LacSubPred: predicting subtypes of Laccases, an important lignin metabolism-related enzyme class, using in silico approaches

Tyler Weirick1,2, Sitanshu S Sahu1,2, Ramamurthy Mahalingam2, Rakesh Kaundal3*

From 11th Annual MCBIOS Conference
Stillwater, OK, USA. 6-8 March 2014

Abstract

Background: Laccases (E.C. 1.10.3.2) are multi-copper oxidases that have gained importance in many industries such as biofuels, pulp production, textile dye bleaching, bioremediation, and food production. Their usefulness stems from the ability to act on a diverse range of phenolic compounds such as o-/p-quinols, aminophenols, polyphenols, polynamines, aryl diamines, and aromatic thiols. Despite acting on a wide range of compounds as a family, individual Laccases often exhibit distinctive and varied substrate ranges. This is likely due to Laccases involvement in many metabolic roles across diverse taxa. Classification systems for multi-copper oxidases have been developed using multiple sequence alignments, however, these systems seem to largely follow species taxonomy rather than substrate ranges, enzyme properties, or specific function. It has been suggested that the roles and substrates of various Laccases are related to their optimal pH. This is consistent with the observation that fungal Laccases usually prefer acidic conditions, whereas plant and bacterial Laccases prefer basic conditions. Based on these observations, we hypothesize that a descriptor-based unsupervised learning system could generate homology independent classification system for better describing the functional properties of Laccases.

Results: In this study, we first utilized unsupervised learning approach to develop a novel homology independent Laccase classification system. From the descriptors considered, physicochemical properties showed the best performance. Physicochemical properties divided the Laccases into twelve subtypes. Analysis of the clusters using a t-test revealed that the majority of the physicochemical descriptors had statistically significant differences between the classes. Feature selection identified the most important features as negatively charges residues, the peptide isoelectric point, and acidic or amidic residues. Secondly, to allow for classification of new Laccases, a supervised learning system was developed from the clusters. The models showed high performance with an overall accuracy of 99.03%, error of 0.49%, MCC of 0.9367, precision of 94.20%, sensitivity of 94.20%, and specificity of 99.47% in a 5-fold cross-validation test. In an independent test, our models still provide a high accuracy of 97.98%, error rate of 1.02%, MCC of 0.8678, precision of 87.88%, sensitivity of 87.88% and specificity of 98.90%.

Conclusion: This study provides a useful classification system for better understanding of Laccases from their physicochemical properties perspective. We also developed a publically available web tool for the characterization of Laccase protein sequences (http://lacsubpred.bioinfo.ucr.edu/). Finally, the programs used in the study are made available for researchers interested in applying the system to other enzyme classes (https://github.com/tweirick/SubClPred).

* Correspondence: rkaundal@ucr.edu
3Bioinformatics Facility, Department of Botany & Plant Sciences, Institute for Integrative Genome Biology (IGIB), University of California, Riverside, California, 92521, USA
Full list of author information is available at the end of the article

© 2014 Weirick et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background
Laccases (EC 1.10.3.2) are the largest sub-group of multicopper oxidases which includes ascorbate oxidases (EC 1.10.3.3), ferroxidases or ceruloplasmins (EC 1.16.3.1) and nitrate reductases (EC 1.7.2.1). Laccases were first discovered in the sap of the Japanese lacquer tree *Rhus vernicifera*. Since then they have been found in many taxa including plants, fungi, bacteria, and metazoa. Laccases are involved in a diverse range of cellular activities such as lignin degradation, lignin biosynthesis, pigment production, plant pathogenesis, melatonin production, spore coat resistance, morphogenesis and detoxification of copper [1-5]. Laccases are also widely used for industrial purposes. For example, Laccases are in paper and pulp, textile, and petrochemical industries for detoxification of industrial effluents [6]. In medicine, Laccases are used for certain medical diagnostics and as catalysts for the manufacture of anti-cancer drugs [6]. They are also used for environmental remediation of herbicides, pesticides and as explosives in soil and cleaning agents for certain water purification systems. In commercial products, they are found in cosmetics, denim bleaching, wine and beer stabilization, fruit juice processing, color enhancement of tea and even baking [6,7]. Laccases are popular in industry for a number of reasons. They are better for the environment, and have fewer non-specific reactions than conventional oxidation technologies. Many Laccases are extracellular enzymes which makes their purification simple. Compared with other oxidative enzymes, these are easier to use as they catalyze reactions with molecular oxygen and do not need reactive oxygen species catalysis [6,8]. Currently, fungal Laccases comprise most widely studied and commercially used Laccases. However, there is much interest in bacterial Laccases also due to their higher temperature stability and ability to operate at different pHs than fungal Laccases. Generally, Laccases are composed of dimeric or tetrameric glycoproteins with each monomer containing a copper containing site. These copper sites may be one of three types: Type-1 or blue copper, Type-2 or normal copper, and Type-3 or coupled-dinuclear centers. These copper binding motifs have been shown to be highly conserved across all Laccases, with a trend towards greater similarity in the N and C terminal domains as these are the copper containing domains. It has been noted that the size of the central binding pockets are larger in bacterial Laccases than in fungal or plant Laccases. These copper binding sites yield significant differences in conserved residues for Laccases of bacteria, fungi, and plants [9].

Fungal Laccases
Fungal Laccases comprise the bulk of experimentally studied Laccases. They occur in many fungal species and are thought to play important roles in morphogenesis, fungal-plant interactions, stress defense, pigment production, and lignin degradation. While typically studied with respect to biomass degradation, most fungi found producing several isoenzymes of different types, enzymatic or physical properties, and expression levels. These can vary even more between species [8]. For example, it has been reported that one of the most efficient lignin degraders, *Phanerochaete chrysosporium* produces a Laccase different than other efficient lignin degrading fungi [10]. While most Laccases are extracellular enzymes, many fungal taxa produce intracellular Laccases [8] also. This is especially interesting when compared with enzymes of similar function such as lignin peroxidases which are strictly extracellular. It is speculated that the cellular localization of Laccases may be connected their function and substrate ranges. This hypothesis still remains elusive due to the majority of studied fungal Laccases coming from wood-rotting basidiomycetes. The enzymatic properties of fungal Laccases vary greatly such as temperatures vary from 25-80°C, pH optiums: 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid) (ABTS) from 2.0-5.0, 2,6-dimethoxyphenol (DMP) from 3.0-8.0, guaiacol from 3.0-7.0, and syringaldazine from 3.5-7.0. Similarly, $K_m$ (µM) ranges vary a lot such as: ABTS from 4-770, DMP from 26-14720, Guaiacol from 4-30000, syringaldazine from 3-4307. Also $K_{cat}$ (S⁻¹) vary in a broad range as: ABTS from 198-350000, DMP from 100-360000, Guaiacol from 90-10800 and syringaldazine from 16800-28000. These properties can further be altered by glycosylation.

Plants Laccases
Traditionally plant Laccases were considered to be only extracellular enzymes involved in the radical-based lignin polymerization. However, a high degree of divergence among Laccases within a single plant species has been observed, such as ryegrass which contains 25 different Laccase genes. Also, it is reported that Laccases lack N-terminal signal peptides for secretion but have signals targeting to other cellular components such as the endoplasmic reticulum or peroxisomes. Another study on poplars showed that Laccase repression had no effect on lignin production. Despite the evidence for novel functions and many known functions in other taxa, the grouping of plant Laccases still remain elusive [11].

Bacterial Laccases
Bacterial Laccases are known to be widespread in prokaryotes; however, only few have been experimentally characterized. To date, bacterial Laccases have been found mostly to be involved in lignin degradation, catabolism of phenolic compounds, cell pigmentation, morphogenesis, and copper defense [12-14]. The best studied bacterial Laccase is CotA and endospore coat protein from *Bacillus*
subtilis which produces a melanin like pigment. This enzyme has generated high amounts of interest due to its extremely high temperature stability. Bacterial Laccases are also unique due to the lack of cellular partitions in prokaryotes. The reactions catalyzed by Laccases can produce quinones and semiquinones as by-products, which are powerful inhibitors of the electron transport chain [5].

Other Laccases
In metazoan, Laccases exist in mammals as well as invertebrates. The roles of Laccases in mammals do not appear to be well understood, however, insect Laccases are known to be involved in cuticle formation [12]. Cuticle tanning also known as sclerotization and pigmentation is the process through which proteins in the exoskeleton are conjugated. This causes the exoskeleton to become insoluble, harder, and darker.

Classification of Laccases: current view
Laccases are currently classified as part of a larger classification scheme for multi copper oxidases [15,16]. This is based on multiple sequence alignments and seems to classify by taxonomical association. The current classification system i.e. “The Laccase Engineering Database” (LccED), classifies multi copper oxidases into eleven classes: basidiomycetes Laccases, ascomycete Laccases, insect Laccases, fungal pigment MCOs, Fungal ferroxidases, fungal and plant ascorbate oxidases, plant Laccases, bacterial CopA proteins, bacterial bilirubin oxidases, bacterial CueO proteins, and SLAC homologs.

Machine learning-based classification systems
As discussed above, the current classification system for Laccases largely follow species taxonomy rather than substrate ranges, enzyme properties, or specific function. Although it has been observed that individual Laccases often exhibit distinctive and varied substrate ranges, and have different functions based on distinguishing pH values among different taxa. We hypothesize that a descriptor-based computational prediction system could be developed to generate a homology-independent classification system for better describing the functional properties of Laccases. In a previous study on feruloyl esterases (EC 3.1.1.73), an unsupervised learning approach was used to create a novel homology independent classification system for this enzyme class. Various bioinformatics tools were used to validate the identified classes [17]. In the present study, we followed a two-way computational strategy to identify and classify various Laccase subtypes by developing a python command line-based implementation of the unsupervised and supervised learning approaches, respectively. Further, we implemented our prediction models as a web-based prediction server to classify novel Laccase subtypes. The tool could be useful to the biofuel researchers and industry as well.

Methods
Dataset generation
Alternate names for Laccases were found via cross referencing with the KEGG database (http://www.kegg.jp/dbget-bin/www_bget?ec:1.10.3.2). To search for Laccase sequences, we combine these names to start as a basic query. Sequences with protein or transcript level evidence were selected to ensure high quality data as well as avoid potentially mislabeled multi-copper oxidases. Then we search UniprotKB for Laccase sequences using some search terms as listed in Table 1. Using the “browse by” option on Uniprot’s GUI the query was checked for possible contaminating sequences. The contaminant sequences were filtered out using NOT conditions (see Table 1). Finally, 329 protein sequences are collected with average sequence length above 200 residues. To further validate the quality of the datasets the protein descriptions of the data were analyzed with the text clustering functionality in Google-Refine version 2.5. A significant variation was found in the protein descriptions but no cases of contamination were found. As a final check of data quality, the lengths of the sequences were calculated and plotted on a bar graph shown in Figure 1. Sequences containing non-standards/ambiguous characters were removed from the data set.

Feature representation of Laccase proteins
It is important to extract better features of protein sequences to improve the performance of the machine learning method. We used several features such as amino acid composition (AAC), Conjoint Triad (CT), Composition-Transition-Distribution (CTD), Dipeptide composition (DIPEP), Geary autocorrelation descriptors, Moran autocorrelation, Moreau-Broto autocorrelation,
physicochemical properties and a composite vector of amino acid composition and physicochemical properties.

Amino acid composition (AAC)
Each protein sequence is represented as a 20-dimensional feature vector with each element corresponding to the percentage of one of the twenty amino acids [18]. For a given protein sequence \( x \), let the function \( f(x_i) \) represent the occurrence of the 20 standard amino acids. Thus, the composition of the amino acids \( P(x) \) in the given sequence can be represented as,

\[
P(x) = [P_1(x), P_2(x), \ldots, P_{20}(x)]
\]

where \( P(x_i) \) is given as,

\[
P(x_i) = \frac{f(x_i)}{\sum_{i=1}^{20} f(x_i)}, \quad i = 1, 2, 3, \ldots, 20
\]

Dipeptide composition (DIPEP)
Dipeptide sequence composition is similar to amino acid composition. However, it considers the percentages of dipeptides occurring in a given protein sequence [18]. Thus, the composition of each dipeptide is given as,

\[
P(x_i, x_j) = \frac{f(x_i, x_j)}{\sum_{i=1}^{20} \sum_{j=1}^{20} f(x_i, x_j)}, \quad i, j = 1, 2, 3, \ldots, 20
\]

where \( P(x_i, x_j) \) is the fraction of number of instances of a specific dipeptide \( f(x_i, x_j) \) and the total number of all dipeptides.

Conjoint triad (CT)
In conjoint triad, in addition to amino acid composition it considers the sequence order effect [19]. It is calculated by grouping the 20 standard amino acids into 7 groups based on physical and chemical similarity [(A,G,V), (I,L,F,P), (Y,M,T,S), (H,N,Q,W), (R,K), (D,E), (C)]. Triads are made from all combinations of three amino acids of these groups, resulting in a vector length of 343 (7 \times 7 \times 7). Thus, a protein sequence is represented as,

\[
P(x_i, x_j, x_k) = \frac{f(x_i, x_j, x_k)}{\sum_{i=1}^{7} \sum_{j=1}^{7} \sum_{k=1}^{7} f(x_i, x_j, x_k)}, \quad i, j, k = 1, 2, 3, \ldots, 20
\]

where \( f(x_i, x_j, x_k) \) is the number of occurrences of a specific triad and \( \sum_{i=1}^{7} \sum_{j=1}^{7} \sum_{k=1}^{7} f(x_i, x_j, x_k) \) is the number of all triads [19].

Composition-transition-distribution (CTD)
In this representation three local descriptors, Composition (C), Transition (T) and Distribution (D) are used in combination to construct the feature vector. These descriptors
are based on the variation of occurrence of functional
groups of amino acids within the primary sequence of pro-
tein [20]. Thus, before computing this feature the twenty
amino acids are clustered into seven functional groups
based on the dipoles and volumes of the side chains [19].
The composition descriptor computes the occurrence of
each amino acid group along the sequence. Transition
represents the percentage frequency with which amino
acid in one group is followed by amino acid in another
group. The distribution feature reflects the dispersion
pattern along the entire sequence by measuring the location
of the first, 25, 50, 75 and 100% of residues of a given
group. Hence, total 63 features (7 composition, 21 transi-
tion and 35 distribution) are constructed to represent a
protein.

**Autocorrelation feature vectors**

Autocorrelation features describe the level of correlation
between two protein sequences in terms of their specific
physicochemical property, which are defined based on
the distribution of amino acid properties along the
sequence. There are 8 amino acid properties used for
deriving autocorrelation descriptors.

**Moran autocorrelation**

The Moran autocorrelation (MAC) descriptor of a pro-
tein is defined as:

\[
D_{MAC}(d) = \frac{1}{N - d} \sum_{j=1}^{N-d} (P_j - \bar{P}) \times (P_{j+d} - \bar{P})
\]

where \(N\) is the length of the protein sequence, \(d = 1,2,...,30\) is the distance between one residue and its
neighbors, \(P_j\) and \(P_{j+d}\) are the properties of the amino acid
at positions \(j\) and \(j+d\) respectively, \(\bar{P} = \frac{1}{N} \sum_{j=1}^{N} P_j\) is the average
of the considered property \(P\) along the sequence.

**Geary autocorrelation**

Geary autocorrelation (GA) descriptor of a protein is
defined as:

\[
D_{GA}(d) = \frac{1}{2(N - d)} \sum_{j=1}^{N-d} (P_j - P_{j+d})^2
\]

\[\bar{P},\ N,\ P_j\text{ and }P_{j+d}\text{ are defined in the same way as above.}\]

**Moreau-Broto autocorrelation**

Moreau-Broto autocorrelation (MBA) descriptor of a
protein is defined as:

\[
D_{MBA}(d) = \sum_{j=1}^{N-d} P_j \times P_{j+d}
\]

**Physicochemical properties**

Physicochemical properties of amino acids have been
used successfully in numerous prediction tools [18]. In
this study, we grouped the amino acids of a protein into
classes based on some physicochemical properties. Also
the theoretical pI, molecular weight, and length of the
protein are used in the feature vector. The non-compo-
sition based values are divided by the length or mass on
the protein titan in order to provide values between one
and zero. Molecular weights were calculated by adding
the weights of the each amino acid in the sequence in a
suitable way related to their chemical activity. A detailed
description of these properties is provided in Table 2.

**Split amino acid composition**

Split amino acid composition aims to capture informa-
tion about signal peptides at their N- or C-terminal
region. The amino acid composition of the N-terminal

| Sr. No. | Physicochemical property            | Amino Acids | # features |
|---------|-------------------------------------|-------------|------------|
| 1       | log10(molecular weight)/7.0        | -           | 1          |
| 2       | log10(Sequence Length)/5.0         | -           | 2          |
| 3       | % Charged Residues                 | DREKH       | 3          |
| 4       | % Hydrophilic and neutral          | NGSTY       | 4          |
| 5       | % basic polar/positively charged   | HKR         | 5          |
| 6       | % acidic or negatively charged     | DE          | 6          |
| 7       | % aliphatic                        | AGLV        | 7          |
| 8       | % aromatic                         | FWY         | 8          |
| 9       | % small                            | DNT         | 9          |
| 10      | % tiny                             | AGPS        | 10         |
| 11      | % large                            | FRWY        | 11         |
| 12      | % hydrophobic and aromatic         | WF          | 12         |
| 13      | % hydrophobic and neutral          | ACGLMFPWW   | 13         |
| 14      | % amide                            | NQ          | 14         |
| 15      | % cyclic                           | P           | 15         |
| 16      | % hydroxylic                       | ST          | 16         |
| 17      | % contains sulfur                  | CM          | 17         |
| 18      | % H-bonding                        | CWNQSTYKRHDE| 18         |
| 19      | % acidic and amide                 | DENQ        | 19         |
| 20      | % ionizable                        | DEHCVKR     | 20         |
| 21      | % sulfur bonding                   | C           | 21         |
| 22      | % pl                               | -           | 22         |
| 23      | Molecular weight/ 400000.0         | -           | 23         |
| 24      | Sequence length / 38000.0          | -           | 24         |
region, Center, and C-terminal region are computed and then concatenated together. The N- and C-terminal regions are the first and last 25 amino acids in the sequence. Thus a protein sample is represented as a 60 element vector as,

\[ P(x) = [\text{AAC}_{N-\text{terminal}} \; \text{AAC}_{\text{Center region}} \; \text{AAC}_{\text{C-\text{terminal}}}] \] (8)

Unsupervised classification

Unsupervised learning organizes the data based on the similarity patterns between them. In this study, clustering was used to group the data into classes sharing same type of similarity not found in other classes. We followed the similar methodology as outlined in the paper [17]. We first used self-organizing map (SOM) to identify the possible number of groups in the dataset and used that information in k-means clustering to divide them in different clusters.

Self-organizing maps (SOM)

SOMs are a type of artificial neural networks used in unsupervised learning to produce low dimensional discrete representations of the vector space represented by some training data [21]. The discrete elements in SOMs are called nodes or neurons. It has been used widely in bioinformatics and computational biology mostly for tasks such as finding gene expression patterns and protein classification [22,23]. The SOM map contains m neurons, where each contains a d-dimensional prototype vector with d as the dimensions of the input vectors. First, initial values were given to each prototype vector. When training begins a vector ‘x’ from the input data is randomly chosen. The distances from ‘x’ to the prototype vectors are computed and the neuron closest to ‘x’ or best matching unit (BMU) is selected. The radius of the neighborhood of the DMU is calculated, any neurons found within the radius are deemed neighbors. The neighbor’s prototype vector is adjusted to be more similar to the input vector. This procedure was then repeated for certain iterations (N) [21]. In this study, SOM of multiple dimensions were studied and N was 10,000 for all dimensions. For the SOM implementation, we used an open source machine learning package ‘Orange.py’ which is freely available at http://orange.biolab.si[24].

K-means clustering

K-means clustering is a class of unsupervised learning algorithms which group input data set into ‘K’ parts or clusters [25] based on similarity measure. K-means is one of the oldest and simplest clustering methods, however still remains a useful tool for cluster analysis. It scales well to large data sets and medium numbers of clusters, however, has the drawback of needing to specify the number of clusters expected. The basic k-mean algorithm begins by initializing k cluster centers (centroids) and iterating to minimize the average distances between centroids and their cluster members. The data which are close to any cluster centroid belong to that cluster. The centroids were pre-computed using the neurons from the SOM. In this study, an open source machine learning library ‘Sci-Kit Learn’ was used to implement the k-means clustering method [26].

SOM for finding K number and centroid locations for K-means clustering

In this study, first an SOM network computed containing N neurons and calculates the Davies-Bouldin index (DBI) of the map treating the neurons as clusters. Then, (N) × (N-1) prototype maps were created by making all combinations of each neuron with the other neurons. The DBI is computed for all prototype maps, and the prototype map with the lowest DBI is selected. If the DBI of this map is lower than the current map the map is changed to other prototype map and the previous steps are repeated until no prototype map with a lower DBI can be found. This reduces the size of the map by one each iteration with the final number of neurons being used as the k value for k-means clustering and the cluster centroids are computed from the vectors belonging to each neuron. The efficiency of k-means clustering is measured using the difference between the inter-cluster and intra-cluster variance and the Davies-Bouldin index. As SOM find the clusters in random fashion, to get the optimum number of clusters, the clustering procedure was run 500 times for each vector type. The optimum number of clusters was chosen by selecting the cluster from the most often occurring cluster number with the largest intercluster and intracluster difference and smallest DBI.

Davies-Bouldin index (DBI)

The DBI is a metric for evaluating overall quality of a given set of clusters originally developed to aid in determining the optimum number of clusters within a dataset [27]. Minimization of the DBI of the clusters within a dataset seems to generally indicate natural partitions of data sets. However, it should be noted that this is a heuristic approach and good values do not always indicate the best clustering arrangement. DBI of a clustering approach is defined as,

\[ DB = \frac{1}{N} \sum_{i=1}^{N} D_i \] (9)

where \( D_i \) is the worst case scenario of all values of \( R_{ij} \),

\[ D_i = \max_{j \neq i} R_{ij} \] (10)
is a measure of the clustering quality, defined as

\[ R_{ij} = \frac{S_i + S_j}{M_{ij}} \]  \hspace{1cm} (11)

The measure of scatter \((S)\) within a given cluster \(i\), is defined as

\[ S_i = \sqrt{\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q} \]  \hspace{1cm} (12)

where \(X_i\) is a \(n\)-dimensional feature vector assigned to the cluster \(C_i\) and \(q\) was kept as two and \(M_{ij}\) is a measure of separation between two clusters defined as

\[ M_{ij} = A_i - A_{ij} \]  \hspace{1cm} (13)

where \(A_i\) is the centroid of cluster \(C_i\) containing samples \(X_1, X_2, \ldots, X_k\) and computed as,

\[ A_i = \frac{X_1 + X_2 + X_3 + \ldots + X_k}{k} \]

**Intra-cluster variance**

Intra-cluster variance was calculated using the Euclidean distances between the points in the cluster and the centroid of the cluster.

**Inter-cluster variance**

Inter-cluster variance was calculated using the Euclidean distance between the centroids of the clusters.

**Co-occurrence matrix analysis**

The cluster numbers returned from the clustering approach is arbitrary which presents a unique problem when trying to access the similarity between runs. Thus, to assess the consistency of belonging of samples in a particular group, a co-occurrence matrix was generated to show the number of times a given data sample in one group occurred with other groups. The higher the numbers of data samples occurring together, the more consistency the clusters in various runs.

**Support vector machine (SVM)**

SVMs are a class of supervised learning algorithms based on the optimization principle from statistical learning theory [28,29]. Support vector machines have been used widely in computational biology in diverse topics such as subcellular localization [18,30-32], protein function prediction [33], secondary structure prediction [34], disease forecasting [35]. SVMs solve classification problems by calculating a hyperplane that separates the training data with a maximum margin. For multi-class classification the classification is transformed into a series of binary classifications. There are numerous strategies for handling a multi-class problem separated into binary classifications and in this study the one-versus-rest approach was used. The SVM Classifiers were developed using the SVM_Light package (https://github.com/daoudclarke/pysvmlight), which is an open source package for SVM implementation [36]. In a preliminary study, the RBF kernel was found to perform best. Therefore, we used RBF kernel in all our SVM classifiers.

**Performance evaluation parameters**

To assess the performance of the developed models, we used a five-fold cross validation test on the training dataset and then tested the models in an independent test. In a five-fold cross-validation procedure, the original sample is randomly partitioned into five equal size subsamples. Of the five subsamples, a single subsample is retained as the validation data for testing the model, and the remaining four subsamples are used as training data. The cross-validation process is then repeated 5 times (the folds), with each of the 5 subsamples used exactly once as the validation data. The results from the five-folds are then averaged to produce a single estimation. The performance is measured by the parameters such as overall sensitivity, specificity, precision, Matthews Correlation Coefficient (MCC) and average accuracy. These parameters are defined as follows:

(i) **Sensitivity or coverage of positive examples:** It is the percent of positive samples correctly predicted,

\[ \text{Sensitivity}(S_n) = \frac{TP}{TP + FN} \]  \hspace{1cm} (14)

(ii) **Specificity or coverage of negative examples:** It is percent of negative samples correctly predicted as positive,

\[ \text{Specificity}(Sp) = \frac{TN}{TN + FP} \times 100 \]  \hspace{1cm} (15)

(iii) **Accuracy:** It is the percentage of correctly predicted samples,

\[ \text{Accuracy}(Acc) = \frac{TP + TN}{TP + FN + FP + FN} \times 100 \]  \hspace{1cm} (16)

(iv) **Error rate:** It is the total percentage of incorrect predictions is calculated as

\[ \text{Error rate (ER)} = \frac{FP + FN}{TP + FN + FP + FN} \times 100 \]  \hspace{1cm} (17)

(v) **Precision:** It is the percentage of positive PPIs those are correct identified true prediction,

\[ \text{Precision} = \frac{TP}{TP + FP} \times 100 \]  \hspace{1cm} (18)
(vi) Matthew’s correlation coefficient (MCC): it is considered to be the most robust parameter of any class prediction method. MCC equal to 1 is regarded as perfect prediction while 0 for completely random prediction.

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

where true positive (TP) is the numbers of positive samples that are predicted correctly; false negative (FN) is the number of positive samples that are predicted to be negative; false positive (FP) is the number of negative samples that are predicted positive and true negative (TN) is the number of negative samples that are predicted correctly as negative.

**Feature scaling**

To have knowledge of most relevant features for classification of Laccase types, a feature scaling approach is conducted. Feature scaling was performed using univariate feature selection using the functions provided by Sci-Kit Learn using the program scale_features.py[26]. Univariate feature selection implemented by considering each element of the descriptor vectors independent from one another and ranking them based on their occurrence between classes.

**Domain map and phylogenetic trees construction**

The program doMosaic was used to create domain maps for visualization of the domains in the initial data and newly generated classes [37]. Interproscan was used to get the information about the domains in the Laccases [38]. To show the relationship between Laccase samples, a phylogenetic tree was generated with the cleaned dataset using Clustal Omega version 1.0.3 [39]. dendroscope version 3.2.10 was used for the visualization of the tree.

**Results and discussion**

We have studied several SOM architectures to see the effect of clustering of the Laccases with many descriptors. The clustering algorithm was run 500 times for each SOM map size. The clustering performance of each descriptor is given the best run with a DBI of 0.37 with an inter-cluster variance of 0.0088 and intra-cluster variance of 0.0015. The performance of the physicochemical descriptor in each SOM dimensions is listed in Table 4. The proteins classified in each group after the clustering approach are listed in Table 5.

Analysis of the taxa in each class revealed that the majority of the classes were dominated by single taxa as reported in Table 5. Several review papers containing large tables of experimentally validated Laccases with various properties were considered to validate the clusters. Unfortunately, these were difficult to draw patterns from as the substrates tested varied widely and heterologously expressed Laccases often have drastically different activities due to different amounts of glycosylation [15,40,41]. To better understand what is driving the distinction of different classes, feature scaling was applied to the physicochemical properties of all classes together, as well as each class against each other. The major contributing features were the percentage of negatively charged amino acids, isolectric point, and the percentage of acidic or amide groups. The detailed information about the significant features is shown in Figure 3. This is particularly interesting as Laccases as a group operate over a wide range of pHs while individual enzymes seem to have fairly specific or broad pH and substrate ranges [41]. Also, it has been reported that different Laccases produced by the fungi *Coriolus versicolor* were easily distinguishable by their isolectric points [42]. The differences between classes in terms of physicochemical properties, the best features were calculated for all classes and shown in Figure 4. This showed that the variation seems to be strongly influenced by acid/base properties, and next to the small residues or aliphatic residues. The isolectric point occurred most often within the top three features with 45 cases, followed by basic amino acids with 34 cases, acidic with 32 cases, ionizable amino acids with 23 cases, acidic and amide with 13 cases, charged residues with 12 cases, \(h\)-bonding and small amino acids both had 8 cases, tiny with 6, neutral and hydrophobic with 4, aliphatic with 4, hydrophilic with 2, and molecular weight with 2.

Additionally, we analyzed the descriptor values for physicochemical properties and amino acid composition between classes with a standard \(t\)-test. The \(t\)-test results of the AAC features between the 12 classes are listed in Additional file 2. It shows that Ala, Cys, Asp, Glu, His, Lys, Met, Asn, Arg, Ser, and Thr vary significantly between the classes. This is particularly interesting as the amino acids which have the highest amounts of statistically significant differences between classes seem to be involved in important aspects of Laccases. For example, the top two amino acids are aspartic acid and lysine with significant differences among 51 of the 66 possible class
Table 3 Performance of different descriptors in clustering of Laccases using various SOM dimensions

| SOM Dimensions | AAC  | C of CTD | CT   | DIPEP | MA   | MBA   | Physico-Chemical Properties | Sequence Order | Coupling |
|----------------|------|----------|------|-------|------|-------|-------------------------------|----------------|----------|
| 5x5            | 13   | 11       | 12   | 5     | 8    | 13    | 12                            | 18             |          |
| 5x6            | 13   | 15       | 5    | 4     | 15   | 13    | 22                            | 22             |          |
| 6x5            | 12   | 14       | 8    | 5     | 5    | 16    | 12                            | 24             |          |
| 6x6            | 13   | 15       | 11   | 6     | 5    | 18    | 13                            | 16             |          |
| 7x7            | 9    | 17       | 7    | 6     | 22   | 11    | 11                            | 39             |          |
| 8x8            | 11   | 13       | 9    | 9     | 11   | 14    | 14                            | 45             |          |
| CDVD           | 8.2E-5 | 2.5E-4 | -0.13 | -1.1E-4 | -2.9E-2 | 4.7E-3 | 6.1E-3                     | -1.3E-4        |          |
| DBI            | 0.89 | 0.48     | 0.95 | 0.61  | 0.62 | 0.42  | 0.37                         | 0.12           |          |

SOM: Self Organizing Map, AAC: Amino Acid Composition, DIPEP: Dipeptide Composition, CT: Conjoint Triad, C of CTD: Composition of CTD, MA: Moran Autocorrelation, MBA: Moreau-Broto Autocorrelation, CDVD: intercluster - intracluster variance, DBI: Davies-Bouldin Index.

Figure 2 Co-occurrence matrix for the 12 clusters. The colors indicate the number of times given sequences in the data set occur in the same cluster. Red values represent high co-occurrence and blue low co-occurrence, both of which indicate a low amount of variation between consecutive runs of the clustering program.
comparisons. Aspartic acid plays an important role in many Laccase catalytic domains such as: assisting in substrate channels in basidiomycete Laccases, affecting Laccase activity of C-terminal domains when mutated in bacterial Laccases, and assisting in the exit of protons from the N-terminal domains of bacterial Laccases [43-45]. Lysine can also be found widely in catalytic domains, for example C-terminal lysines have been implicated in the inactivation of heterologously produced Laccases [46]. Aside from function, lysines are also widely used as a cross linking target to bind Laccases to various materials [47-49]. Glutamic acid had the next most significant differences between classes. This was observed in Leu-Glu-Ala motifs which follow the copper ligating histidines and are thought to be related to Laccases with higher redox potentials [50]. Further, Asparagine closely followed with 41 significant differences. Many Laccases are known to contain asparagines which serve as sites for N-linked glycosylation [51]. These sites have been shown to be involved in regulation of Laccase activity through catalytic sites such as the Leu-Met-Asn motif which often replaces the previously mentioned Leu-Glu-Ala motif [50]. N-Glycosylation has also been found to

| SOM Dimensions | Intercluster-Intracluster Variance | Davies-Bouldin Index |
|----------------|-----------------------------------|----------------------|
| 5 x 5          | 0.005325395                      | 0.385778113          |
| 5 x 6          | 0.006355107                      | 0.365723878          |
| 6 x 5          | 0.006241709                      | 0.363547417          |
| 6 x 6          | 0.0061449                        | 0.372815039          |
| 7 x 7          | 0.006477438                      | 0.332075399          |
| 8 x 8          | 0.006438212                      | 0.3249138            |

Table 5 Distribution of Laccases in different identified clusters under each taxa.

| Cluster Number | Bacteria | Fungi | Metazoa | Plants | UniProt Accessions |
|----------------|----------|-------|---------|--------|-------------------|
| cluster-0      | 1        | 23    | 2       | 5      | Q12541 P17489 Q70K3 Q12542 Q941X2 Q0H9V5 Q8X1W2 B5M6A4 Q4V494 Q68LM0 R9W9K4 R5W576 I0A0Q5 Q7Z9S4 Q6E124 Q2V719 G8A542 G8A560 G8A585 I3P603 G8A529 T1U007 B2K22105 S0E898 Q38757 Q8W905 Q2PA1 J1Q6X0 Q1A6C6 |
| cluster-1      | 1        | 34    | 0       | 5      | Q0DH2L3 Q33936 Q8W9MV Q91278 P33644 Q9P864 Q6W968 Q0K199 G0W6M6 R4J8RR D3K411 Q51800 Q50XG5 B4Y313 Q6U6K8 F2V7P7 C1JCL7 Q9P889 C0JRG9 C0JRG8 Q4GIIH4 Q7Z9S2 H9CA42 D4AIA6 Q69FX1 Q69FW7 Q69FW8 Q1EPM3 G8A545 D2K2601 D2K2701 Q716A3 Q716A2 Q308C0 B0JDP9 C0P5Q0 Q4V126 |
| cluster-2      | 0        | 4     | 0       | 24     | Q10ND7 Q8W3A1 Q0Q1U1 P78722 Q5N0X2 Q02081 Q951Y8 Q9LD22 B3T6L6 C0JRG6 Q5S2A8 Q2PAJO Q2PAIQ Q9AUH9 Q9AUI3 Q9AUI5 Q9AUI0 B9H7K5 Q95FC9 Q9ZQW5 K4LP19 Q4PCQ7 K4N7Z7 K4P1P7 K4NN22 B1F6G7 F4M6L7 M5AN30 |
| cluster-3      | 0        | 0     | 0       | 8      | Q01801 Q9AUJ2 Q24041 Q24040 Q24043 Q24042 Q9ZCQ2 W5AS95 |
| cluster-4      | 3        | 2     | 18      | 2      | P56193 Q2O725 Q8V920 Q8B719 Q99U58 Q49374 Q4U3X4 Q9X011 Q8WPD1 M4Q2Q6 Q8BY11 M9FL93 ASYV0Q Q8BY12 USEN52 D5S7RE D5S7RE1 F6UQ11 U3C6Q1 G9X9R5 F8V189 F8V190 Q2JBQ8 Q9PPB2 |
| cluster-5      | 66       | 0     | 66      | 1      | 076 Q09266 Q02497 Q12379 B8YQ79 Q9U9Q2 11S143 U3M785 Q1W6B1 D3Y58 C9W9P8 Q59944 Q4UX34 Q9X111 Q8WPD1 M4Q2Q6 Q8BY11 M9FL93 ASYV0Q Q8BY12 USEN52 D5S7RE D5S7RE1 F6UQ11 U3C6Q1 G9X9R5 F8V189 F8V190 Q2JBQ8 Q9PPB2 |
| cluster-6      | 26       | 0     | 26      | 1      | 270 Q12729 Q96W9M Q12717 Q8NO5 Q3N7G3 Q5O656 Q9W0T58 R9U111 Q9W9K7 11W1V7 11W1V8 Q7Z9S6 Q577Q0 E7B9Q8 E7B9Q9 H9Z2R2 Q6RYA4 C5XN85 Q2A0D1 Q6H777 C1K7D6 C1K7D7 C1K7D8 USXR0 B5G552 B2CMA7 |
| cluster-7      | 4         | 29    | 10      | 43     | 09Q9YQ2 Q41357 Q536U4 Q2R0L2 P60H11 Q2R0L0 Q20Q20 Q9L6F01 F6N9E7 R9W1U1 G95270 Q5Q766 Q125656 Q85Y10 A7XQ99 B4L6L6 B58R55 D4A459 Q58U13 Q58UI2 10VA6 E9HH10 Q9HH11 Q941Q1 Q94904 A7QQ52 B5EB2D D5S6R9 D5S6RE D5S6RE1 D5EB12 D5EB17 D5EB14 D5EB15 D5EB16 D5EB15 D5EB161 U56A51 Q4Z2Q4 M56EQ3 Q56D52 B5M132 |
| cluster-8      | 1         | 0     | 1       | 11     | 296 Q9Y9F9 Q5LM53 Q2O729 Q9Z9P2 Q5N7B3 G0W9K9 G0WX65 Q58453 Q92PF7 Q2LD62 B9HH7V G9Z9Q4 |
| cluster-9      | 1         | 0     | 4       | 6      | Q95R40 Q6S7E9 Q8W0V6 Q2PAJ2 P93366 Q723W2 |
| cluster-10     | 0         | 15    | 0       | 12     | Q61D8 Q9FL85 Q1PD66 Q9FJD5 Q56Y70 C08434 Q9AUI1 Q9AUI6 Q9AUI4 I3W7E6 K4PCR3 K4P1M3 |
provide protection against proteolysis [51,52]. Other types of glycosylation such as O-linked glycosylation are also major factors, so it comes as no surprise that both serine and threonine are high on the list [52].

In our other statistical analysis, the t-test results of the important physicochemical properties as identified in Figure 3 are listed in Additional file 3. It shows that all the physicochemical properties identified to be important in discriminating between classes are also significant. We believe since the generated classes contain many significant differences in physicochemical properties and the amino acids with high numbers of significant differences also strongly related to Laccase function, these classes may indeed represent different functional classes of Laccases. To investigate the classes further, a cladogram was constructed from a multiple sequence alignment using the sequences used for clustering. We then mapped our clusters and the classes from LccED to the cladogram Figures 5a and 5b respectively [15,16]. Despite many of the clusters being dominated by a single taxa, when mapped to the cladogram they are widely dispersed throughout the taxonomic regions of the cladogram. This contrasts sharply with the LccED classes which largely only follow taxonomy. Many of the neighbors in the tree are composed of enzymes from the same or similar organisms; these could indicate Laccases of different function from within an organism.

Classification framework
To allow for the classification of newly discovered Laccases and Laccases with no experimental evidence, a Support Vector Machine-based classification system was developed. To accomplish this, 90% of the Laccase data collected was used for 5-fold cross-validation and the remaining 10% kept aside for independent testing. As physicochemical descriptors were used to build the classes, physicochemical properties were also used to develop the SVM classifiers. The developed models were further used to classify sequences annotated as Laccases with “homology” or “predicted” level evidence in the UniprotKB database.

5-fold cross-validation
The performance of the classifier in 5-fold cross-validation for all classes is reported in Table 6. The results show that the model achieves the overall accuracy of 99.03%, MCC of 0.9367, precision of 94.20%, sensitivity of 94.20% and specificity of 99.47%. The overall specificity is extremely high indicating a low rate of misclassified sequences. Considering the classes individually, the highest metrics achieved were MCC 1.0 and accuracy, specificity, and sensitivity of 100%. The lowest performance was accuracy of 98.98%, MCC of 0.7252, sensitivity of 80% and specificity of 99.31%.

Independent testing
Performance results on an independent test data are listed in Table 7. The model also provides higher performance with an overall accuracy of 98.98%, error rate of 1.02%, MCC 0.8678, precision of 87.88%, sensitivity of 87.88% and specificity of 98.90%. It should be noted that the MCC of cluster-3 was zero. However, this class contains only one sequence and performs well in cross validation, so we believe it is still credible.

Confusion matrix
Confusion matrices were made in order to better understand which classes are more similar to one another. The confusion matrix for the independent test set is shown in Table 8. According to the confusion matrix, it appears that few proteins in classes 1, 2, 8, 10 and 11 are predicted
as other classes. The results in confusion matrix show the efficiency of the developed classifier in predicting the samples correctly.

ROC curves

ROC curves are important to consider for prediction systems to give an accurate measure of credibility and or reliability. Each point on the curve is based on the confidence score thresholds of a single classifier. Each ROC curves compute the area under the curve (AUC). This indicates the probability of positive sequence having a higher value than a negative sequence when two are selected at random [53]. The more shift of the curve toward left, the more accurate the predictor. We calculated the ROC
curves for each class for 5-fold cross-validation and independent set testing separately. The ROC curve for 5-fold cross-validation is shown in Figure 6 and for independent set in Figure 7. Each contains a line for each class in the prediction system as well as a line showing the average performance of all classes. All classes show excellent performance with lines very close to the left side of the chart, indicating a high rate of correct predictions from these models. Indeed, the overall area under the curve rounds up to 1.00 showing the reliability of our classifier.

To investigate the role of domains in the functional variation between different classes, we generated domain maps for the sequences in each class. Eleven different types of domains were found to exist within the dataset. The frequently occurring domains are PF07732, PF00394, PF07731 and PF02578. The first three are mostly found in plants and fungi and the domain PF02578 found mostly in bacterial or mammalian origins. Class 4 contained a couple of polyphenol oxidase domains and tyrosinase domains.

### Table 6 Performance of physicochemical descriptor classifier in a 5-fold cross-validation test.

| Cluster # | ACC (%) | ERR (%) | MCC    | PER (%) | SEN (%) | SPE (%) | FN | FP | TP | TN |
|-----------|---------|---------|--------|---------|---------|---------|----|----|----|----|
| cluster-0 | 98.63   | 0.35    | 0.91918| 96.15   | 89.29   | 99.62   | 3  | 1  | 25 | 264|
| cluster-1 | 98.98   | 0.69    | 0.72522| 66.67   | 80      | 99.31   | 1  | 2  | 4  | 286|
| cluster-2 | 98.98   | 0.01    | 0.8483 | 100     | 72.73   | 100     | 3  | 0  | 8  | 282|
| cluster-3 | 98.63   | 0.35    | 0.9356 | 97.06   | 91.67   | 99.61   | 3  | 1  | 33 | 256|
| cluster-4 | 98.98   | 1.02    | 0.93961| 89.29   | 98.88   | 0       | 3  | 25 | 265|
| cluster-5 | 100     | 0       | 1      | 100     | 100     | 100     | 0  | 0  | 7  | 286|
| cluster-6 | 100     | 0       | 1      | 100     | 100     | 100     | 0  | 0  | 23 | 270|
| cluster-7 | 98.98   | 0.69    | 0.96881| 96.72   | 98.33   | 99.14   | 1  | 2  | 59 | 231|
| cluster-8 | 99.66   | 0.34    | 0.97797| 96      | 100     | 99.63   | 0  | 1  | 24 | 268|
| cluster-9 | 98.98   | 0.35    | 0.93086| 95.65   | 91.67   | 99.63   | 2  | 1  | 22 | 268|
| cluster-10| 100     | 0       | 1      | 100     | 100     | 100     | 0  | 0  | 39 | 254|
| cluster-11| 99.32   | 0.01    | 0.90134| 80.82   | 100     | 99.47   | 2  | 0  | 9  | 282|
| Overall   | 99.03   | 0.01    | 0.9367 | 94.20   | 94.20   | 99.47   | -  | -  | -  | -  |

ACC: accuracy, ERR: error, MCC: Matthews Correlation Coefficient, PER: Precision, SEN: Sensitivity, SPE: Specificity, FN: False Negatives, FP: False Positives, TP: True Positives, TN: True Negatives.
Table 7 Performance of physicochemical descriptor classifier on an independent test data.

| Cluster # | ACC | ERR  | MCC  | PER | SEN | SPE  | FN  | FP  | TP  | TN  |
|-----------|-----|------|------|-----|-----|------|-----|-----|-----|-----|
| cluster-0 | 100 | 0    | 1    | 100 | 100 | 100  | 0   | 0   | 3   | 30  |
| cluster-1 | 97  | 0.03 | 0.851| 100 | 75  | 100  | 0   | 3   | 29  |
| cluster-2 | 100 | 0    | 1    | 100 | 100 | 100  | 0   | 0   | 3   | 30  |
| cluster-3 | 97  | 0.03 | 0    | 100 | 100 | 0    | 1   | 0   | 0   | 32  |
| cluster-4 | 100 | 0    | 1    | 100 | 100 | 0    | 0   | 0   | 2   | 31  |
| cluster-5 | 97  | 0.03 | 0.909| 100 | 86  | 100  | 1   | 0   | 6   | 26  |
| cluster-6 | 100 | 0    | 1    | 100 | 100 | 100  | 0   | 0   | 3   | 30  |
| cluster-7 | 100 | 0    | 1    | 100 | 100 | 100  | 0   | 0   | 3   | 30  |
| cluster-8 | 97  | 0.03 | 0.851| 100 | 75  | 100  | 1   | 0   | 3   | 29  |
| cluster-9 | 100 | 0    | 1    | 100 | 100 | 100  | 0   | 0   | 1   | 32  |
| cluster-10| 97  | 3.03 | 0.696| 50  | 100 | 97   | 0   | 1   | 1   | 31  |
| cluster-11| 97  | 3.03 | 0.696| 50  | 100 | 97   | 0   | 1   | 1   | 31  |

Overall 97.98 1.02 0.8678 87.88 87.88 98.90 - - - -

ACC: accuracy, ERR: error, MCC: Matthews Correlation Coefficient, PER: Precision, SEN: Sensitivity, SPE: Specificity, FN: False Negatives, FP: False Positives, TP: True Positives, TN: True Negatives.

Table 8 Confusion matrix for the predicted Laccase subtypes from 5-fold cross-validation testing.

| Predicted class | cl-0 (28) | cl-1 (36) | cl-2 (25) | cl-3 (7) | cl-4 (23) | cl-5 (60) | cl-6 (24) | cl-7 (24) | cl-8 (39) | cl-9 (11) | cl-10 (5) | cl-11 (11) |
|-----------------|-----------|-----------|-----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| True Class      |           |           |           |          |           |           |           |           |           |           |           |           |
| cl-0 (28)       | 25        | 0         | 3         | 0        | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| cl-1 (36)       | 1         | 33        | 0         | 0        | 1         | 0         | 0         | 1         | 0         | 0         | 0         | 0         |
| cl-2 (25)       | 0         | 0         | 25        | 0        | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| cl-3 (7)        | 0         | 0         | 0         | 7        | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| cl-4 (23)       | 0         | 0         | 0         | 0        | 23        | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| cl-5 (60)       | 0         | 0         | 0         | 0        | 0         | 58        | 1         | 1         | 0         | 0         | 0         | 0         |
| cl-6 (24)       | 0         | 0         | 0         | 0        | 0         | 24        | 0         | 0         | 0         | 0         | 0         | 0         |
| cl-7 (24)       | 0         | 1         | 2         | 0        | 0         | 0         | 21        | 0         | 0         | 0         | 0         | 0         |
| cl-8 (39)       | 0         | 0         | 0         | 0        | 0         | 0         | 0         | 39        | 0         | 0         | 0         | 0         |
| cl-9 (11)       | 0         | 1         | 0         | 0        | 0         | 0         | 0         | 0         | 9         | 1         | 0         | 0         |
| cl-10 (5)       | 0         | 0         | 0         | 0        | 0         | 0         | 0         | 0         | 1         | 0         | 4         | 0         |
| cl-11 (11)      | 0         | 0         | 1         | 0        | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 8         |

Values in parentheses represent total number of sequences present in each subtype.

Figure 6 ROC curves of different classes in a 5-fold cross-validation test. Area under curve for each enzyme subtype is also depicted.
The domain maps generated for all the classes are shown in Figure S1 in the supplementary material. The majority of the domain maps were highly similar within and between classes with respect to domains present. However, there were some differences between the positions of the domains. We believe that these differences in the relationships between the positions of the domains could also account for functional differences.

Classification of Laccase homologs from UniProtKB

The efficiency of our prediction approach is tested by identifying the Laccases in UniprotKB with homology or predicted level evidence. Out of the 1656 sequences retrieved, 1587 were predicted to one of the 12 classes and reported in Table 9. These annotations could be a good resource to the scientific community working in these areas.

Web tool for classification of Laccases

We have developed a web resource for the classification of the Laccase subtypes by implementing the machine learning models. It will be very useful to the researchers to characterize the newly found Laccase sequences. The tool can be found at http://lacsubpred.bioinfo.ucr.edu/.

Table 9 Classification of UniProt sequences with our method for those annotated as Laccases with homology or predicted level evidence in UniProt KB database.

| Cluster | Accession Numbers |
|---------|-------------------|
| Cluster-0 | Q00HLS Q66554 Q87AR8 P67256 P67257 P33663 Q9PET8 P45496 M5C3X0 Q768Q6 Q8GB87 |
| Cluster-1 | Q768Q6 Q8GB87 |
| Cluster-2 | G7L18S G7L2E0 G7L2E1 B9G57 B9G52 B9HCK0 B9HJC9 |

Figure 7 ROC curves of different classes generated from the best model (physicochemical model) in an independent test. Area under curve for each enzyme subtype is also depicted.
Table 9 Classification of UniProt sequences with our method for those annotated as Laccases with homology or pre-defined level evidence in UniProt KB database.

| Cluster-1 | Cluster-2 | Cluster-3 | Cluster-4 | Cluster-5 | Cluster-6 | Cluster-7 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| A0A011114770 | A0A011114780 | A0A011114790 | A0A011114800 | A0A011114810 | A0A011114820 | A0A011114830 |
| ... | ... | ... | ... | ... | ... | ... |

(Suppl 11):S15

Weirick et al (2014, 15(Suppl)11)  
http://www.biomedcentral.com/1471-2105/15/S11/S15

Page 16 of 19

Weirick et al (2014, 15(Suppl)11)  
http://www.biomedcentral.com/1471-2105/15/S11/S15

Page 16 of 19
Table 9 Classification of UniProt sequences with our method for those annotated as Laccases with homology or predicted level evidence in UniProt KB database. (Continued)

| Cluster-8 | 95 | Q03966 Q12719 Q99055 Q99056 B2W89 1R6Q8 1R6Q9 F8QG75 9KMF9H 9KMEL6 A1Y8E8 M7SP00 M7SC50 H6BQP7 Q50H77 M7SPF1 R7SV61 L7IR77 A2Q6G2 A2Q729 G2W2W4 H4NQ4 V Q13546 L7OR61 K1Y6B2 E1C6GD G7X777 9KMF2J 9NL4G9 G3XU36 G3XCY7 Q6IWB69 9KHHQ9 KHRW5 R8QT5 M7TM62 M7JTC4 N M7DT67 R8PSV5 9G6H11 E1F3F9 F8PB06 81NEY0 G2XU3R E9EY11 G2XLY6 G2XZ63 C9SN11 QSMBH0 A2QXN9 N4UQ06 V4N59 N4VL55 S8ED94 50E275 R8BAAM6 Q75W61 Q75AC0 A1YF0 B6QSM9 A8Y759 Q08AB9 A8N477 H0E3T7 G3FGX5 H0E6M2 N1RP79 B8MH08 G0FX2G F0XJCS Q3KRP1 V2Y238 V2X623 V2XTS V2WXX7 V2X4B3 E7A541 A1Y1E9 M2RE87 L2GCR5 L2FPB4 M2QY3C M2QY2B A5RJG9 K5F5E6 E6Z5D5 B5GS51 F55525 F4RZ56 F4RZ43 F4R9G7 Q5W0J5 Q5K7H5 S3B8V4 |
| Cluster-9 | 61 | Q2RBK2 QOQYS3 G5EGX4 J3PI2O B2VSS7 M5CEU4 R1GKT9 R1EPH7 K9FSV2 F8PY37 9KMF32 9KML70 Q50H78 R7T0U4 R7T27L L7JQ96 A2Q861 L7IKL2 G7XZ1 M7S5R5 M7UW45 B0C729 B0D550 B0DZ27 B0E660 F8WN99 G2XH42 G2Y1J7 81MB93 K9GTR0 N1Q4F7 H0E720 F0X669 B5GS53 Q8TTF2 G5SS55 V2X3Y4 Q13421 C3HL41 G2QZ3S S7P2A1 G2QZ2S S7QC29 G2QZ18 G2QZ14 G2QZ13 5SV1M4 V2RH51 S1RH27 S11G86 T5F168 V3A3E6 T5G4Q5 T5R7F3 T5E666 V5Y8R9 T491M3 T5K4IC |

have also provided the codes used to develop the clustering and classification approach as an open source package available at https://github.com/tweirick/SubClPred.

**Conclusion**

In this work, we present a systematic computational approach to identify Laccase subtypes. First, a novel clustering method is developed to group the Laccase subtypes. Then a classification method is developed based on machine learning approach to generalize the functions of Laccases in each class. These identified groups can be a useful learning approach to generalize the functions of Laccases. A web tool is developed from this study to find the clusters and identify the Laccases in any given sample. The tool is freely available at http://lacsubpred.bioinfo.ucr.edu/.

**Additional material**

Additional file 2: P-values designating the statistical significance of one cluster over the other based on amino acid composition differences; values calculated using the standard t-test.

Additional file 3: P-values designating the statistical significance of one cluster over the other based on protein physicochemical property differences; values calculated using the standard t-test.

Additional file 1: Domain maps for each of the Laccase subtypes cluster generated using doMosaics (http://www.domosaics.net/).

**List of abbreviations used**

ROC, Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient; SOM, Self-Organized Maps; SVM, Support Vector Machines; DB, Davies-Bouldin Index; AAC, Amino Acid Composition; CT, Conjunct Triangle; CTD, Composition-Transition-Distribution; DIPEP, Dipetide Composition; MA, Moran Autocorrelation; MBA, Moreau-Broto Autocorrelation.

**Competing interests**

The authors declare that they have no competing financial interests.

**Authors’ contributions**

TW collected the datasets related to Laccases from public repositories, wrote codes for clustering, developed algorithms and models, performed the calculations, figures and tables, and wrote the draft manuscript. SSS helped in model development, data analysis and tool building. RM helped in biological analysis and in editing the manuscript. RK conceived the study, participated in its design and coordination, and edited the final manuscript. All authors read and approved the final manuscript.

**Declaration**

Funding for the publication of this article has come from the ‘start-up’ funds provided to RK, account A01949-19900-44-CPK1, UCR.

This article has been published as part of BMC Bioinformatics Volume 15 Supplement 11, 2014: Proceedings of the 11th Annual MCBIOS Conference. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S11.

**Authors’ details**

1National Institute for Microbial Forensics & Food and Agricultural Biosecurity (NIMFFAB), Oklahoma State University, Stillwater, Oklahoma, 74074, USA.

**Availability**

LaSubPred, the web resource developed from this study, is freely available at http://lacsubpred.bioinfo.ucr.edu/.

**Additional file 2:** P-values designating the statistical significance of one cluster over the other based on amino acid composition differences; values calculated using the standard t-test.

**Additional file 3:** P-values designating the statistical significance of one cluster over the other based on protein physicochemical property differences; values calculated using the standard t-test.

**Additional file 1:** Domain maps for each of the Laccase subtypes cluster generated using doMosaics (http://www.domosaics.net/).
water in CotA Laccase: assistance during the proton-transfer mechanism. *Acta Crystallographica Section D: Biological Crystallography* 2012, 68(2):186-193.

46. Bleve G, Lezzi C, Spagnolo S, Tasco G, Tufariello M, Casadio R, Mita G, Rampino P, Greco F: Role of the C-terminus of Pleurotus eryngii Ery4 Laccase in determining enzyme structure, catalytic properties and stability. *Protein Engineering Design and Selection* 2013, 26(1):1-13.

47. Yamauchi H, Miyazaki M, Asanomi Y, Maeda H: Poly-lysine supported cross-linked enzyme aggregates with efficient enzymatic activity and high operational stability. *Catalysis Science & Technology* 2011, 1(7):1256-1261.

48. Mikolasch A, Hahn V, Manda K, Pump J, Illas N, Gördes D, Lalk M, Salazar MG, Hammer E, Jülich W-D: Laccase-catalyzed cross-linking of amino acids and peptides with dihydroxylated aromatic compounds. *Amino Acids* 2010, 39(3):671-683.

49. Kurniawan RA, Aulanni M, Sheeh F-K, Chu PP-J: Carbon Nanotube Covalently Attached Laccase Biocathode for Biofuel Cell. *The Journal of Pure and Applied Chemistry Research* 2013, 2(2):79-88.

50. Piontek K, Antorini M, Choinowski T: Crystal Structure of a Laccase from the Fungus *Trametes versicolor* at 1.90-Å Resolution Containing a Full Complement of Coppers. *Journal of Biological Chemistry* 2002, 277(40):37663-37669.

51. Yoshitake A, Koyama Y, Nakamura M, Irmura Y, Kawai S, Morohoshi N: N-linked carbohydrate chains protect Laccase III from proteolysis in *Coriolus versicolor*. *Journal of General Microbiology* 1993, 139(1):179-185.

52. Perry CR, Matcham SE, Wood DA, Thurston CF: The structure of Laccase protein and its synthesis by the commercial mushroom *Agaricus bisporus*. *Journal of General Microbiology* 1993, 139(1):171-178.

53. Lemeshow S, Hosmer D: *Applied Logistic Regression* (Wiley Series in Probability and Statistics: Wiley-Interscience. 2000).