Human Protein Function Prediction from Sequence Derived Features using See5

Manpreet Singh, Gurvinder Singh, Sonam Sharma

Abstract— Drug Discovery is a tedious process and involves lot of iterations and different processes for the final approval. The present work focus on prediction of molecular class of an unknown protein. The sequence data is taken from HPRD (Human Protein Reference Database) and then the different features are explored for each molecular sequence using various online tools. The decision tree was constructed based on training data of 55 sequences and test data of 29 sequences from different type of molecular classes. See5 based on C5 decision tree is used to obtain the results. Continuous data involving the values of sequence derived features for different sequences is given as input. Different advanced options and combinations are tried out of which decision tree powered with boosting and winnowing give the maximum accuracy of 30% for the data under consideration. If the continuous data set for 25 sequences is taken then the accuracy comes out to be 64% with the same technique.

Index Terms— Drug Discovery, sequence derived features, HPRD, decision tree, C5 decision tree, boosting, winnowing, protein function.

1 Decision Trees

Decision Tree is a form of directed tree which consists of a node called root that has no incoming edges. Other nodes with outgoing edges are called testing nodes and nodes at last level are called terminal nodes or decision nodes. Testing nodes are represented by circles while decision nodes are represented by triangles.

Decision tree has the ability to depict the final output among different set of attributes. It refers to hierarchical structure of the problem and their consequences given. The goal is to create a model that predicts the value of a target variable based on several input variables. Decision trees mainly are the structured representation of data that tells us the path followed in obtaining the solution. It follows the "White Box" technique that tells us that what are the steps been followed while obtaining the target Solution. [1], [5]

1.1 C5 Algorithm

C5.0 algorithm was developed by Quinlan in 1987. This algorithm mainly deals with the construction of decision tree which is been formed by the selection of 'best attribute' from the given data. When the attribute is selected from current node, its children nodes are been generated. Best attribute of a node can be selected using following criteria:

1. Random Method : Select any attribute at random
2. Least Value Method : Choose the attribute with the smallest number of possible values
3. Max Value Method : Choose the attribute with the largest number of possible values
4. Max Gain: Choose the attribute that has the largest expected information gain (to select the attribute that will result in the smallest expected size of the subtrees rooted at its children).

Among these, C5.0 algorithm uses the Max-Gain method of selecting the best attribute. [5]

1.1.1 Calculating Information Gain in C5

Step1 : Calculate the Information content present in a sample of data or complete data. Information content can be evaluated using following formula:

$$I(A)= \sum P(x)^*-\log 2\, P(x) \quad (1)$$

Where A is the set of various data samples , x is the variable that range over values to be encoded and P(x) is the probability of occurrence of a value.

Step2: Calculate the remainder that is the weighted sum of the information content of each subset of the attributes that are associated with each child node formed from the sample data by using the formula given in equation 2. Consider that:

A is an attribute with m possible values
Si is the subset of S with value i, i=1,...,m
Pi is the subset of Si that are positive examples
Ni is subset of Si that are negative examples
qi is |Si|/|S| = % of examples on branch i
%Pi is |Pi|/|Si| = fraction of positive examples on branch i
%Ni is |Ni|/|Si| = fraction of negative examples on branch i

$$\text{Remainder (A)} = \sum_{i=1}^{m} q(i) I(\%Pi, \%Ni) \quad (2)$$

Step3: Calculate Information gain using following formula:

$$\text{Gain= I(A)-Remainder(A)} \quad (3)$$

After evaluating all above steps, the attribute whose Information Gain is maximum is selected as the root node in creating decision tree. [4], [6]

1.1.2 Constructing a Decision Tree using Information Gain

A decision tree can be constructed by using top-down methodology from the information gain in the following way:

1. Start at the root node.
2. Calculate the attribute with the highest information gain.
3. Add a child node for each possible value of that attribute.
4. Attach the sample data to the child node where the attribute values of the sample data is identical to the attribute value attached to the node.
5. If data attached to the child node can be classified uniquely add that classification to that node and make it as leaf node.
6. Go to step two if there are unused attributes left, oth-

erwise add the classification.

2 BACKGROUND

2.1 See5 as a classifier

Quinlan proposed See5/C5.0 algorithm which mainly emphasizes on rule-based classifiers because they are easy to understand that means each rule can be separately examined and validated, without having to consider it as a whole. See5/C5.0 is fast with very good performance in just few seconds. It can also generate decision trees, which are useful when there is need to construct the classifiers more quickly. [10]

Arditi, D. et al. (2005) implemented an application of construction litigation using See5. A boosted decision tree system was used to predict the outcome of construction litigation. The study was conducted by using the same 114 Illinois court cases that were used in earlier prediction studies conducted with artificial neural networks in 1998 and case-based reasoning in 1999, augmented by an additional 18 cases that were filed in 1990–2000. All cases were extracted from the Westlaw on-line service. The best prediction result obtained with boosted decision trees was 90%. [2]

Wei-Feng, H. et al. (2011) stated the relationship between the synthetic features and the types of final product are critical for the rational synthesis of zeolite-type open-framework materials. In this paper, a prediction system based on C5.0 combined with a feature selection was proposed. The performance of the method was evaluated using classification accuracy and a receiver operating characteristic (ROC) curve. The results show that the highest area under the ROC curve (90%) and the classification accuracy (88.18%) was obtained for the decision tree model that contains eight input features and some useful rules with high confidence degrees were extracted from the model. [3]

2.2 Methods for Human Protein Function Prediction

Jensen, L. et al. (2002) described the development of entirely sequence-based method that identifies and integrates relevant features that can be used to assign proteins of unknown function to functional classes, and enzyme categories for enzymes. This paper show that strategies for the elucidation of protein function may benefit from a number of functional attributes that are more directly related to the linear sequence of amino acids, and hence easier to predict, than protein structure. These attributes include features associated with post-translational modifications and protein sorting, but also much simpler aspects such as the length, isoelectric point and composition of the polypeptide chain. [6]

Friedberg, I. (2006) stated that not only is the volume of pure sequence and structure data growing, but its diversity is growing as well, leading to a disproportionate growth in the number of uncharacterized gene products. Consequently, established methods of gene and protein annotation, such as homology-based transfer, are annotating less data and in many cases are amplifying existing erroneous annotation. Second, there is a need for a functional annotation which is standardized and machine readable so that function prediction programs could be incorporated into larger workflows. This is problematic due to the subjective and contextual definition of protein function. He emphasized the need to assess the quality of function predictors. [4]

Singh, M. et al. (2007) described that to overcome the problem of exponentially increasing protein data, drug discoverers need efficient machine learning techniques to predict the functions of proteins which are responsible for various diseases in human body. This outline the existing decision tree induction methodology C4.5 uses the entropy calculation for best attribute selection. This paper described that for the same test data, the percentage accuracy of the new HPF (Human Protein Function) predictor is 72% and that of the existing prediction technique is 44%. The data considered in this case is discrete. [8]

Singh, M. et al. (2011) described the cluster analysis as a form of unsupervised learning and cluster analysis is implemented for human protein class prediction. The data is accessed from Human Protein Reference Database (HPRD) which is related to human protein. The sequences related to ten molecular classes are obtained using HPRD. Five amino acid sequences are obtained for each of the molecular class. SDFs (Sequence derived Features) are extracted for each sequence by using various web based tools. On the basis of values of input SDFs and by considering priority of each of the SDF, clusters of the data available in the adjacency matrix are generated. Then those clusters are backtracked to predict the class of the entered sequence. [7]

3 SEE5/C5 : DATA COLLECTION

This C5 implementation aims to predict the molecular class based on different sequences of human protein. Our sample data consists of 55 protein sequences for fifteen molecular classes. Each case consists of 21 attributes of a human protein sequence. These attributes are Number of amino acids, Molecular weight, pI, Number of negative ions, Number of positive ions, Extinction coefficients 1 and 2, Instability index, Aliphatic index, Gravy, T, S, Ser, Thr, Tyr, Mean, D, Probability, ExpAA, Number of helices(PredHel) and ProbN. And fifteen molecular classes under consideration are, 4 cases indicate Defensin class, 4 indicates Acid Phosphatase, 5 for Voltage Gated Channel,6 for DNA Repair Protein, 2 for Decarboxylase, 4 for Heat Shock Protein, 3 Aminopeptidase, 5 for G-Protein, 4 for Water Channel, 2 for Neuraminidase, 4 for Nucleotidyltransferace, 3 for B Cell Antigen Receptor, 4 for Cell Surface Receptor , 2 for Transport Cargo Protein and 3 for RNA Binding Protein. C5 creates 11 rules for predicting molecular class to create the theory of 11 rules. [9], [10]

3.1 Preparing data for C5

C5 predicts the molecular class by using its various properties or sequence derived features. C5 also constructs decision tree for set of classifiers or also generates the rules. The different files used in See5 are described below: [10]

3.1.1 Application Files

C5.0 application has a name called a filestem; we have used the filestem 'sequence' for this illustration. All files read or written by C5.0 for an application have names of the form 'fi-

lestem.extension', where 'filestem' identifies the application and 'extension' describes the contents of the file. Some of the applications used in this case are:

1. sequence.names : this is used to describe the application attributes.
2. sequence.data : this is used to represent the data on which classifiers are been generated.
3. sequence.test : this consists of unseen cases used to construct a classifier.

3.1.2 Names file

The file sequence.names is an essential file that describes the attributes and classes. There are two important subgroups of attributes:

1. The values of an explicitly defined attribute are given directly in data. A discrete attribute has a set of nominal values and continous attribute has a numeric value.
2. The value of implicitly defined attribute is specified by formula.

The first entry in names files specifies the class separated by commas as shown below:
Molecular class: defensin, acid phosphatase, voltage gated channel and so on…
The name of explicitly defined attribute is denoted by colon ':' . Here all the attributes are numeric values therefore it is labeled as 'continous' attribute.

3.1.3 Data file

The second essential file is the data file 'sequence.data' that contains all the sample data upon which C5 creates its rule sets or patterns. It consists of the values of all attributes separated by commas.

3.1.4 Test file

This is an optional file that is used to perform testing over unseen data. Here testing file is 'sequence.test' in which C5 calculates the accuracy of data.

3.1.5 Rulesets

Rulesets are the unordered collection of simple if-then rules.Each rule consists of:

1. Rule number: to identify the rule.
2. Statistics: (n, lift x) or (n/m, lift x) summarize performance of rule.

Here n specifies number of training cases covered by rule, m specifies how many of them do not belong to class predicted by the rule. And lift x determines the result of dividing rule's estimated accuracy by relative frequency of predicted class in training set.

3. One or more conditions to be satisfied if rule is applicable.
4. A class predicted by the rule.
5. Value 0 and 1 shows confidence with which the prediction is made.

Rulesets are generally easier to understand than trees since each rule describes a specific context associated with a class.

3.1.6 Rule Utility Ordering

In this, the rule that most reduces the error rate appears first and the rule that contributes least appears last. Moreover, results are reported in a selected number of bands so that the predictive accuracies of the more important subsets of rules are also estimated.

3.1.7 Boosting

The concept is to generate several classifiers (decision trees or rulesets) instead of one. On classifying a new case, each classifier votes for its predicted class and then the votes are counted to determine the final class. In the first step, a single decision tree or ruleset is constructed as before from the training data. This classifier will usually make mistakes on some cases, like here the first decision tree, gives the wrong class for 14 cases in sequence.data. When the second classifier is constructed, more attention is paid to these cases. As a result, the second classifier will produce different results from the first. It also will make errors on some cases, and these are again constructed by the third classifier. This process continues for a pre-determined number of iterations or trials, but stops when most recent classifiers are either extremely accurate or inaccurate. [3], [4].

3.1.8 Winnowing Attributes

C5 algorithm mainly uses a mechanism to separate the useful attributes from useless attributes and this process ia termed as 'Winnowing'. For example , in our case we have seen that we have total 21 attributes in our sequence file and one attribute of molecular class, but out of 21 only 9 attributes are been used to create decision trees or rulesets. This capability to choose among the predictors adds an advantage to creation of decision tree. This technique is time consuming and is mainly used in large applications. [3], [10].

3.1.9 Advanced Pruning Options

There are three further options of the classifier-generation process. These are best referred to as advanced options. C5.0 constructs decision trees in two phases. A large tree is first grown to fit the data closely and is then `pruned' by removing parts that have a high error rate. Firstly, this pruning process is applied to every subtree and then it is decided whether it should be replaced by a leaf or sub-branch, and after that a global stage looks at the performance of the tree as a whole. [3], [10].

3.1.10 Cross Validation Trials

The performance of a classifier on the training cases provides a poor estimate of its accuracy on new cases. The true predictive accuracy of the classifier can be estimated by sampling, or by using a separate test file.Here the classifier is evaluated on cases that were not used in training data. C5 algorithm construct a different classifier with a lower or higher error rate on the test cases. [3], [10].

4 Results and Discussion

Below Figures depicts the decision trees that are obtained after applying various See5 techniques on sequence file.

1. Fig. 1.shows Decision Tree using Classifier
2. Fig. 2.shows Decision Tree using Rulesets

3. Fig. 3.shows Decision Tree using Sort by Utility
4. Fig. 4.shows Decision Tree using Boosting
5. Fig. 5.shows Decision Tree using Winnowing
6. Fig. 6.shows Decision Tree using Advance Pruning Options

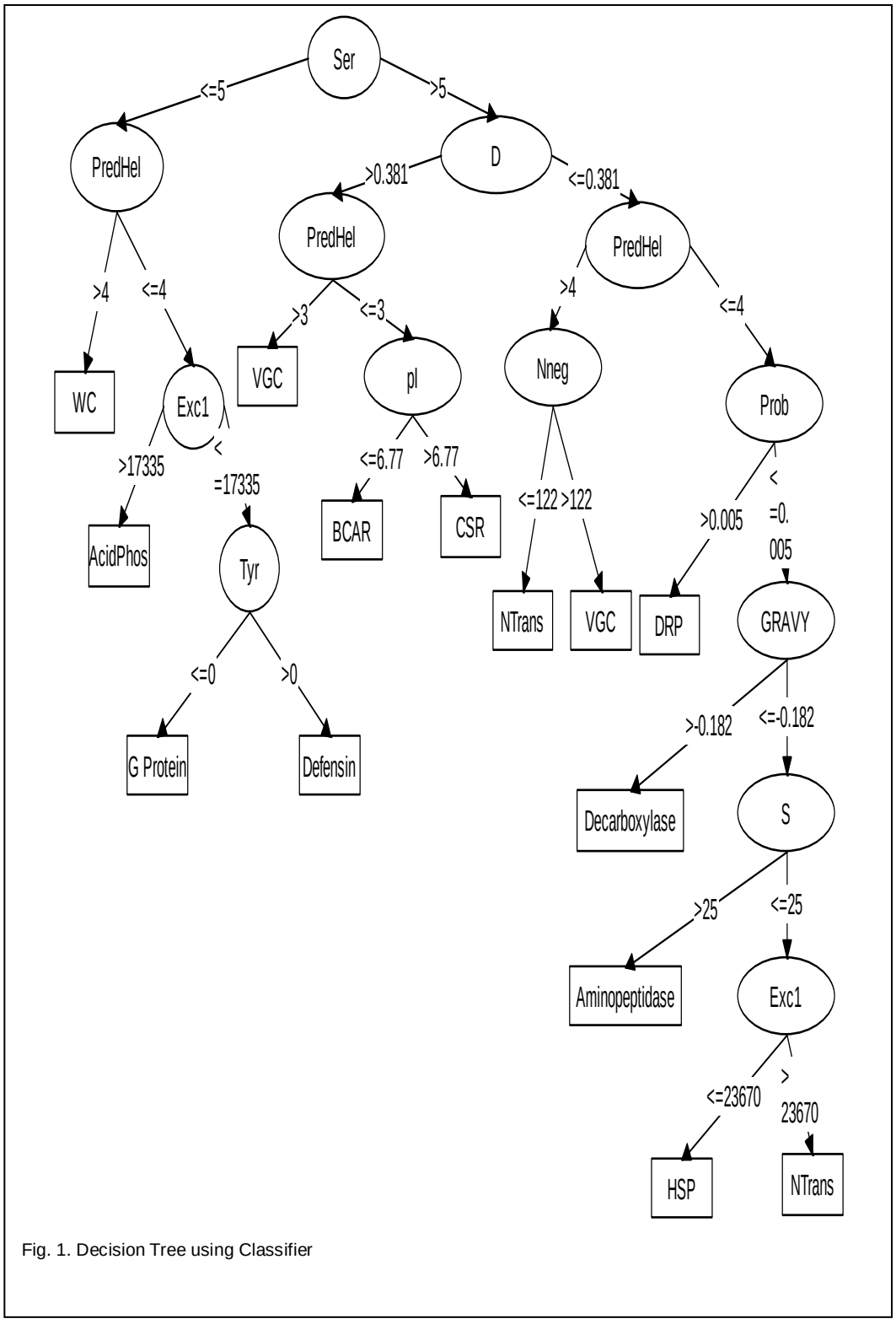

Fig. 1. Decision Tree using Classifier

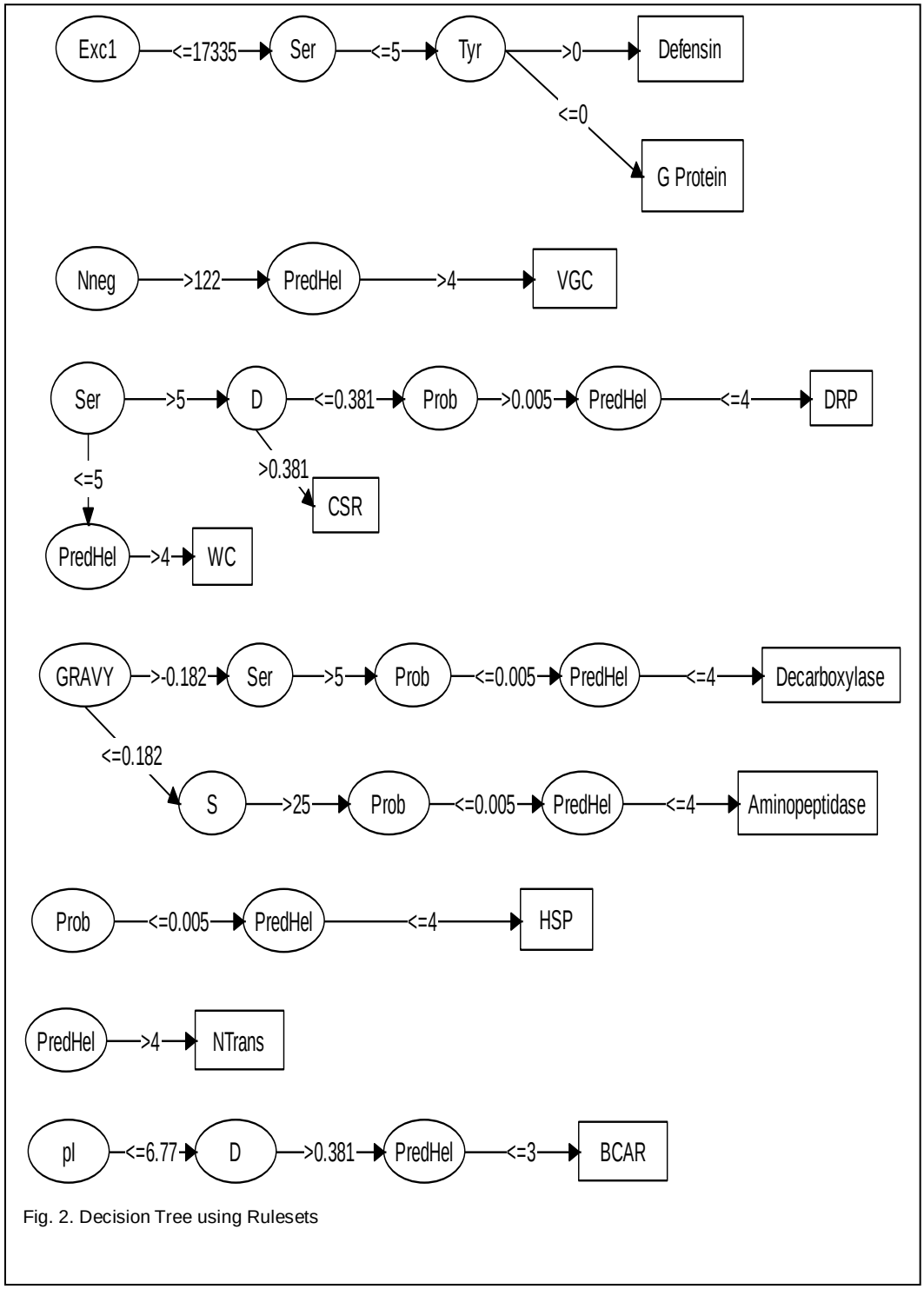

Fig. 2. Decision Tree using Rulesets

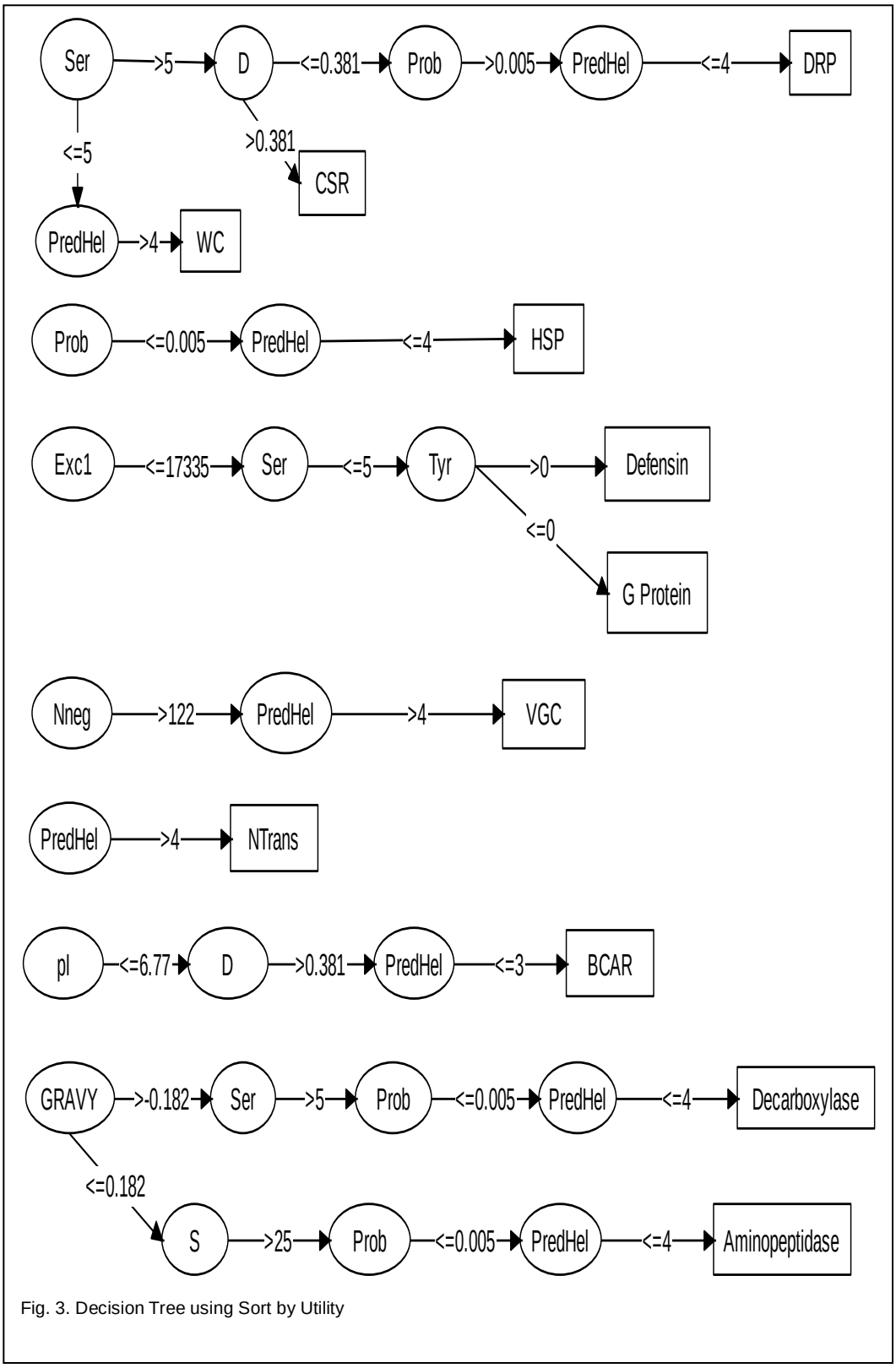

Fig. 3. Decision Tree using Sort by Utility

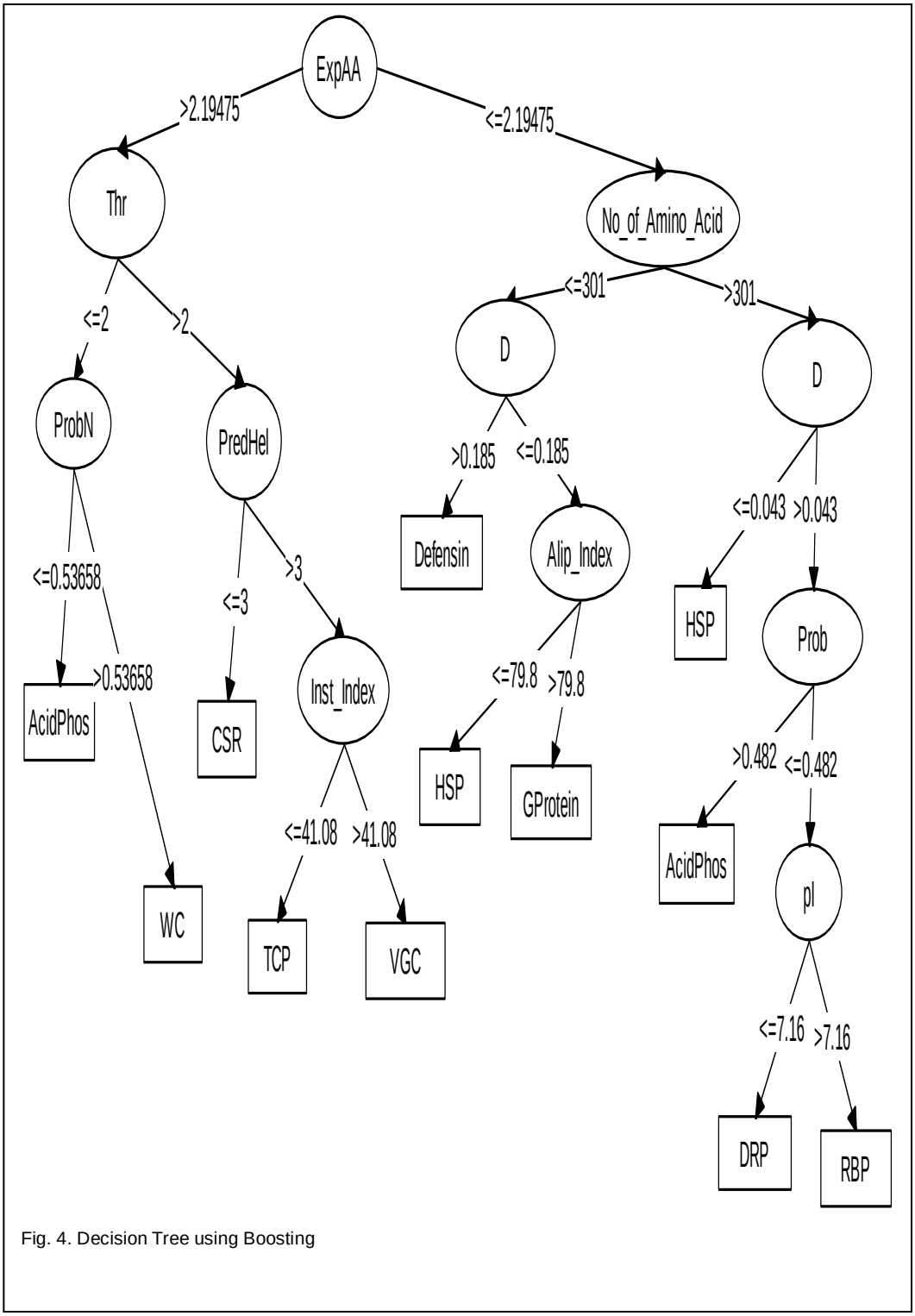

Fig. 4. Decision Tree using Boosting

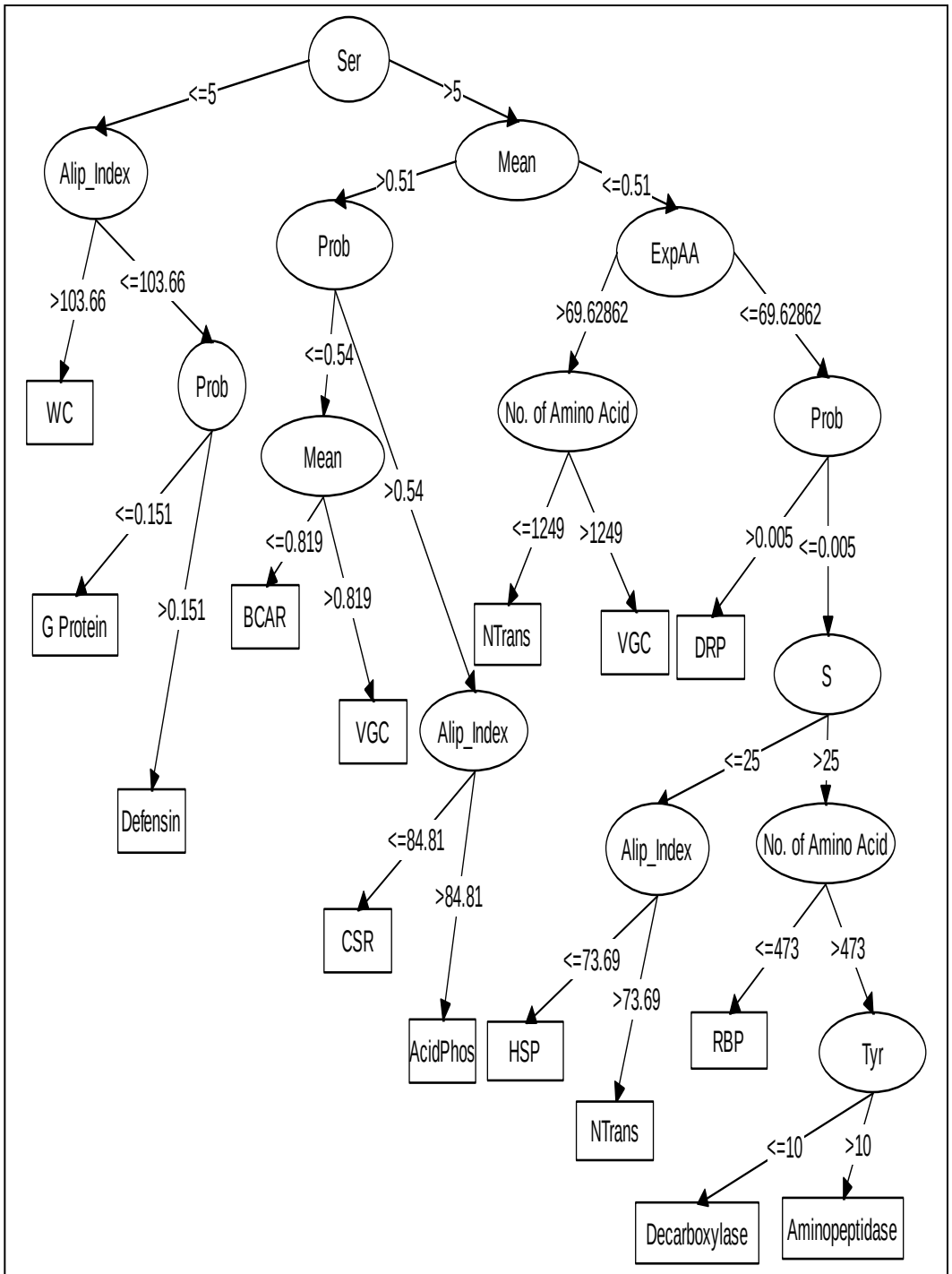

Fig. 5. Decision Tree using Sort by Utility

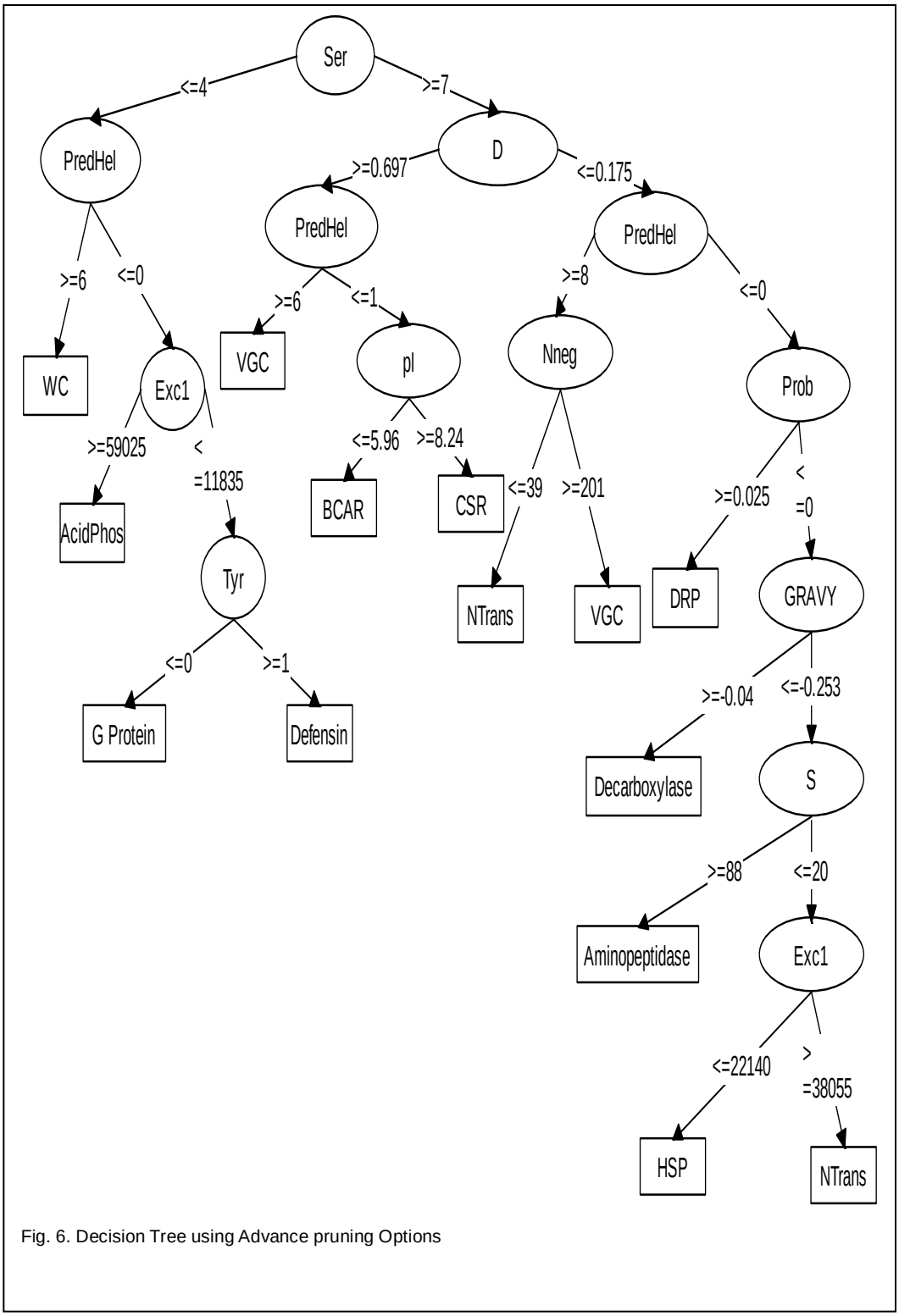

Fig. 6. Decision Tree using Advance pruning Options

5 Conclusion

For the data having 55 sequences and 21 features, the accuracy of the different techniques are shown in table 1. The C5 algorithm with winnowing and advance pruning option provides the maximum accuracy of 30%. If the same number of elements are taken as that of [8], the accuracy comes out to be 64%.

TABLE 1
CALCULATED ACCURACIES OF DIFFERENT TECHNIQUES USED IN SEE 5

Techniques in See5	Calculated Accuracy
Classifier	26.7 %
Rulesets	26.7 %
Rulesets by Utlility Ordering	26.7 %
Boosting	30%
Winnowing	20 %
Advance Pruning Option	30 %

References

[1] B. Bergeron, "Bioinformatics Computing", pp 257-270, 2002.
[2] D. Arditi and T. Pulket " Predicting the outcome of construction litigation using boosted decision trees. ", *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp 387–393, 2005.
[3] H. Wei-Feng, G. Na, Y. Yan, L. Ji-Yang, Y. Ji-Hong, "Decision Trees Combined with Feature Selection for the Rational Synthesis of Aluminophosphate AlPO4-5 ", *National Natural Science Foundation of China*, vol 27, no.9, pp 2111-2117, 2011.
[4] I. Friedberg, "Automated Protein Function Prediction- the Genomic Challenge ", *Briefings in Bioinformatics*, vol 7, no.3, pp 225-242.
[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques", *Morgan Kaufmann Publishers, USA* pp 279-322, 2003.
[6] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames C. Kesmir, H. Nielsen, H.H. Stærfeldt, K. Rapacki, C. Workman C.A.F. Andersen, S. Knudsen, A. Krogh, A.Valencia and S. Brunak , "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features", *Journal of Molecular Biology*, vol. 319, issue 5,pp 1257-1265, 2002.
[7] M. Singh, G. Singh "Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction", *International Journal of Computer Applications (0975 – 8887)*, vol. 20, no.3, pp. 22-27, 2011.
[8] M. Singh, P. K. Wadhwa and P. S. Sandhu , " Human Protein Function Prediction using Decision Tree Induction " *IJCSNS International Journal of Computer Science and Network Security*,vol. 7, no.4, pp. 92-98, 2007.
[9] www.hprd.org
[10] http://rulequest.com/see5-info.html.