Signal processing approaches as novel tools for the clustering of N-acetyl-β-D-glucosaminidases

Mojtaba Mamarabadi[1,2*], Behzad Tokhmechi[3]

[1]Department of Plant Protection, Faculty of Agriculture, Ferdowsi University of Mashhad, P.O. Box 1163, Mashhad, I.R. Iran [2]Faculty of Agriculture, Shahrood University of Technology, P.O. Box 316, Shahrood, I.R. Iran [3]School of Mining, Petroleum and Geophysics Engineering, Shahrood University of Technology, P.O. Box 316, Shahrood, I.R. Iran

Received: 30 May 2011
Accepted: 04 Feb 2012

Abstract
Nowadays, the clustering of proteins and enzymes in particular, are one of the most popular topics in bioinformatics. Increasing number of chitinase genes from different organisms and their sequences have been identified. So far, various mathematical algorithms for the clustering of chitinase genes have been used but most of them seem to be confusing and sometimes insufficient. In the present study, as a first step, different amino acids participating in panoply of chitinases, as a model protein, obtained from the NCBI GenBank, were digitized. Digitized data were normalized to the signal energy. Normalized data decomposed using mother wavelet bior 5.5 to approximation (a_1) and details (d_1), at the first level. Corresponded coefficients have been obtained and cross correlation between normalized, a_1 and d_1 coefficients of amino acid sequences were calculated. Maximum correlation was selected as similarity index and corresponded cladogram trees were made. The results of this study showed that more optimal and reliable cladogram tree can be produced and better discrimination observed from d_1 coefficients compared to normalized sequences and opposed to a_1 coefficients. Using suggested approach, the cladogram tree made from d_1 coefficients not only had more validity but also the drawback of the classic cladogram tree has been improved.

Keywords: Enzyme clustering; Cladogram tree; Mother wavelet; Signal processing; Correlation

*Correspondence to: **Mojtaba Mamarabadi, Ph.D.**
Tel: +98 9153086584; Fax: +98 5118788875
E-mail: mom@shahroodut.ac.ir

INTRODUCTION

Proteins are the most fundamental substance of life, as they are the key component of the protoplasm of all cells. Enzymes, hormones, transcription factors, pumps and antibodies are examples for the diverse functions fulfilled by proteins in a living organism. There are only 20 different types of amino acids, and they can be combined to generate an infinite number of sequences. In reality, only a small subset of all possible sequences appears in nature. The three important attributes of enzymes: are sequence, structure, and function. The sequence is essentially the string of amino acids which comprises the enzymes (Seckbach and Rubin, 2004).

Increasing number of enzyme encoding genes from different organisms and their sequences have been identified. The clustering of enzyme is one of the most popular topics in bioinformatics nowadays. The goal of clustering is to identify distinct groups in a data set and assign a group label to each observation. Observations are partitioned into subsets or clusters, such that observations in one subset are most similar to each other than to observations in different subsets (Bang *et al*., 2010).

There are a wide array of clustering approaches, each with its strengths and weaknesses. Nugent and

Meila (*Edited by* Bang *et al*., 2010) have been presented an overview of clustering method applied in molecular biology. Basically, there are two type of clustering: Attribute base (e.g., K-mean, K-medoids, Model-based, Nonparametric, Simple mean shift, Gaussian blurring mean shift, Dirichler mixture model and Biclustering) and similarity base (e.g., Hierarchical, Spectral and Affinity propagation).

Zainuddin and Pauline applied various clustering algorithms, namely, K-means (KM), Fuzzy C-means (FCM), symmetry-based K-means (SBKM), symmetry-based Fuzzy C-means (SBFCM) and modified point symmetry-based K-means (MPKM) clustering algorithms in choosing the translation parameter of a WNN (Zainuddin and Pauline, 2010).

Two clustering methods named NJ and UPGMA in ClustalW program can be performed. The default method in this internet base program is neighbor-joining (NJ). These methods used to construct the phylogenetic tree: NJ: Neighbor-joining (Saitou and Nei, 1987) and UPGMA: Un-weighted Pair Group Method with Arithmetic Mean (Sneath and Sokal, 1973). The UPGMA algorithm assumes equal rates of evolution, so that branch tips come out equal. The Neighbor-Joining algorithm allows for unequal rates of evolution, so that branch lengths are proportional to amount of change. UPGMA assumes a constant rate of evolution and is not a well-regarded method for inferring phylogenetic trees unless this assumption has been tested and justified for the data set being used. UPGMA was initially designed for use in protein electrophoresis studies, but is currently most often used to produce guide trees for more sophisticated phylogenetic reconstruction algorithms. NJ is a bottom-up clustering method used for the construction of phylogenetic trees. Usually used for trees based on DNA or protein sequence data, the algorithm requires knowledge of the distance between each pair of taxa (Xavier, 2010; Murtagh, 1984).

Chitin is the second most abundant, renewable polysaccharide in nature after cellulose (Li, 2006; Haki and Rakshit, 2003). Chemically, chitin is a β- (1-4)-linked homopolymer of N-acetyl-β-D-glucosamine. Chitin-degrading enzymes, chitinases, which hydrolyze chitin, occur in a wide range of organisms including viruses, bacteria, fungi, insects, higher plants, invertebrate and vertebrate animals like humans (Park *et al*., 1997). The roles of chitinases in these organisms are diverse.

Nomenclature of chitinolytic enzymes is confusing and does not include all known enzymes with chitinolytic activity. A sensible terminology which roughly covers any enzymes that catalyze the cleavage of chitin has been suggested by Lorito (1998). This terminology differentiates chitinolytic enzymes based on the reaction end-products. According to this nomenclature, chitinases can be divided into two major categories: endochitinases and exochitinases. Endochitinases (EC 3.2.1.14) randomly cleave chitin and chitooligomers at internal sites and release a mixture of soluble low molecular mass end-products of different size which are multimers of GlcNAc. Exochitinases can be divided into two subcategories: chitin 1, 4-β-chitobiosidases and β-N-acetylhexosaminidases.

The classification of chitinolytic enzymes based on the similarities of their amino acid has been proposed (Henrissat, 1999). In this classification, the structural features of enzymes have been combined with their three-dimensional structures (Davies and Henrissat, 1995) and chitinolytic enzymes were grouped into families 18, 19 and 20 of glycosyl hydrolases (Henrissat and Bairoch, 1993). Family 18 chitinases are found in bacteria, fungi, yeast, viruses, plant and animals, and hence the family is varied in evolutionary terms. Family 19 members are almost entirely presented in plants. Family 20 consists of the β-N-acetylhexosaminidases or β-N-acetylglucosaminidases from bacteria, fungi and humans (Li, 2006). Therefore, various mathematical algorithms for the clustering of chitinase gene have been proposed but most of them seem to be confusing and sometimes insufficient. They basically make a simple cross correlation among analogous amino acid sequences for the gene from different organisms which sometimes, the interpretation of constructed cladogram is confusing and difficult.

Signal processing approach is a novel tool for constructing the cladogram tree with more discrimination power and higher accuracy. Huge amount of information and sequencing data on chitinolytic enzyme-encoding gene are available but in the present study we only used amino acid sequences for the N-acetyl-β-D-glucosaminidases gene from different organisms which have been obtained from the NCBI gene bank.

MATERIALS AND METHODS

Thirty amino acid sequences for the N-acetyl-β-D-glucosaminidases gene from different fungi, plants and animals which was already used in our previous publication (Mamarabadi *et al*., 2009), retrieved from the

NCBI databases and used as a model protein in the present investigation. The genes and other related information are provided in Table 1. In the present study three techniques for the signal processing of amino acid sequence data were used which have been explained briefly.

Wavelet decomposition: Since there are 20 types of amino acids in protein sequences the number one to 20 have been allocated to digitize each particular amino acid, respectively. For example the number one for *A* (single letter code for Alanine), number 2 for *C* (single letter code for Cysteine) and etc. Each digitized amino acid sequence (DAS) forms a one dimensional data (1D). Therefore a 1D wavelet approach has been used. Eq. (1) defines a discrete wavelet transformer (*DWT*) of a signal x(z) (Jin *et al*., 2008; Mallat, 1989; Daubechies, 1988):

$$DWT_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(z)\psi\left(\frac{z-\tau}{s}\right)dz \qquad (1)$$

This function transforms signal $x(z)$ using mother wavelet $\Psi(z)$ from DAS domain (z) to translation (τ) and scale (s) domains. In Eq. (1), z-τ is the DAS translation. The term $(\sqrt{|s|})^{-1}$ is a normalization factor to remove the scale effect from wavelets with different scales.

Figure 1 shows the procedure of wavelet decomposition. As this Figure shows, in the first step, wavelet decomposes DAS into low and high frequency bands, which are called approximation (a_1) and

Table 1. N-acetyl-β-D-glucosaminidases gene from different organisms.

Sequence Number	Organism	Gene	Accession Number (retrieved from NCBI)
1	*Trichoderma atroviride*	*Ta. nag2*	*AAT70229.1*
2	*Trichoderma virens*	*Tv. nag2*	*AAL84701.1*
3	*Paracoccidioides brasiliensis*	*Pb. pb- nag1*	*AAL14649.1*
4	*Penicillium chrysogenum*	*Pc. nagA*	*AAF00010.1*
5	*Trichoderma virens*	*Tv.nag1*	*AAL84700.1*
6	*Aspergillus oryzae*	*Ao. nagA*	*BAC41255.1*
7	*Aspergillus nidulans*	*An. nagA*	*BAB13330.1*
8	*Trichoderma atroviride*	*Ta. nag1*	*CAC85401.1*
9	*Trichoderma asperellum*	*Ta. exc1y*	*CAC85402.1*
10	*Candida albicans*	*Ca. hex1*	*AAA34346.2*
11	*Clonostachys rosea*	*Cr. cr-nag1*	*ABC73393.1*
12	*Trichoderma asperellum*	*Ta. exc2y*	*ABC95196.1*
13	*Paracoccidioides brasiliensis*	*Pb. pb- nag1*	*AAK94334.1*
14	*Coccidioides posadasii*	*Cp. chit1*	*ABB18373.1*
15	*Coccidioides posadasii*	*Cp. chit2*	*ABU87865.1*
16	*Sclerotinia sclerotiurum*	*Ss. predicted protein*	*XP_001592574.1*
17	*Cryptococcus neoforman*	*Cn. hex*	*XP_571630.1*
18	*Magnaporthe grisea*	*Mg. hp*	*XP_363950.1*
19	*Gibberella zeae*	*Gz. hp*	*XP_382346.1*
20	*Gibberella zeae*	*Gz. hp*	*XP_382115.1*
21	*Gibberella zeae*	*Gz. hp*	*XP_381459.1*
22	*Metarhizium flavoviride*	*Mf. chit*	*CAB44709.1*
23	*Metarhizium anisopliae*	*Ma. chit*	*AAC33265.1*
24	*Ustilago maydis*	*Um. chit*	*AAG35111.1*
25	*Botryotinia fuckeliana*	*Bf. hp*	*XP_001554078.1*
26	*Bacillus subtilis*	*Bs. nag*	*BAA08089.1*
27	*Rhizobium leguminosarum*	*Rl. hex*	*CAK07535.1*
28	*Homo sapience*	*Hs. hex*	*P07686.3*
29	*Caenorhabditis elegans*	*Ce. hex*	*CAO72175.2*
30	*Arabidopsis taliana*	*At. hex*	*NP_176737.2*

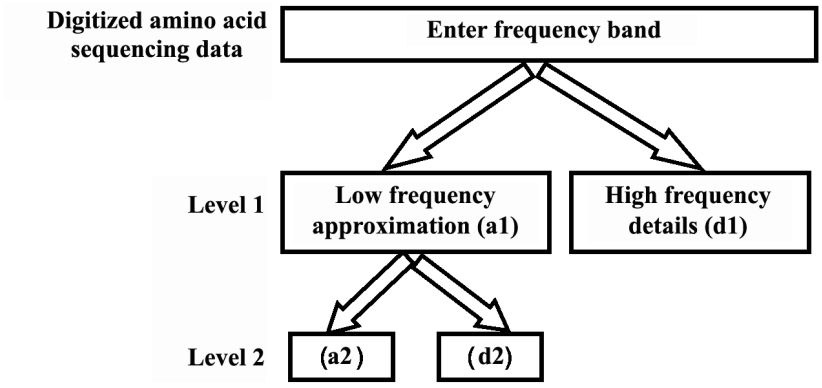


Figure 1. Schematic DAS wavelet decomposition stages.

details (d_1), respectively. In the second step, it decomposes a_1 to low (a_2) and high (d_2) frequency bands. This procedure can be continued for decomposing low frequency bands to higher levels.

Normalization procedure: Normalization of DAS data per DAS energy is a simple procedure as described below:
Calculate DAS energy (E_{DAS}) using Eq. (2):

$$E_{DAS} = \sum_{i=1}^{n} DAS_i^2 \qquad (2)$$

Where n is the number of amino acids for each sequence.
Calculate normalized DAS (N_{DAS}) using Eq. (3):

$$N_{DAS_i} = \frac{DAS_i}{\sqrt{E_{DAS}}} \qquad (3)$$

Where is normalized DAS_i for amino acid i.

Auto and cross correlation: Cross correlation indicates how much two amino acid sequences are similar to each other statistically. In order to measure and

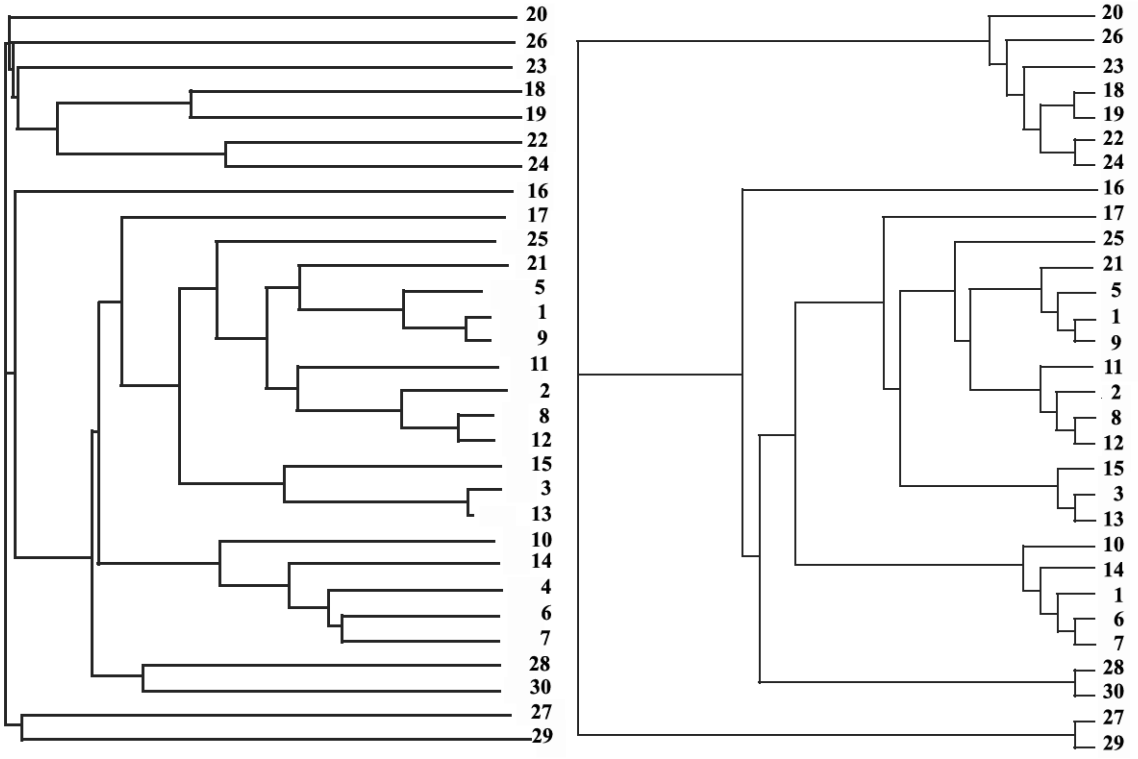
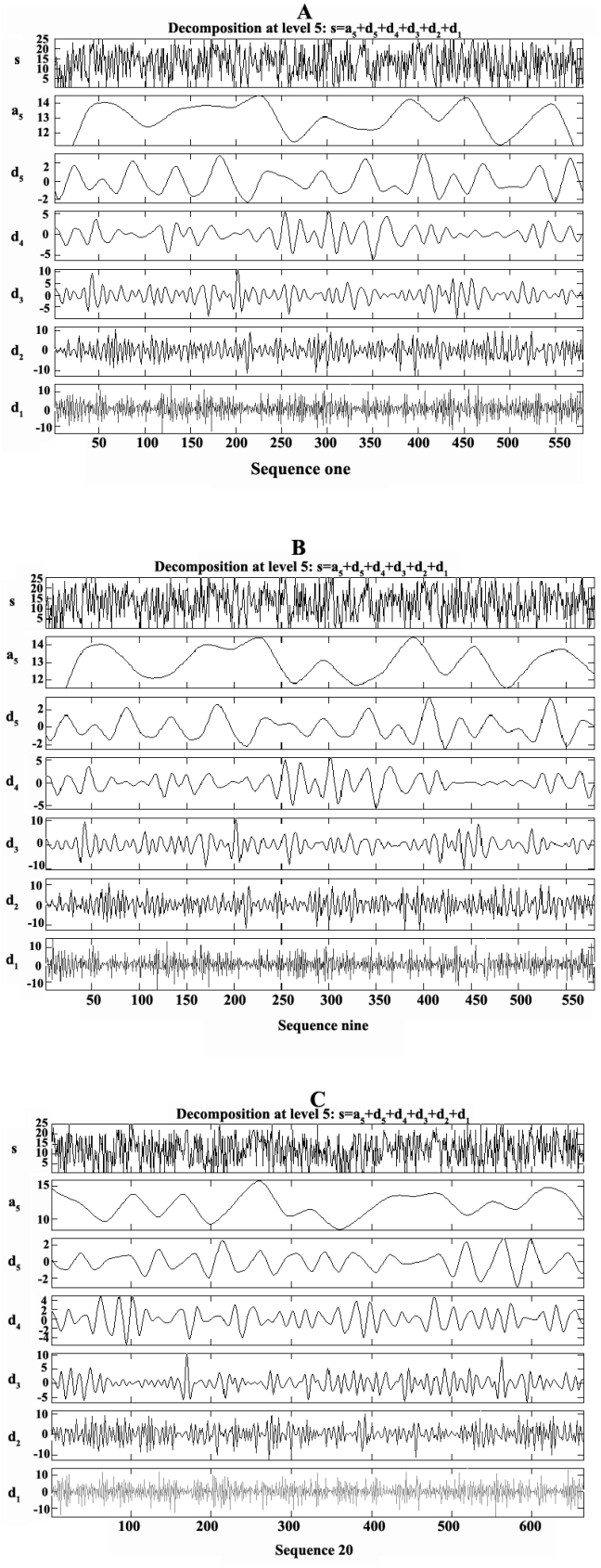


Figure 2. Conventional cladogram trees made for N-acetyl-β-D-glucosaminidases genes from different organisms.



igure 3. Wavelet decomposition to five levels for sequence A: 1, B: 9 and C: 20.

quantify this, four following steps have to be down (Yilmaz, 2001):

Reverse moving of first sequence

Multiply in the vertical direction

Add the product and write as output

Shift the second sequence to the right and repeat steps *b* and *c*.

Cross correlation of an amino acid sequence with itself is known as auto correlation. Cross and auto correlation were performed on normalized DAS. The maximum correlation can be calculated as Eq. (4):

$$\max_{auto} = \sum_{i=1}^{n} \left(N_{DAS_i}\right)^2 = \frac{\sum_{i=1}^{n} DAS_i^{\,2}}{E_{DAS}} = 1 \qquad (4)$$

Therefore maximum cross correlation amount is a similarity index between two amino acid sequences. So that, when that amount is getting close to one, this demonstrates high similarity between them.

RESULTS

In this section conventional cladogram trees and their disadvantages and the results from signal processing approaches for N-acetyl-β-D-glucosaminidases amino acid sequence have been presented, respectively.

Drawback of conventional trees: Conventional cladogram trees drawn for 30 different N-acetyl-β-D-glucosaminidases amino acid sequences from different organisms has been shown in Figure 2. The presented cladogram was made using ClastalW program which has some drawbacks. That means the interpretation of these cladograms is sometimes difficult to explain. For example in the following cases these problems could be observed:

The case number 22 has to be located beside number 23, but in these cladograms it was located beside number 24. Number 22 and 23 are from the same genus (*Metarhizium*) from Ascomycetous fungi, while number 24 (*Ustilago*) is from Basidomycota.

The case number 28, which belongs to human genome, was located beside number 30, which is from a plant named *Arabidopsis thaliana*. These organisms are systemically unrelated. Sequence 28 is also unrelated to the other amino acid sequences.

Amino acid sequence decomposition: All amino acid

sequences were decomposed using *bior5.5* mother wavelet to various levels. For example, two different amino acid sequences from the mycoparasite fungus *Trichoderma* spp., namely sequence number one and sequence number nine which are similar to each other and sequence number 20 from the fungus *Gibberella zeae* which is far from two aforementioned sequences were selected to show the decomposition.

All three amino acid sequences were decomposed to the five levels. The diagrams for approximation five (a5) and details one to five (d_1 to d_5) have been shown in Figure 3. As it can be seen, a5 for sequences one and nine are completely identical, while a5 for sequence 20 is completely different. The same similarity can be seen among d3, d4 and d5 for the sequences one and nine, whereas in sequence 20 the dissimilarity can be observed. As it visually seems in Fig. 3 the similarity of amino acid sequences in different frequency bands are different. For instance, the similarity of a5 between sequences 1 and 9 is more than the similarity of d_1 between two aforementioned amino acid sequences. Therefore, the basic idea is that the similarity of sequences in frequency bands probably is a good index for clustering and it might be used instead of their own amino acid sequence similarity.

Auto and cross correlation: Raw amino acid sequence similarity for the sequences number one, nine and 20 were shown in Figure 4 (section A, B and C). Sequence approximation (section D, E and F) and sequence details (section G, H and I) were also presented. The maximum amount of auto correlation as was expected was equal to 1 and can be observed at the section A, D and G.

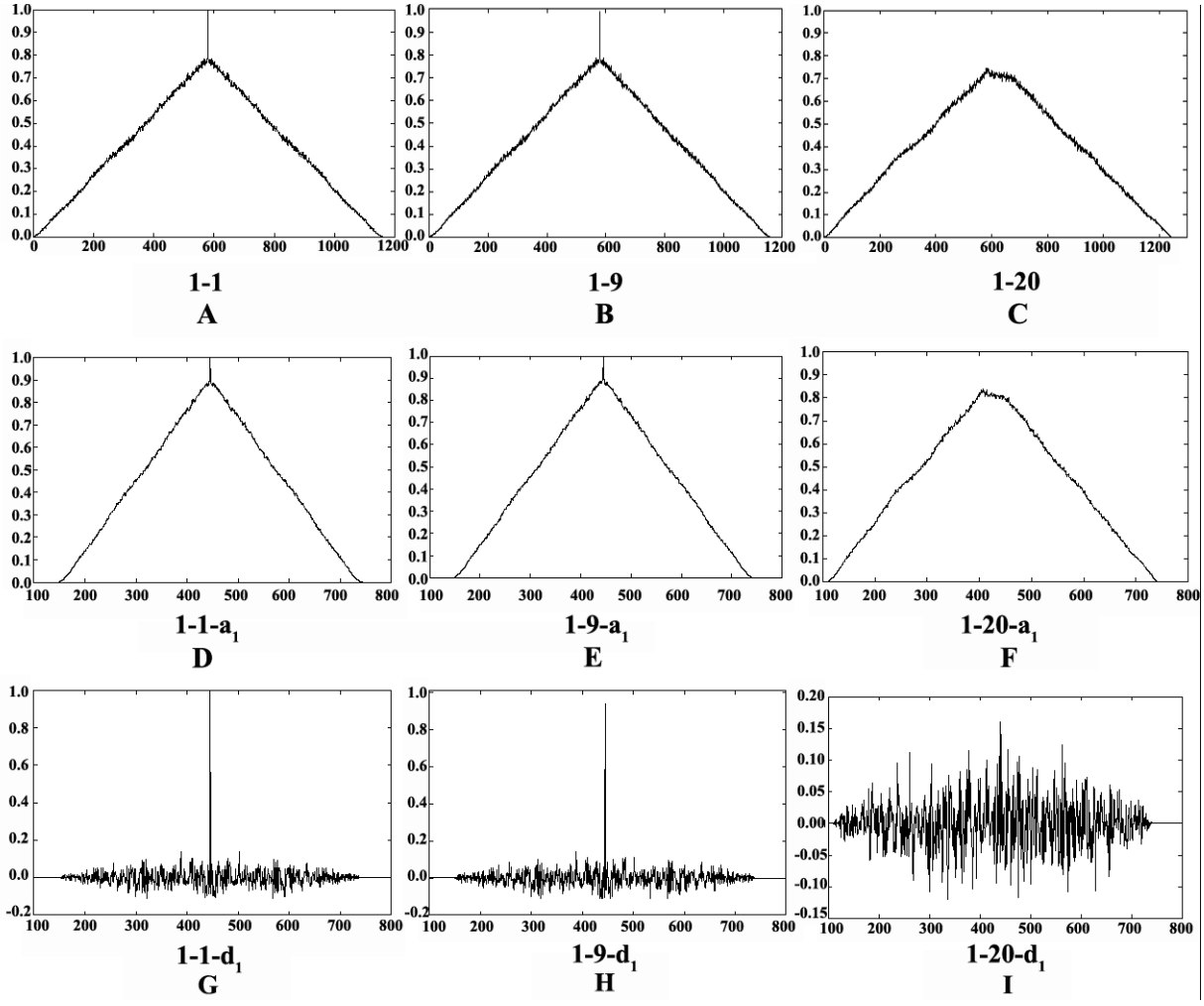


Figure 4. Auto (A, D and G) and cross correlation between raw amino acid data (A, B and C), a_1 (D, E and F) and d_1 (G, H and I) for sequence one, 22 and 28.

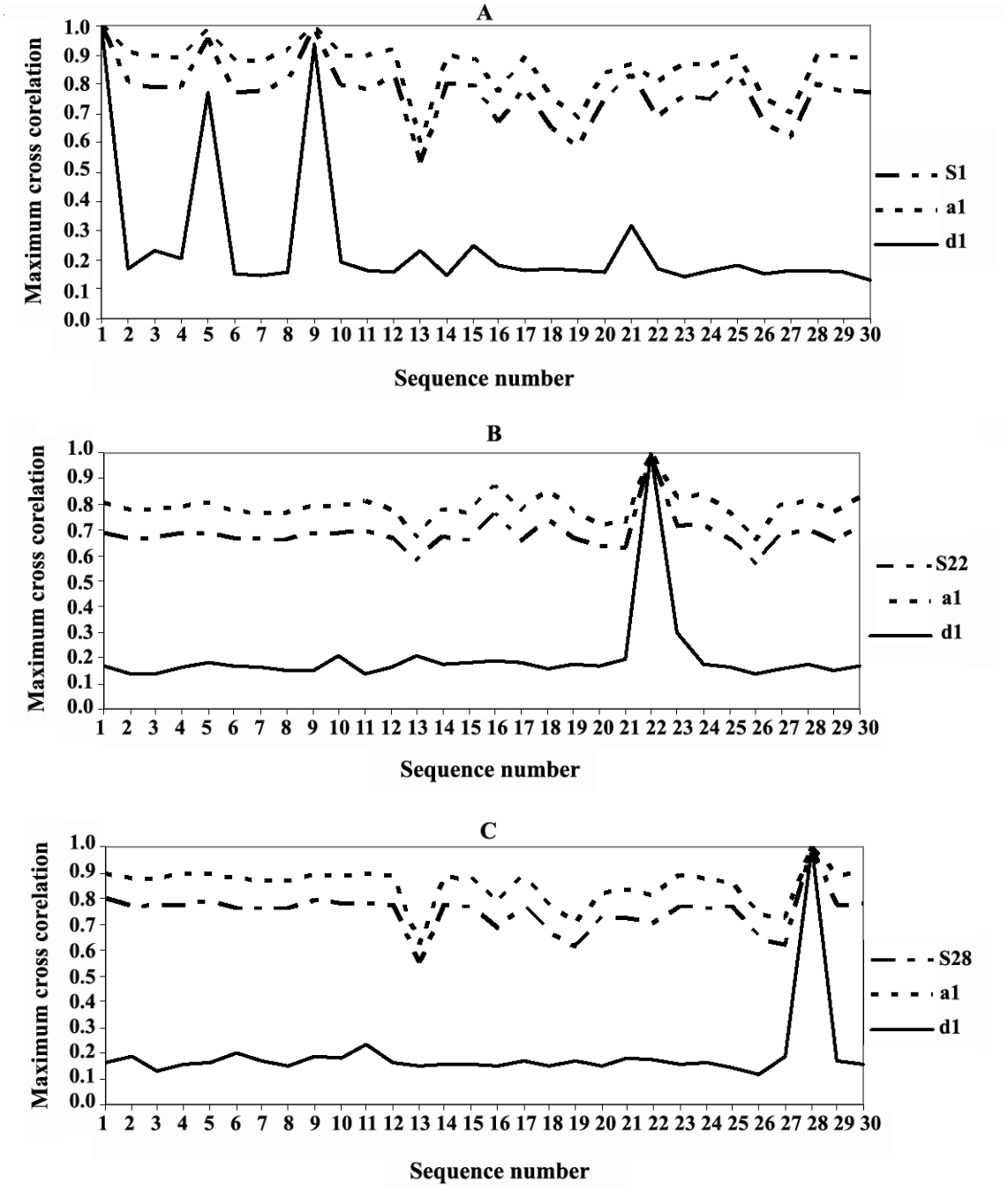


Figure 5. Maximum auto and cross correlation between raw amino acid data, a_1 and d_1 for sequence A: 1, B: 22 and C: 28 compared to the other sequences.

In the section B maximum cross correlation between sequence one and nine was close to one, which means they are similar to each other. On the other hand in section C, maximum cross correlation between sequence one and 20 is about 0.8, which is close to one, and it can not show the dissimilarity between two sequences perfectly. In the section E and F, which were used from the a_1 of three sequences the maximum amount of cross correlation between sequence one and nine and sequence one and 20 are about one and 0.9, respectively. This shows low frequency bands (a_1) of amino acid sequences have been presented a weak discrimination between three sequences.

In contrast, the result of cross correlation performed on d_1 was shown considerable discrimination.

Hence, the maximum cross correlation of d_1 for the sequence one and nine is approaching to one, while this parameter for the sequence one and 20 is less than 0.2. In the section I the amount of dissimilarity between sequence one and 20 can clearly be seen.

In order to find the optimal way for clustering theamino acid sequences the maximum amount of cross correlation between each amino acid sequences and other 29 amino acid sequences and their related a_1, d_1 were drawn. For example in Figure 5A the maximum cross correlation for the sequence one with the other 29 sequences has been presented. As it is clear in the Figure 5, the maximum auto correlation for the sequence one is always equal to one in all used data. Moreover, the maximum cross correlation between sequence one and nine and sequence one and five is close to one. This means all the three sequences are similar to each other and they can be located in the same cluster. On the other hand the maximum cross correlation between sequence one with the other 27 sequences are about 0.8 for the raw amino acid data, 0.9 for a_1 and 0.2 for d_1. Therefore, it is obviously clear that d_1 makes much better discrimination among similar and dissimilar sequences. In fact, sequence decomposition to a_1 and d_1 makes two bunches of data which one of them (a_1) decrease and the other one (d_1) increase the ability of sequence discrimination. In the other word effective indexes for the clustering of amino acid sequences have been concentrated in d_1.

In the Figure 5B, the maximum cross correlation for the sequence 22 with the other 29 sequences has been presented. When we consider the maximum cross correlation from the raw amino acid data or a_1 this sequence shows the most amount of similarity with sequence 24, which are not related to each other. But when the maximum cross correlation for d_1 is considered, the sequence 22 will be located beside sequence 23, which are from the same genus (*Metarhizium*).

Figure 5C shows that sequence 28 from the human genome are close to sequence 30 from *Arabidopsis thaliana,* if we notice the maximum cross correlation from the raw amino acid data or a_1. In fact this sequence has not significant similarity with other 29 sequences, and should be presented as an independent cluster. Once again d_1 makes a good discrimination among different amino acid sequences.

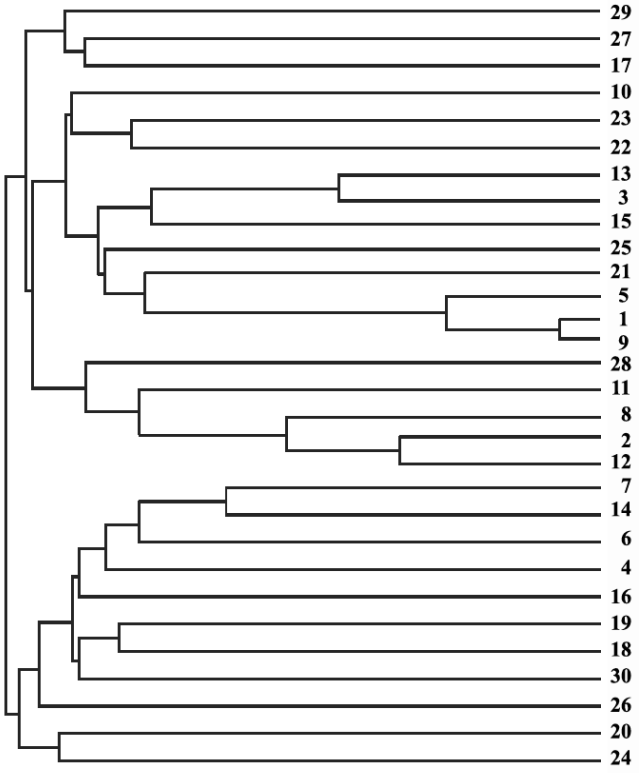


Figure 6. Improved wavelet based cladogram tree for 30 different amino acid sequences from different organisms.

DISCUSSION

An improved wavelet based cladogram tree for the 30 different amino acid sequences from the N-acetyl-β-D-glucosaminidases gene was drawn and presented in Figure 6. This cladogram has all the advantages of conventional cladograms (Fig. 2). For example both cladograms made similar clustering for the sequence one, five and nine. The same similar cluster can be seen for the sequence 2, 8 and 12 in both cladogram trees. In addition using wavelet transform, the drawback of the conventional trees has been improved. For example in conventional cladogarms (Fig. 2) the case number 22 which is from an Ascomycetous fungus named Metarhizium was located beside number 24 which comes from a Basidiomycetes fungus named Ustilago. These two organisms are systematically unrelated and in fact this type of clustering is not true. In the improved wavelet based cladogram tree (Fig. 6) the case number 22 and 23 which are from the same genus (Metarhizium) are located beside each other. From the taxonomical point of view this kind of clustering is much closer to the fact. Another drawback in the conventional tree the case number 28, which belongs to human genome, was located beside number 30, which is from a plant named Arabidopsis thaliana. Again, these organisms are systemically unrelated. As

a matter of fact the case number 28 is unrelated to all other cases and no significant similarity can be seen between this case and other 29 cases, and it should be presented as an independent cluster. In the wavelet based cladogram this drawback has also been improved and this case is presented individually (Fig. 6).

CONCLUSIONS

Increasing number of exo-chitinases gene from the different organisms is becoming identified and different mathematical algorithms have been used for clustering far. But they sometimes seem insufficient and have their own drawback. In the present study, thirty different amino acid sequences for N-acetyl-β-D-glucosaminidases obtained from the NCBI gene bank were digitized and normalized to the signal energy. Normalized data were decomposited using mother wavelet bior 5.5 to approximation (a_1) and details (d_1) afterwards. Cross correlation between normalized, a_1 and d_1 maximum coefficients were calculated and finally the maximum correlation was selected as similarity index and corresponded cladogram trees were made. The results of this study showed that more optimal and reliable cladogram tree can be produced using wavelet based clustering method. Furthermore, better discrimination who observed in the cladogram tree obtained from d_1 coefficients than normalized amino acid sequences and also a_1 coefficients. Interestingly, in the cladogram tree made from d_1 coefficients using this approaches either the drawback of classic cladogram tree has been improved or they have much more validity. The optimization of mother wavelet and also the clustering method in order to make a better discrimination should be further studied.

Acknowledgments

This work was financially support by Shahrood University of Technology which is gratefully acknowledged. We thank Dr. Naser Farrokhi for editing the manuscript.

References

Daubechies I (1988). Orthogonal bases of compactly supported wavelets. *Commun Pure Appl Math*. XLI: 909-96.

Davies G, Henrissat B (1995). Structures and mechanisms of glycosyl hydrolases. *Structure* 3: 853-859.

Jin L, Sen MK, Stoffa PL, Seif RK (2008). Time-lapse seismic attribute analysis for a water-flooded reservoir. *J Geophys Eng*. 5: 210-220.

Haki GD, Rakshit SK (2003). Developments in industrially important thermostable enzymes: a review. *Bio Tech*. 89: 17-34.

Henrissat B (1999). Classification of chitinases modules. EXS. 87: 137-156.

Henrissat B, Bairoch A (1993). New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 293: 781-788.

Li DC (2006). Review of fungal chitinases. *Mycopathologia* 161: 345-360.

Lorito M (1998). In Harman GE, Kubicek CP. Chitinolytic enzymes and their genes, In: *Trichoderma* and *Gliocladium*, Vol. 2. Enzymes, biological control and commercial applications, Taylor and Francis Ltd. London, UK. PP. 73-99.

Mallat S (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intel*. 11: 674-93.

Mamarabadi M, Jensen D F, Lübeck M (2009). An N-acetyl-beta-D-glucosaminidase gene, cr-nag1, from the biocontrol agent *Clonostachys rosea* is up-regulated in antagonistic interactions with *Fusarium culmorum*. *Mycol Res*. 113: 33-43.

Murtagh F (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. Computational Statistics Quarterly. 1: 101-103.

Nugent R, Meila M (2010). In Bang H, Zhou XK, Van Epps HL, Mazumdar M (Ed.) Statistical Methods in Molecular Biology. Humana Press. PP. 349-460.

Park JK, Morita K, Fukumoto I, Yamasaki Y, Nakagawa T, Kawamukai M, Matsuda H (1997). Purification and characterization of the Chitinase (ChiA) from *Enterobacter* sp. G-1. *Biosc Biotech Biochem*. 61: 684-689.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4: 406-425.

Seckbach J, Rubin E (2004). The new avenue in bioinformatics. Kluwer Academic Publisher. Netherland. 281 Pages.

Sneath PHA, Sokal RR (1973). Numerical taxonomy. WH Freeman and Company. San Francisco. 513 pages.

Xavier D (2010). Sequence-based analysis of bacterial population structures. In robinson A, Falush D, Feil EJ. *bacterial population genetics in infectious disease*. John Wiley and Sons. PP. 46-47.

Yilmaz O (2001). Seismic Data Analysis, Processing, Inversion, and Interpretation of Seismic Data, Vol 1. Society of Exploration Geophysics. USA. PP. 39-40.

Zainuddin Z, Pauline O (2010). Improved wavelet neural networks for early cancer diagnosis using clustering algorithms. *Int J Inf Mat Sci*. 6: 31-36.