Genome Clustering Based on Fourier Power Spectrum

Presented to :

Prof. Dr. Sanjeev Narayan Sharma

Presented By :

SIMPLE KUMARI(21BME053),

ASHISH KUMAR SONI(21BEC028),

MANISH KUMAR (21BCS128),

BHAVESH PATIDAR(21BEC034),

SARVAGYA JAIN(21BCS186)

Table of content

Genome Clustering in Bioinformatics ?

- Process of grouping and classification of genomic sequences based on their similarities and dissimilarities.

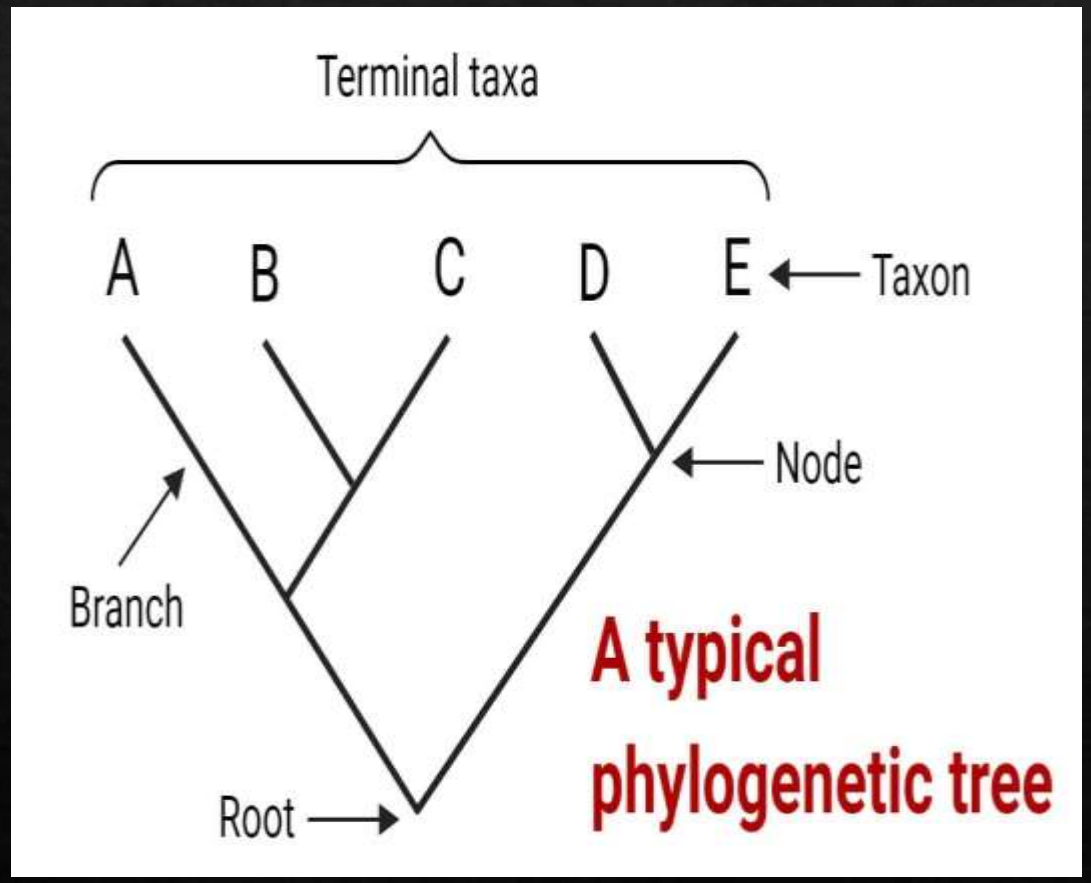
- It reduces the data complexity by organizing genomic sequence into clusters which helps in comparative analysis among the different generations.

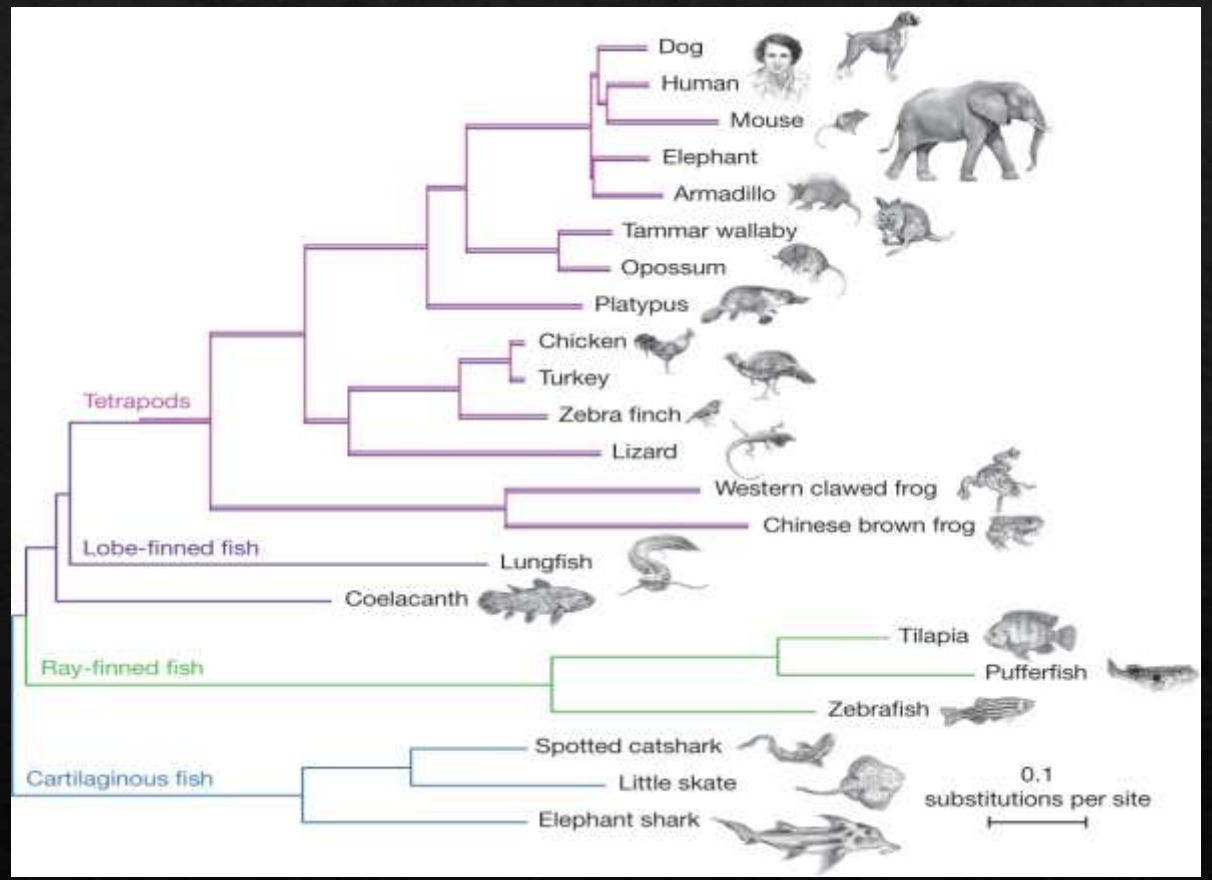
- It helps to **Discover Evolutionary Patterns** Uncover **similarities** across species or individuals, revealing evolutionary relationships and shared genetic traits.

Phylogenecy in Genome Clustering

- Based on Genome clustering determining how genetic material has **evolved** over time and how different species are **related** to each othe
- **Evolutionary Relationships**: helps in constructing evolutionary trees
- **Identification of Homologous Genes**: we can find the relationship between species which share common ancestors.
- **Divergence and Convergence**: by analysing branching patterns of phylogenetic tree we can also find the times at which species diverged from a common ancesto.

Phylogenetic Tree





- **The images representing how the different species are interrelated with their common ancestors And how the evalution in species undergoes with the passage of time from their common ancestors**.

Algorithms used for generating Genome Clustering

- There are many algorithms used for genomic clustering like:

1. K-mer Clustering
2. multiple sequence alignment
3. nucleotide-based Fourier power spectrum (PS)
4. cumulative fourier power spectrum
5. Hierarchical Clustering
6. Mean Shift Clustering:
7. Self-Organizing Maps
8. Agglomerative Clustering

- We will focus on first four algorithms

K-Mer Algorithm

- **breaking down** a longer DNA sequence into **overlapping** subsequences of a fixed length (k). The **frequency** of each unique k-mer is **counted** within the genomic sequence. This results in a frequency vector or histogram representing the distribution of k-mers. based on the similarity of their k-mer representations. **Similar sequences are expected to have similar k-mer profiles**. they are having higher computational **time**.

Multiple Sequence allignment

- in this technique we take **3 or more** sequences {dna ,rna , protein}and allign them to analyse using mathematical computation , its time complexity **depends upon the length** of sequence, for a very large genomic sequence it will take too much time to cluster the genomic sequence

Nucleotide-based Fourier power spectrum (PS) :

- DFT and the moment vectors are used
- Mathematical representation : use indicator function
- Indicator function consist **4 separate sequences** which show the **distribution** of the four **nucleotides** respectively.
- dft is applied on these mathematical numeric vectors to observe **analyse the frequency** properties of these sequences.
- with the help of moment vectors the sequence is transformed into points in space.
- Less storage is taken by this algo. As compared to other.

- Limitation:in this algo. There is possibility of information loss during one one mapping of numeric vectors. That's why the concept of CPF is introduced

Cumulative fourier power spectrum for Genome Clustering why is it considered to be preferable ?

- **recovery of power spectra** from DNA sequences possible, which PS cannot achieve.
- No possibility of **information loss,**ability to **finish calculation in reasonable with time**
- **Reduces the complexity** behind the analysing the sequence and representing it in the numeric vector form

Role of cumulative Fourier power spectrum in genomic clustering

- **Detection of Periodic Patterns:** identify similarity dis similarity
- **Frequency-Based Representation**: unique characterization of genomic sequences
- **Identification of Hidden Patterns:** patterns which are difficult to identify , can be detected based on their corresponding frequency
- **Effective for Non-Linear Structures:** complex, non-linear patterns within DNA sequences. By analyzing frequencies rather than linear sequences,

Stepwise process for genome clustering using the Cumulative Fourier Power Spectrum:

1. Data Preparation:

1. Obtain DNA sequences to be analyzed.
2. Break sequences into fixed-length segments (windows).

2. Fourier Transform:

1. Apply the Fourier Transform to each DNA segment.
2. Calculate the power spectrum for each segment.

3. Cumulative Power Spectrum:

1. Sum the power spectra across segments for each sequence.
2. Create a cumulative power spectrum profile for each sequence.

Stepwise process for genome clustering using the Cumulative Fourier Power Spectrum:

4. Similarity Measurement:

Compare cumulative power spectrum profiles among sequences.
Use similarity metrics (correlation) to quantify similarities/differences.

5. Clustering:

Apply clustering algorithms based on the similarity measures.
Group sequences with similar CPFS profiles into clusters.

6. Validation and Analysis:

1. Validate clusters for coherence and significance
2. Analyze clusters for functional or structural similarities among genes or sequences.

Observation

- "The numbers we get from the Cumulative Fourier Power Spectrum show patterns in DNA. **Similar numbers** mean **similar DNA structures**. Clustering these numbers groups together DNAs that are alike, helping us see how genes might be related or work similarly.

Conclusion:

- "In conclusion, using Cumulative Fourier Power Spectrum (CPS) for genomic clustering helped us find common patterns in the DNA. This gives us a better understanding of how different genetic sequences are organized and grouped together based on their shared features.

THANKYOU