Enhancing Usability of See5 (Incorporating C5 Algorithm) for Prediction of HPF from SDF

[1] **Sunny Sharma,** [2] **Amritpal Singh,** [3] **Dr. Rajinder Singh**

[1, 2, 3] Department of Computer Science,
Guru Nanak Dev University, Amritsar, India

Abstract - Prediction of molecular class of an unknown protein is an area of great relevance for carrying out research in various disease detections and their corresponding drug discovery processes and it is a very tough and challenging task. Some specific approaches were used in the past to increase the accuracy of Human protein Function (HPF) prediction. This research is primarily concentrated on one such approach of HPF prediction with sequence derived features (SDF) using decision trees and there variants implemented using C5 algorithm. More sequence derived features were identified and incorporated, training data was enhanced (Sequence data evolved from HPRD (Human protein reference database)) in terms of number of sequences and the features used to extract the relation towards a specific class. Multiple techniques were tested for accuracy in prediction and a comprehensive comparison was done amongst them and the previous research results.

***Keywords* - HPF, C5, See5, Decision Tree, SDF**.

1. Introduction

A time tried and tested approach of prediction is decision tree based prediction. It is a white-box technique which clearly illustrates the sequence of computations involved at each and every stage. This plus point enables its usage by computational experts even without much knowledge of the concerned domain. Likewise, it enables an expert from the concerned domain to critically examine the steps followed by a computational expert. So it bridges the gap between technical know-how and domain expertise. Decision tree comprises of nodes and edges depicting various functionalities at different levels of computations. A decision tree clearly illustrates the required results or outputs amongst various outcome possibilities. It clearly defines the problem structure and its interpretations in a hierarchical way which is much easier to comprehend. As the model has a unique ability of taking into account various input parameters and reaching a goal.

2. Decision Tree for Class Detection

Decision trees approach is a very potent methodology of supervised learning. Dataset are divided into categories which are as identical as possible regarding consideration of the variable to be identified. A set of classified data is provided as input and a tree that signifies an orientation diagram having each of the leaf nodes as a class i.e. decision and each internal node indicates a test that is obtained as an output. Decision of relation to a class of data is indicated by each of the leaf confirming to all the tests path from the root node to that of the leaf node.

3. C5 Algorithm

Quinlan'sC5.0 algorithm is widely used for classification process. Algorithm primarily focuses on constructing a decision with the identification of most important attributes from the supplied/identified data-set. Once the attribute is finalized from current node, corresponding child nodes are then generated. Best attribute of a node can be selected using following criteria:

- Random Method: Random selection of attribute.
- Least Value Method: Attribute with the least number of possible values is selected.
- Max Value Method: Attribute with the maximum number of possible values is selected.
- Max Gain: Attribute having largest anticipated information gain (selection of attribute resulting in the least possible size of the sub trees rooted at its children).Max-Gain is used for selection of the most suitable attribute.

4. Classification

Rule-based classification is preferred because it is easy to comprehend and the reason lies in examining and validation of every rule individually without bothering about its holistic impact. See5 (Implementing C5) is an excellent tool when performance is taken into account. Decision trees are generated and they are of great use when quick construction of the classifiers is required. [10]. Arditi, D. et al. (2005) examined construction litigation application of See5. A boosted decision-tree system approach was incorporated to predict the results of construction-litigation domain. Same data-sets as used in previous prediction related examination conducted with ANN's earlier and case-based reasoning afterwards were included in the research, augmented by an additional cases that were filed in 1990–2000. All cases were extracted from the Westlaw on-line service. Boosted decision trees provided a superior prediction accuracy of 90%. [2]. Wei-Feng, H. et al. (2011) demonstrated that the link between the synthetic features and the types of final product are very important for the material's rational synthesis. A prediction mechanism was proposed that was C5centric and combined with a feature selection. Classification accuracy and a receiver operating characteristic (ROC) curve determined the performance credential for the proposed methodology. Highest area under the ROC curve (90%) and the classification accuracy (88.18%) was achieved in results for the decision tree model containing 8 input attributes and some important rules with high confidence degrees were extracted from the model [3].

5. Literature Survey

Jensen, L. et al. (2002) focused on developing fully sequence-based method that recognizes and combines important features for the purpose of assigning proteins of unknown function to respective classes and enzyme classification. A number of functional features that are more appropriately related to the linear sequence of amino acids may benefit the strategies for the elucidation of protein function, and hence quite simple to predict, than protein structure. Identified Attributes include features associated with post-translational modifications and protein sorting; also include simpler aspects such as the length, composition of the polypeptide chain and isoelectric point [6]. Friedberg, I. (2006) showed that not only is the volume and diversity of pure sequence and structure data is increasing and resulting to a unequal 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. Also functional annotation is desired which is standardized and machine readable for the requirement of prediction programs implementation on larger workflows. Subjective and contextual definition of protein function is cumbersome in nature. The need to assess the quality of function predictors needs to be stressed upon [4]. Singh, M. et al. (2007) exponential increase in protein data was suggested to solve the problem; drug discoverers need efficient machine learning techniques to predict the functions of proteins which are responsible for various diseases in human body. Decision tree induction methodology used in C4.5 for the selection of best attribute involves the entropy calculation. For the discrete same test data, the correctness of the new HPF (Human Protein Function) predictor was 72% and that of the existing prediction methodology was 44% [8]. Singh, M. et al. (2011) presented 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].

6. Implementation on See5

On the basis of different sequences of human protein the C5 implementation predicts the molecular class. In the sample data 15 Protein classes, with 70 protein sequences is taken and each of them is having 25 attributes or features [9], [10].

6.1 Required Data for See5

Class estimation is done using different features (SDF's) by See5 tool. Decision tree or rules set are got for set of classifiers using See5. Files required in See5 implementation are as follows: Application file, Names file, Data file, Test file, Rule-sets [10].

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.

The file sequence names is an essential file that describes the attributes and classes: The values of an explicitly defined attribute are given directly in data. A discrete attribute has a set of nominal values and continuous attribute has a numeric value. The value of implicitly defined attribute is specified by formula. C5 creates its rule sets or patterns containing all the sample data in a data file. It consists of the values of all attributes separated by commas. Test file is an optional file that is used to perform testing on unseen data. Random collection of simple if-then rules is called Rule-sets. Each rule consists of Rule number and Statistics (summarizing performance of rule).

6.2 Boosting, Winnowing Attributes and Advanced Pruning Options

Boosting is to generate several classifiers (decision trees or rule-sets) instead of one. On classifying a new case, each classifier supports its predicted class and then the support is evaluated to determine the final class. In the first step, a single decision tree or rule-set 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. Other classifier is constructed giving more consideration to the cases. Thus the classifier will provides results variation from the earlier classifier. Errors induced are again rectified by another classifier. It continues for defined iterations/trials and halts once extremely correct classifies is achieved [3], [4].

Winnowing is a mechanism to separate the useful attributes from useless attributes. It provides option to select among the predictors and have an edge to create a suitable decision-tree. However, it's time intensive task and primarily suitable for bigger application domain. [3], [10].

In Advanced pruning technique a massive tree is first allowed to grow to fit the data closely after that it's pruned i.e. error causing segments are removed. Every sub-tree undergoes pruning then replacement by a leaf or sub branch is decided and then a global stage evaluates performance of the tree as a unit. [3], [10].

6.3 Cross-Validation

The real prediction correctness of a classifier can be evaluated by sampling i.e. using different test files rather than relying only on training data. So cross validation is done using unseen data as well and enhances accuracy in the prediction process.

7. Implementation Results

Decision trees obtained with different See5 techniques are shown as follows:
Decision Trees Implemented with:

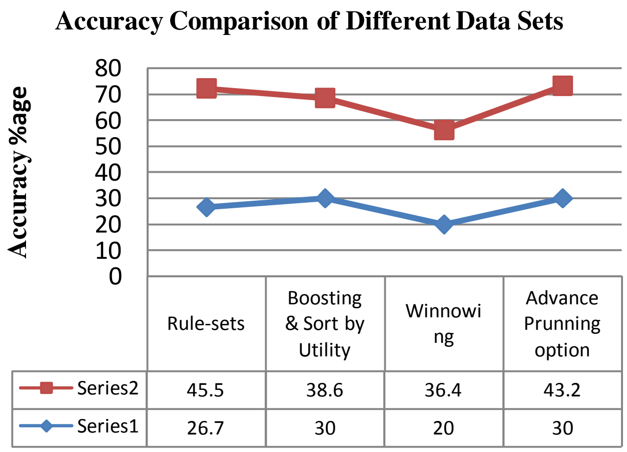
- Winnowing (shown in Figure: 1)
- Boosting & Sort by Utility (shown in Figure: 2)
- Rule-sets (shown in Figure: 3)
- Advance Pruning (shown in Figure: 4)

8. Results & Discussions

The dataset containing 70 sequences and 25 features was examined and the correctness of various techniques are depicted in Table 1. The C5 algorithm with winnowing and advance pruning option provides the maximum accuracy of 45%. If the same number of elements are taken as that of [8], the accuracy comes out to be 64%.

Table 1: Accuracies Details

See5 Classifier	Previous data Set	Improved data Set
Rule-sets	26.7	45.5
Boosting & Sort by Utility	30	38.6
Advance Prunning option	30	43.2
Winnowing	20	36.4



	Rule-sets	Boosting & Sort by Utility	Winnowing	Advance Prunning option
Series2	45.5	38.6	36.4	43.2
Series1	26.7	30	20	30

Fig. 5: Accuracy Comparison

9. Conclusion

Present work focus on usability of see5 tool in HPF prediction and also demonstrate the impact of choosing the right training data. The detailed analysis shows that increasing number of features (5 features) of HPF data increases the accuracy of prediction process (about 16%)but does not necessarily involves the participation of all parameters in decision making process. Some parameters were more dominant than others (like GRAVY

13%, Solubility 8%, Thr 4%) hence they decide the course of prediction. Activities like advanced pruning and winnowing (17 attributes winnowed) help in minimizing the computation time and also help in reaching the most important parameters involved in prediction process (ExpAA came out as most important parameter after winnowing). In future more features can be extracted on more sequences and their relative impact on prediction process can be examined hence it will lead to greater precision in the HPF identification process. Inclusion of comparison feature in See5 tool can be of great importance as it will help researchers in identification of correct rule-set and role of newly incorporated feature for the HPF prediction scenario.

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 Com-bined with Feature Selection for the Rational Synthesis of Aluminophos-phate 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 Chal-lenge", *Briefings in Bioinformatics*, vol 7, no.3, pp 225-242.

[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques", *MorganKaufmann 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.

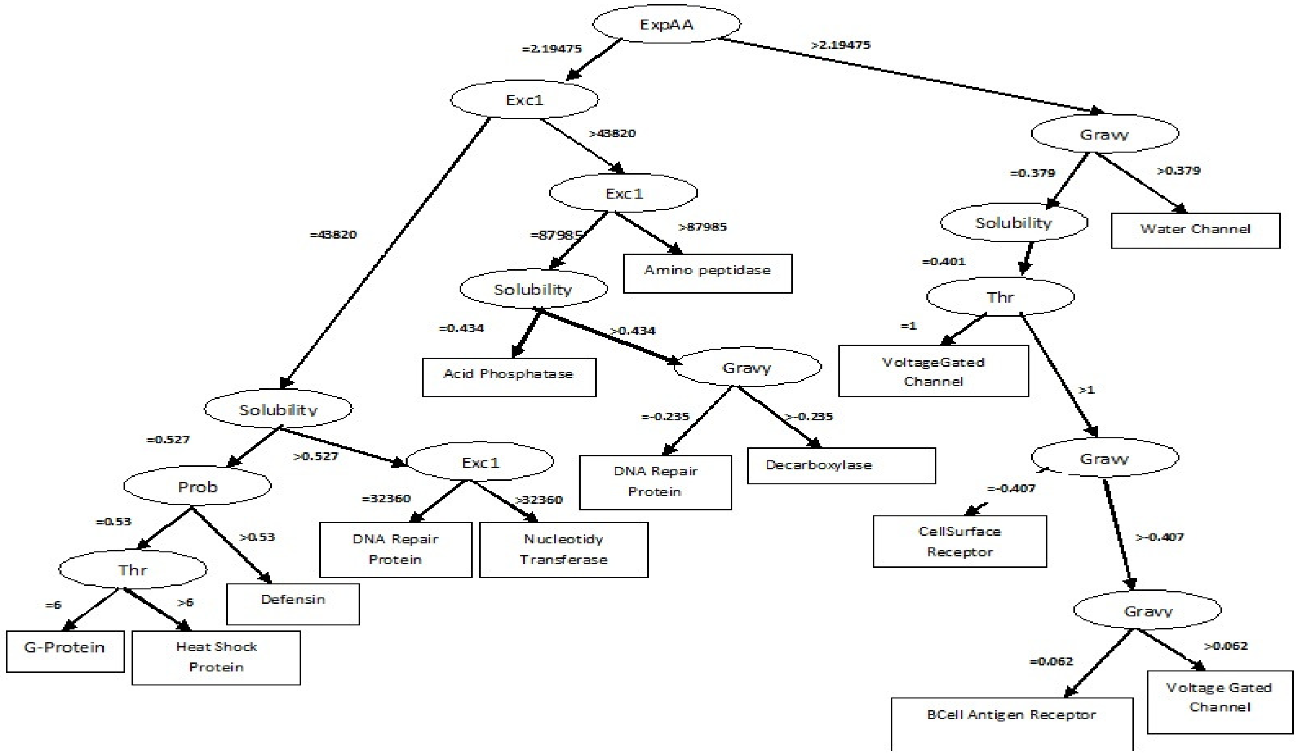


Fig. 1: Decision Tree Implementation with winnowing option

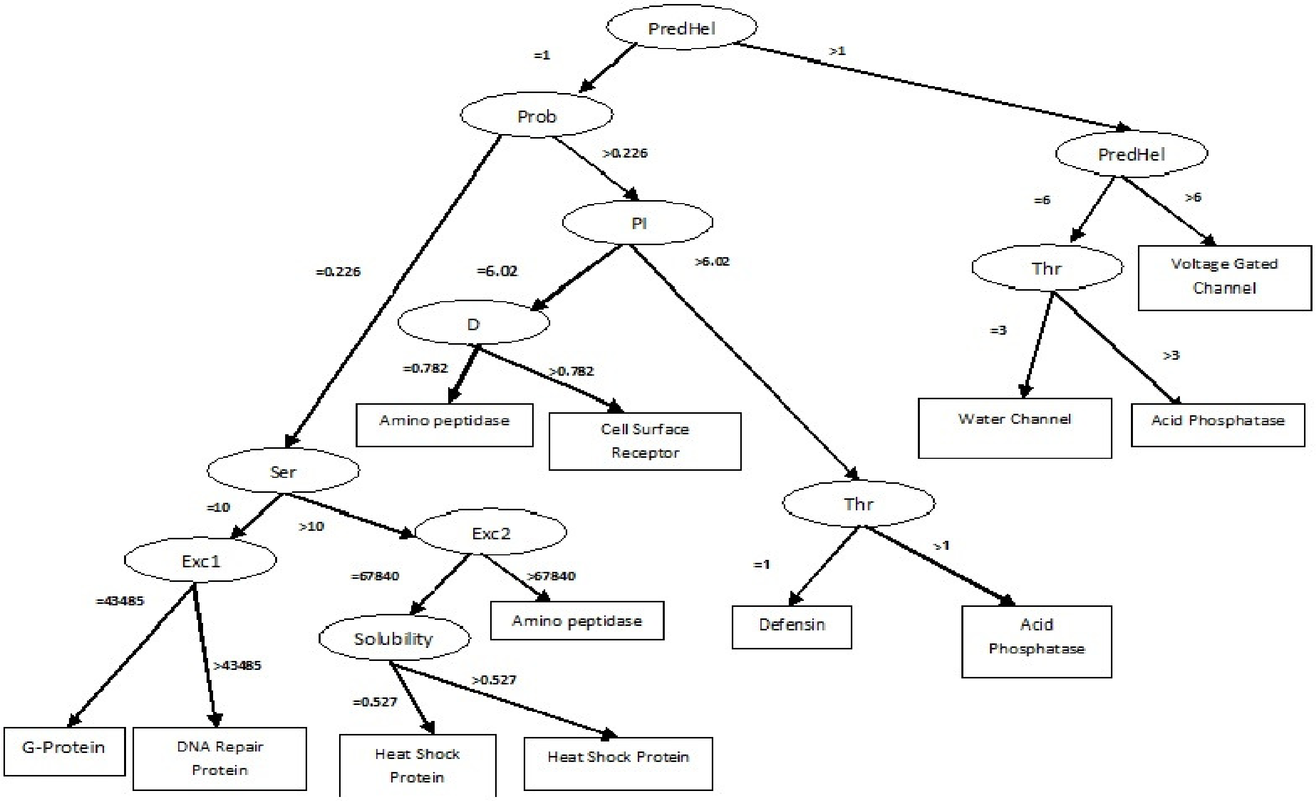


Fig. 2: Decision Tree Implementation with Boosting & Sort by Utility option

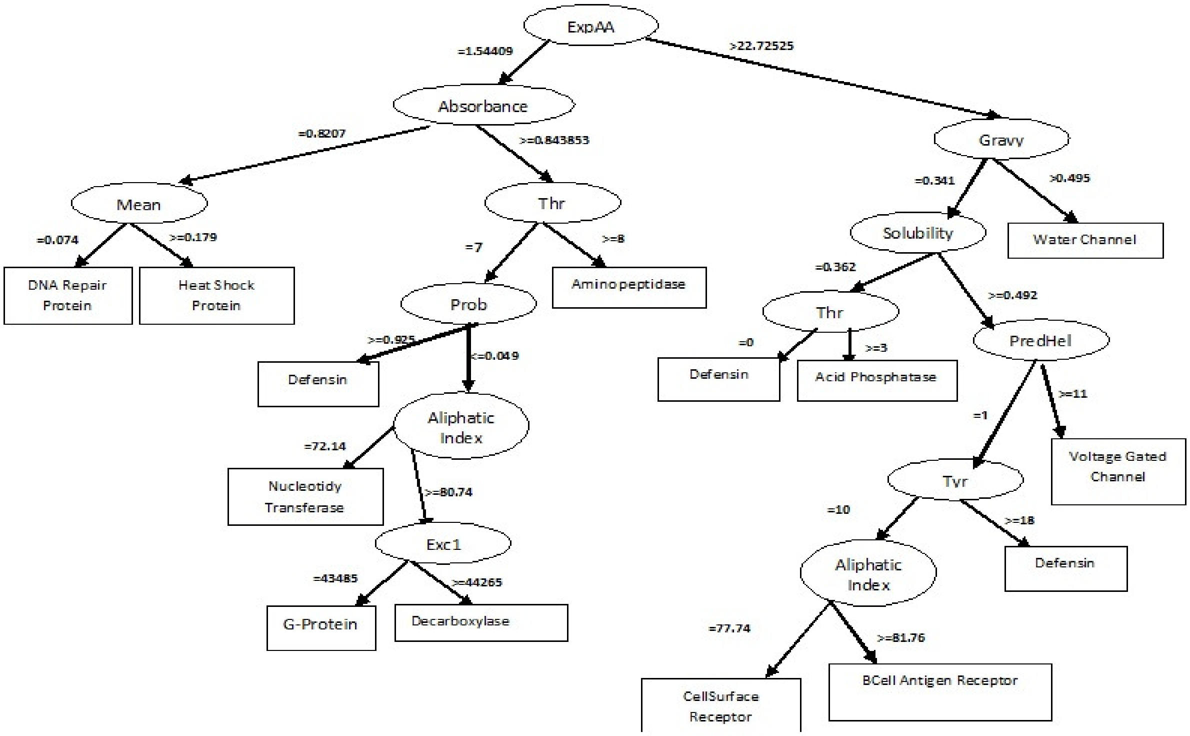


Fig. 3: Decision Tree Implementation with Advance Pruning option

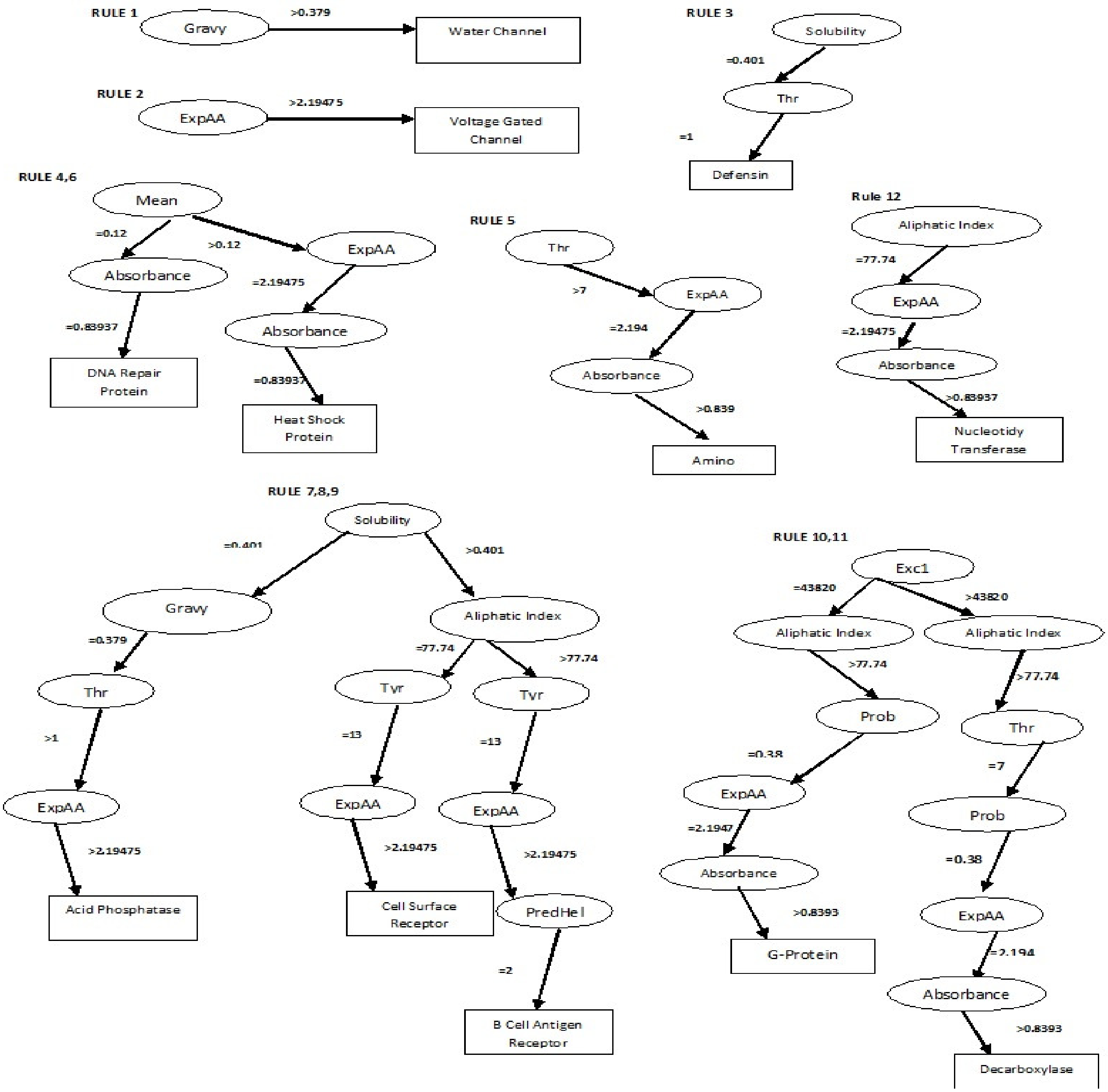


Fig. 4: Decision Tree Implementation with Rule-sets option