DEVELOPMENT OF A DATABASE ON MITOCHONDRIAL GENE ORDER, REARRANGEMENT EVENTS AND THEIR EVOLUTIONARY IMPLICATIONS

Project Report submitted in partial fulfillment of the Degree of

Bachelor of Technology

In

Bioinformatics

Under the Supervision of

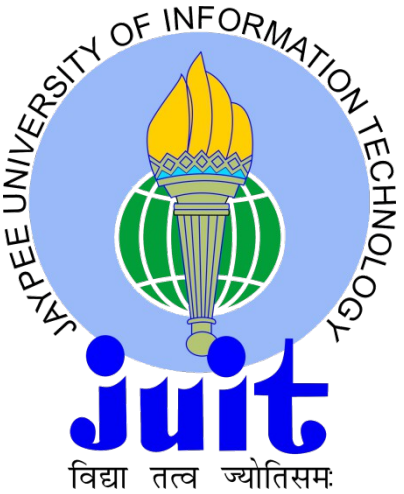
Dr. Tiratha Raj Singh

By

Aakanksha Srivastava (101511)

Harpriya Arora (101503)

To

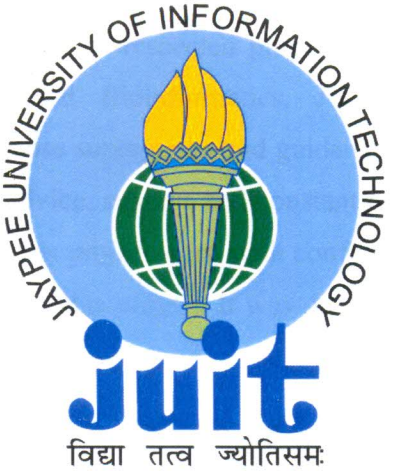


Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology

Waknaghat, Solan-173234, Himachal Pradesh

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, HIMACHAL PRADESH



CERTIFICATE

This is to certify that the work titled "Development of a database on mitochondrial gene order, rearrangement events and their evolutionary implications" submieed by Aakanksha Srivastava(101511) and Harpriya Arora(101503) in partial fulfilment for the award of degree of B. Tech Bioinformatics of Jaypee University of Information Technology, Waknagaht has been carried out under mu supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor .....

Name of Supervisor Dr. Tiratha Raj Singh

Designation Assistant Professor, JUIT

Date 26/05/2014

ACKNOWLEDGEMENT

We express our sincere gratitude to our respected project supervisor Dr. Tiratha Raj Singh, Department Of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat under whose supervision and guidance this work has been carried out. His whole hearted involvement, advice, support and constant encouragement throughout, have been responsible for carrying out this project work with confidence. We are thankful to him for showing confidence in us to take up this project. It was due to his planning and guidance that we were able to complete this project in time.

We are sincerely grateful to Dr. R. S. Chauhan, Professor and Head of Department Of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat for providing all the necessities for the successful completion of our project.

Signature of the student	Aakanksha.	Signature of the student	Harpriya Arora
Name of Student	Aakanksha Srivastava	Name of Student	Harpriya Arora
Date	26.5.14	Date	26.5.14

ABSTRACT

Databases hold vast importance in the field Research and Development. This statement implies that databases are important for the retrieval of data. Any kind of data that the scientific community works upon is usually downloaded from the internet. Hence, we aim to develop such kind of a database that helps the scientific community to extract and utilize the data in an effective way. Mitochondrial Gene Order Database (MGOD) is one such database that has never existed in the past. It aims to capture various details of the mitochondrial genome in organisms. It exhibits manually curated data that is present in a comprehensive manner. It is a public database i.e. it is freely accessible for everyone all over the world. The web address of the database is www.bioinfoindia.org/mgod.

We aim to present a comprehensive database which could help in efficient extraction of information regarding the mitochondrial genome. The reason for developing such kind of a database is that it is the first database to incorporate such information. Plus, mitochondrial genome holds great importance in the field of Research & Development. Mitochondrial genome plays great role in phylogenetic studies. Hence, such kind of a database would be helpful to provide information that could be easily accessed and utilised.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Mitochondrial DNA was discovered in the 1960s by Margit M. K. Nass and Sylvan Nass by electron microscopy as DNase-sensitive thread inside mitochondria, and by Ellen Haslbrunner, Hans Tuppy and Gottfried Schatz by biochemical assays on highly purified mitochondrial fractions

Researchers found that Mitochondrial DNA is important for a number of reasons. 13 of its 37 genes are involved in the process known as oxidative phosphorylation. This is the metabolic pathway that produces adenosine triphosphate (ATP), the main energy source of the cell. The remaining 24 genes are involved in the creation of transfer RNA (tRNA) and ribosomal RNA (rRNA) which help to turn amino acids into proteins.

Mitochondrial DNA also appears to be important for a healthy body as there are a number of genetic disorders associated with changes in mitochondrial genes. Some of these genetic disorders are Cancers, Leber Hereditary Optic Neuropathy, Nonsyndromic deafness.

Mitochondrial DNA is also enjoying increasing importance in forensic science. Criminals can be foiled as it is also used in DNA fingerprinting. It has several useful attributes which are mentioned below -

- There are hundreds of mitochondria in each cell and hundreds of genes as well which provides forensic scientists with abundant source material.
- Mitochondria are well protected in the cell. Hence, they do not suffer from degradation as in case of nuclear DNA.
- As mitochondria are passed solely down the maternal line, any maternal relative can be used as a reference sample. Though this can be a disadvantage in that it is not a unique identifier. This is because relatives within that maternal lineage will share the same mitochondrial DNA.

Because of the above mentioned features of mitochondrial DNA, there was a need for the development of a database on mitochondrial gene order that would help scientific community to fetch data from a single source and use it for various types of studies Plus, the fact that no such database exists, we were inclined towards development of such kind of a database. This database includes advanced search options to search for the organism by Organism Name,

Organism ID and Gene Name. It also takes into account the gene rearrangement events for the organisms.

1.1 MITOCHONDRIAL GENOME

Mitochondria are structures within cells that convert the energy from food into a form that cells can use. Although most DNA is packaged in chromosomes within the nucleus, mitochondria also have a small amount of their own DNA. This genetic material is known as mitochondrial DNA or mtDNA. In humans, mitochondrial DNA spans about 16,500 DNA building blocks (base pairs), representing a small fraction of the total DNA in cells.

Mitochondrial DNA contains 37 genes, all of which are essential for normal mitochondrial function. Thirteen of these genes provide instructions for making enzymes involved in oxidative phosphorylation. Oxidative phosphorylation is a process that uses oxygen and simple sugars to create adenosine triphosphate (ATP), the cell's main energy source. The remaining genes provide instructions for making molecules called transfer RNA (tRNA) and ribosomal RNA (rRNA), which are chemical cousins of DNA. These types of RNA help assemble protein building blocks (amino acids) into functioning proteins.

Mitochondrial genes are among the estimated 20,000 to 25,000 total genes in the human genome.

1.2 STRUCTURE OF MITOCHONDRIAL DNA

In most multicellular organisms, the mitochondrial DNA (mtDNA) is organized as a circular, covalently closed, double-stranded DNA. Whereas in various unicellular (e.g. the ciliate, *Tetrahymena* or the green alga, *Chlamydomonas reinhardtii*) and some multicellular organisms (e.g. in some species of *Cnidaria*) the mtDNA is found as linearly organized DNA. Most of these linear mtDNAs possess telomerase independent telomeres (i.e. the ends of the linear DNA) with different modes of replication, which have made them interesting objects of research, as many of these unicellular organisms with linear mtDNA are known pathogens. In humans (and probably in metazoans in general), 100-10,000 separate copies of mtDNA are usually present per cell (egg and sperm cells being the exceptions). In mammals, each double-stranded circular mtDNA molecule consists of 15,000-17,000 base pairs.

The two strands of mtDNA are differentiated by their nucleotide content with the guanine-rich strand referred to as the heavy strand (or H-strand), and the cytosine-rich strand referred to as the light strand (or L-strand). The heavy strand encodes 28 genes, and the light strand encodes

9 genes for a total of 37 genes. Of the 37 genes, 13 are for proteins (polypeptides), 22 are for transfer RNA (tRNA) and two are for the small and large subunits of ribosomal RNA (rRNA). This pattern is also seen among most metazoans, although in some cases one or more of the 37 genes is absent and the mtDNA size range is greater. Even greater variation in mtDNA gene content and size exists among fungi and plants, although there appears to be a core subset of genes that are present in all eukaryotes (except for the few that have no mitochondria at all). Some plant species have enormous mtDNAs (as many as 2,500,000 base pairs per mtDNA molecule) but, surprisingly, even those huge mtDNAs contain the same number and kinds of genes as related plants with much smaller mtDNAs.

The image given below gives the circular representation of the mitochondrial genome -

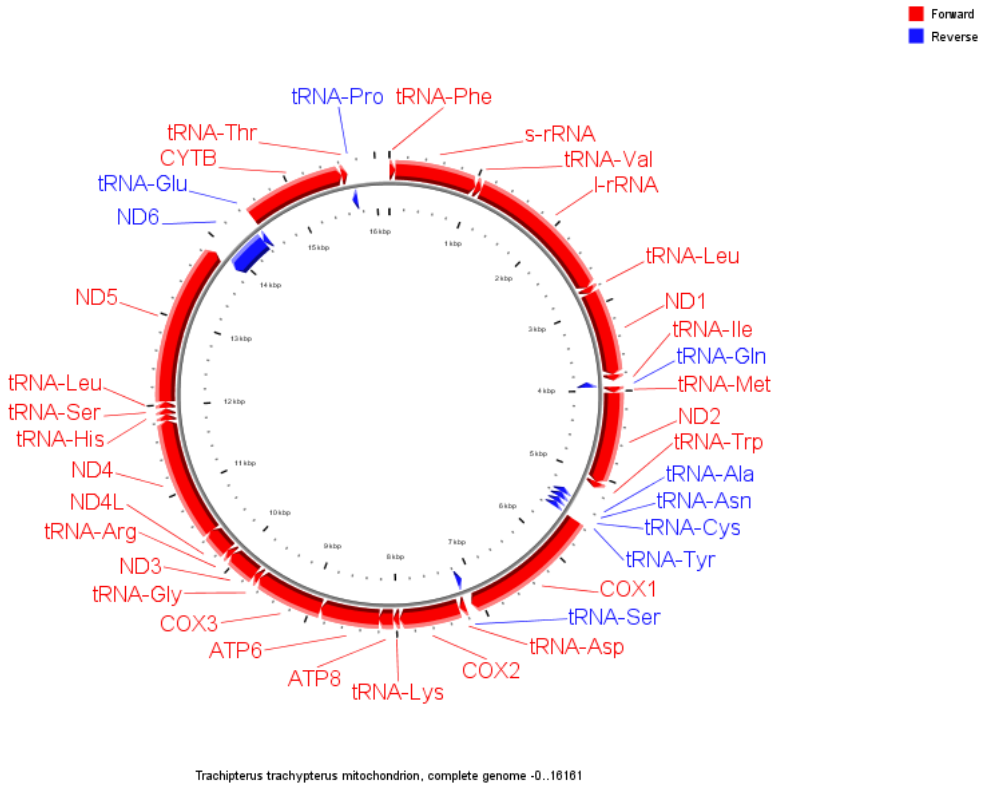


Figure 1.1

1.3 MITOCHONDRIAL GENETICS

Mitochondrial genetics differ from nuclear genetics in three aspects.

- Mitochondrial genes are maternally inherited; they do not follow a Mendelian pattern of inheritance.

- Mitochondrial genome is polyploid. Normally, a state of homoplasmy exists where only one form of mtDNA is present. Mutation can lead to a state of heteroplasmy where two or more forms of mtDNA coexist within a cell.
- Unlike a diploid nuclear gene that can normally only assume three states (homozygous wild type, heterozygous, or homozygous mutant), mtDNA heteroplasmy does not vary by discrete steps. The proportions of mtDNA species can vary with time or through mitotic segregation as cells divide.

1.3.1 MITOCHONDRIAL INHERITANCE

In most multicellular organisms, mtDNA is inherited from the mother (maternally inherited). Mechanisms for this include simple dilution, degradation of sperm mtDNA in the fertilised egg and at least in a few organisms, failure of sperm mtDNA to enter the egg. Whatever the mechanism, this single parent (uniparental) pattern of mtDNA inheritance is found in most animals, most plants and in fungi as well.

1.3.1.1 FEMALE INHERITANCE

In sexual reproduction, mitochondria are normally inherited exclusively from the mother; the mitochondria in mammalian sperm are usually destroyed by the egg cell after fertilization. Also, most mitochondria are present at the base of the sperm's tail, which is used for propelling the sperm cells; sometimes the tail is lost during fertilization. In 1999 it was reported that paternal sperm mitochondria (containing mtDNA) are marked with ubiquitin to select them for later destruction inside the embryo. Some *in vitro* fertilization techniques, particularly injecting a sperm into an oocyte, may interfere with this. Males and females both inherit their mitochondria from their mother but males cannot transmit their mitochondria to subsequent generations.

The fact that mitochondrial DNA is maternally inherited enables genealogical researchers to trace maternal lineage far back in time. (Y-chromosomal DNA, paternally inherited, is used in an analogous way to trace the agnate lineage.) This is accomplished on human mitochondrial DNA by sequencing one or more of the hyper variable control regions (HVR1 or HVR2) of the mitochondrial DNA, as with a genealogical DNA test. HVR1 consists of about 440 base pairs. These 440 base pairs are then compared to the control regions of other individuals (either specific people or subjects in a database) to determine maternal lineage. Most often, the

comparison is made to the revised Cambridge Reference Sequence. Vilà *et al.* have published studies tracing the matrilineal descent of domestic dogs to wolves. The concept of the Mitochondrial Eve is based on the same type of analysis, attempting to discover the origin of humanity by tracking the lineage back in time.

As mtDNA is not highly conserved and has a rapid mutation rate, it is useful for studying the evolutionary relationships—phylogeny—of organisms. Biologists can determine and then compare mtDNA sequences among different species and use the comparisons to build an evolutionary tree for the species examined.

1.3.1.2 MALE INHERITANCE

It has been reported that mitochondria can occasionally be inherited from the father in some species such as mussels. Paternally inherited mitochondria have additionally been reported in some insects such as fruit flies, honeybees, and periodical cicadas.

Evidence supports rare instances of male mitochondrial inheritance in some mammals as well. Specifically, documented occurrences exist for mice, where the male-inherited mitochondrian was subsequently rejected. It has also been found in sheep, and in cloned cattle. It has been found in a single case in a human male.

While many of these cases involve cloned embryos or subsequent rejection of the paternal mitochondria, others document *in vivo* inheritance and persistence under lab conditions.

1.3.2 CHROMOSOMAL REARRANGEMENT

A chromosomal rearrangement is a type of chromosome abnormality involving a change in the structure of the native chromosome. Chromosomal rearrangements encompass several different classes of events: deletions, duplications, inversions; and translocations. Each of these events can be caused by breakage of DNA double helices in the genome at two different locations, followed by a rejoining of the broken ends to produce a new chromosomal arrangement of genes, different from the gene order of the chromosomes before they were broken. Consistent with the origin of chromosomal rearrangements by breakage, rearrangements can be induced artificially by using ionizing radiation. This kind of radiation, of which X rays and gamma rays are the most commonly used, is highly energetic and causes numerous double-stranded breaks in DNA.

1.3.2.1 STRUCTURAL ABNORMALITIES

When the chromosome's structure is altered, it can take several forms:

- **Deletions**

A portion of the chromosome is missing or deleted. Known disorders in humans include Wolf-Hirschhorn syndrome, which is caused by partial deletion of the short arm of chromosome 4; and Jacobsen syndrome, also called the terminal 11q deletion disorder.

- **Duplications**

A portion of the chromosome is duplicated, resulting in extra genetic material. Known human disorders include Charcot-Marie-Tooth disease type 1A, which may be caused by duplication of the gene encoding peripheral myelin protein 22 (PMP22) on chromosome 17.

- **Translocations**

A portion of one chromosome is transferred to another chromosome. There are two main types of translocations:

- **Reciprocal translocation:** Segments from two different chromosomes have been exchanged.
- **Robertsonian translocation:** An entire chromosome has attached to another at the centromere - in humans these only occur with chromosomes 13, 14, 15, 21, and 22.

- **Inversions**

A portion of the chromosome has broken off, turned upside down, and reattached, therefore the genetic material is inverted.

- **Insertions**

A portion of one chromosome has been deleted from its normal place and inserted into another chromosome.

- **Rings**

A portion of a chromosome has broken off and formed a circle or ring. This can happen with or without loss of genetic material.

- **Isochromosome**

It is formed by the mirror image copy of a chromosome segment including the centromere.

Chromosome instability syndromes are a group of disorders characterized by chromosomal instability and breakage. They often lead to an increased tendency to develop certain types of malignancies.

1.4 GENETIC CODE VARIANTS

The genetic code is, for the most part, universal, with few exceptions: mitochondrial genetics includes some of these. For most organisms the stop codons are “UAA”, “UAG”, and “UGA”. In vertebrate mitochondria “AGA” and “AGG” are also stop codons, but not “UGA”, which codes for tryptophan instead. “AUA” codes for isoleucine in most organisms but for methionine in vertebrate mitochondrial mRNA.

There are many other variations among the codes used by other mitochondrial m/tRNA, which happened not to be harmful to their organisms, and which can be used as a tool (along with other mutations among the mtDNA/RNA of different species) to determine relative proximity of common ancestry of related species. (The more related two species are the more mtDNA/RNA mutations will be the same in their mitochondrial genome).

1.4.1 GENE ORDER

Gene order is the permutation of genome rearrangement. Mitochondrial genomes are a valuable source of data for analysing phylogenetic relationships. Besides sequence information, mitochondrial gene order may add phylogenetically useful information as well.

The mitochondrial genetic code (which is used to decipher only 13 different mitochondrial mRNAs on mitochondrial ribosomes) differs slightly from the nuclear genetic code (which specifies perhaps about 70 000–80 000 different mRNAs on cytoplasmic ribosomes). The mitochondrial genome encodes all the ribosomal RNA and tRNA molecules it needs for synthesizing proteins but relies on nuclear-encoded genes to provide all other components (such as the protein components of mitochondrial ribosomes, amino acyl tRNA synthetases, etc.).

As there are only 22 different types of human mitochondrial tRNA, individual tRNA molecules need to be able to interpret several different codons. Eight of the 22 tRNA molecules have anticodons which are each able to recognize families of four codons differing only at the third base, and 14 recognize pairs of codons which are identical at the first two base positions and share either a purine or a pyrimidine at the third base. Between them, therefore, the 22 mitochondrial tRNA molecules can recognize a total of 60 codons [(8 × 4) + (14 × 2)]. The remaining four codons, UAG, UAA, AGA and AGG cannot be recognized by mitochondrial tRNA and act as stop codons.

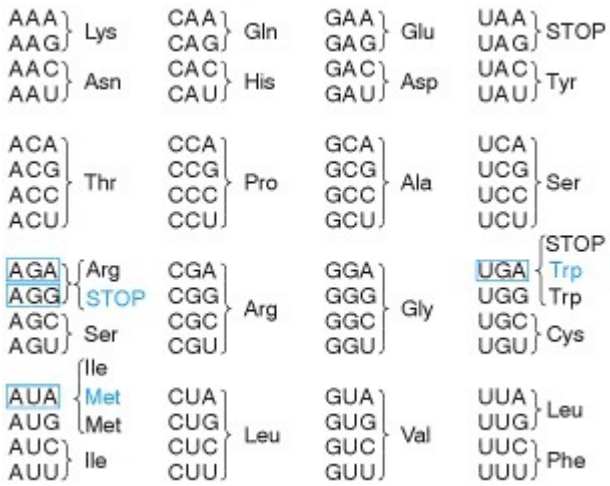


Figure 1.2

The nuclear and mitochondrial genetic codes are similiar but not identical

Blue boxes indicate the four codons which are interpreted differently in the nucleus and mitochondria of mammalian cells, with the mitochondrial interpretation given in blue. Thus, the mitochondrial code has 4 stop codons instead of three (UAA, UAG, AGA, AGG), two Trp codons instead of one (UGA, UGG), four Arg codons instead of six (CGA, CGC, CGG, CGU), two Met codons instead of one (AUA, AUG) and two Ile codons instead of three (AUC, AUU). *Note* that degeneracy of the genetic code most often involves the third base of the codon. Sometimes any base may be substituted (GGN = glycine, CCN = proline, etc., where N is any base). In other cases, any purine (Pu) or any pyrimidine (Py) will do (AAPu = lysine, AAPy = asparagine etc.).

Mitochondrial gene order is highly variable among animal phyla and has been considered a useful phylo-genetic character. Vertebrate mitochondrial gene order was initially considered completely conserved. DNA sequences of the entire mitochondrial genome from mammals and the African clawed frog suggested that all vertebrates shared a common gene order, and sequences from fish later also revealed this order. Three types of deviations from the most common gene arrangement subsequently have been identified: duplication of genes,

rearrangement of genes, and loss of a recognizable origin for replication of the light strand between the genes encoding the asparagine and cysteine tRNAs. According to research papers, the vertebrate mitochondrial genome has found to be comparatively more malleable than it was originally thought to be. For historical reasons, most DNA sequencing being conducted for phylogenetic inference utilizes sequences of single genes or fragments of genes. Recently, tRNA genes have been found to be useful for phylogenetic inference.

1.5 CODING AND NON CODING DNA

The only significant region lacking any known coding DNA is the displacement (D) loop region. This is the region in which a triple-stranded DNA structure is generated by synthesizing an additional short piece of the H-strand DNA, known as 7S DNA (see Figure 4. The replication of both the H and L strands is unidirectional and starts at specific origins. In the former case, the origin is in the D loop and only after about two-thirds of the daughter H strand has been synthesized (by using the L strand as a template and displacing the old H strand) does the origin for L strand replication become exposed. Thereafter, replication of the L strand proceeds in the opposite direction, using the H strand as a template. The D loop also contains the predominant promoter for transcription of both the H and L strands. Unlike transcription of nuclear genes, in which individual genes are almost always transcribed separately using individual promoters, transcription of the mitochondrial DNA starts from the promoters in the D loop region and continues, in opposing directions for the two different strands, round the circle to generate large multigenic transcripts.

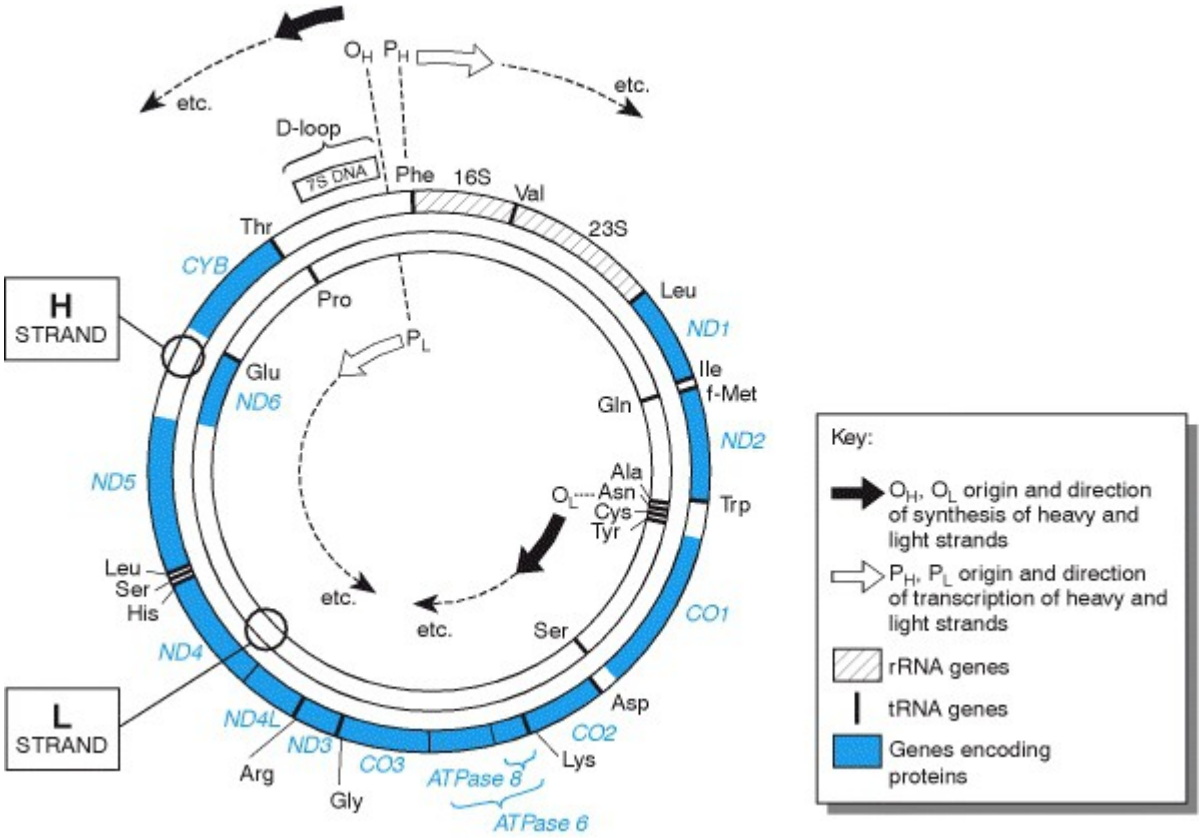


Figure 1.3

1.6 BIOLOGICAL DATABASES

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression AND phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Biological databases can be broadly classified into sequence and structure databases. Nucleic acid and protein sequences are stored in sequence databases whereas structure database only store proteins. These databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting

certain genetic diseases and in discovering basic relationships among species in the history of life.

There is huge importance of biological databases as it helps to access a huge amount of information in a nutshell. It provides various types of information to the scientific community which proves to be of great help to them.

Most biological databases are available through web sites that organise data such that users can browse through the data online. In addition the underlying data is usually available for download in a variety of formats. Biological data comes in many formats. These formats include text, sequence data, protein structure and links.

CHAPTER 2: REVIEW OF LITERATURE

Mitochondrial (mt) genomic study may reveal significant insight into the molecular evolution and several other aspects of genome evolution such as gene rearrangements evolution, gene regulation, and replication mechanisms. Other questions such as patterns of gene expression mechanism evolution, genomic variation and its correlation with physiology, and other molecular and biochemical mechanisms can be addressed by the mt genomics. Rare genomic changes have attracted evolutionary biology community for providing homoplasy free evidence of phylogenetic relationships. Gene rearrangements are considered to be rare evolutionary events and are being used to reconstruct the phylogeny of diverse group of organisms. Mt gene rearrangements have been established as a hotspot for the phylogenetic and evolutionary analysis of closely as well as distantly related organisms.

2.1 DESCRIPTION

Mitochondria are the power house of the cell. They are present in virtually every cell in body. They play an imperative role in metabolism, apoptosis, disease and aging. They are the site of oxidative phosphorylation, essential for the production of ATP, as well as for other biochemical functions. Mitochondria have a genome separate from the nuclear genome referred to as mitochondrial DNA (mtDNA). Animal mt DNA is a small (~16 kb), compact, economically-organized circular molecule, composed of 37 genes coding for 22 tRNAs, 2 rRNAs and 13 proteins, with few exceptions specifically in invertebrates The thirteen proteins are mainly involved in electron transport and oxidative phosphorylation of the mitochondria.

Recent advances in sequencing techniques have made available a great deal of data on whole genome basis. Complete mt genome sequences are available for thousands of organisms. The order of the genes in the mitochondrial DNA molecule in a wide variety of organisms has begun to be disclosed during last two decades. The gene order is highly conserved in vertebrates except for the region around the control region (D-loop), which is more prone to gene rearrangement. Maximum variability has been found in the gene order in invertebrates. The control region of the mitochondrial genome is frequently used in population studies due to the high variability in its nucleotide sequence, while protein-coding genes, such as cytochrome *b* (Cyt *b*), are generally used for phylogenetic analysis of taxa

Genome level character comparison can address a large number of evolutionary branch points. A small number of such comparisons have provided strong resolution of some evolutionary relationships which were controversial earlier. This suggests the reliability and confidence in their usage as markers. Mitochondrial gene rearrangements are considered to be rare evolutionary events. In principle, 'rare genomic changes' , such as changes in gene order, can retain phylogenetic information for long periods of time, even when primary sequence data must have become randomized due to the involvement of long time periods.

Differences in mitochondrial gene order have been useful for phylogenetic resolution of some groups of species, for example Arthropoda being monophyletic, and within this Crustacea grouping with Hexapoda to the exclusion of Myriapoda and Onychophora. The universal gene order for living vertebrates is not followed by birds. In some recent studies based on complete mitochondrial genomes in birds, it has been suggested that gene orders evolved independently more than once. Also mapping of these gene orders on the avian phylogenetic tree suggested that one of these gene orders is an ancestral while the other is a derived form. The most acceptable model for the mitochondrial gene rearrangements till date is the tandem duplication followed by the deletion

With the increasing amount of mitochondrial and nuclear sequence data there will be new outcomes for phylogenetic studies at several levels and will soon emerge as the knowledge base for molecular evolution. As the understanding of the secondary structure of the control region, non coding region and ribosomal genes and their role in the mitochondrial genome will improve, these sequences will be helpful for use in phylogeny. Several studies and debate are still in progress on the structure of tRNAs and their specific role in gene regulation and evolution. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have an insightful effect on the functions and existence of the genes in the organism. Mitochondrial gene rearrangements are likely to reveal structural, functional and evolutionary aspects of the biology distinct from those evident from homologous sequence comparisons. It will also help in the prediction and reconstruction of homoplasy free phylogenetic relationships, and will also be used to develop models for the phylogenetic analysis of remotely related organisms and to better understand the evolutionary history of organisms.

2.2 MOLECULAR PHYLOGENETICS

Molecular phylogenetics is the study of evolutionary relationships among organisms by using molecular data such as DNA and protein sequences, insertions of transposable elements, or other molecular markers. It is one of the areas of molecular evolution that have generated much interest in recent years, mainly because in many cases phylogenetic relationships are difficult to assess any other way. The objectives of phylogenetic studies are to reconstruct the correct genealogical ties among biological entities, to estimate the time of divergence between organisms (i.e., the time since they last shared a common ancestor), and to chronicle the sequence of events along evolutionary lineages

2.2.1 IMPACT OF MOLECULAR DATA ON PHYLOGENETIC STUDIES

The study of molecular phylogeny began before the turn of the century, even before Mendel's laws were rediscovered in 1900. Immunochemical studies showed that serological cross-reactions were stronger for closely related organisms than for distantly related ones. The evolutionary implications of these findings were used by Nuttall (1902, 1904) to infer the phylogenetic relationships.

Since the late 1950s, there has been extensive use of molecular data in phylogenetic research. In particular, the study of molecular phylogeny progressed tremendously in the 1960s and 1970s as a result of the development of protein-sequencing methodologies. Less expensive and more expedient methods such as protein electrophoresis, DNA-DNA hybridization, and immunological methods, though less accurate than protein sequencing, were extensively used to study the phylogenetic relationships among populations or closely related species

The rapid accumulation of DNA sequence data since the late 1970s has had a great impact on molecular phylogeny. DNA sequence data are more abundant and easier to analyze than protein sequence data.

2.2.2 ADVANTAGES OF MOLECULAR DATA IN PHYLOGENETIC STUDIES

Molecular data, particularly DNA and amino acid sequence data are much more suitable for evolutionary studies than morphological and physiological data. First, DNA and protein

sequences are strictly heritable entities. This may not be true for many morphological traits that can be influenced to varying extents by environmental factors. Second, the description of molecular characters and character states is unambiguous. Third, molecular traits generally evolve in a much more regular manner than do morphological and physiological characters and therefore can provide a clearer picture of the relationships among organisms. Fourth, molecular data are often much more amenable to quantitative treatments than are morphological data. Fifth, homology assessment is easier with molecular data than with morphological traits. Sixth, some molecular dataccan be used to assess evolutionary relationships among very distantly related organisms.

Molecular data is much more abundant than morphological data. This abundance is especially useful when working with organisms such as bacteria, algae, and protozoa, which possess only a limited number of morphological or physiological characters that can be used for phylogenetic studies.

2.3 GENETIC CODES

2.3.1 The standard genetic code is given below:

Table 2.1

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly

2.3.2 THE VERTEBRATE MITOCHONDRIAL CODE

Table 2.1

TTT F Phe	TTT F Phe	TTT F Phe	TTT F Phe
TTC F Phe	TTC F Phe	TTC F Phe	TTC F Phe
TTA L Leu	TTA L Leu	TTA L Leu	TTA L Leu
TTG L Leu	TTG L Leu	TTG L Leu	TTG L Leu
CTT L Leu	CTT L Leu	CTT L Leu	CTT L Leu
CTC L Leu	CTC L Leu	CTC L Leu	CTC L Leu
CTA L Leu	CTA L Leu	CTA L Leu	CTA L Leu
CTG L Leu	CTG L Leu	CTG L Leu	CTG L Leu
ATT I Ile i	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile i	ATC I Ile i	AAC N Asn	AGC S Ser
ATA M Met i	ACA T Thr	AAA K Lys	AGA * Ter
ATG M Met i	ACG T Thr	AAG K Lys	AGG * Ter
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val i	GCG A Ala	GAG E Glu	GGG G Gly

2.4 GENE ORDER REARRANGEMENTS

2.4.1 GENE ORDER REARRANGEMETS IN BACTERIA

Because bacteria are severely restricted in recombination and because many of their genes function as units (operons), Ochman and Wilson (1987) suggested that bacterial evolution is characterized by spatial and temporal stability of gene order. With the completion of the sequencing of the first bacterial genomes, however, it became clear that gene order in bacteria is anything but conserved (Mushegian and Koonin 1996b).

2.5 GENE ORDER AS A PHYLOGENETIC CHARACTER

Evolutionary biologists are increasingly drawn to structural features of the genome, such as gene order, as a phylogenetic marker. Unfortunately, the rapid rate of gene order rearrangements and the rate variation among evolutionary lineages essentially spells bad news for phylogenetic reconstructions based on gene order. For example, in a study of mitochondrial gene order in 137 species of birds, Mindell et al. (1998) showed that parallel evolution in gene order is quite common, and that the same gene order may arise independently several times even in closely related taxa. Because of this unsavory phylogenetic property and the difficulty in reconstructing the sequence of events leading to gene order rearrangements, phylogenetic

trees based on gene order frequently contain glaring errors. For example, by using synteny conservation and chromosome rearrangements in nine species of mammals, Ehrlich et al. (1997) inferred an absurd phylogenetic tree in which baboons are closer to humans than humans are to chimpanzees, and carnivores are paraphyletic. A more "optimistic" view on the utility of gene order in phylgenetic reconsruction.

CHAPTER 3: METHODOLOGY

The methodology carried out for the development of database includes the following:

- Data download
- Manual Curation of Data
- Conversion of data into correct format
- Circular Diagram Generation
- Development of Graphical User Interface
- Database Connectivity
- Gene Rearrangement Events

3.1 DATA DOWNLOAD

The data has been downloaded from National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The data obtained from NCBI was in raw format which had to be converted into appropriate format in order to be uploaded in our database. The original format of the data was in XML format and it was converted into .ptt format for appropriate analysis.

PTT File (Protein Table File)

It is a tab delimited file containing a list of all the proteins for their genomes. It corresponds with the CDS annotations from the GenBank file and can be created by parsing the GenBank files and writing the appropriate output.

A Protein Table File contains the following columns:

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product

3.2 MANUAL CURATION OF DATA

The downloaded data had to be manually curated to be used further for uploading in the database. This is an important step while developing the database as redundant entries leads to certain problems in the database and the extraction of data.

Data obtained from NCBI contained obsolete information. Such kind of information had to be deleted for efficient usage of the data.

3.3 CONVERSION OF DATA TO CORRECT FORMAT

The data downloaded from NCBI had to be converted to .ptt format in order to be uploaded in the database. It was carried out with the help of following program.

```
use strict;
use warnings;


my $input="C:\\mint\\study material\\MGOD\\dwnld 2504\\5834843.txt";
my $output="C:\\mint\\study material\\MGOD\\dwnld 2504\\5834843.ptt";

open (IN, "$input")|| die "can not open $output\n";
open (OUT, ">$output") || die "can not open $output\n";

my (%data, $i, $def);
### to parse data ###
$i=0;
while (my $line =<IN>) {
	if ($line =~ /  <GBSeq_definition>(.*)<\/GBSeq_definition>/) {
	$def = $1;
		print                                                                                                   OUT
"\n\n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n\n$def\n\n\nLocation\tStrand\tLength\tPID\tGene\tSynonym\tCode\t
COG\tProduct\n";
	}

	if ($line =~ /      \<GBFeature_location\>(\d+)\.\.(\d+)\<\/GBFeature_location\>/) {
		my $from=$1;
		my $to=$2;
		my $len=$to-$from;
		my $strand="+";
		my $range="$from\.\.$to";
		if (not exists $data{$range}) {
			&getdata($range);
			print OUT "$range\t$strand\t$len\t-\t$data{$range}{'gene'}\t-\t-\t-\t-\n";
		}
	}

	elsif ($line =~ /      \<GBFeature_location\>complement\((\d+)\.\.(\d+)\)\<\/GBFeature_location\>/) {
		my $from=$1;
		my $to=$2;
		my $len=$to-$from;
		my $strand="-";
		my $range="$from\.\.$to";
		if (not exists $data{$range}) {
			&getdata($range);
			print OUT "$range\t$strand\t$len\t-\t$data{$range}{'gene'}\t-\t-\t-\t-\n";
		}
```

```
        }
}


sub getdata {
        my $start=$_[0];
        while (my $nline = <IN>) {


                if ($nline =~ /         \<GBQualifier_value\>(.*)\<\/GBQualifier_value\>/) {
                        $data{$start}{'gene'}=$1;

                        return;

                }

        }

}
```

The data was manually curated to remove certain discrepancies and redundancies for its effective usefulness.

3.4 CIRCULAR DIAGRAM GENERATION

We generated the circular diagram of the genome with the help of software- Circular Genome Viewer (CGView). It is a java package for generating high quality, zoom able maps of circular genomes. Its primary purpose is to serve as a component of sequence annotation pipelines, as a means of generating visual output suitable for the web. Feature information and rendering options are supplied to the program using an XML file, a tab delimited file, or an NCBI ptt file. CGView converts the input into a graphical map (PNG, JPG, or Scalable Vector Graphics format), complete with labels, a title, legends, and footnotes. In addition to the default full view map, the program can generate a series of hyperlinked maps showing expanded views.
CGView works on command prompt and generates circular diagrams with the help of commands run on the same. The basic command is given below:

>java -jar cgview.jar -i cybercell.xml -o cybercell.png -f png
The command mentioned above excludes various parameters that can be used to change the font size, labelling, colours, etc of the image.

The command prompt window for generating the image of the mitochondrial genome is given below:

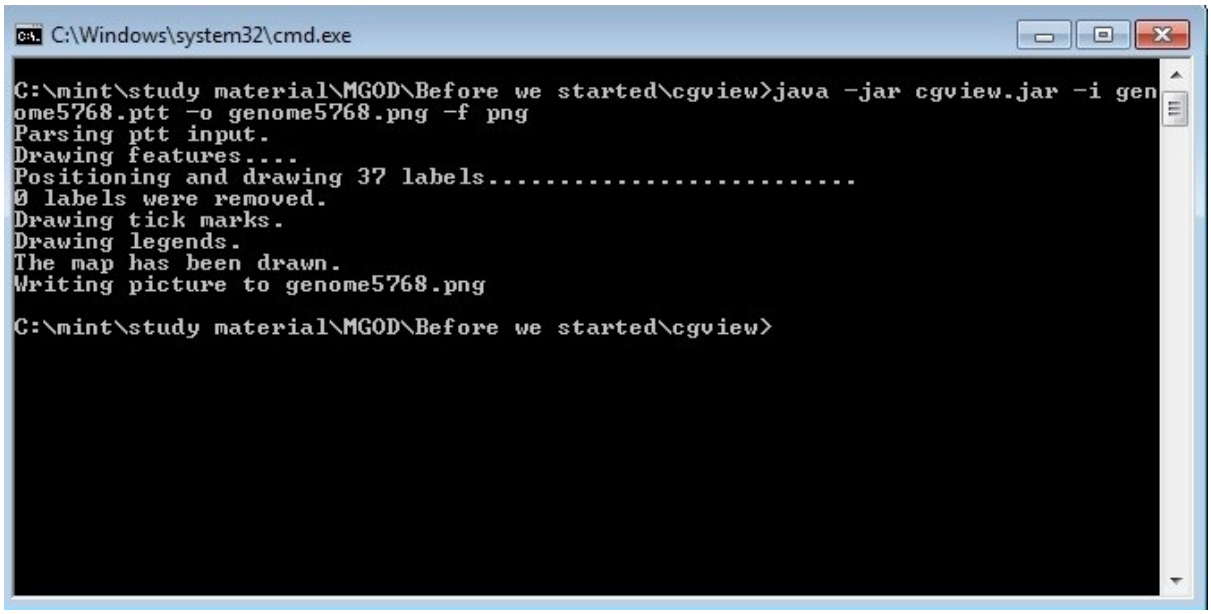


Figure 3.1

There by the generated diagram looked as depicted below-

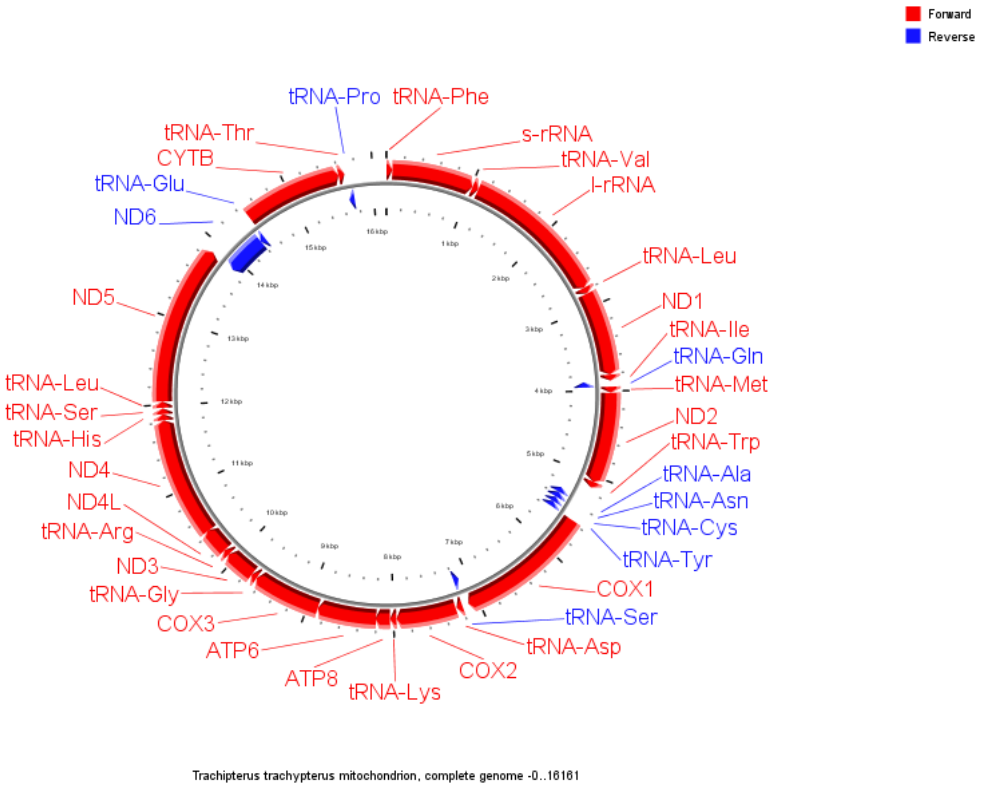


Figure 3.2

3.5 DEVELOPMENT OF GRAPHICAL USER INTERFACE

The development of Graphical User Interface required coding in HTML and connecting the same via certain codes in PHP. The website menu includes the following options:

- Home
- About
- Advanced Search Option
- Contacts

3.6 DATABASE CONNECTIVITY

In order link our Graphical User Interface and the database; we used WAMPserver as a web development environment and phpMyAdmin to manage the database. phpMyAdmin is a free and open source tool written in PHP intended to handle the administration of MySQL with the use of a web browser. It helps in various tasks such as-creating, modifying and deleting databases, tables, fields and rows; executing SQL statements; managing users and permissions. T here were certain codes developed to facilitate database connectivity-

3.6.1 For the database

```
<?php
$dbhost='localhost';
$dbuser='root';
$a=mysql_connect($dbhost,$dbuser);
if(!$a)
{
echo "connected sucessfully";
}
$sql=mysql_query('CREATE database april');
if(!$sql)
{
echo mysql_error();
}
?>
```

3.6.2 For the insertion of table

```
<?php
$dbhost='localhost';
```

```
$dbuser='root';
$a=mysql_connect($dbhost,$dbuser);
$a=fopen("table.csv","r");
$b=fgets($a);
$d=explode(',',$b);
mysql_select_db('april');
while(($c = fgetcsv($a,1000000,','))!=FALSE)
{
$sql=mysql_query("INSERT INTO basic_table1(Organism_Id,Organism,Sequence_Length,NCBI_ID,Location,Strand, Length, Gene) VALUES('$c[0]','$c[1]','$c[2]','$c[3]','$c[4]','$c[5]','$c[6]','$c[7]')");
}
if(!$sql)
{
echo mysql_error();
}
?>
```

3.6.3 For the organism table

```
<?php
$dbhost='localhost';
$dbuser='root';
$a=mysql_connect($dbhost,$dbuser);
mysql_select_db('april');
$sql=mysql_query("CREATE table basic_table1(Organism_Id varchar(100), Organism varchar(10000),Sequence_Length varchar(100),NCBI_ID varchar(100), Location varchar(200), Strand varchar(500),
 Length varchar(500),Gene varchar(2000), PRIMARY KEY (Organism_Id))");
if(!$sql)
{
echo mysql_error();
}
?>
```

3.7 GENE REARRANGEMENT EVENTS

To depict the possible rearrangement events in the genetic makeup of the organism, Cytochrome B was considered as the first gene in the genome. It was followed by the remaining genes that have been present in the organism. This shows the possible genetic rearrangement events that could have taken place in the genome of the organism. Furthermore, Cytb is supposed to be present in the genomes of all organisms. Hence, being the most

important and omnipresent gene, we considered it to be reference gene for gene rearrangement events.

CHAPTER 4: DATASET

4.1 ORIGINAL DATA

The dataset that we downloaded from the internet was obtained in XML format. It had to be converted to The Protein Table Format for effective usage.

The .ptt file is given below:

Table 4.1

Cepaea nemoralis mitochondrion, complete genome								
Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
1..14100	+	14099	-	Cepaea nemoralis	-	-	-	-
1493..1552	+	59	-	tRNA-Val	-	-	-	-
1548..2762	+	1214	-	16S ribosomal RNA	-	-	-	-
2767..2826	+	59	-	tRNA-Leu	-	-	-	-
2827..2890	+	63	-	tRNA-Ala	-	-	-	-
2884..3377	+	493	-	ND6	-	-	-	-
3378..3442	+	64	-	tRNA-Pro	-	-	-	-
3444..5129	+	1685	-	ND5	-	-	-	-
5138..6020	+	882	-	ND1	-	-	-	-
6024..6261	+	237	-	ND4L	-	-	-	-
6260..7382	+	1122	-	CYTB	-	-	-	-
7383..7438	+	55	-	tRNA-Asp	-	-	-	-
7439..7499	+	60	-	tRNA-Cys	-	-	-	-
7501..7561	+	60	-	tRNA-Phe	-	-	-	-
7561..8214	+	653	-	COX2	-	-	-	-
8207..8266	+	59	-	tRNA-Tyr	-	-	-	-
8260..8319	+	59	-	tRNA-Trp	-	-	-	-
8320..8380	+	60	-	tRNA-Gly	-	-	-	-
8377..8434	+	57	-	tRNA-His	-	-	-	-
8435..8491	-	56	-	tRNA-Gln	-	-	-	-
8489..8549	-	60	-	tRNA-Leu	-	-	-	-
8543..8704	-	161	-	ATP8	-	-	-	-
8705..8767	-	62	-	tRNA-Asn	-	-	-	-
8789..9347	-	558	-	ATP6	-	-	-	-
9388..9448	-	60	-	tRNA-Arg	-	-	-	-
9445..9508	-	63	-	tRNA-Glu	-	-	-	-
9515..10224	-	709	-	12S ribosomal RNA	-	-	-	-
10198..10259	-	61	-	tRNA-Met	-	-	-	-
10260..10638	-	378	-	ND3	-	-	-	-
10260..10602	-	342	-	ND3	-	-	-	-
10604..10652	-	48	-	tRNA-Ser	-	-	-	-
10653..10712	-	59	-	tRNA-Thr	-	-	-	-

10713..11526	-	813	-	COX3	-	-	-	-
11685..11744	+	59	-	tRNA-Ser	-	-	-	-
11775..13026	+	1251	-	ND4	-	-	-	-
13031..13090	+	59	-	tRNA-Ile	-	-	-	-
13116..14013	+	897	-	ND2	-	-	-	-
14013..14075	+	62	-	tRNA-Lys	-	-	-	-

4.2 DATA FOR UPLOADING IN THE DATABASE

Once the data had been curated, an excel file had been generated to capture the required fields. The excel file contained the following headings-

- Organism ID
- Organism Name
- Length of The Genome
- NCBI ID
- Location of a particular gene
- Length Range
- Gene Name

The excel file as uploaded in the database is given below –

Table 4.2

Organism_Id	Organism Name	Length	NCBI_ID	Location	Length	Gene
atm16692	Acipenser transmontanus	166931	30065662	1..68	67	tRNA-Phe
				69..1029	960	s-rRNA
				1030..1100	70	tRNA-Val
				1101..2801	1700	l-rRNA
				2350..2512	162	PMC31832P1
				2802..2876	74	tRNA-Leu
				2877..3851	974	ND1
				3861..3931	70	tRNA-Ile
				3931..4001	70	tRNA-Gln
				4001..4070	69	tRNA-Met
				4071..5115	1044	ND2
				5116..5188	72	tRNA-Trp
				5190..5258	68	tRNA-Ala
				5260..5332	72	tRNA-Asn
				5365..5431	66	tRNA-Cys
				5432..5502	70	tRNA-Tyr
				5504..7057	1553	COX1
				7065..7135	70	tRNA-Ser
				7144..7215	71	tRNA-Asp
				7230..7920	690	COX2
				7921..7994	73	tRNA-Lys
				7996..8163	167	ATP8
				8154..8836	682	ATP6

				8837..9621	784	COX3
				9622..9694	72	tRNA-Gly
				9695..10043	348	ND3
				10044..10113	69	tRNA-Arg
				10114..10410	296	ND4L
				10404..11784	1380	ND4
				11785..11853	68	tRNA-His
				11854..11909	55	tRNA-Ser
				11922..11994	72	tRNA-Leu
				11995..13836	1841	ND5
				13833..14354	521	ND6
				14355..14424	69	tRNA-Glu
				14427..15567	1140	CYTB
				15568..15640	72	tRNA-Thr
				15644..15713	69	tRNA-Pro
				15714..16692	978	

4.3 GENE REARRANGEMENT

The data considered for gene rearranged events is given below:

Table 4.3

Organism_Id	Organism Name	NCBI_ID	Gene Original	Gene Rearrangement
amm17239	Achalinus meiguensis	212725550	tRNA-Phe s-rRNA tRNA-Val l-rRNA ND1 tRNA-Ile control region 2; tRNA-Leu tRNA-Gln tRNA-Met ND2 tRNA-Trp tRNA-Ala tRNA-Asn putative light- strand replication origin tRNA-Cys tRNA-Tyr COX1 tRNA-Ser tRNA-Asp COX2 tRNA-Lys ATP8 ATP6 COX3 tRNA-Gly ND3 tRNA-Arg ND4L ND4 tRNA-His tRNA-Ser tRNA-Leu ND5 ND6 tRNA-Glu CYTB tRNA-Thr tRNA-Pro control region 1;	CYTB tRNA-Thr tRNA-Pro tRNA-Phe s-rRNA tRNA-Val l-rRNA ND1 tRNA-Ile tRNA-Leu tRNA-Gln tRNA-Met ND2 tRNA-Trp tRNA-Ala tRNA-Asn tRNA-Cys tRNA-Tyr COX1 tRNA-Ser tRNA-Asp COX2 tRNA-Lys ATP8 ATP6 COX3 tRNA-Gly ND3 tRNA-Arg ND4L ND4 tRNA-His tRNA-Ser tRNA-Leu ND5 ND6 tRNA-Glu

CHAPTER 5: RESULTS

A view of the Graphical User Interface is given below –

5.1 MENU OPTIONS

HOME PAGE

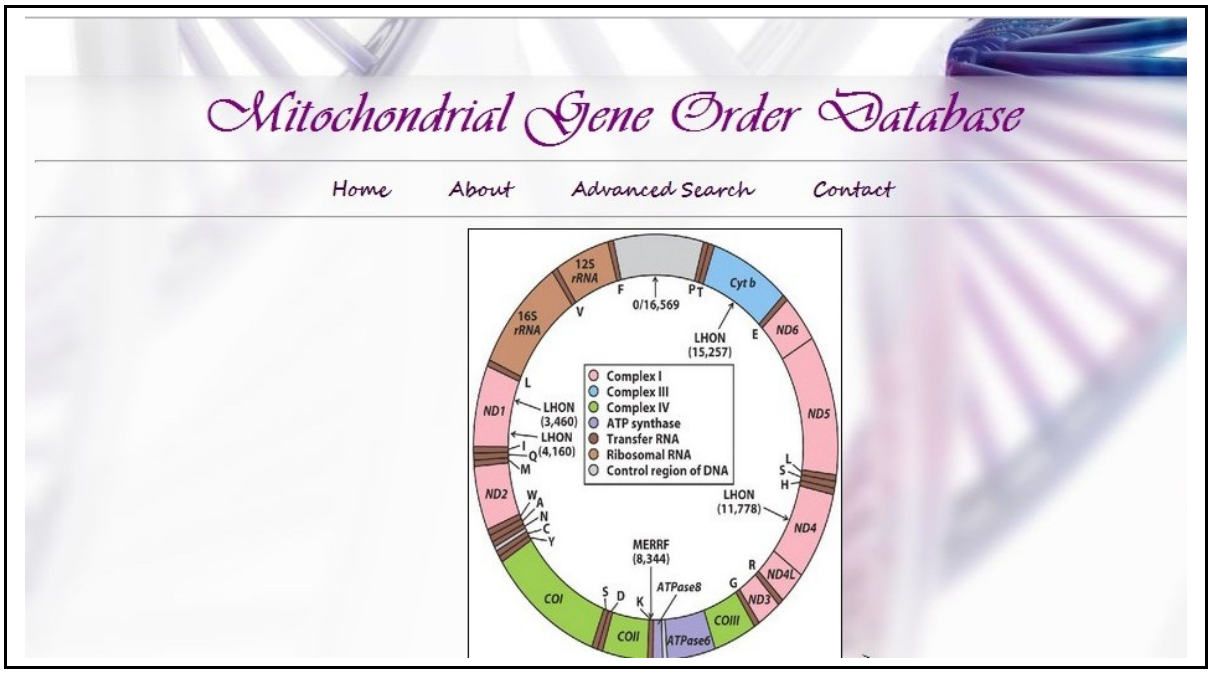


Figure 4.1

ABOUT

This page consists of the details with respect to the database. It also links various other pages that gives various information regarding the mitochondrial genome. These links carry different types of information regarding mitochondrial genome.

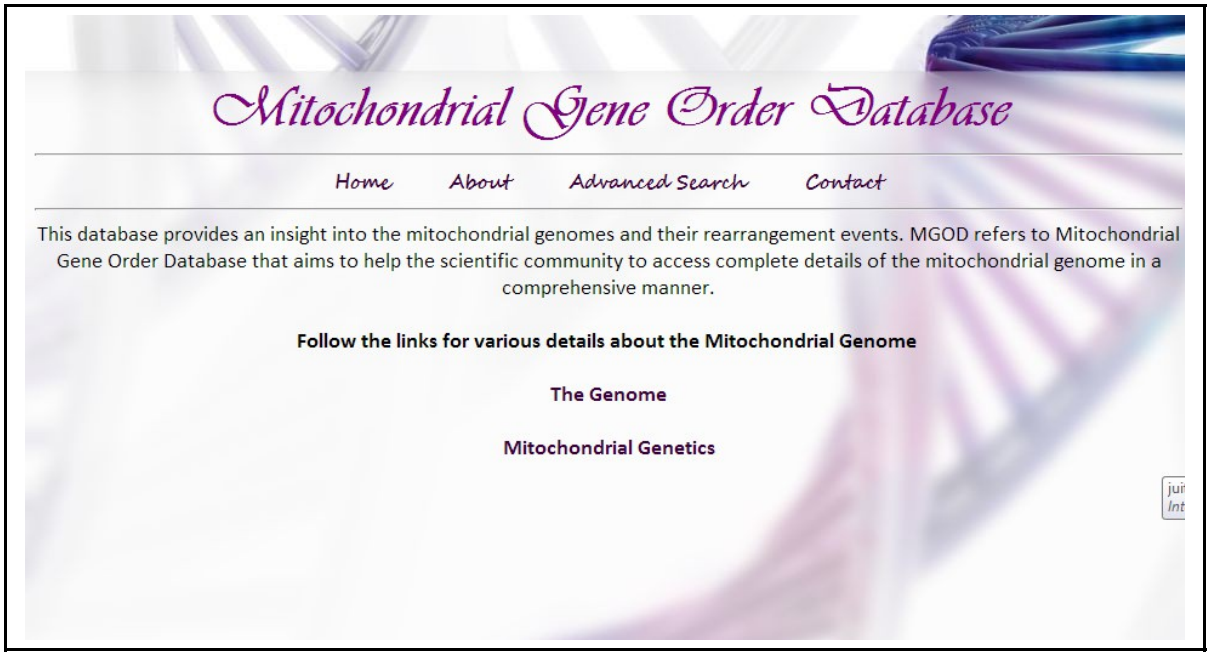


Figure 4.2

ADVANCED SEARCH OPTION

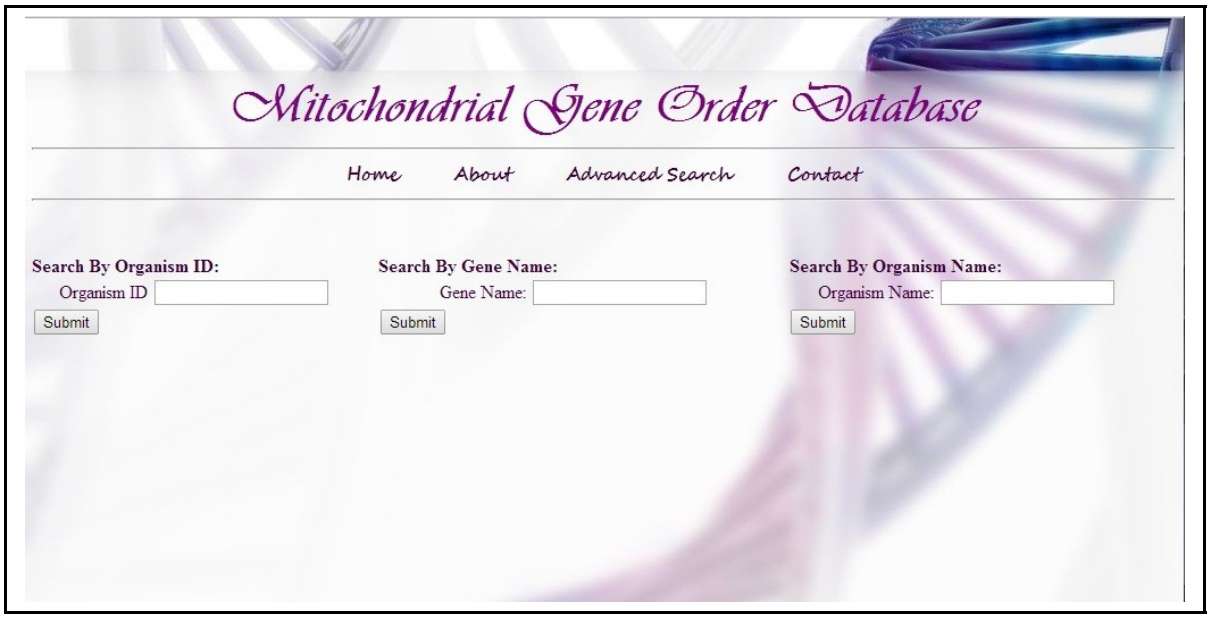


Figure 4.3

This page helps in refining the result by Organism ID, Organism Name and Gene Name.

CONTACTS

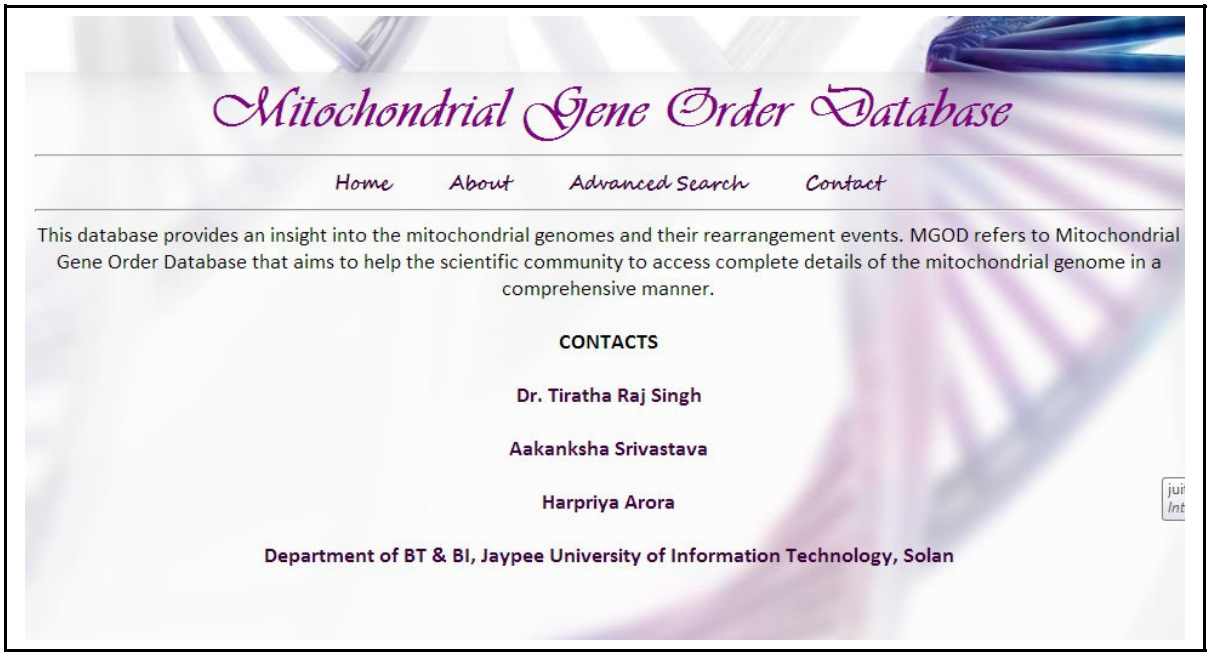


Figure 4.4

CHAPTER 6: CONCLUSION

Mitochondrial Gene Order Database (MGOD) is a public database that stores curated information of various genomes and their possible gene rearrangement events. This database provides information about the organism name, its sequence length, genes with respect to sequence positions and circular diagram of the genome. The database is available at www.bioinfoindia.org/mgod.

The Graphical User Interface gives the Search Option in order to search for the required entity in the database. The visual representation of the Graphical User Interface is shown in the results. Various details of the organism is displayed on the page once the user enters his desired input. The gene rearrangement events give the changes that could have taken place in the genome of the organism. This gene rearrangement event has been compared with the original genome structure of the organism thereby helping in efficient comparison of both the genes.

REFERENCES

1. B. Franz lang, michael w. Gray, and gertraud burger (1999). "Mitochondrial genome evolution and the origin of Eukaryotes", Annual Review of Genetics, Vol. 33: 351-397

2. Tiratha Raj Singh (2008), "Mitochondrial gene rearrangements: new paradigm in the evolutionary biology and systematics", Bioinformatician, 2008; 3(2):95-97

3. Francisco J Iborra, Hiroshi Kimura and Peter R Cook, "The Functional Organisation of Mitochondrial genomes in human cells", BMC Biology 2004, 2:9

4. "Mitochondrial DNA: The Eve Gene". Bradshaw Foundation. Retreieved 5 November 2012.

5. Penman, Danny(2002), "Mitochondrial can be inherited from both parents". NewScientist.com. Retrieved 2008-02-05.

6. Schwartz M, Vissing J (2002). "Paternal inheritance of mitochondrial DNA". N. Engl. J. Med. 347 (8): 576–80.

7. Sykes, B (2003). "Mitochondrial DNA and human history". The Human Genome. Wellcome Trust. Retrieved 5 February 2012.

8. Iborra FJ, Kimura H, Cook PR (2004). "The functional organization of mitochondrial genomes in human cells". BMC Biol. 2: 9.

9. Wallace, Douglas C. "Mitochondrial genetics: a paradigm for aging and degenerative diseases?" Science 256.5057 (1992): 628-632.

10. Coen, D., et al. "Mitochondrial genetics. I. Methodology and phenomenology." Symposia of the Society for Experimental Biology. Vol. 24. 1969.

11. Avner, P. R., et al. "Mitochondrial genetics." Molecular and General Genetics MGG 125.1 (1973): 9-52.

12. Chinnery, Patrick F., et al. "Clinical mitochondrial genetics." Journal of medical genetics 36.6 (1999): 425-436.

13. Shoffner, John M., and Douglas C. Wallace. "Mitochondrial genetics: principles and practice." American journal of human genetics 51.6 (1992): 1179.

14. Michaelis, Georg, Eric Petrochilo, and Piotr P. Slonimski. "Mitochondrial genetics." Molecular and General Genetics MGG 123.1 (1973): 51-65.

15. Dujon, B., et al. "Mitochondrial genetics." Molecular and General Genetics MGG 143.2 (1976): 131-165.

16. DiMauro, Salvatore, and Eric A. Schon. "Mitochondrial DNA mutations in human disease." American journal of medical genetics 106.1 (2001): 18-26.

17. Linnane, AnthonyW, et al. "Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases." The Lancet 333.8639 (1989): 642-645.

18. Walker, A. C., S. D. Smith, and S. D. Smith. "Mitochondrial DNA and human evolution." Nature 325 (1987): 1-5.

19. Ingman, Max, et al. "Mitochondrial genome variation and the origin of modern humans." Nature 408.6813 (2000): 708-713.

20. Taanman, Jan-Willem. "The mitochondrial genome: structure, transcription, translation and replication." Biochimica et Biophysica Acta (BBA)-Bioenergetics 1410.2 (1999): 103-123.

21. Lang, B. Franz, Michael W. Gray, and Gertraud Burger. "Mitochondrial genome evolution and the origin of eukaryotes." Annual review of genetics 33.1 (1999): 351-397.
22. Macey, J. Robert, et al. "Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome." Molecular Biology and Evolution 14.1 (1997): 91-104.
23. Sankoff, David, et al. "Early eukaryote evolution based on mitochondrial gene order breakpoints." Journal of Computational Biology 7.3-4 (2000): 521-535.