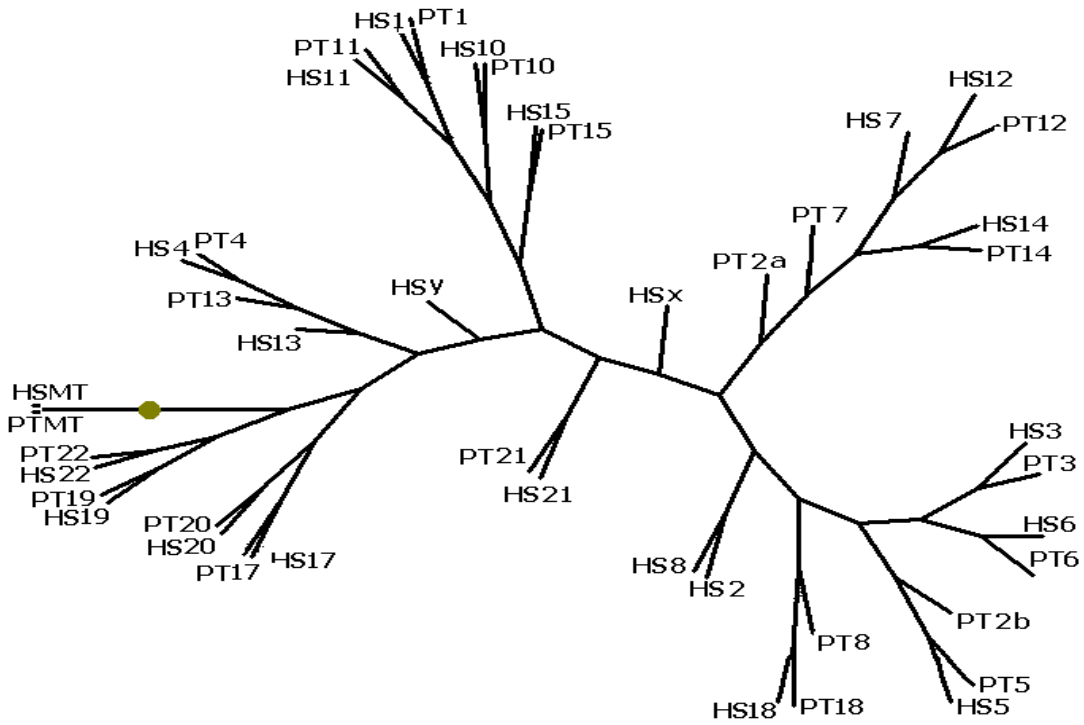


Figure 3.12 - Tree based on 3rd order FCGR of Human (HS) & Chimpanzee (PT) chromosomes

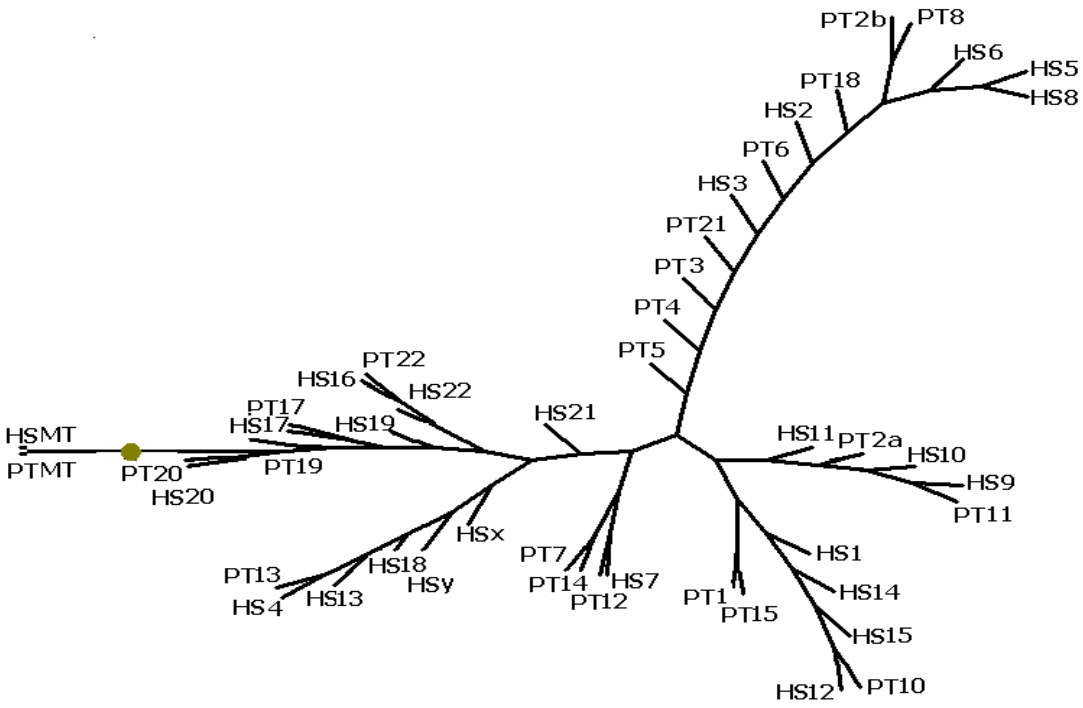


Figure 3.13 - Tree based on 10th order FCGR of Human (HS) & Chimpanzee (PT) chromosomes

3.5.2.4 FCGR order variation in Bacterial phylogenetic tree

The phenomenon of varying tree topology in accordance to varying FCGR orders was further observed in the case of bacterial phylogeny too. Sixty six bacterial genomes belonging to various classes were taken for building a phylogenetic tree. The results when using 3rd order FCGR based distance show clustering of organisms belonging to the same class in a single branch (Figure 3.14).

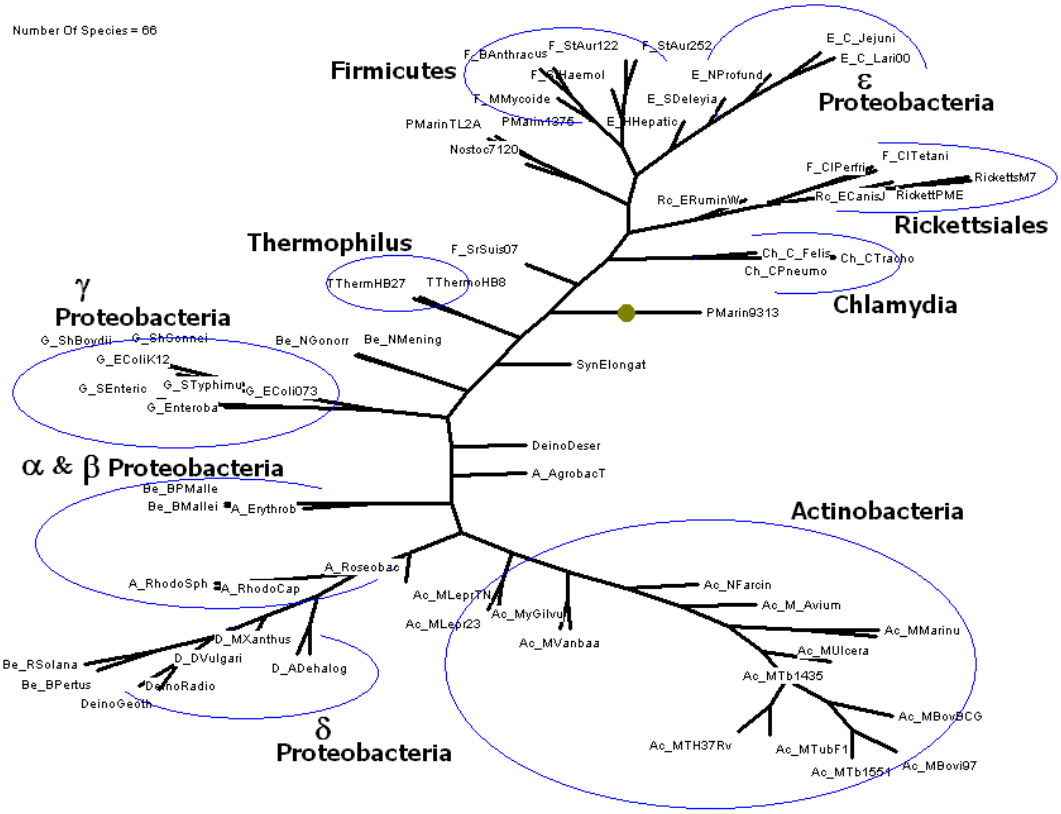


Figure 3.14 - Bacterial Tree based on 3rd order FCGR

When higher orders are used (e.g. 10th order) the overall accuracy of the tree decreases, in the sense that Rickettsiales, Epsilon-proteobacteria and Gamma-proteobacteria which formed separate clusters in the 3rd order tree, does not form such clusters in the 10th order FCGR tree (Figure. 3.15). On the other hand, higher orders give a better resolution in the case of very closely related species. The distance between actinobacteria is negligibly small in lower orders. In particular, the distance between two strains of 'Mycobacterium Leprae' is obtained 'zero' in the 3rd order FCGR distance chart while its non-zero in higher orders (> 6-mer).

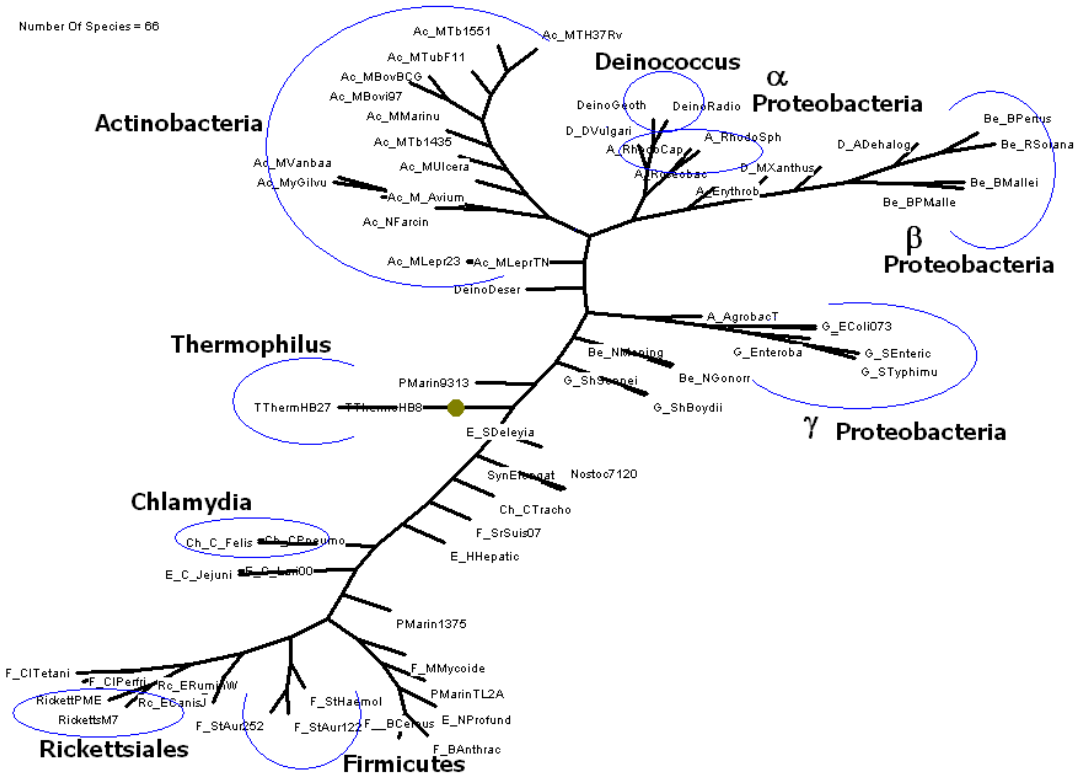


Figure 3.15 - Bacterial Tree based on 10th order FCGR

In summary we find that different orders of the FCGR provide different levels of resolution of the phylogenetic relationships. Higher order FCGRs carry more species-specific information but lose information on inter-species relatedness. Lower order FCGRs carry more information on inter-species relationships but can fail to resolve closely related species into separate clusters. Compensating for monomer bias improves the phylogenetic classification in some cases.

3.6 Conclusion

The phylogenetic signal in the genomic signature is explored in this chapter to utilize FCGR as a phylogenetic tree construction tool. The potential of this method is validated since trees obtained using this method is seen to exhibit many features of established phylogenies. However there are discrepancies also. The advantage of the

method is in the alignment free comparison property. The trees based on higher order FCGR show more resolved phylogenetic relationships between closely related species whereas the lower order FCGR produces trees with better overall inter-species relatedness. Another interesting result is the FCGR order variation on chromosomal trees of closely related organisms. The lower order FCGRs paired the 'counterpart' chromosomes, while the higher order FCGRs cluster the chromosomes organism-wise. These characteristics of FCGR based trees make it a significant complementary tool investigating into phylogenies from a novel view point.

3.7 References

1. Agosti D, Jacobs D and DeSalle R (1996) On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics* 12: 65–82
2. Almeida JS, Carrico JA, Maretzek A, Noble PA and Fletcher M (2001) Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics,* 17(5): 429--437
3. Bergsten J (2005) A review of long-branch attraction, *Cladistics* 21(2): 163--193
4. Bush EC and Lahn BT (2006) The evolution of word composition in metazoan promoter sequence, *PLoS Comput. Biol.* 2 (11): e150
5. Campbell A, Mrázek J, Karlin S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA, *Proc Natl Acad Sci USA* 96(16): 9184--9189
6. Chapus C, Dufraigne C, Edwards S, Giron A, Fertil B and Deschavanne P (2005) Exploration of phylogenetic data using a global sequence analysis method, *BMC Evol. Biol.* 5: 63

7. Conant GC and Lewis PO (2001) Effects of nucleotide composition bias on the success of the parsimony critierion in phylogenetic inference, *Molecular Biology and Evolution* 18: 1024--1033
8. Dayhoff MO, Eck RV and Park CM (1972) A model of evolutionary change in proteins, In: Dayhoff MO (ed.) *Atlas of Protein Sequence and Structure,* NBRF 5: 75--84,
9. Dehnert M, Plaumann R, Helm WE and Hutt MT (2005) Genome phylogeny based on short-range correlations in dna sequences, *J. Comput. Biol.* 12 (5): 545--553
10. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.* 16(10):1391—1399
11. Deschavanne P, Giron A, Vilain J, Vaury A, Fertil B (2000) Genomic signature is preserved in short DNA fragments, *IEEE International Symposium on Bioinformatics and Biomedical Engineering* (BIBE'00): 161--167
12. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* 33: e6
13. Edwards SV, Fertil B, Giron A and Deschavanne PJ (2002) A genomic schism in birds revealed by phylogenetic analysis of dna strings, *Syst. Biol.* 51: 599--613
14. Feng DF, Doolittle RF (1987) Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees, *J Mol Evol* 25: 351--360
15. Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positively misleading, *Syst. Zool.* **27**: 401--410

16. Gentles AJ and Karlin S (2001) Genome-scale compositional comparisons in eukaryotes, *Genome Res.* 11(4): 540--546

17. Gupta RS and Griffiths E (2002) Critical issues in Bacterial Phylogeny, *Theoretical Population Biology* 61(4): 423--434

18. Haibin WEI, Ji QI and Bailin HAO (2004) Prokaryote phylogeny based on ribosomal proteins andaminoacyl tRNA synthetases by using the compositional distance approach, *Science in China Ser. C Life Sciences* 47(4): 313--321

19. Holder MT and Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches, *Nature Reviews Genetics* 4: 275-284

20. Jacob F (1977) Evolution and tinkering, *Science* 196: 1161--1166

21. Jacob F and Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins, *J. Mol. Biol.* 3, 318–356

22. Jermiin LS, Ho SYW, Ababneh F, Robinson J and Larkum AWD (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be Underestimated, *Syst Biol* 53(4): 638-- 643

23. Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature, *Trends Genet*. 11: 283--290

24. Karlin S and Ladunga I (1994) Comparisons of eukaryotic genomic sequences, *Proc. Natl. Acad. Sci. U.S.A*. 91: 12832--12836

25. Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW (1992) Substitutional bias confounds inference of cyanelle origins from sequence data, *J Mol Evol* 34:153-162

26. Makalowski W (2003) Genomics. Not junk after all, *Science* 300:1246--1247

27. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P and Rigoutsos I (2006) Accurate phylogenetic classification of variable-length DNA fragments, *Nature Methods* 4: 63 -- 72

28. Miyamoto MM and Fitch WM (1995) Testing the covarion hypothesis of molecular evolution, *Mol. Biol. Evol.* 12:503–513

29. Monod J and Jacob F (1961) General conclusions- teleonomic mechanisms in cellular metabolism, growth, and differentiation, *Cold Spring Harb. Symp. Quant. Biol.* 26, 389–401

30. Naylor GJ and Gerstein M (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins, *J. Mol. Evol.* 51: 223--233

31. Pagani F, Raponi M and Baralle FE (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution, *Proc. Natl Acad. Sci. USA* 102: 6368--6372

32. Qi J, Wang B and Hao BI (2004) Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach, *J. Mol. Evol.* 58: 1--11

33. Perrière G, and Gouy M (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364-369.

34. Rosenberg MS and Kumar S (2003) Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference,.*Molecular Biology and Evolution* 20(4): 610--621

35. Sandberg R, Winberg G, Branden C, Kaske A, Ernberg I, Coster J, (2001) Capturing whole-genome characteristics in short sequences using a naive bayesian classifier, *Genome Research* 11, 1404– 1409

36. Sandersson MJ and Driskell AC (2003) The challenge of constructing large phylogenetic trees, *Trends in Plant Science* 8: 374--379

37. Sankoff D, Morel C, Cedergren RJ (1973) Evolution of 5S RNA and the non-randomness of base replacement, *Nature New Biology* 245:232--234

38. Sarfaty CK, Mi Oh J, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity, *Science* 315: 525--528

39. Simmons MP (2004) Independence of alignment and tree search. *Mol Phylogenet Evol* 31(3): 874--879

40. Strait DS, Grine FE (2004) Inferring hominoid and early hominid phylogeny using craniodental characters: the role of fossil taxa. *J Hum Evol* 47(6):399—452

41. Wang Y, Hill K, Singh S and Kari L (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation, *Gene* 346: 173--185

42. Wiens JJ (2001) Character analysis in morphological phylogenetics: problems and solutions. *Syst Biol* 50(5): 689--99.

43. Wray GA (2007) The evolutionary significance of cis-regulatory mutations, *Nat. Rev. Genet* 8: 206—216

Chapter 4

Evolutional ancestry of mitochondria computed using FCGR

4.1 Introduction

Having established in the last chapter that the genome signature represented by FCGR has potential for complementing traditional methods for deducing phylogenetic relationships, here we focus on the exploration of an important problem of evolutionary biology namely, the evolutionary origin of the eukaryotic organelles, mitochondria and chloroplasts using FCGR.

The evolutionary origin of the eukaryotic cell is an intensely debated topic. Intimately tied up with this is the origin of the eukaryotic organelles, mitochondria and chloroplasts. The established view on the origin of these organelles is that they evolved from endosymbiotic bacteria which were ingested by an ancestral eukaryote - an alpha proteobacterium evolving into mitochondria and a cyanobacterium evolving into chloroplasts. Identification of the bacterial ancestor of mitochondria as an alpha proteobacterium rests on the remarkable similarity of the amino acid sequences of

mitochondrial proteins to the corresponding α-proteobacterial proteins. Here we explore whether the same relationships can be deduced using the whole genome signature as the phylogenetic signal. We find that the whole genome signature tells a different story from that evidenced by the amino acid sequence alignments of specific proteins. Intrigued by this discrepancy, we explore whether alignments of the nucleotide sequences that code for the proteins support the evidence from the amino acid sequence alignments. We find that the evidence from the nucleotide sequence alignments support the evidence from whole genome signatures rather than that from the amino acid sequence alignments. Based on this we propose a plausible alternate hypothesis for the origin of mitochondria which we support further with arguments like parsimony, timing of geological events, selectional advantages and structural and functional similarities.

4.2 Current view of the origin of Mitochondria and Chloroplasts

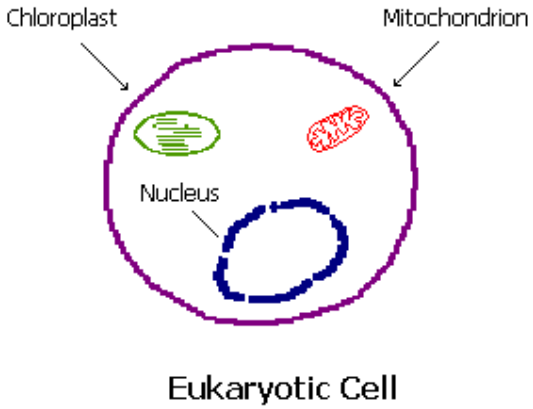


Figure 4.1

Mitochondria are rod-shaped organelles essential to all respiring eukaryotes. They provide the energy a cell needs to move, divide, produce secretory products,

contract - in short, they are the power centers of the cell. They are the sites where organic compounds are oxidized to carbon dioxide and water with a high yield of chemical energy in the form of ATP. This is such an effective process that it often is regarded as the prerequisite for multicellular life. Chloroplasts are organelles found in plant cells and eukaryotic algae that conduct photosynthesis. Chloroplasts absorb sunlight and use it in conjunction with water and carbon dioxide gas to produce food for the plant. Chloroplasts capture light energy from the sun to produce the free energy stored in ATP and NADPH through a process called photosynthesis. Both organelles are surrounded by a double celled composite membrane with an intermembrane space, have their own DNA and are involved in energy metabolism. Both mitochondria and chloroplasts have reticulations, or many infoldings, filling their inner spaces. Mitochondria replicate independently from the nucleus, arising only from a preexisting mitochondria. Unlike mitochondria, however, chloroplasts replicate at the same time as the host cell, however replication of chloroplast DNA and cell DNA are not synchronized.

The Serial Endosymbiosis Theory (SET) was first articulated by the Russian botanist Konstantin Mereschkowski (1905) and was later popularized by Lynn Margulis (1981) in her book Symbiosis in Cell Evolution is currently the most accepted theory explaining the origin of the eukaryotic organelles, mitochondria and chloroplasts. According to SET, a set of anaerobic, heterotrophic cells known as proto- eukaryotes ingested and developed a mutually beneficial relationship symbiosis with an aerobic bacterium, which evolved into mitochondria, and later with a photosynthetic bacterium, which evolved into chloroplasts.

The currently accepted view (Yang et al., 1985; Andersson et al., 1998; Gray and Doolittle, 1982; Gray et al., 2001) is that mitochondria and chloroplasts originated from two separate endosymbiont events – the engulfment of an aerobic respirer belonging to the α-subdivision of proteobacteria leading to formation of mitochondria, and the subsequent engulfment of a photosynthetic cyanobacterium leading to formation of chloroplasts. The alpha proteobacterial ancestry of mitochondria is based on the similarity of the amino acid sequences of mitochondrial proteins to corresponding α-proteobacterial proteins. Andersson et al. (1998) did phylogenetic analysis of concatenated aminoacid sequences of ribosomal proteins and concatenated sequences of NADH proteins of mitochondria, alpha-proteobacteria and other bacteria. Both the trees show close evolutionary relationship between mitochondria and α-proteobacteria and in particular with Rickettsia Prowazekii.

4.3 Results

4.3.1 Genome signature relationships between cyanobacteria, α-proteobacteria, and the eukaryotic organelles

As mentioned in the previous chapter, more and more evidence is turning up that selectional constraints work not only at the amino acid level but also at the underlying nucleotide level itself. The so-called "silent mutations" producing synonymous codons are proving to be not really *silent*. Further, the non-coding regions face evolutionary constraints because of their involvement in the regulation of gene expression. This points out the necessity to support the amino-acid sequence-based evidence with nucleotide sequence-based evidence at the level of the whole genome and at the level of individual genes.

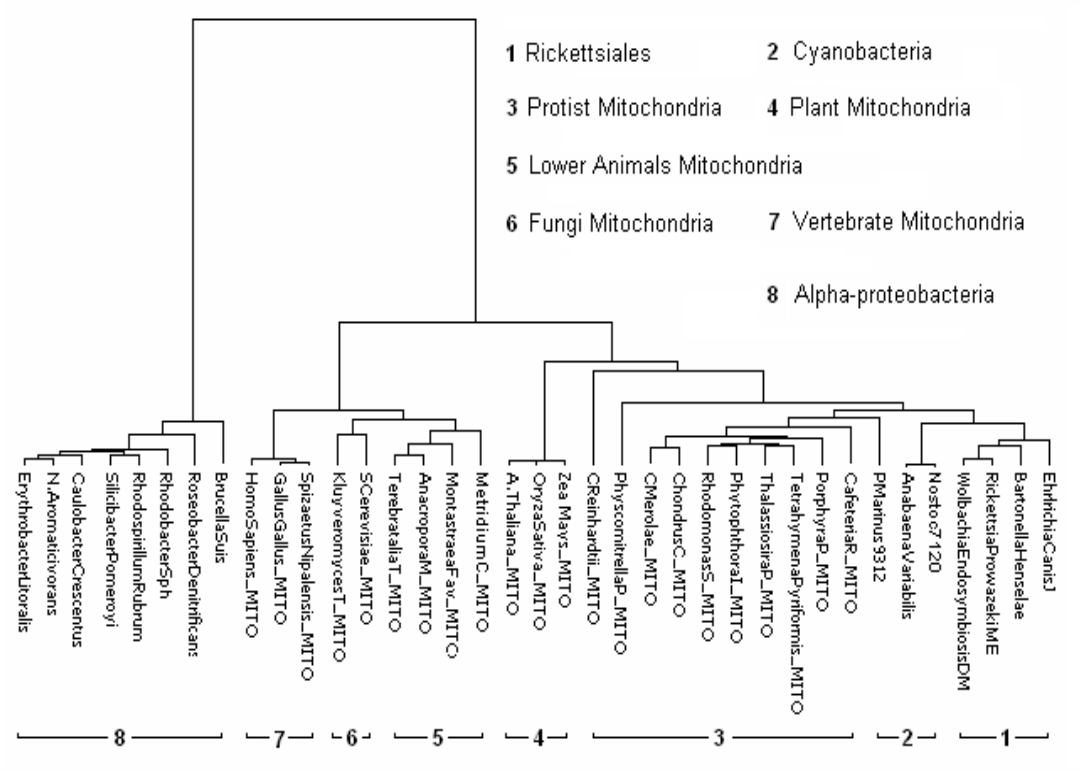


Figure 4.2 : Tree based on 3rd order FCGR corrected for base composition showing positions of cyanobacteria, α-proteobacteria and mitochondria

Figure 4.2 shows the genome signature-based distance tree of the whole genomes of cyanobacteria, α-proteobacteria and mitochondria from different organisms. The 3rd order FCGR corrected for base composition is taken as representing the genome signature and the distance between the genome signatures is evaluated as a statistical distance based on weighted Pearson correlation. The method is described in the previous chapter. The mitochondria cluster into clearly defined groups of plants (Group 4), protists (Group3), fungi (Group 5), simple animals (Group 6) and vertebrates (Group 7) showing the presence of a strong phylogenetic signal in the genome signatures. Cyanobacteria (Group 2) and Rickettsiales (Group 1) cluster together in groups close to the protist mitochondria (Group 3). All mitochondria are

seen to be distant from the cluster of alpha proteobacteria excluding the Rickettsiales (Group8).

It has already been established that the Rickettsiales cannot be ancestral to mitochondria and that mitochondria and Rickettsiales have a common ancestor. (Andersson et al., 2003, Emelyanov, 2003) The closeness of mitochondria and the Rickettsiales to cyanobacteria raises the question: Could a Cyanobacterium have been the ancestor of mitochondria?

4.3.2 Genome signature distances between Mitochondria, Chloroplasts and Nuclear Genomes

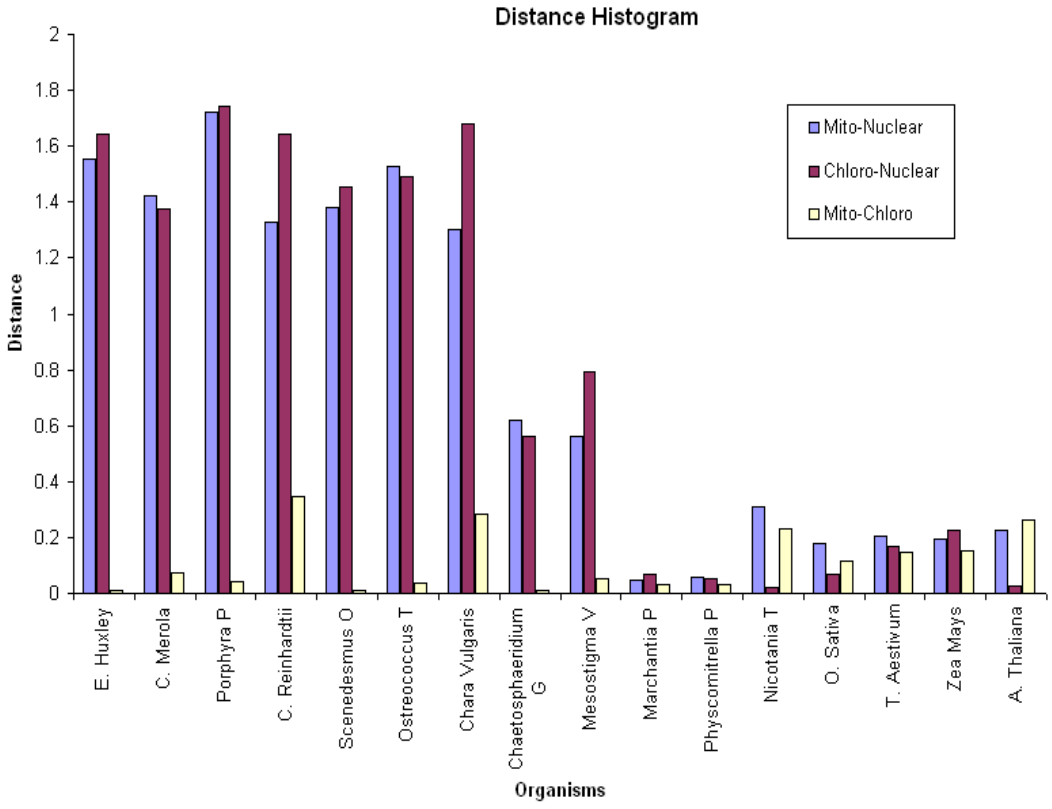


Figure 4.3: The CGR distance histogram showing mito-nuclear and chloro-nuclear distances decreasing from lower to higher organisms (from left to right).

Figure 4.3 is a bar graph that shows the genome signature-based "distance" between chloroplasts and mitochondria as well as the distance of the organellar genomes from the respective nuclear genomes of twelve lower and higher photosynthetic eukaryotes. We find that, in the lower eukaryotes the organellar signatures are very distinct from the nuclear signatures while the two organellar signatures are strikingly close to each other. **The closeness of the two organelles is indicative of the organelles having originated from a single endosymbiont.** The large dissimilarity of signature between the organelles and the nuclear genome in the lower organisms indicate that the **ancestor of the organelles had distinctly different signature from that of the host organism that engulfed it**. Later because of massive gene transfer between the organelles and the nucleus, the nuclear signature could have come closer to the organelles signature as indicated by our finding that in higher plants the nuclear signature is close to the organellar signature. Cyanobacteria are the only bacteria that have the capability of both photosynthesis and aerobic respiration. Therefore Cyanobacteria are the only bacteria that could evolve into both chloroplast and mitochondria. Thus if the two organelles had a common ancestor it could only be a cyanobacterium

4.3.3 Comparisons of nucleotide sequences of genes based on multiple sequence alignment

To see whether the nucleotide-based results are consistent with the amino acid sequence based evidence we performed phylogenetic analyses with the same genes

that are used in one of the classical works that showed the closeness of mitochondrial proteins to α-proteobacterial proteins (Andersson et al., 1998). Multiple sequence alignment was performed on the nucleotide sequences using ClustalW (Thompson et al., 1994). The alignment thus obtained was given as input to the logdet program of PHYLIP package to compute the logdet distance matrix. Logdet distances were computed since sequences were taken from organisms with widely varying mutation rates. The tree was drawn from the distance matrix using the Kitsch program of PHYLIP (Felsentin, 1989). The unrooted trees are visualized using the Phylodraw (Choi et al., 2000) package. The rooted trees are visualized using the njplot program (Perrière, 1996).

The following figures show phylogenetic trees obtained using nucleotide sequence of concatenated ribosomal genes, concatenated respiratory genes and Iron-Sulphur cluster assembly gene of mitochondria, chloroplasts, α-proteobacteria, and cyanobacteria. It is seen that, the cluster containing cyanobacteria and chloroplasts is significantly closer to the mitochondrial cluster than to the rest of the α-proteobacterial cluster. The only alpha proteobacteria that clustered with the mitochondria were of the Rickettsiales family.

Figure 4.4 show the unrooted tree obtained using the concatenated genes of ribosomal proteins. The mitochondria and the chloroplast genes can be seen to cluster together with the cyanobacteria and the Rickettsiales group. The alpha-proteobacteria are seen to be distant from both. Figure 4.5 show the tree obtained using the same genes, but this time using an archaea, a Thermoplasma, as an out group. Here also the alpha-proteobacterial group can be seen to cluster distantly from the organelles + cyanobacteria + Rickettsiales group .

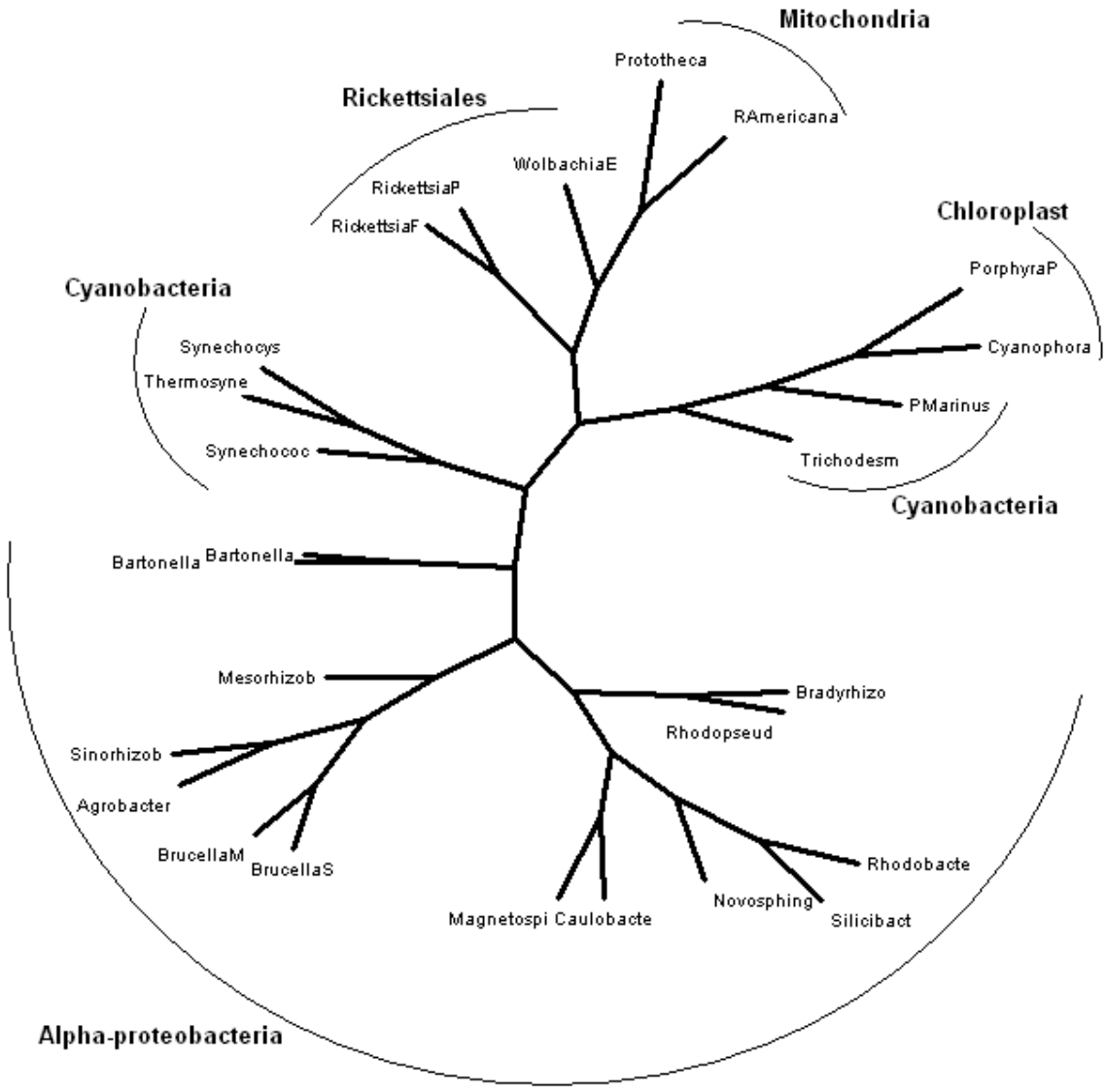


Figure 4.4: Unrooted tree of concatenated ribosomal genes S2, S3, S7, S10, S11, S12, S13, S14, S19, L5, L6 and L16

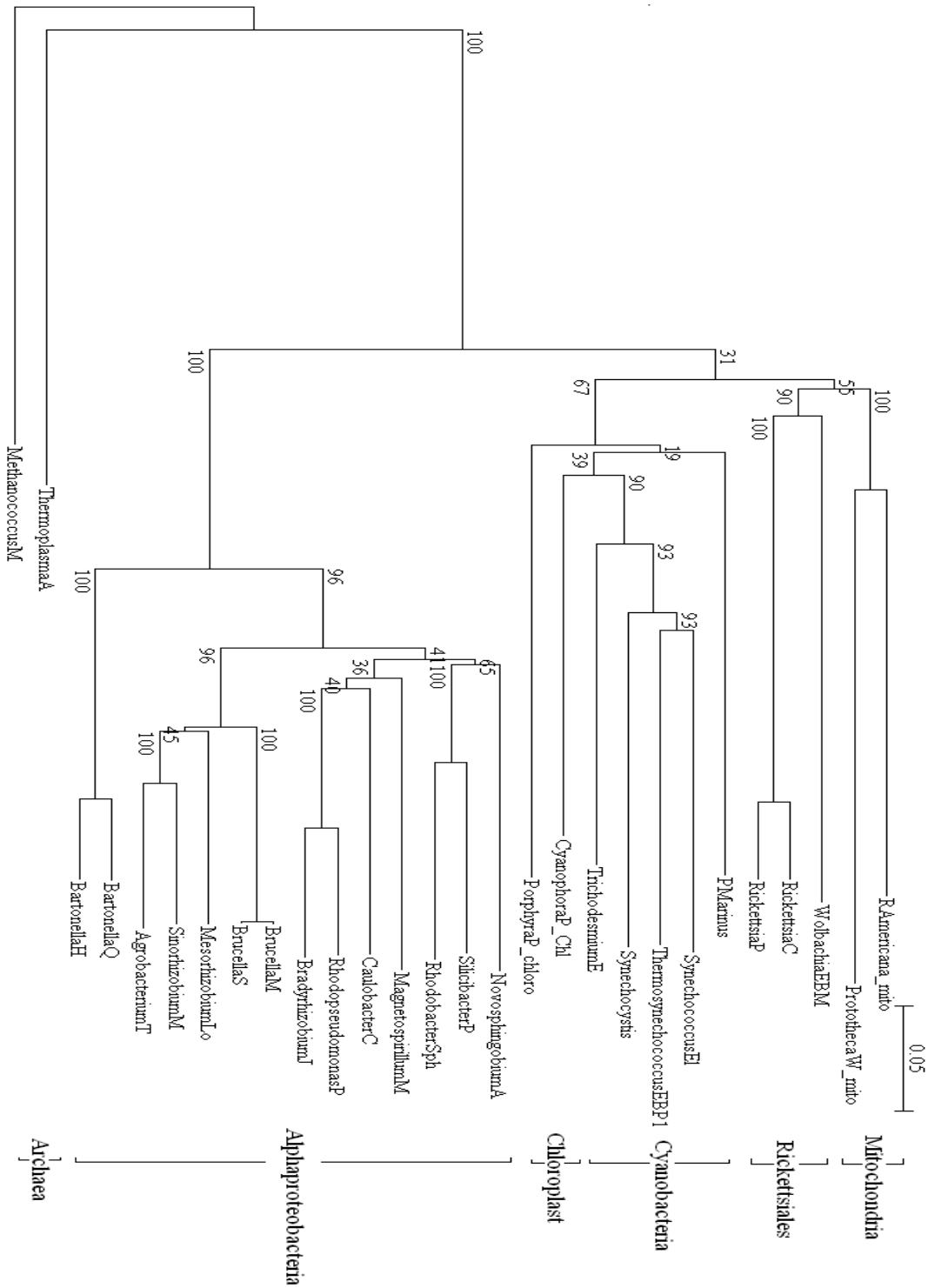


Figure 4.5 Rooted tree of concatenated ribosomal genes S2, S3, S7, S10, S11, S12, S13, S14, S19, L5, L6 and L16

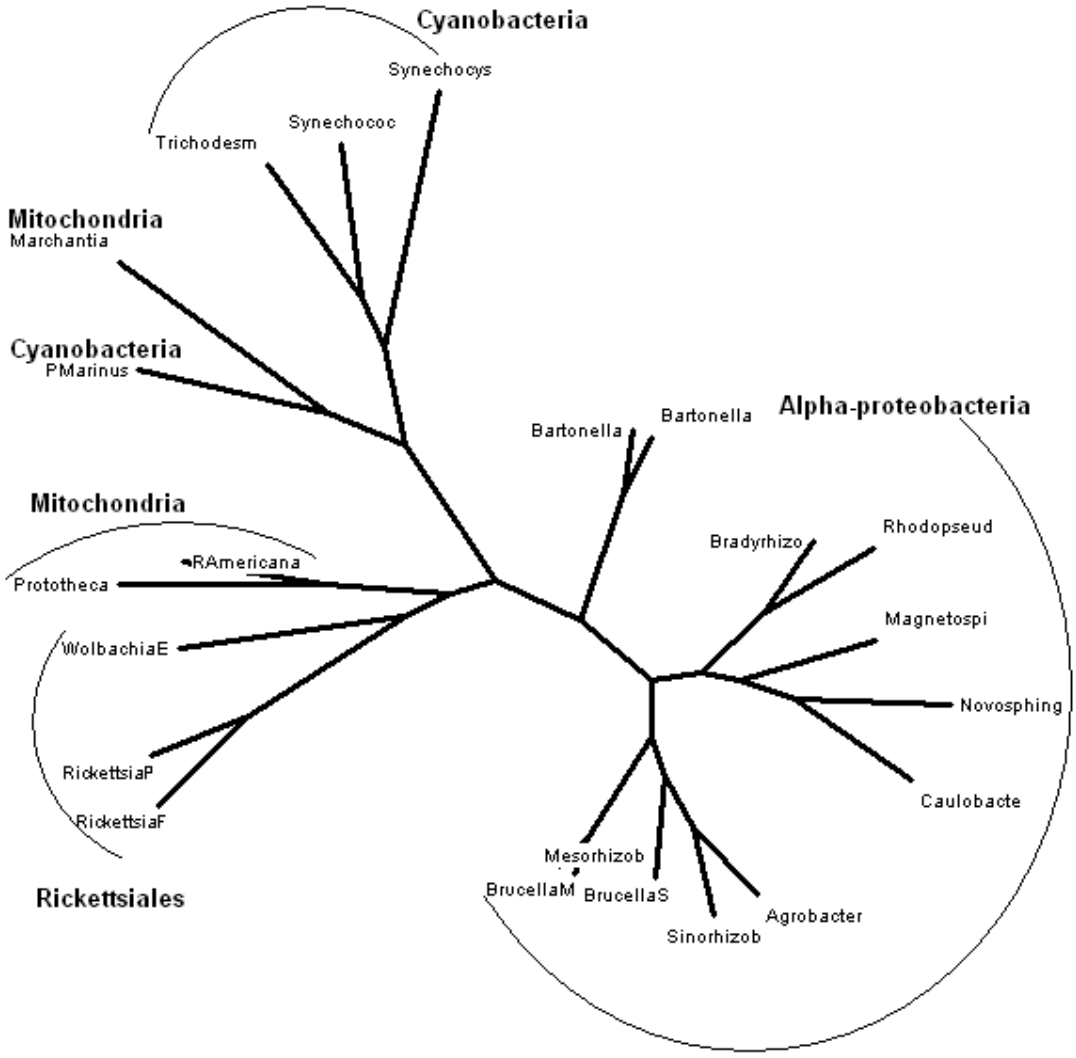


Figure 4.6 : Unrooted tree with logdet distances of concatenated NADH genes nuoA, J, K, L, M and N.

Figure 4.6 show the unrooted tree obtained using the concatenated NADH genes. The mitochondrial genes and can be seen to cluster together with the cyanobacteria and the Rickettsiales group. The alpha-proteobacteria is seen to be distant from both. Figure 4.7 show the tree obtained using the same genes, but this time using a Mycobacterium, as an out group. The alpha-proteobacterial group can be seen to cluster distantly from the other group. High boot strap values are seen to support the phylogeny obtained.

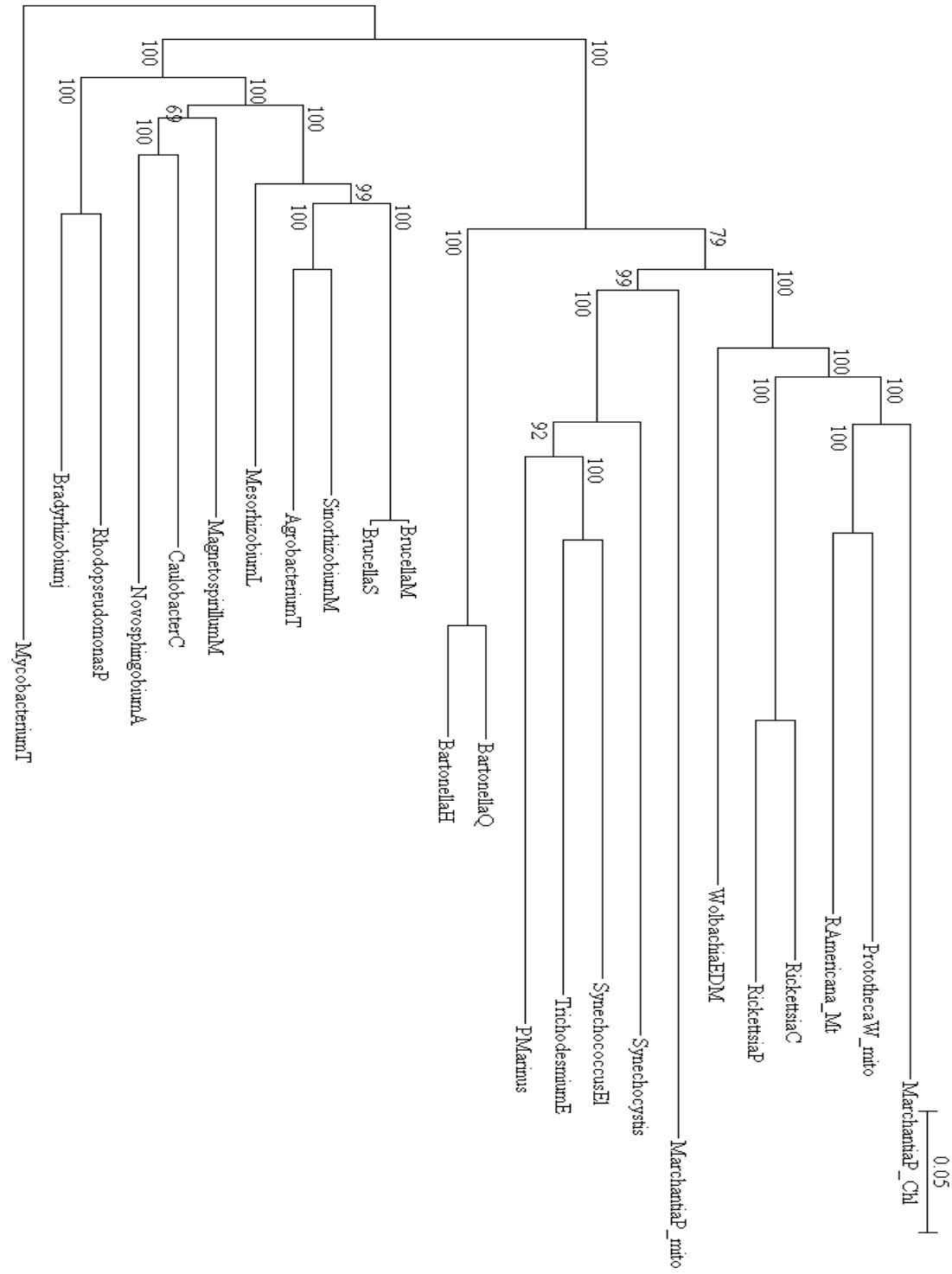


Figure 4.7 - Rooted tree with Mycobacterium outgroup of NADH genes nuoA, J, K, L, M, N.

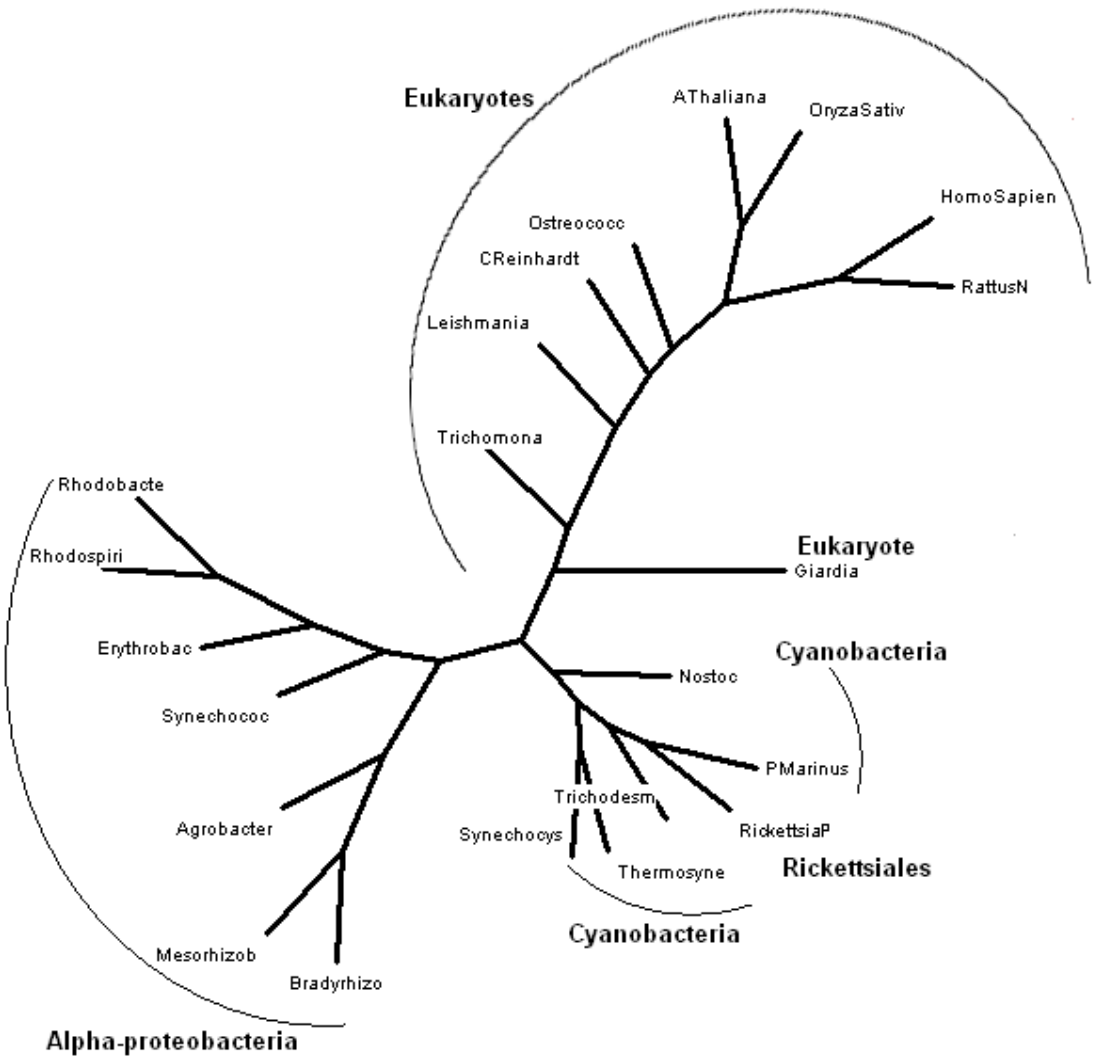


Figure 4.8 : Unrooted tree of iron-sulphur cluster assembly genes, IscS

The assembly of iron–sulfur clusters is considered one of the most essential functions of the mitochondrial compartments. Iron-Sulphur Cluster Assembly gene IscS is found in all eukaryotic lineages including the a-mitochondrial ones containing hydrogenosomes and mitosomes. Figure 4.8 shows a phylogenetic tree based on the nucleotide sequences of IscS genes of eukaryotes, α-proteobacteria, and cyanobacteria. We find that the eukaryotic genes cluster together and that this cluster is closest to a cluster containing both cyanobacteria and Rickettsiales while the rest of the α-proteobacteria are considerably farther apart.

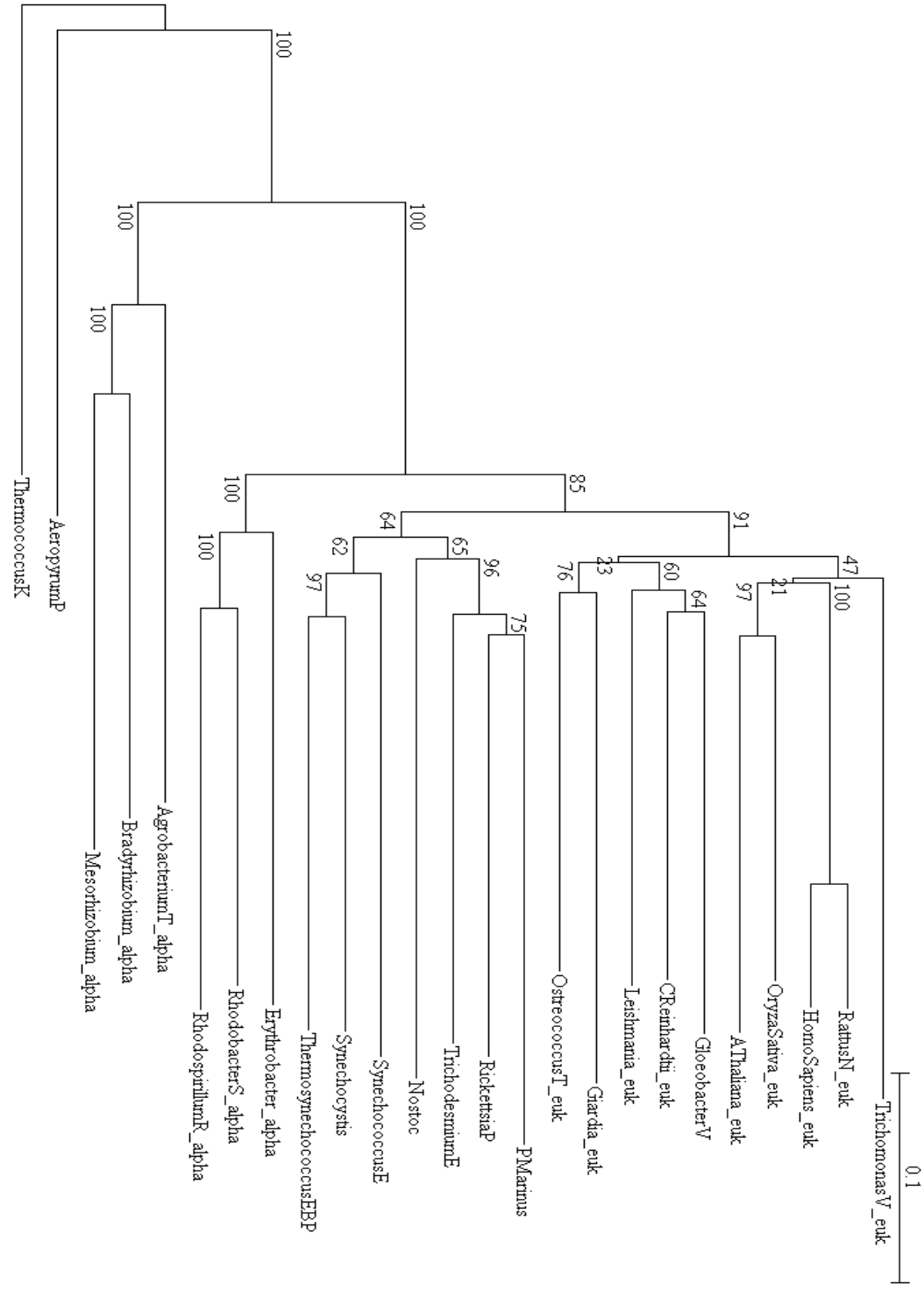


Figure 4.9 Rooted IscS tree with an archaea outgroup

The nuclear encoded mitochondrial aminoacyl-trna synthetases (aaRSs) occupy a position in the tree that is not close to any of the currently sequenced /alpha-proteobacterial genomes (Brindefalk et al., 2006). Corresponding to the gene of each aminoacyl-trna synthetase a phylogenetic tree was formed. The phylogenetic tree based on arginyl tRNA synthetase gene is seen in figure 4.10. A rooted tree with an archaea as an outgroup is drawn. The mitochondrial tRNA synthetases can be seen to cluster with cyanobacteria while the alpha-proteobacteria form another cluster. However, all the tRNA synthetase trees does not conform to this topology. Figure 4.11 show the phylogenetic tree based on methionyl tRNA synthetase gene. A rooted tree is drawn with an archaea as an outgroup. The mitochondrial tRNA synthetase can be seen equally close to cyanobacterial and alpha-proteobacterial genes.

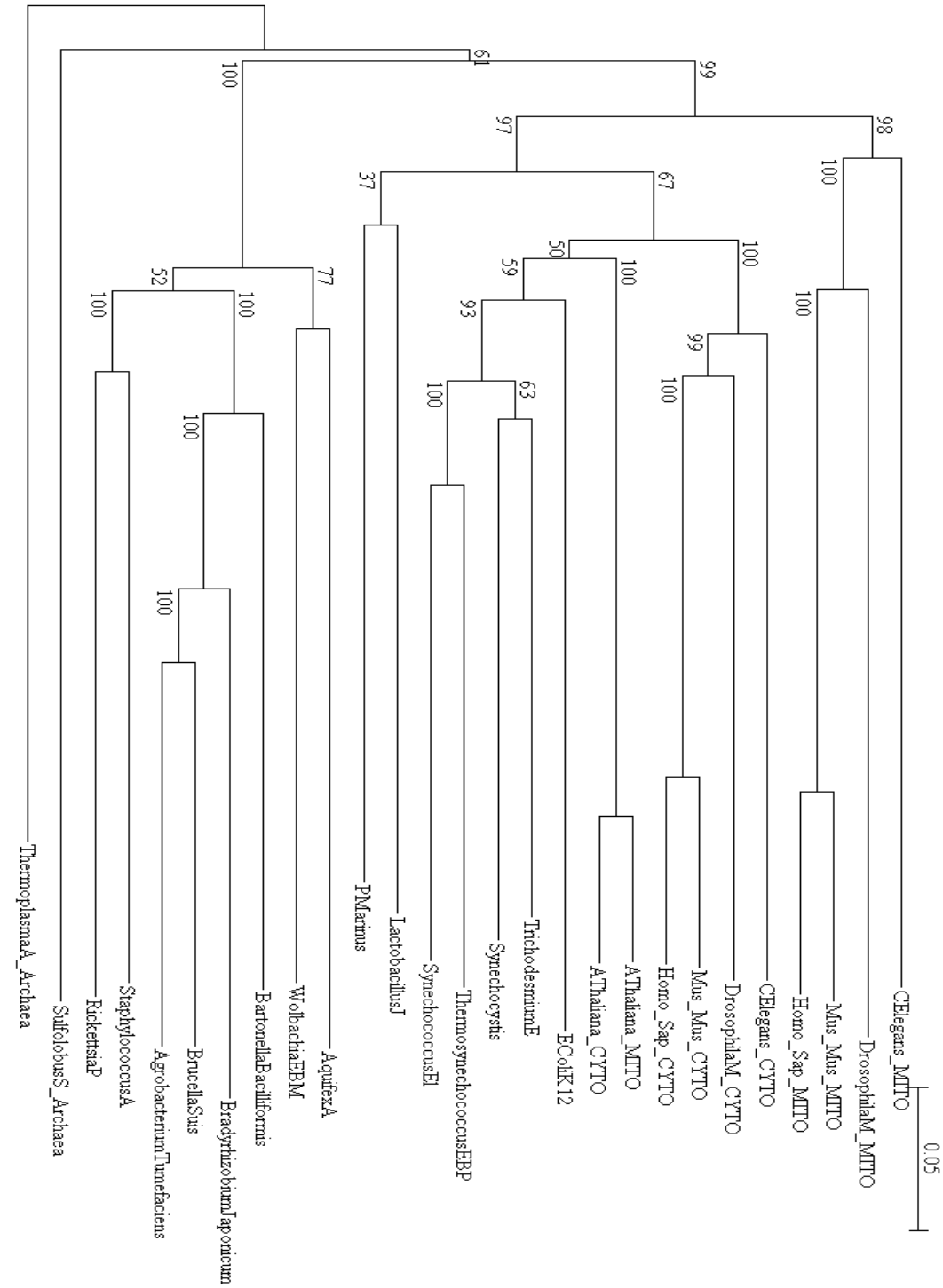


Figure 4.10 - Arginyl tRNA synthetase genes tree with outgroup archaea

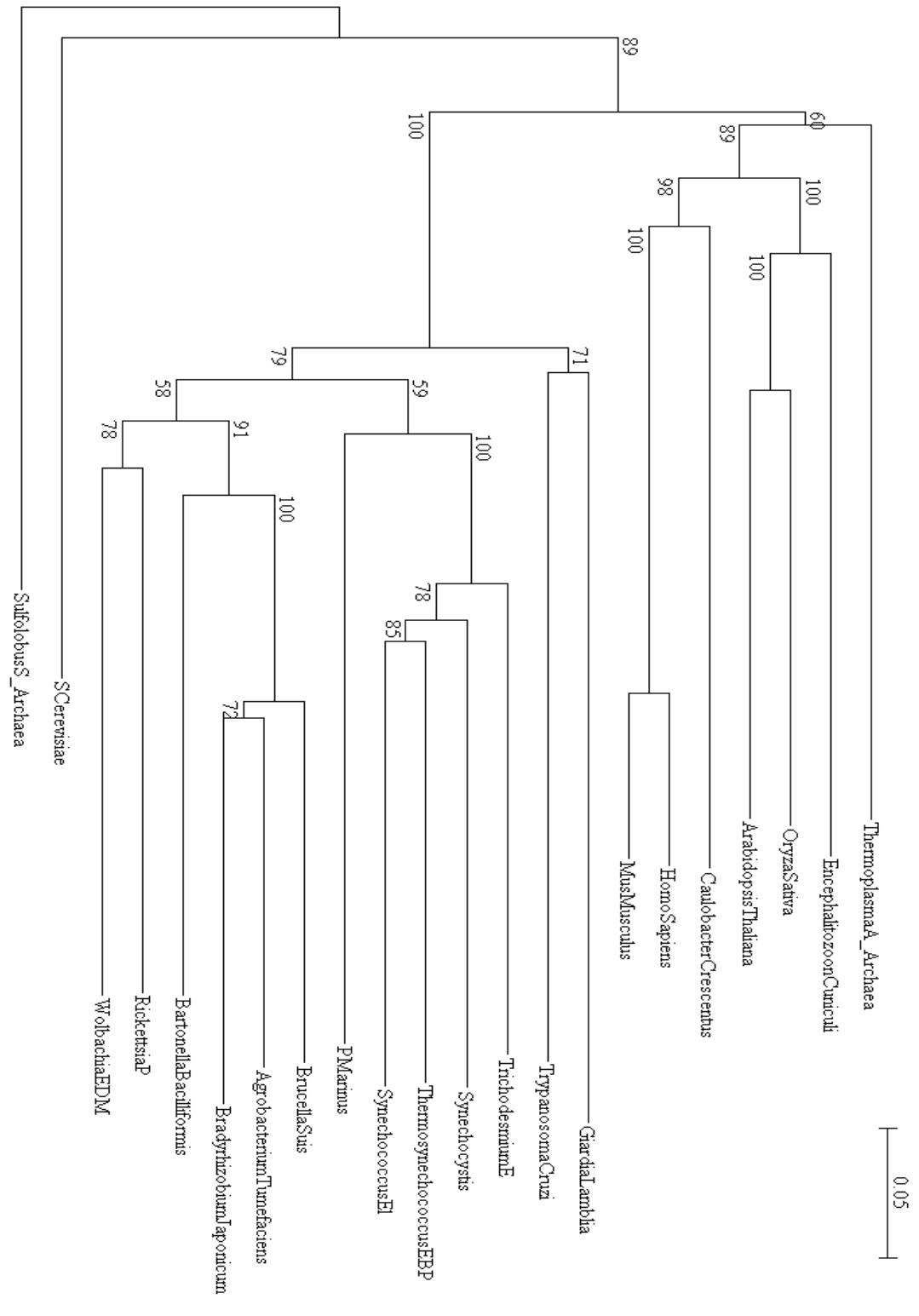


Figure 4.11 Methionyl tRNA synthetase genes tree with archaea outgroup

4.4 Discussion

4.4.1 An alternate hypothesis

Based on the results of our comparisons of nucleotide sequences at whole genome level and individual gene level we put forth an alternate hypothesis, namely that both mitochondria and chloroplasts originated through the single endosymbiont event of a host cell engulfing a cyanobacterium. We propose that a cyanobacterium could have been engulfed by a proto-eukaryote in a rare-chance event very early in evolution (~ 2.4 Ga ago). Possessing the unique capability of both aerobic respiration and oxygenic photosynthesis the endosymbiotic cyanobacterium conferred such a selective advantage to the host that this "lucky host" out-competed all other proto-eukaryotes and thus emerged as the ancestor of eukaryotes. The ingested cyanobacterium functioned initially as a "chloromitochondrion" and later separated into two separate organelles the chloroplast and the mitochondrion performing photosynthesis and respiration separately.. Later some of the hosts lost chloroplasts and became non photosynthetic eukaryotes Evolving in an atmosphere of progressively increasing oxygen the early primitive cyanobacterium-derived mitochondria developed more and more efficient systems for oxygen respiration and evolved into the present day mitochondria.

In the following sections we put forth several arguments that support this hypothesis

4.4.2 The Timing of Events

The fossil record of cyanobacteria has been argued to extend to 3.5 Ga backwards but reliable biomarker evidence at 2.7 - 3.2 Ga ago is usually considered to

be the earliest undisputable micro-fossil record of cyanobacteria (Barghoorn and Schopf, 1965,1966; Schopf, 1970).

The date of emergence of eukaryotes is still debated and estimates vary widely, from as early as 3.5 billion years ago to no more than 0.9 – 1.3 billion years ago. The estimates (< 1.3 Ga) of the origin of eukaryotes rest mostly on the lack of undisputed fossil evidence, which is not a strong argument by itself (de Duve, 2007), since it is possible that the earliest eukaryotes did not leave fossils at all, or that they are yet to be found. On the other hand **there are a number of arguments that strongly favor an early origin of eukaryotes.** There are a large number of apparently ancient eukaryotic innovations that do not have a prokaryotic counterpart (de Duve, 2007). The characteristics that are unique to eukaryotes seem to show that the proto-eukaryotes were essentially anaerobic organisms that had developed long before the atmosphere became oxygenated (~ 2.4 Ga ago). The earliest reliable biomarker evidence of eukaryotes is reported at 2.7 Ga ago (Brocks et al., 1999) and the earliest fossils are reported to be 2.1 Ga old (Han and Runnegar, 1992). A genomic timescale fixes the origin of eukaryotes around 2.6 Ga (Hedges et al., 2001). **All these findings have strengthened the view that modern eukaryotic and prokaryotic cells had long followed separate evolutionary trajectories** (Kurland et al., 2006)**.** A number of studies have suggested that the divergence of the archaeal, bacterial, and eukaryotic lineages is ancient (Forterre, 2001, Sicheritz-Ponten and Andersson, 2001) and, what seems most important in the present context, **that the divergence of the eukaryotic lineage predates the divergence of α-proteobacteria** (Canback et al., 2002).

Alpha-proteobacteria has been shown to be a rather late diverging group within bacteria by evidence provided by signature sequences in a number of proteins (Gupta RS, 1998). The α-proteobacterial endosymbiosis is considered to have taken place around 1.8 Ga ago (Farquhar et al., 2007), 600 million years after the atmosphere became oxygen-rich. **This late acquisition of mitochondria raises questions on how the anaerobic proto-eukaryotes survived until they met their α-proteobacterial rescuers and became aero-tolerant and even aerobic** (de Duve, 2007)**. It seems more plausible that eukaryotic aero-tolerance was accomplished much earlier than 1.8 Ga**. If alpha proteobacteria were the ancestors of mitochondria, an early origin of eukaryotes implies that eukaryotes stayed a-mitochondriate for a long time after their origin. The question then is how they survived the toxic onslaught of oxygen during this period. **Thus an early origin of eukaryotes is inconsistent with the alpha proteobacterial origin of mitochondria** (Cavalier-Smith, 2006). Giardia Lamblia which is considered to have once possessed mitochondria, emerged around 2.2 Ga according to the genomic timescale for eukaryotic evolution calculated by Hedges et al (2001). The secondarily a-mitichondriate nature of Giardia becomes consistent with their estimated origin around ~2.2Ga once the constraint of alpha proteobacterial origin of mitochondria after 1.6Ga is removed.

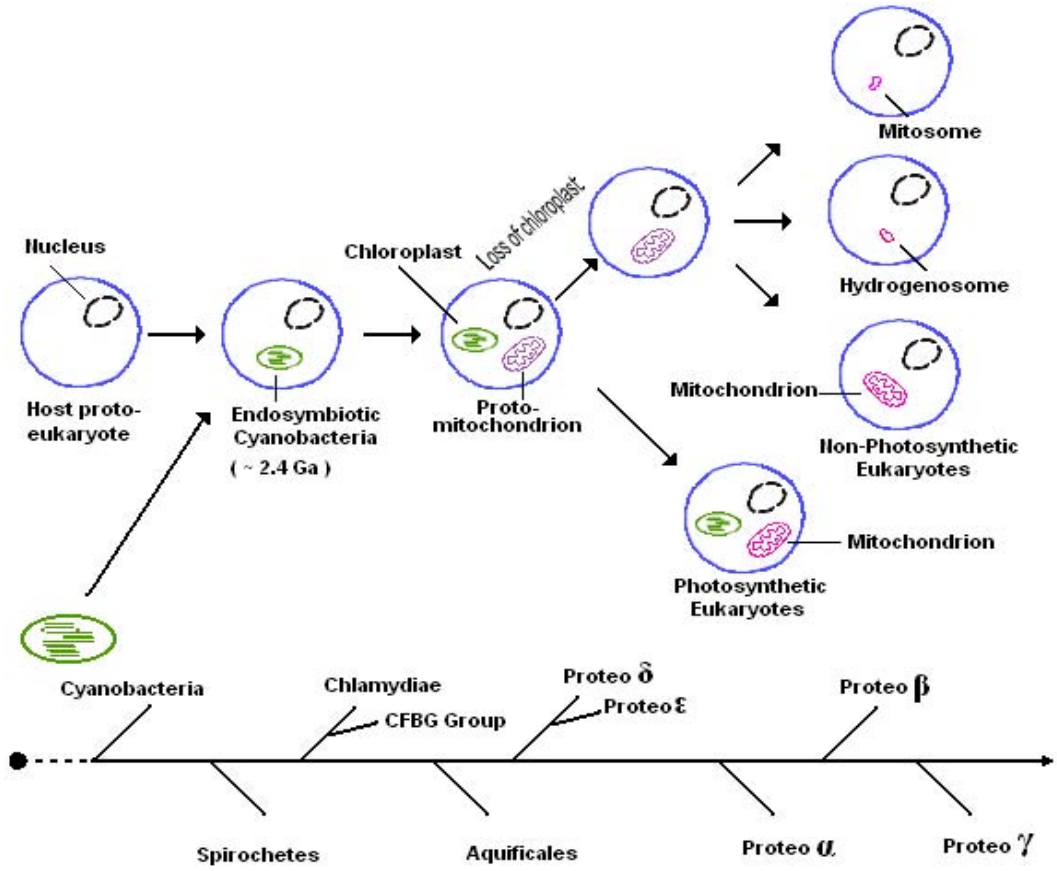


Figure 4.12 Proposed sequences of events leading to eukaryotic organelles. (Parallel evolution of (free-living) bacteria also shown schematically).

Cyanobacteria possessing aerobic respiratory systems were already around from the early days of atmospheric oxygenation. **Endosymbiosis of a cyanobacterium by an anaerobic proto-eukaryote before 2.4 Ga ago and subsequent evolution of the respiratory system under increasing oxygen is more consistent with an ancient origin of eukaryotes**.

It is generally believed that mitochondria were acquired before chloroplasts because most extant eukaryotes possess some form of mitochondria while only a subset of them possess chloroplasts. However, recent evidence suggests that a number

of today's non-photosynthetic eukaryotic microbes have evolved from a photosynthetic ancestor by loss of pigments and/or chloroplasts. In the light of all newly accumulated knowledge, the primary endosymbiont event leading to chloroplasts occurred much earlier in eukaryotic evolution than currently envisaged (Andersson and Roger, 2002, Purton, 2002). Phylogenetic data on the origin of chloroplast Hsp90 suggest that the common ancestor of animals and plants once harbored chloroplasts (Emelyanov, 2002), and the recent discovery of vestiges of photosynthetic structures ("thylakosomes") in chemoheterotrophic protists such as *Psalteriomonas lanterna* and some representatives of the parasitic genus *Apicomplexa* (Hackstein et al., 1997) would rather conform to this type of reasoning. Even some peroxisomal enzymes from animals have been shown to have originated from cyanobacteria (Gabaldon et al., 2006, Tabak et al., 2006). **Therefore, it is well possible that the ancestors of all the current non-photosynthetic organisms had once harbored chloroplasts in the form of cyanobacteria.**

Stanier (1970) proposed that chloroplast endosymbiosis took place first because oxygenic photosynthesis must have preceded aerobic respiration. Cavalier Smith (1987) proposed a simultaneous origin of chloroplasts and mitochondria by (nearly) simultaneous endosymbiont events around 0.9 Ga ago, that should have led to uptake of both the progenitors (of chloroplasts and mitochondria) by a proto-eukaryotic host possessing phagocytic capability. He argued that, as soon as eukaryotes acquired the phagocytic machinery, all kinds of symbionts could be taken up and **it is unlikely that photosynthetic ones would be taken up appreciably earlier or later than respiratory ones**. In a more recent paper (Cavalier Smith, 2006) he reported that TOM70, a protein of crucial importance in the import of

proteins into mitochondria, shows a clear cyanobacterial origin thereby supporting his hypothesis that the primordial host also harbored symbiotic cyanobacteria along with the proto-mitochondrial endosymbiont.

4.4.3 Parsimony and Selective Advantage of a Single Primordial Cyanobacterial Endosymbiosis

Cavalier Smith's hypothesis of late endosymbiosis between a proto-eukaryote with well-developed phagocytic capability and some α-proteobacterium does not go well with the **monophyletic nature of mitochondria**. If "all kinds of symbionts could be taken up" (Cavalier Smith, 1987), it is difficult to see why one and only one combination could out-compete all the others. The monophyletic nature of mitochondria suggests that the mitochondrial endosymbiosis was a highly unusual and unlikely event. **Evidence that other organelles like mitosomes and hydrogenosomes are also derived from mitochondria is another indication of the extreme parsimony involved in the primary endosymbiont event.** Therefore, the event must have occurred very early in evolution at a time when the mechanism for attaining a stable endosymbiotic system was not yet well established.

In view of the physiological unlikeliness of the so-called respiration-early hypothesis (Paumann et al., 2005) it is logical to hypothesize that the trigger for the emergence of aerobic respiration came from the molecular oxygen first released by the cyanobacterial oxygenic photosynthesis and that **the aerobic respiratory chain was derived from the photosynthetic electron transfer chain, as extensively discussed by the so-called conversion hypothesis** (Broda, 1975; Broda and Peschek, 1979; Peschek 1996 a,b, 2004, 2005, 2008; Paumann et al., 2005). Initially the

function of the respiratory chain must have been the detoxification of O_2 and there must have been a long intermediate phase before it developed into the full-fledged respiratory chain currently found in mitochondria and aerobic bacteria (Paumann et al., 2005). **As the inventors of oxygenic photosynthesis cyanobacteria must naturally have been the first organisms to face the challenge of molecular oxygen which they generated within themselves. They are thus likely to have been the first organisms to have developed oxygen-detoxifying systems via aerobic respiration.** A wide variety of detoxifying enzymes for O_2 and its even more dangerous partially reduced intermediates have been identified in cyanobacteria (Regelsberger et al., 2002; Paumann et al., 2005; Bernroitner et al, 2009).

We propose that, when primitive organisms faced the challenge of toxic oxygen, one of them, **a proto-eukaryote made a lucky break-through by managing to enslave a cyanobacterium and availed the enormous selective advantage conferred by simultaneous acquisition of oxygenic photosynthesis and aerobic respiration, thus taking the famous quantum leap in evolution.** A dramatic increase in atmospheric oxygen levels known as the Great Oxidation Event took place around 2.4 Ga ago though the cyanobacteria had emerged at least 300 million years before already (Kasting, 2006). This time lag has not yet received absolute explanation, though several hypotheses have been offered (Lenton et al, 2004; Goldblatt et al, 2006; Kasting, 2006). We propose that the efficient transfer of photosynthetic capability to eukaryotes made photosynthesis extremely wide-spread and thus contributed significantly to the dramatic rise in oxygen. **This would place the timing of cyanobacterial endosymbiosis around 2.4 Ga ago.**

4.4.4 Separation of the Organelles

During replication of the cyanobacterial endosymbiont within the host, some of the progeny could have lost the thylakoid membrane possibly by mutation of a critical gene involved in biosynthesis of the membranes. This part of the off-spring was transformed into proto-mitochondria specializing in the function of aerobic respiration using the respiratory electron transfer chain that had been situated in the cyanobacterial plasma membrane from the very beginning (Peschek et al., 2004; Paumann et al., 2005). The main function of the proto-mitochondrion must have been oxygen detoxification by aerobic respiration. As the atmosphere became progressively richer in oxygen, the proto-mitochondria developed more efficient and refined mechanisms for aerobic respiration as nowadays found in full-fledged mitochondria. In some of the eukaryotes occupying anaerobic niches, the proto-mitochondria degenerated into hydrogenosomes and mitosomes. It might also be speculated that peroxisomes are vestiges of the proto-mitochondria still contributing to the scavenging of oxygen free radicals.

The original cyanobacterium probably survived longer in its endosymbiont form and was converted to a chloroplast at a later stage (Cavalier Smith, 2006). Somewhere along the way a lineage of non-photosynthetic eukaryotes emerged by secondary loss of pigments and/or the photosynthetic function as mentioned earlier in this chapter.

4.4.5 The majority of the mitochondrial proteome does not show alpha proteobacterial origin

A large fraction of eukaryotic nuclear-encoded genes with a prokaryotic homolog are of eubacterial origin and these are generally assumed to have originated

from the mitochondrial endosymbioint. **However only a surprisingly small fraction of these genes can be traced specifically to alpha proteobacteria** (Gabaldon and Huynen, 2003). Even more surprising is the fact that **less than 20% of the mitochondrial proteome which could be even more legitimately assumed to have originated from the mitochondrial endosymbiont**, **has close homologues in alpha proteobacteria** (Kurland and Andersson, 2000; Martin et al.,2002). A recent work on the origin and evolution of the nuclear-encoded mitochondrial aminoacyl-tRNA synthetases (aaRS) (Brindefalk et al., 2007 and figures 4.11 and 4.12) shows that while all the 20 aaRSs considered originate from within the bacterial clade, not a single one is seen to originate from alpha proteobacteria. Irrespective of the method used and the aaRS analyzed, the results consistently place the node of the mitochondrial divergence within the bacterial domain, but distinct from the a-proteobacterial clade. The mitochondrial aaRS do not show affinity to any specific bacterial clade. The same is the case with glycolytic genes (Canback et al., 2002). This is indicative of the fact that we are dealing with such ancient phenomena that the phylogenetic signals have become hopelessly blurred.

In contrast to non-photosynthetic eukaryotes, the origin of a large percentage of plant nuclear genes can be clearly traced to cyanobacteria (Martin et al., 2002). Using a novel supertree-based phylogenetic signal-stripping method Pisani et al. (2007) show that the strongest phylogenetic signals in eukaryotic genomes link eukaryotes with the cyanobacteria. The results of Brindefal et al. (2007) show that the majority of the organelle aaRSs from plants cluster with cyanobacteria. In the case of glycolytic genes from green plants several of the phylogenetic reconstructions showed close relationship with cyanobacteria. Now the fact that the plant genes show a clear

origin from cyanobacteria has been interpreted as evidence that these are genes that have been transferred to the nucleus from the chloroplast. An alternate explanation is that the nucleotide substitution rates in plant mitochondria are much slower than their non-plant counterparts (Lynch et al., 2006) and therefore ancient phylogenies are shown up much better with plant genes than those using non-photosynthetic eukaryote genes. Therefore the phylogenies shown up by plant genes could be interpreted as showing the true ancestral relationships while the phylogenetic signals have become too blurred in the highly mutated non-plant genes to show up a relationship to any particular eubacterial clade.

4.4.6 Structural and Functional Characteristics of Cyanobacterial and Mitochondrial membranes

Mitochondria, in their inner membrane, contain the highly sophisticated system of chemiosmotic oxidative phosphorylation which, inherently dependent on membrane-bound electron transport, must be the result of a long process of evolution under aerobic conditions. The cyanobacterial respiratory system, though basically similar to the mitochondrial one, is still more primitive and simpler, showing signs of having evolved under an atmosphere poorer in oxygen (Peschek et al., 2004).

Like other Gram-negative bacteria, cyanobacteria have a cell envelope consisting of an outer membrane, a peptidoglycan layer, and a plasma membrane. In addition, these organisms possess an elaborate internal system of intracellular (thylakoid) membranes that host a dual-function photosynthetic–respiratory electron transport chain (Peschek, 1996 a,b). Their plasma membrane carries a pure respiratory chain without the photosynthetic reaction centers (Peschek et al., 2004). Perhaps one of the critical evolutionary innovations of the cell membrane in cyanobacteria was its

ability to invaginate to create a space between the cell membrane and the cell wall (Koning, 1994). The "invaginations", which form contact points between plasma and thylakoid membranes, are sometimes called "mesosomes" or thylakoid centers (Hinterstoisser et al., 1993). These areas of the cell membrane are rich in respiratory electron transport proteins. It has been debated whether the thylakoid membranes themselves are invaginations of the plasma membrane, or if they form separate compartments within the interior of the cyanobacterial cell. Recent electron microscopic studies indicate that the thylakoid membranes are physically discontinuous from the plasma membrane. (Liberton et al., 2006). They may, however, be physically connected with plasma membrane through thylakoid centers (Hinterstoisser et al., 1993).

Mitochondria and chloroplasts are both surrounded by double membranes. Chloroplasts contain, in addition, an internal (thylakoid) membrane system which, in cyanobacteria, carries both respiratory and photosynthetic electron transport chains. The mitochondrial respiratory chain is contained in the inner mitochondrial membrane which would thus be analogous (and homologous?) to the cyanobacterial plasma membrane, and which exhibits numerous invaginations called cristae. Both the mitochondrial inner membrane and the chloroplast thylakoid membrane form closed, osmotically autonomous (thus chemiosmotically competent) compartments and both membranes are stuffed with a reversible ATP synthase of appropriate orientation to catalyze the phosphorylation of ADP to ATP. **Thus physiologically the chloroplast structure is virtually identical to that of cyanobacteria whereas mitochondria could be considered to resemble cyanobacteria without thylakoid membranes.**

The respiratory electron transport chain carried by the cyanobacterial membranes (Figure 4.13) has seen shown to be quite similar to that carried by the mitochondrial inner membrane, and in general – as is one of the main conclusions of the conversion hypothesis (see before) – functionally speaking the electron transport components of both respiratory and photosynthetic chains are strikingly similar to each other (Peschek, 2008).

The electron transport sequence (Peschek et al., 2004; Paumann et al., 2005; Peschek, 2008) will be briefly discussed in the following in order to realize the remarkable similarity of electron transport systems in cyanobacteria, chloroplasts, and mitochondria:

Complex I - NAD(P)H dehydrogenase : Cyanobacteria possess either a multi-subunit "mitochondrial" energy-transducing NDH-1 enzyme or a 1-subunit non-mitochondrial, non-energy transducing NDH-2 enzyme.

Complex II - (SDH) Succinate-dehydrogenase

The occurrence of succinate dehydrogenase has been firmly established in both cyanobacterial membranes. Interestingly, just like the mitochondrial SDH, the cyanobacterial enzyme is inhibited by thenoyltrifluoro acetone (TTFA), and it cross-reacts with monospecific antibody against mitochondrial SDH.

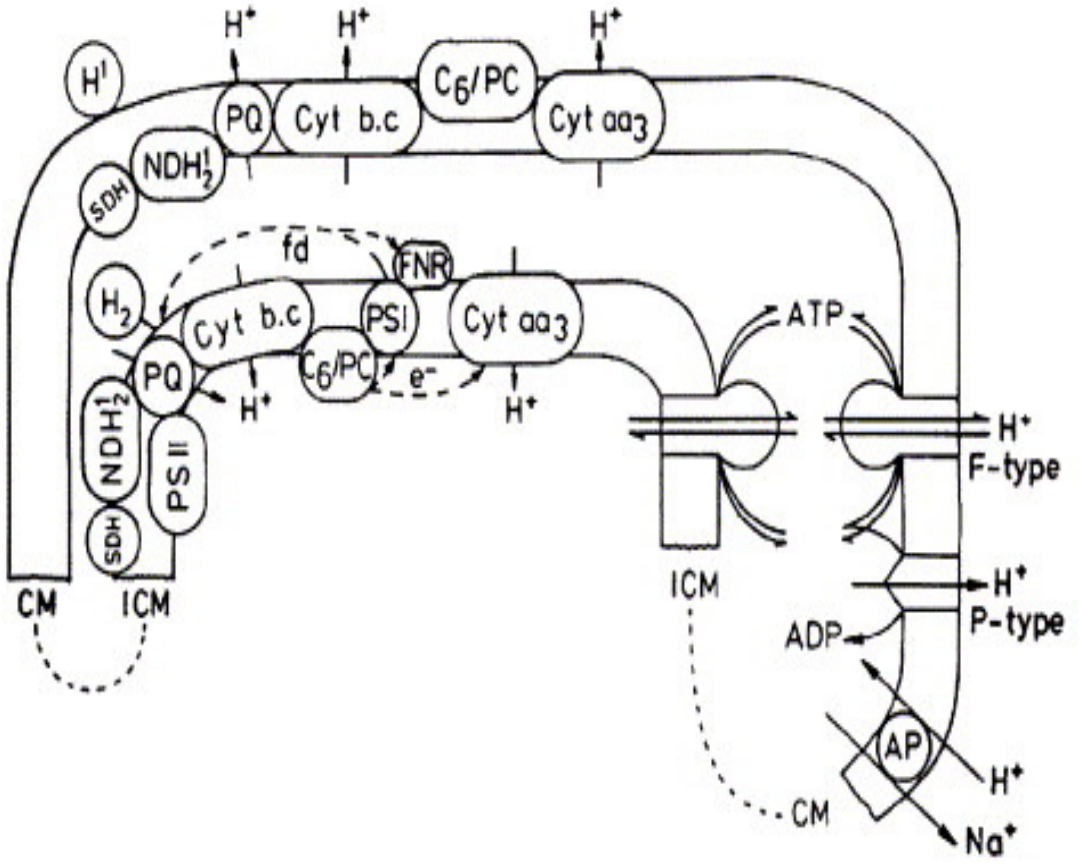


Figure 4.13 Details of the photosynthetic and respiratory electron transport systems in a cyanobacterium

Lipid-soluble mobile carrier

This pool component is ubiquinone in mitochondria but plastoquinone in all cyanobacteria. Redox potentials of ubi- and plasto-quinone (both benzoquinones) are almost the same and they can functionally substitute for each other. **Side chains in positions 3 and 4 are methoxy- in ubiquinone, but methyl- in plasto-quinone which may indicate that ubiquinone replaced plastoquinone as the atmosphere became richer in oxygen.**

Complex III: The cytochrome b.c complex

Functionally speaking this complex is basically the same in cyanobacteria, chloroplasts, and mitochondria. The cytochrome ***b_6f*** complex in chloroplasts and cyanobacteria, and the ***$b.c_1$*** complex in mitochondria, have exactly the same function, viz. as a quinol:cytochrome *c*/PC oxidoreductase.

Water-soluble mobile carrier

In cyanobacteria this can be cytochrome c_6 or PC or, as an electron donor to COX, also cytochrome c_M (Bernroitner et al., 2008b). In chloroplasts of higher plants this carrier is the blue copper protein PC while in mitochondria it inariably is cytochrome *c*, functionally and topologically very similar to cyanobacterial cytochrome c_6. In prokaryotic and eukaryotic algae cytochrome c_6 and PC may be physiologically interchangeable according to availability of Cu in the medium

The terminal respiratory oxidase (TRO)

While there may be several TRO-types in cyanobacteria, e.g. aa_3-type, bo_3-type, *bd*-type, etc., the only TRO unequivocally characterized as a functional protein in up to 35 different cyanobacteria so far is a canonical aa_3-type cytochrome *c* oxidase. The mitochondrial TRO also is an aa_3-type cytochrome *c* oxidase. O_2-affinity, redox properties and inhibition profiles of mitochondrial and cyanobacterial TRO are identical. Like other bacterial TROs its main redox-group-carrying subunit I shows the well-known property of the promiscuity of heme groups. The characteristic difference is that the enzyme is composed of 13 protein subunits in mitochondria but only 3-4 in cyanobacteria (and other bacteria), and that the TON (turnover number) of the cyanobacterial enzyme is lower by a factor of almost 100. Therefore, in spite of its

high O_2-affinity it is obviously not as well adapted to aerobic respiration in fully oxic conditions as is the mitochondrial TRO.

4.4.7 Summary of arguments

The evidences supporting the alternate hypothesis are discussed in detail in the above five sections. Our arguments can be summarized as follows:

- Cyanobacteria, atmospheric oxygen and eukaryotes existed a long time before the origin of alpha-proteobacteria. If eukaryotes acquired mitochondria only after the advent of alpha proteobacteria, it is difficult to see how they survived the toxic onslaught of oxygen till then
- Cyanobacteria being the first producers of oxygen were the first to develop oxygen detoxification capability through aerobic respiration. Early chance endosymbiosis of a cyanobacterium by a proto-eukaryote would have conferred the enormous advantages of photosynthesis and aerobic respiration to this lucky proto-eukaryote.
- The argument that since all eukaryotes possess mitochondria but only a subset possesses chloroplasts need not be considered as a valid argument to show that mitochondrial endosymbiosis took place earlier than chloroplast endosymbiosis. It is possible that earliest eukaryotes had both functions and some lost the photosynthetic function later.
- The monophyletic nature of mitochondria points to endosymbiosis being a rare chance event. Therefore a single event leading to both the organelles is a more parsimonious hypothesis.
- The structural and functional similarity of both mitochondrial and cyanobacterial membranes supports the hypothesis further. It is reasonable to

propose that the primitive cyanobacterial oxygen de-toxification systems developed into efficient respiratory systems as the endosymbiont evolved into mitochondria under the pressure of increasing atmospheric oxygen.

4.4.8 Explanation for the similarity of alpha-proteobacterial proteins to mitochondrial proteins

The primary reason that the search for the bacterial ancestor of mitochondria pointed to α-proteobacteria was that because, among the bacterial clades, the α-proteobacteria were found to be the closest identified relatives of mitochondria on the basis of sequence similarities of several protein-encoding genes on the mitochondrial genomes. However, **the inference that the most closely related bacterial clade contains the mitochondrial ancestor is true only if we are sure that this bacterial clade existed before the emergence of mitochondria**. According to our hypothesis the cyanobacterial endosymbiosis took place around 2.4 Ga, viz., long before the emergence of α-proteobacteria. **Then it can be considered that the proto-mitochondria and the ancestors of α-proteobacteria evolved together in an oxygen rich environment.** This hostile environment was equally new to both and as a result both had to adapt to the novel conditions. **This would have led to the formation of functionally similar proteins by a process of convergent evolution rather than diverging from a common ancestor.** Thus the similarity in their amino acid sequences can be primarily attributed to convergent evolution. Further, the predominantly parasitic nature of α-proteobacteria adds to the similarity of the environments in which mitochondria and α-proteobacteria evolved. It has been suggested that there could also have been some amount of "lateral gene transfer" from

the parasitic α-proteobacteria to mitochondria (Brindefalk et al., 2007). Symbionts belonging to the Rickettsiales have been found in the mitochondria of animal cells (Beninati et al., 2004) and some species of Rickettsiales have been observed to transfer their genes into the nuclear genomes of their hosts (Kondo et al., 2002). Therefore, some of the "well-conserved" mitochondrial genes, including the rRNA-genes, could also have been acquired much later in evolution from parasitic α-proteobacteria.

4.4.9 Importance of nucleotide sequence analysis

This above mentioned possibility of a convergent evolution and the evolutionary plausibility of an alternate hypothesis, increases the significance for a rechecking of the existing phylogeny based on protein sequences with a nucleotide sequence based phylogenetic comparison. Unless the phylogeny based on amino acid sequence variations are well supported by those of nucleotide sequences, there is always room for legitimate doubt. This is especially true when chances of convergent evolution are high. Our results based on genome signatures as well as aligned nucleotide sequences of several genes consistently throw doubts on the currently accepted hypothesis on the origin of mitochondria.

4.5 Conclusion

This chapter describes how a genome signature based phylogenetic analysis led us to a new hypothesis on an important biological question. Intrigued by the discrepancy between phylogenetic analysis based on nucleotide sequences and amino acid sequences we went deeply into factors other than molecular phylogeny which

could complement the results of molecular phylogenetic analysis. We put together a number of arguments to establish the plausibility of the alternate hypothesis. However, as with most hypotheses regarding ancient evolutionary events it is difficult to offer a conclusive proof of this hypothesis. We have, however, shown that the "proof" based on amino acid similarity of the proteins of α-proteobacteria and mitochondria becomes open to doubt when we consider the underlying nucleotide sequences. This case shows an example of the potential of CGR for investigating into "established" phylogenetic relationships from an alternate point of view.

4.6 References

1. Almeida JS, Carrico JA, Maretzek A, Noble PA and Fletcher M (2001) Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics* 17(5): 429--437
2. Andersson JO and Roger AJ (2002) A cyanobacterial gene in nonphotosynthetic protists-An early chloroplast acquisition in Eukaryotes? *Curr. Biol.* 12: 115--119
3. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson AS, Winkler HH and Kurland CG (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria, *Nature* 396 (6707): 133—140
4. Andersson SGE, Karlberg O, Canback B and Kurland CG (2003) On the origin of mitochondria: a genomics perspective, *Phil. Trans. R. Soc. Lond.. B* 358; 165--179
5. Barghoorn ES and Schopf JW (1965) Microorganisms from the late Precambrian of Central Australia, *Science* 150: 337--339
6. Barghoorn ES and Schopf JW (1966) Microorganisms three billion years old from the precambrian of South Africa, *Science* 152: 758—763

7. Beninati T, Lo N, Sacchi L, Genchi L, Noda H and Bandi C (2004) A novel alpha-proteobacterium resides in the mitochondria of ovarian cells of the tick Ixodes ricinus, *Appl. Environ. Microbiol.*70: 2596--2602
8. Bernroitner M, Tangl D, Lucini C, Furtmüller PG, Peschek GA and Obinger C (2008a) Cyanobacterial cytochrome c_M: Probing its role as electron donor for Cu_A of cytochrome *c* oxidase, *Biochim. Biophys. Acta* 1787(3): 135--143
9. Bernroitner M, Zamocky M, Pairer PG, Furtmüller PG, Peschek GA and Obinger C (2008b) Heme-copper oxidases and their electron donors in cyanobacterial respiratory electron transport, *Chem Biodivers.* 5(10): 1927--1961
10. Bernroitner M, Zamocky M, Furtmüller PG, Peschek GA and Obinger C (2009) Occurrence, phylogeny, structure and function of catalases and peroxidases in Cyanobacteria, *J. Exp. Bot.* 60(2): 423--440
11. Blankenship RE (1992) Origin and early evolution of photosynthesis, *Photosynth. Res.* 33: 91--111
12. Blankenship RE and Hartman H (1998) The origin and evolution of oxygenic Photosynthesis, *Trends in Biol. Sci.* 23: 94--97
13. Brindefalk B, Viklund J, Larsson D, Thollesson M and Andersson SGE (2007) Origin and Evolution of the Mitochondrial Aminoacyl-tRNA Synthetases, *Mol. Biol. Evol.* 24 (3): 743--756
14. Brocks JJ, Logan GA, Buick R and Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes, *Science* 285: 1033--1036
15. Broda, E. (1975) *The evolution of the bioenergetic processes*, Pergamon Press, Oxford

16. Broda E and Peschek GA (1979) Did respiration or photosynthesis come first? *J. Theor. Biol.* 81: 201–212

17. Bush EC and Lahn BT (2006) The evolution of word composition in metazoan promoter sequence, *PLoS Comput. Biol.* 2 (11): e150

18. Canback B, Andersson SGE and Kurland CG (2002) The global phylogeny of glycolytic enzymes, *Proc. Natl. Acad. Sci. U.S.A.* 99(9): 6097--6102

19. Cavalier-Smith T (1987) The simultaneous symbiotic origin of mitochondria, chloroplasts, and microbodies, *Ann. N. Y. Acad. Sci.*503: 55--71

20. Cavalier-Smith T (2006) Origin of mitochondria by intracellular enslavement of a photosynthetic purple bacterium, *Proc. R. Soc. B* 273: 1943--1952

21. Chapus C, Dufraigne C, Edwards S, Giron A, Fertil B and Deschavanne P (2005) Exploration of phylogenetic data using a global sequence analysis method, *BMC Evol. Biol.* 5: 63

22. Choi JH , Jung HY, Kim HS and Cho HG (2000) PhyloDraw: a phylogenetic tree drawing system, *Bioinformatics* 16 (11): 1056--1058

23. Christian de Duve (2007) The origin of eukaryotes: a reappraisal, *Nat. Rev. Genet.* 8: 395--403

24. Dehnert M, Plaumann R, Helm WE and Hutt MT (2005) Genome phylogeny based on short-range correlations in dna sequences, *J. Comput. Biol.* 12 (5): 545--553

25. Edwards SV, Fertil B, Giron A and Deschavanne PJ (2002) A genomic schism in birds revealed by phylogenetic analysis of dna strings, *Syst. Biol.* 51: 599--613

26. Emelyanov VV (2002) Phylogenetic relationships of organellar Hsp90 homologs reveal fundamental differences to organellar Hsp70 and Hsp60 evolution, *Gene* 299 (1-2): 125--133

27. Emelyanov VV (2003) Common evolutionary origin of mitochondrial and rickettsial respiratory chains, *Archives of Biochemistry and Biophysics* 420 (1) 130 -- 141

28. Farquhar J, Peters M, Johnston DT, Strauss H, Masterson A, Wiechert U and Kaufman AJ (2007) Isotopic evidence for Mesoarchaean anoxia and changing atmospheric sulphur chemistry, *Nature* 449: 706--709

29. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2), *Cladistics* 5: 164--166

30. Forterre P (2001) Genomics and early cellular evolution. The origin of the DNA world, *C. R. Acad. Sci. Ser. III* 324: 1067--1076

31. Gabaldon, Huynen MA (2003) Reconstruction of the proto-mitochondrial metabolism, *Science* 301: 609

32. Gabaldón T, Snel B, Frank van Zimmeren, Hemrika W, Tabak H and Huynen MA (2006) Origin and evolution of the peroxisomal proteome, *Biology Direct* 1: 8

33. Goldblatt C, Lenton TM and Watson AJ (2006). The Great Oxidation at ~2.4 Ga as a bistability in atmospheric oxygen due to UV shielding by ozone, *Geophys. Res. Abstr.* 8: 00770

34. Gray MW and Doolittle WF (1982) Has the endosymbiont hypothesis been proven? *Microbiol. Rev.* 46: 1--42

35. Gray MW, Burger G and Lang BF (2001) The origin and early evolution of mitochondria, *Genome Biology Reviews* 2(6): 1018.1–1018.5

36. Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes, *Microbiol. Mol. Biol. Rev.* 62: 1435--1491

37. Gupta RS (2000) The phylogeny of Proteobacteria: relationships to other Eubacterial phyla and eukaryotes, *FEMS Microbiol. Rev.* 24: 367--402

38. Gupta RS (2003) Evolutionary Relationships among Photosynthetic Bacteria, *Photosynth. Res.* 76: 173--183

39. Hackstein JHP, Schubert H, Rosenberg J, Mackenstedt M, Berg Mvd, Brul S, Derksen J and Matthijs HCP (1997) Plastid-like organelles in anaerobic mastigotes and parasitic Apicomplexans, In: Schenk HEA, Herrmann RG, Jeon KW, Müller NE and Schwemmler W (eds) *Eukaryotism and Symbiosis. Intertaxonic Combination versus Symbiotic Adaptation*, pp 49--55. Springer Verlag, Berlin

40. Han TM and Runnegar B (1992) Megascopic eukaryotic algae from the 2.1-billion-year-old negaunee iron-formation, Michigan, *Science* 257: 232--235

41. Hartmann, H. (1998) Photosynthesis and the origin of life, *Origins of Life and the Evolution of the Biosphere (OLEB)* 28: 515-521

42. Hedges SB, Chen H, Kumar S, Wang DYC, Thompson AS and Watanabe H (2001) A genomic timescale for the origin of eukaryotes, *BMC Evol. Biol.*1: 4

43. Hinterstoisser B, Cichna M, Kuntner O and Peschek GA (1993) Cooperation of plasma and thylakoid membranes for the biosynthesis of chlorophyll in cyanobacteria: The role of the thylakoid centers, *J. Plant Physiol.*142: 407--413

44. Jacob F (1977) Evolution and tinkering, *Science* 196: 1161--1166

45. Jeffrey H.J (1990) Chaos game representation of gene structure, *Nucleic Acids Res.* 18: 2163--2170

46. Jeon KW (1995) Bacterial endosymbiosis in amoebae, *Trends Cell Biol*. 5:137--140

47. Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature, *Trends Genet.* 11: 283--290

48. Karlin S and Ladunga I (1994) Comparisons of eukaryotic genomic sequences, *Proc. Natl. Acad. Sci. U.S.A.* 91: 12832--12836

49. Kasting JF (2006) Ups and downs of ancient oxygen, *Nature* 443: 643--645

50. Kondo N, Nikoh N, Ijichi N, Shimada M and Fukatsu T (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect, *Proc. Natl. Acad. Sci. U.S.A.* 99: 14280--14285

51. Koning and Ross E (1994) Cyanophyta. Plant Physiology Information http://plantphys.info/plant_biology/cyanophyta.shtml (June 22, 2006)

52. Kurland CG and Andersson SGE (2000) *Microbiology And Molecular Biology Reviews*, 786–820

53. Kurland CG, Collins LJ and Penny D (2006) Genomics and the irreducible nature of Eukaryote cells, *Science* 312: 1011--1014

54. Lenton TM, Schellnhuber HJ and Szathmáry E (2004) Climbing the co-evolutionary ladder, *Nature* 431: 913

55. Liberton M, Berg HR, Heuser J, Roth R, and Pakrasi HB (2006) Ultrastructure of the membrane systems in the unicellular cyanobacterium Synechocystis sp. strain PCC 6803, *Protoplasma* 227: 129--138

56. Lynch M, Koskella B and Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture, *Science* 311: 1727--1730

57. Margulis L (1981) *Symbiosis in Cell Evolution : Life and its environment on the early earth,* W.H. Freeman, San Francisco

58. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M and Penny D (2002) *Proc Natl Acad Sci U.S.A.* 99(19):12246--12251

59. Mereschkowski C (1905) Über Natur und Ursprung der Chromatophoren im Pflanzenreiche, *Biol Centralbl* 25: 593—604

60. Paumann M, Regelsberger G, Obinger C and Peschek GA (2005) The bioenergetic role of dioxygen and the terminal oxidase(s) in cyanobacteria, *Biochim. Biophys. Acta (BBA) – Bioenergetics* 1707 (2-3): 231--253

61. Peschek, GA (1996a) Cytochrome *c* oxidase and the *cta* operon of cyanobacteria, *Biochim. Biophys. Acta* 1275: 27--32

62. Peschek GA (1996b) Structure-function relationships in the dual-function photosynthetic-respiratory electron transport assembly of cyanobacteria, *Biochem. Soc. Trans.* 24: 729--733

63. Peschek GA (2005) Cyanobacteria viewed as free-living chloromitochondrtia, In: Est Avd. And Bruce D. (eds) *Photosynthesis: Fundamental Aspects to Global Perspectives*, pp 746—749, The International Society of Photosynthesis, Toronto, Canada

64. Peschek GA (2008) Electron transport chains in oxygenic cyanobacteria, In: Renger G (ed) *Primary Processes of Photosynthesis: Principles and Applications*,

2: 383--415, European Society of Photobiology, The Royal Society of Chemistry, Great Britain

65. Peschek GA, Obinger C and Paumann M (2004) The respiratory chain of blue-green algae (cyanobacteria), *Physiologia Plantarum* 120(3): 358—369
66. Perrière, G. and Gouy, M. (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78: 364--369.
67. Pisani D, Cotton JA, and McInerney JO (2007) Supertrees Disentangle the Chimeric Origin of Eukaryotic genomes, *Mol. Biol. Evol.* 24(8):1752–1760
68. Pringsheim EG (1949) Colorless algae, *Bact. Rev.* 13: 47--56
69. Pringsheim EG (1963) *Farblose Algen*, Gustav Fischer Verlag, Stuttgart
70. Pringsheim EG and Wiessner W (1960) Colorless phototrophs? *Nature* 188: 919--920
71. Purton S (2002) Going green: the evolution of photosynthetic eukaryotes. *Microbiology Today* 29: 126--128
72. Qi J, Wang B and Hao BI (2004) Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach, *J. Mol. Evol.* 58: 1--11
73. Regelsberger G, Jakopitsch C, Plasser L, Schwaiger HJ, Furtmüller PG, Peschek GA, Zamocky M and Obinger C (2002) Occurrence and biochemistry of hydroperoxidases in oxygenic phototrophic prokaryotes (cyanobacteria), *Plant Physiol. Biochem.* 40: 479--490
74. Sarfaty CK, Mi Oh J, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity, *Science* 315: 525--528
75. Schopf JW (1970) Precambrian microorganisms and evolutionary events prior to

the origin of vascular plants, *Biol. Rev.* 45: 319--352

76. Sicheritz-Ponten T and Andersson SG (2001) A phylogenomic approach to microbial evolution, *Nucleic Acids Res.* 29: 545--552
77. Simmons MP, Carr TG and Neill KO (2004) Relative character-state space, amount of potential phylogenetic information and heterogeneity of nucleotide and amino acid characters, *Mol. Phylogenet. Evol.* 32: 913--926
78. Stanier RY (1970) Some aspects of the biology of cells and their possible evolutionary significance, *Symp. Soc. Gen. Microbiol.* 20: 1--38
79. Tabak HF, Hoepfner D, Zand Avd, Geuze HJ, Braakman I and Huynen MA (2006) Formation of peroxisomes: Present and past, *Biochim. Biophys. Acta (BBA) - Molecular Cell Research* 1763(12): 1647--1654
80. Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.* 22: 4673– 4680
81. Wang Y, Hill K, Singh S and Kari L (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation, *Gene* 346: 173--185
82. Woese CR, Magrum LJ and Fox GE (1978) Archaebacteria, *J. Mol. Evol.* 11: 245--252
83. Xiong J, Fischer WM, Inoue K, Nakahara M and Bauer CE (2000) Molecular evidence for the early evolution of photosynthesis, *Science* 289: 1724--1730
84. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ and Woese CR (1985) Mitochondrial origins, *Proc. Natl. Acad. Sci. U.S.A.* 82 (13): 4443—4447

Summary and Scope for future work

Summary

The preceding chapters present our attempts to develop the technique of Chaos Game Representation from a mere sequence visualization algorithm to a unique tool with versatile abilities to draw out the information content present in colossal biological sequence databases.

The first chapter gives a comprehensive introduction to CGR and explains the construction of a Frequency Chaos Game Representation (FCGR), which is in fact a square matrix representing the frequency of different oligonucleotides in a sequence. The inception of CGR in 1990 by HJ Jeffrey and the works by various authors raising the stature of CGR to a valid sequence analysis technique is discussed in detail in the chapter. The distance measure between two sequences, proposed by Almeida et al. in 2001, computed from FCGR and based on Pearson correlation coefficient is described here. The concept of genome signature brought out by Deshavanne et al. in 1999, shows that subsequence of a genome exhibits characteristics of the whole genome. The characteristic which is preserved here is the short oligonucleotide composition and hence FCGR becomes an ideal tool in analyzing genome signature.

A key property of CGR is that, the source sequence can be retraced upto any desired level from the coordinates of the CGR points. This previously unutilized characteristic of CGR is made use in the second chapter to build an alignment based comparison of genomes. Here, an algorithm for identifying all local alignments

between two long DNA sequences using the sequence information contained in CGR points is developed. This is done by defining the distance between the CGR points (x_i, y_i) and (x_j, y_j) of two positions of the two sequences respectively as $d(i,j) = \max(abs(X_i - X_j), abs(Y_i - Y_j))$. Further it is found that for k identical nucleotides in both sequences the equation becomes $d(i,j) = (0.5)^k d(i-k, j-k)$. The length k of identical sequence is thus found out and its location is further marked to identify the region of local alignment. The alignments thus found are depicted graphically in a dot-matrix plot or in text form. The program execution time is further made fast by using an anchored alignment approach similar to that used in other programs such as FASTA. The anchoring is done by taking similar 'k-mers' in both the sequences, where the k-mers are further found by FCGR. Neighbouring local alignments with a gap, in-del or mismatch separating them are then chained together. Inversions of segments in the sequences are also accounted for by taking the reverse complement of the second sequence and comparing with the first sequence. Genomes of several closely related microbial species are compared and the results are illustrated in the chapter.

The phylogenetic signal contained in genome signature is the focal theme of the third chapter. The inherent difficulties in the traditional molecular phylogenetic methods based on sequence alignment are discussed carefully. The distance measure mentioned in chapter one, based on Pearson correlation coefficient between the FCGRs of two sequences, is utilized to compute the pairwise distance between two sequences. The distance matrix thus obtained is used to construct phylogenetic trees. The results thus obtained on evolutionarily related data sets clearly indicate the

presence of phylogenetic signal in genome signature. Established phylogenetic relationships could be recreated to a fair extent using this alignment free technique. The fact that correlation of two data sets is scale independent and since this distance measure is based on the correlation of FCGRs, two sequences of varying dimensions can be compared meaningfully using this technique.

The effect of varying the order of the FCGR on the resulting trees is studied in this chapter. It can be observed in general that trees based on lower order FCGR is more suited for comparing sequences of distantly related organisms, while higher order FCGR is more suited for closely related organisms. Another interesting result is the 'chromosome wise' tree of organisms. The lower order FCGRs caused the intermingling of chromosomes of closely related organisms, that too pairing of 'counterpart' chromosomes in the case of the Human – Chimpanzee chromosome distance tree. However, the higher order FCGRs classifies the chromosomes into separate groups, but at the cost of losing inter-species relationships. The tree based on bias-compensated third order FCGR, i.e. each trinucleotide frequency divided by its component monomer frequencies to nullify the effect of base composition, displayed the twin advantage of species specificity and preservation of inter-species relation. A distinctiveness of the signature method to be noticed here is that a chromosome wise distance tree would not make any sense in an alignment based method. The presence of a distinct phylogenetic signal in genome signature can be inferred from the results in this chapter.

This distinct phylogenetic signal in FCGR is utilized in the fourth chapter in re-examining the currently accepted hypothesis on the origin of the eukaryotic

organelle, mitochondria. Molecular phylogeny based on aminoacid sequences is the main result in support of the hypothesis that an endosymbiosis of an alpha-proteobacteria led to the formation of mitochondrion. However, genome signature results depict that the signature of mitochondria is closer to cyanobacteria and chloroplasts than to its hypothesized ancestor, the alpha-proteobacteria. Further, taking into account the unique capability of cyanobacteria to perform both oxygenic photosynthesis and aerobic respiration, an alternate more parsimonious hypothesis is proposed; that a single endosymbiotic uptake of a cyanobacterium could have led to the evolution of both the organelles. We bring together arguments such as parsimony, timing of evolutionary and geological events, structural and functional similarity of cyanobacterial membranes to both the organellar membranes, so as to create the plausibility of this alternate hypothesis. Multiple sequence alignment done on the nucleotide sequences of ribosomal proteins and NADH dehydrogenase genes also supported the new hypothesis. Moreover, the signature based distance between the mitochondria and chloroplasts is relatively small while the distance from the organelles to the nuclear genome is large for the early eukaryotes (protists), which again points to a common endosymbiotic ancestry for both the organelles. Molecular phylogenetics is a powerful tool for probing ancestral relationships, but it is known not to be infallible. Hence, the discrepancies in the results obtained by amino acid based methods and nucleotide sequence based methods must be considered seriously. If this alternate hypothesis finds acceptance, it would imply that we take a serious re-look at other 'established' phylogenetic relationships based solely on amino acid sequence similarity.

In conclusion this work considerably enhances the current repertoire of applications of Chaos Game Representation and positions it as an important genome sequence analysis tool .

Scope for future work

We hope that this work will spawn renewed interest in Chaos Game Representation, both in terms of using the currently available applications to address important biological questions as well as to develop new applications.

Biologists can use phylogenetic analysis based on CGR as a tool for having a re-look into other 'established' phylogenies. For a given evolutionary tree to be well established trees based on different construction methods should complement each other. Whenever there is a contradiction among the results, there is space for an alternate evolutionary relationship. The effect of varying the order of FCGR on the resulting distance trees is to be further studied and representations that combine the advantages of different orders can be thought of.

Information scientists can further go in depth by utilizing the sequence (data) retracing property from a single CGR point. The technique can be employed in fields beyond biological sequences. The information content in a single CGR point makes the technique a likely candidate for lossless file compression techniques. The computing of frequencies of oligomers of various lengths in a single run of the program can be utilized to make FCGR a faster pattern-searching algorithm. An efficient web-based software tool for CGR based sequence analysis would be a very useful addition to Bioinformatics tools.

List of Publications

1. Chaos game representation for comparison of whole genomes. **Jijoy Joseph** and Roschen Sasikumar, BMC Bioinformatics, 2006, 7: 243

2. An alternate hypothesis for the origin of mitochondria. Roschen Sasikumar, **Jijoy Joseph** and G.A. Peschek, The Bioenergetic Processes of Cyanobacteria – From Evolutionary Singularity to Ecological Diversity, Springer International, New York (Ed.) Guenter.A. Peschek, University of Vienna, *In press.*

Papers Presented at Conferences

1. Investigation of endosymbiont theory using genome signatures. **Jijoy Joseph**, George Titus, K.Rajeev, Mariamma Cherian, Roschen Sasikumar (Presented at *Kerala Science Congress*, January 2006)
2. Genome signature analysis throws new light on the origins of DOX genes, mitochondria and chloroplasts. Roschen Sasikumar, **Jijoy Joseph** (Presented at International *Conference on Bioinformatics, InCoB-2006,* New Delhi, December 2006)
3. Origin of mitochondria re-investigated using Genome Signatures. Roschen Sasikumar, **Jijoy Joseph** (Presented at *Bioinformatica Indica – 08*, University of Kerala, January 2008)