Evaluating the Significance of Global and Local Features in Expressed Sequence Tag: A Clustering Quality Perspective

Keng-Hoong Ng, Somnuk Phon-Amnuaisuk, and Chin-Kuan Ho

***Abstract*—Clustering of expressed sequence tag (EST) plays an important role in gene analysis. Alignment-based sequence comparison is commonly used to measure the similarity between sequences, and recently some of the alignment-free comparisons have been introduced. In this paper, we evaluate the role of global and local features extracted from the alignment free approaches i.e., compression-based method and generalized relative entropy method, in the quality of EST clustering perspective. Our evaluation shows that the local feature of EST yields much better clustering result compares to the global feature.**

***Index Terms* - sequence clustering, alignment-free, similarity measure, grammar-based distance, generalized relative entropy**

I. Introduction

Expressed sequence tags (ESTs) were introduced in the early 90's and they represent a significant advancement in modern biology [1]. This high-throughput technology provides the continuous flow of EST data that forms one of the richest resources for discoveries in genetics. An Expressed sequence tag is a tiny portion of an entire gene, it is produced by one-shot sequencing of a cDNA clone [2]. The cDNA clone is produced from a mRNA library. ESTs are easy to produce and they are valuable resources for different kinds of gene analysis e.g. gene identification, analysis of gene expression and gene structure analyis.

The characterisitics of EST are low quality, high redundancy and short sequences. Therefore, unprocessed ESTs will not give any important information on gene analysis [3]. Clustering is usually the first step in EST data mining. It is a process of grouping ESTs that originate from the same gene. The goal is to to construct gene indices, where all expressed data are partitioned into index classes such that expressed data are put into the same index class if and only they represent the same gene [4]. The ESTs in one cluster can be assembled to generate one or more consensus sequences [5]. Publicly available databases such as Unigene (http://www.ncbi.nlm.gov/unigene) and the Institute of Genome Research (http://www.tigr.org) accumulate and store the clustered EST data for gene research.

Methods employ in sequence clustering are commonly based on sequence comparison by alignment, which assumes conservation of contiguity between homologous segments. This alignment approach generates a similarity score, and this score can be calculated using BLAST [6] or FASTA [7]. TIGR uses this method for EST clustering, where it identifies all sequence overlaps using BLAST and FASTA [8].

Manuscript received January 7, 2009. K. H. Ng, S. P. Amnuaisuk, and C. K. Ho are with the Faculty of Information Technology, Multimedia University, Cyberjaya, Selangor, Malaysia (email: khng@mmu.edu.my; somnuk.amnuaisuk@mmu.edu.my; ckho@mmu.edu.my).

Unigene [9] is another established player that uses the alignment method, where sequences are compared with the Smith-Waterman algorithm.

Although the alignment method gives satisfactory solutions, it is unfeasible to use it for long sequences because the computational load escalates as a power function of the sequence length [10]. The second drawback is the approach only considers local mutations of the genome, therefore it is not suitable to measure events and mutations that involve longer segments of genomic sequences [11]. For this reason alignment free sequence comparison has been recently introduced. We outline several non-alignment based clustering algorithms that are currently available.

d2_cluster [12] is a well-known agglomerative clustering method used to cluster ESTs. It is considered as a non-alignment based scoring method. The method begins with every sequence in a singleton cluster, and the clusters will be merged based on a series of similarity comparisons. d2_cluster performs clustering according to the minimal linkage or transitive closure rules. The latter rule means that sequence *A* and sequence *B* are in the same cluster even if they share no similarity but there exists a sequence *C* with enough similarity to both *A* and *B*. The criterion for joining clusters is based on the word matching percentage within a window size. The clustering process finishes after *n* (number of sequences) iterations of merging.

Another clustering algorithm that is alignment free is Xsact [13]. It uses a suffix array, a lexicographically ordered array of all suffixes of the EST sequences. Radix sort is used to generate the suffix array, and then it will be used to find pairs of ESTs with long common substrings. Xsact calculates a score by finding the longest set of consistent matching substrings between each pair of EST. The clustering starts with the highest scoring pairs, where EST pairs above a certain similarity score are merged into a single cluster hierarchically. Clusters are then split according to the clustering threshold. The performance of this algorithm in terms of clustering quality is comparable to d2_cluster and alignment-based clustering, but it requires higher memory for the suffix construction.

Clustering algorithm such as ESTmapper [14] reads genome sequence and converts it into an eager WOTD (write only, top down) suffix tree. Each EST is mapped using the generated suffix tree, where it finds the long common substrings with the genome. The algorithm examines the list of common substrings and locations, and then combines substrings into a single gapped matching region if two common substrings are adjacent when mapped onto the genome. The longest matching region is used to determine the mapping of all ESTs to a location in the genome. ESTs are clustered if their sequences overlap or at nearby location in the genome. ESTmapper is efficient since ESTs can be compared to a suffix tree in linear time but its drawback is the consumption of large amount of memory.

II. RELATED WORK

In this paper we survey and focus on some recent approaches used to define alignment-free distance measures of sequences. These features can be used to perform clustering of sequences. A good distance measure is expected to give satisfactory results with most of the available clustering algorithms. We review alignment-free sequence comparison methods based on counting the word frequencies, based on information theory and also based on data compression technique. We briefly describe some of the clustering methods at the end of this section.

A. Methods based on word frequencies

These methods transform a sequence into an object on which the analytical tools available in Linear Algebra and Statistical Theory can be applied. It starts with the mapping of sequences to vectors defined by the *k-tuple* counts. The obtained vectors represent the original sequence with the fixed word length *k*. The basic idea for this sequence comparison is that similar sequences will share common words, and then it can be quantified by many techniques. Blaisdell [15] is the pioneer who published sequence comparison report based on *k-tuple* counts, where the difference between two sequences was quantified by the Euclidean distance calculated between their word frequencies. For each word length *k* (or resolution), the Euclidean distance between sequence *P* and *Q* is defined as:

$$D_L(P,Q) = \sum_{i=1}^{j} (c^P_{L,i} - c^Q_{L,i})^2 \quad (1)$$

The c^P_L and c^Q_L represent the word counts for the sequences and *j* is the number of possible *k-tuple* for the resolution *k*. For instance, the maximum number of *k-tuple* for word length = 3 is 64 (4^3). Even though this approach is alignment-free, but it is still length dependent in the sense that sequence comparisons are made for a fixed word-length. In fact, it can be recognized as local alignments between identical segments of sequences [11]. In [16], the distance based on word frequencies is regarded as a filtration method for sequence alignment algorithms. It increases the efficiency of the latter because it eliminates low similarity sequences which will directly reduce the input to the dynamic programming algorithm for sequence alignment.

Once this distance measure is established in sequence comparison, several methods originate from *k-tuple* frequencies are also quickly proposed. In [17], classification of proteins is based on di-peptide frequencies. It calculates the linear correlation coefficient between two sequences, from *k-tuple* frequencies and uses the conventional Pearson formalism. Mahalanobis distance is also introduced in sequence comparison, where it takes into account the data covariance structure [18].

B. Methods based on information theory

In this method, the distance between two sequences is measured based on the *k-tuple* vectors and an information theory based metric is used to quantify the dissimilarity between them. In [19], the Kullback-Leibler discrepancy is proposed and it is computed from the *k-tuple* frequencies between two sequences *P* and *Q*. The equation of KL discrepancy is

$$D_k^{KL}(P,Q) = \sum_{i=1}^{n} f^P_{k,i} \times \log_2\left(f^P_{k,i} / f^Q_{k,i}\right) \quad (2)$$

where $f^P_{k,i}$ is the *k-tuple* frequency of sequence *P*, integer *n* is the number of possible *k-tuples* with resolution *k*. The paper concludes that the KL discrepancy is preferred over the Mahalanobis distance and standard Euclidean distance in terms of computational efficiency, but it is not a good performing metric compared to the latter in the aspect of selectivity and sensitivity.

C. Methods based on compression

The method is based on the basic idea that the more two sequences are similar, the more succinctly one sequence can be described given the other. It means that two sequences are considered close if one sequence is significantly compressible given the information contained in the other sequence. Similarities between sequences can be computed based on the well-known Lempel-Ziv parsing algorithm. In [20], the author introduces a measure of relative entropy between two sequences and it is a variant of the Lempel-Ziv parsing algorithm. Given two sequences *x* and *y*,

$$ZM(y|x) = (w_1, w_2, ..., w_m, w_{m+1}, \ldots, w_n) \quad (3)$$

where $y = w_1 w_2 \ldots w_n$, the block w_m is the longest prefix of $w_m w_{m+1} \ldots w_n$ which occurs as factor in *x*. If such a prefix is different from the empty word, and w_m is the first character of $w_m w_{m+1} \ldots w_n$, otherwise. The integer *n* is the complexity of *y* relative to *x*. The idea is that the number of elements in ZM(*y*|*x*) will be smaller if *x* and *y* are more similar.

Otu and Sayood [21] also introduce a distance measure based on the LZ parsing. Given two sequences *P* and *Q*, consider the sequence *PQ* and its exhaustive history. The number of component needed to build *Q* when appended to *P* is $c(PQ) - c(P)$, where $c(PQ)$ and $c(P)$ denote the number of components in the exhaustive history of sequence *PQ* and *P*. The number will be less than or equal to $c(Q)$, and it is dependent on the degree of similarity between *P* and *Q*. The closer between the two sequences, the fewer steps will be used to build *Q* in the production process of *PQ*. The paper shows that the algorithm constructed consistent phylogenies successfully.

D. Clustering approaches in DNA sequence

We outline three common clustering algorithms that have been used to cluster sequence data. They are hierarchical clustering [22] that operates in a bottom up manner, *k*-means [23] and self-organizing map [24]. Recent works include graph based clustering [25] where it can be naturally cast as a graph optimization problem, and ant-based clustering [26] that treats one gene as a node, every edge is associated with a certain level of pheromone intensity. The co-expression level between two genes determines the pheromone intensity of the edge. Then minimum spanning tree algorithm is used to break the linkages in order to generate clusters.

III. PROPOSED METHODS

We are motivated by the problems encountered in EST clustering and the alignment-free similarity distance measures proposed in some research papers we highlighted

in the above section. Hence, we propose a method to compare and evaluate the performance of derived global feature and local feature of EST in terms of clustering quality. To our best knowledge, there has never been any published work on this so far.

First, we download the dataset containing 850 EST sequences from the Unigene database in the National Center for Biotechnology Information (NCBI) website. The dataset contains ESTs from the cardiac muscle of heart organ in the organism named *Meleagris gallopavo* (turkey). These EST sequences have been grouped into 11 clusters by the Unigene and therefore it is a reliable source to be used for experiment. We use grammar-based distance [27] which is based on LZ compression to represent the global feature, while local feature is extracted from the generalized relative entropy method.

A. Grammar-based sequence distance

In this distance measure, it uses the fact that sequences share commonalities in their sequence structure if they have similar biological properties. In [27], it is used to perform multiple sequence alignment in proteins and promising result is claimed by the authors. Figure 1 gives an overview of the calculation of the grammar-based distance. It starts with the creation of LZ dictionaries for each EST sequence. Initially, the dictionary (G_p) for sequence P is empty; a fragment $f^1 = s_p(1)$ is set to the first residue of the corresponding sequence and it is visible to the algorithm. At ith iteration of the process, if fragment f^i is not reproducible from $s_p(1, \ldots, i-1)$, then f^i will be added to the dictionary $G^i_p = G^{i-1}_p + f^i$, and the fragment is reset. On the contrary, if the current dictionary contains enough rules to produce the current fragment, i.e. $G^i_p = G^{i-1}_p$, then it will not be added to the dictionary. The process continues until the visible sequence is equal to the entire sequence. For example, the dictionary for sequence $P = AACGTACC$ is $\{A, AC, G, T, ACC\}$.

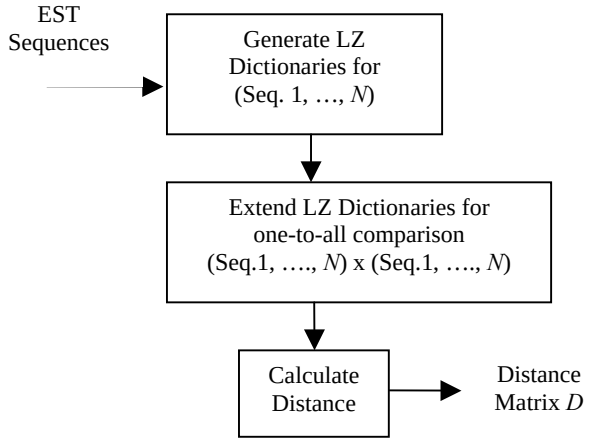

Figure 1. Steps involve in the calculation of grammar-based distance.

Each sequence is compared with all other sequences in the next step to generate the N x N size dictionaries. In this case, consider the comparison of sequence P and R. First, let the dictionary $G^1_{p,r} = G_p$, a fragment $f^1 = s_r(1)$ is set to the first residue of the sequence R, and the visible sequence is all rules in the dictionary of P. The algorithm operates as mentioned above. When it is complete, the new dictionary size will be smaller for sequences with higher similarity.

The final step is the calculation of the distance using the dictionary sizes. The distance measure is based on one of the five suggested methods in [21]

$$d_{p,r} = \frac{H_{p,r} - H_{p,p} + H_{r,p} - H_{r,r}}{\frac{1}{2}(H_{p,r} + H_{r,p})} \quad (4)$$

where $p, r \in \{1, \ldots, N\}$ are the two sequences being compared, and H denotes the dictionary size of a sequence. The matrix distance D is generated from the calculation. This method compresses and builds the dictionary of an EST sequence based on the parsing of entire sequence string. Therefore, this method produces global feature for EST sequences.

B. Distance based on generalized relative entropy

This algorithm is one of the statistical distance measures used in protein or nucleotide sequences. Relative entropy has been explored as similarity measures such as *KLD* (Kullback-Leibler discrepancy) and *SimMM* (Similarity of Markov Models) to compare biological sequences. The drawback of *KLD* is when some entries of vectors are equal to 0 or 1, it becomes unsuitable. We adopt the generalized relative entropy described in [28] as the distance measure for EST sequences. It is denoted by *gre.k* and the following shows the calculation of the *gre.*k distance between sequence P and Q.

$$gre.k(P,Q) = \sum_{i=1}^{n} f^P(w_{k,i}) \times \log_2 \left[\frac{2 \times f^P(w_{k,i})}{f^P(w_{k,i}) + f^Q(w_{k,i})} \right] \quad (5)$$

The $f^P(w_{k,i})$ and $f^Q(w_{k,i})$ are the k-word frequencies of sequence P and Q. The generalized relative entropy can deal with all kinds of k-word frequencies, including 0 and 1. We use this distance measure on several word sizes (k), ranging from 5 to 7 and a distance matrix will be generated for each of them. We use the average *gre.k* distance between two sequences because of the generated distances are not symmetric, i.e. $gre.k\ (P, Q) \neq gre.k\ (Q, P)$. This approach is based on the statistical measures of word frequencies in sequences and therefore it is regarded as local feature of EST sequences.

C. EST Clustering

Visualization is a powerful method for profiling clusters. By plotting the distance matrix into a 2D image [29], cluster can be seen in the image if there are a group of sequences with smaller distance among them. We use the hierarchical clustering algorithm to perform EST clustering based on these distance matrices. The clustering quality is then measured with the non-weighted version of F-measure stated in [30].

IV. RESULT AND DISCUSSION

In this paper, our goal is to evaluate the significance of the global and local features in EST sequences, in the perspective of clustering quality. We assess the benchmark dataset that contains 11 clusters with alignment free methods i.e., grammar-based distance and generalized relative entropy. Initially, we visualize and compare the two methods based on the images plotted from the generated distance

matrices. We further investigate their contributions in EST clustering by performing hierarchical clustering, and the clustering quality of each method is shown in F-measure value.

A. Initial evaluation of features via visualization

We perform an initial evaluation of both methods based on the plotted images. Figure 2 shows the grammar-based distance between sequences while figure 3 displays the generalized relative entropy distance with word size set to 5 and 7. The comparison of the images indicates that the latter distance measure performs better than the former, it is because we can see the 11 physically formed clusters in the images especially for *k*-words equal to 5 and 7. A square or rectangle shape object in the image represents one cluster. These objects are formed due to the distance between EST sequences are small, which are shown in colour i.e. blue indicates small distance (less than 0.3). Furthermore, they also display larger distance with sequences from all other clusters. When comparing the images from figure 3, we can claim that the generalized relative entropy with word size 7 will give better clustering result compares to word size 5. It is because the former not only exhibits small distance among sequences in a cluster, but it also shows larger inter-cluster distance (red colour indicates distance between 0.8 – 1.0) compares to the latter with inter-cluster distance between 0.4 – 0.6 (light green).

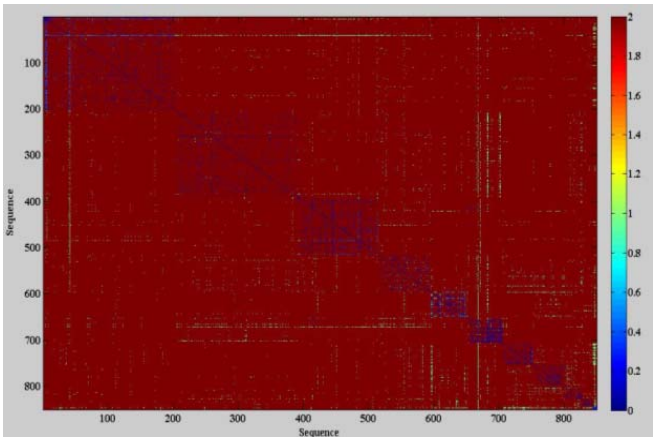

Figure 2. Distance matrix calculated from the grammar-based method.

The initial evaluation via visualization implies that the generalized relative entropy method will outperform the grammar-based method in terms of quality at the clustering stage. Furthermore, it also gives us a hint that the clustering quality in generalized relative entropy with larger word size will be higher compares to the smaller word size. It is because we discover two common things in the method with larger word size i.e. (i) smaller intra-cluster distance, and (ii) larger inter-cluster distance.

B. Evaluation with hierarchical clustering algorithm

The visualization results are verified by the hierarchical clustering algorithm, and then their outputs are evaluated using the F-measure method. Table 1 shows the clustering result of both methods in F-measure value, it is confirmed that the generalized relative entropy method outperforms the grammar-based method. The former obtains 0.8650 and 0.8202 respectively for word size 6 and 7. We did not extend the word size further due to the constraint of computational load. From the result, we can say that the local feature (in this case, the *gre.k*) in EST sequences plays a more important role towards the clustering quality as compares to the global feature (grammar-based method) in EST sequences. The generalized relative entropy with word size 6 gives the best result among all others in terms of clustering quality.

TABLE I. EVALUATION OF BOTH METHODS WITH F-MEASURE

Methods	F-Measure value
Grammar-based	0.1127
Generalized relative entropy (*gre.k*) with	
k-word = 5	0.5650
k-word = 6	0.8650
k-word = 7	0.8202

We further investigate the reasons for the poor performance of the grammar-based distance measure in ESTs. Basically, the ESTs are sequenced from the cDNA library and they are not the complete representation of the parental cDNA [31]. Their length can be varying from sequence to sequence even though they originate from the same cDNA clone. As a result, the variance in length might affect the compression outcome since the EST sequence with larger length tends to produce richer LZ dictionary. Thus, this measure produces unreliable distance among the EST sequences. Another possible reason is the start position for the parsing, where different start positions for the same sequence give other versions of the dictionaries.

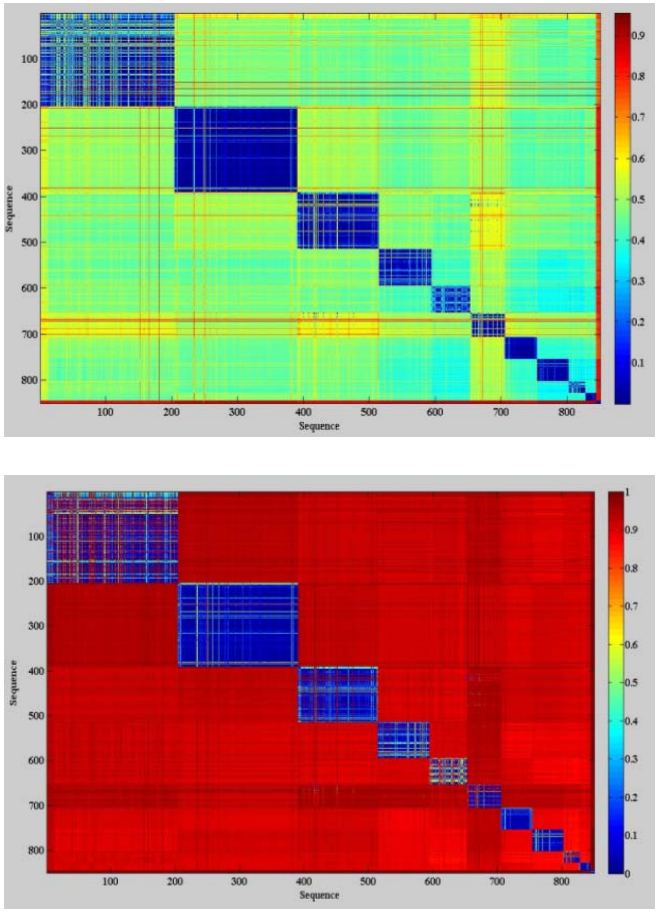

Figure 3. Distance matrices for generalized relative entropy (*gre.k*), with *k*-word is set to 5 (top), and *k*-word is set to 7 (down) respectively.

V. CONCLUSION

In this paper, we presented a method to evaluate the significance of global and local features in ESTs based on the alignment-free distance measures i.e. grammar-based method and generalized relative entropy method. We conclude that the local feature extracted from the generalized relative entropy method outperforms the global feature derived from the former method in terms of clustering quality. In future work we will continue to enhance the EST clustering quality by exploring more alignment-free techniques and clustering algorithms.

REFERENCES

[1] A. Ptitsyn, and W. Hide,"CLU: A new algorithm for EST clustering", *BMC Bioinformatics*, 6, 2005, doi: 10.1186/1471-2105-6-S2-S3.

[2] K. Malde, E. Coward, and I. Jonassen, "A graph based algorithm for generating EST consensus sequences", *Bioinformatics*, vol. 21(8), 2005, pp. 1371 – 1375.

[3] W. Hide, R. Miller, A. Ptitsyn, J.Kelso, C.Gopallakrishnan and A. Christoffels, EST clustering tutorial, SANBI, 1999.

[4] J. P. Burke, H. Wang, W. Hide, and D. Davison, "Alternative gene form discovery and candidate gene selection from gene indexing projects", *Genome Research*, vol. 8, 1998, pp. 276 290.

[5] S. A. Haas, T. Beissbarth, E. Ribals, A. Krause and M. Vingron, "GeneNest: automated generation and visualization of gene indices", *Trends Genet.*, vol.16, 2000, pp. 521 – 523.

[6] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "A basic local alignment search tool", *J. Mol. Biol*, vol.215, 1990, pp. 403 – 410.

[7] D. J. Lipman, and W. R. Pearson, "Improved tools for biological sequence comparison", *Proc. Natl. Acad. Sci. USA*, vol.85(8), 1988, pp. 2444 – 2488.

[8] G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, "TIGR assembler: A new tool for assembling large shotgun sequencing projects", *Genome Sci. Technol*, vol.1, 1995, pp. 9 -18.

[9] M. S. Boguski, and G. D. Schuler, "ESTablishing a human transcript map", *Nat. Genet*, vol.10, 1995, pp. 369 -371.

[10] S. Vinga, and J. Almeida, "Alignment-free sequence comparison – a review", *Bioinformatics*, vol.19(4), 2003, pp. 513 -523.

[11] S. Mantaci, A. Restivo, and M. Sciortino, "Distance measures for biological sequences: Some recent approaches", *Internat. J. Approx. Reason*, 47, 2008, pp. 109 – 124.

[12] J. Burke, D. Davison, and W. Hide, "d2_cluster: A validated method for clustering EST and full length cDNA sequences", *Genome Research*, 9, 1999, pp. 1135 – 1142.

[13] K. Malde, E. Coward, and I. Jonassen, "Fast sequence clustering using a suffix array algorithm", *Bioinformatics*, vol. 19(10), 2003, pp. 1221 – 1226.

[14] X. Wu, W.J. Lee, D.Gupta, and C.W. Tseng, "ESTmapper:Efficiently clustering EST sequences using genome maps", *Proc. of the 19th IEEE International Parallel and Distributed Proceesing Symposium*, 2005, pp. 196a, doi:10.1109/IPDPS:2005.204.

[15] B. E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment", *Proc. Natl. Acad. Sci. USA*, 83, 1986, pp. 5155 – 5159.

[16] P. A. Pevzner, "Statistical distance between texts and filtration methods in sequence comparison", *Comput. Appl. Biosci*., 8, 1992, pp. 121 – 127.

[17] P. Petrilli, "Classification of protein sequences by their dipeptide composition", *Comput. Appl. Biosci.*, 9, 1993, pp. 205 – 209.

[18] T. J. Wu, J. P. Burke, and D. B. Davison, "A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words", *Biometrics*, 53, 1997, pp. 1431 – 1439.

[19] T. J. Wu, Y. C. Hsieh, and L. A. Li, "Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition", *Biometrics*, 57, 2001, pp. 441 – 448.

[20] J. Ziv, N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification", *IEEE Trans. Inform. Theor.*, 39(4), 1993, pp. 1270 – 1279.

[21] H. H. Otu, and K. Sayood, "A new sequence distance measure for phylogenetic tree construction", *Bioinformatics*, 19(16), 2003, pp. 2122 – 2130.

[22] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display genome-wide expression patterns", *Proc. Nat. Acad. Sci. USA*, vol. 95(25), 1998, pp. 14 863 – 14 868 .

[23] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M.Church, "Systematic determination of genetic network architecture", *Nat. Genet.*, vol. 22(3), 1999, pp. 281 – 285.

[24] P. Tamayo, D. Slonim, J. Mesiov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation", *Proc. Nat. Acad. Sci. USA*, vol. 96(6), 1999, pp. 2907 – 2912.

[25] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees", *Bioinformatics*, vol. 18(4), 2002, pp. 536 – 545.

[26] D. Zhou, Y. He, C. K. Kwoh, and H. Wang, "Ant-MST: An ant-based minimum spanning tree for gene expression data clustering", *LNBI*, vol. 4774, 2007, pp. 198 - 205.

[27] D. J. Russell, H. H. Otu, and K. Sayood, "Grammar-based distance in progressive multiple sequence alignment", *BMC Bioinformatics*, 9:306, 2008, doi: 10.1186/1471-2105-9-306.

[28] Q. Tai, and T. Wang, "Comparison study on k-word statistical measures for protein: From sequence to sequence space", *BMC Bioinformatics*, 9:394, 2008, doi: 10.1186/1471-2105-9-394.

[29] R. J. Hathaway, and J. C. Bezdek, "Visual cluster validity for prototype generator clustering models", *Pattern Recognition Letters*, 24, 2003, pp. 1563 – 1569.

[30] G. Dong, and J. Pei, Sequence data mining, vol. 33, Springer US, 2007, pp. 47 – 65, doi: 10.1007/978-0-387-69937-0.

[31] S. Rudd, "Expressed sequence tags: alternative or complement to whole genome sequence?", *Trends in Plant Science*, vol. 8(7), 2003, pp. 321 – 329.