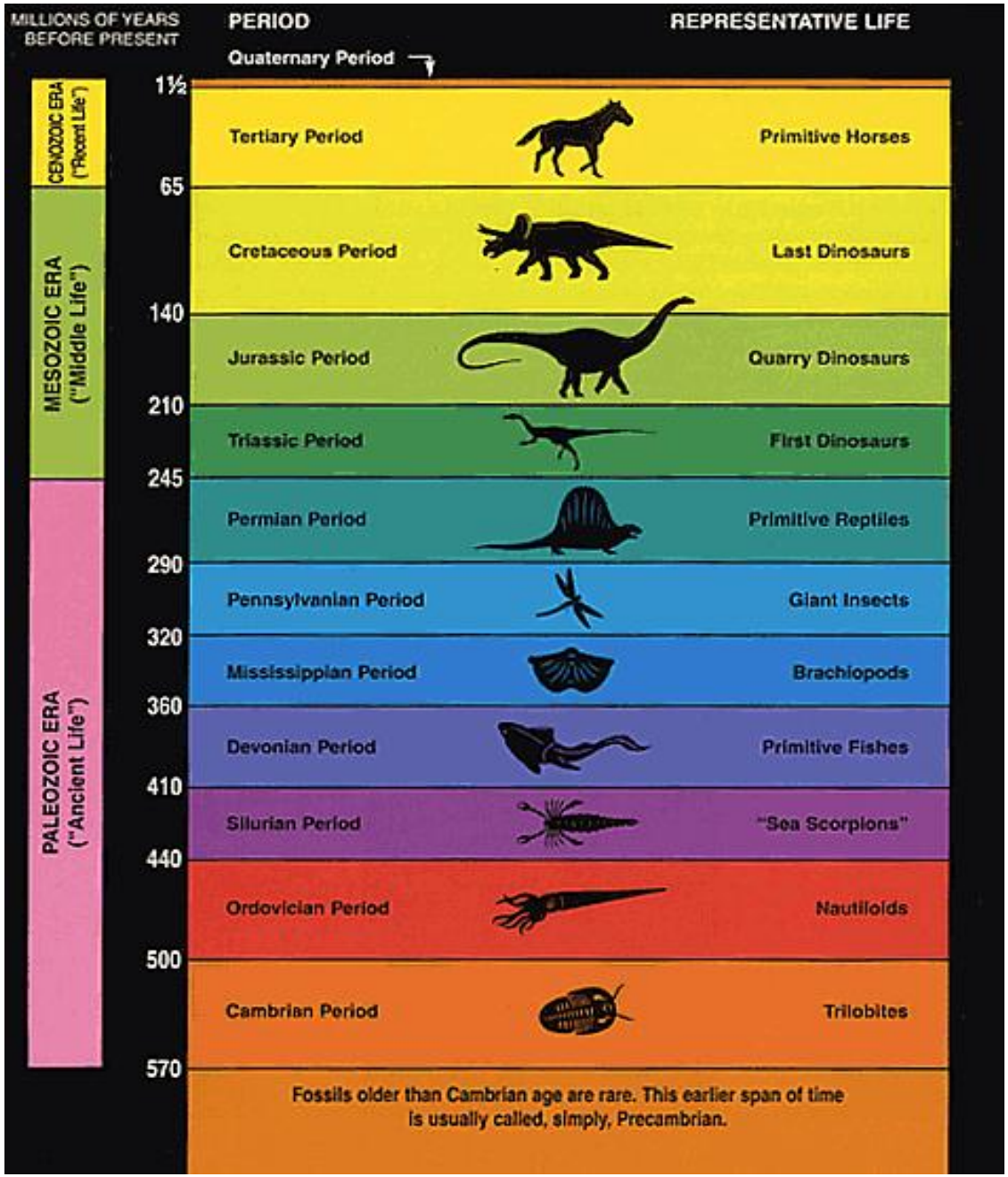
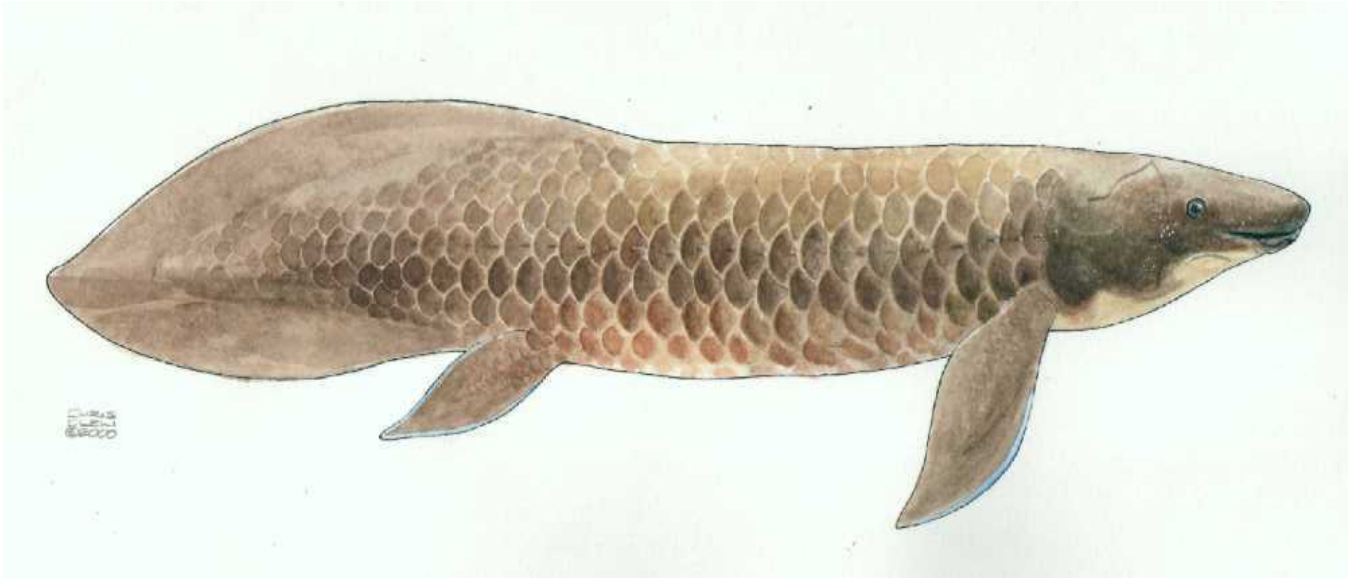
Statistical Learning Based on Distributions of Oligonucleotides in DNA Sequences

Probal Chaudhuri

Indian Statistical Institute, Calcutta


MILLIONS OF YEARS BEFORE PRESENT
PERIOD
REPRESENTATIVE LIFE
Quaternary Period
1½
CENOZOIC ERA ("Recent Life")
Tertiary Period
Primitive Horses
65
MESOZOIC ERA ("Middle Life")
Cretaceous Period
Last Dinosaurs
140
Jurassic Period
Quarry Dinosaurs
210
Triassic Period
First Dinosaurs
245
PALEOZOIC ERA ("Ancient Life")
Permian Period
Primitive Reptiles
290
Pennsylvanian Period
Giant Insects
320
Mississippian Period
Brachiopods
360
Devonian Period
Primitive Fishes
410
Silurian Period
"Sea Scorpions"
440
Ordovician Period
Nautiloids
500
Cambrian Period
Trilobites
570
Fossils older than Cambrian age are rare. This earlier span of time is usually called, simply, Precambrian.

Australian Lungfish



Coelacanth, the living fossil

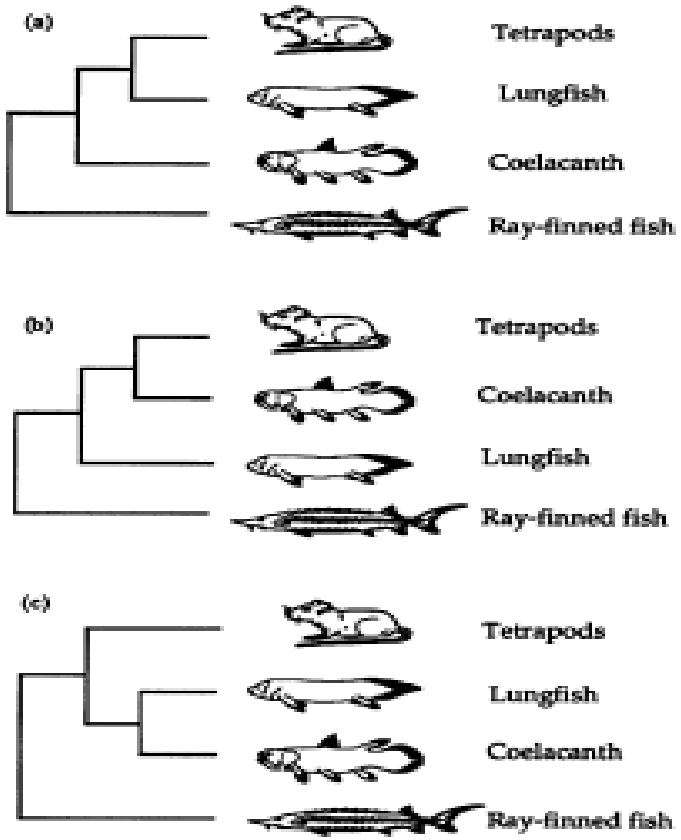


FIG. 1. Alternative hypotheses of sister group relationships between sarcopterygii and tetrapods. (*A*) Lungfish as the sister group of tetrapods. (*B*) Coelacanth as the closest living relative of tetrapods. (*C*) Coelacanth and lungfish equally closely related as sister groups of tetrapods.

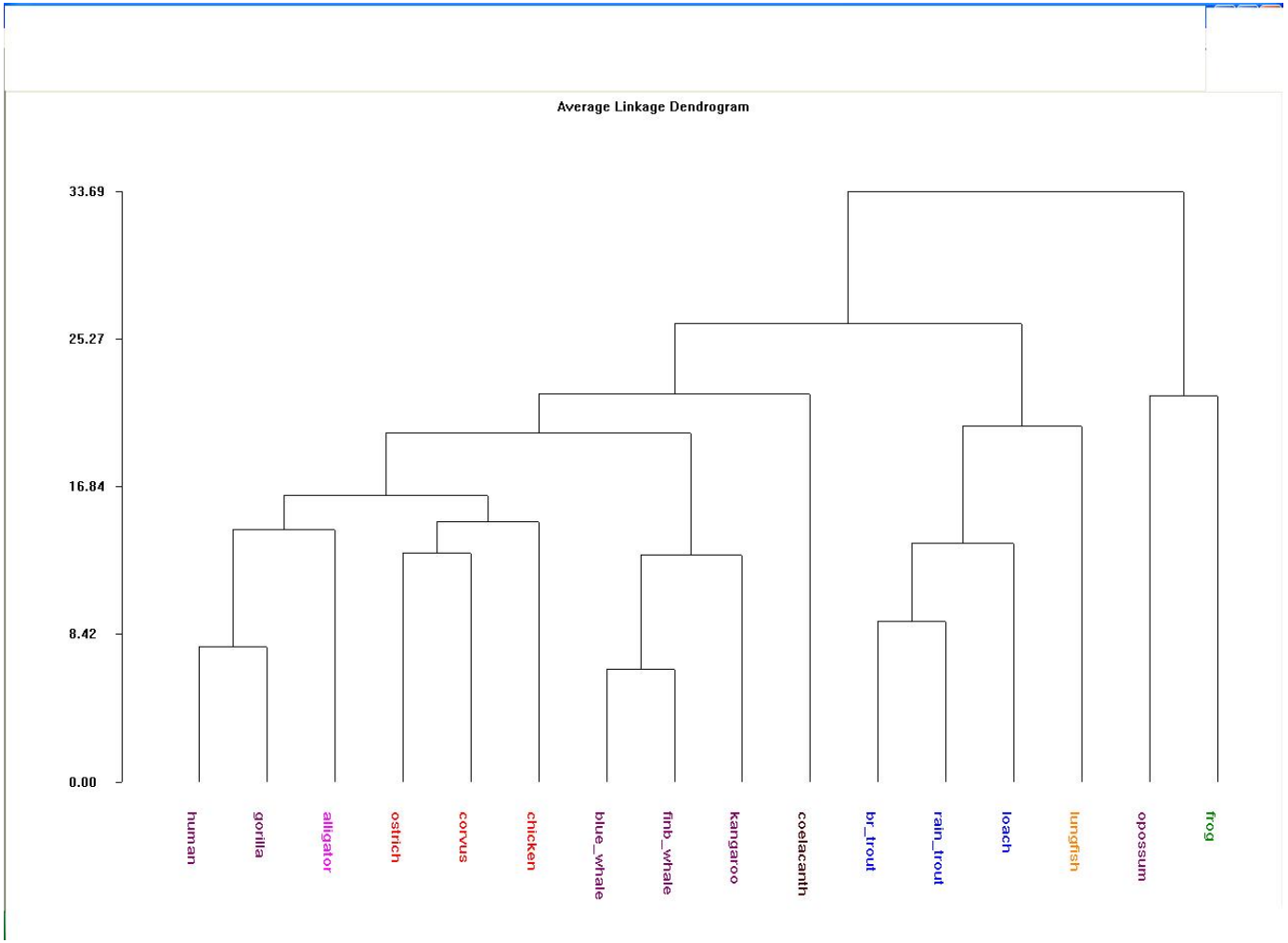
Analysis of Mitochondrial Genomes

- DNA sequences of complete mitochondria of several vertebartes.
- Lengths of sequences varying from 16K to 18K nucleotide bases.
- The sample includes a few mammals, a few birds and a few standard types of fish. It also contains a reptile and an amphibian animal.

Statistical Analysis of DNA Sequences

- Oligo-nucleotide frequencies are computed for each mitochondrial sequence.
- These frequencies can be used as numerical measurements or variables. They form multivariate observations associated with the given DNA sequences.
- One can carry out cluster analysis with this multivariate data.

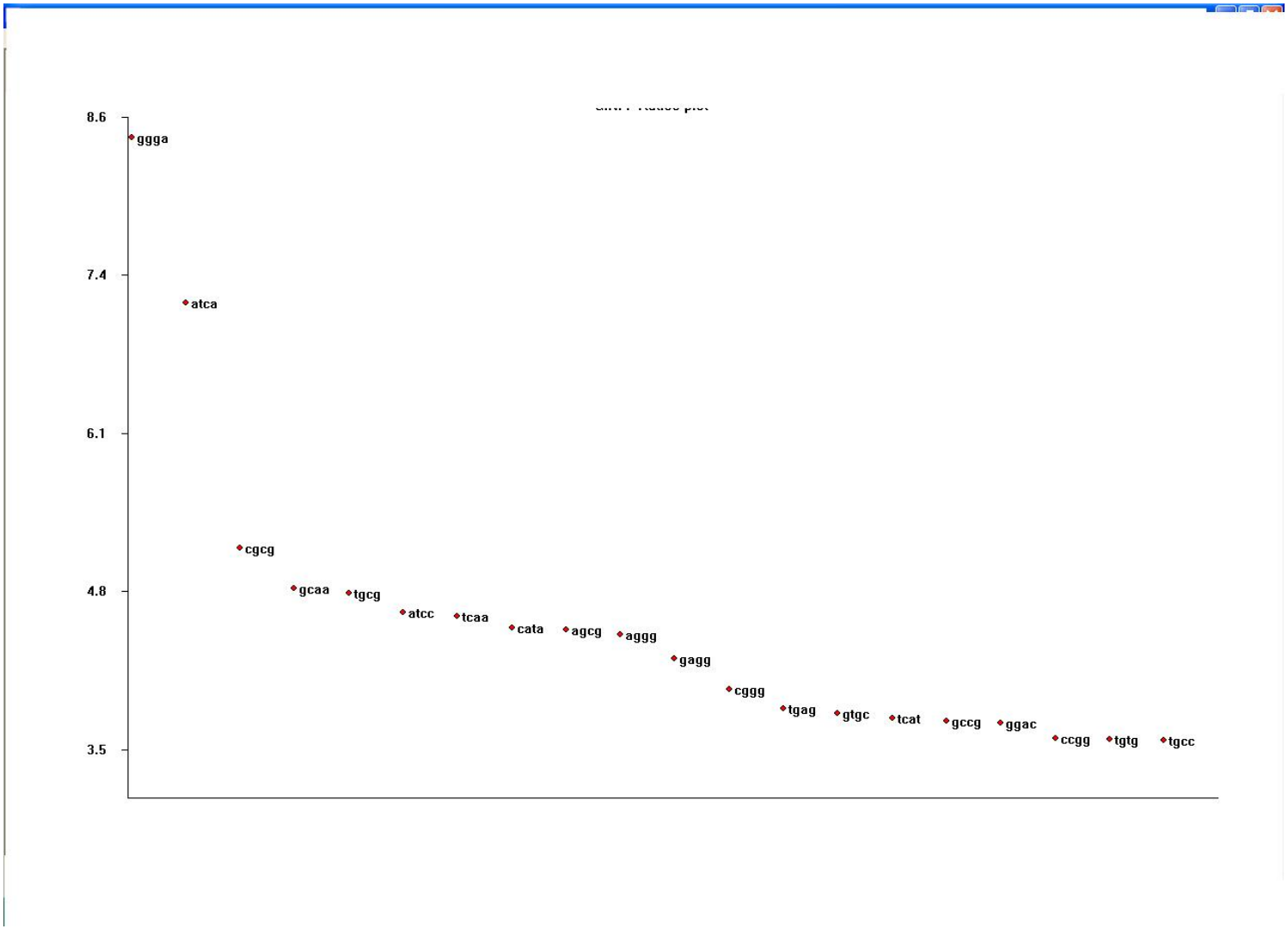
Result of cluster Analysis



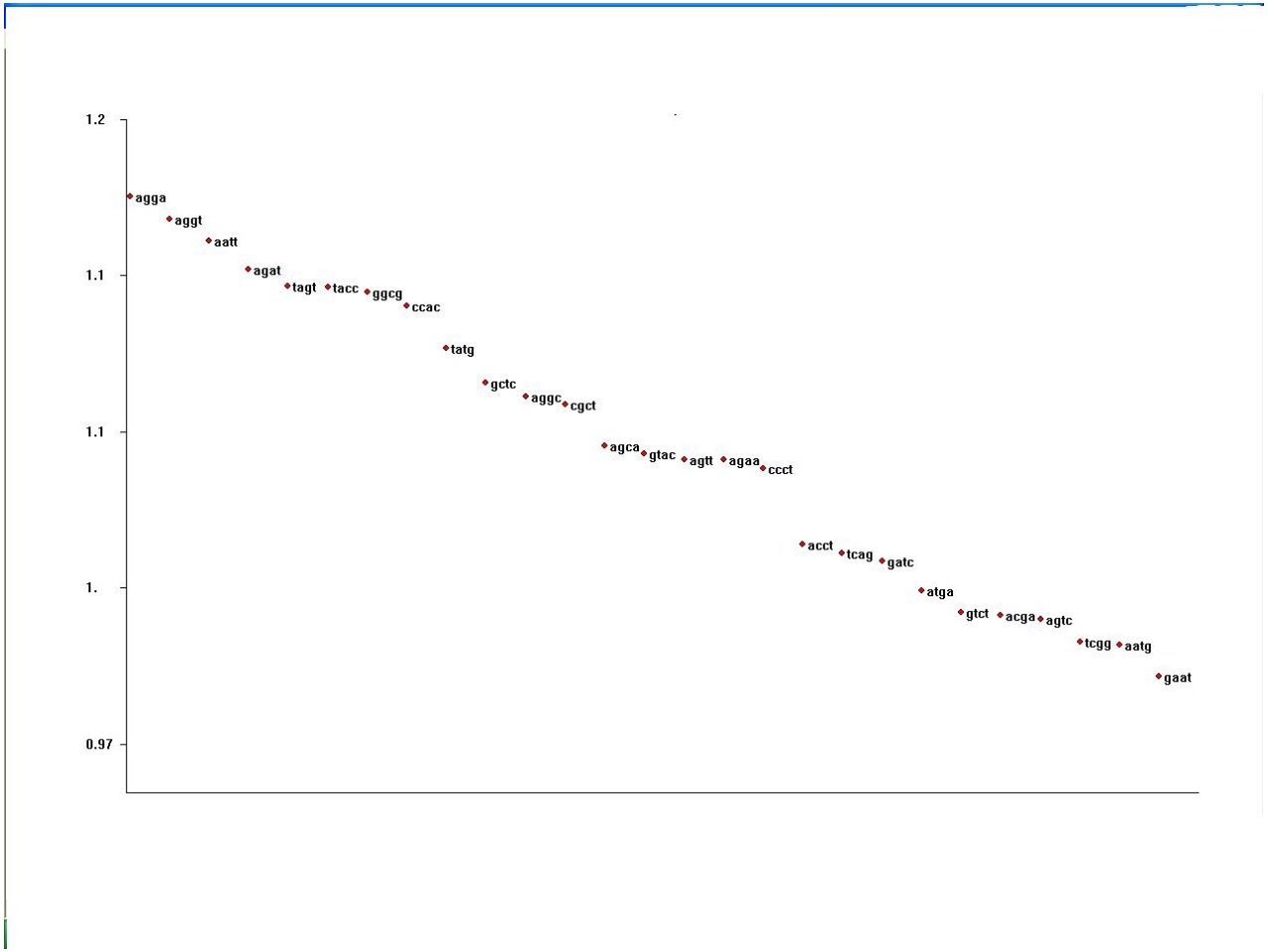
Identification of Informative Oigonucleotides

- The sample contains mammals, birds and standard fish, which form three distinct groups.
- Compute within group and between group variations of each oligonucleotide frequency.
- Compute the ratio of those two variations for each oligonucleotide.
- Plot the ratios for different oligonucleotides.

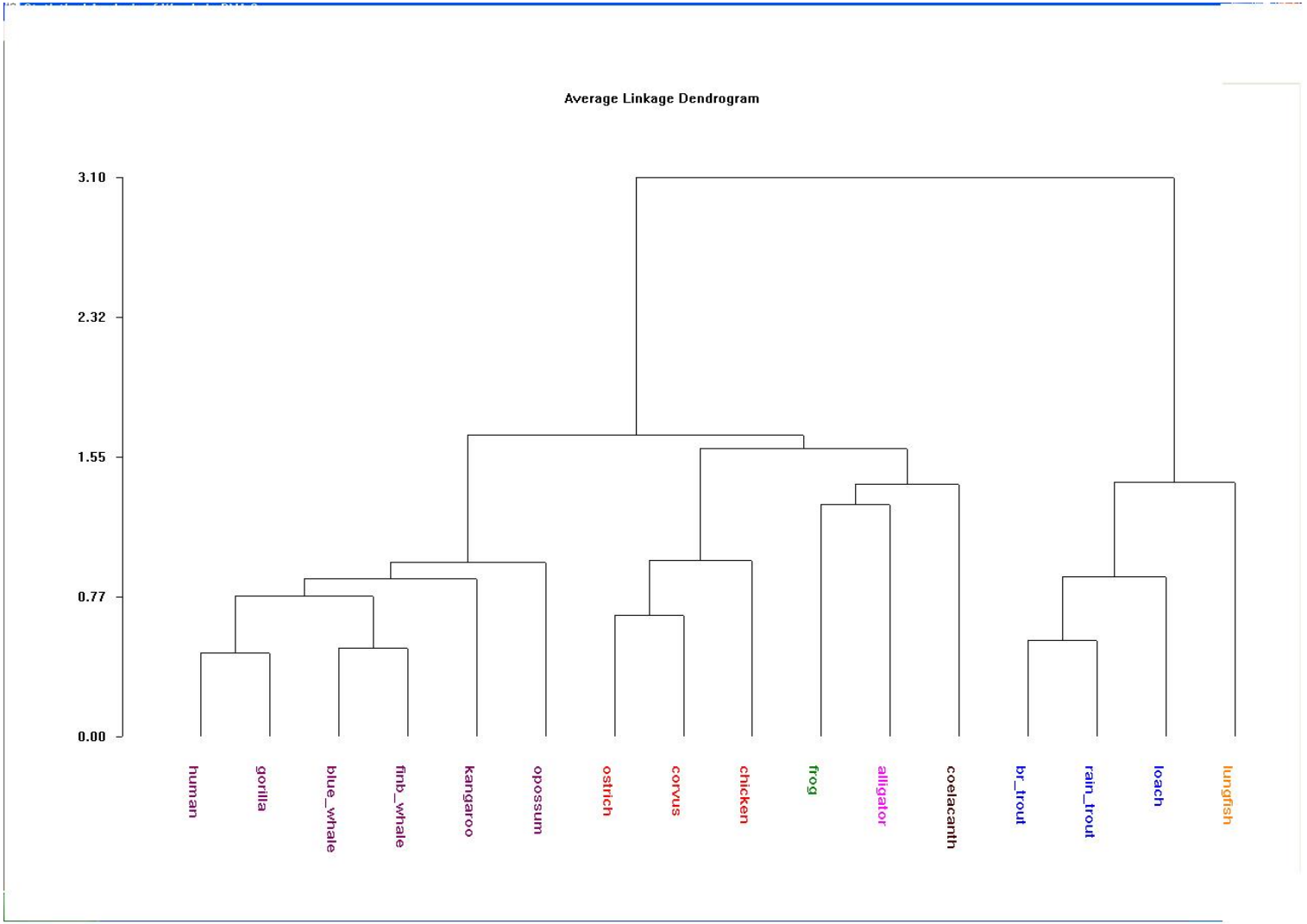
Oligonucleotides with Large Ratios



Oligonucleotides with Small Ratios



Refined Cluster Analysis



Identification of Genomic Islands

- Genomic Islands (GIs) are those parts of a bacterial genome that have originated through horizontal gene transfer.
- In pathogenic bacteria some of the GIs are often associated with pathogenic activities.
- GIs may also contain genes that are responsible for symbiotic activities and may help the bacteria in the process of adaptation.

Since GIs are acquired by the bacterial genome by horizontal transfer, one may expect that it will have oligonucleotide composition different from that in other parts of the genome.

Fig.1

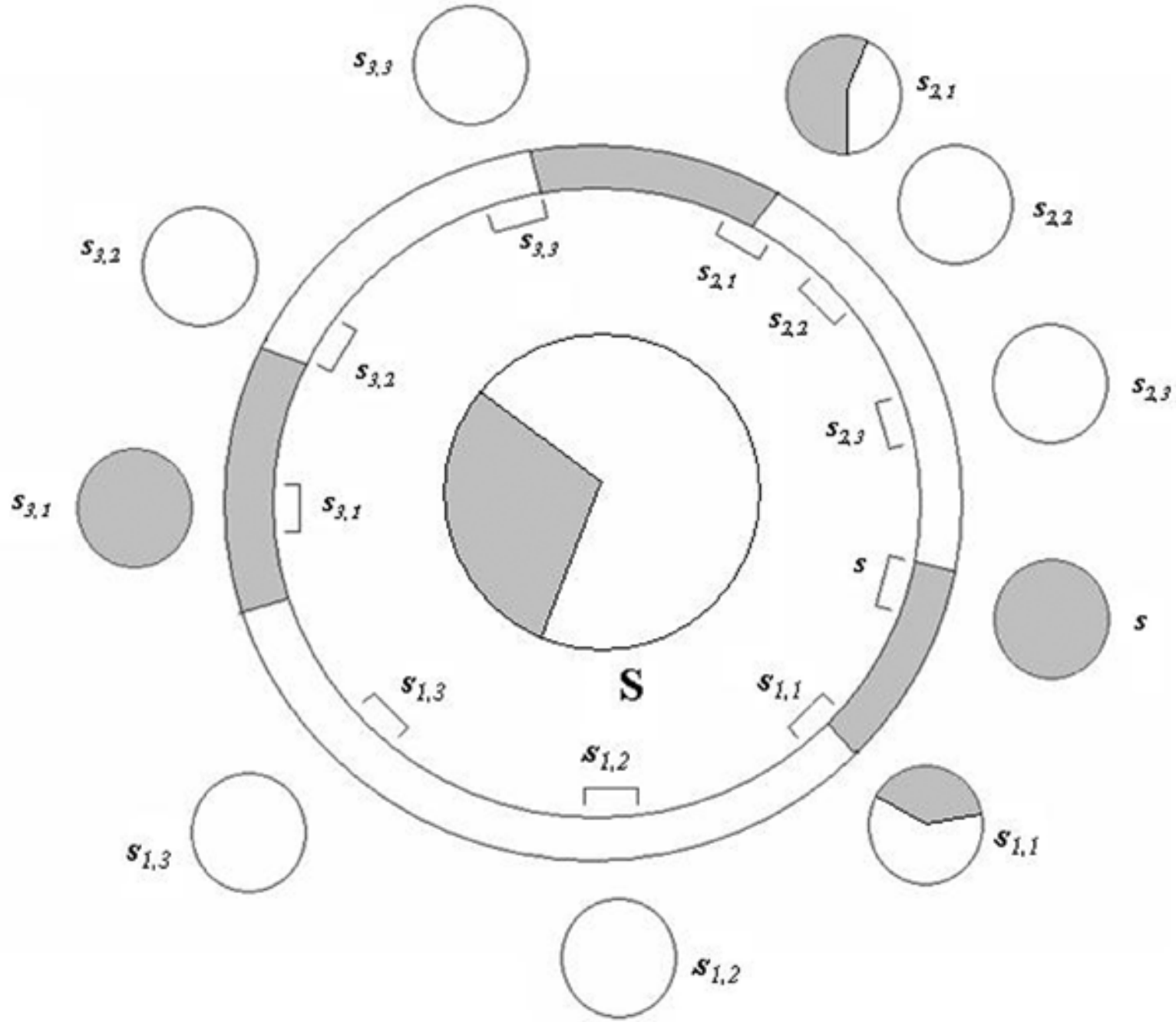


Fig.3A

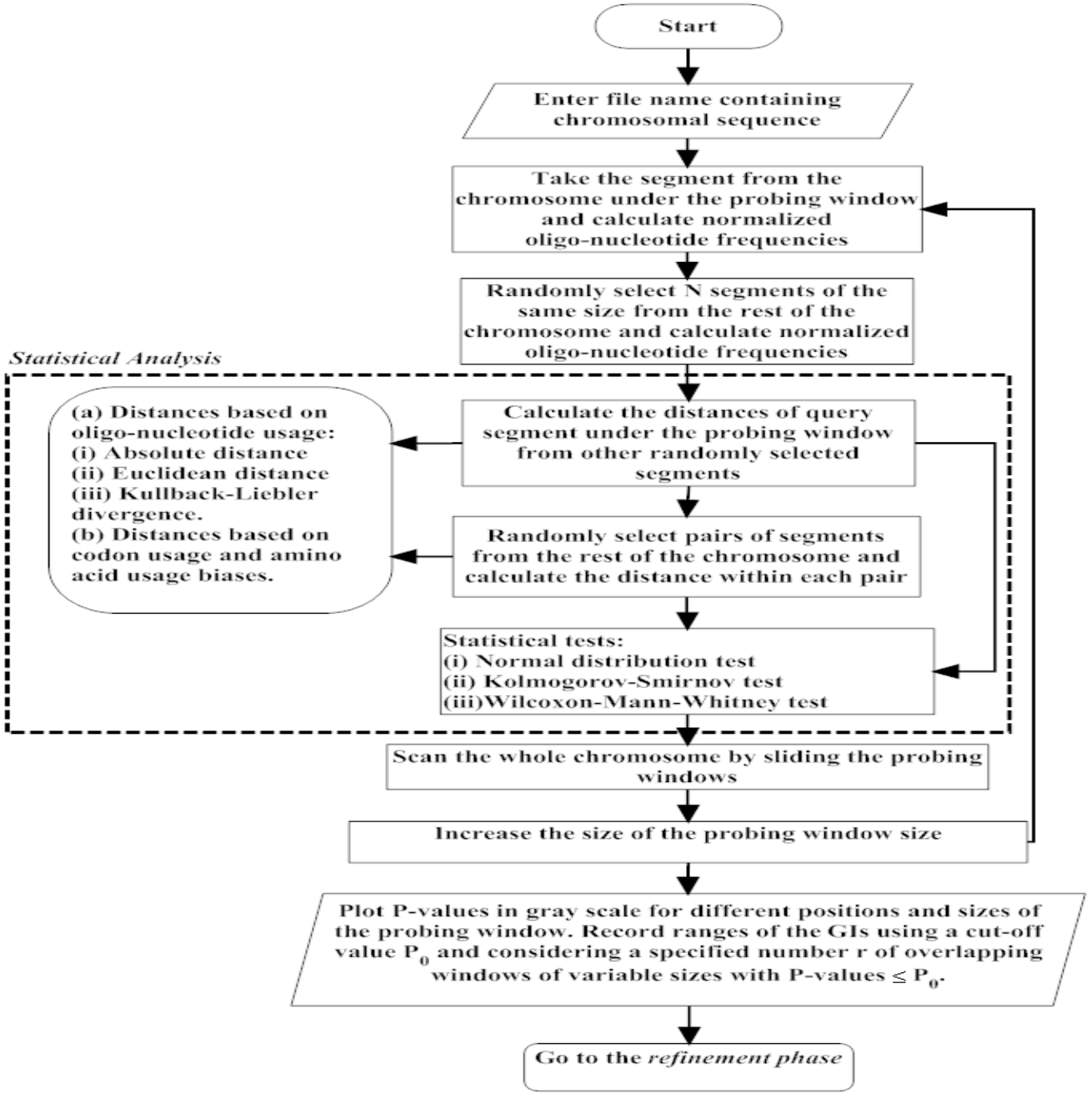
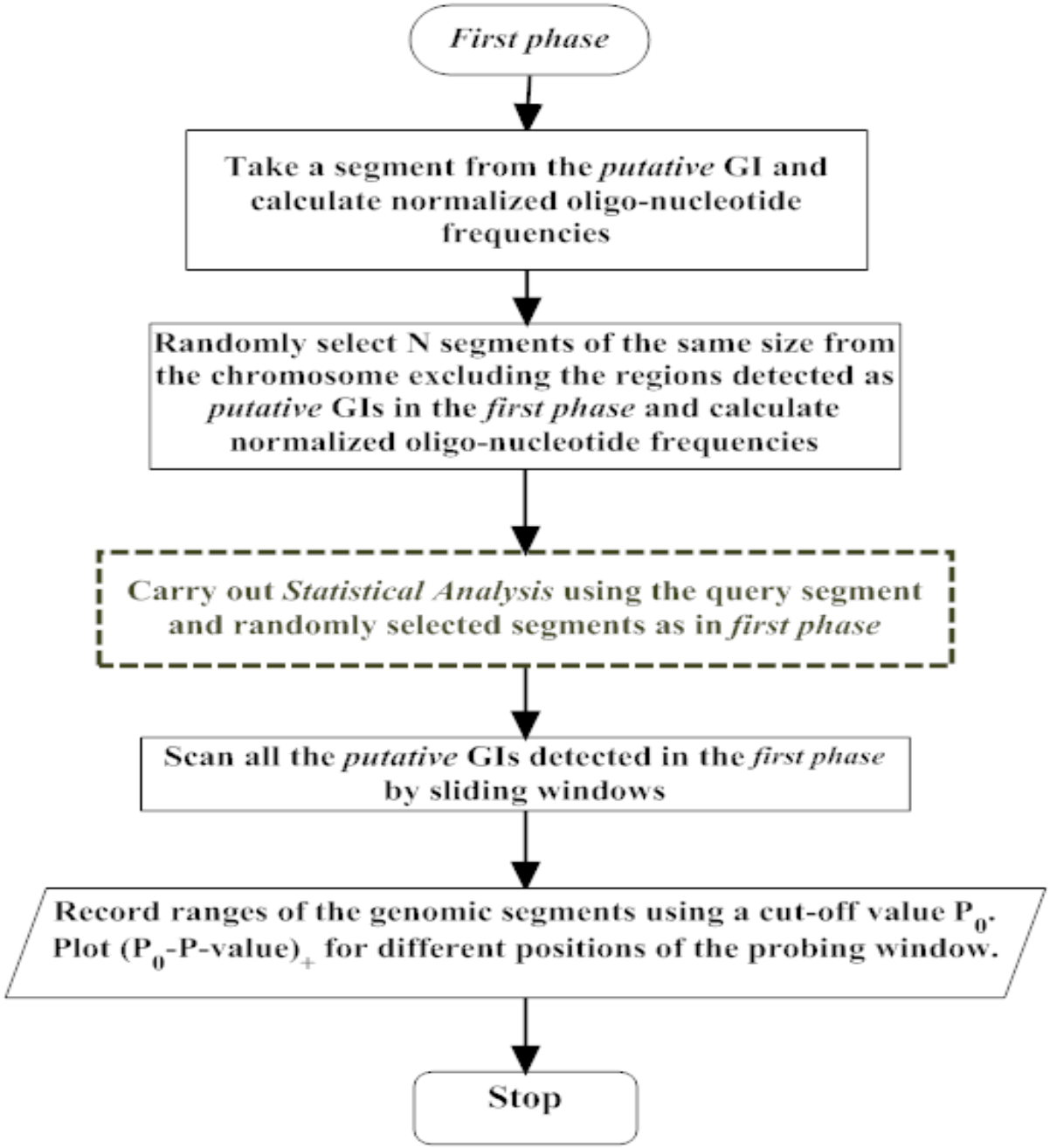

Start
Enter file name containing chromosomal sequence
Take the segment from the chromosome under the probing window and calculate normalized oligo-nucleotide frequencies
Randomly select N segments of the same size from the rest of the chromosome and calculate normalized oligo-nucleotide frequencies
Statistical Analysis
(a) Distances based on oligo-nucleotide usage:
(i) Absolute distance
(ii) Euclidean distance
(iii) Kullback-Liebler divergence.
(b) Distances based on codon usage and amino acid usage biases.
Calculate the distances of query segment under the probing window from other randomly selected segments
Randomly select pairs of segments from the rest of the chromosome and calculate the distance within each pair
Statistical tests:
(i) Normal distribution test
(ii) Kolmogorov-Smirnov test
(iii)Wilcoxon-Mann-Whitney test
Scan the whole chromosome by sliding the probing windows
Increase the size of the probing window size
Plot P-values in gray scale for different positions and sizes of the probing window. Record ranges of the GIs using a cut-off value P_0 and considering a specified number r of overlapping windows of variable sizes with P-values $\leq P_0$.
Go to the *refinement phase*

Fig.3B


First phase
Take a segment from the putative GI and calculate normalized oligo-nucleotide frequencies
Randomly select N segments of the same size from the chromosome excluding the regions detected as putative GIs in the first phase and calculate normalized oligo-nucleotide frequencies
Carry out Statistical Analysis using the query segment and randomly selected segments as in first phase
Scan all the putative GIs detected in the first phase by sliding windows
Record ranges of the genomic segments using a cut-off value P_0. Plot $(P_0\text{-P-value})_+$ for different positions of the probing window.
Stop

Salmonella typhi CT18

- A pathogenic bacteria causing typhoid.
- The genome is known to contain several pathogenic islands.
- Many of these islands are reported to contain genes from bacteriophages.

Fig.4A(i)

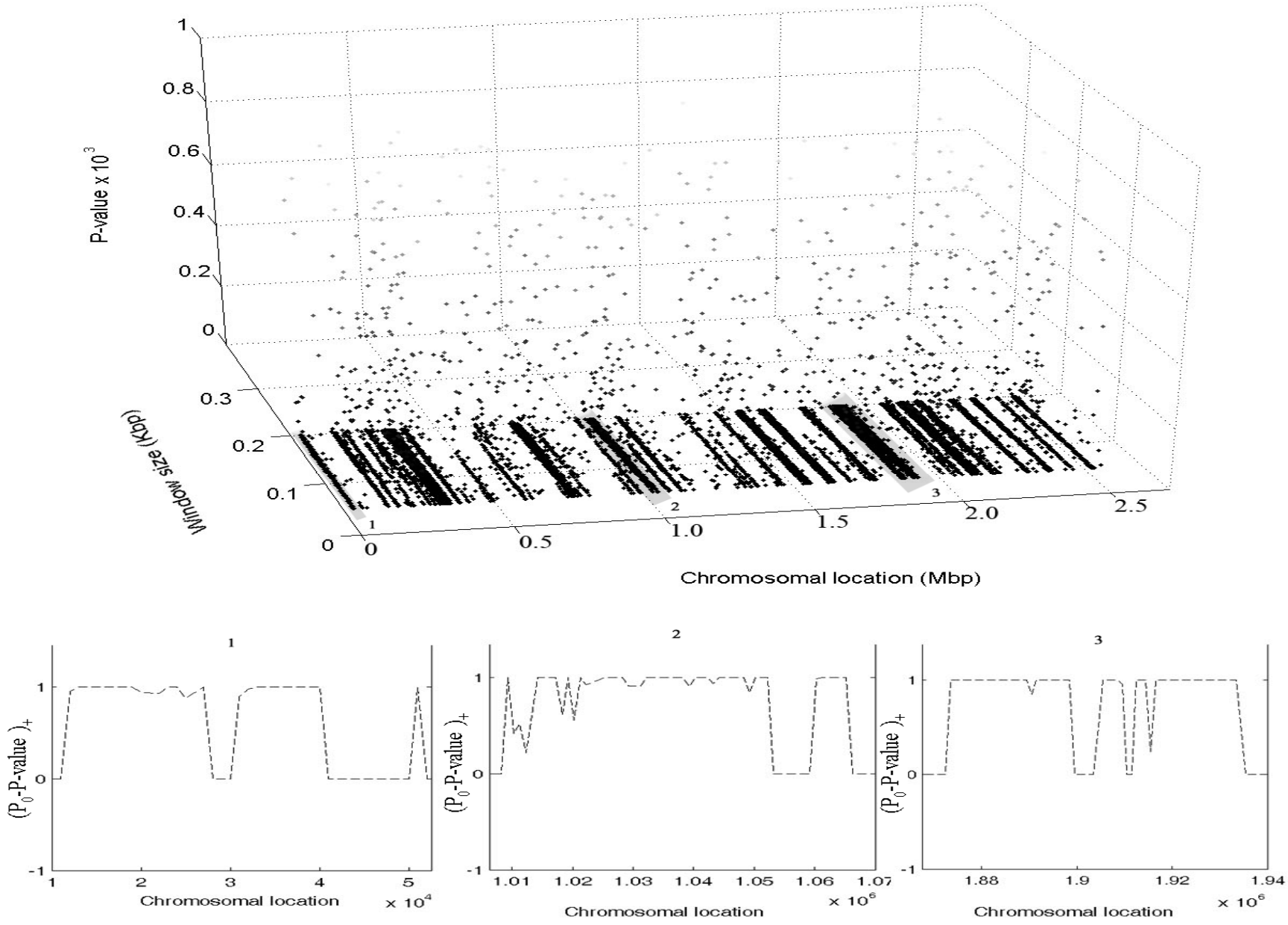
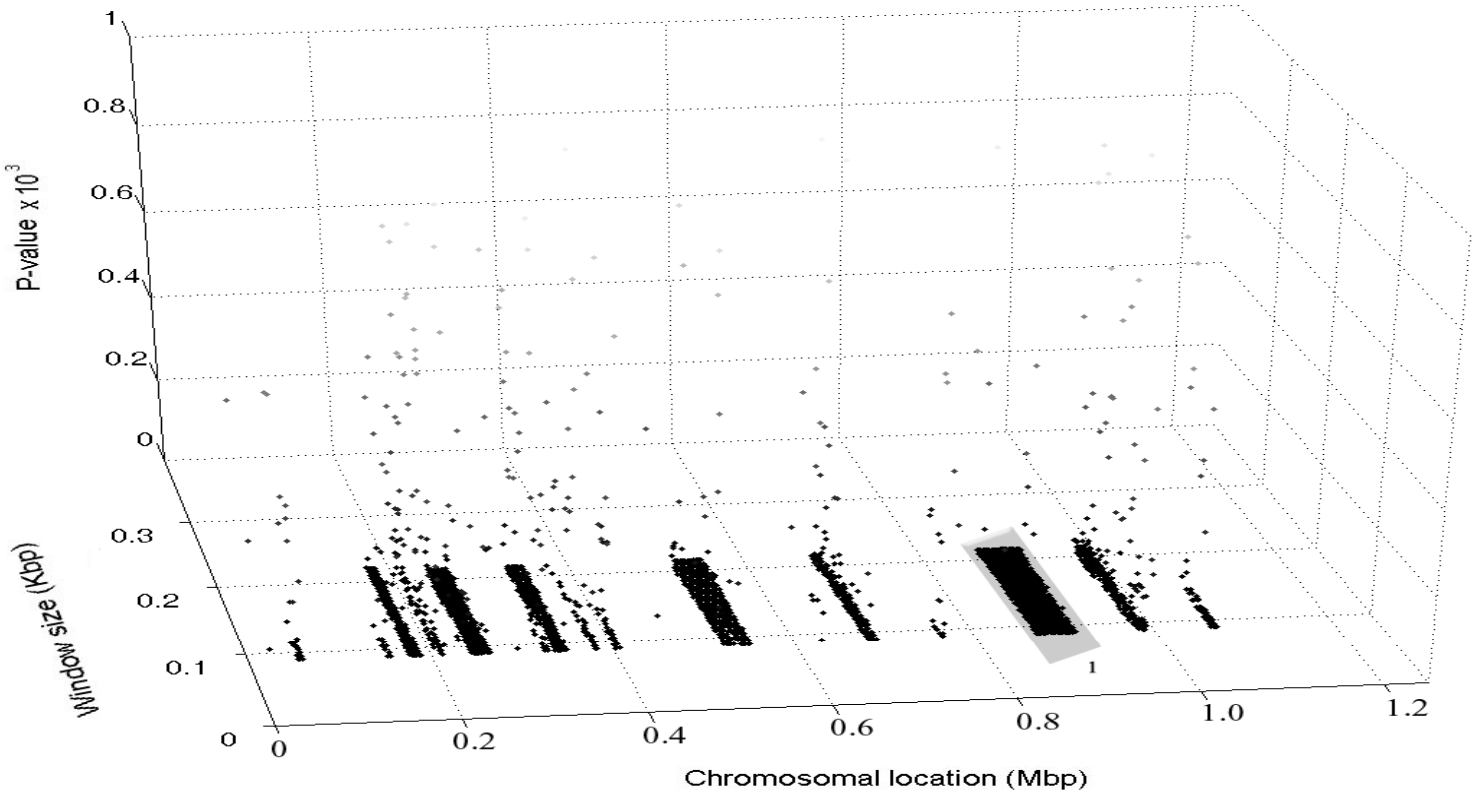
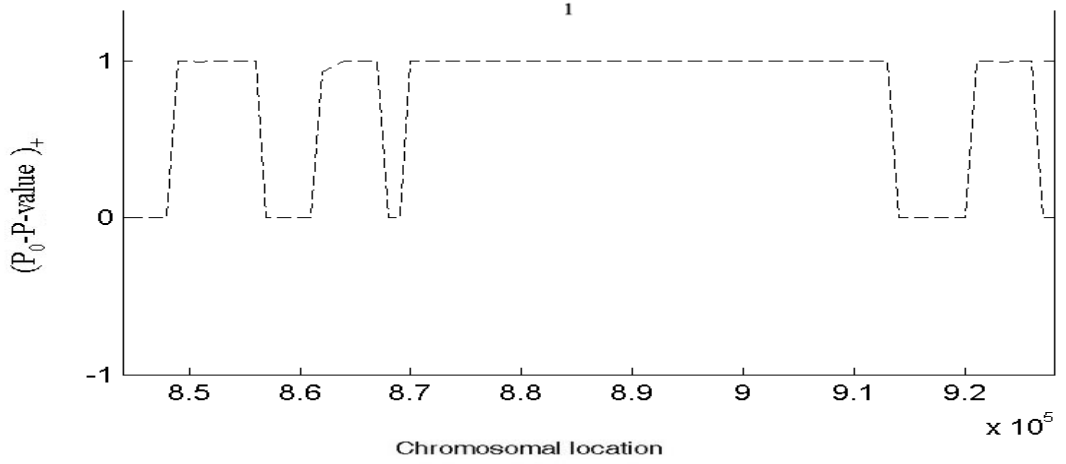

P-value x 10^3
Window size (Kbp)
Chromosomal location (Mbp)
1
2
3
(P0-P-value)+
Chromosomal location
x 10^4
x 10^6

Fig.4B(i)


P-value x 10^-3
Window size (Kbp)
Chromosomal location (Mbp)
1
0.8
0.6
0.4
0.2
0
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1.0
1.2

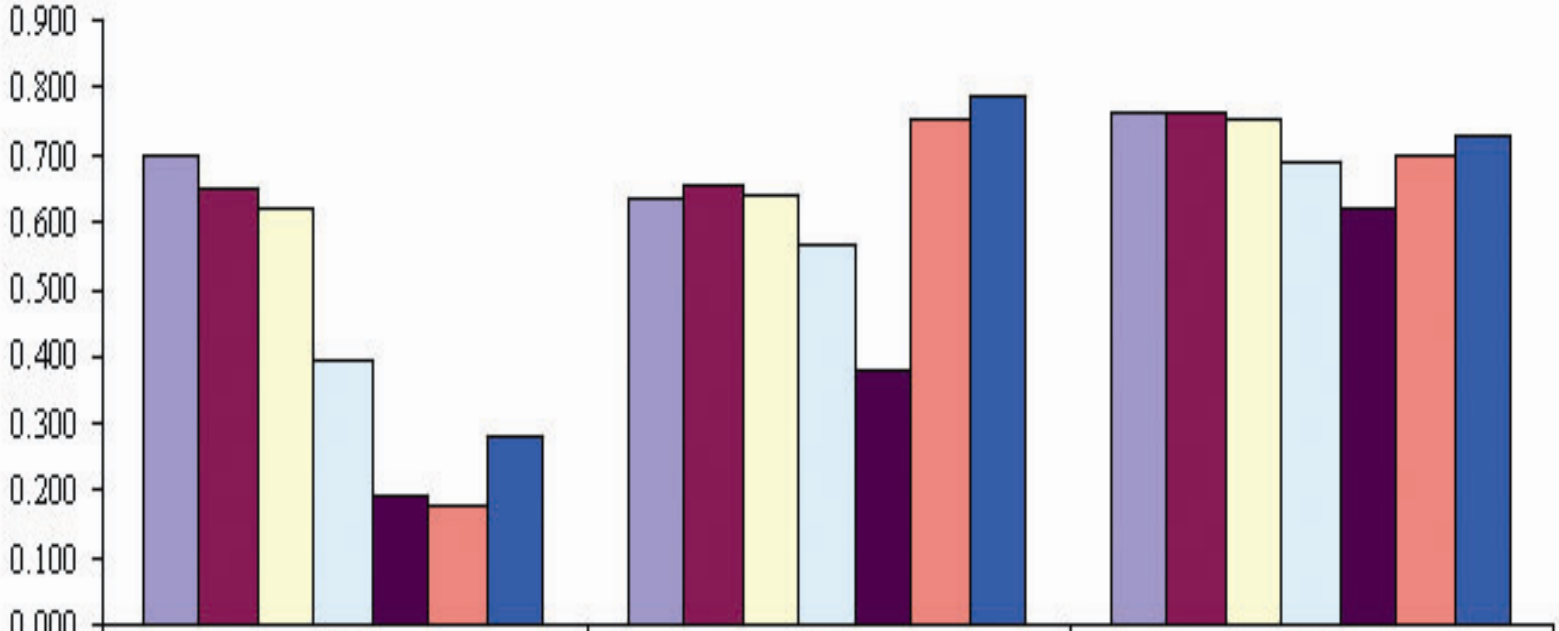

(P0-P-value)+
Chromosomal location
x 10^5
1
0
-1
8.5
8.6
8.7
8.8
8.9
9
9.1
9.2

Salmonella typhi CT18 (contd.)

- *S. typhimurium* LT2 is considered a *sister lineage* to *S. typhi* CT18 and *E. coli* K12 is considered an *outgroup* (Vernikos and Parkhill, 2006, *Bioinformatics*).
- Genes present in all three genomes form a set of *Core Genes.*
- A data set consisting of 1560 manually curated **putative horizontally transferred genes**.

Fig.5

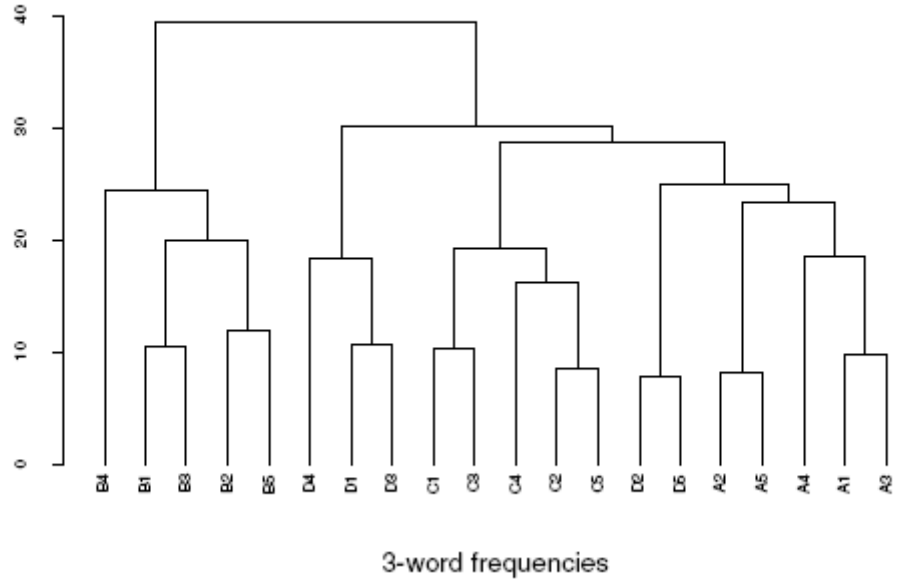
Salmonella typhi CT18



	SN	SP	AC
Design-Island	0.700	0.635	0.762
IVOM	0.649	0.653	0.764
W8	0.620	0.643	0.754
IslandPath(DINUC)	0.392	0.567	0.692
IslandPath(DIMOB)	0.193	0.380	0.619
Islander	0.176	0.755	0.701
HGT-DB	0.279	0.789	0.730

Data on bacteriophage genomes

- Four different single stranded lytic phages : *ΦX174* (host : *E. coli*), *G4* (host : *E. coli*), *F1* (host : *E. coli*) and *PF3* (host : *P. aeruginosa*)
- The sizes of these phage genomes lie between 5.3K to 6.5K nucleotide bases.
- We consider five segments of each of these phage genomes, and carry out cluster analysis based on oligonucleotide frequencies.



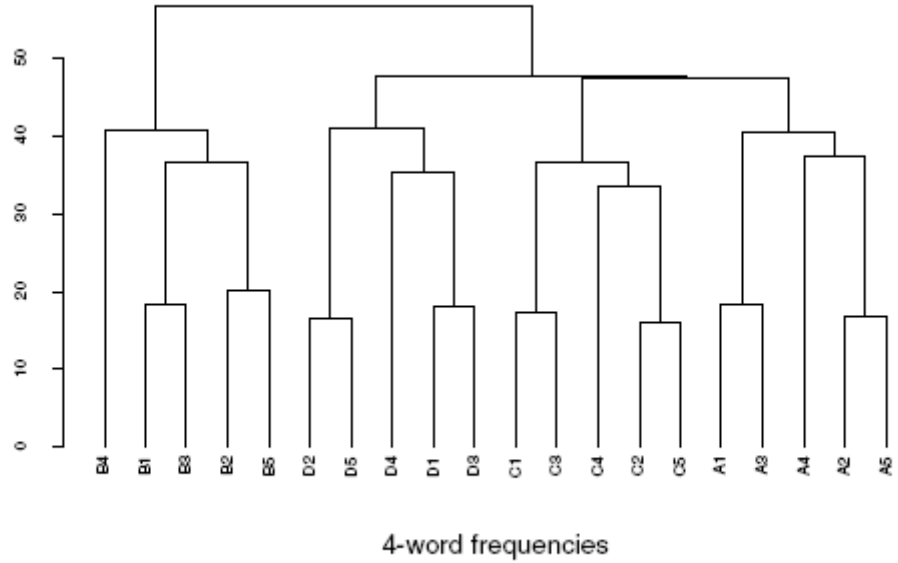
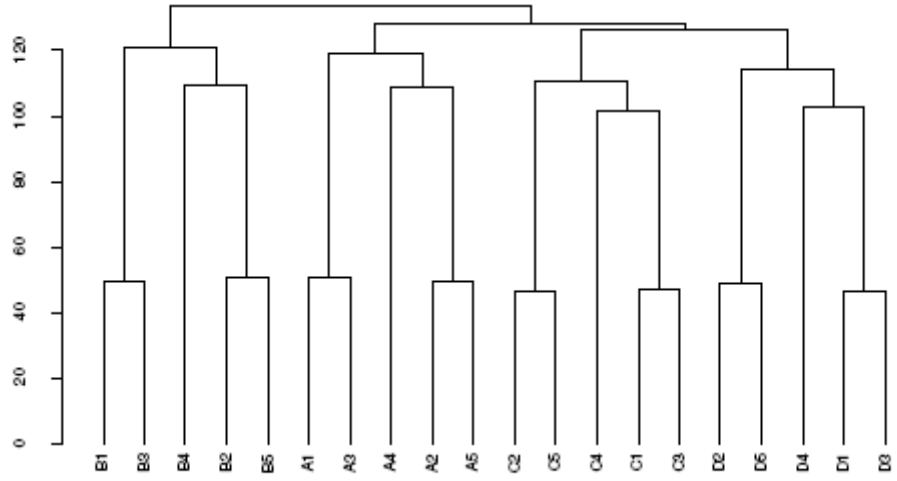
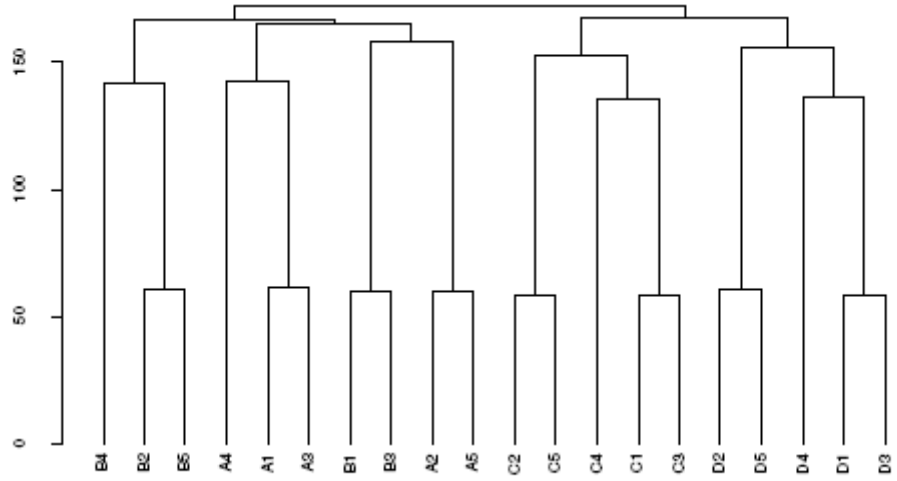


Fig. 2.5. Dendrograms for average linkage cluster analysis of fragments of phage genomes.



Dendrogram for 6-word frequencies



Dendrogram for 7-word frequencies

Fig. 2.6. Average linkage cluster analysis of fragments of bacteriophage genomes.

References

- P. Chaudhuri and S. Das (*Current Science*, 2001)
- S. Basu, D. P. Burma and P. Chaudhuri (*Journal of Mathematical Biology*, 2003)
- G. S. Vernikos and J. Parkhill (*Bioinformatics*, 2006)
- S. Nag, R. Chatterjee, K. Chaudhuri and P.Chaudhuri (*Sadhana*, 2006)
- R. Chatterjee, K. Chaudhuri and P. Chaudhuri (*BMC Genomics*, 2008)