Journal of AI and Data Mining
Vol 3, No 2, 2015, 209-215. 10.5829/idosi.JAIDM.2015.03.02.10

Propensity based classification: Dehalogenase and non-dehalogenase enzymes

R. Satpathy[1*], V. B. Konkimalla[2] and J.Ratha[1]

1. School of Life Sciences, Sambalpur University, Burla, Sambalpur, India.
2. Department of Biological Sciences, National Institute of Science Education & Research (NISER), Bhubaneswar, India.

Received 15 March 2015; Accepted 8 August 2015
**Corresponding author: rnsatpathy@gmail.com (R.Satpathy).*

Abstract
The present work was designed to classify and differentiate between the dehalogenase enzyme and non–dehalogenases (other hydrolases) by taking the amino acid propensity at the core, surface and both the parts. The data sets were made on an individual basis by selecting the 3D structures of protein available in the PDB (Protein Data Bank). The prediction of the core amino acids were predicted by IPFP tool and their structural propensity calculation was performed by an in-house built software, Propensity Calculator which is available online. All datasets were finally grouped into two categories, namely dehalogenase and non-dehalogenase using Naïve Bayes, J-48, Random forest, K-means clustering, and SMO classification algorithm. By making the comparison of various classification methods, the proposed tree method (Random forest) performs well with a classification accuracy of 98.88 % (maximum) for the core propensity data set. Therefore, we proposed that, the core amino acid propensity could be approved as a novel potential descriptor for the classification of enzymes.

Keywords: *Core Propensity; Classification Algorithm; Random Forest; Protein Data Bank; Dehalogenase and Non- dehalogenases*

1. Introduction

Microbial dehalogenases are unique enzymes produced by a microbe that dehalogenates halogenated substances (toxic) by breaking the C-Cl bonds, thus, making a biotechnologically important enzyme group (Arand et al, 1994; Kovalchuk & d'Itri, 2004). In general, these enzymes are classified as hydrolases along with other hydrolytic enzymes that catalyse hydrolytic bond cleavage for C-N, C-P, C-O bonds (Koonin & Tatusov, 1994). The mechanism of bond cleavage is quite similar across all hydrolases regardless of the binding atoms. Apart from the classical enzyme classification techniques, data mining methods are helpful in analyzing large sets of sequences and information retrieval for these enzymes (Borro et al, 2006; Nasibov & Kandemir-Cavas, 2009). Various classification methods are being applied in different areas, as there is no unique classifier that can best classify them due to the presence of various data. For data mining problem especially in case of proteins, classifiers can help achieve full accuracy by considering suitable features such as enzyme physicochemical and structural properties (Banerjee et al, 2010; King et al, 2001). To address this enzyme grouping problem, many data mining approaches were implemented earlier. The most common method followed is clustering enzymes based on their sequence and structural similarity (Fayech et al, 2009). However, these approaches sometimes fail especially incase of proteins (enzymes) where many of them can perform the same function in spite of dissimilarity in their sequence and/or structure. Another significant task for researchers in bioinformatics is to classify these proteins into families based on their structural and functional properties, thereby predicting the functions of these new protein sequences (Krishna et al, 2003). Over few years, new computational methods as well as novel protein features have been developed and implemented to expand the knowledge about protein classification. Some of them are global and local structural alignment algorithms that

trace out conformational similarities between proteins indicating functional similarities (Holm & Sander, 1993; May & Johnson, 1994; Taylor, 1999). More practically, computational methods that utilize three-dimensional (3D) structures of proteins are more efficient compared to the sequence-based function prediction. This is due to the fact that protein structures are more conserved than sequences during evolution. Various categories of structural information of proteins such as folding pattern, amino acids forming active sites along with their conformation and interactions pattern with ligands have been used for data mining purposes (Ivanciuc et al, 2002; Oldfield, 2002). Notwithstanding, the availability of high-resolution structural data of target proteins or their homologs, however, remains the major limitation of this methodology. For performing an accurate and efficient classification, a robust strategy in data mining technology is essential and needs a specific dataset that ultimately classifies and improves predictions for unclassified data. Several typical types of classification techniques are available in the literature such as Decision Trees, Naïve Bayesian methods, Sequential Minimal Optimization (SMO), etc. (Delen et al, 2005; Ramesh & Ramar, 2011; Wisaeng, 2013; Wei X et al, 2014). For various data mining purposes, Weka is used as a good simulating software that integrates several data mining features as data pre-processing tools, learning algorithms and performance evaluation methods. Additionally, the graphical user interfaces (GUI) provide an excellent environment for inferring classification details (Amini et al, 2013; Frank et al, 2004).

The primary goal of this work is to further classify the dehalogenase class of an enzyme from other hydrolases based on its structural amino acid propensities. For this purpose, a protocol is initially developed to find the amino acids present in the core/surface of a protein and their propensities were calculated. Further, various other methods are employed to separate the two groups of enzymes to examine the different classifiers using Weka tool in order to know which classification algorithms perform better by analyzing different parameters.

2. Materials and methods

All the works were performed using PC having OS Windows 7, Dual Core, RAM 2 GB, 250 GB HD, and 1.76GZ processing speed.

2.1. Data set preparation

Protein Data Bank (PDB) is the primary repository for experimentally determined 3D protein structures. The protein structures available for dehalogenase was retrieved by querying the PDB for structures that are single chained and less than 400 amino acids. The search yielded 90 protein structures, where 45 are dehalogenase and rest 45 are non-dehalogenase (other hydrolases) whose structures are determined using X-ray crystallography (Berman et *al*, 2000).

2.2. Calculation of core and surface residues and propensities

Calculations for core and surface amino acids in a given PDB file were performed using IPFP tool, available on line (Satpathy et *al*, 2014). This tool first computes the accessible surface area of all the residues by calculating the atomic accessible surface defined by a rolling probe of given size around a van der Waals surface explained by Lee and Richards (Hubbard & Thornton, 1993; Lee & Richards, 1971). Here, a probe size of 1.4 Å was chosen. Further, from the accessible surface area of all amino acids, the core amino acids are predicted; since, those amino acids having non polar accessible surface area is zero and the rest of the amino acids are predicted to be on the surface. The following equations are used to calculate the propensity from individual PDB file. The propensity was calculated automatically from a Matlab script that prepares an input file for the classification. The missing value for the amino acid propensity was assigned zero in the input files. The script in the form of a windows executable is freely available online. Here, the propensity was computed by providing the core amino acids and the total amino acid information as input. The Propensity calculator tool computes the surface exposed propensity (SP) and core propensity (CP) as presented as below (1 and 2) and described by Reddy et al. (1998) and Shambhu Malleshappa Gowder, et al., (2014).

$$\mathbf{SP} = \frac{\mathbf{N_{Soli}} / \mathbf{T_{Soli}}}{\mathbf{T_i} / \mathbf{Total}} \quad (1)$$

$$\mathbf{CP} = \frac{\mathbf{N_{Buri}} / \mathbf{T_{Buri}}}{\mathbf{T_i} / \mathbf{Total}} \quad (2)$$

In (1), N_{Soli} indicates number of residue i, that are solvent, exposed. T_{Soli} indicates the total number of specific residues that is solvent exposed.T_i indicates about the total number of i residues present in the protein. Total is for the total number of residues present in the protein. Similarly, in (2), N_{Buri} indicates, number of residue i , that are buried in the core region. T_{Buri} indicates the total number of specific residues that are buried. T_i indicates about the total number of i residues

present in the protein. Total is for the total number of residues present in the protein.

2.3. Data mining approaches by utilizing propensity feature

The entire propensity computed data were divided into three parts; core alone, surface alone and a combination of two for the complete data set. The classification experiments were conducted in WEKA 3.6.10 environment. Weka is a java-based open-source collection of machine learning algorithms developed at the University of Waikato, New Zealand. For classification purposes of the current data sets, following classification algorithms were used (Mark Hall et *a*l, 2009).

2.4. Naïve bayes classifier

Bayesian networks represent a probabilistic approach and are potentially used for classification purpose. A naïve bayes classifier computes the conditional probability of a class (C_i) as $P(C_i/X)$ by assuming that all the attributes are conditionally independent.

$$P(C_i/X) = \frac{P(X/C_i)P(C_i)}{P(X)} \quad (3)$$

The main advantage of using this classifier is that, they are probabilistic models; hence, it can perform better even there is presence of any noise and missing value in the data also if the sample size is small (De Ferrari & Aitken, 2006).

2.5. J48

J48 Decision tree classifier algorithm needs to create a decision tree based on attribute values in the available training data. Basically, a decision tree is a flow chart-like tree structure in which the topmost node in a tree is called a root node, each internal node denotes a test on an attribute, while each branch represents the outcome of the test and every terminal node (leaf node) holds a class label (Kotsiantis 2007). J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in the tree. Given a set T of total instances (training set), the following steps are used to construct the tree structure.

1. Select a test based on a single attribute with at least two or greater possible outcomes.
2. Then consider this test as a root node of the tree with one branch of each outcome of the test.
3. Partitioning of T into corresponding T1, T2, T3 ... Tn, according to the result for respective cases, and the same may be applied in recursive way to each sub node.

2.6. Random forest

Random Forest (RF) is a method of classification, which is based on the gathering of a large number of decision trees. More precisely, it is a combination of decision trees constructed from a training data set, which is validated to generate a prediction of response from the given predictors for future observations. The basic step of the algorithm is as follows:

Sample data → training of data → feature selection → splitting of data by best predictor (growing of trees) → estimate error → Random forest (Collection of all trees)

Usually this method combines tree predictors such that each tree depends on the values of a random vector with the same distribution for all trees in the forest. The generalization error for forests behaves as limited functions as the number of trees in the forest becomes large. The algorithm works iteratively until the specific numbers of trees are obtained (Yao *et al*, 2013).

2.7. K-means clustering

K-means is one of the simplest and oldest unsupervised learning algorithms that solve the well known clustering problem (Jain 2010). The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) to be fixed first. The main idea is to define k centroids, one class for each cluster. The basic step of k-means clustering is given below:

1. Determine the number of cluster (k) that represents centroid/center of the clusters.
2. Take any random objects as the initial centroids and assign the closest one.
3. After assignment of all objects, re-calculate the position of the centroid.
4. Repeat the step 2 and 3 whenever there is no further movement of the centroid.
5. Separate objects based on number of k.

2.8. SMO

Sequential Minimal Optimization (SMO) is an algorithm for training of Support Vector Machines (SVMs). SVM is a learning machine for two group classification problems that transform the attribute space into multidimensional feature space using a kernel function to separate dataset instances by an optimal hyperplane. As SVM accuracy depends mostly on selection of attributes; hence, a proper attribute selection in a data set increases the performance of the SMO

algorithm (KR 2011). SMO algorithm basically works iteratively for solving the optimization problem by breaking a problem into a series of smallest possible sub-problems which are solved analytically.

2.9. Performance evaluation for the classifiers

The correct classifications were evaluated comparatively. Basically the performance is based on the True positive (TP) that is correctly identified, False positive (FP) that is incorrectly identified, True negative (TN) that is correctly rejected and False negative (FN) related to incorrectly rejected. The best classification methods obtained were re-evaluated by following performance analysis (Table 4, Table 5, and Table 6).

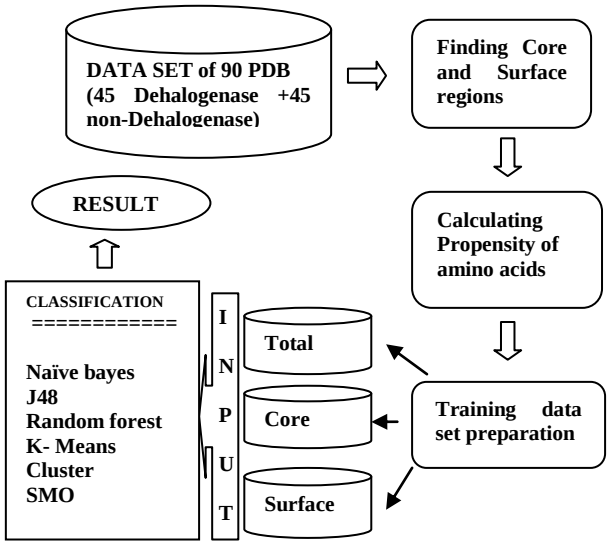

Figure 1. Schematically representation of steps followed in this work.

2.9.1. True positive rate (TPR)

The true positive rate (TPR) is the probability of correctly predicting the positives.

$$TPR = TP/(TP + FN) \quad (4)$$

2.9.2. False positive rate (FPR)

The false positive rate (FPR) is the probability of incorrectly predicting the negatives.

$$FPR = FN/(TP + FN) \quad (5)$$

2.9.3. Precision

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is the proportion of units correctly predicted as faulty.

$$Precision = TP/(TP + FP) \quad (6)$$

2.9.5. F-measure (FM)

FM is a combination of recall and precision. It is also defined as harmonic mean of precision and recall and the Recall measures the proportion of actual positives which are correctly identified

$$F - Measure = (2 \times Recall \times Precision)/(Recall + Precision) \quad (7)$$

2.9.6. ROC area

ROC (Receiver Operating Characteristics) is a tool for comparing different data models. ROC measures the impact of alterations on the probability threshold and tends to forecast the percent of correct classification. Normally, the value for ROC lies between 0 to 1. The probability threshold is the decision point used by the model for categorization. For the two class classification, the default probability threshold is 0.5. When the probability of a prediction is 50% or more, then the model predicts that class and the result are considered as within true positive regions.

3. Results and discussions

In the preliminary approach for classification of dehalogenase and other hydrolase enzymes, all 90 proteins structures (45 for each group) were selected from the Protein Data bank (PDB). All of them belong to hydrolase class; however, dehalogenase cleaves C-Cl bond and other hydrolases cleaves C-N, C-P, ester bond etc. The entry details considered from the PDB IDs are given in table 1.

Table 1. PDBID of considered proteins for data set preparation.

Group	PDB ID	REMARK
Group I	1B6G,1EDB,1EDD,1G42,1G4H,1G5F,1IZ7,1K5P,1K63,1K6E,1NRW, 1PWZ,1QQ6,1QQ7,2BFN,2DHD,2DHE,2GFH,2GO7,2HCF,2NO5,2O2H,3FWH,3G9X,3HLT,3L5K,3QNM,3QUQ,3QUT,3QYP,3R3U,3R3V,3R40,3RK4,3SD7,3SK0,4DCC,4DFD,4EFR, 4EZE, 4F5Z, 4F71, 4IXT, 4IXW, 4IY1	Dehalogenase
Group II	1AID,1B5V,1BA1,1C2K,1HJO,1NL2,1O2T,1QBO,1R54,1U2P,1VL9, 2AOM,2BG2,2BJE,2BO4,2E1E,2G4W,2G7F,2JFR,2OKB,2OSN,2OUD,2RB4,2VND,2VYO,2WTA,2YV5,3BHN,3DAI,3F9O,3H7K,3IR2,3LEZ,3MDQ,3NH4,3QU1,3QU5,3QYP,3RDR,3S1Y,3UQ9,3UWB,4EKD,4HKY,4I69	Non-dehalogenases (Other hydrolases)

For every protein structure in a group, we calculated the core residues and surface residues followed by determination of propensity as explained above in the material and methods section. The residue that is not present in the core or surface is assigned a zero. The core and surface

data set contains 20 attributes (for amino acid) and 90 rows (90x20) that correspond to each PDB. Similarly, (90x40) pattern file was generated for total dataset containing both surface and core propensities for a particular PDB. In this way, we used a one-dimensional representation of 3D protein structures based on calculated regional propensity properties, to train with different classification algorithms for automatic enzyme classification purpose. To perform this, all these data were applied to the classifier as training data separately.

Among all the classifiers, classification of enzyme propensity data sets was classified into two groups, dehalogenase and non-dehalogenase, respectively. For determination of core propensity in all 3 categories of data set, the Random forest algorithm was found suitable in both classification accuracy and execution time (Table 2, Table 3 and Table 4).

Table 2. Classification accuracy of the different algorithms in total data set.

Algorithm	Classification accuracy (%)	Kappa statistic	Time taken in Seconds	Root mean squared error
Bayesian	76.6	0.53	0.05	0.3953
J48	97.7	0.95	0.01	0.1291
Random forest	93.3	0.86	0.01	0.2174
k-means cluster	67.7	0.35	0.01	0.5676
SMO	75.5	0.51	0.15	0.4944

Table 3. Classification accuracy of the different algorithms in surface data set.

Algorithm	Classification accuracy (%)	Kappa statistic	Time taken in Seconds	Root mean squared error
Bayesian	62.2	0.24	0.02	0.5358
J48	95.5	0.91	0.01	0.1939
Random forest	98.8	0.97	0.02	0.2085
k-means cluster	61.1	0.22	0.01	0.6306
SMO	75.5	0.51	0.07	0.4944

Table 4. Classification accuracy of the different algorithms in core data set.

Algorithm	Classification accuracy (%)	Kappa statistic	Time taken in Seconds	Root mean squared error
Bayesian	80	0.6	0.02	0.385
J48	96.6	0.93	0.01	0.1725
Random forest	**98.8**	**0.97**	**0.01**	**0.1856**
k-means cluster	66.6	0.33	0.02	0.57
SMO	65.5	0.31	0.03	0.5869

Compared to Bayesian, clustering and SVM based methods (SMO), the tree based classifier methods yielded good results. *k*-means clustering methods did not perform well to classify the three experimental dataset compared to other methods. During calculation of various statistical parameters *Kappa* static value of 0.97 indicates strong statistical dependence from Random forest algorithm (bold in Table 7).

Based on the *Kappa* statistics criteria, the accuracy of this classification system is substantial. In case of performance of Random forest algorithm, there is considerably high value of TPR and low value of FPR indicating consistency of the algorithm. Similarly, the Random forest based classification results also resulted in excellent value for Recall, F-measure and ROC area for all type of data set (Table 3). This indicates that Random forest model could very advantageously applied for classifying enzymes taking propensity value as training. There are several literatures available in comparative analysis where Random forest algorithm holds good (Chen& Liu, 2005; Hamby & Hirst 2008; Jain & Hirst 2010) for classification purpose.

Table 5. Performance parameters measure (weighted average value) of different algorithms for total dataset.

Algorithm	TPR	FPR	Precession	F-measure	ROC-area
Bayesian	0.767	0.233	0.773	0.765	0.91
J48	0.978	0.022	0.978	0.978	0.998
Random forest	0.933	0.067	0.933	0.933	0.994
k-means cluster	0.678	0.322	0.727	0.659	0.678
SMO	0.756	0.244	0.756	0.756	0.756

Table 6. Performance parameters measure (weighted average value) of different algorithms for surface dataset.

Algorithm	TPR	FPR	Precession	F-measure	ROC-area
Bayesian	0.622	0.378	0.624	0.621	0.593
J48	0.956	0.044	0.956	0.956	0.987
Random forest	0.989	0.011	0.989	0.989	0.997
k-means cluster	0.611	0.389	0.614	0.609	0.611
SMO	0.756	0.244	0.783	0.749	0.756

Table 7. Performance parameters measure (weighted average value) of different algorithms for core dataset.

Algorithm	TPR	FPR	Precession	F-measure	ROC-area
Bayesian	0.8	0.2	0.801	0.8	0.893
J48	0.967	0.033	0.967	0.967	0.989
Random forest	**0.989**	**0.011**	**0.989**	**0.989**	**0.999**
k-means cluster	0.667	0.333	0.708	0.649	0.667
SMO	0.656	0.344	0.658	0.654	0.656

4. Conclusions

Both core and surface residues are responsible for many features of the protein like substrate binding, thermo-stability, protein folding and several other functions. The core region amino acids in case of a protein are basically conserved during evolution. Also during protein folding, the specific arrangement of these residues forms a '*topological pattern*' that provides functional implications to the proteins (enzymes). Hence, the core feature of amino acids is an important feature for protein classification purpose. By analyzing derived results, we conclude that accuracy of Random forest is best in comparison with any other considered algorithms. Here, it is also inferred that the *propensity quantitative feature* at the core region of the protein can be used as one of the excellent and novel descriptors for the classification of enzymes. In future, it is aimed to implement this novel feature to classify other proteins/enzymes.

References

[1] Arand, M., Grant, D. F., Beetham, J. K., Friedberg, T., Oesch, F., & Hammock, B. D. (1994). Sequence similarity of mammalian epoxide hydrolases to the bacterial haloalkane dehalogenase and other related proteins: implication for the potential catalytic mechanism of enzymatic epoxide hydrolysis. FEBS letters, vol.338, no. 3, pp. 251-256.

[2] Kovalchuk, V. I. & d'Itri, J. L. (2004). Catalytic chemistry of chloro-and chlorofluorocarbon dehalogenation: from macroscopic observations to molecular level understanding. Applied Catalysis A: General, vol. 271, no. pp. 13-25.

[3] Koonin, E. V. & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity: application of an iterative approach to database search. Journal of Molecular Biology, vol. 244, no. 1,pp. 125-132.

[4] Borro, L. C., et al. (2006). Predicting enzyme class from protein structure using Bayesian classification. Genetics and molecular research: GMR, vol. 5, no. 1, pp. 193-202.

[5] Nasibov, E. & Kandemir-Cavas, C. (2009). Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. Computational biology and chemistry, vol.33, no. 6, pp. 461-464.

[6] Banerjee, A. K., Sunita, M., Naveen, M., & Murty, U. S. (2010). Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties. Bioinformation,vol. 4, no. 10, pp. 456-462.

[7] King, R. D, Karwath, A., Clare, A. & Dehaspe, L. (2001). The utility of different representations of protein sequence for predicting functional class. Bioinformatics, vol.17, no. 5, pp. 445-454.

[8] Fayech, S., Essoussi, N. & Limam, M. (2009). Fayech S, Essoussi N, Limam M (2009) Partitioning clustering algorithms for protein sequence data sets. BioDataMining, vol.2, no.1, pp. 3.

[9] Krishna, S. S, Majumdar, I. & Grishin, N. V. (2003). Structural classification of zinc fingers Survey and Summary. Nucleic Acids Research, vol. 31, no. 2, pp. 532-550.

[10] Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology, vol. 233, no.1, pp. 123-138.

[11] May, A. C. & Johnson, M. S. (1994). Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. Protein Engineering, vol. 7, no.4, pp.475-485.

[12] Taylor, W. R. (1999). Protein structure comparison using iterated double dynamic programming. Protein Science, vol. 8, no. 3, pp. 654-665.

[13] Ivanciuc, O., Schein, C. H. & Braun, W. (2002). Data mining of sequences and 3D structures of allergenic proteins. Bioinformatics, vol. 18, no.10, pp. 1358-1364.

[14] Oldfield, T. J. (2002). Data mining the protein data bank: residue interactions. Proteins, vol. 49, no. 4, pp. 510-528.

[15] Delen, D., Walker, G. & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, vol. 34, no.2, pp. 113-127.

[16] Ramesh, V. & Ramar, K. (2011). Classification of agricultural land soils: a data mining approach. Agricultural Journal,vol. 6, no. 3, pp.82-86.

[17] Wisaeng, K. (2013). An Empirical Comparison of Data Mining Techniques in Medical Databases. International Journal of Computer Applications, vol. 77, no. 7, pp. 23-27.

[18] Wei, X., et al. (2014). Identification of biomarkers that distinguish chemical contaminants based on gene expression profiles. BMC genomics, vol 15, no. 1, pp. 248.

[19] Amini, L., et al. (2013). Prediction and Control of Stroke by Data Mining. International journal of preventive medicine, vol. 4, no., pp. S245-S249.

[20] Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, vol. 20, no. 15, pp. 2479-2481.

[21] Berman, H. M., et al. (2000). The protein data bank. Nucleic Acids Research, vol. 28, no. 1, pp. 235-242.

[22] Satpathy, R., Konkimalla, V. S. B., & Ratha, J. (2014). IPFP: An Integrated Software Package for Automated Protein Feature Prediction. International Journal of Applied Research on Information Technology and Computing, vol.5, no. 3, pp.223-227.

[23] Hubbard, S. J. & Thornton, J. M. (1993). Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London, 2(1).

[24] Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility.Journal of Molecular Biology, vol.55, no. 3, pp. 379-IN4.

[25] Reddy, B. V., Datta, S. & Tiwari, S. (1998). Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. Protein Engineering, vol. 11, no. 12, pp. 1137-1145.

[26] Malleshappa Gowder, S., Chatterjee, J., Chaudhuri, T. & Paul, K. (2014) Prediction and Analysis of Surface Hydrophobic Residues in Tertiary Structure of Proteins, ScientificWorld Journal, vol. 2014, pp. 1-7.

[27] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, vol. 11, no.1, pp. 10-18.

[28] De, Ferrari, L., & Aitken, S. (2006). Mining housekeeping genes with a Naive Bayes classifier. BMC Genomics, vol. 7, no. 1, pp. 277.

[29] Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. Informatica, vol. 3, no. 3, pp. 249-268.

[30] Yao, D., Yang, J. & Zhan, X. (2013). An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis. Open Electrical & Electronic Engineering Journa, vol.7, no.1, pp. 62-70.

[31] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666.

[32] Seeja, K. R. (2011). Microarray Data Classification Using Support Vector Machine. Journal of Biometrics and Bioinformatics, vol. 5, no. 1, pp. 10-15.

[33] Chen, X. W., & Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. Bioinformatics, vol. 21, no. 24, pp. 4394-4400.

[34] Hamby, S. E., & Hirst, J. D. (2008). Prediction of glycosylation sites using random forests. BMC Bioinformatics, vol. 9, no. 1, pp. 500.

[35] Jain, P. & Hirst, J. D. (2010). Automatic structure classification of small proteins using random forest. BMC Bioinformatics, vol. 11, no. 1, pp. 364.

نشریه هوش مصنوعی و داده کاوی

دانشگاه شاهرود

طبقه‌بندی بر اساس میل ترکیبی: آنزیم‌های دهالوژناز و غیر دهالوژناز

R. Satpathy[1,*]، V. B. Konkimalla[2] و J.Ratha1[1]

[1] دانشکده علوم، دانشگاه سامبالپور، بورلا، سامبالپور، هند.

[2] گروه علوم زیستی، موسسه ملی آموزش و پرورش علوم و تحقیقات (NISER)، بوبانسور، هند.

ارسال ۲۰۱۵/۰۳/۱۵؛ پذیرش ۲۰۱۵/۰۸/۰۸

چکیده:

این پژوهش برای طبقه‌بندی و تمایز بین آنزیم دهالوژناز و غیر دهالوژناز با در نظر گرفتن میل ترکیبی اسید آمینه در هسته، سطح و هر دو قسمت طراحی شده است. مجموعه داده به صورت فردی با انتخاب ساختارهای سه‌بعدی از پروتئین موجود در بانک اطلاعات پروتئین (PDB) ساخته شده است. این صورت پیش‌بینی از اسیدهای آمینه هسته‌ای توسط ابزار IPFP پیش‌بینی شده بود و محاسبه میل ترکیبی ساختاری آنها توسط یک نرم‌افزار داخلی برای محاسبه میل ترکیبی که به صورت آنلاین در دسترس است انجام شد. همه مجموعه داده در نهایت با استفاده از بیزین، J-48، جنگل‌های تصادفی، خوشه‌بندی کا-میانگین و الگوریتم دسته‌بندی SMO به دو دسته تقسیم می‌شود که دهالوژناز و غیر دهالوژناز نامیده می‌شوند. با ایجاد مقایسه روش‌های مختلف طبقه‌بندی، روش درخت پیشنهاد شده (جنگل تصادفی) با دقت طبقه‌بندی ۹۸,۸۸٪ (حداکثر) برای مجموعه داده متمایل هسته به خوبی انجام شد. بنابراین، پیشنهاد می‌شود که، میل ترکیبی هسته اسید آمینه می‌تواند به عنوان یک توصیفگر بالقوه جدید برای طبقه‌بندی آنزیم ها مورد تایید باشد.

کلمات کلیدی: میل ترکیبی هسته، الگوریتم دسته‌بندی، جنگل‌های تصادفی، بانک داده پروتئین، دهالوژناز و غیر دهالوژناز.