# Revisiting spectral clustering: assessments and extended capabilities

Johny Matar
*LaRRIS, Faculty of Sciences*
*Lebanese University*
Fanar, Lebanon
johny.matar@ul.edu.lb

Hicham El Khoury
*LaRRIS, Faculty of Sciences*
*Lebanese University*
Fanar, Lebanon
hkhoury@ul.edu.lb

Jean-Claude Charr
*DISC Laboratory, Femto-ST Institute, UMR 6174 CNRS*
*Université de Bourgogne Franche-Comté*
Besançon, France
jean-claude.charr@univ-fcomte.fr

Christophe Guyeux
*DISC Laboratory, Femto-ST Institute, UMR 6174 CNRS*
*Université de Bourgogne Franche-Comté*
Besançon, France
christophe.guyeux@univ-fcomte.fr

***Abstract*—Many techniques are being used for biological sequence analysis. Among them, the efficiency of Clustering techniques has emerged in many recent research works. Spectral clustering in general, and Gaussian Mixture Models (GMMs) in particular, have a demonstrated efficiency in clustering biological sequences having unknown similarity ratios. The pipeline of a tool, implementing the spectral clustering technique, consists of the following steps: i- sequence alignment, ii- pairwise affinity computation of the sequences, iii- Laplacian Eigenmap embedding of the data and GMM-based clustering. The choice of a certain tool or method, for two initial steps, can affect the quality of the resulting clustering. In the present paper, we assess the effect of using different sequence alignment tools and applying different affinity matrix types on the accuracy of the results. Inputting heterogeneous sequences, or ones that are subjected to Horizontal Gene Transfers (HGTs), a natural factor that can lead to high divergence between related sequences, is also experimented with. Our main contribution is showing the most appropriate use cases for different alignment tools and affinity matrices. Moreover, the ability of the spectral clustering technique to handle heterogeneous sequences and HGTs has been demonstrated.**

***Index Terms*—Biological sequence clustering, Clustering quality analysis, Spectral clustering, Gaussian Mixture Model, Sequences alignment, Affinity matrices, Horizontal Gene Transfer.**

## I. Introduction

The huge number of newly sequenced genes or genomes allows researchers and practitioners, in the bioinformatics field, to study the relationships between the newly discovered ones. Many tools were developed to analyze the sequenced data. In particular, clustering packages were implemented to compare a set of sequences and regroup them into clusters according to their similarity.

Linking the large number of sequences, that get discovered on a daily basis, to their ancestors and variants is a challenging and paramount target. This target requires robust clustering tools to achieve an accurate linkage especially when the similarity index between a newly discovered sequence and its ancestors is unknown. Indeed, predicting such similarity is challenging due to the significant difference in mutation rates [1], [2] among different species. Moreover, Horizontal Gene Transfer [3], a phenomenon where two organisms could exchange parts of their DNA, adds an extra layer of complexity to the predictability of the similarities. However, the mostly used clustering tools [4], which use hierarchical algorithms, heavily rely on user input of the challenging similarity or identity parameter to deliver their result.

New tools that use a spectral clustering algorithm, instead of a hierarchical one, may prove a better efficiency based on recent works [5]–[8]. Indeed, some methods, adopting spectral clustering techniques for biological sequences, started emerging [8]–[10] in the last few years. In [8], [10], [11], various numerical validation experiments demonstrated the relevance of using Gaussian Mixture Models for clustering of biological sequences having an unknown inter-cluster index of similarity. To our knowledge, a single spectral clustering-based tool, which implements the methods in [8], [10], [11], is publicly released.

Since the sequence alignment and the pairwise affinity calculation represent two core steps for applying spectral clustering to biological sequences, studying the different options for these steps improves the performance of spectral clustering in this field. In addition, the proven efficiency of spectral clustering in handling unpredictable levels of inter-cluster similarity could make it a potential solution for handling sequences that were subject to HGTs. The contributions in this work include the impact assessment of using different sequence alignment tools and different affinity matrix types with spectral clustering. It extends to a study of the performance of the spectral clustering in handling HGTs.

The remainder of this paper is organized as follows. A set of widely used sequence alignment tools, in addition to the different affinity types, are presented in Section II. Section III details our experimental protocol. The results of

our experiments are presented in Section IV. Finally, Section V concludes this work and states our future perspectives.

## II. Literature Review

### A. Sequence Alignment Tools

One of the methods for computing the distance between a pair of sequences requires obtaining their alignment first. Comparing two aligned sequences allows easy visualization of the dissimilarities by disclosing the occurrences of mutations, insertions, and deletions that differentiate the sequences. Therefore, many efficient algorithms were proposed for aligning the sequences and computing the pairwise distances, such as Needleman-Wunsch, Sankoff and Sellers [12]. Indeed, many alignment tools, that rely on such algorithms, are publically available, and the following are a few examples of the mostly used ones:

- MUSCLE [13] uses $kmer$ counting in its fast distance estimation. It then progressively aligns the sequences before refining the initial alignment by using tree-dependent restricted partitioning. As stated in [13], MUSCLE promises a superior quality alignment when compared to its rivals. Conversely, its implementation does not provide a multi-threading or a multi-processing option to accelerate its computations.
- MAFFT [14] is another multiple sequence alignment tool that applies a pipeline of five steps: i- performing a pairwise alignment, ii- calculating a pairwise distance matrix, iii- constructing a guide tree, iv- progressively aligning from the leaves to the root, and v- refining the results by iterating over the previous steps. MAFFT also supports multi-threading for further acceleration.
- DECIPHER [15] is an R language package. Its modules are written in both C and R languages. DECIPHER accelerates the sequence alignment by using secondary structure prediction algorithms. The accuracy of these algorithms increases as more sequences are used in the prediction.
- CLUSTALX [16] has in its algorithm a few common steps with its rivals. It implements the following three steps: i- performing pairwise alignment using the progressive alignment method, ii- creating a guide tree or using a user-provided one, and iii- computing the multiple sequence alignment by using the created guided tree. CLUSTALX does not support multi-threading or multi-processing.

The alignment speed and accuracy represent two major differentiating aspects between these tools that might influence the clustering quality. Therefore, it is crucial to investigate the effects of the alignment on the spectral clustering technique in order to enhance the quality of the produced clustering.

### B. Affinity matrix types

Following the alignment, the pairwise affinity is computed in two steps: i- computing a pairwise similarity matrix from the pairwise distance[1] between each couple of sequences, ii- Applying a certain mathematical transformation to the distance matrix to obtain the affinity matrix. In [17] and [8], the affinity matrix was computed as a Random Walk Normalized Laplacian and it proved to be relevant for the clustering of biological sequences. Nevertheless, various alternative matrices have been proposed for spectral clustering [18]–[21], such as the Non-normalized Laplacian, Modularity [19] and the Bethe Hessian (Deformed Laplacian) [22]. These matrices are defined as follows:

- **Non-normalized Laplacian:**

$$L = D - A,$$

where $A$ is the adjacency matrix between the sequences and $D$ is its diagonal matrix of degrees.

- **Random Walk Normalized Laplacian:**

$$L^{rw} = D^{-1}L,$$

where $D$ is the degrees matrix of the adjacency matrix and $L$ is the Non-normalized Laplacian matrix. The Laplacian matrix is symmetric and positive semidefinite.

- **Modularity:**

$$M = \frac{1}{K}\left(A - \frac{1}{K}kk^T\right),$$

where $A$ is the adjacency matrix, $k$ is the degrees vector of $A$, and $K$ is the total degree of $A$. High values for this quality function reveal the possible existence of strong communities.

- **Bethe Hessian:**

$$H_r = (r^2 - 1)I + D - rA$$

where $I$ is the identity matrix, $D$ is the degrees matrix of the adjacency matrix $A$, and the constant $r$ is the square root of the average degree of the graph, as suggested in [20].

## III. Experimental protocol

### A. The data sets

Twelve different biological sequence datasets, containing either real sequences or simulated ones, were assembled to conduct our experiments:

- A first set of $HIV-1$ type B virus[2] sequences, comprising 78 complete genomes.
- A second set of $NADH$ dehydrogenase 3 (ND3) mitochondrial gene, containing 100 sequences.
- A third set from the A/H1N1 strain of the $Influenza$ virus[3], containing 24 different nucleoprotein (NP) sequences.

[1] using a metric such as the Needleman-Wunsch distance, the similarity equlas 1-(distance/length of the shorter sequence).

[2] downloaded from https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html

[3] downloaded from https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi

The resulting clustering for the first three datasets will be assessed based on a reference clustering that will be generated via a novel and dynamic method described in citematar2021spclustv2. This method relies on phylogenetic trees where well-formed clusters should contain children, siblings, or parents from the same branch of the tree. A description of these datasets is presented in Table I.

| Dataset | Seqs count | Max length | Avg length | Min similarity % | Max similarity % | Avg similarity % |
|---|---|---|---|---|---|---|
| HIV | 78 | 8272 | 8167 | 86 | 99.4 | 89.6 |
| NADH | 100 | 369 | 341 | 46.2 | 99.7 | 62.8 |
| Influenza | 24 | 498 | 498 | 97.4 | 99.8 | 98.8 |

TABLE I
STATISTICAL DESCRIPTION OF THE REAL DATASETS.

Moreover, supplementary datasets were taken into account to assess the capability of the spectral clustering technique in accurately handling the heterogeneous datasets and the HGT.To achieve these objectives, a complete genome of SARS-CoV, along with the entire set of segments associated with the genomes of Influenza A and D strains, was obtained from viruSITE[4]. Subsequently, five more SARS-COV genomes were generated through the simulation of a 2% mutation of the original genome along with randomly introducing a comparable proportion of gaps and insertions. Likewise, nine supplementary genomes were derived from each complete genome of Influenza A and Influenza D, utilizing the assembled segments that were retrieved.

The process of simulating horizontal gene transfer was conducted in the following manner: i- two gene segments were randomly extracted from an HIV genome retrieved from the first dataset, ii- two genomes each of Influenza A, Influenza D, and SARS-COV were chosen from the previously created ones, specifically the root sequence along with one of its direct descendants, iii- the first HIV gene segment was inserted between segments 1 and 2, while the second was placed between segments 6 and 7 in the selected Influenza A and Influenza D genomes, iv- both HIV gene segments were also added at two random positions in the chosen SARS-COV sequences. The six newly created genomes replaced the original genomes within the Influenza and SARS-COV datasets. Although it is unlikely for this exact gene transfer to happen in-vivo, it remains theoretically possible in-vitro. For example the research in [23], has demonstrated that a human HeLa cell is capable of concurrently incubating and producing both Influenza and HIV-1 viruses.

The 26 genomes generated from the prior simulations, comprising 10 Influenza A, 10 Influenza D, and 6 SARS-CoV genomes, along with 9 HIV genomes randomly chosen from the first dataset, have been utilized to create four distinct biological datasets, each incorporating a diverse assortment of genomes:

- A fourth set of 20 complete genomes, comprising 10 genomes of Influenza A and 10 genomes of Influenza D.
- A fifth set comprising 26 genomes has been assembled, which includes the 20 Influenza genomes from the preceding set along with 6 SARS-CoV genomes. Both this set and the fourth set encompass pathogens that target the same body region.
- A sixth collection comprising 29 genomes has been assembled, which includes the 20 Influenza genomes from the fourth set and 9 HIV genomes selected randomly from the first set.
- A seventh set consisting of 35 genomes that encompasses all the genomes from the two preceding collections. Both this seventh set and the sixth one comprise pathogens that exhibit distinct zones of infection.

Furthermore, to validate the spectral clustering technique on larger datasets, five sets were generated via simulated mutations and starting from 6 variants of a chloroplast gene. These sequences have a minimum pairwise similarity of 60.6%, a maximum of 84%, and an average of 70.99%. Each one of the simulated datasets typically contains 6 clusters (mutations produced from each one of the 6 source sequences). Table II shows the detailed properties of the last five datasets.

TABLE II
STATISTICAL DESCRIPTION OF THE SIMULATED DATASETS.

| Dataset # | Seqs count | Min inter-cluster similarity % | Avg inter-cluster similarity % |
|---|---|---|---|
| 8 | 300 | 66.5 | 81.27 |
| 9 | 600 | 63.7 | 81.23 |
| 10 | 900 | 62.6 | 78.37 |
| 11 | 1200 | 64.8 | 81.45 |
| 12 | 1500 | 64 | 81.08 |

All these assembled datasets are publically hosted on an online repository[5].

### *B. The experiments*

The primary objective of the initial series of experiments is to evaluate the effectiveness of the alignment tool by substituting MUSCLE with other widely used software. This evaluation includes an analysis of both the quality of clustering and the efficiency of the resultant pipeline. The experiments were carried out in the following manner:

1) The first three sets of sequences are aligned using the following packages: MUSCLE, MAFFT [14], DECIPHER [15], and CLUSTALX [16].
2) The resulting aligned sets are clustered using the EM-GMM spectral clustering technique; specifically, the implementation of the "BestBIC" algorithm proposed in [10].

The second series of experiments aims to assess the quality of clustering by utilizing various types of affinity matrices. A comparative analysis will be performed between the Non-normalized Laplacian, Modularity, and Bethe Hessian, alongside the Normalized Laplacian previously employed in

[4]http://www.virusite.org/archive/2021.1/genomes.fasta.zip

[5]https://github.com/johnymatar/SpCLUST-V2/tree/master/src/datasets

SpCLUST. This experimental framework will also be applied to the three previously utilized datasets.

The third set of experiments aims to evaluate the capability of the spectral clustering technique in handling sets of different pathogens and HGTs. These experiments will be conducted over the fourth, fifth, sixth, and seventh datasets.

Finally, the last series of experiments aims to demonstrate the scalability of this spectral clustering technique and its ability to handle larger datasets. The last five simulated datasets will be used for these experiments.

## IV. Experiments results

### A. Impact of the alignment tools

This section examines the possible impact of the sequence alignment technique on the resulting clustering. The sequences under consideration were aligned utilizing four advanced alignment tools: ClustalX, Biostarts Decipher, MAFFT, and MUSCLE. Subsequently, the aligned sequences were subjected to clustering through the application of the "BestBIC" algorithm proposed in [10]. Table III displays the ARI for each clustering, along with the number of obtained clusters compared to the one in the reference clustering.

TABLE III
External clustering validation with regards to the alignment tools.

| | HIV | | | NADH | | | Influenza | | |
|---|---|---|---|---|---|---|---|---|---|
| | Nb. Clusters | | | Nb. Clusters | | | Nb. Clusters | | |
| | ref. | gen. | ARI | ref. | gen. | ARI | ref. | gen. | ARI |
| CLUSTALX | 4 | 3 | 0.690 | 4 | 3 | 0.955 | 2 | 2 | 1 |
| DECIPHER | 3 | 3 | 0.759 | 4 | 3 | 0.982 | 2 | 2 | 0.833 |
| MAFFT | 1 | 2 | - | 2 | 2 | 0.960 | 2 | 2 | 1 |
| MUSCLE | 3 | 3 | 0.828 | 4 | 3 | 0.839 | 2 | 2 | 1 |

MUSCLE achieved the highest average ARI of 0.889, followed closely by ClustalX at 0.881, Decipher at 0.858, and MAFFT at 0.653. Notably, MAFFT generated a smaller number of clusters within the HIV and NADH datasets, whereas Decipher was responsible for the most parsimonious clustering of the Influenza nucleoproteins. Consequently, MUSCLE consistently demonstrates superior performance compared to its counterparts. An examination of the alignment within the Influenza proteins dataset, where all sequences are of uniform length, reveals that DECIPHER failed to identify SNPs, misclassifying the mutations between sequences as insertions and deletions.

The speed of alignment is another important factor in evaluating the efficacy of alignment tools. For the purpose of examining this variable, the $HIV$ dataset has been chosen due to its substantial size compared to the previously analyzed datasets. Table IV presents the recorded alignment times on a system featuring a 3.4GHz Intel Core i7-6700 processor, with 8GB of RAM. An attempt was made to assess these tools using a larger dataset, specifically 7MB in size; however, only Decipher successfully executed the alignment, whereas the other tools necessitated a memory larger than 8GB.

TABLE IV
Alignment duration for $HIV$ sequences using i7-6700 3.4GHz processor.

| Alignment tool | MUSCLE | CLUSTALX | MAFFT @ 1 thread | MAFFT @ 4 threads | DECIPHER |
|---|---|---|---|---|---|
| Time (seconds) | 844 | 8027 | 1753 | 735 | 115 |

The results of the experiment indicate that Decipher is the most efficient alignment tool in terms of speed. It is succeeded by MAFFT when utilizing four threads, along with MUSCLE. In contrast, ClustalX emerged as the slowest alignment tool in the study. In conclusion, while Decipher demonstrates superior speed, its application remains problematic because:

1) the clusters generated result in a lower average ARI compared to the other methods;
2) it consists of a function in a library of R-language, i.e., not an easily integrable standalone executable for other packages.

Regarding MAFFT, there is a standalone package available for both Linux and Windows platforms. Additionally, the clustering process was notably rapid. Nevertheless, the clusters generated exhibited the lowest average ARI score, while MUSCLE scored the highest average ARI. Therefore, MUSCLE was used in the next set of experiments. Finally, the following conclusions can be deduced from the experiments:

- MUSCLE produces the most accurate results for small datasets where it is deemed the best suited;
- A multi-threaded execution of MAFFT is recommended for medium-sized datasets where it performs faster, but it might produce fewer clusters than the other tools;
- Decipher requires significantly fewer resources and is advised for large datasets.

### B. Impact of the used affinity types

As introduced previously, Non-normalized Laplacian, Modularity, and Bethe Hessian are also relevant types of affinity matrices. In this section, these affinity matrices are compared while using the same implementation of the "BestBIC" algorithm that was used in the previous set of experiments. The experiment's results are presented in Table V. The computed ARIs reveal that the Non-normalized Laplacian matrix produced poor clustering results for the $HIV$ dataset, evidenced by a notably low ARI of 0.057 and a limited number of clusters. Conversely, the application of the Modularity and Bethe Hessian matrices yielded the most effective clustering for this dataset, achieving an ARI of 0.831. Additionally, the Normalized Laplacian matrix also demonstrated commendable clustering performance, attaining an ARI of 0.828.

In the $NADH$ case, the usage of the Modularity and Bethe Hessian matrices produced the best ARI of 0.968. The Normalized Laplacian scored an ARI of 0.839, while a single cluster was produced using the Non-normalized Laplacian. This last clustering failed in detecting the different communities among the highly divergent elements of this dataset. Therefore, this matrix is not suitable for clustering highly divergent sequences. Pure clusters (all the elements of a cluster have similar labels)

TABLE V
ADJUSTED RAND INDEX WITH REGARDS TO THE USED AFFINITY MATRIX.

| | HIV | | | NADH | | | Influenza | | |
|---|---|---|---|---|---|---|---|---|---|
| | Nb. Clusters | | | Nb. Clusters | | | Nb. Clusters | | |
| | ref. | gen. | ARI | ref. | gen. | ARI | ref. | gen. | ARI |
| Non-normalized Laplacian | 7 | 2 | 0.057 | 1 | 1 | - | 2 | 2 | 1 |
| Modularity | 4 | 3 | 0.831 | 4 | 3 | 0.968 | 3 | 3 | 0.857 |
| Bethe Hessian | 4 | 3 | 0.831 | 4 | 3 | 0.968 | 2 | 2 | 1 |
| Normalized Laplacian | 3 | 3 | 0.828 | 4 | 3 | 0.839 | 2 | 2 | 1 |

are produced by using the Modularity or Bethe Hessian matrix which outperforms the Normalized Laplacian.

Finally, the Modularity matrix yielded the lowest, yet still satisfactory, ARI in the analysis of the $Influenza$ nucleoprotein dataset. In contrast, the other matrices produced clusterings that were nearly identical, achieving a perfect ARI score. Notably, the clustering derived from the Modularity matrix resulted in a greater number of clusters compared to the others, with only one misclassified element, which facilitated the identification of additional hidden communities. On the other hand, the alternative matrices, while achieving a perfect ARI, demonstrated lower sensitivity in detecting these communities. Thus, the Modularity matrix proved to be more effective in enhancing detection sensitivity when clustering data that exhibit high similarity.

To summarize, it is acceptable to use the Non-normalized Laplacian only for clustering highly similar sets. The other matrices produce performed equally well on divergent data sets. Moreover, the Modularity matrix allows the detection of more clusters among highly similar sequences.

### C. Clustering heterogeneous datasets

In the previous experiments, each of the three datasets comprised distinct sequences of the same pathogen or gene. This study aims to assess the effectiveness of the spectral technique in clustering the last four heterogeneous datasets, which include genomes from various pathogens that impact either the same or different areas of the human body, while also simulating scenarios involving horizontal gene transfer. The implementation of the BestBIC algorithm was also used in this experiment. According to the conclusions from the previous experiments, MAFFT was used for the alignment of the medium-sized genomes of the fifth and seventh sets. This set of experiments involved the usage of the Normalized Laplacian (NL), Bethe Hessian (BH), and Modularity (Mod) affinity matrices. For the sake of comparison with this approach, the clustering of these four datasets is attempted by using state-of-the-art competitors, namely CD-HIT and UCLUST. Table VI displays the expected number of clusters, the generated number of clusters, and the calculated ARI for each produced clustering.

Regardless of the used affinity matrix, the spectral clustering technique, using the bestBIC algorithm, accurately clustered all the datasets. The genomes under consideration were categorized into distinct clusters based on their respective pathogen types. Utilizing CD-HIT, a minimum similarity threshold of 0.8 was supported, resulting in the division of pathogen sequences into two separate clusters. UCLUST was assessed using various identity thresholds, commencing with a threshold of 0.9 and subsequently decreasing by 0.1 in each trial. Notably, the most favorable outcomes were achieved with identity thresholds ranging from 0.5 to 0.9, as illustrated in Table VI. A successful grouping for the sequences of the fourth and fifth datasets was also obtained by UCLUST at this threshold range. In contrast, the HIV sequences identified in the sixth and seventh sets were categorized as singletons. Conversely, the Influenza sequences were inaccurately classified when thresholds below $0.5$ were applied: specifically, at a threshold of $0.4$, three Influenza D sequences were erroneously included in the Influenza A cluster, and at a threshold of $0.3$, all Influenza sequences were grouped into a single cluster, with the exception of one sequence that was assigned to the SARS-COV cluster. At higher identity thresholds ($\geq 0.91$) the correct clusters were split.

Finally, Figure 1 illustrates the impact of simulated horizontal gene transfer in the seventh dataset on the phylogenetic signal.

For clearer legibility, the naming of the sequences is made as follows: the names of the Influenza A sequences begin with $FA$ and, in a similar pattern, the names of the Influenza D, SARS-COV, and HIV sequences start respectively with $FD$, $Co$, and $HIV$. The ancestors or descendants of a sequence can be identified by the numbers that follow the leading letters in their names, and which are separated by hyphens. For example, the sequence named $FA1\_3$ is the ancestor of $FA00\_1\_3\_5$, that is, in its turn, the ancestor of $FA0000\_1\_3\_5\_7$, etc. In addition, the sequences that received gene transfers are highlighted and marked with an additional $M$ in the alphabetic part leading their names. Since the transferred genes are identical, the pairwise similarity between the affected sequences raised and disrupted the phylogenetic signal. These sequences are incorrectly positioned as leaves on a branch in the subtree of their respective species. This phylogenetic tree is naturally expected to contain four large subtrees that separate the four involved species (HIV, SARS-COV, Influenza A, and Influenza D), while the Influenza subtrees are expected to be adjacent or share the same parent node. In contrast, the computed phylogenetic tree shows the Influenza D sequences spread over three distant subtrees. It also shows that the HIV, SARS-COV, and Influenza A sequences share the same subtree, incorrectly indicating that the Influenza A sequences are more related to the HIV and SARS-COV sequences than to the Influenza D sequences. Subsequently, an accurate visual identification of the clusters on this tree became impossible. This further highlights the superiority of the results obtained by using the proposed algorithms which successfully handled the HGT cases. Moreover, the clustering with the proposed algorithms proved to be much faster than the computation of phylogenetic trees.

This assessment demonstrated that the spectral clustering technique is more capable than the state-of-the-art tools

TABLE VI
ADJUSTED RAND INDEX WITH REGARDS TO THE USED CLUSTERING TOOL.

| | dataset 4 | | | dataset 5 | | | dataset 6 | | | dataset 7 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nb. Clusters | | | Nb. Clusters | | | Nb. Clusters | | | Nb. Clusters | | |
| | exp. | gen. | ARI | exp. | gen. | ARI | exp. | gen. | ARI | exp. | gen. | ARI |
| BestBIC-NL | 2 | 2 | 1 | 3 | 3 | 1 | 3 | 3 | 1 | 4 | 4 | 1 |
| BestBIC-BH | 2 | 2 | 1 | 3 | 3 | 1 | 3 | 3 | 1 | 4 | 4 | 1 |
| BestBIC-Mod | 2 | 2 | 1 | 3 | 3 | 1 | 3 | 3 | 1 | 4 | 4 | 1 |
| CD-HIT (id=0.8) | 2 | 4 | 0.479 | 3 | 6 | 0.570 | 3 | 6 | 0.587 | 4 | 8 | 0.627 |
| UCLUST (id=0.5 till 0.9) | 2 | 2 | 1 | 3 | 3 | 1 | 3 | 11 | 0.775 | 4 | 12 | 0.816 |

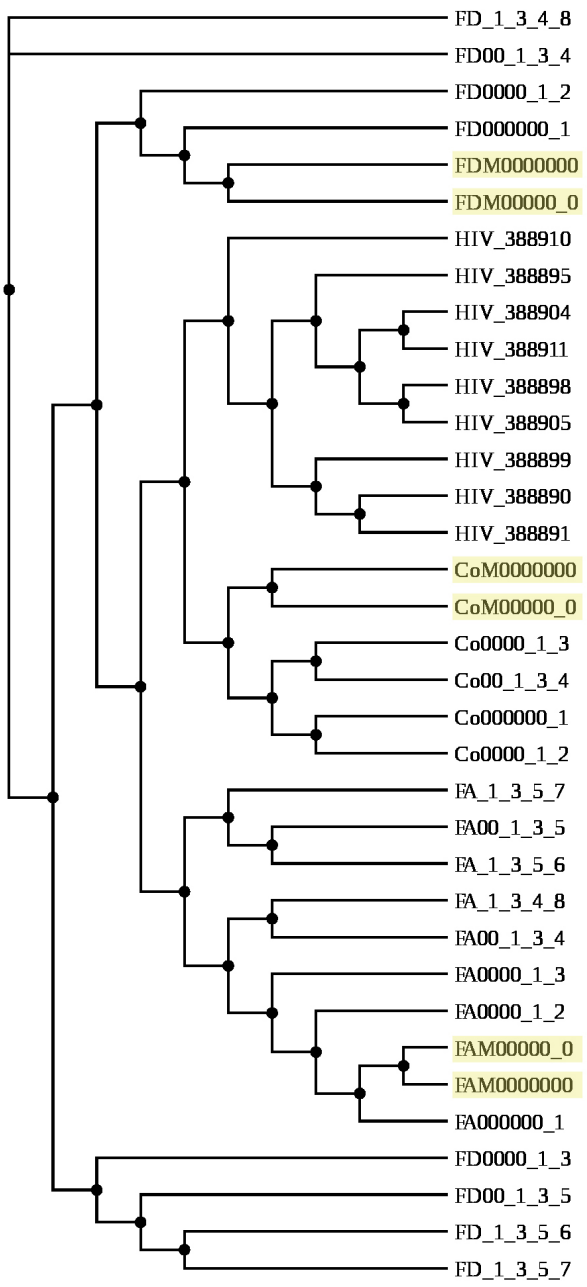


Fig. 1. Phylogenetic tree of the last hybrid set.

in clustering heterogeneous datasets. The experiments also proved the relevancy of all the affinity matrices experimented, in the spectral clustering pipeline of heterogeneous sequences as well as those subjected to HGTs.

### *D. Clustering larger datasets of divergent sequences*

The previous capabilities of the spectral technique were experimented on datasets smaller than 100 sequences. In this experiment, we will scale the size of our datasets to several hundred in order to discover the efficiency of the spectral technique over larger datasets of divergent sequences. We recall that the inter-cluster similarity scored as low as 62.6% for these last sets of data. Moreover, the pairwise similarity scored as high as 84% between their source sequences. Therefore, it can be deduced that the resulting clusters are overlapping as well. Table VII shows the number of generated clusters and the Adjusted Rand Index calculated for each produced clustering.

The spectral technique, using the implementation of the BestBIC algorithm, accurately clustered the eighth and ninth datasets, by using both the Normalized Laplacian and the Modularity matrices. It also produced the best clustering for the tenth dataset by using the Modularity matrix. Conversely, the clustering quality drastically degraded for the last two datasets containing over 1000 sequences. Indeed, when a nearly similar inter-cluster similarity is maintained, and combined with a higher number of elements, the overlapping parts among the clusters become more dense. Therefore, accurately clustering the overlapping regions becomes harder and the spectral clustering technique starts merging some clusters. This fact is clearly reflected by the diminishing number of the produced clusters, in addition to the single-cluster result for the 1500-sequence dataset when the Normalized Laplacian was used.

Although UCLUST failed to produce any accurate clustering in this experiment, it scored a fairly good ARI for its results on the last two datasets where the spectral clustering failed. A closer look on the clusterings shows that UCLUST produced at least four clusters out of six which were highly pure[6]. The two remaining clusters randomly grouped the remaining elements. These highly pure clusters contributed to the good ARI scores for UCLUST. However, in order to reach these results, the identity threshold had to be carefully selected

[6]The large majority of the elements form a sub-group of a same cluster from the ground truth.

TABLE VII
ADJUSTED RAND INDEX FOR THE CLUSTERING OF THE SIMULATED DATASETS.

| | dataset 8 | | dataset 9 | | dataset 10 | | dataset 11 | | dataset 12 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Nb. Clusters | ARI | Nb. Clusters | ARI | Nb. Clusters | ARI | Nb. Clusters | ARI | Nb. Clusters | ARI |
| BestBIC-NL | 6 | 1 | 6 | 1 | 6 | 0.681 | 5 | 0.369 | 1 | - |
| BestBIC-BH | 5 | 0.772 | 6 | 1 | 4 | 0.426 | 4 | 0.410 | 4 | 0.314 |
| BestBIC-Mod | 6 | 1 | 6 | 1 | 7 | 0.948 | 4 | 0.269 | 2 | 0.014 |
| UCLUST (0.54<id<0.59 - carefully selected) | 7 | 0.810 | 7 | 0.901 | 8 | 0.768 | 6 | 0.864 | 6 | 0.711 |
| CD-HIT (id=0.8 - the minimum supported) | 20 | - | 28 | - | 39 | - | 36 | - | 49 | - |

for each dataset with a strict precision of 0.1. Conversely, CD-HIT which supports the lowest identity threshold of 0.8, failed to produce any clustering close to the ground truth.

This last assessment demonstrated that the spectral clustering technique proves superior in clustering highly divergent sequences. Its success might be limited to a maximum of a few hundred sequences. It also does not require careful, sometimes uncertain, and variable parameter selection compared to the threshold selection of the traditional tools.

## V. CONCLUSION AND FUTURE PERSPECTIVES

In this work, the effects of using four different alignment tools, in the initial stage of the spectral clustering pipeline, were discussed in terms of accuracy, speed, and capacity to handle large sequences. The relevance of using three additional affinity matrices was studied. In contrast with some state-of-the-art tools, the efficiency of the spectral clustering technique, in handling datasets containing different types of sequences, including sequences subjected to Horizontal Genes Transfers, was also proved. Finally, an additional validation of the spectral technique was performed by using larger datasets.

Finally, some possible future extensions to this work include comparing the efficiency of the spectral clustering techniques with the emerging deep learning techniques that are being introduced in the field of clustering biological sequences. Moreover, finding more efficient algorithms that can handle both highly divergent and large data sets can tackle the highlighted drawback of the GMMs. Furthermore, exploring the capabilities of the GMMs on different types of data input, such as network datagrams, could lead to interesting results.

## REFERENCES


[1] M. Lynch, "Evolution of the mutation rate," *TRENDS in Genetics*, vol. 26, no. 8, pp. 345–352, 2010.

[2] T. A. Sun and P. A. Lind, "Distribution of mutation rates challenges evolutionary predictability," *Microbiology*, vol. 169, no. 5, p. 001323, 2023.

[3] S. M. Soucy, J. Huang, and J. P. Gogarten, "Horizontal gene transfer: building the web of life," *Nature Reviews Genetics*, vol. 16, no. 8, pp. 472–482, 2015.

[4] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in bioinformatics*, vol. 21, no. 1, pp. 1–10, 2020.

[5] W. Pentney and M. Meila, "Spectral clustering of biological sequence data," in *AAAI*, vol. 5, pp. 845–850, 2005.

[6] A. Paccanaro, J. A. Casbon, and M. A. Saqi, "Spectral clustering of protein sequences," *Nucleic acids research*, vol. 34, no. 5, pp. 1571–1580, 2006.

[7] X. Hu and I. Yoo, "Cluster ensemble and its applications in gene expression analysis," in *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, pp. 297–302, Australian Computer Society, Inc., 2004.

[8] J. Matar, H. E. Khoury, J.-C. Charr, C. Guyeux, and S. Chrétien, "Spclust: Towards a fast and reliable clustering for potentially divergent biological sequences," *Computers in biology and medicine*, vol. 114, p. 103439, 2019.

[9] H. Xiong and Z. Li, "Clustering validation measures," in *Data clustering*, pp. 571–606, Chapman and Hall/CRC, 2018.

[10] J. Matar, H. ElKhoury, J.-C. Charr, C. Guyeux, and S. Chrétien, "Optimized spectral clustering methods for potentially divergent biological sequences," *Scientific Enquiry and Review*, vol. 8, no. 3, pp. 58–87, 2024.

[11] J. Matar, H. ElKhoury, J.-C. Charr, C. Guyeux, and S. Chrétien, "Biological sequence clustering: novel approaches and a comparative study." Preprint on https://www.researchgate.net/publication/355184531_Biological_sequence_clustering_novel_approaches_and_a_comparative_study, 2021.

[12] M. S. Waterman, "Efficient sequence alignment algorithms," *Journal of theoretical biology*, vol. 108, no. 3, pp. 333–337, 1984.

[13] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[14] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[15] E. S. Wright, "Decipher: harnessing local sequence context to improve protein multiple sequence alignment," *Bmc Bioinformatics*, vol. 16, no. 1, p. 322, 2015.

[16] M. Larking, G. Blackshields, N. Brown, R. Chenna, G. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, *et al.*, "Clustalw and clustalx version 2," *Bioinformatics*, vol. 23, no. 21, pp. 2947–8, 2007.

[17] M. Bruneau, T. Mottet, S. Moulin, M. Kerbiriou, F. Chouly, S. Chretien, and C. Guyeux, "A clustering package for nucleotide sequences using laplacian eigenmaps and gaussian mixture model," *Computers in Biology and Medicine*, vol. 93, pp. 66 – 74, 2018.

[18] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[19] R. Langone, C. Alzate, and J. A. Suykens, "Modularity-based model selection for kernel spectral clustering," in *The 2011 International Joint Conference on Neural Networks*, pp. 1849–1856, IEEE, 2011.

[20] A. Saade, F. Krzakala, and L. Zdeborová, "Spectral clustering of graphs with the bethe hessian," in *Advances in Neural Information Processing Systems*, pp. 406–414, 2014.

[21] L. Dall'Amico, R. Couillet, and N. Tremblay, "Optimized deformed laplacian for spectrum-based community detection in sparse heterogeneous graphs," *arXiv preprint arXiv:1901.09715*, 2019.

[22] L. Dall'Amico, R. Couillet, and N. Tremblay, "Revisiting the bethe-hessian: improved community detection in sparse heterogeneous graphs," in *Advances in Neural Information Processing Systems*, pp. 4037–4047, 2019.

[23] S. Khurana, D. N. Krementsov, A. de Parseval, J. H. Elder, M. Foti, and M. Thali, "Human immunodeficiency virus type 1 and influenza virus exit via different membrane microdomains," *Journal of virology*, vol. 81, no. 22, pp. 12630–12640, 2007.