A New Decentralized Alignment-free Visualization Tool for Analyzing the Segmented Genome of a Virus in Feature Spaces

Mosaab Daoud [1], *Benjamin Aziz* [2]

[1]8-925 Rosemont Blvd, Montreal, Quebec, Canada.
[2]School of Computing, University of Portsmouth, Portsmouth,U.K.

Abstract

In this paper, we propose a novel approach to analyze and visualize genetic variation of a segmented genome of a virus in feature spaces embedded in $\mathbb{R}^p$ and/or higher feature spaces embedded in $\mathbb{R}^{p'} \ni p' > p$. The segmented genome of a virus is considered as a heterogeneous sequence-set with different mutation rates. In this approach, the dispersion maps of a virus are computed in a decentralized manner using an arbitrary set of substrings of specific length (e.g. 2-grams). The new approach is considered as new decentralized alignment-free visualization tool. The new decentralized alignment-free visualization tool, shows effectiveness in capturing and analyzing genetic variation that causes the genetic biodiversity of the segmented gnome of a virus. The new alignment-free visualization tool is expected to be a useful tool in biomedical sector, specifically with a network of mobile laboratories.

1 Introduction

Sequence-set analysis is a developing research discipline in the area of sequence-analysis [5, 8, 7]. The sequence-set analysis is focusing on analyzing sets of sequences in data space ($\Sigma^* - \epsilon$, where Σ is an alphabet and ϵ is the empty string) or feature space ($\mathbb{R}^p$) or distance space ($\mathbb{R}^+ \cup \{0\}$) using different approaches or techniques. There are two types of sequence-sets, those types are: (i) homogeneous sequence-sets, and (ii) heterogeneous sequence-sets. A homogeneous sequence-set is defined as a set of sequences, where sequences have the same nucleotide composition and share common biological features (e.g. common ancestor). A heterogeneous sequence-set is defined as a set of sequences, where sequences have different nucleotide compositions and different biological features. In fact, there is no evidence that sequences of a heterogeneous sequence-set have a common ancestor. In addition, for each type of sequence-sets, different approaches can be implemented to analyze the extracted information from sequence-sets. Sequence-sets are embedded in data space, and therefore, to analyze sequence-sets in feature space, feature extraction techniques are required to map sequence-sets from data space to feature space.

The segmented genomes of a virus are heterogeneous sequence-sets (e.g. flu virus has eight segments, which can be encoded into 10-11 proteins, and each protein has a different function [5]). They are changing rapidly with different mutation rates. Therefore, the changes in segmented genomes of a virus have negative impacts on different life forms (e.g. human health). Monitoring the changes in segmented genomes of a virus is considered as a complicated analysis process. In this paper, we aim to propose a decentralized alignment-free visualization tool that can be effectively used in analyzing and visualizing genetic variation of a virus in different feature spaces.

The rest of this paper is organized as follows. We shall present: research problem under consideration in section 2, related research works in Section 3, new decentralized alignment-free sequence-set visualization tool in Section 4, results and discussion in Section 5, and finally, conclusions and future work in Section 6.

2 Research problem under consideration

In this paper, the research problem under consideration is a complicated problem due to the following facts. Each data point is a data set (set of sequences). Each sequence has different base compositions. Stochastically, each sequence has different stationary probability distributions. The variation among data points is complicated to be captured and visualized by ordinary data analysis techniques. In addition, data points are distributed among several network-nodes. In fact, we can consider this problem as one of new research problems in the field of Big Data mining [4, 24].

Recent years, we witnessed new research directions in data science. One of those new research directions is Big Biological Data Mining (or Massive Biological Data Mining)[17]. Developing new methods, algorithms, approaches, and techniques to use them in analyzing Big Biological data sets is the most promising and optimistic research developments in Data Science [5]. Developing the applicability of the existing data analysis methods, approaches, algorithms, and techniques to use them in analyzing Big data sets is a new challenge in Data Science.

We define *Big Data* or *Massive Data* as large or complex data sets that are collected using different data acquisition procedures with the following properties: (1) size of data sets is either variable or constant, (2) data types may vary in terms of varieties and variations [24], (3) velocity of collecting data sets is dynamically increasing/decreasing (in real time t) or static. In an abstract sense, suppose that we have data sets $\{d_1, d_2, ..., d_n\}$ of segmented genomes (i.e. sets of sequences) of a well-known virus. Let $d_i = \{S_1, S_2, S_3, ..., S_{u_i}\}$ $((i = 1, ..., n)$ and $(|d_i| = u_i))$ be i-th set of sequences, and S_{u_i} be u_i-th sequence. Suppose that $\{d_1, d_2, ..., d_n\}$ are distributed over m nodes (i.e m databases). In addition, suppose that there exists query node (main node) among the m nodes. We aim to capture, visualize, and compare the genetic variations of each data set of the segmented genomes of a given virus in feature spaces by using the following generic decentralized computational concept(DCC): move code of computations to data instead of moving data to code of computations in order to reduce the volume of transferred raw data over the network, and implicitly to preserve the privacy and sensitivity of data under consideration. Moreover, we aim to capture, visualize, and compare the genetic variations of S_{i_1} and S_{i_2} by using DCC, such that $S_{i_1} \in d_{j_1}$ and $S_{i_1} \in d_{j_2}$, d_{j_1} and d_{j_2} are located at two different nodes. Therefore, in this paper, we propose the following new research objective: *to design a new statistical computing algorithm or model (i.e. new alignment-free visualization tool) in biological data mining that can be implemented in asynchronous and autonomous manner.*

The main question that arises in this context is: why we consider the data under consideration as Big Biodata? The answer is simply due to the following: (i) The data points are datasets, (ii) the accumulating process of the data points (i.e. segmented genome of a virus) is a continuous process or near-real time process, specifically during spreading, and (iii) the datasets or the data points under consideration are stored in different distributed databases (or at different nodes).

After we described the research problem under consideration, in the next section, we shall present related research works.

3 Related research works

Functionally, Data-visualization tool is a necessary complementary phase to data-analysis tool. In this section, we present the existing alignment-free sequence-set visualization tools (SSVT) and the existing alignment-free sequence visualization tools (SVT). Moreover, we consider the existing alignment-based sequence visualization tools and alignment-based sequence-set visualization tools as unrelated research works, therefore, more details about alignment-based sequence-visualization tools and alignment-based sequence-set visualization tools can be found in [8, 7, 18].

It should be noted that all the existing SSVT and SVT are designed using the centralized computational concept(CCC). Sequences can be compared using either alignment-based algorithms (Pairwise Alignment and Multiple Alignment) [12, 20, 11] or alignment-free algorithms (Euclidian Distance, Standardized Euclidean Distance, Mahalanobis Distance, Correlation Coefficient, Largest Generalized Eigenvalue-based Distance, Entropy Measure, Kolmogorov Complexity, and Markov Chain models) [5, 22, 26, 25, 6]. Multiple sequence alignment algorithms are very useful algorithms in aligning and analyzing any homogeneous set of sequences[12, 20]. In next part of this paper, we focus on reviewing the existing statistical information-based visualization algorithms that can be used in visualizing results of sequence analysis algorithms.

A statistical information-based visualization algorithm is an integrated algorithm to statistical analysis algorithm (see [2]). The two algorithms can be combined to create two-phase process: (1) extracting statistical information from sequences (analysis phase) and, (2) visualizing the extracted statistical information (visualization phase). The extracted statistical information are denoted by the following: probability distribution of nucleotide compositions and its statistical parameters (e.g. mean, variance, standard deviation), variations and co-variations between/among sequences, statistical clustering of sequences, statistical classification of sequences, and analysis of outliers.

From modeling point of view, any sequence is linear in time. Information can be extracted from any given sequence using a feature extraction algorithm. For example, counting the occurrences of n_1-grams in a sequence is a well-known statistical language modeling algorithm. The algorithm has the following computational step. Without loss of generality, suppose that we have the following sequence: ACGACT. The algorithm simply converts ACGACT to the following fixed-length feature vector: 2 AC, 1 CG, 1 GA, 1 CT, or $(2111)'$. The vector represents the occurrences of the following 2-grams: AC, CG, GA, and CT (i.e. frequency of AC, CG, GA, CT). If we consider all possible 2-grams, then we have to add 12-zeros to $(2111)'$. In this context, the mapping represents a fixed-length vector for a sequence of symbols. It is a data-vector that results from a sequential discretization process. The relative-frequency of AC, CG, GA, CT represents the normalized vector of $(2111)'$. The feature vector can be computed using two different mechanisms: (1) sliding a fixed-length window on a given sequence from one end to another end, and computing the feature vector for each instance of window, or by (2) computing the feature vector for the entire sequence (i.e. window-length = sequence-length). There are various statistical analysis algorithms that can be implemented to extract various statistical information from extracted feature vector(s). The next step is to visualize the extracted statistical information using different statistical graphs (descriptive statistics). In this context, data points are sequences, which can be generalized to sets of sequences (see [5, 8, 7, 9]). As we previously mentioned, there are two types of sequence-sets: (i) homogeneous sequence-sets, and (ii) heterogeneous sequence-sets. We define both types of sequence-sets from two different angles: (i) mathematically or statistically, and (ii) biologically. It is not always true that mathematical definitions are perfectly identical to biological definitions. The differences always exist between mathematical and biological definitions, and yet mathematical definitions are proved to be powerful in modeling any biological phenomenon.

One of the well-known bioinformatics tool boxes is the MATLAB toolbox: bioinformatics[18]. There are various built-in functions that can be used by end-users to visualize the statistical analysis output of any given sequence (i.e. statistical sequence analysis). For example: pie-chart, bar-chart, and codon-map (see Figure 4). In Figure 4, the outputs are created using mechanism 2 (entire sequence). In Figures (3a), the outputs represent the probability density of all possible 1-grams (upper subplot), and the probability density of all possible 2-grams (lower subplot) in a given sequence respectively. In addition, the outputs are created using mechanism 2. In Figure (3b), the outputs represent the probability density of all possible 1-grams (upper subplot), and the probability density of {AT,CG} (lower subplot) in a given sequence respectively. In addition, the outputs are created using mechanism 1 (sliding-window based mechanism). The cluster analysis of sequences can be visualized using dendogram plot (see Figure 5c). Dendogram is a built-in visualization tool in MATLAB-bioinformatics toolbox. In this example, the default distance measure is the euclidean distance. The distance values are computed between every possible pair of sequences in a set of sequences. Precisely, a distance value represents the distance between a pair of feature vectors, where each feature vector represents a sequence (mechanism 1). Finally, it should be noted that there are non-statistical based visualization tools that can be used to analyze and visualize the extracted information from sequences. For example, the dot-plot of a pair of sequences (identicalness between two sequences, see Figure 3c).

To be consistent with the objectives of this paper, the non-statistical based visualization tools are excluded from this section. Sequence-set analysis is a developing research direction. The data points under consideration are sets of sequences. Sequence-set analysis has two sub-directions: (1) Alignment-free Sequence-set Analysis, and (2) Alignment-based Sequence-set Analysis. Analyzing sequence-sets in feature spaces requires feature extraction techniques to extract observed feature vectors from sequence-sets (e.g. n_1-grams model). M. Daoud [5] proposed a new variance-covariance structure-based statistical pattern recognition system for solving the sequence-set proximity problem under the homology-free assumption. The system is designed upon using the difference between two variance-covariance matrices, where each variance-covariance matrix represents a sequence-set. The variance-covariance matrix is a well known matrix in multivariate analysis [13, 1]. The key point of the proposed system is [5]: the system has the capability in estimating the distance between any two sequence-sets (i.e. two sets of sequences), such that there is no prior knowledge about homology-assumption. In terms of time complexity and complexity of data points under consideration [5], the proposed system shows robustness in performing the following processes on sets of sequences without alignment: (i) classification, (ii) clustering, (iii) variability detection, and (iv) sequence-set based searching.

The Outputs from visualization tools included in the analysis phase of the proposed system are illustrated in Figures (5a) and (5b). The outputs are the integrated phase of analyzing patterns in sets of sequences using variance-covariance matrices to perform classification and clustering algorithms. In Figure (5b), the output represents scatter diagram of sequence-sets embedded in a high dimensional feature space (classification), whereas, In Figure (5a), the output represents a dendogram of sequence-sets embedded in a high dimensional feature space (clustering). In addition, both tools offer a visual assessment for the extracted information from sequence-sets embedded in a high dimensional feature space. Moreover, the extracted knowledge is expected to be undetectable in lower dimensional feature spaces/data space, or are detectable differently from lower dimensional feature spaces/data space. It should be noted that each data point represents a set of sequences with the following condition: no prior knowledge about homology-assumption.

In this section, we presented the related research work that focuses on analyzing and visualizing sequences and sequence-sets. In the next section, we shall present the proposed algorithm to analyze sets of heterogeneous sequences (e.g. segmented gnome of flu virus) in feature space using decentralized computational concept (or model).

4 New decentralized alignment-free sequence-set visualization tool

In this section, we present a new decentralized statistical computing algorithm to analyze and visualize sets of heterogeneous sequences (e.g. segmented gnome of flu virus) in feature spaces. Given data sets $\{d_1, d_2, ..., d_n\}$ of segmented genomes (i.e. sets of heterogeneous sequences) of a well-known virus. Let $d_i = \{S_1, S_2, S_3, ..., S_{u_i}\}$ $(i = 1, ..., n)$ $(|d_i| = u_i)$. Suppose that $\{d_1, d_2, ..., d_n\}$ are distributed over m nodes (i.e m databases), and let $\Omega^{(j)} = \{\omega_1^{(j)}, \omega_2^{(j)}, ..., \omega_p^{(j)}\}$ be sets of strings, $j = 1, 2, ..., l$. To illustrate the abstract concept via an example, suppose that we aim to send one or more software agent(s) to mine distributed databases over a network. The network consists of five nodes and five links (see Figure 2). In this example, the main node is (v_1). Assume that dataset d_3 is hosted by node two (v_2), datasets d_1 and d_2 are hosted by node three (v_3), and datasets d_4 is hosted by node four (v_4). Suppose that we dispatch a software agent with the following task T: the main task is to mine all the distributed datasets independently in a decentralized manner and send the outputs to the main node.

Definition 1 *[10] A mobile agent is a software entity that has the capability to roam the network from one node to another to accomplish a given task T on behalf of a user. At the time of its dispatching, a mobile agent has a specified route to a accomplish a given task T. During agent's mission, the route is either static or dynamic.*

To design an approach for analyzing sets of heterogeneous sequences in a feature space using decentralized computational concept, we have to define a feature extraction phase. In other words, we have to map each sequence in a given sequence-set into a feature space using a well-defined feature vector $\boldsymbol{X}$. Let $\boldsymbol{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, ..., X_p^{(j)})'$ be a $(p \times 1)$ real-valued feature vector, where each feature variable $X_r^{(j)}$ $(r = 1, 2, ..., p)$ represents the occurrences of the string $\omega_r^{(j)} \in \Omega^{(j)}$ of length n_1 (n_1-gram) in a sequence. In order to associate feature variables $X_1^{(j)}, X_2^{(j)}, ..., X_p^{(j)}$ with biological features, the feature selection phase is an essential pre-processing phase that can be used in reducing the dimensionality of any feature space without loosing or damaging the essential information required for the decision making phase. Now, the question that arises in this context can be formed as follows: Which feature vector is useful in analyzing sequence-based datasets in order to detect dissimilarities that are undetectable in other feature spaces?

The previous research question is a very complicated question. In this paper, we aim to map the heterogeneous sequence-based datasets into feature space in order to increase the capability of discrimination analysis techniques in capturing differences and/or hidden differences that are undetectable in feature space or data space. The advantage behind capturing various types of differences is to be able to associate the feature variables with the biological features, which will give virologists and epidemiologist the opportunity to understand segmented genomes of viruses from two different angles: (1) biological features (e.g. type, subtype, host), and (2) evolution (e.g. biodiversity).

Let (m_1) be an alphabet-size, and let (n_1) be a substring-length. Consequently, the number of all possible feature variables is ${n_1}^{m_1}$. To select p mathematical features from a set of ${n_1}^{m_1}$ feature variables, therefore a feature selection procedure is required. In the literature of sequence analysis field, there exists a few number of feature selection procedures, for example, J. T. L. Wang et al. [23] proposed a feature selection procedure to select relevant feature variables using the maximum likelihood approach, in other words, the probability of observing the selected feature variables in a given target class of sequences must be greater than the probability of observing the selected feature variables in a given non-target class of sequences in order to stochastically maximize discrimination between two classes of sequences. The proposed procedure is a stochastic assumptions-based procedure, which it can be implemented on homogeneous sequence-sets. In this paper, homogeneous sequence-sets are defined as classes of sequences, such that (1) sequences in each class have common biological features, and (2) the nucleotide compositions of sequences in each class have common stochastic characteristics.

In order to clarify the main reason behind inapplicability of the previous feature selection procedure (or feature selection phase) in analyzing a segmented gnome of a virus, the probabilistic analysis of occurrences of substrings of length n_1 (i.e. n_1-grams) in a segmented gnome of influenza virus is illustrated in Figure 3a. Figure 3a illustrates the relative frequency of all possible 1-grams, and 2-grams, in each sequence of a segmented genome of influenza virus. In case of 2-grams, we have 16 possible 2-grams, therefore, we have 16 lines, where each line represents the relative frequency of one of the possible substrings of length 2.

In case of 1-grams, we have 4 possible 1-grams, therefore, in Figure (3a)(upper subfigure), we have 4 lines, where each line represents the relative frequency of one of the possible substrings of length 1 (i.e.1-grams). To select a set of n_1-grams from the set of all possible n_1-grams, it is trivial to choose the maximum likelihood criterion (i.e to choose n_1-grams with highest relative frequency in all sequences of a heterogeneous sequence-set). From Figure (3a), it is clear that the maximum likelihood criterion is inapplicable (e.g.: in Figure (3a), upper subfigure, red line at sequence 9). Hence, in case of heterogeneous sequence-sets, we have to select substrings of length n_1 from the set of all possible substrings of length n_1 in a random manner (see [5]). Consequently, the power of analyzing heterogeneous sequence-sets is directly depends on sequence-set analysis phase. Now, the research question that arises in this context can be summarized as follows: Can we compose a dispersion map for a given heterogeneous sequence-set using the observed feature vectors that are embedded in feature space?

To maximize the power of analyzing heterogeneous sequence-sets, after mapping each sequence in a heterogeneous sequence-set into feature space, we have to define another mapping function in order to convert the extracted feature vectors to a dispersion map. A dispersion map represents relations (variations and co-variations) between all possible pairs of feature variables rather than relations between all possible pairs of sequences. Hence, a dispersion map is conceptually different from a score matrix. A score matrix is an output from a multiple sequence alignment algorithm, and it can be used to evaluate the degree of similarity (or dissimilarity) between any two sequences in a set of sequences. Moreover, the limitation of using the existing multiple sequence alignment algorithms can be specified as follows: the homology assumption is expected to be violated by any heterogeneous set of sequences. It should be noted that the proposed approach is an alignment-free sequence-set analysis approach. To compose a dispersion map for any heterogeneous sequence-set, variance-covariance matrices are computed for the extracted feature vectors that are embedded in feature space (i.e. $Cov(\boldsymbol{X}^{(j)})$). The variance-covariance matrix is a symmetric positive semidefinite matrix, where its diagonal elements represent variances, and its off-diagonal elements represent covariances (see [5, 1]). The variance-covariance matrix of a feature vector $X_1^{(j)}, X_2^{(j)}, ..., X_p^{(j)}$ is denoted by $Cov(\boldsymbol{X}^{(j)})$, and it is defined as [1]:

$$Cov(\boldsymbol{X}_n^{(i)}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{1}$$

where $\sigma_{i_1i_1}=$ E $({X_{i_1}^{(j)}}^2)$, $\sigma_{i_1j_1}=$ E $(X_{i_1}^{(j)}X_{j_1}^{(j)})$ for i_1, j_1 = 1,2,...,p, such that E$(X^{(j)})$=0. The sample covariance matrix is denoted by $\widehat{Cov(X^{(j)})}$, and it is defined as:

$$\widehat{Cov(\boldsymbol{X}^{(j)})} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{bmatrix} \tag{2}$$

where $\hat{\sigma}_{i_1i_1}=\frac{\sum_{r=1}^{l}{(x_{ri}^{(j)}-\bar{x_i}^{(j)})}^2}{l}$, $\hat{\sigma}_{i_1j_1}=\frac{\sum_{r=1}^{l}(x_{ri}^{(j)}-\bar{x_i}^{(j)})(x_{rj}^{(j)}-\bar{x_j}^{(j)})}{l}$ for i_1, j_1 = 1,2,...,p. In addition, l denotes the number of observed values of the feature variable $X_i^{(j)}$ (i_1 = 1,2,...,p).

For each dataset d_{i_2} hosted by node v_{i_3}, we have to compute the distance matrix $D^{d_{i_2}} = [D_{i_1j_1}(\gamma_1)]$ of d_{i_2} as follows:

$$D_{i_1j_1}(\gamma_1) = |\gamma_1'(Cov(\boldsymbol{X}_{i_1}^{(j)}) - Cov(\boldsymbol{X}_{j_1}^{(j)}))\gamma_1| = |\lambda_1| > 0 \tag{3}$$

such that $i_1, j_1 = 1, 2, 3, ..., |d_{i_2}|$ and $i_1 \neq j_1$. Now, we have to compute the sorted eigenvalues of the computed distance matrix $D^{d_{i_2}} = [D_{i_1j_1}(\gamma_1)]$ (more details about $D_{i_1j_1}(\gamma_1) = |\lambda_1| > 0$ can be found in [5]). Now, one vector of sorted eigenvalues can be sent from the current node to the main node by the dispatched agent. The vector of sorted eigenvalues is the dispersion map of d_{i_2} or the dispersion map is defined as a vector representation. We shall now discuss the core concept of the proposed decentralized statistical computing algorithm. The datasets under consideration are complex and distributed datasets (i.e hosted by different nodes). Hence, by sending the code of computations to datasets under consideration, we can locally mining those datasets by mapping them into vectors of sorted eigenvalues (transformation with minimum loss of information), and consequently by sending those vectors via the network, we can easily minimize the amount of information to be transfered via the network. The dispersion vector has the following distinguished feature: the dispersion vector offers the opportunity for the researchers in the field to conclude the biological or environmental factors that may cause the genetic variability and the genetic diversity in asynchronous and autonomous manner. Figure 1 illustrates the phases of computational process of the proposed Decentralized Alignment-free Visualization Tool (DA-fVT).

Algorithm 1: The software-agent based data visualization algorithm for distributed sets of sequences

```
input : Given data sets $\{d_1, d_2, ..., d_n\}$ of segmented genomes (i.e. sets of sequences) of a
        known virus. Let $d_i = \{S_1, S_2, S_3, ..., S_{u_i}\}$ $(i = 1, ..., n)$ $(|d_i| = u_i)$. Suppose that
        $\{d_1, d_2, ..., d_n\}$ are distributed over $m$ nodes (i.e $m$ databases), and let
        $\Omega^{(j)} = \{\omega_1^{(j)}, \omega_2^{(j)}, ..., \omega_p^{(j)}\}$ be sets of strings, $j = 1, 2, ..., l$.
output: The Eigen Analysis Chart

At the main node(node$_{main}$), Dispatch $m$ software agents $(Agent_l, l = 1, 2, 3, ..., m)$ that
  have the instance of the following sub-algorithm, to perform the required computations at
  each node node$_i$$(i = 1, ..., m)$.
while true do
    foreach $d_{i_2}$ at node$_{i_3}$ do
        foreach $S_{i_1} \in d_{i_2}$ do
            Let $\boldsymbol{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, ..., X_p^{(j)})'$ be a real-valued feature vector of
              dimensionality $p \times 1$, where the feature variable $X_r^{(j)}$ $(r = 1, 2, ..., p)$
              represents the occurrences of the string $\omega_r^{(j)} \in \Omega^{(j)}$. Map each sequence in the
              sequence-set into the feature space using the feature vector $\boldsymbol{X}^{(j)}$. The
              observed feature vectors are: $\boldsymbol{x}_1^{(j)}$, $\boldsymbol{x}_2^{(j)}$, $\boldsymbol{x}_3^{(j)}$,...,$\boldsymbol{x}_u^{(j)}$.
            For each set of sequence in $S_i$, Find $\hat{Cov}(\boldsymbol{X}^{(j)})$ as in equation 2.
        end
        Find the distance matrix $D^{d_{i_2}}$ as in equation 2.
        Find the eigenvalues of $D^{d_{i_2}}$, and send the sorted eigenvalues of $D^{d_{i_2}}$ to the main
          node (node$_{main}$).
    end
    Deactivate the software agent after sending the sorted eigenvalues, and set the
      condition to false.
end
At the main node node$_{main}$, Plot the Eigen-Analysis Chart after receiving the sorted
  eigenvalues for all data sets under consideration $\{d_1, d_2, ..., d_n\}$.
```

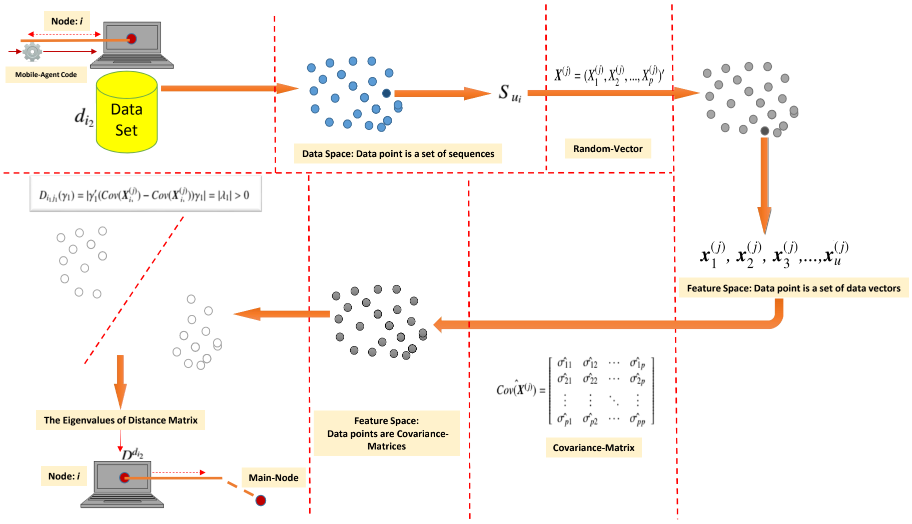

Node: i
Mobile-Agent Code
Data Set
d_{i_2}
Data Space: Data point is a set of sequences
S_{u_i}
$X^{(j)} = (X_1^{(j)}, X_2^{(j)}, ..., X_p^{(j)})'$
Random-Vector
$x_1^{(j)}, x_2^{(j)}, x_3^{(j)}, ..., x_u^{(j)}$
Feature Space: Data point is a set of data vectors
$D_{i_1 j_1}(\gamma_1) = |\gamma_1'(Cov(X_{i_1}^{(j)}) - Cov(X_{i_1}^{(j)}))\gamma_1| = |\lambda_1| > 0$
$\hat{Cov}(X^{(j)}) = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{bmatrix}$
Covariance-Matrix
Feature Space: Data points are Covariance-Matrices
The Eigenvalues of Distance Matrix
$D^{d_{i_2}}$
Node: i
Main-Node

Figure 1: The computational process of Decentralized Alignment-free Visualization Tool (DA-fVT)

In this section, we presented the proposed approach to analyze and visualize sets of heterogeneous sequences (e.g.segmented genome of flu virus) in feature space using asynchronous and autonomous manner. In the next section, we present results and discussion.

5 Results and discussion

As we previously mentioned, the data points under consideration are sets of heterogeneous sequences, and therefore, in this section, we implement the proposed algorithm (visualization tool) using real sets of heterogeneous sequences (i.e. real data points). Hence, in order to motivate the way of testing the proposed algorithm, we select the segmented genomes of two viruses: (i) Influenza A virus, and (ii) Ebola Virus. The following part of this section, in brief, describes the two selected viruses.

The influenza virus infection is considered as serious public health problem in many countries. The influenza virus has highly mutation rates, which implicitly indicates the following: the virus can change rapidly and spread quickly. The virus has negative impacts on human health, especially for the elderly group. The influenza virus is classified under the family Orthomyxoviridae [21, 3, 14]. The genome of influenza virus is a segmented genome, and it has eight segments ([9]). Each segment is encoded into either one or two proteins ([9]). The eleven RNA-proteins of influenza virus genome are: PB1 (Polymerase protein), PB2 (Polymerase protein), PA (Polymerase protein), HA (Haemagglutinin protein), NP (Nucleoprotein), NA (Neuraminidase), M1 (Matrix protein), M2 (Matrix protein), NS1 (non-structural protein), and NS2 (non-structural protein). The evidence of variability is embedded in the genetic text of the two surface proteins: (i) haemagglutinin (HA) and (ii) Neuraminidase (NA) [15, 16]. The identification of influenza sub-type can be accomplished using the variability of HA and NA proteins.

One of the most highly virulent viruses is Ebola virus. The Ebola virus is a negative-sense RNA virus, and it is classified under the family Filoviridae [19]. The genome of Ebola virus is a segmented genome. The seven RNA proteins of Ebola virus genome are: Nucleoprotein (NP), Nucleocapsid protein (VP35), Matrix protein (VP40), Glycoprotein (GP), Nucleocapsid protein (VP30), Nucleocapsid protein (VP24), Polymerase protein (L).

The following experiments represent different expected implementations for the proposed algorithm:

1. Suppose that we have four distributed datasets. Each data point is a segmented genome of influenza A virus (i.e. set of heterogeneous sequences). Those datasets are hosted by different nodes. The four datasets have different sample sizes: 40, 30, 17, and 17 respectively. The sizes of the four datasets are (in bytes): 1123290, 845460, 490194, and 485544 respectively. The output of the proposed algorithm is depicted in Figure 6a. The output has the form of line graphs. The line graphs represent the inner variation structures of the distributed datasets. It is clearly verified that the datasets under consideration have different inner variation structures. The algorithm successfully reduces the amount of information to be transfered over the network by 99.97174381420471% (i.e. only 0.02825618579529% of the amount of information is expected to be transfered over the network).

2. The second implementation of the proposed algorithm is very useful in analyzing any feature extraction process that is based on using the occurrences of 1-grams, 2-grams and 3-grams. Comparing the variation of a dataset, such that each data point is a set of heterogeneous sequences, by considering different feature variables that can be defined as the occurrences of n_1-grams ($n_1 = 1, 2, and 3$). Suppose that we have two datasets: (i) d_1: 40 segmented genome of influenza A virus (i.e. sets of heterogeneous sequences), (ii) d_2: 34 segmented genome of Ebola, and we aim to compare the variation of three feature vectors $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, and $\boldsymbol{X}^{(3)}$ that are defined as the occurrences of all possible 1-grams, 2-grams and 3-grams respectively. In Figures 6e and 6f present the comparisons of the variability of $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, and $\boldsymbol{X}^{(3)}$ in d_1(influenza A) and d_2(Ebola) respectively. It is clearly verified that 3-grams have the highest variation compared with 2-grams and 1-grams. Next, we shall present another implementation.

3. Now, we present a useful implementation about monitoring the variation in the process of accumulating data points. The jump points in statistical variation can be used to indicate the existence of jump points in biological variation. Suppose that we have two datasets: (i) d_1: 40 segmented genome of influenza A virus (i.e. sets of heterogeneous sequences), (ii) d_2: 34 segmented genome of Ebola. Figures 6b and 6c illustrate the outputs of the proposed algorithm, specifically, by considering the data points sequentially (i.e. one data point each time, initially, starting with 2 data points). d_1 has more jump points than d_2. In this way, we can simply monitor the existence of the jump points in statistical variation. Another way of monitoring the statistical variation can be achieved by applying the proposed algorithm at ordered points in time (e.g. $t_1 < t_2 < t_3$). The output is illustrated in Figure 6d. It is verified that the jump points in statistical variation are clearly identified.

During outbreaks, suppose that we distribute mobile biomedical labs in specific geographical areas to collect data about a biological phenomenon. Building a new platform to analyze datasets at each distributed node without transferring data to the main node will minimize the amount of transferred data via the network, communication errors and failures, and consequently will maximize data security and privacy. The algorithm can be implemented as a mobile application. Therefore, specialists in medical sector can analyze and visualize distributed big datasets from their mobiles or laptops or tablets, specifically, those devices have limitations in power of computations and data-storage.

At to this point, we remark the following: the implementations of the proposed algorithm are dealing with the way of minimizing the amount of information to be transfered over the network, and the way of projecting datasets with data-complexity(i.e. big datasets and each data point is a dataset) into real vectors of eigenvalues. Next, we shall present conclusions and future work.

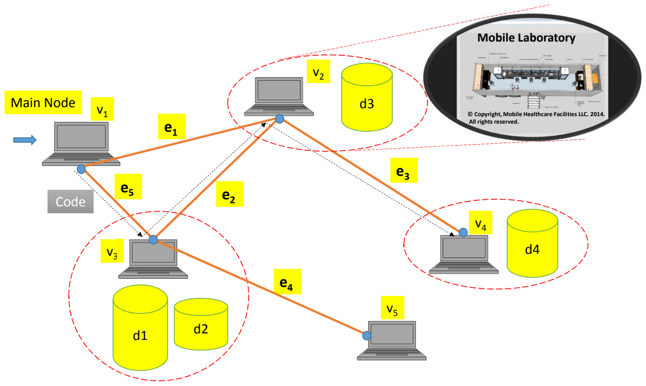


Figure 2: The distributed data sets: (1) Data sets@node 3: d1,d2, (2) Data sets@node 2: d3, (3) Data sets@node 4: d4

6 Conclusions and future work

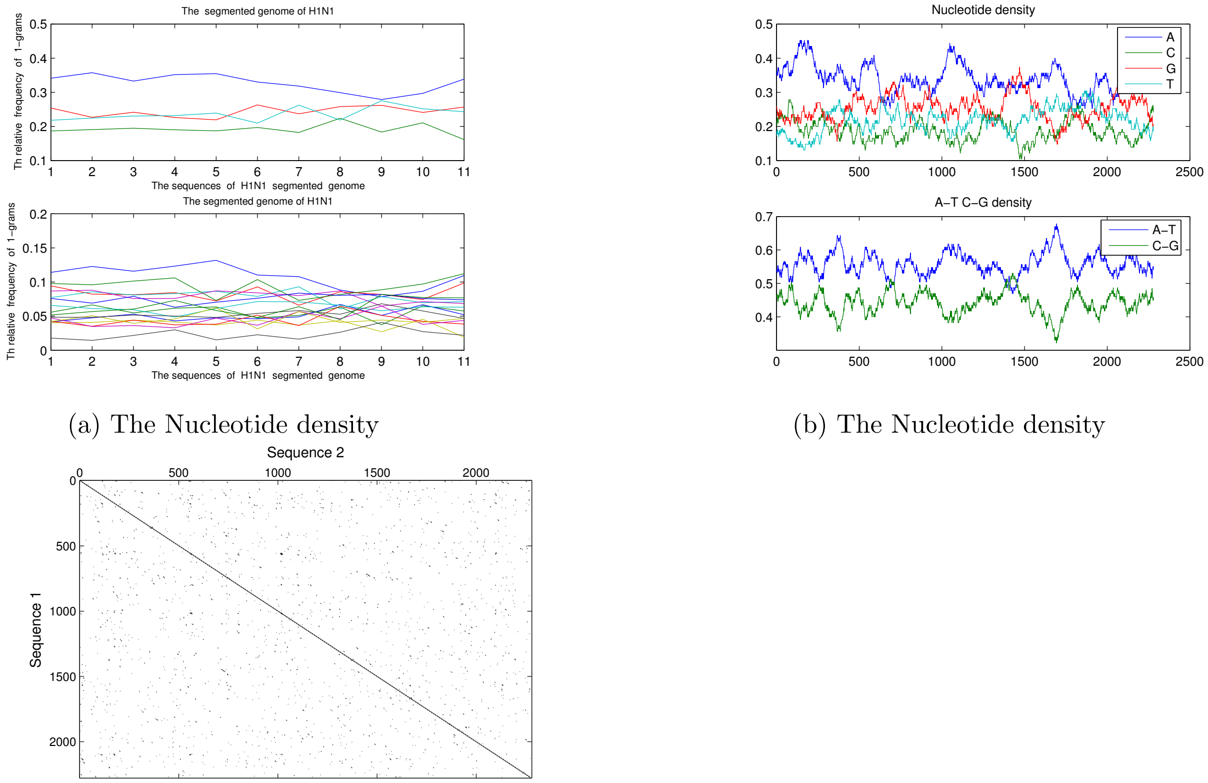
Recently, we deal with new terminology or new research fields, for example, Big Data Mining, Biological Data Mining, and Big Biological Data Mining however, the research problem under consideration is tagged under a new research direction, which is Decentralized Big Biological Data Mining. The new algorithm is a visualization tool that can be used to visualize and analyze Big-BioData (sets of sequences) in a decentralized manner. The proposed algorithm is designed based on analyzing the distributed datasets locally, and map them into vectors of generalized eigenvalues in a decentralized manner. The generalized eigenvalues represent the distance among the data points under consideration.The proposed algorithm has two significant contributions: (1) minimizing the amount of information to be transfered over the network, and (2) projecting datasets with complexity (i.e. Big Datasets) into real vectors of eigenvalues in a decentralized manner. The experiments effectively and significantly showed the robustness of the proposed algorithm. In the future work, we aim to propose other various approaches, techniques and algorithms to deal with distributed big datasets such that the each data point is a dataset (i.e the concept of the data point is generalized to a dataset).

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley, Hoboken, NJ, 3rd edition edition, 2003.

[2] Daniela G. Calo. Italian contributions on some recent research topics in cluster analysis. *STATISTICA*, 72(3):pp.271–286, 2012.

[3] Alan Cann. *Principles of Molecular Virology.* Academic Press, 4 edition, 2005. ISBN: 9780080470726.

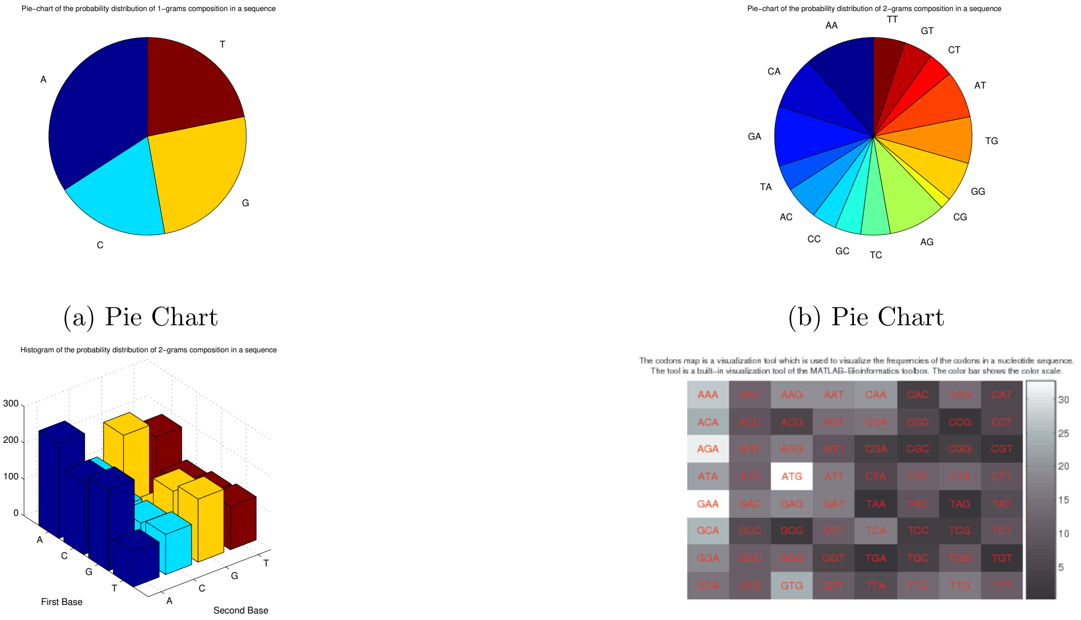


(a) The Nucleotide density

(b) The Nucleotide density

(c) The dot plot

Figure 3: The built-in visualization tools of Bioinformatics toolbox$^{(MATLAB)}$



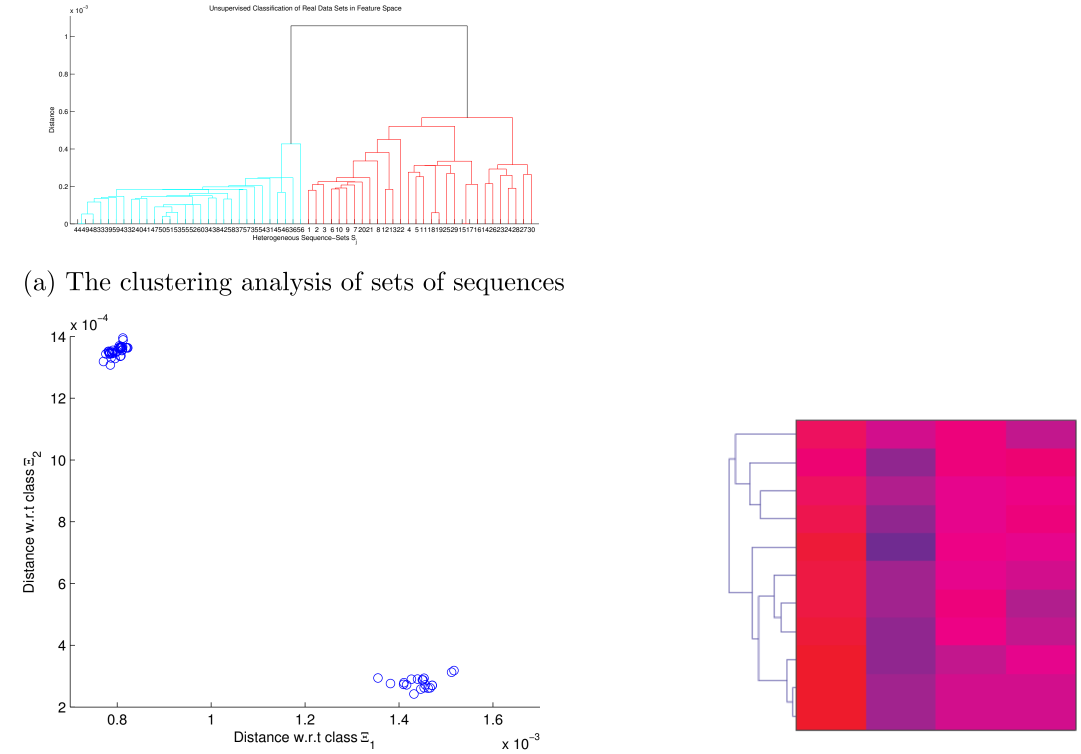
(a) Pie Chart

(b) Pie Chart

(c) Histogram

(d) Map of codons

Figure 4: The built-in visualization tools of Bioinformatics toolbox$^{(MATLAB)}$ for the probability distribution of n-grams.



(a) The clustering analysis of sets of sequences

(b) The classification analysis of sets of sequences

(c) The clustering analysis of sequences

Figure 5: Visualization tools

[4] Dianne Cook, Eun-Kyung Lee, and Mahbubul Majumder. Data visualization and statistical graphics in big data analysis. *Annual Review of Statistics and Its Application*, 3(1), 2016.

[5] Mosaab Daoud. *A New Variance-covariance Structure-based Statistical Pattern Recognition System for Solving the Sequence-set Proximity Problem Under the Homology-free Assumption.* PhD thesis, University of Guelph, Ontario, Canada, 2010. http://search.proquest.com/docview/815574781.

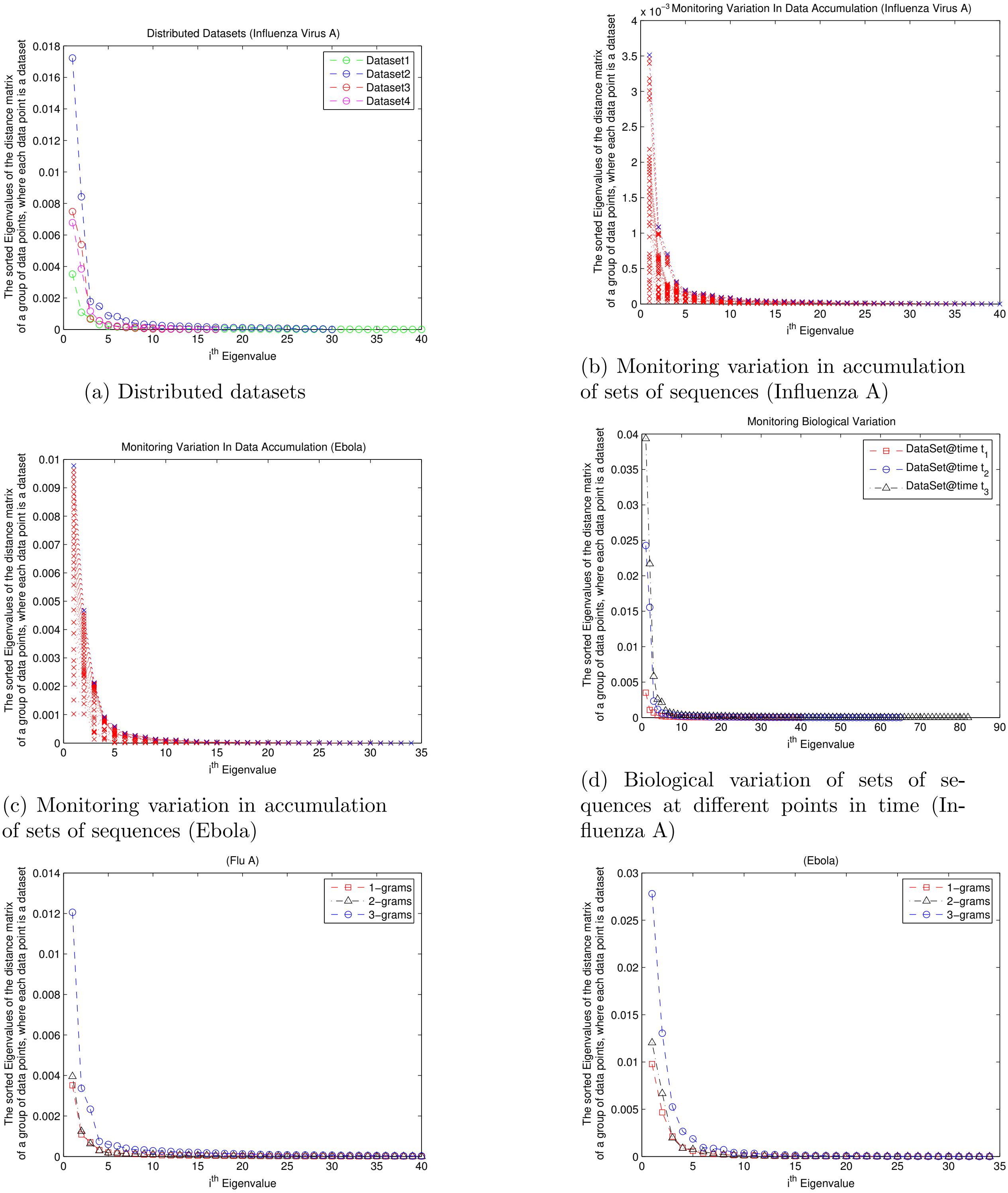
[6] Mosaab Daoud. Quantum sequence analysis: A new alignment-free technique for analyzing sequences in feature space. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, pages 702–703, New York, NY, USA, 2013. ACM.

[7] Mosaab Daoud and Stefan C. Kremer. Detecting similarities between families of bio-sequences using the steady-state of a pca-neural network. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB06)*, pages 179–185, 2006. http://dx.doi.org/10.1109/CIBCB.2006.330988.

[8] Mosaab Daoud and Stefan C. Kremer. Neural and statistical classification to families of bio-sequences. *International Joint Conference on Neural Networks, 2006 (IJCNN '06)*, pages 699–704, 2006. http://dx.doi.org/10.1109/IJCNN.2006.246752.

[9] Mosaab Daoud and Stefan C. Kremer. A new distance distribution paradigm to detect the variability of the influenza-a virus in high dimensional spaces. In *Bioinformatics and Biomedicine Workshop, 2009. BIBM 2009. IEEE International Conference on*, pages 32 –37, nov. 2009. http://dx.doi.org/10.1109/BIBMW.2009.5332133.

[10] Mosaab Daoud and Qusay H. Mahmoud. Monte carlo simulation-based algorithms for estimating the reliability of mobile agent-based systems. *J. Network and Computer Applications*, 31(1):19–31, 2008.



(a) Distributed datasets

(b) Monitoring variation in accumulation of sets of sequences (Influenza A)

(c) Monitoring variation in accumulation of sets of sequences (Ebola)

(d) Biological variation of sets of sequences at different points in time (Influenza A)

(e) Comparing variation of 1-grams, 2-grams, and 3-grams in sets of sequences (Influenza A)

(f) Comparing variation of 1-grams, 2-grams, and 3-grams in sets of sequences (Ebola)

Figure 6: The outputs of the proposed algorithm

[11] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[12] R.C. Edgar and S. Batzoglou. Multiple sequence alignment. *Current Opinion in Structural Biology*, (16):368–373, 2006.

[13] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3rd edition edition, 2007.

[14] Maurice R. Hilleman. Realities and enigmas of human viral influenza: Pathogenesis, epidemiology and control. *Vaccine*, 20:3068–3087, 2002.

[15] E. Kurstak and A. Hossain. *Virus Variability, Epidemiology and Control*, volume 2, chapter 1, pages 1–7. Springer, 1990. ISBN 978-0-306-43359-7.

[16] Robert A. Lamb and Robert M. Krug. *Fields of virology*, volume 2, chapter 46, pages 1487–1579. Lippincott Williams and Wilkins, 2001. ISBN/ISSN: 9781451105636.

[17] Yixue Li and Luonan Chen. Big biological data: Challenges and opportunities. *Genomics, Proteomics & Bioinformatics*, 12(5):187 – 189, 2014.

[18] MathWorks. *Bioinformatics Toolbox: Users Guide*, 2013. http://www.mathworks.com/help/bioinfo/index.html.

[19] NCBI. Ebolavirus resource. viralzone, 2015. http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola

[20] C. Notredame. Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):1–14, 2002.

[21] B. Schweiger, I. Zadow, and R. Heckler. Antigenic drift and variability of influenza viruses. *Med Microbiol Immunol*, 191:133–138, 2002.

[22] Susana Vinga and Jonas Almeida. Alignment-free sequence comparisona review. *Bioinformatics*, 19(4):513–523, 2003.

[23] Jason Tsong-Li Wang, Qicheng Ma, Dennis Shasha, and Cathy H. Wu. New techniques for extracting features from protein sequences. *IBM Systems Journal*, 40(2):426–441, 2001.

[24] Patrick J. Wolfe. Making sense of big data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45):18031–18032, 2013.

[25] Tiee-Jian Wu, John P. Burke, and Daniel B. Davison. A measure of DNA sequence dissimilarity based on mahalanobis distance between frequencies of words. *Biometrics*, 53(4):1431–1439, 1997.

[26] Tiee-Jian Wu, Ya-Ching Hsieh, and Lung-An Li. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, 57(2):441–448, 2001.

Mosaab Daoud
mdaoud@alumni.uoguelph.ca

Benjamin Aziz
benjamin.aziz@port.ac.uk