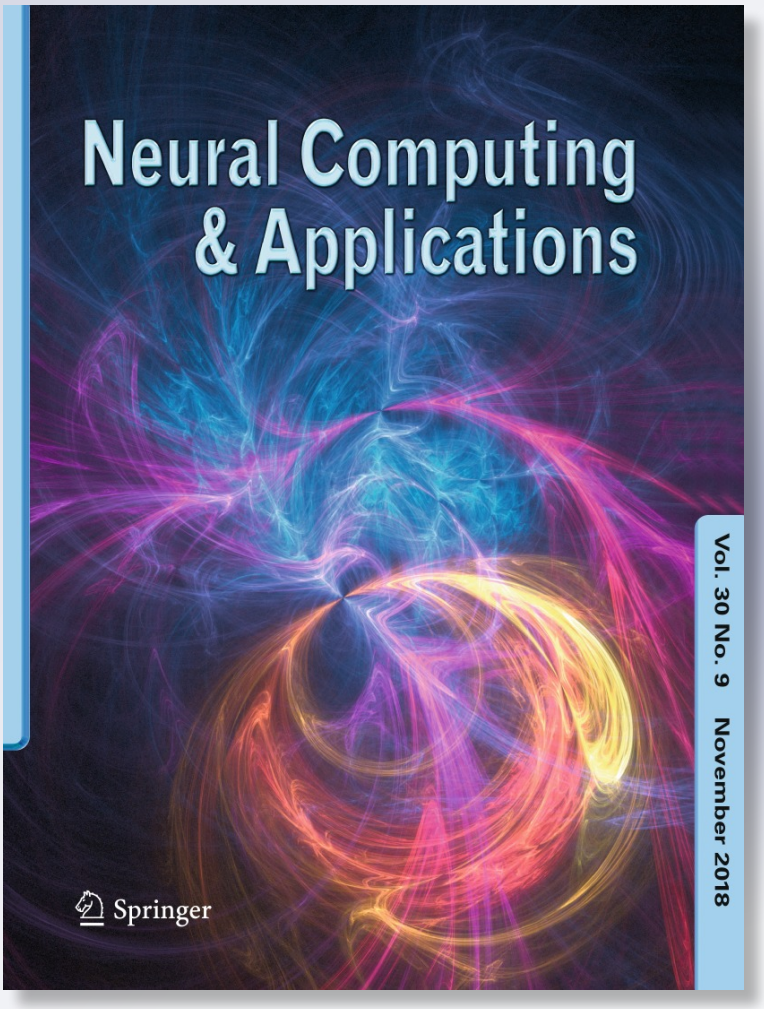
Clustering of fungal hexosaminidase enzymes based on free alignment method using MLP neural network

Mojtaba Mamarabadi & Abbas Rohani

Neural Computing and Applications

ISSN 0941-0643
Volume 30
Number 9

Neural Comput & Applic (2018)
30:2819-2829
DOI 10.1007/s00521-017-2876-0



Neural Comput & Applic (2018) 30:2819–2829
https://doi.org/10.1007/s00521-017-2876-0

ORIGINAL ARTICLE

Clustering of fungal hexosaminidase enzymes based on free alignment method using MLP neural network

Mojtaba Mamarabadi[1] · Abbas Rohani[2]

Received: 23 November 2015 / Accepted: 10 February 2017 / Published online: 17 February 2017

Abstract Studies of biological evolution have generally focused on nucleotide or amino acid sequences of certain genes related to specific enzymes. Most phylogenetic tree constructions have been carried out using amino acid sequences and are used as a predictor to show evolutionary relationships. Phylogenetic analysis is usually performed based on multiple sequence alignment of a gene from different organisms including fungi. A number of programs have been introduced for gene clustering and phylogenetic analysis. For example, the most popular web-based program is Clustal Omega which is commonly used by biologists. When the number of uploaded sequences increases, this program not only works slowly but also the final constructed cladogram is confusing and incorrect from evolutionary point of view. In the present study, we used fungal hexosaminidases which are extracellular enzymes with a lot of applications in biotechnology but extremely varied and confusing in evolutionary terms. A standard taxonomy-based phylogenetic tree was constructed for 835 FH amino acid sequences retrieved from National Center for Biotechnology Information (NCBI) on March 16, 2015. Then a supervised multilayer perceptron (MLP) neural network was used to discriminate FH sequences. Based on relative frequency of amino acid in FH sequences, 41 neural networks were designed for seven levels from the phylum to family. Minimum accuracy of the neural network was equal to 99% at all seven discrimination levels. As a final step, an additional evaluation was performed on the designed model with 143 new released FH sequences extracted on July 1, 2015. The clustering results have shown a proper match with fungal taxonomy to show evolutionary relationships.

Keywords Fungal hexosaminidases · Clustering · MLP neural network

Availability and Implementation: Our suggested software and other related information are freely available on the web at the following link: https://www.dropbox.com/s/q2irc46g0wsj43k/soft%20ware.zip?dl=0

✉ Mojtaba Mamarabadi
mamarabadi@um.ac.ir

[1] Department of Plant Protection, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Department of Biosystem Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

1 Introduction

The enzymes *β*-*N*-acetyl-D-hexosaminidases (EC 3.2.1.52, Hex) belong to the glycoside hydrolase family 20 and catalyze the removal of *N*-acetyl-D-glucosamine (GlcNAc) or *N*-acetyl-D-galactosamine (GalNAc) from the non-reducing ends of a sort of physiological substrates, like oligosaccharides, glycoproteins and glycolipids. These enzymes are present in numerous species of various organisms such as bacteria, fungi, yeasts, plants, actinomycetes, arthropods and humans in which they play different physiological roles [5, 7, 14]. Among the hexosaminidase family, fungal hexosaminidases are extracellular enzymes which like other chitinases have lots of applications in biotechnology. This includes bio-conversion of chitin to useful products such as fertilizer, the production of non-allergenic, non-toxic, biocompatible and biodegradable materials and development of insecticides and fungicides. Possible future applications of hexosaminidase are as food additives to increase shelf life, therapeutic agent for osteoarthritis, asthma and chronic

rhinosinusitis, as an antifungal pesticide, an anti-tumor drug and as a general ingredient to be used in protein engineering [5]. Unfortunately, these enzymes are extremely varied and confusing in evolutionary terms.

Several mathematical algorithms have been proposed for the clustering of these enzymes, but most of them seem to be confusing and sometimes create incorrect results. They basically make a simple cross-correlation among analogous amino acid sequences based on their multiple sequence alignment, and finally, the interpretation of constructed cladogram is confusing from the evolutionary point of view [10].

Studies of biological evolution have generally performed by construction of phylogenetic tree using nucleotide or amino acid sequences of certain genes [15]. Phylogenetic trees are mathematically constructed based on comparative similarity among different samples using several multiple sequence alignment software such as ClustalX, ClustalW and Clustal Omega [8, 13]. Mathematical calculations sometimes lead to coincidental similarity while there are no genetic relationships among independent or even related samples. Sometimes, related genes from related organisms were unexpectedly placed in the wrong or unrelated cluster. This is a consequence of the used algorithms and not a logical error. This problem could also be emerged when alignment base phylogenetic trees are constructed for one specific enzyme from different organisms [10].

Many efforts have been concentrated on the potential of neural networks for enzyme and protein clustering using their amino acid sequences. For example, a neural network was trained to identify the catalytic residues found in enzymes, based on an analysis of the structure and sequence [4]. Their neural network output was then used to predict the location of the active site in enzymes.

In some studies, the potential of support vector machines (SVM) has been suggested for the enzyme family classification and for facilitating protein function prediction. The classification accuracy for enzymes families was in the range of 50.0–95.7% [2]. Moreover, the performance of SVM classifiers and neural network was accurately predicted lipid-binding proteins irrespective of sequence homology [1]. Their combination was successfully used for classification of different LBPs classes (about 92% in average).

CLUSS was another tool which has been developed for clustering of protein sequences to meet the needs of biologists in terms of phylogenetic analysis and also prediction of biological functions [6]. They have claimed their method accurately highlighted the functional characteristics of the clustered families compared to the other clustering methods in that time.

Recent publications were more focused on using biological information and alignment-free methods for enzyme clustering. For example, an unsupervised gene clustering algorithm has been proposed based on the integration of external biological knowledge, such as gene ontology annotations, into expression data. Therefore, two genes are considered close if they have both similar expression and similar functional profiles at the same time [17].

Classification of proteins (CLAP) was another alignment-free software for automatic classification of protein sequences. It is utilized a pattern-matching algorithm that assigned local matching scores (LMS) to residues which were a part of the matched patterns between two sequences being compared [3].

Recently, a two-dimensional graphical representation of protein sequences has been introduced based on the two physicochemical indexes (hydrophobicity and hydrophilicity of amino acids); meanwhile, a numerical characteristic has been proposed to compute the distance of different sequences for analysis of sequence similarity/dissimilarity on the basis of this graphical representation [18].

Finally, pattern recognition strategy was used to unravel the evolution of *Nanog*, which is a key transcription factor involved in self-renewal of undifferentiated embryonic stem cells [11]. They have extracted 47 *Nanog* genes sequences from various species, and two datasets of features were computationally extracted from these sequences. They used various data mining algorithms such as decision tree models which were applied on these datasets to find the evolutionary pathways of *Nanog* diversion. The outcomes of their study unraveled the importance of particular genomic features in *Nanog* gene evolution [11].

In this paper, we suggest a free alignment clustering method in order to construct a robust and uncomplicated cladogram for all the 835 known fungal hexosaminidases which have been submitted in the National Center for Biotechnology Information (NCBI) gene bank so far. Our suggested clustering approach is able to show a good evolutionary relationship among fungal hexosaminidase genes with minimum error even at the family level. We also believe this method could be used for the other fungal hexosaminidases which will be released in the future.

2 Materials and methods

2.1 Amino acid sequences

All known fungal hexosaminidases (FH) which were about 835 amino acid sequences from the diverse fungi retrieved

from NCBI (http://www.ncbi.nlm.nih.gov/) databases on March 16, 2015, are used in the present study. Some of these sequences were already used in our previous publication [10]. The FASTA format of these amino acid sequences and their accession numbers are presented in complimentary material where could be down loaded at: https://www.dropbox.com/s/q2irc46g0wsj43k/soft%20ware.zip?dl=0.

2.2 Clustal analysis

As a first step in this study, all 835 FH amino acid sequences were submitted in Clustal Omega [13] which is a new multiple sequence alignment program for proteins that uses seeded guide trees and HMM profile–profile techniques to generate alignments between three or more sequences. As it has been mentioned in the program literature, this program is able to produce the biologically meaningful multiple sequence alignments of divergent sequences and evolutionary relationships can be seen via viewing cladograms or phylograms [9]. The constructed tree made by this software has been presented in complementary material (abovementioned link). This tree was compared to our constructed tree in the result section.

2.3 Taxonomy base clustering

We have suggested a clustering method based on recent fungal taxonomy which has been used in NCBI and other databases for fungal classification. In the other words, clustering of fungal hexosaminidase enzymes was performed based on fungal taxonomy. We put 835 amino acid sequences belonging to fungal hexosaminidase in their fungal taxonomical group using a classifier neural network (Fig. 1).

2.4 Classifier neural network

The structure of neural network with two hidden layers used in this study is presented in Fig. 2. The duty of each neuron (node) is the calculation of its inputs total weight and passing them from a soft nonlinear function. Log activation sigmoid function and hyperbolic tangent sigmoid were chosen for the first and second hidden layers, respectively, and a linear function was chosen for the output layer. Back propagation with Declining Learning Rate Factor (BDLRF) algorithm based on total sum square error (TSSE) was used for the network learning and for the finding of proper weights between layers [12]. Multilayer perceptron (MLP) neural network works in the forward and backward phases. The features X insert to the network in the forward phase. The output (y) and the net output for each layer are calculated based on Eqs. 1 and 2:

$$\text{net} = \sum XW \tag{1}$$

$$y = \frac{1}{1 + e^{-\text{net}}} \tag{2}$$

The weights W_1, W_2 and W_3 were updated in the backward phase. The weights in each layer were similarly updated using Eq. 3:

$$W(n+1) = W(n) - \eta \frac{\partial E}{\partial W} + \alpha(W(n) - W(n-1)) \tag{3}$$

where η and α are learning rate and momentum factor, respectively. Their amount will be different from 0 to 1. W is the weight matrix between the neurons.

2.5 Features extraction

Fungal hexosaminidases like all other enzymes are made of amino acids where each amino acid is bonded to the next by chemical bonds. The vast majority of enzymes including

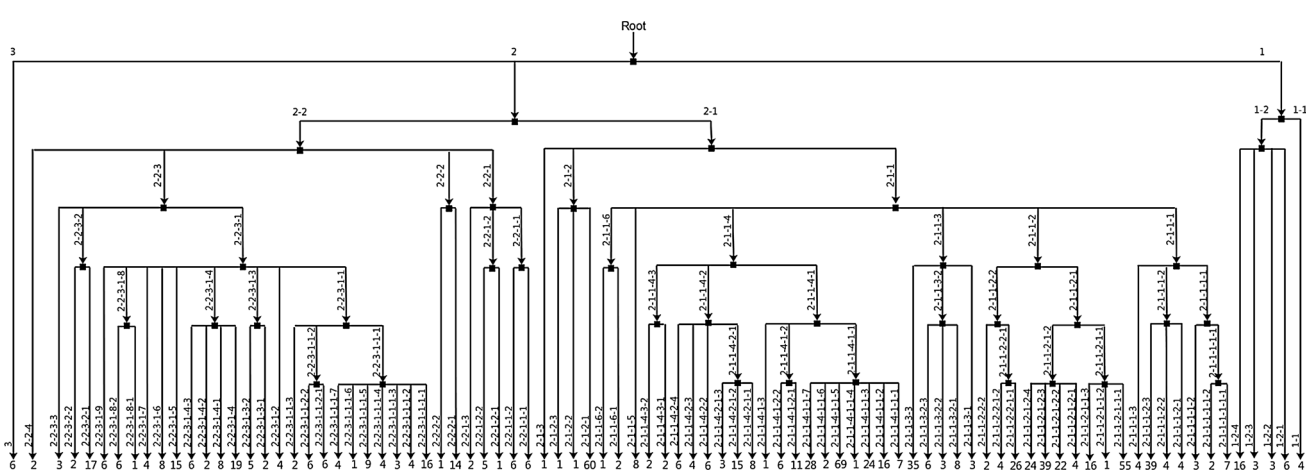


Fig. 1 Fungal hexosaminidase enzymes clustering based on fungal taxonomy

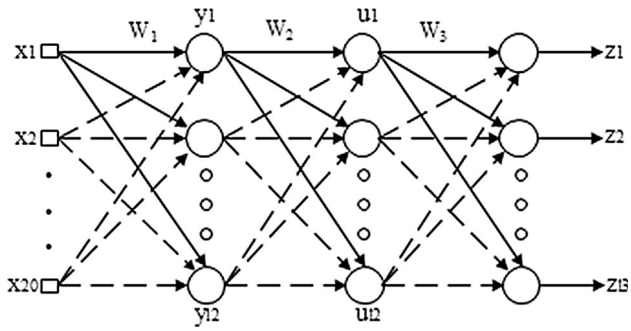


Fig. 2 Configuration of the MLP with two hidden layers

FH are made of only 20 different kinds of amino acid (A R N D C Q E G H I L K M F P S T W Y V) where the structure and function of the enzyme are determined by the order of the amino acids. As 835 FH sequences retrieved from NCBI had different length, the number of amino acid was also different in each sequence. The following equation was used for calculation of sequence features based on the number of amino acid in each sequence:

$$x = 100 \times \left[\frac{L_A}{L_{\text{seq}}} \frac{L_R}{L_{\text{seq}}} \cdots \frac{L_V}{L_{\text{seq}}}\right]^{\text{T}}_{1\times 20} \tag{4}$$

where L_A is the number of amino acids A and L_{seq} is the total number of amino acids in each sequence. As there were 20 extracted features in each sequence, therefore the input number of neural network was also equal to 20. To optimize neural network behavior during the learning process, the extracted features were normalized with the following equation:

$$x_n = \frac{2(x - x_{\min})}{x_{\max} - x_{\min}} + 1 \tag{5}$$

Finally, 41 MLP neural networks were designed which had 1, 2, 3, 5, 9, 12 and 9 neural networks from the phylum to family level, respectively (Fig. 1). The responsibility of each neural network was discrimination and recognition of FH sequences at the same levels from the phylum to family. For training of each network at every level, their related patterns were separated from the total patterns. The used sequences in each network were independently selected among 835 FH amino acid sequences. 80% of the total patterns were used for network training, and the rest of them were used for the network validation and test. The steps of this study are illustrated in Fig. 3 to make a better perception for readers.

3 Results

We assumed that taxonomy-based clustering could be considered as a standard method for all known fungal hexosaminidases which have been identified so far. We also believe that this method could be used for the other fungal hexosaminidases which will be released in future. Our designed software and its guide are provided in supplementary information to give users an opportunity to try. The link for software is also freely available on the web at: https://www.dropbox.com/s/q2irc46g0wsj43k/soft%20ware.zip?dl=0.

As it has been presented in Fig. 1, the constructed tree has three main branches (number 1, 2 and 3) at the beginning level. As branches go forward, a numerical code was dedicated to each sub-branch and these codes get longer and longer and finally all fungal hexosaminidases were distributed in 86 clusters at the last level of clustering. Each cluster might be attributed to one or more fungal hexosaminidase sequences from different fungal families. The number at the end of branches shows the number of sequences which are located in each cluster. Black square nodes indicate the positions on the constructed tree where clustering has to be performed from these places. In fact, each position indicates a classifier neural network model.

3.1 Clustal Omega versus neural network clustering

First of all, if numerous sequences (like 978 FH) being uploaded in Clustal Omega, the processing time for the clustering and phylogenic analysis will be too long, while the responding time for our suggested software is less than 10 s for processing of 978 FH sequences. Furthermore, Clustal Omega works online while our software package is able to work offline. Discrimination of FH at the phylum level was conducted based on the percent of amino acid content in each sequence using MLP neural network with three outputs. In constructed tree, each output shows the main branches number 1, 2 and 3 which were Fungi incertae sedis, Dikarya and Glomeromycota, respectively. Randomized selection of sequences was carried out in the two working steps of neural network. The result is presented in Table 1 and confirms that the designed neural network has properly been able to separate the FH sequence in three phyla. As Glomeromycota has only one subdivision, this phylum was completely identified in this step. Since the weight of each of neural network layers and also the used sequences in each working step of neural network were randomly chosen, the result of neural network will be different. The standard deviation for correct recognition percent at the phylum level was equal to 2% for 20 different runs. Therefore, at the phylum clustering step to the tolerance of neural network versus mutability could be trusted. The neural network convergence diagram for the phylum level recognition has been presented in Fig. 4. Neural network training was stopped based on Sum Squared Error (SSE) scale in 12th epoch.

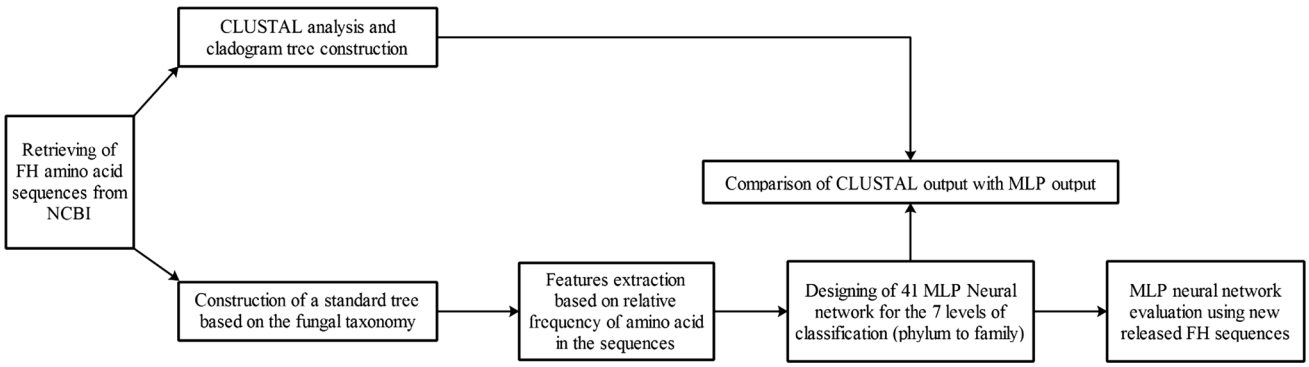


Fig. 3 Steps of the present study

Table 1 Number of FH sequences at the phylum level with the identified number in train and test phase including their recognition percent

Kingdom	Phylum	Number of sequences	*n* (% Recognition)[a]		Clustal Omega
			Train phase	Test phase	
Fungi	Fungi incertae sedis	32	29 (100)	3 (100)	0 (0)
	Ascomycota and Basidiomycota (Subkingdom of Dikarya)	797	637 (100)	160 (100)	728 (91.34)
	Glomeromycota	6	4 (100)	2 (100)	0 (0)

[a] *n* the number of patterns, *% Recognition* the percent of correct recognition

Different clustering result was obtained using Clustal Omega software (presented in complementary information at abovementioned link). Constructed tree for 835 FH made by Clustal Omega showed that although FH sequences were clustered in three main phyla, the result was entirely incorrect. In Clustal Omega analysis, FH belonging to Fungi incertae sedis and Glomeromycota were incorporated to Dikarya where as some of FH belongs to Dikarya were situated in the position of Fungi incertae sedis and Glomeromycota so that Dikarya (Ascomycota and Basidiomycota) was properly clustered by Clustal Omega with 91.34% accuracy at the subkingdom level. Furthermore, by going forward to the lower level, the number of branches was incredibly increased and this caused much more error. Therefore, the cladogram made by Clustal Omega had a little similarity with Fig. 1.

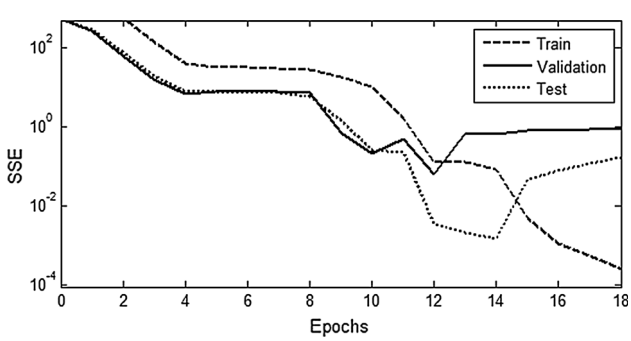


Fig. 4 Neural network convergence diagram for the phylum level identification

At the phylum level, subkingdom of Dikarya was divided into Ascomycota and Basidiomycota but 5.97% of 201 FH sequences belonging to Basidiomycota were correctly clustered in their right phylum. The noticeable point is that 2 FH sequences belonging to Basidiomycota were placed as an individual phylum and 30 FH sequences again belonging to Basidiomycota were located as another individual phylum. In the real life, the phylum of Basidiomycota has four subphyla while only two subphyla have been predicted for that by Clustal Omega. Moreover, in the fungal taxonomy Ascomycota has three subphyla but only two subphyla have been considered for Ascomycota by the Clustal analysis of FH enzyme. In the other words, the subphylum Taphrinomycotina was merged in Saccharomycotina and Pezizomycotina by Clustal analysis.

In general, when we move forward from the subphylum and class to the order and family level, the number of families becomes more and more and far from the reality, so that FH enzymes were located at 225 families by Clustal analysis while they have to be placed in 87 fungal families based on taxonomy. The result of correct recognition percent with the number of patterns at each of seven levels is presented in Table 2. All 41 neural networks have independently been trained and tested. As it could be observed, the accuracy of correct recognition is about 100% at all cases (Table 2). Because of the lacking sufficient pattern for network training, the recognition percent became less than 100% at the lower level (Family).

Table 2 Number of FH sequences from the phylum to family level with the identified total number and their recognition percent

Phylum	Subphylum	Class	Subclass	Order	Family
Fungi incertae sedis 32 (100)	Mortierellomycotina 4 (100)	–	–	Mortierellales	Mortierellaceae
	Mucoromycotina 28 (100)	–	–	Mucorales	Cunninghamellaceae 6 (100)
					Lichtheimiaceae 3 (66.7)
					Mucoraceae 3 (100)
					Rhizopodaceae 16 (100)
Glomeromycota 6 (100)	–	Glomeromycetes	–	Glomerales	Glomeraceae
Dikarya-Ascomycota 596 (100)	Pezizomycotina 533 (100)	Dothideomycetes 64 (100)	Dothideomycetidae 13 (100)	Capnodiales 10 (100)	Mycosphaerellaceae 7 (100)
					Teratosphaeriaceae 3 (100)
				Dothideales3 (100)	Aureobasidiaceae
			Pleosporomycetidae 47 (100)	Pleosporales	Leptosphaeriaceae 4 (100)
					Phaeosphaeriaceae 4 (100)
					Pleosporaceae 39 (100)
			Dothideomycetes incertae sedis 4 (10)	Botryosphaeriales	Botryosphaeriaceae
		Eurotiomycetes 192 (100)	Eurotiomycetidae 160 (100)	Eurotiales 71 (100)	Aspergillaceae 55 (100)
					Thermoascaceae 1 (100)
					Trichocomaceae 15 (100)
				Onygenales 89 (100)	Onygenaceae 4 (100)
					Ajellomycetaceae 22 (100)
					Arthrodermataceae 39 (100)
					Mitosporic onygenales 24 (100)
			Chaetothyriomycetidae 32 (100)	Chaetothyriales 30 (100)	Herpotrichiellaceae 30 (100)
					Cyphellophoraceae 2 (100)
				Verrucariales 2 (100)	Verrucariaceae
		Leotiomycetes 55 (98.18)	–	Erysiphales 3 (100)	Erysiphaceae
				Helotiales 17 (100)	Sclerotiniaceae 8 (100)
					Helotiaceae 3 (100)
					Dermateaceae 6 (100)
				Leotiomycetes incertae sedis 35 (100)	Pseudeurotiaceae
		Sordariomycetes 211 (100)	Hypocreomycetidae 165 (100)	Hypocreales 147 (100)	Mitosporic hypocreales 7 (85.71)
					Cordycipitaceae 16 (100)

Table 2 continued

Phylum	Subphylum	Class	Subclass	Order	Family
					Clavicipitaceae 24 (100)
					Bionectriaceae 1 (100)
					Nectriaceae 69 (100)
					Ophiocordycipitaceae 2 (100)
					Hypocreaceae 28 (100)
				Glomerellales 17 (100)	Glomerellaceae 11 (100)
					Plectosphaerellaceae 6 (100)
				Microascales 1 (100)	Microascaceae
			Sordariomycetidae 39 (97.44)	Sordariales 23 (100)	Chaetomiaceae 8 (100)
					Sordariaceae 12 (100)
					Lasiosphaeriaceae 3 (100)
				Magnaporthales 6 (100)	Magnaporthaceae
				Ophiostomatales 4 (100)	Ophiostomataceae
				Calosphaeriales 6 (100)	Calosphaeriaceae
			Xylariomycetidae 4 (100)	Xylariales	Diatrypaceae 2 (100)
					Amphisphaeriaceae 2 (100)
		Orbiliomycetes 8 (100)	–	Orbiliales	Orbiliaceae
		Pezizomycetes 3 (100)	–	Pezizales	Tuberaceae 2 (100)
					Pyronemataceae 1 (100)
	Saccharomycotina 62 (100)	Saccharomycetes	–	Saccharomycetales	Debaryomycetaceae 60 (100)
					Pichiaceae 1 (100)
					Saccharomycetales incertae sedis 1 (100)
	Taphrinomycotina 1 (100)	Taphrinomycetes	–	Taphrinales	Taphrinaceae
Dikarya-Basidiomycota 201 (99)	Pucciniomycotina 20 (100)	Pucciniomycetes 12 (100)	–	Pucciniales	Pucciniaceae 6 (100)
					Melampsoraceae 6 (100)
		Microbotryomycetes 6 (100)	–	Microbotryales	Microbotryaceae
		Mixiomycetes 2 (100)	–	Mixiales	Mixiaceae
	Ustilaginomycotina 15 (100)	Ustilaginomycetes 14 (100)	–	Ustilaginales	Ustilaginaceae
		Exobasidiomycetes 1 (100)	–	Georgefischeriales	Tilletiariaceae

Table 2 continued

Phylum	Subphylum	Class	Subclass	Order	Family
	Agaricomycotina 162 (100)	Agaricomycetes 141 (100)	Agaricomycetidae 55 (100)	Agaricales 41 (100)	Agaricaceae 16 (100)
					Psathyrellaceae 4 (100)
					Strophariaceae 3 (100)
					Tricholomataceae 4 (100)
					Marasmiaceae 9 (100)
					Pleurotaceae 1 (100)
					Schizophyllaceae 4 (100)
				Boletales 12 (100)	Coniophoraceae 6 (100)
					Serpulaceae 6 (100)
				Jaapiales 2 (100)	Jaapiaceae
			–	Auriculariales 4 (100)	Auriculariaceae
				Cantharellales 7 (100)	Botryobasidiaceae 2 (100)
					Ceratobasidiaceae 5 (100)
				Polyporales 35 (100)	Polyporaceae incertae sedis 19 (100)
					Polyporaceae 8 (100)
					Meruliaceae 2 (100)
					Phanerochaetaceae 6 (100)
				Hymenochaetales 15 (100)	Hymenochaetaceae
				Gloeophyllales 8 (100)	Gloeophyllaceae
				Russulales 4 (100)	Stereaceae
				Sebacinales 7 (85.71)	Sebacinales group B 1 (100)
					Bondarzewiaceae 6 (100)
				Corticiales 6 (100)	Punctulariaceae
		Tremellomycetes 19 (100)	–	Tremellales	Tremellaceae 17 (100)
					Mitosporic tremellales 2 (100)
		Dacrymycetes 2 (100)	–	Dacrymycetales	Dacrymycetaceae
	Basidiomycota incertae sedis 4 (100)	Wallemiomycetes	–	Wallemiales	Wallemiales incertae sedis

Therefore, according to the obtained results, a neural network was created which had 41 MLP neural models (Fig. 1). This network could be used for clustering of FH sequences based on fungal taxonomy from the phylum to family level.

3.2 New released FH clustering

In the last stage, each neural network was independently trained and tested for the recognition of 835 FH from the phylum to family level. Beside the last sequences, 143 new

FHs were released by NCBI from the March 16, 2015, to July 1, 2015 (presented in complementary information at abovementioned link). The new released sequences were used to the final test of neural network model. Among the new FHs, 85 sequences which had some background in the last sequences were clustered compatible with Fig. 1 but 58 sequences had no background and were completely new from the subclass to the family level. The percent of correct recognition in each clustering phase has been presented in Table 3. Obviously, the percent of correct recognition has been decreased from the phylum to family level.

4 Discussion

Clustering of enzymes into phylogenetically correct groups is a difficult dilemma, especially for those whose alignment is not biologically validated and not definitively performed [6]. Our results and experiences have shown that the amino acid frequency could be an appropriated feature for FH sequences clustering at the all seven levels from the phylum to family. Furthermore, the neural network recognition accuracy was highly dependent on the number of patterns which have been used for the network training in each level. Verbanck et al. proposed an unsupervised gene clustering algorithm based on the integration of external biological knowledge. They have claimed the simulation of dataset was varied according to the number of samples [17]. As the number of FH sequences is getting fewer from the higher (phylum) to lower (family) levels, the neural network extendibility will also be decreased at the test phase. Clustering by supervised method is based on real taxonomy, and the prediction will just be according to what the network has been trained. Therefore, unlike to Clustal Omega the number of branches in the cluster would not illusively be increased. Anyway, for the some levels such as class or order, correct predictions would be occurred and wrong prediction might happen afterward.

Although Clustal Omega software could discriminate fungal hexosaminidase in three main groups, this discrimination was not shown as a proper match with real fungal taxonomy, whereas our approach makes a good discrimination result which have had a good match with fungal taxonomy for showing evolutionary relationships. Along with our study, a comparative study was performed on 874 protein attributes of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms to create a link between protein characteristics and the organism [16]. They applied various weighting and modeling algorithms to determine how structural protein features change between organisms. Their result showed that within different tested models, the C5.0 model was the most efficient and precise model for discrimination of organism type, based on ammonium transporter sequence, with the precision of 94.85%. They have claimed that dissecting a large number of structural protein characteristics through data mining algorithms provides a novel functional strategy for studying evolution and phylogeny [16].

The potential of SVM for enzyme family classification and lipid-binding proteins prediction was reported in different studies [1, 2]. The classification accuracy for enzymes families and lipid-binding proteins was in the range of 50.0–95.7% and around 92%, respectively, which are comparable to our 100% recognition accuracy.

As some of FH sequences have been submitted as partial coding sequences in NCBI, their recognition was more difficult by neural network and lower percent of recognition was observed for them. Moreover, the cladogram made by Clustal Omega had numerous branches and each branch included a few FH sequences but in our suggested cladogram (Fig. 1) there were sometimes about 70 FH sequences which have correctly been located in the correct branch from the phylum to family level. Concerning 143 new released FH clustering, the percent of correct recognition has been decreased from the phylum to family level (Table 3). As the higher levels of neural networks (like phylum level) gained more training experience from the numerous patterns, they show more resistance against value range of feature extraction obtained from the new sequences. Therefore, they have shown much more accuracy than lower level (like subclass to family level). Furthermore, error propagation from the higher to the lower level could increase the error in the lower level. However,

Table 3 Network clustering results for the new released FH sequences

	Number of new released FH sequences	Correct recognition %					
		Phylum	Subphylum	Class	Subclass	Order	Family
Total	143	98.62	90.28	88.19	67.36	50.69	45.14
With background	85	100	98.82	95.29	77.64	69.41	65.88
Without background	58	96.61	77.96	77.96	52.54	23.72	0
Retrain ANN	143	100	100	100	*	*	*

reduction of correct recognition percent at the phylum level was arisen from lack of correct recognition for the two completely new sequences. This fact became clearer when 85 sequences were clustered. As these sequences had some background in the last sequences, the recognition percent for them was equal to 100% at the phylum level (Table 3). There were 58 completely new released FHs among 143 sequences. Fortunately, an acceptable percent of correct recognition has been calculated for them which was about 78% at the phylum to class level. It means the result of clustering could well be trusted. Because of lack of family definition for 58 FH sequences in the last step, the percent of correct recognition for them was zero. Therefore, Fig. 1 has to be improved and new information needs to be inserted on it. As it has been known, artificial neural networks were inspired by the human brain. Human needs knowledge and experiences to make a decision. Neural network should also be obtained some experiences to confront with new conditions. Therefore, the last networks were again trained by 978 FH sequences for the three levels (phylum, subphylum and class). The trained and new experienced neural networks were able to fully recognize all 143 new FH sequences at three levels (100% recognition). Therefore, neural networks could develop their knowledge by getting experience from the new sequences and they might be used in the future prediction with much more reliance. We believe that our suggested method can become an effective tool for clustering other enzymes to meet the needs of biologists in terms of phylogenetic analysis and evolutionary relationships. Finally, the tool can also be easily adapted to cluster other types of enzymes.

5 Conclusion

Neural networks have this potential to be used as powerful and rapid tools for clustering of fungal hexosaminidases and most likely other type of enzymes. Our software is presented in complementary information and could also be easily adapted and used to cluster new released FH enzymes. The clustering results and constructed tree have demonstrated a suitable match with fungal taxonomy to show evolutionary relationships. Although the emphasis of this paper was on supervised neural network for FH clustering, two other options could also be suggested for this purpose. The first option is to use supervised methods based on free alignment techniques such as neural networks, support vector machines and the second one will be based on the combination of supervised and unsupervised alignment-based techniques for enzyme clustering. For example, in the case of FH, supervised methods could be used till class level and after that unsupervised method will be used for the lower level which has fewer patterns.

Acknowledgements Financial support by the vice president for research and technology, Ferdowsi University of Mashhad, is gratefully acknowledged. We thank Mr. Meisam Nazari for editing the manuscript.

References

1. Bakhtiarizadeh MR, Moradi-Shahrbabak M, Ebrahimi M, Ebrahimie E (2014) Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. J Theor Biol 356:213–222
2. Cai CZ, Han LY, Ji ZL, Chen YZ (2004) Enzyme family classification by support vector machines. Proteins 55:66–76
3. Gnanavel M, Mehrotra P, Rakshambikai R, Martin J, Srinivasan N, Bhaskara RM (2014) CLAP: a web-server for automatic classification of proteins with special reference to multi-domain proteins. BMC Bioinform 15:343
4. Gutteridge A, Thornton JM, Bartlett G (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. Biochemistry 37:11940–11948
5. Hamid R, Khan MA, Ahmad M, Ahmad MM, Abdin MZ, Musarrat J, Javed S (2013) Chitinases: an update. J Pharm BioAllied Sci 5:21–29
6. Kelil A, Wang S, Brzezinski R, Fleury A (2007) CLUSS: clustering of protein sequences based on a new similarity measusre. BMC Bioinform 8:286–305
7. Kulik N, Slámová K, Ettrich R, Křen V (2015) Computational study of *β-N*-acetylhexosaminidase from *Talaromyces flavus*, a glycosidase with high substrate flexibility. BMC Bioinform 16:28
8. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
9. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucl Acids Res 43:W1
10. Mamarabadi M, Tokhmechi B (2012) Signal processing approaches as novel tools for the clustering of *N*-acetyl-*β*-D-glucosaminidases. Iran J Biotechnol 10(3):175–183
11. Pashaiasl M, Khodadadi K, Kayvanjoo AH, Pashaeiasl R, Ebrahimie E, Ebrahimi M (2016) Unravelling evolution of Nanog, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics. Gene 578:194–204
12. Rohani A, Abbaspour Fard MH, Abdolahpour S (2011) Prediction of tractor repair and maintenance costs using artificial neural network. Expert Syst Appl 38(7):8999–9007
13. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539
14. Slámová K, Bojarová P, Petrásková L, Křen V (2010) *β-N*-Acetylhexosaminidase: what's in a name…? Biotechnol Adv 28:682–693
15. Sorimachi K, Okayasu T (2013) Phylogenetic tree construction based on amino acid composition and nucleotide content of complete vertebrate mitochondrial genomes. IOSR J Pharm 3(6):51–60

16. Tahrokh E, Ebrahimi M, Ebrahimie E, Ebrahimi M, Zamansani F, RahpeymaSarvestani N, Mohammadi-Dehcheshmeh M (2011) Comparative study of ammonium transporters in different organisms by simultaneous study of a large number of protein features using data mining algorithms. Genes Genom 33:561–571
17. Verbanck M, Le S, Pages J (2013) A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. BMC Bioinform 14:42
18. Zhang YP, Sheng YJ, Zheng W, He PA, Ruan JS (2015) Novel numerical characterization of protein sequences based on individual amino acid and its application. BioMed Res Int 2015:1–8