

2D-Representation of Gene Sequence Using Binary Codes in Spiral Conformation



**CAPSTONE PROJECT (BME499)
School of Science and Technology**

**A final year project report presented
in partial fulfillment
of the requirements for the
Bachelor of Science (Hons) in Biomedical Engineering**

**BY
LIM JIXIANG
N0401510
SIM UNIVERSITY 2007**

TABLE OF CONTENT

1 - Abstract

2 – Introduction

2.1 - DNA The Building Blocks of Life

2.2 - DNA Sequencing

2.3 - DNA Comparison methods

2.4 – Direct Linear DNA Comparison

2.5 - Visual Comparison of Sequences: The Dot Plot

2.6 - Visual Comparison of Sequences: The U-lam Spiral

2.7 – Ulam Spiral and it's Role in Previous Gene Sequencing Techniques

2.8 - Graphical Representation of Gene Sequences

2.9 - Aims of the Explored Method

3. Methodology

3.1 - Concept

3.2 - DNA Sample Used

3.3 – Summary of Methodology

3.4 – Graphical Representation Graphs Methodology

3.4.1 – The ‘X.1. 1’ Graphs

3.4.2 – The ‘X.1.2’ Graphs

3.5 - Comparison Plot Methodology

3.5.1 – The ‘X.3.1’ and ‘X.3.2’ Comparison Graphs

3.6 - Comparison of Animal Sequences

3.7 - Similarity and Dissimilarity on Comparison Plots

3.8 - List of Combinations Used

3.9 - Determining the best BB/ BW/ WW/ WB combination for Individual Clusters

3.10 - Producing Graphical Comparison Plots (Using the Best BB/ BW/ WW/ WB Combination Determined from Clusters 1 and 2)

4 - Results

4.1 - Labelling of Graphs

4.2 – List of Graphs for Cluster 1

4.3 - Results for Cluster 1: Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee Set

4.4 – List of Graphs for Cluster 2

4.5 - Results for Cluster 2: Bovine – Goat Set

4.6 – List of Graphs for Cluster 3

4.7 - Results for Cluster 3: Human-Animal Set

5 – Interpretation of Results

5.1 - Best BB/ BW/ WW/ WB combination for Clusters 1 and 2

5.2- Results of Comparison of Human Gene Sequence with 10 Other Animals Using ‘the Best BB/ BW/ WW/ WB combination for Clusters 1 and 2: Combination 1a’

5.3 - Specific Advantages of Explored Method

5.4 - Specific Disadvantages of Explored Method

6 – Conclusion

7 - Further Research Suggestions

8 - Critical Review

9 - References

10 – Graphs

10.1 - Graphs for Cluster 1 – Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee Comparison (Graphs 1-12)

10.1.1 Graphs for BW/WB/WW/BB Combination 1

10.1.2 - Graphs for BW/WB/WW/BB Combination 1a

10.1.3 - Graphs for BW/WB/WW/BB Combination 2

10.1.4 - Graphs for BW/WB/WW/BB Combination 3

10.2 - Graphs for Cluster 2 – Goat – Bovine Comparison (Graphs 13-16)

10.3 - Graphs for Cluster 3 – Human–Animal Comparison (Graphs 17- 24)

11- Appendix

11.1 – Definition of Terms

11.2 – Gantt Progress Chart

12 - Notes

LIST OF FIGURES USED

Fig 1 - The four bases found of DNA

Fig 2 – An example of BLAST’s linear comparison of a pair of DNA sequences

Fig 3 - A DNA dot plot of a certain transcription factor is shown above

Fig 4 –The Ulam Spiral

Fig 5 - Four-color map representation of DNA sequences by Randic, Nella, Dejan, Basak and Balaban

Fig 6 – The Ulam Spiral Diagram to be used in the project.

Fig 7 – Black and white graphical representation of a particular DNA sequence after plotting is done.

Fig 8 – Excel Spreadsheet with the yellow box representing the starting point of the plot.

Fig 9i– The ‘X.1.1’ black and white plots after re-plotting is done

Fig 9ii – The ‘X.1.2’ surface graph for sequences

Fig 10 – Typical ‘X.3.1’, ‘X.3.1’ comparison graph before plotting

Fig 11 – The ‘X.3.1’ comparison graph after plotting

Fig 12 – A typical ‘ X.3.2 graph ’

Fig 13 – A typical ‘X.3.2 graph ’

Fig 14 –A plot showing the number of crosses in animals - human sequence comparison using the 1a combination.

Fig 15 - One method of numerical characterization

LIST OF TABLES USED

Tables 1 - The first exon of α -globin genes belonging to eight species (including human), which differ in length from 86 to 93 bases.

Table 3i - Table of results for Cluster 1- Human-Gorilla, Gorilla-Chimpanzee and Human- Chimpanzee set.

Table 3ii – Table of results for Cluster 2 – Bovine- Goat set.

Table 3iii - Table of results for Cluster 3 – Human-Animal set

Acknowledgement

I would like to sincerely thank Dr Lim Teik Cheng, my Project Supervisor for his invaluable guidance and advice in the entire course of this project. Without Dr Lim's foresight and expertise, this project would not have been a success.

I would also like to thank the University for assistance to enhance our understanding of what is required of us in this project.

1 - Abstract

Graphical techniques have emerged as a very powerful tool for the visualization and analysis of biological. Besides enabling storage of data, 2-D graphical representations have a great advantage: it enables convenient visual inspection of data and therefore aid the discrepancy recognition of almost-similar genetic code sequences.

The main aim of this project is to introduce a method by adopting the “blanks-filling” approach rather than the “path-walking approach”. We hope it will allows the physical mapping of gene sequences, so as to provide a type of 2-D graphical representation of DNA (and RNA) using binary codes such as 0-0, 0-1, 1-0 and 1-1, or other combinations (e.g. black and white cubes) to represent Adenine (*A*), Cytosine (*C*), Guanine (*G*) and Thymine (*T*).

One extension of these 2-D graphical representations will be to compare 2 of these graphs so that a third comparison graph can be plotted so as to investigate and determine (quantitatively), how similar or dissimilar the human DNA sequence from other animals.

2 - Introduction

2.1 – The Background of the Project

DNA is the basic building block of life. Every cell in an organism has a set of chromosomes containing the gene, which in turn contains genetic material - the DNA molecule.

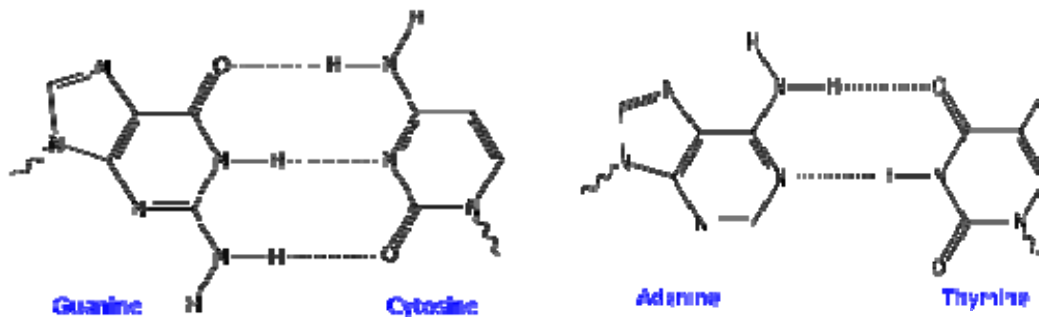


Fig 1 - Adenine/Thymine Watson and Crick base pair. This figure shows the four bases of DNA, and how they match with one another (13)

The four bases found in DNA are adenine (abbreviated *A*), cytosine (*C*), guanine (*G*) and thymine (*T*). Each type on one strand forms a bond with just one type of base on the other strand. In DNA, *A* bonds only to *T*, and *C* bonds only to *G*.

Each strand will consist of a particular arrangement of *A*, *T*, *G* and *C*, and this The sequence basically encodes the heritable genetic information that forms the basis in which living cells conduct themselves for the survival of all living things. Study of sequences is therefore useful in the understanding of biological processes as well as diagnostic and forensic research.

2.2 - DNA Sequencing

DNA sequencing are biochemical methods used to determine the order of nucleotide bases adenine, guanine, cytosine and thymine. Modern technology has been vital in enabling large-scale sequencing of the human genome in the Human Genome Project, or HGP. The HGP has one noble aim: to correctly identify and map out the physical chromosomal location of every single human gene and more importantly, determine the exact chemical sequence of each gene. Animal and plant DNA have also been extensively plotted, and studied to find out its link to diseases and how the processes actually take place.

Unfortunately, research on DNA sequences do not stop at successful plotting of exact chemical sequences; Project like these will inevitably involve mathematical (and visual) analysis of large volume genomic DNA sequence. One method we are interested in is the comparing of DNA strands.

2.3– DNA Comparison Methods

Graphical representation of data is one way to make these jobs easier. Therefore, the more straightforward the characterization is, the better and easier the method. Indeed, graphical representation of pictorial data can be transformed into digital format, which allows computer storing, search, and processing of such data with a lot of convenience and ease.

2.4 – Direct Linear DNA Comparison

An alignment program compares the sequence homology between DNA sequences. The BLAST (Basic Local Alignment Search Tool) from the National Centre for Biotechnology Information basically determines the best match (based on database) between the two sequences submitted, as shown in Fig 2.

```
seq1 > 1 ggccctctgcctaatacacacagat-ctaacaggattatttc
          ||||| ||||| || ||||| || ||||| |||||
seq2 > 1 ggccctctgccttattacacaaatcttaacaggactatttc
```

Fig 2 – An example of BLAST’s linear comparison of a pair of DNA sequences (6)

Occasionally gaps need to be introduced to make the two sequences align. Such analysis helps in the studying of evolutionary links between related species. It can also be used to look for mutations in genes. (6)

For database searches such as BLAST, statistical methods are used to determine the possibilities of specific alignment between sequences or sequence segments arising by chance. This largely depends on the size and composition of the database being searched. These values can vary significantly depending on the search space and the available data.

In particular, the likelihood of finding a given alignment by chance increases if the database consists only of sequences from the same organism as the query sequence. Repetitive sequences in the database or query can adversely affect both the search results and the assessment of statistical significance. Hence, one of BLAST’s abilities is to automatically filter such repetitive sequences in the search to avoid apparent hits that are statistical artifacts.

2.5 - Visual Comparison of Sequences: The Dot Plot

Besides methods that use direct linear comparison, methods like ours can be visual in nature, like the dot matrix diagram shown below.

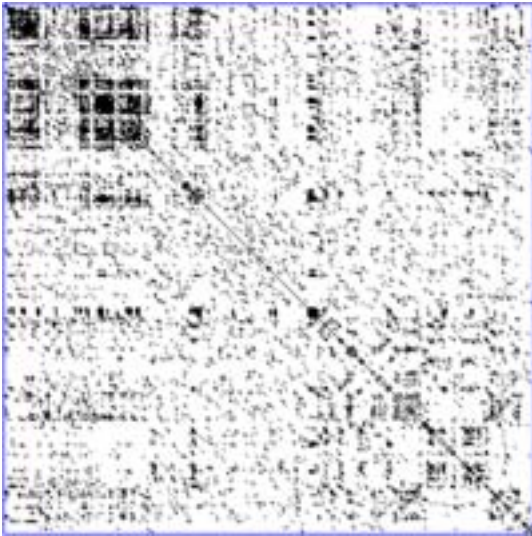


Fig 3 - A DNA dot plot of a certain transcription factor is shown above (14)

Figure 3 shows the dot-matrix approach, which implicitly produces a series of alignments for individual sequence regions. Though it can be tedious to analyze it on a large scale, this diagram is simple to observe. It is easy to visually identify certain sequence features—such as insertions, deletions, repeats, etc from such a dot-matrix plot.

Some implementations can vary the size or intensity of the dot depending on the degree of similarity of the two characters, thus being able to accommodate conservative substitutions. Moreover, the dot plots of extremely close-related sequences will appear as a single line along the matrix's main diagonal.

Dot plots can also be used to assess repetitiveness in a single sequence. An example is when a protein consists of multiple similar structural domains

2.6 - Visual Comparison of Sequences: The U-lam Spiral

Why did we use the Ulam spiral for our project? Firstly let's look at what the Ulam spiral is.

Ulam spiral, better known as the prime spiral is a simple graphical representation of prime numbers in a spiral. It basically consists of a grid of numbers, starting with 1 at the center, and spiralling outwards (fig. 1). One will note that non-random patterns will often appear, sometime in form of diagonal lines no matter how many numbers are plotted.

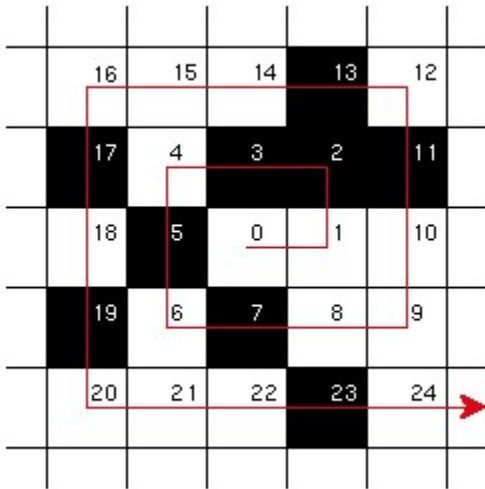


Fig 4 –The Ulam Spiral

All prime numbers (except 2) are odd numbers and will often end with 1, 3, 7, or 9 (except for 2 and 5). As adjacent diagonals are made up of odd and even numbers alternating with one another, all prime numbers, being odd numbers, lie in alternate diagonals of the Ulam spiral. Indeed, it has been observed that there is a tendency of prime numbers to lie on some diagonals more than others while a random distribution is expected. What is interesting, too, is that when viewed at a distance from the centre, horizontal and vertical lines will also be clearly visible.

The Ulam spiral has some pretty intriguing properties. Many mathematicians have observed that somehow, the prime numbers tend to cluster along diagonal lines, thereby suggesting the non-random quality of prime numbers through such 2-D representation of prime numbers.

2.7 – Ulam Spiral and it’s Role in Previous Gene Sequencing Techniques

In 2005, a group of researchers filled up a grid with genetic code sequences starting from the middle and spiralling outward. By incorporating corresponding colours to the nucleotide base, a 4-color map was obtained (in Fig. 3). As expected, the use of colours has revealed a Ulam-spiral observation.

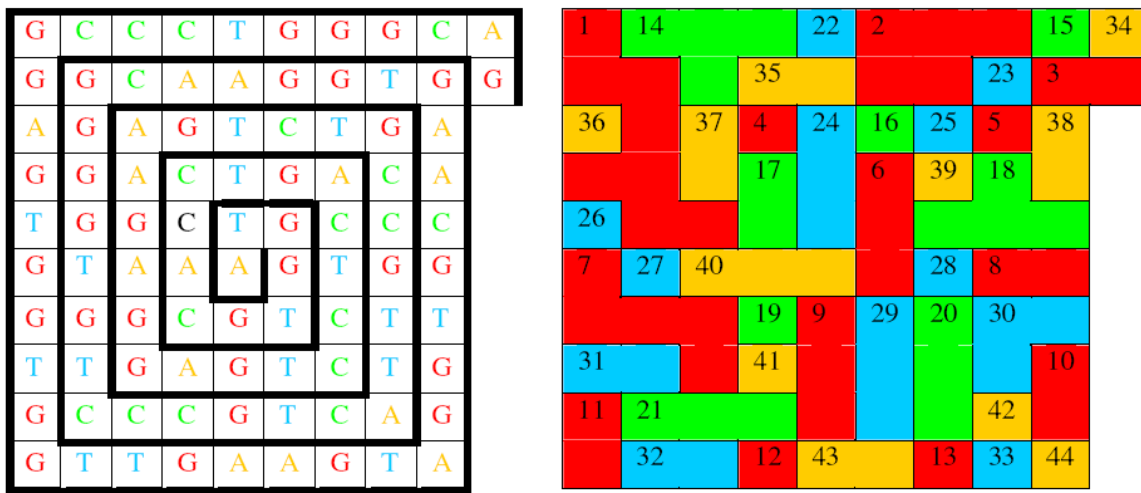


Fig 5 - Four-color map representation of DNA sequences by Randic, Nella, Dejan, Basak and Balaban (11)

The first exon in the beta-globin for human was previously plotted by others. Each nucleotide is given a unique colour. As shown above, the grid is filled up with genetic code sequences starting from the middle and spiralling outward (11).

Some of these methods can be as straight forward as manual plotting, like BLAST. However in order to demonstrate the patterns visually, scientists use complex 2D, 3D, 4D mapping that involve complex mathematical formulas.

Indeed, many 2D, 3D, and even 4D graphical representations of DNA sequences have been outlined by scientists. (12)

2.8 - Graphical Representation of Gene Sequences

Graphical representation of DNA sequence have long been instrumental in providing a simplified way of viewing, sorting and comparing anything from protein (like Feng-lan Bai, Ying-zhao Liu and Tian-ming Wang representation of proteins by star-like graphs) to DNA sequences.(9) Besides BLAST which compares sequences linearly, we can also compare sequences using graphical methods.

2.9 – The Aims of Explored Method

The aims of this project is as follows:

- (1) Introducing a graphical representation method for DNA sequences. This is done by devising an EXCEL spreadsheet-based program which characterizes the sequences by carrying out physical mapping of gene sequences using binary black and white codes to represent the 4 basic components of DNA. This program will then plot of 2D graphical representation of the sample nucleotide bases.
- (2) To investigate and determine (quantitatively) how similar human DNA sequence is compared to other mammals.
- (3) To establish a method in which will give the most recognizable pattern for most genetic code sequences over different species.

3 - Methodology

3.1 - Concept

Currently, there exist many ways that enable visual mapping of the DNA sequences.

The main concept behind the method we proposed is build upon previous methods devised to study DNA sequencing. Conceptually it is similar to BLAST, but we chose to represent the sequences two-dimensionally in an Ulam spiral like configuration instead of linearly like BLAST.

The major part of this project will be to physically map gene sequences of various animals using numerical codes like 0 1 and -1 to represent the 4 basic components of DNA.

As an extension of this, we shall aim to establish a method in which will give the most recognizable pattern for most genetic code sequences over different species, and to compare the spiral plots to investigate and determine (quantitatively) how similar human DNA sequence is compared to other mammals.

3.2 - DNA Sample Used

We shall use the first exon of α -globin genes shown in Table 1 to form the basis of our project.

Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT TACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGG CTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTG CTGAGGCCCTGGGCAG
Opossum	ATGGTGCACCTGACTTCTGAGGAGAAGAACTGCAT CACTACCATCTGGTCTAAGGTGCAGGTTGACCAGA CTGGTGGTGAAGGCCCTTGGCAG
Gallus	ATGGTGCACCTGGACTGCTGAGGAGAAGCAGCTCAT CACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAAT GTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGT CACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAG TTGGTGGCGAGGCCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTG TCTCTTGCCTGTGGGGCAAAGGTGAACCCCGATGAA GTTGGTGGTGAAGGCCCTGGGCAGG

Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGT CACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAG TTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTT AGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGT TGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT TACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGC CTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTG GTGAGGCCCTGGGCAG
Chimpan -zee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT TACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

Table 1: The first exon of α -globin genes belonging to eight species (including human), which differ in length from 86 to 93 bases.

Previous methods have often used this same portion of DNA. By selecting the same portion for our project, we hope to be able to compare our method with the other methods. The DNA samples are obtained from the National Centre for Biotechnology Information, Washington, D.C.).

3.3 –Summary of Methodology

Based on a specific combination of binary codes, DNA sequences will then be mapped out by one after another in using an EXCEL spreadsheet.

Instead of the 4 colours used in the previous research Randic, Nella, Dejan, Basak and Balaban (as illustrated in Fig 2), we will instead use binary codes to represent Adenine (*A*), Cytosine (*C*), Guanine (*G*) and Thymine (*T*) on graph paper> These codes will then be plotted on the EXCEL spreadsheet to allow the plotting of comparison graphs.

Because of the way these codes are being plotted from centre outwards, a Ulam spiral will emerge.

3.4 – Graphical Representation Graphs Methodology

3.4.1 – The X.1. 1' Graphs

1. We assign 4 binary codes white white, white black, black white and black black (to be referred to as WW/ WB/ BW/ BB from this point onwards) for each of the 4 nucleotides *A*, *T*, *C* and *G*. In the first Graph Set, Graph Set 1: *A* = **white white**, *C* = **white black**, *G* = **black white** and *T* = **black black**.
2. We plot each nucleotide on the Ulam spiral diagram starting from the centre in the middle, and continue outwards in a clockwise fashion. The plotting of each nucleotide is done one after another, in accordance to the DNA sequence given for the subject. We shall use 2 consecutive cells in the Ulam Spiral diagram in Figure 6 to represent EACH nucleotide. All these plotting are done on graph paper.

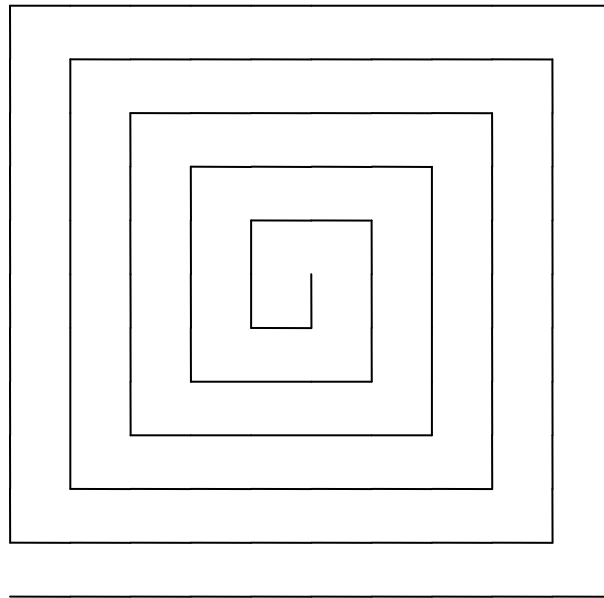
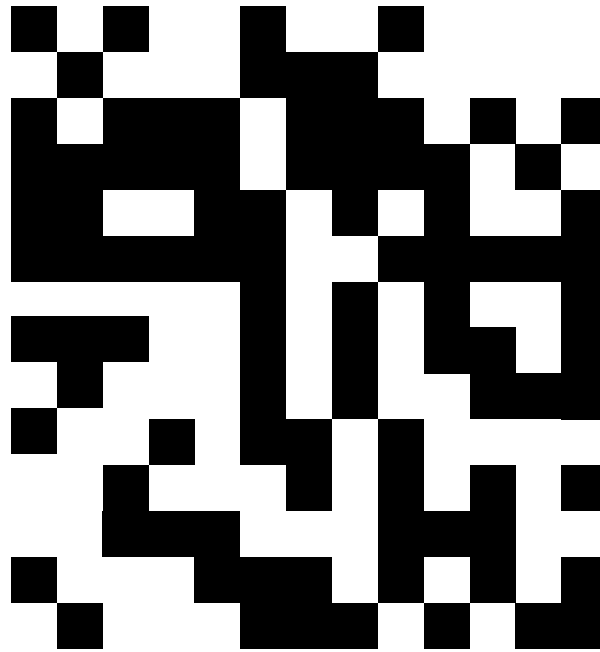


Fig 6 – The Ulam Spiral Diagram to be used in the project

3. After this plot is completed, the Ulam Spiral diagram consists only of black and white cubes on the graph paper, which look like what is shown in Fig 7.



where

A = white white

C = white black

G = black white

T = black black

Fig 7 – Black and white graphical representation of a particular DNA sequence after plotting is done. The Ulam Spiral is then removed for easy viewing.

4. These black/white cells on the graph paper are replaced by numerical digits ‘1’ or ‘0’ and transferred onto the Excel spreadsheet as shown on Fig 8. The digit ‘1’ represents the black squares, and ‘0’ represents the white squares. In the end, we will have a

diagram that looks like Fig 9i. We call these the 'X.1.1' Graphs where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc).

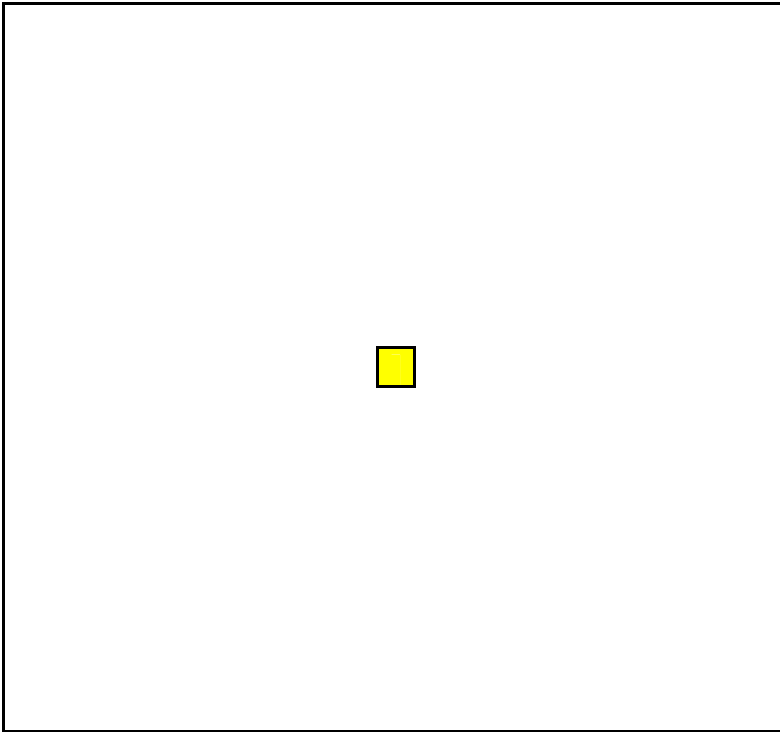


Fig 8 – Excel Spreadsheet with the yellow box representing the starting point of the plot.

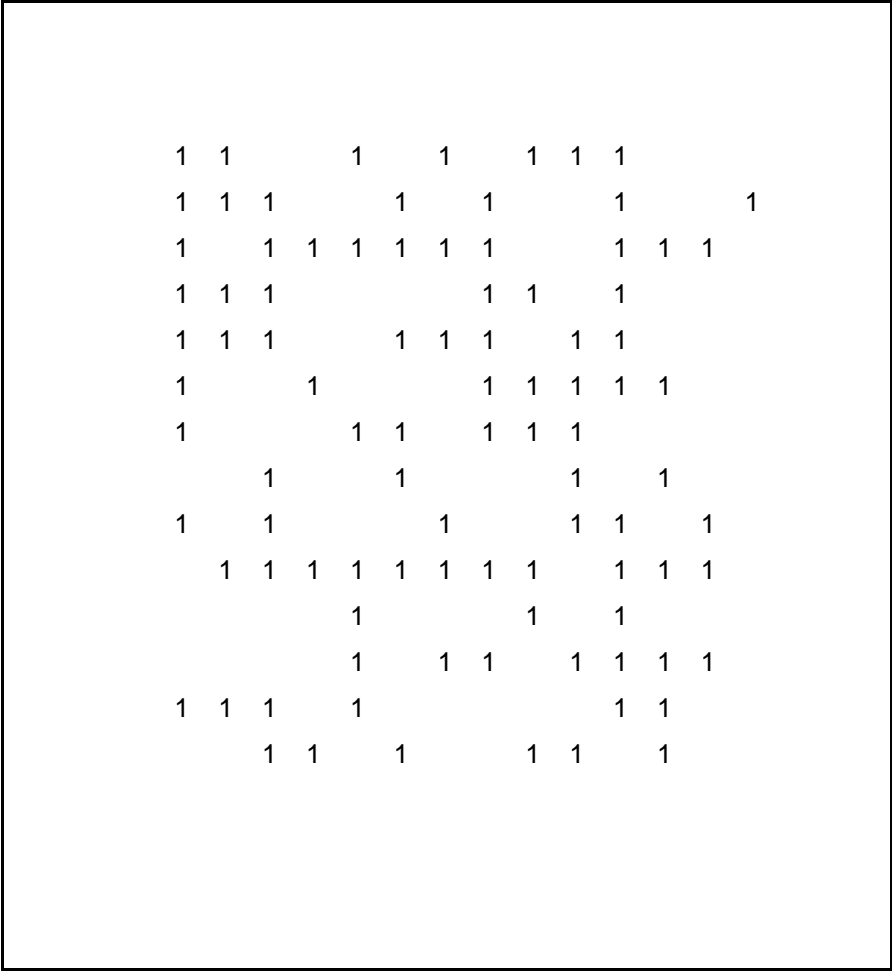


Fig 9i– The 'X.1.1' Graphs where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc). The black and white plots from Figure 7 will look like the above shown when re-plotting is done using '1' to represent black cubes, and '0' to represent white cubes. Note that the '0's in Figure 9i have been omitted for clarity.

3.4.2 – The ‘X.1.2’ Graphs

5. Surface graphs for the ‘X.1.1’ graphs will then be plotted. This will produce 2 surface graphs that resemble Figure 11ii. We call them the ‘X.1.2’ graphs.

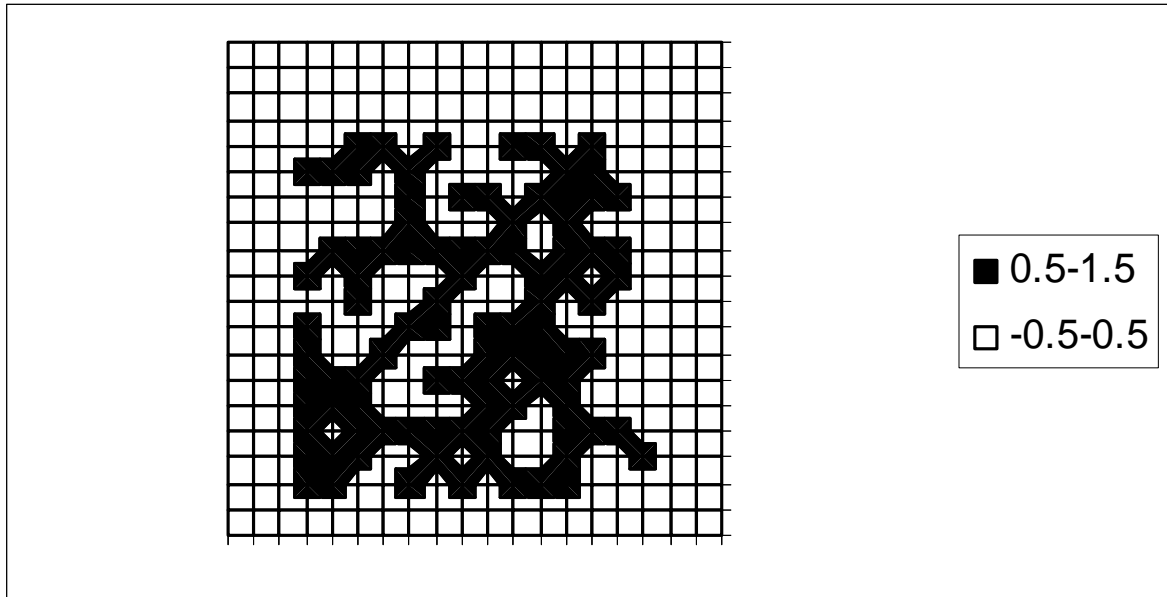


Fig 9ii – A typical ‘X.1.2’ surface graph. They are the surface graph plots of their respective ‘X.1.1’ graphs.

7. Each of the *black and white cubes* in the cells of Fig 7 will be converted to digits '1', '0', or '-1' (so that Figure 7 becomes something like as shown in Figure 11). On Figure 11, '0' means that the nucleotides on similar locations on the plots are the same. '1' or '-1' indicates the nucleotides on similar locations on the plots are different.

8. These 3 digits will enable us to see, how similar or different the 2 sample plots are, and where these similarities and/or differences are located within the sequence.

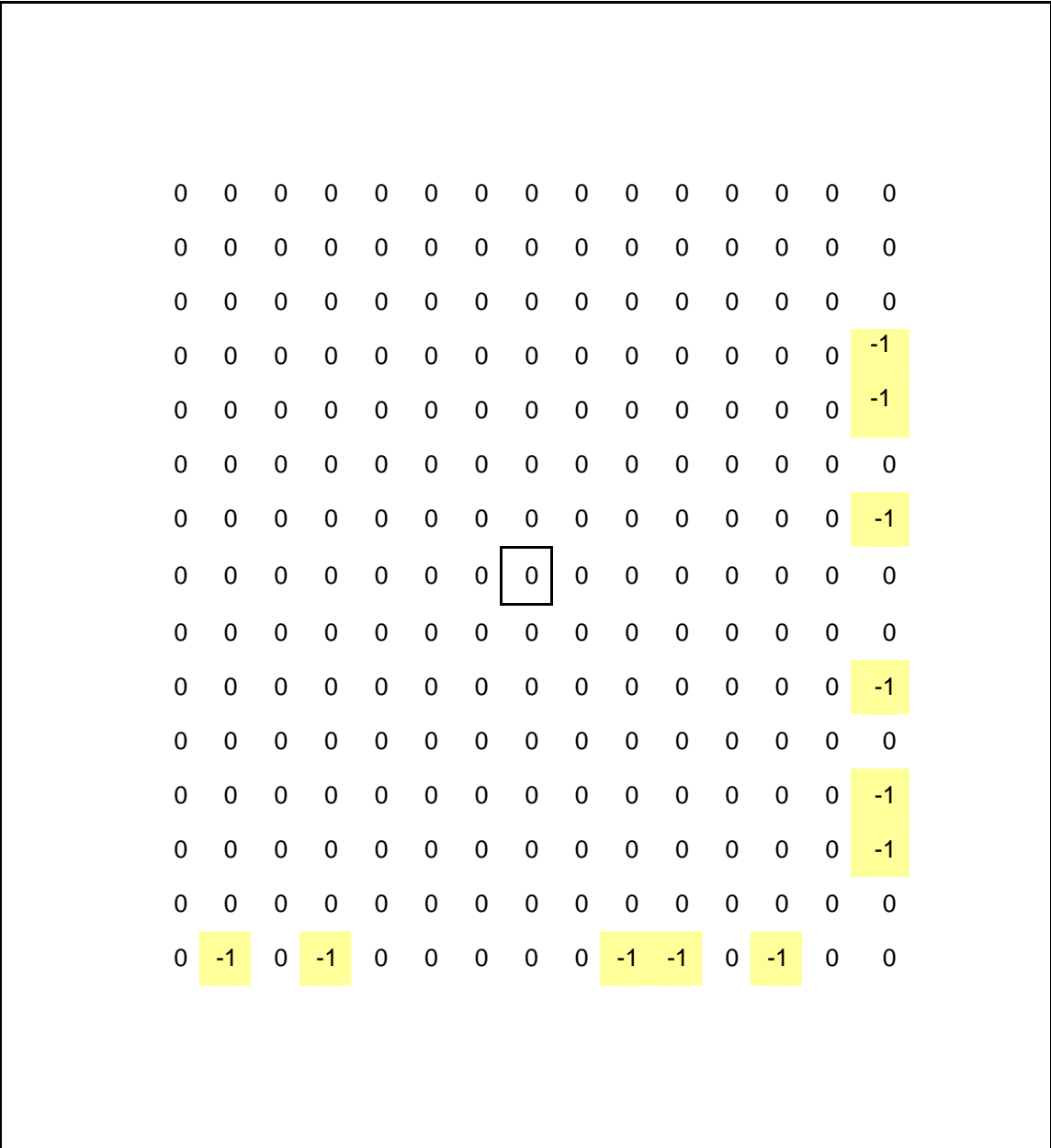


Fig 11 – The ‘X.3.1’ comparison graph after given samples of DNA sequences is plotted. The black and white cubes in Figure 9 are replaced when it was re-plotted using ‘1’, ‘0’ and ‘-1’.

9. Once the plots are done (for each set), we use surface graphs to represent the digits '1', '0' and '-1' on the 'X.3.1' comparison graph (Figure 11). This comparison graph will produce a new surface graph in the form of X.3.2 (Figure 12), where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc), The image on the X.3.2 graphs will then allow us to examine and compare the plotted sequence patterns more closely.

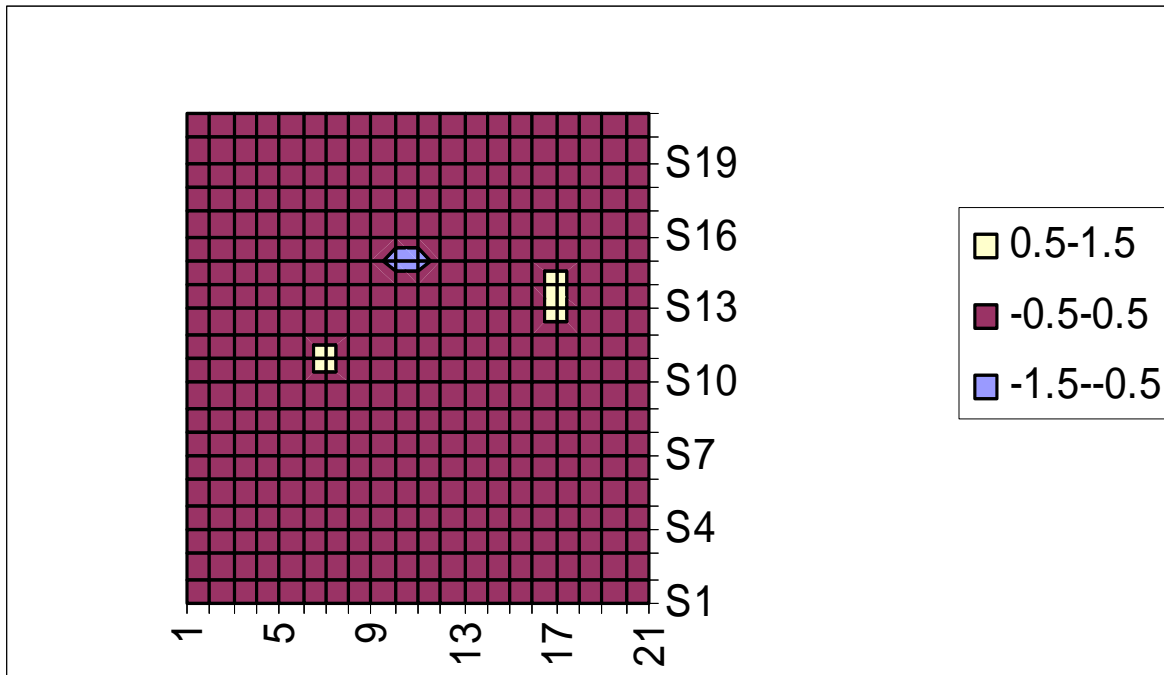


Fig 12 – A typical 'X.3.2' graph of their respective 'X.3.1' graphs.

10. We then use another different combination (of BW/ WW/ BB/ WB) for the 4 nucleotides, and repeat steps 1 to 9.

11. In the end, we aim to get 4 plots (each using a different BW/ WW/ BB/ WB combination) for each pair of animals we chose to pair up with. We will be using 4 different pairs of animal for our studies (namely **chimpanzee-gorilla**, **gorilla-human**, **human-chimpanzee**, and **bovine-goat** pairs).

12. Then from these 2 clusters we will pick a combination that will then be used to compare the human gene sequence with that of the other animals given in Table 1.

3.6 - Comparison of Animal Sequences

On the X.3.2 graphs, 'crosses' are defined as any shade of blue or yellow region that lies on the '+' of the graph. The lesser the number of crosses on the X.3.2 graphs, where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc), the more the graphs are similar to each other in terms of genetic sequences.

The comparison of the plots will be done via the use of the X.3.2 graphs, where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc).

After all the X.3.2 plots are constructed, we try to determine which BW/ WW/ BB/ WB combination gives us the least number of crosses for each pair in the following clusters listed below:

- i) Cluster 1 (Graph Number 1- 12 in Table 3i) – **Human-Gorilla, Gorilla-Chimpanzee and Human- Chimpanzee** Comparison, and
- ii) Cluster 2 (Graph Number 13- 16 in Table 3ii) – **Bovine-Goat** Comparison

Note that each Graph Number may contain more than 1 graph.

Next, we select the a particular BW/ WW/ BB/ WB combination based on the observations seen from Cluster 1 and 2, and proceed to compare the human DNA sequence with that of the 10 other species of animals using this combination gathered from Cluster 1 and 2.

Before we proceed, it is important to note that this BW/ WW/ BB/ WB combination is NOT the best combination available. It is simply a combination that is the closest to the best combination of BW/ WW/ BB/ WB that will give the best results for the plot comparison. And this best combination is an unknown at present.

We name the graphs in this group Cluster 3 – **Human – Animal** Comparison

3.7 - Similarity and Dissimilarity on Comparison Plots

The number of differences for each set is calculated to see and record the quantitative similarities and dissimilarities between 2 plots. **‘Dissimilarity’ or differences is defined quantitatively by the number of crosses the comparison plots (X.3.2 graphs) record.**

In the X.3.2 graphs, only the yellow and blue regions represent the differences, as shown in Figure 13.

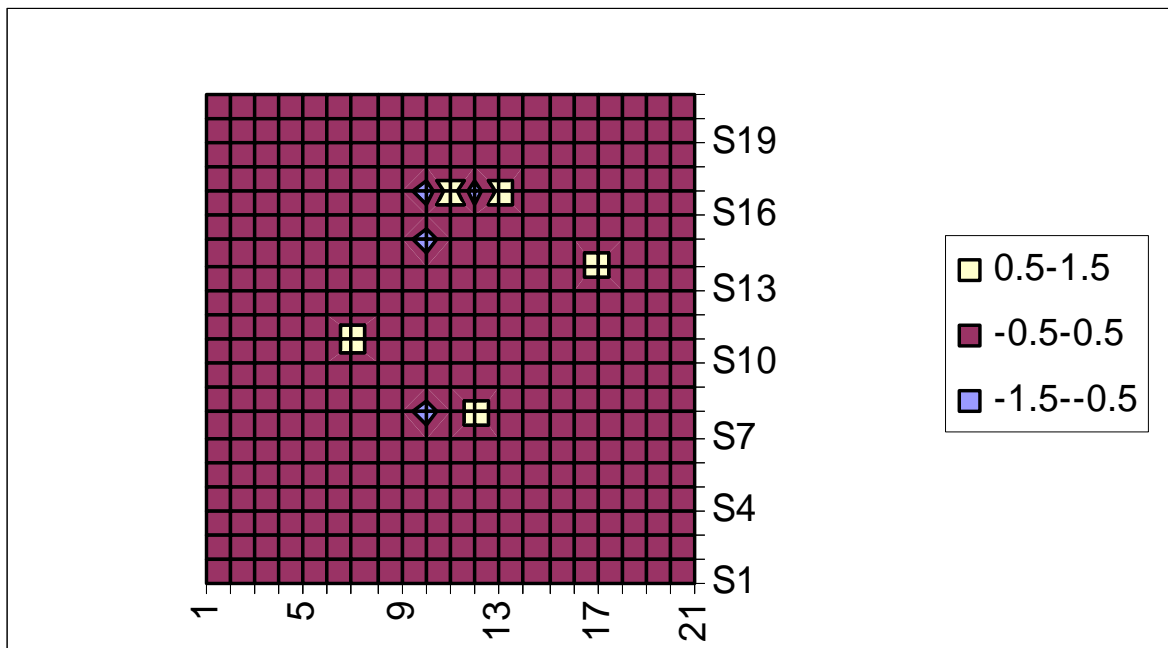


Figure 13 – A typical ‘X.3.2 graph’, and X is the graph number. Here, the yellow and blue regions representing the differences between 2 sample sequences.

It is also useful to note that the sequences to be compared may not be of similar lengths. For instance the first exon of $\hat{\alpha}$ -globin genes for chimpanzee is much longer than first exon of $\hat{\alpha}$ -globin genes for humans. So inevitably, when the 2 individual plots are

compared, the excess length of the chimpanzee gene will be manifested as dissimilarities, or '1's and '-1's.

This will contribute to the crosses on the surface graph plots. Hence we shall include all these when we do the quantitative recording of the number of crosses (that is the number of dissimilarities).

A graph is then constructed to graphically show the magnitude of the differences .

3.8 - List of Combinations Used

The list of BB, BW, WB and WW used will be as follows:

Set 1 –

C – White White

A - Black Black

T – Black White

G – White Black

Set 2 –

T – White White

G - Black Black

C – Black White

A – White Black

Set 3 –

G – White White

C - Black Black

A – Black White

T – White Black

Set 1a –

A – White White

C - Black Black

T – Black White

G – White Black

Note that the numbering of the ‘Sets’ in Table 3i and 3ii, and the graphs plotted are done in accordance to this list. So, for example ‘Set 2’ represents graphs that are plotted using BB/BW/WW/WB combination 2.

3.9 - Determining the closest BB/ BW/ WW/ WB combination (to the best combination) for Individual Clusters

1. After getting the results in Tables 3i and 3ii, we look at the total number of 'crosses' for each graph.
2. Based on the number of crosses for each graph set, we select the BB/ BW/ WW/ WB combination is able to be used to compare human gene sequence and the gene sequences of 10 other animals.

3.10 - Producing Graphical Comparison Plots (Using the Closest BB/ BW/ WW/ WB Combination Determined from Clusters 1 and 2)

3. The final portion of this project will be to use the BB/ BW/ WW/ WB combination (determined in Point 2) to compare the human gene sequence with the gene sequences of 10 other animals given in Table 1, namely Opossum, Gallus Lemur, Mouse, Rabbit, Rat, Gorilla, Bovine, Chimpanzee and Goat.
4. The number of crosses for each human-animal pair is complied.
5. Lastly, we will construct a graph to chart and explain our conclusions based on the observations seen in point 4 above.

4 - Results

4.1 - Labelling of Graphs

Note that the labelling of the ‘Sets’ in Table 3i, 3ii and 3iii are done in accordance to the list as given below.

i) Set 1 – Represents graphs that are plotted using BB/BW/WW/WB combination 1.

C – White White

A - Black Black

T – Black White

G – White Black

ii) Set 2 – Represents graphs that are plotted using BB/BW/WW/WB combination 2.

T – White White

G - Black Black

C – Black White

A – White Black

iii) Set 3 – Represents graphs that are plotted using BB/BW/WW/WB combination 3.

G – White White

C - Black Black

A – Black White

T – White Black

iv) Set 1a – Represents graphs that are plotted using BB/BW/WW/WB combination 1a.

A – White White

C - Black Black

T – Black White

G – White Black

4.2 – List of Graphs for Cluster 1

Graphs for Set 1)

- i) Graph 1: Graphical representation of Gorilla and Chimpanzee genes
- ii) Graph 2: Graphical representation of Human and Chimpanzee genes
- iii) Graph 3: Graphical representation of Human and Gorilla genes

Graphs for Set 2)

- i) Graph 7: Graphical representation of Gorilla and Chimpanzee genes
- ii) Graph 8: Graphical representation of Human and Chimpanzee genes
- iii) Graph 9: Graphical representation of Human and Gorilla genes

Graphs for Set 3)

- i) Graph 10: Graphical representation of Gorilla and Chimpanzee genes
- ii) Graph 11: Graphical representation of Human and Chimpanzee genes
- iii) Graph 12: Graphical representation of Human and Gorilla genes

Graph for Set 1a)

- i) Graph 4: Graphical representation of Gorilla and Chimpanzee genes
- ii) Graph 5: Graphical representation of Human and Gorilla genes
- iii) Graph 6: Graphical representation of Human and Chimpanzee genes

4.3 - Results for Cluster 1: Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee Set

	Set 1			Set 2			Set 3			Set 1a		
BW/WB/WW/BB Coding	1/4			2/4			3/4			1a/4		
Graph Number	1	2	3	7	8	9	10	11	12	4	5	6
Number of 'crosses' in each graph	18	19	4	12	14	2	11	10	3	10	1	11

Table 3i - Table of results for Cluster 1- Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee set. The number of crosses on the 3rd graph that belong to each graph number X.3.2, where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc) will determine the similarities and dissimilarities between 2 DNA sequence samples. 'Crosses' are defined as any shade of blue or yellow region that lies on the '+' of the graph. The highlighted region shows the set that has the LEAST number of differences

4.4 – List of Graphs for Cluster 2

Graphs for Set 1)

Graph 13: Graphical representation of Goat and Bovine genes

Graphs for Set 2)

Graph 14: Graphical representation of Goat and Bovine genes

Graphs for Set 3)

Graph 15: Graphical representation of Goat and Bovine genes

Graph for Set 1a)

Graph 16: Graphical representation of Goat and Bovine genes

4.5 - Results for Cluster 2: Bovine – Goat Set

	Set 1	Set 2	Set 3	Set 1a
BW/WB/WW/BB Coding	1/4	2/4	3/4	1a/4
Graph Number	13	14	15	16
Number of 'crosses'	9	5	5	3

*Table 3ii –Table of results for Cluster 2 – **Bovine- Goat** set. The number of crosses on the 3rd graph that belong to each graph number X.3.2, where X is the graph number (example **1.3.2**, **2.3.2**, **3.3.2**, etc) will determine the similarities and dissimilarities between 2 DNA sequence. 'Crosses' are defined as any shade of blue or yellow region that lies on the '+' of the (13.1.3, 13.2.3, etc) graphs. The highlighted region shows the set that has the LEAST number of differences.*

4.6 – List of Graphs for Cluster 3

Graphs for Set 3)

- i) Graph 17: Graphical representation of Human and Gallus genes
- ii) Graph 18: Graphical representation of Human and Opossum genes
- iii) Graph 19: Graphical representation of Human and Lemur genes
- iv) Graph 20: Graphical representation of Human and Mouse genes
- v) Graph 21: Graphical representation of Human and Rabbit genes
- vi) Graph 22: Graphical representation of Human and Rat genes
- vii) Graph 23: Graphical representation of Human and Bovine genes
- viii) Graph 24: Graphical representation of Human and Goat genes
- ix) Graph 11: Graphical representation of Human and Chimpanzee genes
- x) Graph 12: Graphical representation of Human and Gorilla genes

4.7 - Results for Cluster 3: Human-Animal Set

	Set 3									
BW/WB/WW/BB Coding	1a/4									
Graph Number	17	18	19	20	21	22	23	24	5	6
Number of 'crosses'	31	31	30	94	11	11	75	80	1	11

*Table 3iii - Table of results for Cluster 3 – **Human-Animal** set. The number of crosses on the 3rd graph that belong to each graph number X.3.2, where X is the graph number (example **1.3.2**, **2.3.2**, **3.3.2**, etc) will determine the similarities and dissimilarities between 2 DNA sequence samples. 'Crosses' are defined as any shade of blue or yellow region that lies on the '+' of the graph. The highlighted region shows the set that has the **LEAST** number of differences.*

This table includes data from graphs 11 and 12 that has been recorded in Table 3i earlier.

5 – Interpretation of Results

5.1 - Best BB/ BW/ WW/ WB combination for Clusters 1 and 2

From Table 3i, it is obvious that the 1a/ 4 combination:

Set 1a) –

A – White White

C - Black Black

T – Black White

G – White Black

gives the least number of crosses (Actual value = $10 + 1 + 11 = 22$). That is to say, among the 4 combinations used, the 1a combination is the closest to the best BB/ BW/ WW/ WB combination for the majority of pairs in Cluster 1.

From Table 3ii, it is obvious that the 1a/ 4 combination:

A – White White

C - Black Black

T – Black White

G – White Black

gives the least number of crosses (Actual value =3). That is to say, among the 4 combinations used, the 1a combination is the closest to the best BB/ BW/ WW/ WB combination for the bovine-goat comparison in Cluster 2.

Comparing the results, the combination that gave Cluster 1 the least number of crosses is the same BB/ BW/ WW/ WB combination that will give Cluster 2 the least number of crosses.

In Cluster 1, all the Gorilla-Human plots (Graphs 3, 5, 9, and 12) gave the lowest number of crosses, regardless of the BB/ BW/ WW/ WB combination used.

5.2- Results of Comparison of Human Gene Sequence with 10 Other Animals Using ‘the Best BB/ BW/ WW/ WB combination for Clusters 1 and 2: Combination 1a’

The more similarities in the genetic sequence of the first exon of α -globin genes of humans and the other animals (as listed in Table 1), the smaller the number of crosses.

The list of animals and their differences with the human genome is given in Figure 14.

Number of crosses

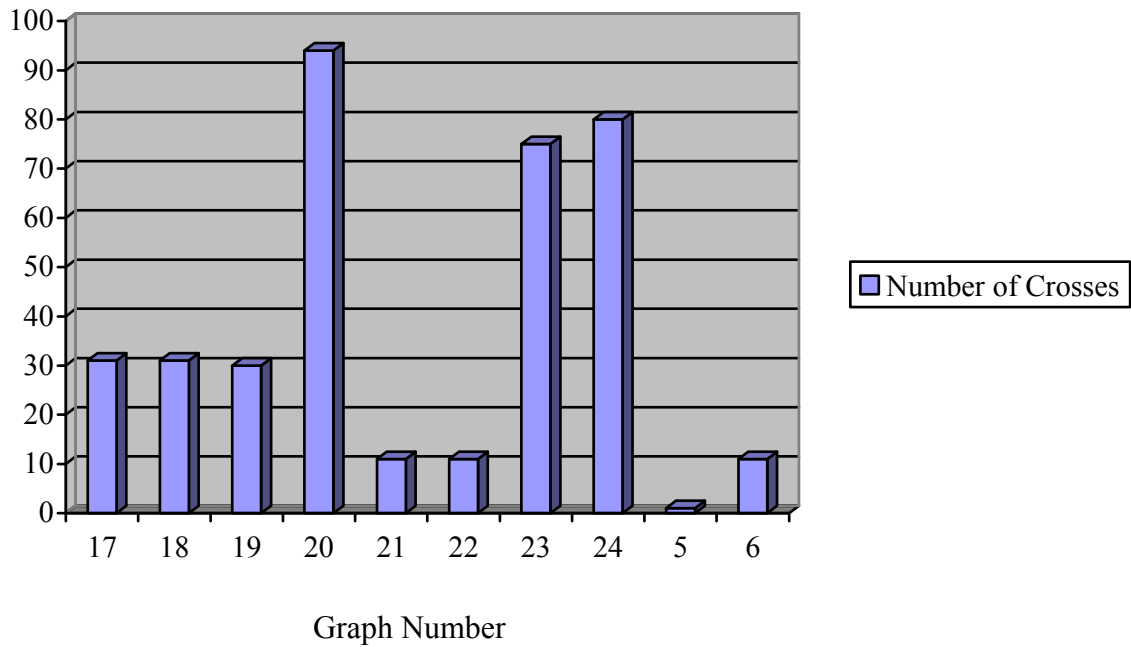


Figure 14 - A plot showing the number of crosses when the gene sequences of the 10 animals are compared to the human gene sequence using the 1a combination.

From Table 3iii and Figure 14, it is obvious that gorilla bears the most similarities with the human gene (highlighted in YELLOW, graph 5, Number of crosses = 1).

Also, we noted that mouse (highlighted in ORANGE, graph 20, Number of crosses = 94), bears the least similarities with humans.

Hence from the list of animals given in Table 1, we conclude that gorilla bears the most similarities with the human gene while the mouse bears the least similarities with the human gene.

5.3 - Specific Advantages of Explored Method

- 1) **Relatively low lost of information:** It prevents loss of information during the transfer of data from the actual DNA sequence to its graphical representation because it only involves a very simple direct conversion of DNA nucleotides to BB, BW, WB or WW combination, followed by conversion to 1 0 -1 codes.

- 2) **Easy to use:** Our representation is compact and easy to use. It allows visual inspection of physical similarities/ differences between DNA sequences (In terms of DNA sequences only).

- 3) **Infinite comparison possibilities:** While we are using a small segment for this project, we must bear in mind that scientists are likely to use larger, more complex segments so as to study relations between supposedly related DNA samples. Our method allows large scale comparison of DNA sequences between infinite no of plots, and with relatively little effort needed. All that is needed is to change the math formula that enables numerical characterization of the plots. That is to say, even if we were to map out few thousand nucleotides using the EXCEL spreadsheet we devised, we can still effectively compare this huge plot with another equally huge plot. Though tedious, the visual graphics of this spreadsheet still enable us to spot similarities and differences quite easily. This method is also applicable to any DNA, even RNA sequences.

5.4 - Specific Disadvantages of Explored Method

- 1) **Combination of BB/BW/WB/WW affects results:** Not all combinations for C, A, T and G will give good plots, graphs and comparison for 2 subjects whose DNA sequence may be quite similar to. In the BW plots the degree of visual similarity differs (for each set) according to which combination of BW, BB, WB and WW are used for each nucleotide, besides due to the difference in DNA sequences between 2 samples. Some BB BW WB WW combinations are likely to give better plots and graphs that shows a closer relationship between say 2 plots, while some will give a 'less match' result between the same 2 subject; Hence if specific combinations of BB BW WB WW do not show a good match between 2 samples, we cannot make any valid conclusion about the degree of similarities or dissimilarities till we compare results obtained via usage of few different combinations.
- 2) Last but not least, DNA sequences are only one line of evidence illuminating evolutionary relationships. For example, human and chimpanzee DNA is 98% identical, and genetic sequencing can tell us exactly where in the genome those few DNA differences are — but anatomical, behavioural, and developmental studies are also crucial in deeply understanding our differences, similarities, and shared evolutionary history.

6 - Conclusion

Our 2-D graphical representation is useful for visualizing of DNA sequences. Graphical representations have the potential means of facilitating the discovery of similar/dissimilar sequences between any number of DNA segments.

Is our method a more accurate model? It largely depends on what kind of information we want to obtain. If one is only interest in how gene sequences differ from another sample graphically using EXCEL plots/graphs, then the method devised in this project will be adequate.

If we are not concerned about the form, or its repetitive nature of the DNA sample chosen, then this method should not be used on its own.

It is also vital to note that this visual comparison method will be of less practical use as the differences between the samples increases. Hence it is necessary for us to have a method to compare the sequences quantitatively. Numerical characterization of our plots may be created.

Indeed our unique 2-D representation of sequences provides different approaches for both computational scientists and molecular biologists to analyses DNA sequences efficiently.

For any specific similarities/ dissimilarities data to be of use, knowledge obtained from this project should form the basis of facilitate ease or produce potential for further research. In other words, it should be utilised together with existing gene sequencing methods that is able to compare the results obtained with our method to other available methods, such that we can easily determine which set of DNA samples are worth studying. However some 2-D and 3-D graphical representation are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself (7) (8).

We hope future developments in our project will facilitate comparative studies of DNA and open new directions for representation and characterization of DNA primary sequences.

This is especially vital when it comes to improving efficiency when studying virulent germs with high mutation rate, like H5N1, where it can be a potentially daunting task to map, the compare specific sections of DNA samples. There also has been suggestions that this method we devised can be used to investigate segments from the same strand from different virus strains (2) (3).

7 - Further Research Suggestions

- 1) A good way to test this method is to do representations for the entire genomic sequence for a subject in question. However at present for this project, we only present a simple demonstration of our method of graphical representation, and how this method compares with similar methods in existence, so the DNA sequences we used are those that have also been used by other scientists.
- 2) A possible extension of this project will to determine the best BB BW WB WW combination that can give the best visual representation and best match between a certain number of samples.
- 3) Indeed, even the simple 2-D graphical representation of DNA sequences should provide different approaches for both computational scientist sand molecular biologists to adequately analyze DNA sequences efficiently with changes in parameter. Also, by numerically characterizing the data obtained from our project, important features of DNA can be captured and studied more thoroughly and effectively.

For instance in Randic's *Four-color map representation of DNA or RNA sequences and their numerical characterization*, matrices (for instance E matrix, D/D matrix, L/L matrix and their "high order" matrices) are used to the numerically characterize the four colour maps devised so as to gain deeper insight into similarities/ dissimilarities between the sequences (8). Similar methods can be used to characterizing our graphs.

Not only do they have the potential to provide considerable information on new global sequence homologies, repeated structures, relative base abundances, provide useful insights into local characteristics and the occurrences, variations and repetition of the nucleotides along a sequence which are not as easily obtainable by other methods, they also can be used to study probable evolutionary paths/evolutionary divergences, reveal differences in the features of exon/ intron segments, and protein coding/non-coding regions of eukaryotic sequences.

4) For some models devised, the differences in length have created some problems in the process of sequence comparison. A base by base comparison would be of little use when sequences have different lengths. Finding the best correspondence between sequences (as outlined by Kruskal involves various strategies, such as trace (linking the same elements in the two sequences), alignment, or matching (spacing elements using blank spaces between) may be easy, but the resultant potting may involve some complex math like construction of D/D matrices and calculation of its respective eigenvalues.

For instance, in Milan Randic and Marjan Vracko study, again, numerical characterization is done via the use of eigenvalues of so-constructed matrices, which will then be used to construct numerical sequences, and subsequently used for similarity/dissimilarity analysis. These will likely to create some inaccuracy due to lost of information associated with eigenvalues, and we have to use the results with some caution (8).

While such numerical representations allow visual inspection of data, they are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself. There are still means to overcome such problems, but we shall not discuss them in this project.

Even when we are simply using our very simple graphical representation, the best is to combine such information with similar information obtained by some other characterization of the set of structures.

How methods are there to allow for fast, hassle-free identification and retrieval of relevant sequences, and efficient comparisons between different DNA samples? To do this, these sequences are likely to be graphically represented and characterized using specific methods. Numerous options (for example E, D/D, L/L matrix) are available to numerically characterized, then represent the sequences in a digitized pictorial form or in a table form with calculated values so that fast and efficient comparisons between different DNA samples can be carried out (12). Alternative methods use a set of

invariants of biological sequences, rather than directly using sequence alignment like BLAST.

Another particular graphical representation of DNA sequences can involve something like the 'four line' graphical representation by Randić, Vračko, Lerš, Plavšić in 2003 (11). The 'four line' graphical representation of DNA is transformed into numerical sequence that reduced comparisons of different sequences to arithmetic manipulations with DNA primary sequences.

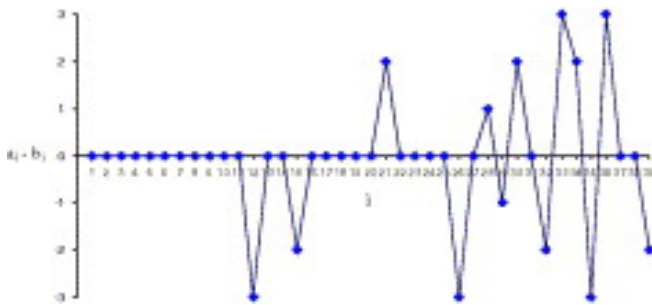


Fig 15 - It is shown here one method of numerical characterization (11)

4) Computer programs that automatically transfer DNA sequence data to the '1', '0', '-1' plots can also be created so that manual plotting is avoided. This will make the plotting process much easier and less prone to error.

These offer potential insight and discoveries, as the first exon of DNA from subject A may have 99% similarity to the same segment of DNA from subject B, but the third exon of DNA from subject A may not have the same EXACT similarity to the same segment of DNA from subject B.

8 - Critical Review

When Dr Lim and I chose this project, we are charting into an unknown territory when we chose to representation gene sequences using binary codes in spiral conformation to look for method, that it, will for one, give us the closest WW/ WB/ BB /WW combination to best represent the genes of for most genetic code sequences over different species. The best combination is of course an unknown, and the methods we devised is something that is not guaranteed to produce any results at all. That is why I'm particularly proud, that after completion of the project, Dr Lim and I still managed to get some useful data from this project.

One regrettable thing is that for this 'closest' BW/ WW/ BB/ WB combination there is no time to determine how close we are to the best combination available for this group of animals. If time permits, more plots could have been created, compared and studied. If not for he time constraint, we will sure to get a BW/ WW/ BB/ WB combination that is even closer to the best combination available for this group of animals. With more time, we could have been able to quantitatively determine how far the combination we found is from the so-called best combination, which till this day remains an unknown.

Another difficulty encountered is the sheer amount of graphs to be plotted. Each graph for each animal have to be plotted at least 4 times, one nucleotide after the other in accordance to the DNA sequences given. There are 11 animals in this project, so I have to plot at least 40 graphs. Plotting itself is no easy feat. Despite its simplicity, plotting can be tedious when one has to plot many as 40 over graphs.

Moreover one single mistake, whether it is the use of the wrong coding, or missing out on a particular nucleotide) can create a major error in the subsequent comparison plots (like the X.3.2 plots). If these small errors in the individual DNA sequence plots are not detected in time, it will be carried forward and create large, erroneous plotting that will inevitably misinform us and affect proper interpretation of the results obtained.

Therefore, after the first plot, the graphs must be checked at least twice to ensure that no errors are present. This means plotting as many as 80 graphs.

I had an unpleasant experienced when a small error in the initial comparison plots of Graphs 1 and 3 resulted in 'the closest BB/ BW/ WW/ WB combination for Clusters 1 and 2' being determined incorrectly. The difference in the number of crosses (between graphs of the correct and incorrect combination) is only 2, yet, I have to re-do at least 10 plots for that single error, not to mention changing a lot of data and write-ups to reflect that change.

However despite it all, this project is still a worthwhile venture, given the results we have obtained.

9 - References

- (1) Ashesh Nandy: *Novel Method for Discrimination of Conserved Genes through Numerical Characterization of DNA Sequences*
- (2) Ashesh Nandy, Subhash C. Basak, and Brian D. Gute: *Graphical Representation and Numerical Characterization of H5N1 Avian Flu Neuraminidase Gene Sequence*
- (3) Jay E. Gee, Claudio T. Sacchi, Mindy B. Glass, Barun K. De, Robbin S. Weyant, Paul N. Levett, Anne M. Whitney, Alex R. Hoffmaster, and Tanja Popovic: *Use of 16S rRNA Gene Sequencing for Rapid Identification and Differentiation of Burkholderia pseudomallei and B. mallei*
- (4) Qi Dai, Xiao-qing Liu and Tian-ming Wang : *C(i,j) matrix: A better numerical characterization for graphical representations of biological sequences*
- (5) Milan Randić, Jure Zupan, Dražen Vikić-Topić and Dejan Plavšić : *A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences*
- (6) Anonymous *BioWeb* © 2002 The Board of Regents of the University of Wisconsin System
- (7) Jun Wang , Yi Zhang: *Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation*
- (8) Yu-hua Yao and Tian-ming Wang: *A class of new 2-D graphical representation of DNA sequences and their application*
- (9) Wilson, Sarich, Sibley, and Ahlquist: *Genetic Similarities*
- (10) Milan Randić, Jure Zupan, Dražen Vikić-Topić and Dejan Plavšić: *A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequence*

(11) Randic, Nella, Dejan, Basak and Balaban: Four-color map representation of DNA or RNA sequence and their numerical characterization

(12) Bo Liao, MingShu Tan, Kequan Ding: *A 4D representation of DNA sequences and its application*

(13) Isilanes: 'Deoxyribonucleic Acid' from Wikipedia, the Free Encyclopaedia

(14) Anonymous: 'Sequence Alignment' from Wikipedia, the Free Encyclopaedia

10 – Graph List

It is useful to note that

- i) X.1.1 and X.2.1 Graphs refer to a graph that is produced when black white binary codes used to map out DNA sequences are changed to digits '1', '0', with '1' replacing the black cubes, and '0' replacing the white cubes. In these graphs, X is the graph number (for example 1.1.1, 1.2.1, 2.1.1, 3.1.1, etc plots belong to this category).

- ii) X.1.2 and X.2.1 Graphs are the surface graph plots of their respective 'X.1.1' graphs.

- iii) X.3.1 Graphs refer to a new graph that is produced so that individual sequence graphs of 2 animals can be compared visually (in terms of digits '1', '0' and '-1'. In these graphs, X is the graph number (for example 1.3.1, 2.3.1, 3.3.1, etc plots belong to this category).

- iv) X.3.2 Graphs are surface graph plots of their respective 'X.3.1' graphs.

10.1 Graphs for Cluster 1 – Human-Gorilla, Gorilla-Chimpanzee and Human-Chimpanzee Comparison (Graphs 1-12)

10.1.1 Graphs for BW/WB/WW/BB Combination 1

Graph 1 - Graph for Set 1): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Gorilla and Chimpanzee (font 9)

- i) *Graph 1.1.1 – Graph showing numerical representation of genes of Gorilla*
- ii) *Graph 1.1.2 –Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iii) *Graph 1.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iv) *Graph 1.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- v) *Graph 1.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*
- vi) *Graph 1.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*

Graph 2 - Graph for Set 1): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Chimpanzee

- i) *Graph 2.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- ii) *Graph 2.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iii) *Graph 2.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iv) *Graph 2.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- v) *Graph 2.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*
- vi) *Graph 2.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*

Graph 3 - Graph for Set 1) : Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla

- i) *Graph 3.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- ii) *Graph 3.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iii) *Graph 3.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iv) *Graph 3.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- v) *Graph 3.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Gorilla and Human*
- vi) *Graph 3.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Gorilla and Human*

10.1.2 - Graphs for BW/WB/WW/BB Combination 1a

Graph 4 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Gorilla and Chimpanzee

- i) *Graph 4.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- ii) *Graph 4.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iii) *Graph 4.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iv) *Graph 4.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- v) *Graph 4.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*
- vi) *Graph 4.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*

Graph 5 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Gorilla and Human.

- i) *Graph 5.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- ii) *Graph 5.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iii) *Graph 5.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 5.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 5.3.1– Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*
- vi) *Graph 5.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*

Graph 6 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human.

- i) *Graph 6.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- ii) *Graph 6.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iii) *Graph 6.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 6.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 6.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*
- vi) *Graph 6.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*

10.1.3 - Graphs for BW/WB/WW/BB Combination 2

Graph 7 - Graph for Set 2): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Gorilla and Chimpanzee

- i) *Graph 7.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- ii) *Graph 7.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iii) *Graph 7.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iv) *Graph 7.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- v) *Graph 7.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*
- vi) *Graph 7.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*

Graph 8 - Graph for Set 2): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Chimpanzee

- i) *Graph 8.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- ii) *Graph 8.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iii) *Graph 8.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 8.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 8.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*
- vi) *Graph 8.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*

Graph 9 - Graph for Set 2): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla

- i) *Graph 9.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- ii) *Graph 9.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iii) *Graph 9.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iv) *Graph 9.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- v) *Graph 9.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*
- vi) *Graph 9.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*

1.1.4 - Graphs for BW/WB/WW/BB Combination 3

Graph 10 - Graph for Set 3): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Gorilla and Chimpanzee

- i) *Graph 10.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- ii) *Graph 10.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iii) *Graph 10.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iv) *Graph 10.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- v) *Graph 10.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*
- vi) *Graph 10.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Gorilla*

Graph 11 - Graph for Set 3): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Chimpanzee

- i) *Graph 11.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- ii) *Graph 11.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee*
- iii) *Graph 11.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 11.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 11.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*
- vi) *Graph 11.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Chimpanzee and Human*

Graph 12 - Graphs for BB/ WW/WB/WW combination 3 (Set 3): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla

- i) *Graph 12.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- ii) *Graph 12.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iii) *Graph 12.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- iv) *Graph 12.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gorilla*
- v) *Graph 12.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*
- vi) *Graph 12.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gorilla*

10.2 Graphs for Cluster 2 – Goat – Bovine Comparison (Graphs 13-16)

Graph 13 - Graph for Set 1): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Bovine and Goat

- i) *Graph 13.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- ii) *Graph 13.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- iii) *Graph 13.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- iv) *Graph 13.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- v) *Graph 13.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*
- vi) *Graph 13.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*

Graph 14 - Graph for Set 2) : Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Bovine and Goat

- i) *Graph 14.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- ii) *Graph 14.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- iii) *Graph 14.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- iv) *Graph 14.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- v) *Graph 14.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*
- vi) *Graph 14.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*

Graph 15 - Graph for Set 3) : Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Bovine and Goat

- i) *Graph 15.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*
- ii) *Graph 15.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*
- iii) *Graph 15.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- iv) *Graph 15.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- v) *Graph 15.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- vi) *Graph 15.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*

Graph 16 - Graph for Set 1a) : Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Bovine and Goat

- i) *Graph 16.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- ii) *Graph 16.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- iii) *Graph 16.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- iv) *Graph 16.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- v) *Graph 16.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*
- vi) *Graph 16.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat and Bovine*

10.3 Graphs for Cluster 3 – Human–Animal Comparison (Graphs 17- 24)

Graph 17 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Gallus

- i) *Graph 17.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Gallus*
- ii) *Graph 17.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Gallus*
- iii) *Graph 17.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 17.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 17.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gallus*
- vi) *Graph 17.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Gallus*

Graph 18 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Opossum

- vii) *Graph 18.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Opossum*
- viii) *Graph 18.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Opossum*
- ix) *Graph 18.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- x) *Graph 18.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- xi) *Graph 18.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Opossum*
- xii) *Graph 18.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Opossum*

Graph 19 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Lemur

- i) *Graph 19.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Lemur*
- ii) *Graph 19.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Lemur*
- iii) *Graph 19.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 19.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 19.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Lemur*
- vi) *Graph 19.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Lumur*

Graph 20 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Mouse

- i) *Graph 20.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Mouse*
- ii) *Graph 20.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Mouse*
- iii) *Graph 20.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 20.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 20.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Mouse*
- vi) *Graph 20.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Lumur*

Graph 21 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Rabbit

- i) *Graph 21.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Rabbit*
- ii) *Graph 21.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Rabbit*
- iii) *Graph 21.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 21.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 21.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Rabbit*
- vi) *Graph 21.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Rabbit*

Graph 23 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Bovine

- i) *Graph 23.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- ii) *Graph 23.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- iii) *Graph 23.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- iv) *Graph 23.2.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Human*
- v) *Graph 23.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Bovine*
- vi) *Graph 23.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Bovine*

Graph 24 - Graph for Set 1a): Graphical representation of first exon of $\hat{\alpha}$ -globin genes of Human and Goat

- i) *Graph 24.1.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- ii) *Graph 24.1.2 – Surface graph showing graphical representation of the first exon of $\hat{\alpha}$ -globin genes of Bovine*
- iii) *Graph 24.2.1 – Graph showing numerical representation of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- iv) *Graph 24.2.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Goat*
- v) *Graph 24.3.1 – Graph showing numerical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Goat*
- vi) *Graph 24.3.2 – Surface graph showing graphical representation for comparison of the first exon of $\hat{\alpha}$ -globin genes of Human and Goat*

11 - Appendix

11.1 – Definition of Terms

- i) X.1.2 Graphs: They are the surface graph plots of their respective 'X.1.1' graphs.

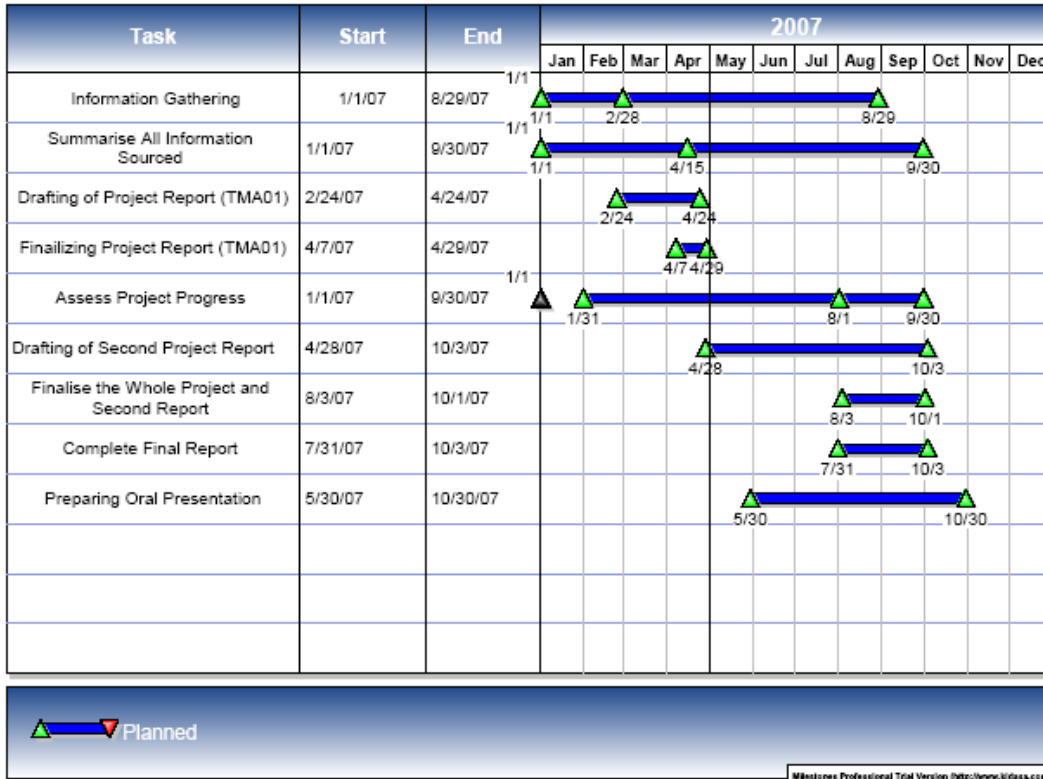
- ii) X.3.1 Graphs: They refer to a new graph that is produced so that individual sequence graphs of 2 animals can be compared visually (in terms of digits '1', '0' and '-1'. In these graphs, X is the graph number (for example 1.3.1, 2.3.1, 3.3.1, etc plots belong to this category).

- iii) X.3.2 Graphs: They refer to a new surface graph plots of their respective 'X.3.1' graph

- iv) Crosses: On the X.3.2 graphs, they are defined as any shade of blue or yellow region that lies on the '+' of the graph. The lesser the number of crosses on the X.3.2 graphs, where X is the graph number (example 1.3.2, 2.3.2, 3.3.2, etc), the more the graphs are similar to each other in terms of genetic sequences.

- v) Dissimilarity: It is defined quantitatively by the number of crosses on the X.3.2 graphs, where X is the graph number.

11.2 – Gantt Progress Chart



12 – Notes

END OF REPORT