Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100432

New Hierarchical Clustering Algorithm for Protein Sequences Based on Hellinger Distance

Gamil Abdel-Azim[*]

Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

Received: 10 Sep. 2015 , Revised: 23 Oct. 2015, Accepted: 24 Oct. 2015
Published online: 1 Jul. 2016

Abstract: Protein sequences clustering based on their sequence patterns has attracted lots of research efforts in the last decade. The principal idea of most clustering systems is how to represent and interpret protein sequences, which principally determines the performance of classifiers. In this paper, we proposed a new methodology, that definite a new descriptor to represent and interpret each sequence using its Probability Densities Functions (PDF). The Hellinger distance is used to measure the similarity between the sequences. Afterward, a hierarchical algorithm is applied to clustering proteins sequences using the Hellinger distance. Two of protein data sets are using for the experiments; the first is a mixed between Influenza and Ebola virus and the second is a set of Influenza. We compare between a two Hierarchical Clustering Algorithms, The first based on similarity measure is to use methods with sequences alignments (HCAWSA). The second is the proposed approach to the similarity measure is to use methods without sequences alignments.(HCAWOSA). The experiments result show that the proposed methodology is feasible and achieves good accuracy.

Keywords: Hierarchical clustering; Probability Densities Function; Hellinger distance; Protein sequences; Classification

1 Introduction

The increase in biological sequences information resulting from the development of advanced biotechnology and quantity of genetic information is more rapidly than the speed at which it can analyze [1]. Clustering techniques offer a workable solution to handling and analysis of these growing rapidly genetic data. Clustering algorithms separate the sequences into different significant groups biologically, thereby facilitating prediction of the sequences functions such that proteins function and genes functions [1,2,3].

When a new protein assigned to a cluster, the biological properties of this cluster can be attributed to this protein with high confidence. On the other hand, clustering of protein sequences can also help analyze the evolutionary relationships between the sequences in a cluster [2,3].

Clustering sequences of proteins need a computing of similarity between sequences. There are two approaches for clustering according to the measure of similarity used in a method of clustering. The first based on the sequence alignment. The similarity between two sequences of proteins measured by scores derived from an alignment algorithm such as BLAST [4] or FASTA [5]. While the alignment of the sequence yields good solutions, it is relatively difficult to assemble a large number of sequences because it's computational complexity. Also, if the sequences vary in length, satisfactory alignment is difficult to realize, resulting in low the clustering accuracy. The second approach to the similarity measure is to use methods without alignments- [6,7,8,9,10,11]

In recent years, several measures without alignments have been proposed for more information see [12,13,14, 15]. Different evaluating of the similarity between two vectors is used. We site for example the Euclidean distance [16], the Mahalanobis distance [17], Kullback-Leibler divergence [18], the cosine distance [19] and the correlation coefficient of Pearson [20]. Major algorithms used in biological sequences clustering can divide into two categories according to the result format: partition clustering algorithms and hierarchical clustering algorithms [21].

Hierarchical classification widely used for detecting clusters in genomics data. It generates a set of partitions that form a cluster hierarchy. According to linkage criteria, there are three hierarchical clustering methods including single-linkage clustering (SL), complete linkage clustering (CL) and average linkage clustering

[*] Corresponding author e-mail: gazim3@gmail.com

(AL) [22]. With SL, clusters can be merged due to sequence only be close to the other, although most of the sequences of each group may be very distant from each other [23]. With CL, all sequences in the cluster similar to each other. AL viewed as an intermediate between the complete linkage clustering and the single-linkage, resulting in more homogenous than those obtained by single-linkage. BlastClust [24] and GeneRage [25] are employed the single linkage clustering approach. SWORDS [26] based on the profile of word frequencies to merge clusters hierarchically; and Uchiyama [27] use the average linkage clustering algorithm to classify DNA sequences. The clusters are formed such that the objects in the same cluster are very similar, and the objects in different clusters are very distinct. The similarity measure between objects must be selected and criterions function, which minimize the similarity between the objects that belonging to the same cluster and maximize the similarity between the objects of different clusters. [28,29,30].

The different hierarchical clustering methods differ the way they define the distance between already computed clusters, or between clusters and individual sequences. Thus, we have: Nearest neighbor (single linkage), the furthest neighbor (complete linkage) and an average neighbor (average linkage)[29,30].

The basic hierarchical clustering algorithm (BHCA) proceeds as follows:

1. Compute all pair-wise distances between the sequences
2. Merge the sequences that are closest (most similar) to each other
3. Compute the distance between the newly created cluster and all other sequences/clusters
4. Repeat 2.

In this article, a new similarity measure is defined based on a new descriptor without alignments. We consider the Probability Densities Functions (PDF) of each sequence as a descriptor to present the sequence. Then, the Hellinger distance between the PDF's of the sequence is calculated to measure the similarity between the sequences.

The paper is organized as follows, In Section 2 we explain the Hellinger distance, the definition of the descriptor for each sequence, a new similarity (Computing the PDF) and a general methodology of hierarchical clustering methods. Section 3, experimental results that contains datasets description, evaluation of similarity measure of the proposed algorithm, and discussion In Section 4, the article finishes with the conclusion.

In the following section, we are presenting Hellinger distance and the computing of Probability Densities Functions (PDF) for each sequence.

2 Distance Metrics- Hellinger distance

The concept has been developed to provide a metric for the distance between two different discrete probability distributions P and Q. See [31,32,33].

Define as follows:

$$D^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i}-\sqrt{q_i})^2 \tag{1}$$

Note that P and Q are described as N-tuples (vectors) of probabilities, where $P = p_{1,}p_{2,\ldots,}p_N$ and $Q = q_{1,}q_{2,\ldots,}q_N$, p_i and q_i are assumed to be non-negative real numbers such that $\sum_i p_i = 1$ and $\sum_i q_i = 1$.

Hellinger distance is a metric quantity, which means that it has the properties of non-negativity, the identity, and symmetry, besides, to obey the triangle inequality. See [31,32,33]. The Hellinger distance between two variables can be computed between two variables if we have explicit knowledge of the probability distributions. In general, these probabilities are not known. There are various methods for estimating the probability densities from observed data. See [31,32,33]. In this paper, we are calculating the exact probability densities functions for every protein sequence.

Given a series x_i and y_i of n simultaneous observations for two random variables X and Y. Let $f_X(i)$ denote the number of observations i in X. The probabilities then estimated as:

$$p_i = \frac{f_X(i)}{n} \tag{2}$$

Let $f_Y(j)$ denote the number of observations of j in Y. The probabilities are estimated as:

$$q_i = \frac{f_Y(j)}{n}, \tag{3}$$

where, Hellinger distance is computed using discrete probabilities. Then Hellinger distance between X and Y is computed as:

$$D^2(X,Y) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i}-\sqrt{q_i})^2 \tag{4}$$

2.1 Sequence descriptor (Probability Densities Functions (PDF))

We are calculating the PDF for every sequence, and then defined the distance between the sequences using equation (4). Where PDF defined as follows:

$f : P_r \to [0,1]^n$, $f(s) = (p_m,\ m = 1,2,\ldots,20)$ and $\sum_m p_m = 1$, where P_r is the set of proteins sequences see Ref. [33]. Firstly we describe the PDF for the sequence as the following:

Compute the PDF of sequence S
Find $p_m = \frac{N_m}{\ln(S)}, m = 1, 2, ..., 20.$
Where $\ln(S)$ is the length of the sequence S,
N_m is the number of letter m in the sequence,
m belongs to the proteins alphabets.

Any clustering task characterized by three principles pattern representation, definition of a similarity measurement and clustering algorithm [34]

2.2 A new similarity measure

Hellinger distance between the PDF's sequences S_i and S_j are computed as following:

Hellinger distance computation (H-distance)
Find PDF of the two sequences S_i and S_j
Calculate the Hellinger distance between S_i and S_j using equation 4

That is reducing the dimensionality of the features that represent the protein sequence. We are applying the following strategy for implementing Hierarchical clustering techniques.

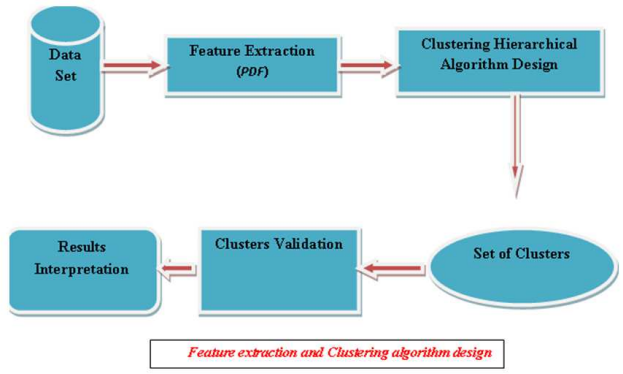

Fig. 1: The proposed Hierarchical clustering Strategy

derived from human, birds and pigs, collected during 1918-2014 years in Canada downloaded from Influenza Virus Resource database National Center for Biotechnology Information (NCBI). Up to 2014 the database of Influenza Virus Resource consists of about 300000 sequences (full genomes, sequences of RNA, proteins). The Ebola virus proteins sequences are collected and downloaded from (NCBI).

The proposed algorithm implemented in MATLAB program developed by the author. Also, the MATLAB dendrogram function is used to compute and display a hierarchy of clusters that depends on the Hellinger distance. Fig 2 and Fig.3 illustrate the PDF of two proteins sequences.

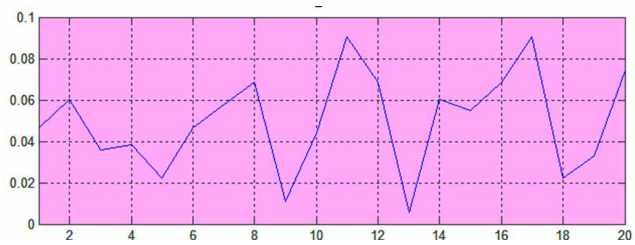

Fig. 2: PDF of protein sequence where the length =230

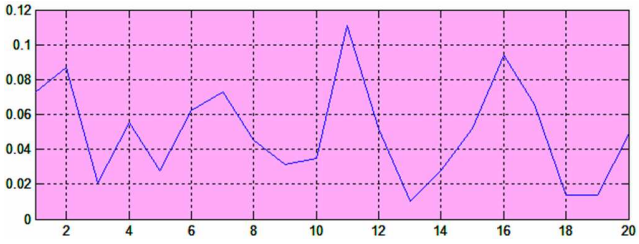

Fig. 3: PDF of protein sequence where the length =288

3 Experimental Results and Discussion

To evaluate the proposed similarity measure, we used two datasets Influenza virus families and mixed dataset from Influenza and Ebola virus.

3.1 Data Description

The dataset I: Mixed from Influenza and Ebola virus consists of 2102 sequences. The dataset I consists of 1417 Influenza and 685 Ebola virus.

Dataset II: Influenza virus families. Amino acid sequences of subtypes of hemagglutinin influenza A

3.2 Evaluation of similarity measure

Classifications presented in the figure of hierarchical trees also called a dendrogram. A dendrogram is a graphical representation of a ultrametric matrix (= cophenetic); then the dendrograms can compare with each other by comparing their cophenetic matrices [35].

Given the original data $\{X_i\}$ that has been modeled using a clustering method to produce a dendrogram $\{T_i\}$. Set the following distance measurements $x(i, j) = |X_i - X_j|$ is the ordinary Euclidean distance

Table 1

Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
0.916	0.641	0.929	0.629	0.644	0.751	5.467

Table 2: title

Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
0.992	0.929	0.984	0.979	0.834	0.720	7.326

Table 3

Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
0.957	0.997	0.911	0.924	1.000	0.560	8.295

Table 4

Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
0.894	0.995	0.937	0.943	1.000	0.565	8.274

between i and j^{th} observations. The $t(i,j)$ = dendrogrammatic distance between T_i and T_j model points. This distance is the height of the node at which these two points first joined. Letting x be the mean of $x(i,j)$ and t is the average of the $t(i,j)$, the correlation coefficient cophenetic c is defined as follows [36].

$$c = \frac{\sum_{i<j} (x(i,j)-x)\,(t(i,j)-t)}{\sqrt{\left\{\sum_{i<j} (x(i,j)-x)^2\right\}\left\{\sum_{i<j} (t(i,j)-t)^2\right\}}} \qquad (5)$$

Since its introduction by Sokal and Rohlf [37], the cophenetic correlation coefficient has been widely used as criteria of the effectiveness of different clustering techniques [38].

The dendrogram function MatLab can display any number of points; However, dendrograms of data sets with more than 30 points may be incomprehensible for reading. Only 30 nodes (sequences) used in the display of the dendrograms for illustrating the examples figs. 4, 5, 6 and 7, and Tables 1, 2, 3 and 4.

Coph1, Cor1, and TWOA are the Cophenetic coefficients, correlation coefficient, and execution time for HCAWOSA respectively. Coph2, Cor2, and TWA are the Cophenetic coefficients, correlation coefficient, and time execution for HCAWSA respectively. B-coefficient is the similarity index between the two clustering algorithms

Table 5

Data set I	S1	Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
Average Method	10	0.97	0.90	0.93	0.93	0.86	0.81	0.85
	15	0.96	0.87	0.93	0.92	0.83	0.89	2.15
	20	0.96	0.86	0.94	0.93	0.84	0.83	3.58
	40	0.95	0.83	0.94	0.91	0.80	0.85	14.59
Single Method	10	0.97	0.91	0.93	0.93	0.89	0.81	0.87
	15	0.95	0.86	0.92	0.92	0.85	0.87	2.14
	20	0.93	0.83	0.91	0.90	0.83	0.82	3.46
	40	0.92	0.81	0.92	0.89	0.83	0.88	15.70
Data set II	S9	Coph1	coph2	Cor1	Cor2	B	TWOA	TWA
Average Method	10	0.93	1.00	0.90	0.93	0.93	0.62	1.01
	15	0.92	1.00	0.90	0.94	0.92	0.62	2.48
	20	0.90	1.00	0.90	0.90	0.90	0.60	4.30
	40	0.88	0.99	0.88	0.94	0.88	0.61	17.99
Single Method	10	0.90	1.00	0.89	0.94	0.92	0.60	0.97
	15	0.89	0.99	0.89	0.94	0.93	0.61	2.43
	20	0.87	0.99	0.88	0.94	0.90	0.58	4.24
	40	0.82	0.99	0.85	0.94	0.86	0.62	17.93

Table 5 that regroup the average of 100 runs for all coefficients comparison (Coph1, coph2, Cor1, Cor2, B, TWOA, and TWA).

The following Fig s 8, 9,10 and 11 resume the average of 100 runs of the execution time for the two algorithms HCAWOSA and HCAWSA. TWA and TWOA are the execution time for HCAWSA and HCAWOSA respectively.

According to Table 5 and Figs 8,9,19 and 11 the proposed algorithms gives the best results. Both in terms of execution time and quality clustering which based on the high value of B coefficients.

3.3 Discussion

In this study, we compare between two Hierarchical Clustering Algorithms:

- The first based on similarity measure, which need to sequences alignments for calculating the distance between two sequences (HCAWSA))
- The second is the proposed algorithm based on a new similarity, which need to calculate PDF for calculating Hellinger distance between two sequences, without sequences alignments. (HCAWOSA)). The two algorithms implemented with single and average methods.

The comparison based on Cophenetic coefficients, *B*-coefficients and execution times. *B*-coefficients is the similarity index proposed in [39], B belongs to $[0\ 1]$ when $B = 1$ that meaning perfect matching between the two

Example 1:

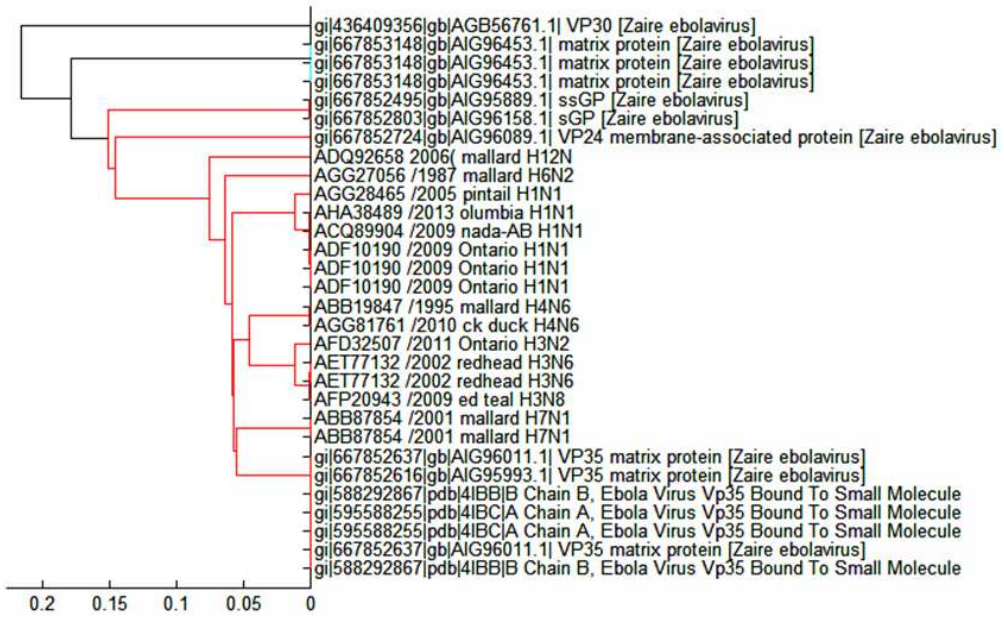

Fig. 4: 30 sequence randomly taken from dataset I with single method

Example 2:

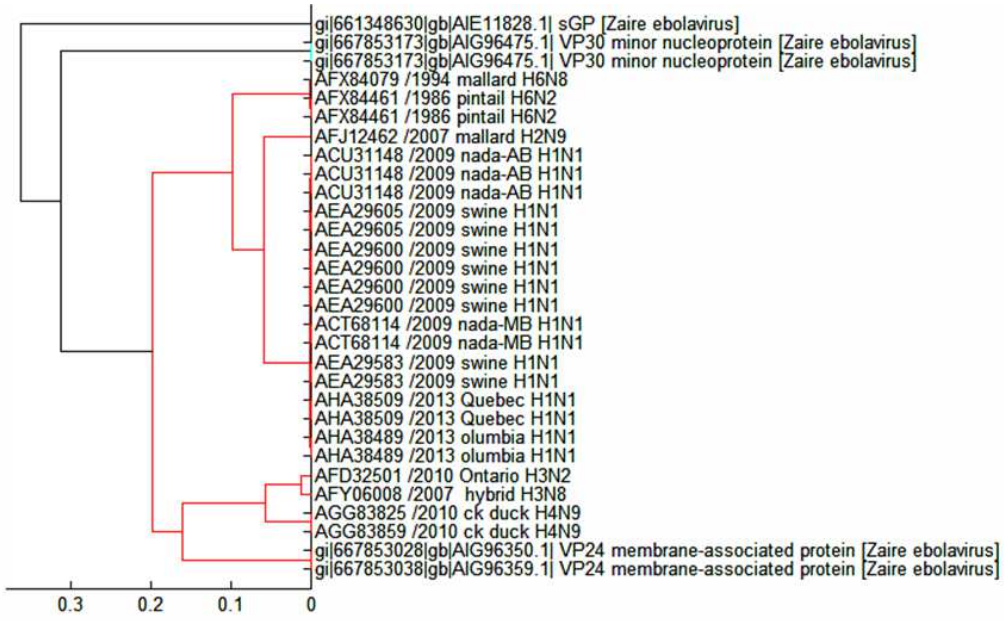

Fig. 5: 30 sequence randomly taken from dataset I with average method

partitions and $B = 0$ no matching between the two partitions. The two algorithms applied for different numbers of sequences $n = 10, 15, 20, 30$ and 40. And for each size the program run 100 times. See Table 5 that regroup the average of all runs for all coefficients comparison.From the Figs 8,9,19 and 11 we see that the HCAWOSA (proposed algorithm) gives the best results of execution time that is a linear, where the HCAWSA is exponential. From Table 5, the quality of clustering of HCAWOSA is best and much closed to HCAWSA, which based which based on the high value of B coefficients.

Example 3:

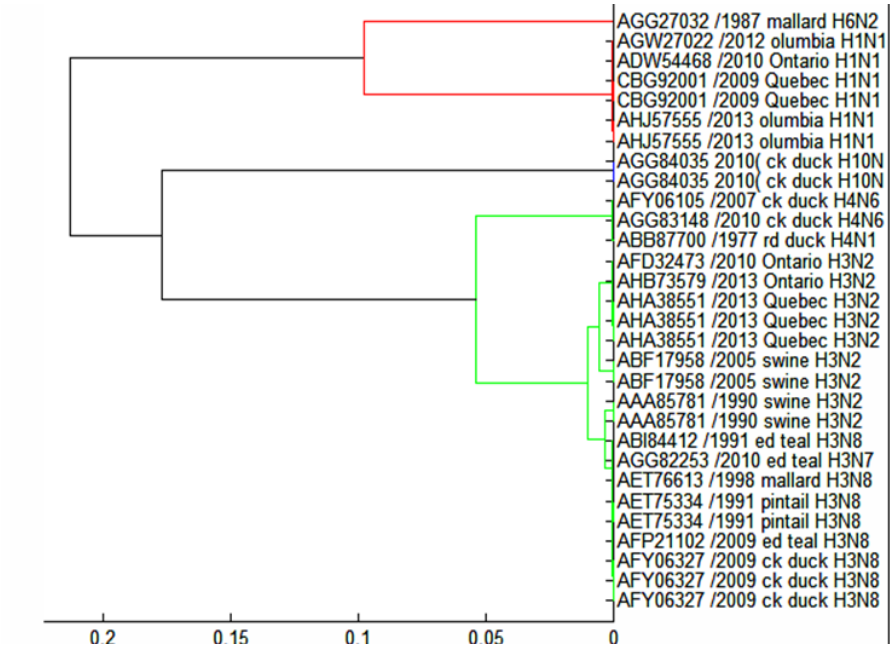

Fig. 6: 30 sequence randomly taken from dataset II with average method

Example 4:

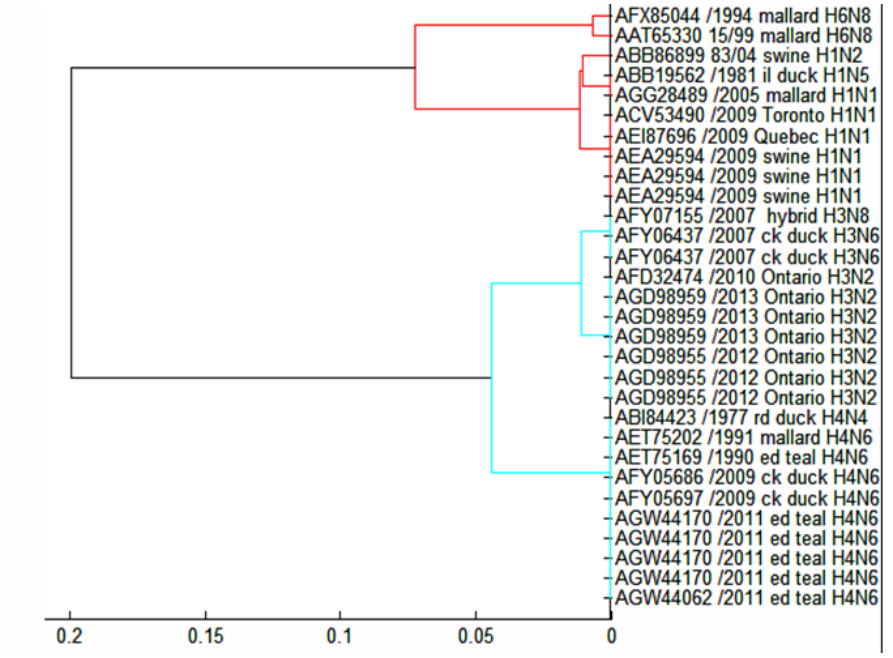

Fig. 7: 30 sequences randomly taken from dataset II with single method

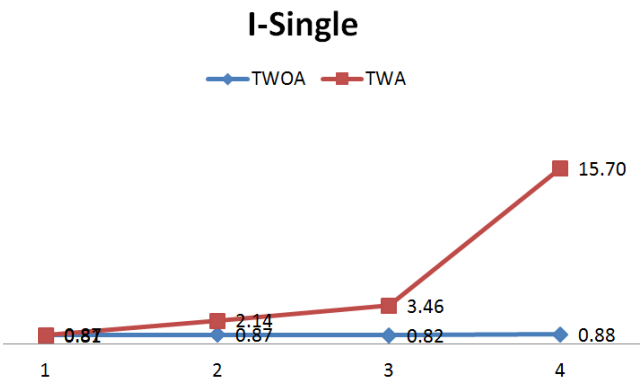

Fig. 8: Execution time average for 100 runs ($N = 10, 15, 20, 40$) dataset I and Single method

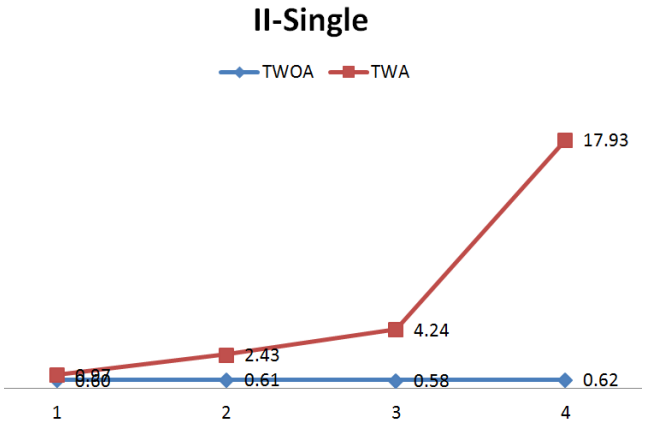

Fig. 10: Execution time average for 100 runs ($N = 10, 15, 20, 40$) dataset II and Single method

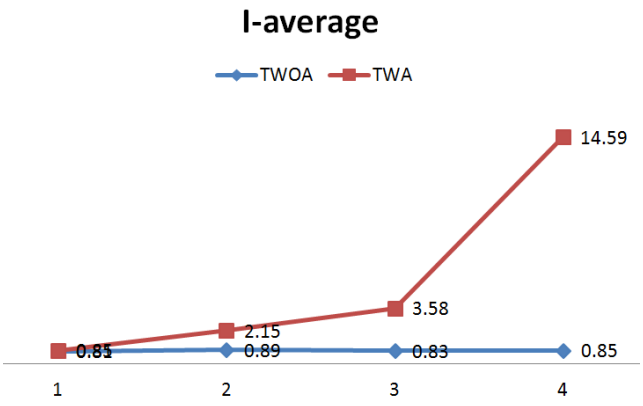

Fig. 9: Execution time average for 100 runs ($N = 10, 15, 20, 40$) dataset I and average method

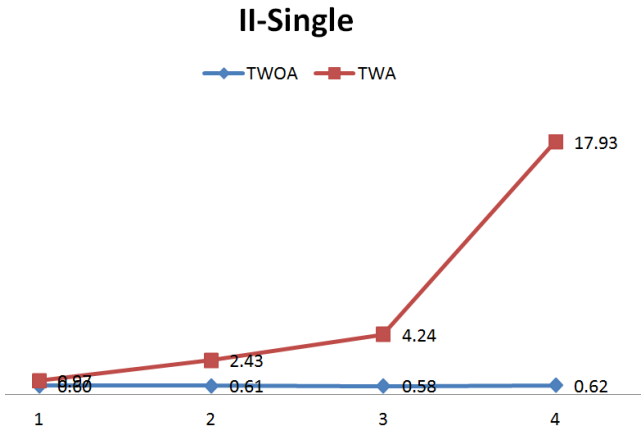

Fig. 11: Execution time average for 100 runs ($N = 10, 15, 20, 40$) dataset II and average method

4 Conclusion

We propose a clustering algorithm that based on a new sequence similarity measure. It is effective in classifying proteins sequences with similar biological characteristics. We are implementing the proposed algorithm for classifying the subtypes of hem agglutinin influenza A using more than 2,000 viral sequences proteins. Two different algorithms ('average' and 'single') utilized PDF and Hellinger distance achieve a good performance in the Hierarchical classification of influenza virus proteins and Ebola Virus. We are observing that a good separation. The results of the proposed algorithm show that the effectively for the classification of proteins sequences in terms of execution time and quality clustering. We compare between a two Hierarchical Clustering Algorithms, The first based on similarity measure is to use methods with sequences alignments (HCAWSA). The second is the proposed approach to the similarity measure is to use methods without sequences alignments (HCAWOSA). The experiments result prove that the proposed methodology is feasible and achieves good accuracy.

References

[1] Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW: The evolution of mammalian gene families. PLoS One 2006, 1:1–10.

[2] Zhao B, Duan V, Yau SS: A novel clustering method via nucleotide-based Fourier power spectrum analysis. JTheor Biol 2011, 279:83–89

[3] Dan Wei, Qingshan Jiang, Yanjie Wei2 and Shengrui Wang, A novel hierarchical clustering algorithm for gene sequences, BMC Bioinformatics 2012, 13:174.

[4] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: A basic local alignment search tool. JMol Biol 1990, 215:403–410.
[5] Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. ProcNatlAcad Sci USA 1988, 85(8):2444–2488.
[6] Vinga S, Almeida J: Alignment-free sequence comparison-a review. Bioinformatics 2003, 19(4):513–523.
[7] Haubold B, Reed FA, Pfaffelhuber P: Alignment-free estimation of nucleotide diversity. Bioinformatics 2011, 27(4):449–455.
[8] Liu Z, Meng J, Sun X: A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. Biochem Biophys Res Commun 2008, 368(2):223–230.
[9] Domazet-Loso M, Haubold B: Efficient estimation of pairwise distances between genomes. Bioinformatics 2009, 25(24):3221–3227.
[10] Domazet-Loso M, Haubold B: Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics 2011, 27(11):1466–1472.
[11] Kelil A, Wang S, Brzezinski R, Fleury A: CLUSS: Clustering of protein sequences based on a new similarity measure. BMC Bioinformatics 2007, 8:286.
[12] Reinert G, Chew D, Sun FZ, Waterman MS: Alignment-free sequence comparison (I): statistics and power. JComput Biol 2009, 16(12):1615–1634.
[13] Dai Q, Liu X, Yao Y, Zhao F: Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. JTheor Biol 2011, 276(1):174–180.
[14] Lu G, Zhang S, Fang X: An improved string composition method for sequence comparison. BMC Bioinformatics 2008, 9(Suppl 6):S15.
[15] Aita T, Husimi Y, Nishigaki K: A mathematical consideration of the wordcomposition vector method in comparison of biological sequences. BioSystems 2011, 106:67–75.
[16] Blaisdell BE: A measure of the similarity of sets of sequences not requiring sequence alignment. ProcNatlAcad Sci USA 1986, 83:5155–5159.
[17] Wu TJ, Burke JP, Davison DB: A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 1997, 53(4):1431–1439.
[18] Wu TJ, Hsieh YC, Li LA: Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 2001, 57(2):441–448.
[19] Stuart GW, Moffett K, Baker S: Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 2002, 18(1):100–108.
[20] Fichant G, Gautier C: Statistical method for predicting protein coding regions in nucleic acid sequences. ComputAppl Biosci 1987, 3(4):287–295.
[21] Dong G, Pei J: Classification, clustering, features and distances of sequence Data. Sequence Data Mining 2007, 33:47–65
[22] Sokal RR, Rohlf FJ: Biometry: The Principles and Practice of Statistics in Biological Research. 3rd edition. New York: W. H. Freeman and Company; 1995.
[23] Everitt BS, Landau S, Leese M: Cluster Analysis. London: Oxford University Press; 2001.
[24] National Center for Biotechnology Information (NCBI): Documentation of the BLASTCLUST-algorithm. ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html.
[25] Enright A J, Ouzounis CA: GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 2000, 16(5):451–457.
[26] Chaudhuri P, Das S: SWORDS: A statistical tool for analyzing large DNA sequences. J Biosci 2002, 27(1):1–6.
[27] Uchiyama I: Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res 2006, 34(2):647–658.
[28] Richard C. Dubes and Anil K. Jain, (1988), Algorithms for Clustering Data, Prentice Hall.
[29] Anderberg MR (1973). Cluster Analysis for Applications. Academic Press, New York.
[30] Everitt BS, Landau S, Leese M, Stahl D (2011). Cluster Analysis. 5th edition. John Wiley & Sons.
[31] Donoho, D. & Liu, R. (1988). The 'automatic' robustness of minimum distance functionals, Annals of Stat., Vol. 16, (1988), pp. 552-586, ISSN 0090-5364.
[32] Giet, L. & Lubrano, M. (2008). A minimum Hellinger distance estimator for stochastic differential equations: an application to statistical inference for continuous time interest rate models, Comput. Stat. & Data Anal., Vol. 52, No. 6, (Feb. 2008), pp. 2945-2965, ISSN: 0167-9473.
[33] Azim, G. A, Aboubekeur Hamdi-Cherif, Mohamed Ben Othman and Z.A. Abo-Eleneen "Protein Progressive MSA Using 2-Opt Method" Systems and Computational Biology - Bioinformatics and Computational Modeling 2011 InTech September 2011
[34] Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A review. ACM Computing Surveys 31 (1999)
[35] Lapointe, FJ, Legendre, P: Comparison tests for dendrograms: a comparative evaluation. J. Classif. 1995, 12, 265-282
[36] Sinan Sarali, Nurhan Dogan and Ismet Dogan, Comparison of hierarchical cluster analysis methods by cophenetic correlation, Journal of Inequalities and Applications 2013, 2013:203.
[37] Sokal, RR, Rohlf, FJ: The comparison of dendrograms by objective methods. Taxon 11, 33-40 (1962)
[38] Farris, JS: On the cophenetic correlation coefficient. Syst. Zool. (1969) 18(3), 279-285
[39] E. B. Fowlkes & C. L. Mallows ,A Method for Comparing Two Hierarchical Clusterings, Journal of the American Statistical Association, 1983: 78:383, 553-569

Gamil Abdel Azim was born in Beni Seuf, Egypt. He received his BSc in Mathematics from Cairo University and a DEPS (Diplôme des Etudes Pratiques Supérieures) from Poitiers University, France. He received MSc. And Ph.D. degrees in Computer Science from Paris Dauphine University, France, in 1988 and 1992, respectively. Recently,

He received DESS Diploma in Bioinformatics, the University of Québec at Montreal (UQAM), 2014. He is currently an Associate Professor in the Department of Computer Sciences, College of Computers and Informatics, Suez Canal University, Egypt. His Current research interests have evolved into three categories that are interconnected. The three areas of interest are Optimization and Neural Networks (Combinatorial, linear and non-linear programming), Bioinformatics, Computational Intelligence and pattern recognition (Identification-clustering-Classification-Medical images Segmentation, Machine learning). Dr. Gamil is Member of IEEE.