Plant Gene Classification and Functionality Prediction

Problem Statement

In the area of computational biology, it is very interesting to know about the geneclass and prediction of the function of gene. Gene classification or identification refers to recognize the area through certain computational process where genomic DNA can found. Each DNA sequence has certain functionality and if DNA sequence is available with us then the objective is to predict the use of function of gene from the basic data.

Any biological information includes gene expression data, protein interaction data, genome sequence etc. It is an important step toward the prediction of the gene network in the cell. This different type of information can be analyzed through computation to know about the class of gene and corresponding functions and even the hidden property can be predicted**.** It is having mush of the probability every specific data type possess its own strengths and weaknesses in discovering specific relation. Literature proposes a new method to cluster genes optimally in such a way that it can be used to predict the role of undiscovered genes.This finding will be based on analysis of multiple data sources. The basic idea is tofind the maximum of the similarity gain function among all individual clusters.

Background

J. Jung and M. R. Thon. [“Automatic annotation of protein functional class from sparse and imbalanced data sets”]**.** Proposes automated gene annotation methods using hierarchical nature vocabulary. This approach can greatly simplify the classification problem of genes. However, several authors proposed that hierarchical information can be effectively used for gene function annotation.The training set including hierarchical information can outperform similar where hierarchical nature of the data can be ignored.

In other paper Shahbaba and Neal [“Gene function classification using Bayesian models with hierarchy based priors”] describe three multinomial models. It uses simple tree-like structure where each node belongs to only one parent.

King et al. [“Predicting Gene Function from Patterns of Annotation”] describe the process predict gene function prediction based on the relation between GO(Gene Ontology) annotation. It uses decision trees with Bayesian network. Bayesian network can also be effectively used for the purpose of devising a multilabel annotation, so this process can easily reduce the disadvantages of inconstancy annotations of parent child as this prevent the Childs from being annotated.

Methodology

Bayesian network is used as underlying architecture for this approach where directed acyclic graph of the Gene Ontology (GO) is constructed. To further enhance the expected outcome data is taken from experts of concerned area. After collecting the training data of proteins with annotations, prior probability of each node is computed. Using the InterProScan application referred in literature, where InterPro terms associated and assigned to proteins automatically. Thus, Conditional probability of each InterPro term is computed at each node level in the network. Then, Bayesian probability for each GO term is calculated as:

$$P(X_{1\in\{F,T\}},\cdots,X_{i\in\{F,T\}}) = \prod_{i=1}^{v} P(X_i|Par(X_i)) \quad (1)$$

Where Xi is Gene ontology term in given network. Par is referred to as probability parent. The conditional probability can computed as given above as P(Xi).

Presence of every GO term is with annotation is tested with the Bayesian Probilbity computation and know whether it is without root term or with root term. If the presence probability is more than conditional probability is computed recursively child annotation of proteins genes. A list of occurrence is pairs are developed for all proteins in the training set.

Experimental design

The modified experimental setup can be used where hierarchal structure of gene ontology is exploited. This paper proposes an improved HMC (Hierarchal multi-label classification) approach. The HMC pattern is represented with help of support vector machines. Then, two measures are applied to provide better result. Firstly SMOTE is applied (Synthetic Minority over-Sampling technique (SMOTE) to preprocess the unbalanced training subsets SVM classification. Further, a modified TPR approach is used to get binary support vector classifiers probabilistically. Thus it improved classification results to considerable extent.

In this approach an instance x can be represented as C = {c1, c2, ⋯, cm}.

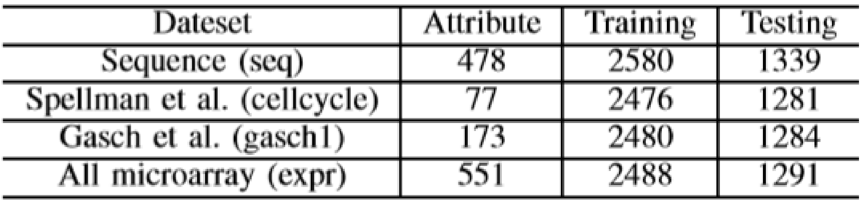
A vector y = (y1,y2,⋯,y_m) ∈ $\{0,1\}^m$ is used for labels. i.e. if xi belongs of Ci then Yi will be 1 else it will be zero. All the classes are represented as rooted tree. Here nodes represents the class and paths from individually represent the relation between classes.For x dataset item, parents and child of a specific node ci, the corresponding classifier will follow the following rules:

$$\begin{cases} d_i = 1 \Rightarrow d_{par(i)} = 1 \\ d_i = 0 \Rightarrow d_{child(i)} = 0 \end{cases}$$

here d_i is label for the classifier.

Predictions achieved at each node level are stored by classifier and positive decisions are communicated from down to upward direction.

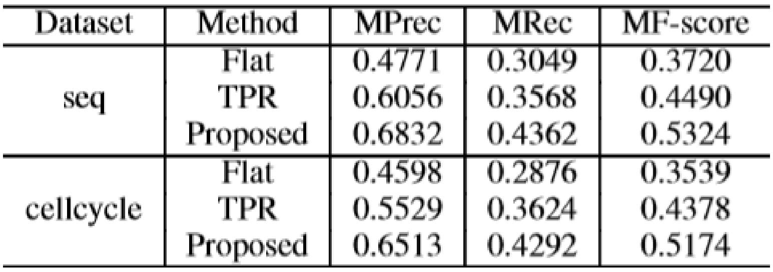
Properties of dataset used by author.



Dateset	Attribute	Training	Testing
Sequence (seq)	478	2580	1339
Spellman et al. (cellcycle)	77	2476	1281
Gasch et al. (gasch1)	173	2480	1284
All microarray (expr)	551	2488	1291

Classification results of given data set produced

CLASSIFICATION RESULTS ON THE FOUR FUNCAT DATASETS.



Dataset	Method	MPrec	MRec	MF-score
seq	Flat	0.4771	0.3049	0.3720
	TPR	0.6056	0.3568	0.4490
	Proposed	0.6832	0.4362	0.5324
cellcycle	Flat	0.4598	0.2876	0.3539
	TPR	0.5529	0.3624	0.4378
	Proposed	0.6513	0.4292	0.5174

Result and Discussion: The above dataset describes different aspects of genes in and the different aspects of gene function predicted. The Flat ensemble technique used here does not consider the hierarchal structure. Second approach refers Basic hierarchal TPR Ensemble where parent wt. is fixed to specific value and the last approach result are obtained by applying the approach discussed in this report which comparatively better than previous two approaches.