AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1

Chaos Game Representation of DNA Sequences

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

Contents

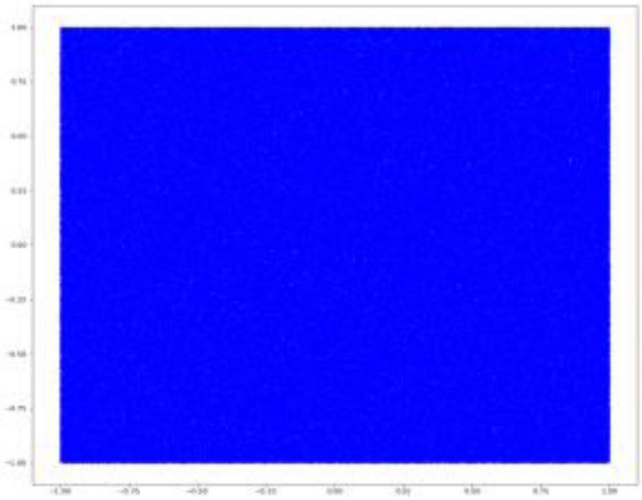
- CGR for sequence comparison
- CGR for COVID-19
- Frequency Chaos Game Representation(FCGR)
- FCGR probability matrix
- CGR Centroid method
- FCGR Probability Matrix distance
- CGR Centroid distance
- Hierarchical Agglomerative Clustering (HAC) analysis.

CGR for sequence comparison

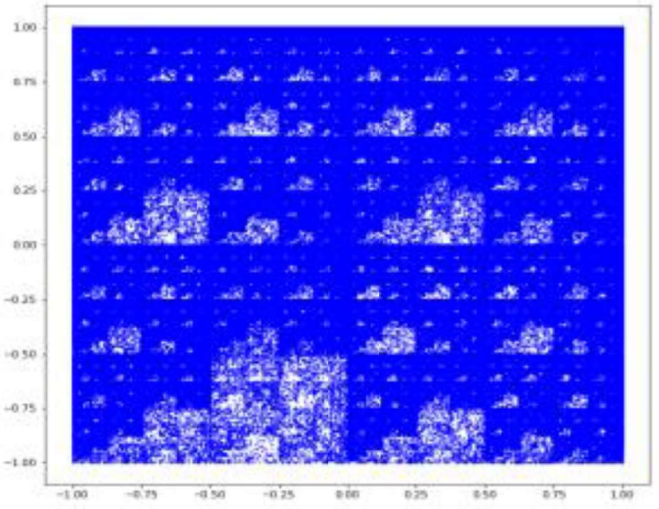
- Different and unique patterns in the CGR emerge from different sequences.
- CGR patterns of DNA segments have been proposed as a method for the classification and Identification of genomic sequences.
- CGR of DNA sequences coming from various species show pattern such as squares, parallel lines, rectangles, triangles, and also complex fractal patterns.
- Random sequence of DNA will not generate such a pattern
- CGR of same species (eg., virus genomes) look very similar visually.

CGR for sequence comparison

- While CGR can create fractals, it is not guaranteed to occur with every organism.
- Patterns differ between organisms as well
- Patterns formed by the CGR of prokaryotes differs from that of eukaryotes due to less variation in DNA



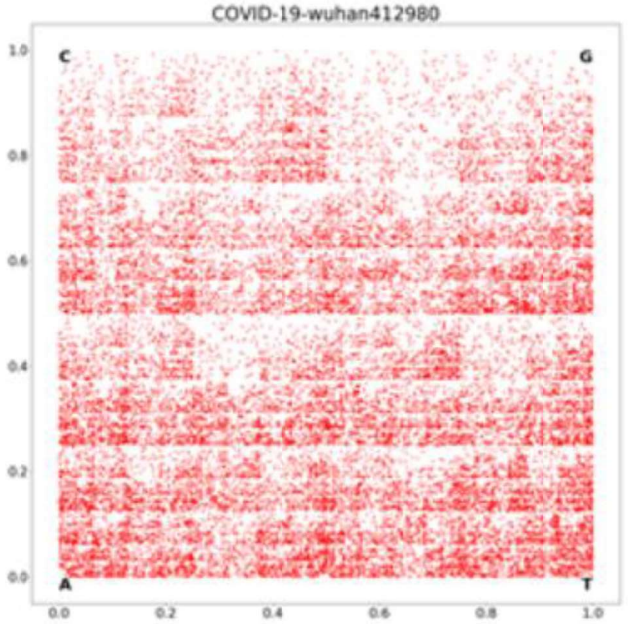
(a) Random Sequence CGR

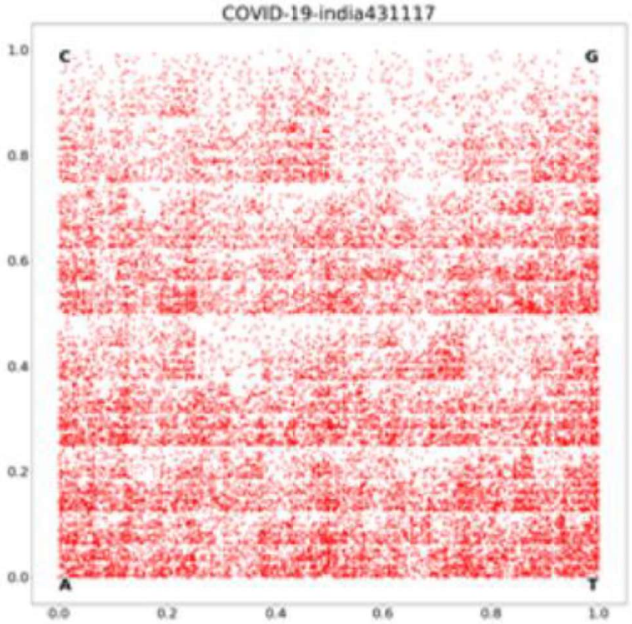


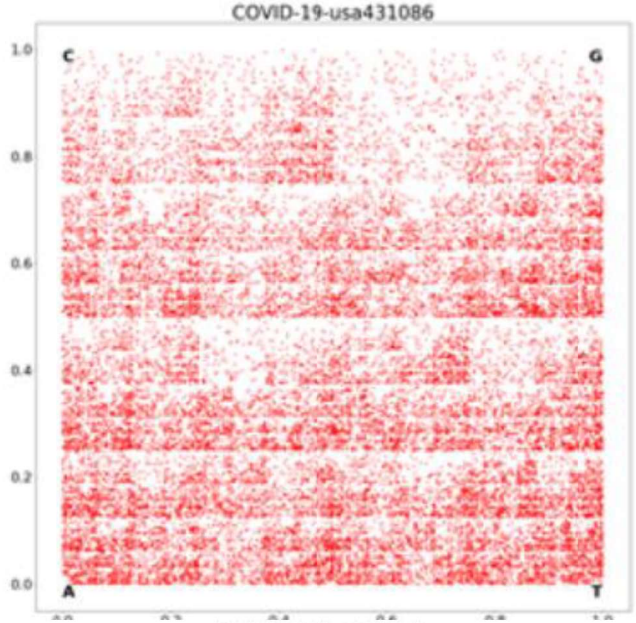
(b) Human Chromosome 21 CGR

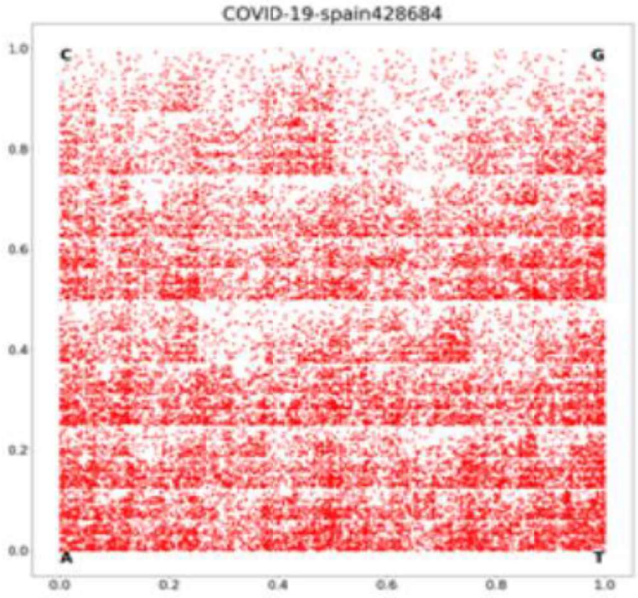
https://www.scirp.org/journal/paperinformation?paperid=101153

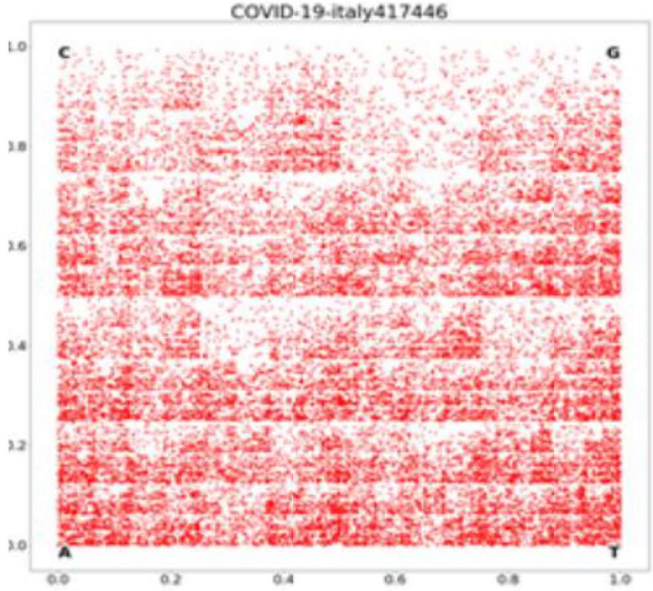
CGR for COVID-19

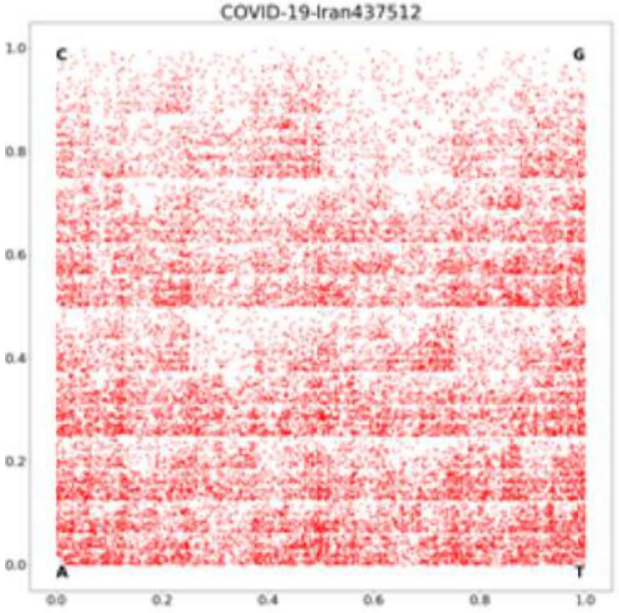












https://www.scirp.org/journal/paperinformation?paperid=101153

AMRITA VISHWA VIDYAPEETHAM

CGR for sequence comparison

- Step1: Create CGR of Sequences
- Step2: Compute the FCGR Probability Matrix
- Step3: Calculate the centroid of each grid
- Step4: Measure the distance between two CGR images
  - FCGR Probability Matrix distance
  - CGR Centroid distance
- Step5: Create the dendrogram of the distance matrices using Hierarchical Agglomerative Clustering (HAC) analysis.

Frequency Chaos Game Representation(FCGR)

- A k-th order FCGR is a $2^k \times 2^k$ matrix that can be constructed by dividing the CGR plot into a $2^k \times 2^k$ grid, and defining the element |aij| as the number of points that are situated in the corresponding grid square.

First-order FCGR

$$FCGR_1(s) = \begin{pmatrix} N_C & N_G \\ N_A & N_T \end{pmatrix}$$

Second-order FCGR

$$FCGR_2(s) = \begin{pmatrix} N_{CC} & N_{GC} & N_{CG} & N_{GG} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AA} & N_{TA} & N_{AT} & N_{TT} \end{pmatrix}.$$

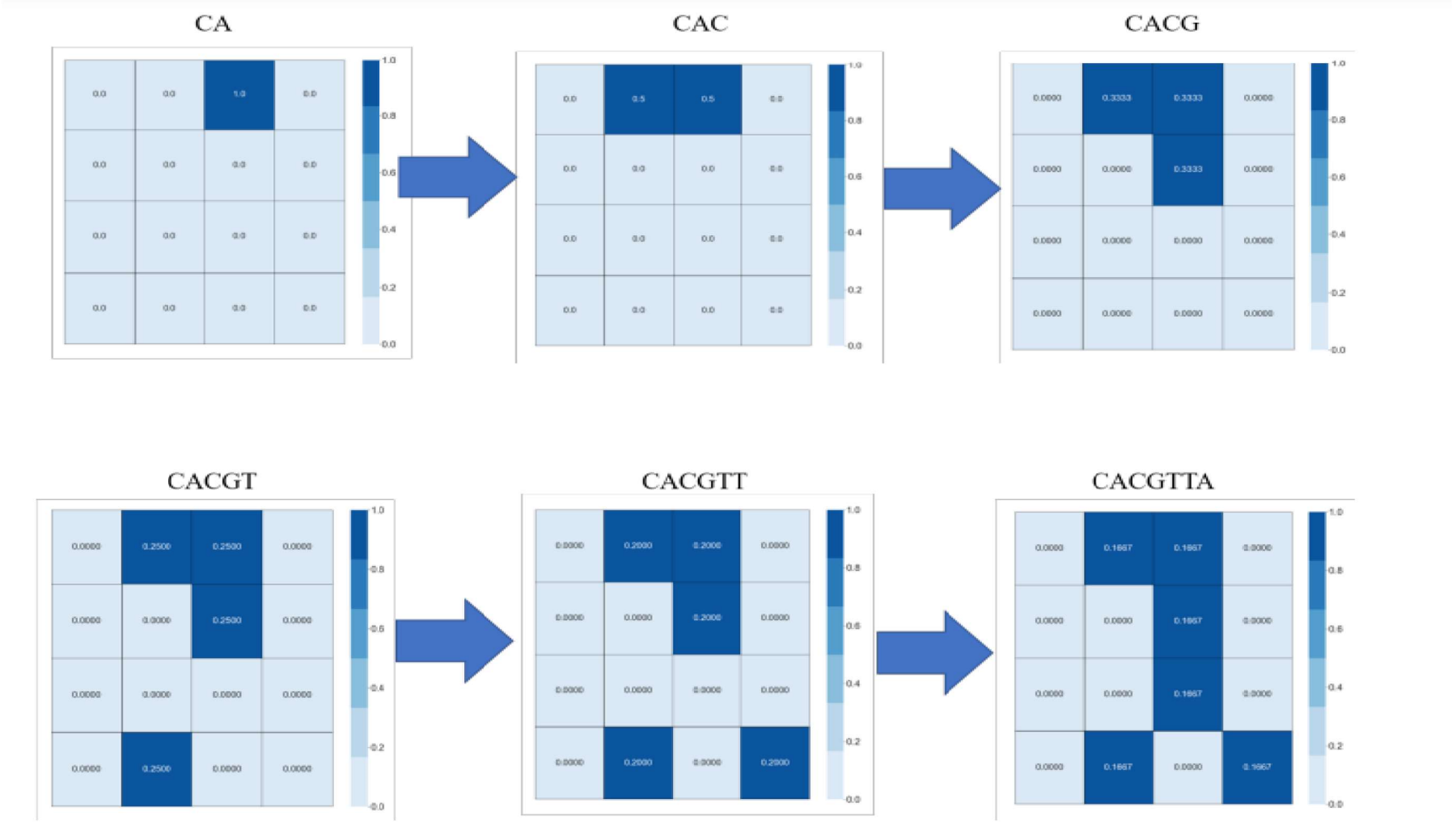
- $(k+1)^{th}$ order $FCGR_{k+1}(s)$ can be obtained by replacing each element N_X in $FCGR_k(s)$ with four elements $\begin{pmatrix} N_{CX} & N_{GX} \\ N_{AX} & N_{TX} \end{pmatrix}$ where X is a sequence of length k over the alphabet {A, C, G, T}

FCGR probability matrix

- If the generated square image has a size of $2^k \times 2^k$ pixels, then every pixel represents a distinct k-mer.
- For each $k \geq 1$, define a probability matrix of $FCGR_k(s)$ by taking each entry of $FCGR_k(s)$ dividing by the total counts of all k-mers.
- Probability matrix can be interpreted as probability of distribution.

$$(P_{ij}), 1 \leq i, j \leq 2^k. \qquad \sum_{i,j} P_{ij} = 1$$

FCGR probability matrix



An example of FCGR representation of the sequence CACGTTA

CGR Centroid method

- Centroid of a cluster of points is the mean of those points
- Let (x_k,y_k) be the coordinates of a point in the CGR. We define the centroid in each of the $2^k \times 2^k$ grid as follows:

$$C_{ij} = \left(\frac{\sum_{k=1}^{n} x_k}{n}, \frac{\sum_{k=1}^{n} y_k}{n} \right)$$

$$1 \leq i,j \leq 2^k$$

centroid is then calculated for each grid

FCGR Probability Matrix distance

- For two FCGR probability matrices (p_{ij}) and (p'_{ij}) , The distance between the two probability matrices

$$D_{PM} = \sum_{i=1}^{2^k} \sum_{j=1}^{2^k} d_{ij}.$$

where

$$d_{ij} = \left| p_{ij} - p'_{ij} \right|.$$

CGR Centroid distance

- For two centroids $c_{ij}=(x_{ij},y_{ij})$ and $c'_{ij}=(x'_{ij},y'_{ij})$ respectively for $1 \le i,j \le 2^k$, the centroid distance is

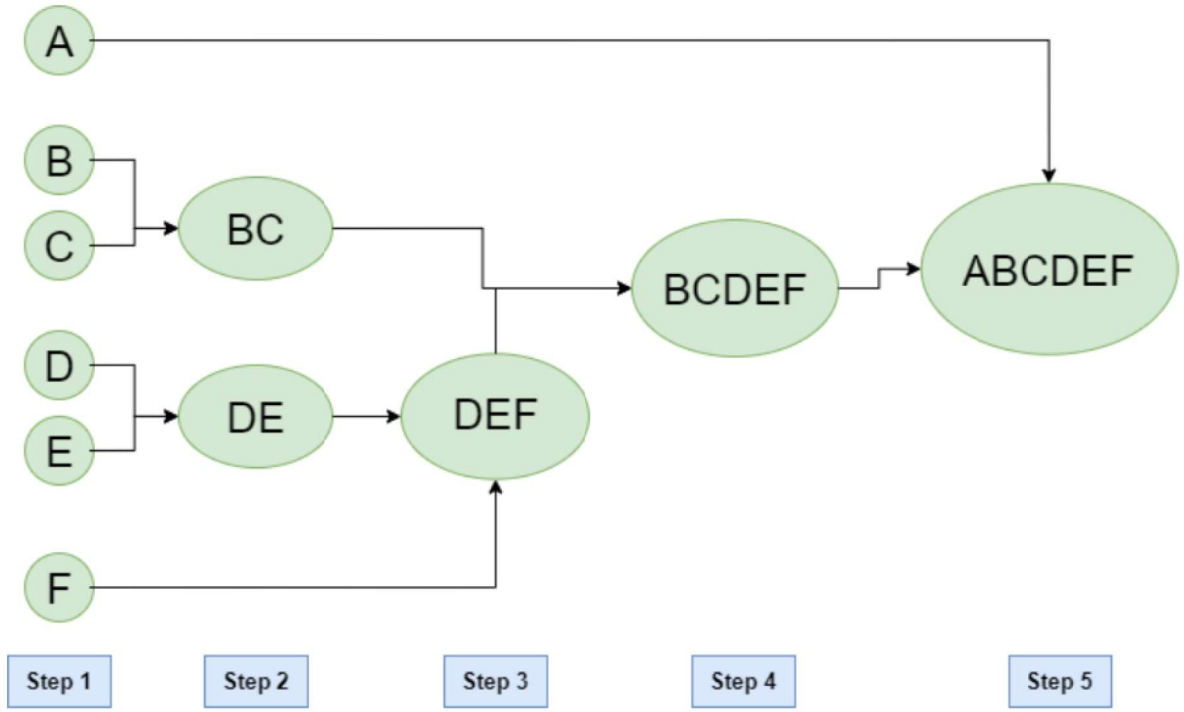
$$D_{cd} = \sum_{i=1}^{2^k} \sum_{j=1}^{2^k} d_{ij}.$$

where

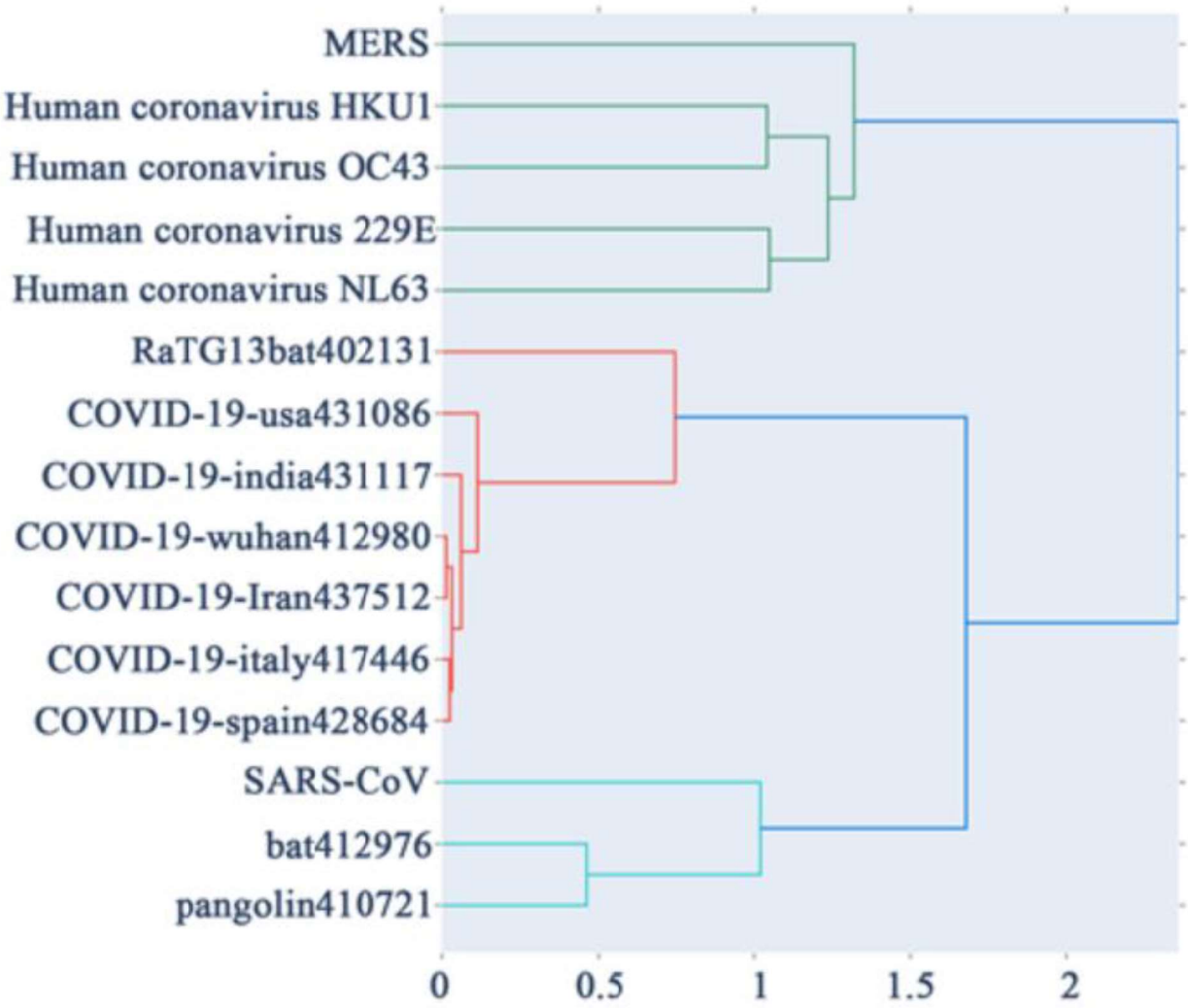
$$d_{ij} = \sqrt{(x_{ij} - x'_{ij})^2 + (y_{ij} - y'_{ij})^2}.$$

Hierarchical Agglomerative Clustering

- This clustering algorithm does not require us to prespecify the number of clusters.
- This Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all data.

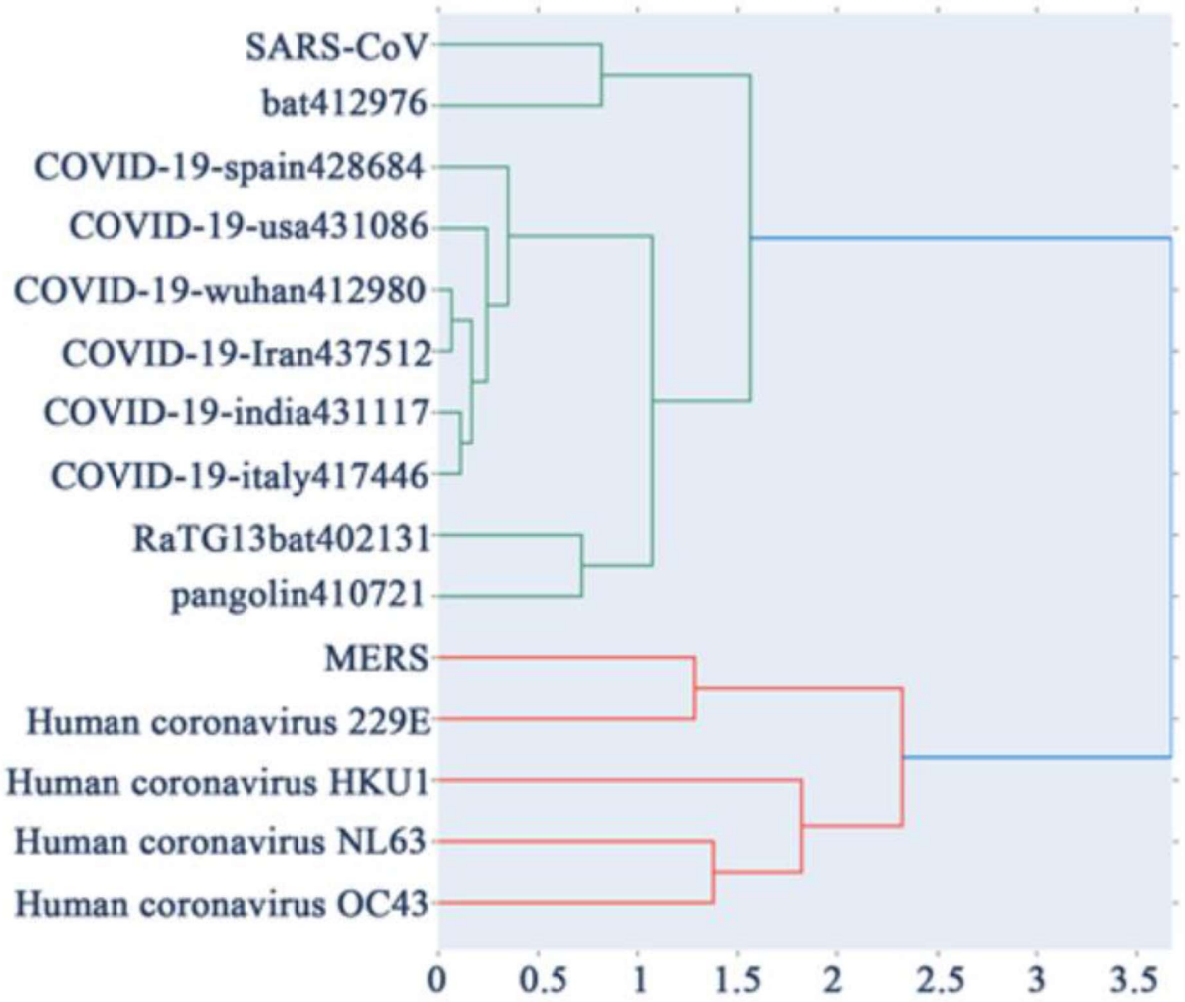


HAC Analysis



HAC phylogenetic tree using probability matrix distance.

HAC Analysis



HAC phylogenetic tree using CGR centroid distance

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1