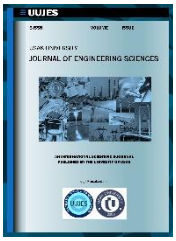


Usak University

Journal of Engineering Sciences

An international e-journal published by the University of Usak

Journal homepage: dergipark.gov.tr/uujes

Research article

THE PHYLOGENETIC RELATIONSHIPS OF LIPASE ENZYMES ACCORDING TO THEIR THERMAL STABILITY BY A COMPUTATIONAL APPROACH

Çağın Kandemir-Çavaş, Türkan Arıt*

Department of Computer Science, Faculty of Science, Dokuz Eylül University, Turkey

Received: 25 April 2018 Revised: 12 May 2018 Accepted: 22 May 2018 Online available: 25 June 2018
Handling Editor: Kemal Mazanoğlu

Abstract

Molecular Phylogenetic Analysis studies conducted to reveal the evolutionary relationships of biological sequences, use two basic methodologies. In the study of distance-based methodologies, phylogenetic trees were obtained by clustering Lipase enzymes according to their thermal stability. The biological sequences to be used in the study were obtained from the NCBI Genbank database. Methods were coded in the Java programming language and distance matrices were obtained. Phylogenetic trees were constructed using the R language. As a result, Lipase enzymes were effectively clustered using a distance-based method without alignment, according to their thermal stability.

Keywords: *Molecular Phylogeny; distance based methodology; enzyme clustering.*

1. Introduction

Bioinformatics is the science of storing, analyzing and interpreting biological data through computer science, mathematics, statistics, and informatics sciences [1]. With the development of DNA sequencing technologies and finally with the completion of the human genome project, the need to analyze and interpret large quantities of biological data is needed. The fact that the size of the data generated in biological researches is too great has led to the need for computer science in order to reach the right knowledge.

Cells are the building blocks of all living organisms. Each cell contains a nucleus in which DNA (deoxyribonucleic acid) molecules are contained in small structures called chromosomes. DNA carries information about how an organism develops from a single cell.

*Corresponding author: Çağın Kandemir-Çavaş
E-mail: cagin.kandemir@deu.edu.tr

The integrity of this information is called the genome. DNA has double-stranded helical structure and consists of 4 different bases, {A, T, C, G}. DNA molecules can be billions of base pairs long. The human genome contains about 3 billion base pairs. But about 3% of the DNA in a chromosome represents the basic units of inheritance, which are called genes. Proteins are molecules responsible for the formation of an organism. Most of the functions in a cell are performed by proteins. The building blocks of proteins are amino acids. DNA sequences include triple bases called codons encoded by amino acids. Proteins are formed by converting codons into amino acid sequences.

Enzymes are molecules in the protein structure that contribute to metabolic activities by increasing the rate of chemical reactions. Phylogenetic Analysis is the removal of the evolutionary relationship between organisms. Molecular phylogenetic studies are based on determining the rate and nature of changes in DNA and proteins, thus exploring the evolutionary history of genes and organisms [2].

In the study, it was aimed to cluster lipase enzymes according to their thermal stability by a distance-based method without alignment. The data used were obtained in the FASTA format from the NCBI Genbank database [3].

1.1. Comparison of biological sequences

The distance between two sequences is calculated by comparing the sequences with several methods and algorithms to obtain a numerical output expressing dissimilarities (distance). Biological sequences belonging to organisms with common evolutionary roots have higher similarity than organisms without close common roots [4].

Sequence comparison reveals the evolutionary, structural, or functional relationships of biological sequences such as DNA and protein sequences. When a new biological sequence is discovered, functionality and function in biological databases can be compared to known sequences to obtain information about its functionality.

Conventional sequence comparison methods are performed using alignment algorithms based on matching similar regions of arrays using a space character. However, since alignment requires consideration of local mutations and appropriate parameter selection according to the data set used, studies on methods that do not include alignment have gained speed. Most non-alignment methods are based on the principle of common sub-array search and frequency calculation between arrays.

As a result of the comparison of the sequences, the distance matrix of the sequences in the data set is obtained. The phylogenetic tree is constructed by applying the clustering algorithms in the obtained distance matrix.

1.1.1. Biological sequence alignment

The basic approach to sequence alignment is to identify the most similar regions of different DNA or protein sequences. The similarity of regions means that there is a functional, structural or evolutionary relationship between the series [5]. The degree of similarity of amino acids in a given position in sequence alignment of proteins can be taken as a rough measure of how conserved a particular region or sequence motif is for these proteins. The absence of substitutions in a particular region of the directory is an indication that this region has a structural or functional prefix [6]. In phylogenetic analysis studies, the conserved regions obtained by the alignment step in protein, enzyme classification are frequently used [2,7].

Aligned DNA or protein sequences are typically represented as rows of a matrix. Spaces between letters representing nucleotides or amino acids are provided so that the same or similar letters in consecutive columns are aligned.

Unaligned states are interpreted as mutations that have occurred since the sequences have diverged from each other. The spaces in the alignment are interpreted as insertions or deletions. Alignment algorithms use reward scores (m) in matching, penalty scores non-matching (s) and gap (g) situations. According to selected parameters for $S = AGGCTAGTT$ and $Q = AGCGAAGTTT$ sequences possible alignment score can be computed as:

$$(\#\text{Matching}) \times \text{m} - (\#\text{Non-matching}) \times \text{s} - (\#\text{Gap}) \times \text{g}$$

As a result, the alignment with the highest scoring is selected. As you can see, the effect of the parameter selection on the alignment score is great. Choosing the appropriate alignment parameter is a necessary and difficult step in data sets where there are many differences in sequence lengths and diversity. Alignment Algorithms are divided into Local and Global Alignment Algorithms. While Global Alignment Algorithms are preferred in data sets consisting of arrays belonging to similar organisms, Local Alignment Algorithms are preferred in data sets consisting of arrays belonging to distant organisms [8].

Clustal Omega Multiple sequence alignment program can be used to find similarities between large proteins, DNA data sets in bioinformatics studies [9]. Clustal Omega [9] provides a comprehensive analysis of protein or DNA sequences compared to sequences in the sequence database.

2. Methods and data set

Algorithm-based methods have been successful in phylogenetic studies, but studies on methods that do not contain alignments have been accelerated because of the influence of different sequence lengths in the data set, alignment algorithm suitable for the data set's characteristic and parameter selection. In non-alignment methods, the basic approach is based on the calculation of key (invariant) region search and subsequence frequency, scatter statistics between the arrays. This approach makes the method an automatic step that is independent of the sequence lengths and does not require preprocessing. Genomic sequences are stored as linked lists containing categorical values {A, T, G, C}. Therefore, categorical statistical analytical-based scientific derivation is necessary to examine the similarity relations between the sequences [10]. For non-alignment methods, the goal is to get a vector that represents subsequence distributions in the array.

2.1. Clustering methods

Clustering algorithms receive a distance matrix entry specifying the distance between each biological sequence in the dataset. As a result of the method, a phylogenetic tree is formed which also defines the pattern of branches in the tree and in some cases.

Hierarchical clustering methods is divided into two parts as agglomerative and divisive. In Agglomerative method, each sequence is treated as a single cluster, then all sequences have been merged into a single cluster by taking into account the lowest distance between each sequence. In divisive method, all sequences that are initially treated as in the same cluster, are partitioned until each sequence exists in a single cluster [11].

2.2. K-mer natural vector method

The subsequences formed by n consecutive bases in genetic sequences are called k-mers [12-14].

If the length of each biological sequence in the data set is defined as *L*, the *k-mers* seen in a sequence can be obtained by shifting the sequence by *L-k* **+** 1*th* indices by increasing the index 1 at each time while maintaining the *k* length. There are 4^k different *k-mers* in a DNA sequence that can be seen for any *k* value [15]. The resulting *k-mers* can be represented by notation [1], [2], ..., [4^k].

The *k-mer* count vector indicated by $n^{(s,k)}$ notation is the vector that gives the frequency of the *k-mers* in the sequence. $n^{s[i]}$ is the number of occurrences *k-mer* [*i*] in the sequence,

$n^{(s,k)} = (n_{s[1]}, n_{s[2]}, \cdots, n_{s[4^k]})$

The k-mer mean distance vector represented by notation ($\mu_{[1], [2], \ldots, } \mu_{[4^k]}$) gives the average distance from the first base to each *k-mer* seen in the sequence. If *k-mer* [*i*] is not seen in the array, $\mu_{[i]}$ is defined as zero. As a result, the normalized central moment vector ($D_2^{[1]}$, $D_2^{[2]}$, ..., $D_2^{[4k]}$) is obtained as follows,

$$D_m^{[i]} = \sum_{j=1}^{n_{[i]}} \frac{\left(s_{[i][j]} - \mu_{[i]}\right)^m}{n_{[i]}^{m-1}(L-k+1)^{m-1}}, \quad m = 1,2,\ldots,n_{[i]} \tag{1}$$

where $n_{[i]}$ is the number of occurrences of *k-mer* [*i*] in the sequence, $s_{[i][j]}$ is the distance of j$^{\text{th}}$ *k-mer* [*i*] to the first base in the array. Applying the formula yields a normalized center-moment vector representing the distribution of *k-mers* in the array. Finally, the central moment vector obtained by using the natural parameters associated with the distribution of the *k-mers* in the biological sequence.

- When *k* = 1, the *k-mer* native vector is identical to the original native vector. Thus, the *k-mer* natural vector is the generalization of the original natural vector model. The natural parameters used in the method ensure that a vector representing the subsequence distributions of the biological sequences is well represented.
- It is very important to determine the value of *k* because the parameter *k* used has a great influence on the results of phylogenetic analysis and computational complexity. It is suggested that the *k* parameters in the method must be in the range [floor (*log*4 *min*(*L*)), ceil (*log*4 *max*(*L*))].

Cosine similarity is used to determine the distance between the vectors obtained because the similarity between the textual data is calculated via the cosine function in trigonometry [16].

$$d(s_1, s_2) = 1 - cos(v_1, v_2) = 1 - \frac{v_1 . v_2}{|v_1||v_2|} \tag{2}$$

Lipase enzymes with wide use in the industrial field are frequently used in biotechnological studies. In the study, it was aimed to cluster the thermophilic and mesophilic lipase enzymes according to their active temperature ranges. Six mesophilic, nine thermophilic lipase enzymes were used to generate the data set [17].

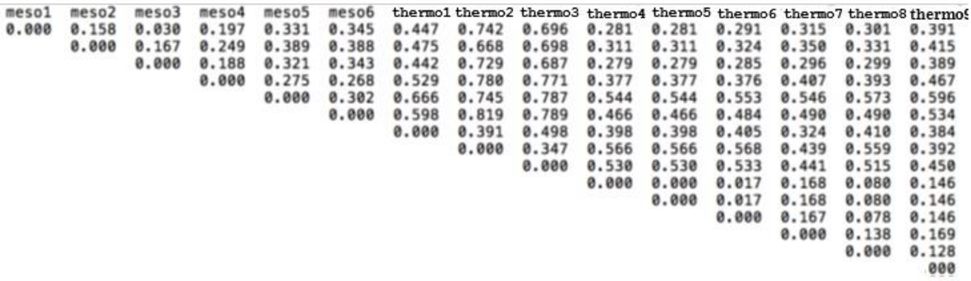
Thermophilic enzymes are enzymes capable of growing at 43°C-55°C temperature. Mesophilic enzymes grow at temperatures of 30°C-42°C [18]. Lipase enzymes were clustered using the k-mer natural vector method, a method that does not involve alignment and pre-processing.

The length of the shortest biological sequence in the dataset is 633, the length of the longest biological sequence is 1854, and the optimal k-length range {4,7} is obtained as a result of iterations.

3. Results

The clustering of lipase enzymes is of great importance in predicting the presence of catalytic domain or disulfide bonds, in determining the secretion mechanism and specific lipase-based folds, in describing relationships with other enzyme families [19].

Biological sequences to be used in the study were obtained from NCBI Genbank database and the methods were coded in Java programming language to obtain distance matrices and phylogenetic trees were constructed using R language. The distance values between the sequences are calculated with the *k-mer* algorithm as in Fig.1. As seen from Fig. 1, the distance values for thermo4 and thermo5 are equal. Therefore, they are in the same branch of phylogenetic trees.



meso1	meso2	meso3	meso4	meso5	meso6	thermo1	thermo2	thermo3	thermo4	thermo5	thermo6	thermo7	thermo8	thermo9
0.000	0.158	0.030	0.197	0.331	0.345	0.447	0.742	0.696	0.281	0.281	0.291	0.315	0.301	0.391
	0.000	0.167	0.249	0.389	0.388	0.475	0.668	0.698	0.311	0.311	0.324	0.350	0.331	0.415
		0.000	0.188	0.321	0.343	0.442	0.729	0.687	0.279	0.279	0.285	0.296	0.299	0.389
			0.000	0.275	0.268	0.529	0.780	0.771	0.377	0.377	0.376	0.407	0.393	0.467
				0.000	0.302	0.666	0.745	0.787	0.544	0.544	0.553	0.546	0.573	0.596
					0.000	0.598	0.819	0.789	0.466	0.466	0.484	0.490	0.490	0.534
						0.000	0.391	0.498	0.398	0.398	0.405	0.324	0.410	0.384
							0.000	0.347	0.566	0.566	0.568	0.439	0.559	0.392
								0.000	0.530	0.530	0.533	0.441	0.515	0.450
									0.000	0.000	0.017	0.168	0.080	0.146
										0.000	0.017	0.168	0.080	0.146
											0.000	0.167	0.078	0.146
												0.000	0.138	0.169
													0.000	0.128
														000

Fig. 1 Distance matrix calculated for Lipase enzymes with parameter k=4.

In Fig. 2, a phylogenetic tree was constructed using a *k-mer* distance-based neighboring algorithm based on the thermal stability of Lipase enzymes. The phylogenetic tree for Lipase enzymes was obtained using the Clustal Omega Alignment program as in Fig.3. In both clustering methods, thermophilic and mesophilic lipase enzymes are clustered in separate branches of phylogenetic trees.

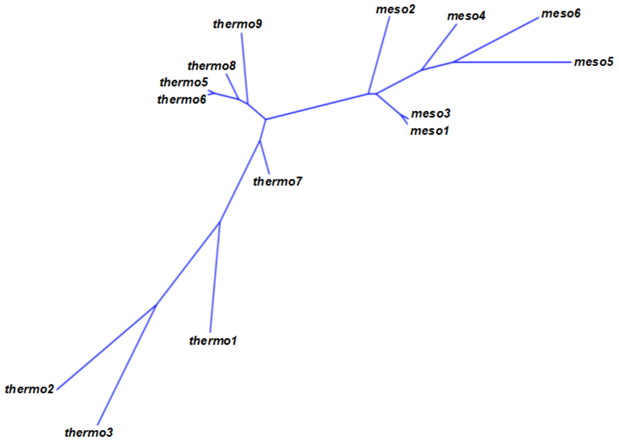


Fig. 2 Phylogenetic tree constructed using $k = 4$ parameters for Lipase enzymes using neighbor-joining algorithm.

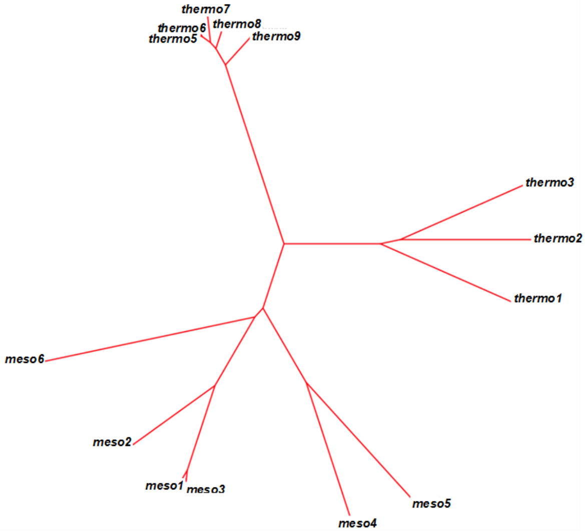


Fig. 3 Phylogenetic tree obtained by applying Clustal Omega Alignment Program for Lipase enzymes.

4. Conclusion

A successful clustering based on a functional property of the enzyme sequences, without any preprocessing on the enzyme sequences was achieved with the *k-mer* natural vector method that is a distance-based method. The study differs from other studies [2,20] in the literature because the natural parameters used together represent subsequence distributions (*k-mers*) of the biological sequences. We suggest that the *k-mer* method can be used as an alternative in phylogenetic studies.

References

1. Lesk AM. Introduction to bioinformatics. 2nd edition. New York: Oxford University Press; 2005.
2. Nasibov E and Kandemir-Cavas C. OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees. Expert Sys. Appl., 2011;38:12684-12690.
3. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. Nucleic Acids Research, 2016;4(44 Database issue):D7-D19.
4. Apostolico A, Guerra C, Landau GM, and Pizzi C. Sequence similarity measures based on bounded hamming distance. Theor. Comput. Sci., 2016;638:76-90.
5. Mount DM. Bioinformatics: Sequence and Genome Analysis. 2nd edition. New York: Cold Spring Harbor Laboratory Press; 2004.
6. Ng PC and Henikoff S. Predicting deleterious amino acid substitutions, Genome Res., 2001;11(5):863-74.
7. Albayrak A and Sezerman OU. Lempel-Ziv Complexity scores for clustering mesophilic and thermophilic Lipases. The International Enzyme Engineering Symposium 2008, 2008, Kuşadası, Aydın.
8. Krane DE and Raymer ML. Fundamental concepts of bioinformatics. San Francisco, USA: Pearson Education; 2008.
9. Clustal Omega [Document on the Internet]. 2017 [cited 2017 April]. Available from: http://www.ebi.ac.uk/Tools/msa/clustalo/

10. Amiri S and Dinov D. Comparison of genomic data via statistical distribution. J. Theor. Biol., 2016;407:318-327.
11. Mitra S and Acharya T. Data mining: multimedia, soft computing and bioinformatics. NJ: Wiley; 2003.
12. Melsted P and Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics, 2011;12:333.
13. Hashim EKM and Abdullah R. Rare k-mer DNA: Identification of sequence motifs and prediction of CpG island and promoter. J. Theor. Biol., 2015;387:88-100.
14. Fiannaca A, La Rosa M, Rizzo R, and Urso A. A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network. Artif. Intell. Med., 2015;64:173-184.
15. Wen J, Chan RH, Yau SC, He RL, and Yau SS. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Gene, 2014;546(1):25-34.
16. Al-Anzi FS and AbuZeina D. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University – Int. J. Comput. Inf. Sci., 2014;29:189-195.
17. Royter M, Schmidt M, Elend C, Höbenreich H, Schäfer T, Bornscheuer UT, and Antranikian G. Thermostable lipases from the extreme thermophilic anaerobic bacteria Thermoanaerobacter thermohydrosulfuricus SOL1 and Caldanaerobacter subterraneus subsp. Tengcongensis, Extremophiles, 2009;13:769–83.
18. Anaerobic Digestion-Mesophilic Vs. Thermophilic [Document on the Internet]. 2017 [Cited 2017 April]. Available from: https://www.theecoambassador.com/Anaerobic-Digestion-Temperature.html
19. Arpigny JL and Jaeger KE. Bacterial lipolytic enzymes: classification and properties, Biochem. J., 1999;343:177-183.
20. Weyenberg G and Yoshida R. Phylogenetic Tree Distances, Ref. Module Life Sci. Encyc. Evol. Biol., 2016;285-290.