Journal of Phylogenetics & Evolutionary Biology

ISSN: 2329-9002

Satpathy et al., J Phylogen Evolution Biol 2013, 1:2
DOI: 10.4172/2329-9002.1000108

Research Article Open Access

Computational Phylogenetic Study and Data Mining Approach to Laccase Enzyme Sequences

Raghunath Satpathy*, Rashmiranjan Behera, Susant Ku Padhi, Rajesh Kumar Guru

Department of Biotechnology, MIRC Laboratory, MITS Engineering College, Rayagada, Odisha, India

Abstract

Currently data mining is an essential tool to discover the hidden data and important patterns from a large data set. The present work is a pilot study that compares the result of sequence based phylogenetic study and some of the physicochemical and structural feature based clustering of Laccase enzyme sequences. Total of 50 homologous sequences were obtained specific to each of the organism like plant, fungi and bacteria. Multiple sequences alignment of sequences was performed followed by phylogenetic tree construction and consistency study also to observe the major clusters. Again the major domain and motif analysis was done to support the study in the divergence pattern of Laccase enzyme sequences. There after 13 numbers of physicochemical and structural features were computed for each enzyme sequences. Then data normalisation and k-means clustering technique revealed that the fungi, bacteria and plant were obtained in three distinct clusters. The analysis indicates that the result of sequence based classification is in a good agreement with physicochemical basis of classification of proteins. The methods can be further optimised for different clustering algorithm to obtain specific physicochemical features that would help to classification of proteins.

Keywords: Data mining; Phylogenetic study; Clustering; Physicochemical features; Laccase enzyme**;** Motif and Domain

Introduction

Laccases are a group of multi-copper containing enzymes that catalyze the oxidation of phenolic compounds by reduction of oxygen to water [1].These enzymes having a broad natural substrate range, which is a major attractive feature of Laccases to biotechnological/ industrial applications [2]. Laccase have been found in organisms from the bacteria to plants and fungi that are present in a wide spectrum of environments. The occurrence of the conserved domain and motif was found in both pro and eukaryotic proteins that include a variety of enzymes led to the concept that sequence and structure pattern is responsible for the common function of the enzyme under a great variety of environmental impact. More than 60 fungal strains, belonging to various classes such as Ascomycetes, Basidiomycetes and Deuteromycetes, have been observed to produce Laccase [3,4]. Laccases are originally discovered in the exudates of *Rhus vernicifera*, the Japanese lacquer tree and subsequently demonstrated as a fungal enzyme as well [5]. There are many plant species in which the Laccase enzyme has been detected includes lacquer, mango, mung- bean, peach, pine, prune, and sycamore [6]. Laccase has also been discovered in a number of bacteria including *Bacillus subtilis*, *Caulobacter crescentus*, *Escherichia coli* etc. [7]. In the presence of different mediators, Laccases are widely used in many industrial processes and environmental bioremediations purposes. Their commercial applications are found in the pulp and paper industry, bio-bleaching, biosensing and beverage refining [8]. Various methods have been adopted to classify the Laccase enzyme sequences and one of the common methods is phylogenetic based classification, which is a sequence based clustering method by Multiple Sequence Alignment (MSA) [9].

Sequence analysis of proteins which are shared by diverse taxonomic groups provides the information about their divergence. Comparison of the amino acid sequences in between different species having functionally similar proteins has been used to estimate the amount of genetic similarity between species [10]. The usual methods of protein based phylogeny are based on multiple alignments of protein sequences and calculation of distances (insertions, deletions and mutations) between these sequences. From the distance matrix the appropriate clustering method is used to obtain the phylogenetic tree. Basically the phylogenetic analysis of enzyme sequences is a powerful tool for organization and interpretation of the taxa. With even a very basic understanding of general principles and conventions, it is possible to obtain clear valuable information about the origin, evolution and possible function of the proteins from a phylogenetic tree [11]. But most of the time the output obtained from the multiple alignment method usually fluctuates with the alignment parameters (number of matches, mismatches and gaps) [12]. So there is a haunt for suitable methods, which are to be adopted for obtaining a reliable alignment among sequences. Methods have been proposed for clustering of biological sequences based on their physicochemical properties [13,14]. The identification of similar groups or clusters of data showing similar behaviour is an important aspect of classification [15]. Hence the clustering methods play a major role which has been extensively applied specifically in sequence analysis to group homologous sequences into gene or protein families [16,17].

The aim of present work is to analyse the Laccase enzyme sequences from different sources of organism by both phylogenetic analysis and data mining approach. This ultimately aims to cluster various physicochemical and structural parameters of the sequences despite of their origin.

Materials and Methods

Phylogenetic study and Motif/domain computation

Laccase enzyme representative sequences for plants, bacteria and fungi were retrieved from uniprot data base. The individual sequences were further analysed by PSI-BLAST, which was carried out to find their

***Corresponding author:** Raghunath Satpathy, Department of Biotechnology, MIRC Laboratory, MITS Engineering College, Rayagada, Odisha-765017, India, E-mail: rnsatpathy@gmail.com

Received January 25, 2013; **Accepted** March 11, 2013; **Published** March 20, 2013

Citation: Satpathy R, Behera R, Padhi SK, Guru RK (2013) Computational Phylogenetic Study and Data Mining Approach to Laccase Enzyme Sequences. J Phylogen Evolution Biol 1: 108. doi:10.4172/2329-9002.1000108

group specific homologs. All the group of sequences were combined to form a common data set. Phylogenetic analysis was carried out by MEGA 4 software [18]. Initially an unrooted tree was obtained by NJ method and a particular out group was chosen and added based on branch length. Further accordingly the root was placed based on out group and consistency of tree was analysed by NJ methods [19]. This Neighbour Joining (NJ) algorithm does not make the assumption of molecular clock and adjust for the rate of variation among branches. It begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. The pair that yields the smallest sum is considered the closest neighbours and is thus joined. Then a new branch is inserted between them and the rest of the tree and the branch length is recalculated. This process is repeated until only one terminal is present. Similarly UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative or hierarchical based clustering method used commonly in bioinformatics for the creation of phenetic trees in a stepwise manner [20].

Physicochemical feature retrieval

For calculation of various physicochemical features of the protein sequences Protparam tool was used and secondary structure was predicted GORV software The types of physicochemical parameters like no. of amino acids, no. of atoms, molecular weight, isoelerctric point, number of negatively and positively charged amino acids, extinction coefficient, instability index, aliphatic index and GRAVY (grand average hydropathy) were calculated. In addition to this secondary structures were calculated by GORV software and added to the existing computed features. Details data have been given in supplementary material. The computational methods used to calculate the physiochemical and structural features by the server is given as below. The pH at which a protein carries no charge and exists as zwitterion is termed as Isoelectric point (pI). The instability index which gives clue about the stability of a protein in vitro can be calculated using the following formula: i=L-1. Instability index=(10/L) *Sum DIWV(x(i)x(i+1)) at i=1. Where L denotes length of sequence, DIWV(x(i)x(i+1)) is the instability weight value for the dipeptide starting in position i. The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic side chains. Aliphatic index=X(Ala)+a*X(Val)+b*X(Leu)+b*X(Ile) where X (Ala), X (Val), X (Ile) and X (Leu) are the amino acid compositional fractions. The Grand Average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. Similarly the secondary structure features was predicted using GOR V server. The GOR V algorithm computes secondary structure from the sequence information by combining the information theory, Bayesian statistics and evolutionary information. The GOR in its fifth version has been achieved an accuracy of prediction Q_3 of 73.5% [21].

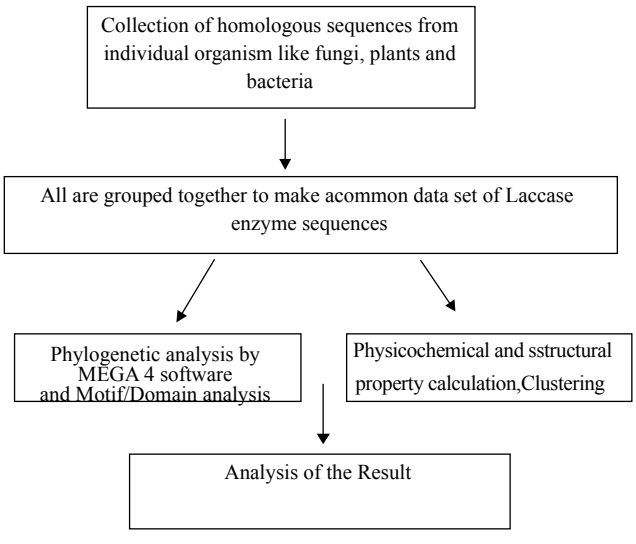

Figure 1: The detailed methodology followed for the present work.

Data normalisation and clustering

Data normalisation is done to obtain an unbiased result while clustering the data. All the computed data were normalised to a linear manner so that all the values will remain between 0 and 1. For normalisation of the data following formulae was used. Further *k*-means clustering of the computed features was done by Genesis tool [22].

$$\text{Normaliseddata} = \frac{\text{Original data value} - \text{Minimum data value}}{\text{Maximum data value} - \text{Minimum data value}}$$

To the normalised data, *k*-means clustering approach was performed. In data mining technology, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \le n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares.

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu \|^2$$

Where, μ_i is the mean of points in S_i and $\| X_j - \mu_i \|^2$ is a chosen distance measure between a data point X_j and the cluster centre μ_i.

The detailed procedure of the present work has been given in a schematic manner as below (Figure 1).

Results and Discussion

Analysis of phylogenetic consequences

All total 50 sequences from all species were selected for the present study. The sequence data about 50 sequences were found to be reliable for phylogenetic analysis as well as for data mining approach [23-25]. After complete alignment of the sequences by Clustal–W tool integrated with MEGA 4 software, boot- strapping was performed for 1000 times. Further Neighbour-Joining (NJ) and Unweighted pair of arithmetic means (UPGMA) methods was used to construct the unrooted and rooted phylogenetic tree. The phylogenetic tree shows a taxonomic clustering through the major taxa (Planate, Fungi, and Bacteria) (Figure 2). Again to check the reliability an out group sequence *Halobacterium sp. DL1* was chosen based on the PSI-BLAST score and then clustering was made to observe the out group position in the clusters of sequences (Figure 3). The multiple sequence analysis shows, a total of 18 different conserved amino acid positions. This suggests that these conserved amino acid residues have an important function in case of Laccase sequences and its evolution from lower organisms (bacteria) to higher organisms (fungi and plants). The information may also be useful to design PCR (polymerase chain reaction) primers for the Laccase gene isolation purpose.

The taxonomic relationship between plant, bacteria and fungi

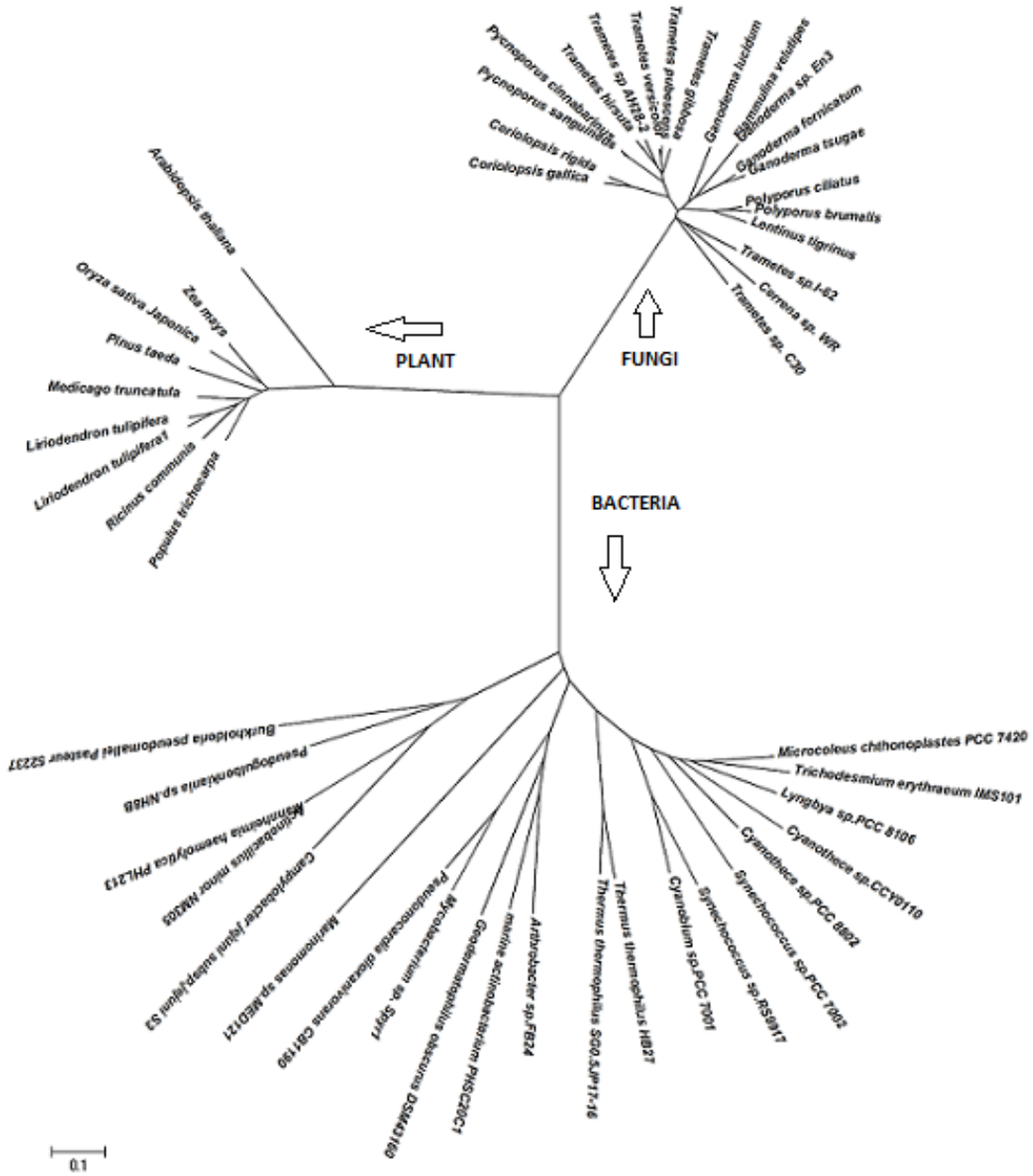

Figure 2: Unrooted phylogenetic tree constructed by NJ method by MEGA 4 tool after 1000 bootstrap.

species based on Laccase enzyme sequences was revealed by constructing the unrooted tree with and without out groups (Figure 4).

Further to root the tree, it is necessary to add an outgroup, which is a (unrelated) group of species or single species that is not included in the group of species under the study [26]. The outgroup *Halobacterium sp. DL1* was selected on the basis of scores of the sequence alignment. Placing of the out group taxa onto the phylogeny by connecting them somewhere below the ancestor for the entire taxa group was performed (Figure 5). To analyse the consistency of the phylogenetic tree bootstrap method is used [27]. The bootstrap is statistical procedures that resample the data and determines how strongly supported are different nodes on the tree. It is a measure of the internal consistency of the data. Boot strap values range from 100%, which indicates strongest support, up to ≥ 90% indicates very strong support also the values <50% indicates that the branch is less or not even supporting. In this case these samples were resampled the data with 1000 replicates. In this analysis, it was found that except few branches almost all branches having its bootstrap support is >70% (Figures 4 and 5). Generally bootstrap values of 70% and higher indicate the real groupings which have been proved [28].

Motif and Domain level analysis

A total 5 conserved prosite motif signatures were found in the sequences also 9 prodom domains and 5 pfam domains have been obtained. Multi cupper oxidase signature domain 1 is present in fungi alone and both signature 1 and 2 are present in plants and bacteria. Similarly Cu-oxidase, Cu-oxidase 2 and Cu-oxidase 3 pfam computed domains were obtained for almost all sequences.

Detail data are available in supplementary material 2. Combined

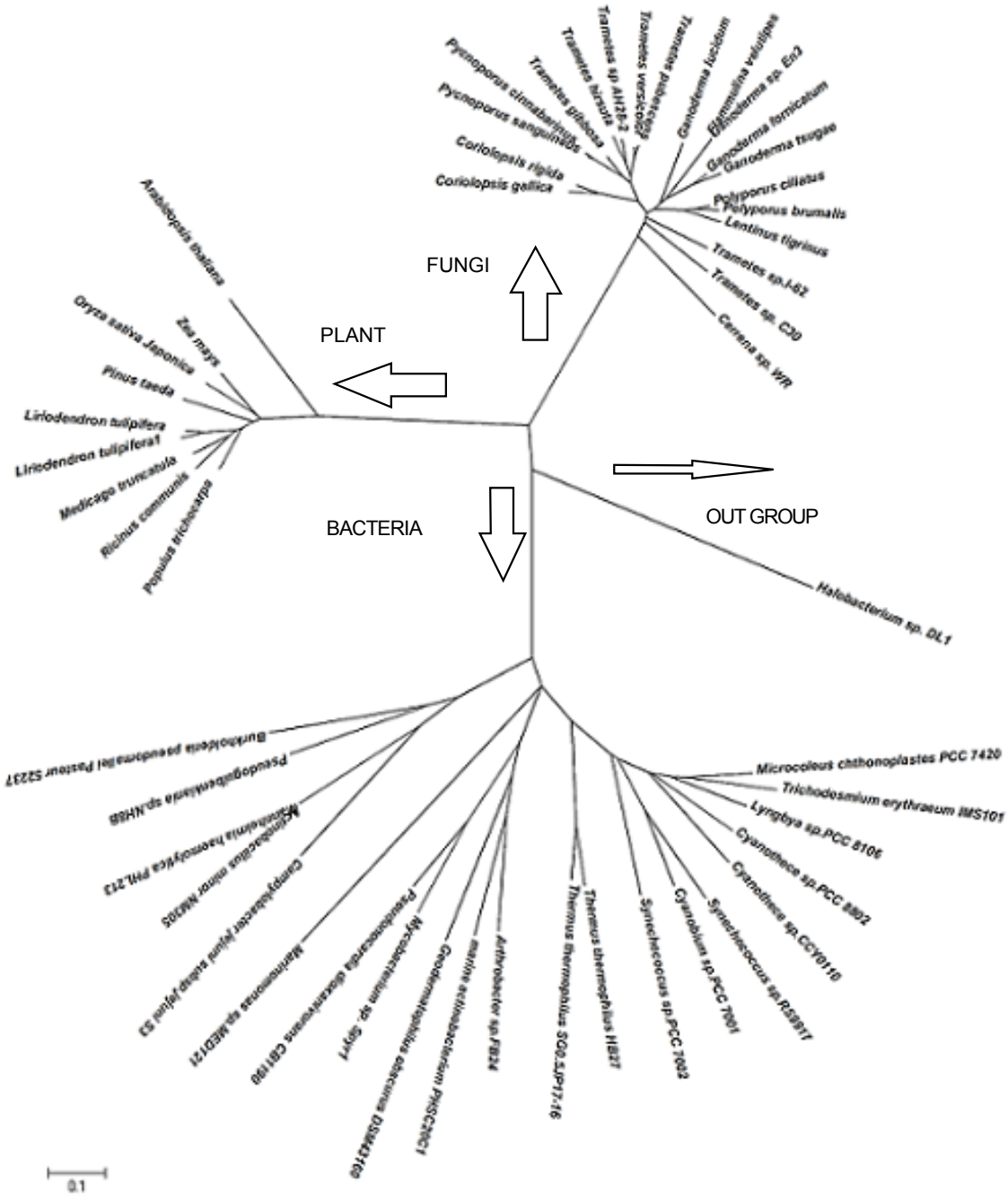

Figure 3: Unrooted phylogenetic tree with out-group sequence *Halobacterium sp. DL1* constructed by NJ method by MEGA 4 tool after 1000 bootstrap.

with the results of phylogenetic analysis presented above confirm that the three different group of organisms are having distinct phylogeny based on both sequence and structure. Our comparative analyses have also identified a specific number of motif and domain distributions in whole protein sequence of Laccsae enzymes from different organisms, which are distinctive characteristics of the species.

Analysis of *k*-means clustering method

K-means clustering is an algorithm among several methods, which attempts to find groups in the data [20]. The clustering was done for k=3 as three different group of organisms were studied; also 3 major clusters were obtained as done by the phylogenetic analysis (Figures 2-5). The clustering approach resulted 3 different clusters of the three organisms cluster 1 consists of fungal Laccase sequences, followed by plants and bacteria in cluster 2 and 3 respectively (Figure 6). The conventional method for classification is sequence based phylogenetic analysis, also considering various structural features like motif, domains etc [29]. But the physicochemical and structural descriptors of the proteins also can be used to classify proteins. Various computational approaches including clustering methods has been successfully used in many cases for efficient classification of protein sequences as well as structure and function [30]. After clustering result was obtained, based on the cluster the original data set was again analysed to find out the causes of

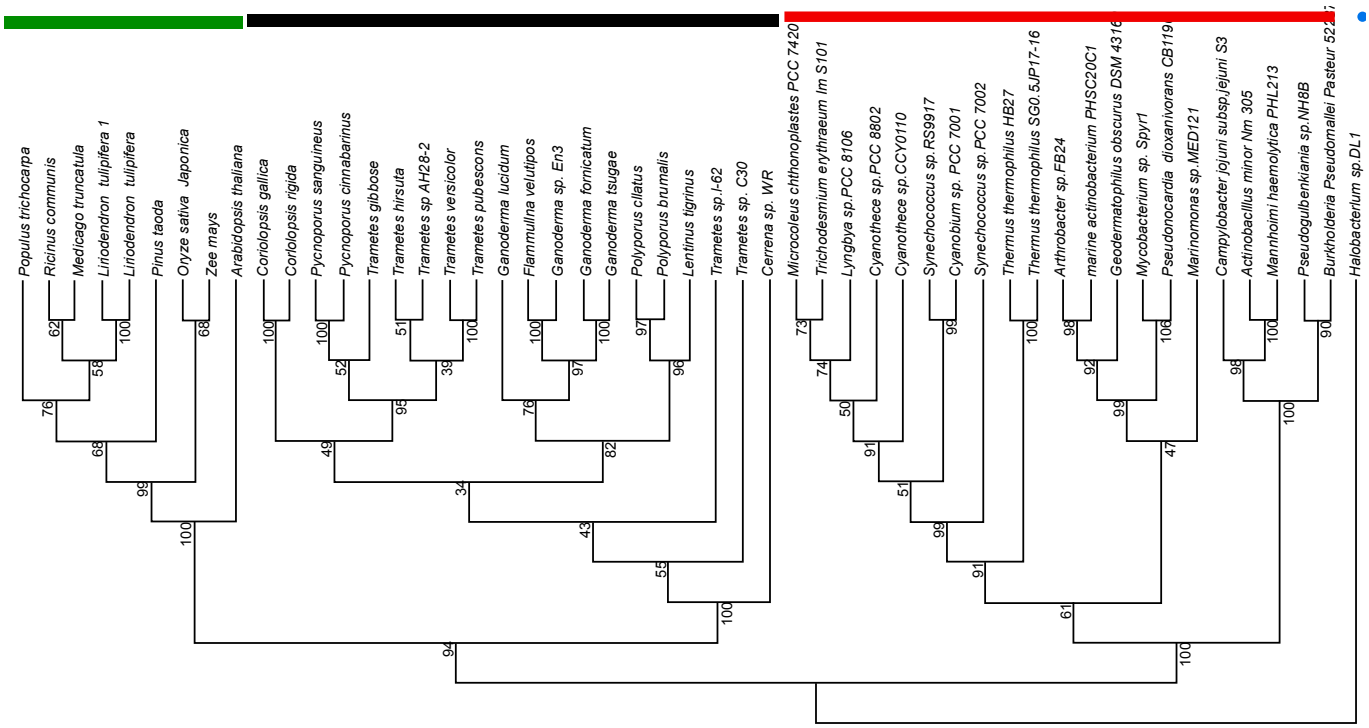

Figure 4: Phylogenetic tree constructed by NJ method (with out-group) that show the relationship among 50 sequences of Laccase enzymes. There are 3 major clusters; red colour indicates bacteria, black colour fungi and green colour indicate plant and the out group indicated as blue dots.

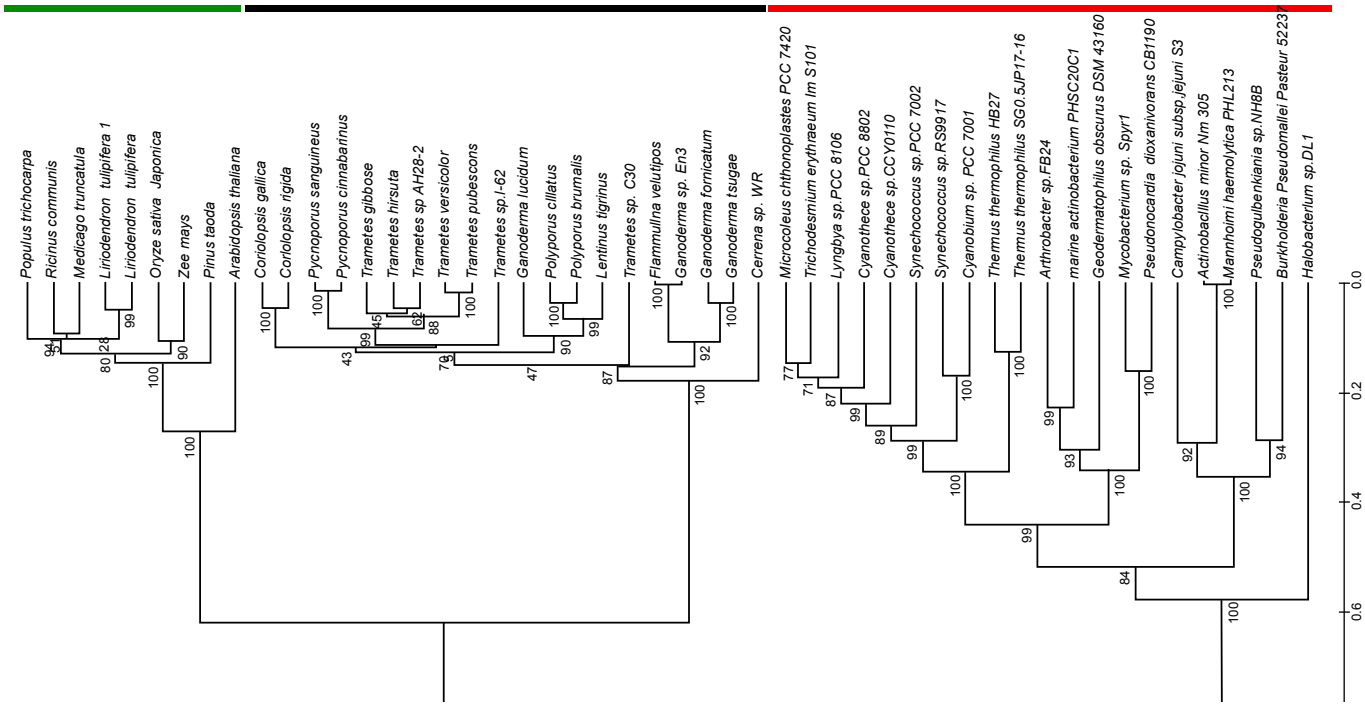

Figure 5: Phylogenetic tree (rooted) was constructed by UPGMA method (without out group) also show the three major clusters; red colour indicates bacteria, black colour fungi and green colour indicate plant.

variations that makes them into a group (cluster). The major features obtained were tabulated (Table 1).

There are many phylogenetic analysis has been performed to analyse the protein sequences by considering the alphabet. Whereas based on the physicochemical and structural information can also be applicable to find out the relationship among the taxa.The results obtained in this *insilico* analysis indicates that, the fungal, bacterial and plant derived Laccase can be classified both by sequence based cluster and also by

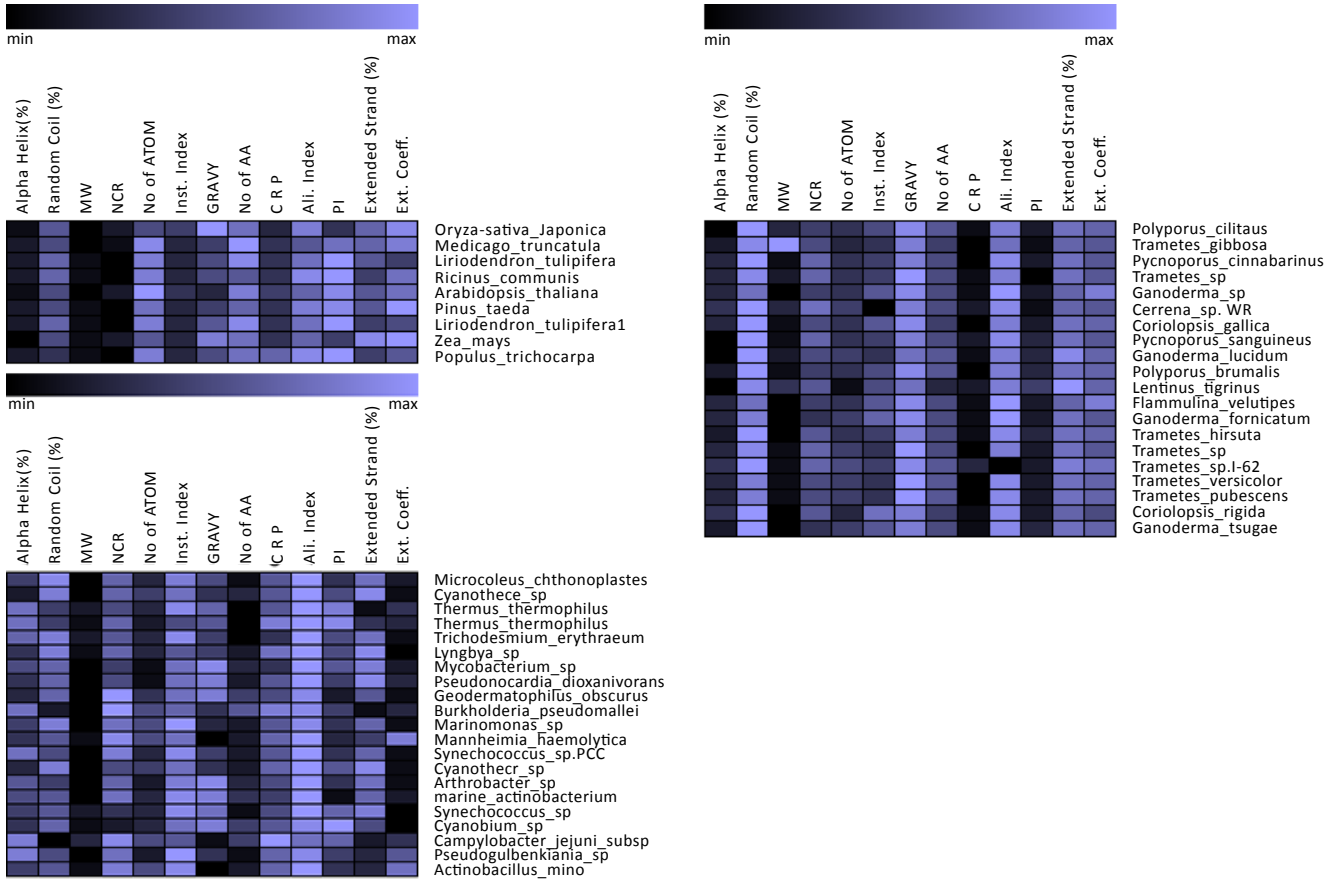

Figure 6: Clusters obtained after clustering the normalised data by k- means clustering method. The abbreviations used are, MW (molecular weight), NCR (negatively charged residues), CRP (positively charged residues), PI (isoelectric point), AA (amino acids), Ext. Coefficient (extinction co-efficient), GRAVY (grand average hydropathy).

Clusters	No. of sequences	Charged Residues	Secondary Structure (%)	Instability Index	pI range
Cluster 1 (Fungi)	20	No. of negative charge is more	Coil > Sheet>Helix	All are stable	4.21-5.54
Cluster2 (Plants)	9	No. of positive charge is more	Coil > Sheet>Helix	All are stable	9.09-9.84
Cluster3 (Bacteria)	21	No. of negative charge is more	Coil > Sheet>Helix	All are stable except one sequence	4.89-8.81

Table 1: Major feature variations obtained from clustered sequences.

physicochemical and structural properties i.e. the taxa in terms of Laccase enzyme sequences that are distinct not only in phylogenetic terms but also in molecular, structural and physicochemical terms (Figure 6).

Analysing properties of the proteins experimentally is a difficult task. But due to application of data mining technology we are able to obtain the important information that well suitable for classification of a group of proteins from different sources. This classification can significantly contribute in the understanding of the evolutionary relations between the species at molecular level. Due to the considerable importance of Laccase enzymes, more contribution is expected for the detailed investigation of the activity and functional analysis of enzymes. The data mining (clustering method) used in this work shows how to find out the similar group of the sequences without go for aligning them. Moreover the data mining approach is a useful tool to extract new features like "*physicochemical fingerprint*", which make it enable for classification of a large number of protein sequence data set.

Acknowledgment

We are thankful to Chief executive of MITS Engineering College, Rayagada, and Odisha for his encouragement and providing us MIRC LAB for computing facility.

References

1. Gianfreda L, Xu F, Bollag JM (1999) Laccases: A useful group of oxidoreductive enzymes. Biorem J 3: 1-26.
2. Machado KMG, Matheus DR (2006) Biodegradation of remazol brilliant blue R by ligninolytic enzymatic complex produced by *Pleurotus ostreatus*. Braz J Microbiol 37: 468-473.
3. Dwivedi UN, Singh P, Pandey VP, Kumar A (2011) Structure–function relationship among bacterial, fungal and plant laccases. Journal of Molecular Catalysis B: Enzymatic 68: 117-128.
4. Singh Arora D, Kumar Sharma R (2010) Ligninolytic fungal laccases and their biotechnological applications. Appl Biochem Biotechnol 160: 1760-1788.
5. Giardina P, Faraco V, Pezzella C, Piscitelli A, Vanhulle S, et al. (2010) Laccases: a never-ending story. Cell Mol Life Sci 67: 369-385.
6. Morozova OV, Shumakovich GP, Gorbacheva MA, Shleev SV, Yaropolov AI (2007) "Blue" Laccases. Biochemistry (Mosc) 72: 1136-1150.

7. Endo K, Hosono K, Beppu T, Ueda K (2002) A novel extracytoplasmic phenol oxidase of Streptomyces: its possible involvement in the onset of morphogenesis. Microbiology 148: 1767-1776.
8. D'Souza-Ticlo D, Sharma D, Raghukumar C (2009) A thermostable metal-tolerant laccase with bioremediation potential from a marine-derived fungus. Mar Biotechnol (NY) 11: 725-737.
9. Hoegger PJ, Kilaru S, James TY, Thacker JR, Kües U (2006) Phylogenetic comparison and classification of Laccase and related multicopperoxidase protein sequences. FEBS J 273: 2308-2326.
10. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1: e45.
11. Mallika V, Sivakumar KC, Soniya EV (2011) Evolutionary Implications and Physicochemical Analyses of Selected Proteins of Type III Polyketide Synthase Family. Evol Bioinform Online 7: 41-53.
12. Harrison CJ, Langdale JA (2006) A step by step guide to phylogeny reconstruction. Plant J 45: 561-572.
13. Banerjee AK, Arora N, Murty USN (2008) Classification and Regression Tree (CART) Analysis for Deriving Variable Importance of Parameters Influencing Average Flexibility of CaMK Kinase Family. Electronic Journal of Biology 4: 27-33.
14. Bakis Y, Otu HH, Sezerman OU (2012) Inferring phylogenies from physico-chemical properties of DNA. American Journal of Bioinformatics Research 2: 1-6.
15. Barile BB (2012) more work on K -Means Clustering Algorithm: The Dimensionality Problem. Int J Comp Appl 44: 23-30.
16. Fayech S, Essoussi N, Limam M (2009) Partitioning clustering algorithms for protein sequence data sets. BioData Min 2: 3.
17. Sugár IP, Sealfon SC (2010) Misty Mountain clustering: application to fast unsupervised flow cytometry gating. BMC Bioinformatics 11: 502.
18. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599.
19. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.
20. Zhong W, Altun G, Harrison R, Tai PC, Pan Y (2005) Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Trans Nanobioscience 4: 255-265.
21. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A (2005) GOR V server for protein secondary structure prediction. Bioinformatics 21: 2787-2788.
22. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18: 207-208.
23. Murty USN, Banerjee AK, Arora N (2009) An In Silico Approach to Cluster CAM Kinase Protein Sequences. J Proteomics Bioinform 2: 97-107.
24. Kumar V, Singh G, Verma AK, Agrawal S (2012) In silico characterization of histidine Acid phytase sequences. Enzyme Res 2012: 845465.
25. Gupta RS, Gao B (2009) Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus *Clostridium sensu stricto* (cluster I). Int J Syst Evol Microbiol 59: 285-294.
26. Baldauf SL (2003) Phylogeny for the faint of heart: a tutorial. Trends Genet19: 345-351.
27. Soltis PS, Soltis DE (2003) Applying the bootstrap in phylogeny reconstruction. Statistical Science 18: 256-267.
28. Jeffrey R, Eric CR (2007) Review of Phylogenetic Tree Construction University of Louisville Bioinformatics Laboratory Technical Report Series 1-7.
29. Mohammed A, Guda C (2011) Computational Approaches for Automated Classification of Enzyme Sequences. J Proteomics Bioinform 4: 147-152.
30. Mocz G (1995) Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins. Protein Sci 4: 1178-1187.