International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:2, No:3, 2008

Predicting Protein Function using Decision Tree

Manpreet Singh, Parminder Kaur Wadhwa, and Surinder Kaur

***Abstract*—**The drug discovery process starts with protein identification because proteins are responsible for many functions required for maintenance of life. Protein identification further needs determination of protein function. Proposed method develops a classifier for human protein function prediction. The model uses decision tree for classification process. The protein function is predicted on the basis of matched sequence derived features per each protein function. The research work includes the development of a tool which determines sequence derived features by analyzing different parameters. The other sequence derived features are determined using various web based tools.

***Keywords*—**Sequence Derived Features, decision tree.

I. Proteins and Their Role

PROTEINS are the primary components of living things, and they play many roles. Proteins are the molecular machinery that regulates and executes nearly every biological function [4]. Proteins provide structural support and the infrastructure that holds a creature together; they are enzymes that make the chemical reactions necessary for life possible; they are the switches that control whether genes are turned on or off; they are the sensors that see and taste and smell, and the effectors that make muscles move; they are the detectors that distinguish self from oneself and create an immune response.

Proteins have a variety of roles that they must fulfill:

- They are the enzymes that rearrange chemical bonds.
- They carry signals to and from the outside of the cell, and within the cell.
- They transport small molecules.
- They form many of the cellular structures.
- They regulate cell processes, turning them on and off and controlling their rates.

Despite their radical differences in function, all proteins are made of the same basic constituents: the amino acids. Each amino acid shares a basic structure, consisting of a central carbon atom (C), an *amino* group (NH3) at one end, a *carboxyl* group (COOH) at the other, and a variable side chain (R), as shown in Fig. 1.

Chains of amino acids are assembled by a reaction that occurs between the nitrogen atom at the amino end of one amino acid and the carbon atom at the carboxyl end of another, bonding the two amino acids and releasing a molecule of water. The linkage is called a *peptide bond*, and long chains of amino acids can be strung together into polymers, called *polypeptides*, in this manner. All proteins are polypeptides.

When a peptide bond is formed, the amino acid is changed (losing two hydrogen atoms and an oxygen atom), so the portion of the original molecule integrated into the polypeptide is often called a *residue*. The sequence of amino acid residues that make up a protein is called the protein's *primary structure*.

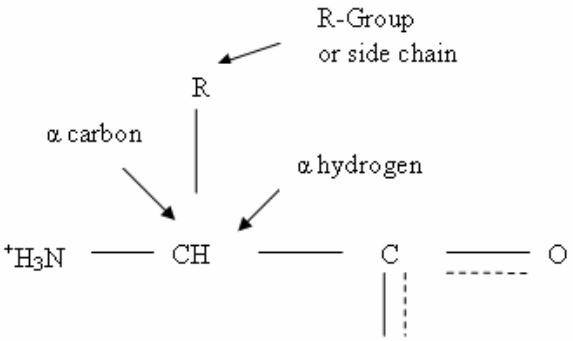

Fig. 1 Basic Chemical Structure of Amino Acid

II. Protein Function

The definition of biological function is ambiguous, and the exact meaning of the term varies based on the context in which it is used. It is obvious that the biological function of a protein has more than one aspect. Take for example a protein kinase; in the biochemical aspect, a kinase's function would be the phosphorylation of a hydroxyl group of a specific substrate. The scope of interest implied by this definition does not require any more than a 'disembodied' protein performing alone in vitro. However, proteins perform their function within an organism, and this has consequences ranging from the subcellular to the whole-organism level. In a physiological aspect, the same kinase may be part of a signaling pathway, where a protein both phosphorylates, and is phosphorylated by, interacting partners. A mutation in this kinase might cause a disease, so yet another aspect is a phenotypic or medical one. Therefore, when speaking of a protein's function, we must always specify the aspect or aspects of the functional description [2].

Manpreet Singh is with the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: mpreet78@yahoo.com).

Parminder Kaur Wadhwa is with the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: parminder_k_engg@yahoo.com).

Surinder Kaur is with Department of CSE, Institute of Engineering and Technology, Bhaddhal, Ropar (e-mail: saini_seema_in@yahoo.com).

III. Applications of Function Prediction in Cancer Research

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge [7].

Page et al. presents the most extensive study to date of the protein expression map (PEM) of the normal human breast, which can be compared to the PEMs of breast cancer cells in further studies [5]. Normal human luminal and myoepithelial breast cells were used in two-dimensional gel proteome studies. A total of 43,302 proteins were detected across 20 samples, and a master image for each cell type comprising a total of 1,738 unique proteins was derived. Differential analysis detected 170 proteins that were elevated two-fold or more between the cell types—including muscle-specific enzyme isoforms and contractile intermediate filaments as well as a large number of cytokeratin subclasses and isoforms—and 51 of these were annotated by tandem mass spectrometry. A further 134 nondifferentially regulated proteins were also annotated from the two breast cell types.

IV. Sequence Derived Features

Sequence derived features are the various features of protein which are used to predict human protein function. Sequence derived features are very important in protein prediction as these are the input to the HPF predictor as labeled vector. SDF's can be derived from a given set of amino-acid (protein) sequences.

V. SDF Tool

SDF Tool uses equations to calculate the following sequence derived features by taking amino acid sequence as input. Some of the features are described here.

A. Extinction Coefficient ($E_{protein}$)

The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it. The value of the Extinction coefficient can be determined from the composition of Tyrosine, Tryptophan and Cystine.

$$E_{protein} = N_{tyr} * E_{tyr} + N_{trp} * E_{trp} + N_{cys} * E_{cys} \quad (1)$$

where E_{tyr}, E_{trp}, E_{cys} are the Extinction coefficients of the individual amino acid residues.

B. Absorbance (Optical Density)

The absorbance can be determined by the ratio of Extinction coefficient and the molecular weight of the protein.

$$\text{Absorbance} = E_{protein} / \text{Molecular Weight} \quad (2)$$

C. Number of Negatively Charged Residues (N_{neg})

This can be calculated from the composition of Aspartic acid and Glutamic acid.

D. Number of Positively Charged Residues (N_{pos})

This can be calculated from the composition of ARginine and Lysine.

E. Aliphatic Index (AI)

Aliphatic Index can be calculated from the mole percentages of Alanine, Valine, Isoleucine and Leucine.

$$AI = X_{ala} + a * X_{val} + b * (X_{ile} + X_{leu}) \quad (3)$$

where X_{ala} , X_{val} , X_{ile} and X_{leu} are the mole percentages of alanine, valine, isoleucine and leucine respectively. Coefficients a and b are the relative volume of valine side chain and of leu/ile side chains to the side chain of alanine i.e. a = 2.9 and b = 3.9.

VI. Decision Trees

Decision trees are supervised algorithms which recursively partition the data based on its attributes, until some stopping condition is reached [1]. This recursive partitioning gives rise to a tree-like structure. Decision trees are white boxes as the classification rules learned by them can be easily obtained by tracing the path from the root node to each leaf node in the tree.

Decision trees are very efficient even with the large volumes data. This is due to the partitioning nature of the algorithm, each time working on smaller and smaller pieces of the dataset and the fact that they usually only work with simple attribute-value data which is easy to manipulate. The decision tree classifier is one of the possible approaches to multistage decision making. The most important feature of DTC's is their capability to break down a complex decision – making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret [6].

VII. Proposed Model

Determined sequence derived features are then processed in a form suitable for input to classifier by placing them in particular value ranges. The classifier for protein function prediction is the decision tree induction technique that finds features match for the prediction. The working model for the HPF Predictor is shown in Fig. 2.

VIII. Results and Discussion

Sequence derived features are determined from SDF tool and some are obtained using different web based tools. These sequence derived features will be given as input to the HPF predictor. Values of all features are stored in ranges corresponding to each sequence of each class in the database. For the given values of all features, this technique will match the value of each feature with the corresponding ranges in the database to calculate the total matched features for each function as shown in Table I. If the number of matched features of any function is greater than or equal to 75% of number of features given as input, that function will be displayed as result.

TABLE I
Matched Features for Each Protein Function

Serial No.	Features Name	Protein Functions				
		Defensin	Cell Surface Receptor	DNA Repair Protein	Heat Shock Protein	Voltage Gated Channel
1	Nneg	1	1	0	0	0
2	Npos	3	1	0	0	0
3	Exc1	1	1	1	0	1
4	Exc2	1	1	1	0	1
5	Instability Index	1	1	2	2	1
6	Aliphatic Index	2	2	1	1	0
7	GRAVY	1	0	3	0	1
8	Ser	5	1	2	0	1
9	Thr	5	3	3	3	1
10	Tyr	5	3	4	5	1
11	Mean S	4	3	0	0	1
12	D	2	1	0	0	0
13	Probability	5	4	0	0	1
14	ExpAA	4	0	5	5	0
15	PredHel	0	4	0	0	2
Total Features Matched		14	13	9	5	10

References

[1] Clare, A. "*Machine learning and data mining for yeast functional genomics*", Ph.D. thesis, University of Wales, vol. 11, February, 2003, pp. 112-183.

[2] Iddo Friedberg "*Automated Protein Function Prediction-the genomic challenge*" Briefings in Bioinformatics Vol 7. No. 3., January 2006, pp. 225-242.

[3] L. Jenson "Prediction of Protein Function from Sequence Derived Protein Features", Ph.D. thesis, Technical University of Denmark, 2002, pp. 1-570.

[4] D. Krane and RM. Raymer "*Fundamental Concepts of Bioinformatics*", Benjamin Cumming, 2006, pp. 1-203.

[5] Page M. J. Page, B. Amess, R. R. Townsend et al. "*Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties*" Proc. Natl. Acad. Sci. 1999, pp. 12589-12594.

[6] S. R. Safavian and D. Landgrebe (1991) "*A Survey of Decision Tree Classifier Methodology*", IEEE Trans.Systems, Man and Cybernetics, vol. 21, issue 3, May/June, 1991, pp. 660-674.

[7] T. R. Golub, D. K. Slonim, P. Tamayo et al. "*Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*" Science, Vol. 286. no. 5439, 15 October 1999, pp. 531 – 537.

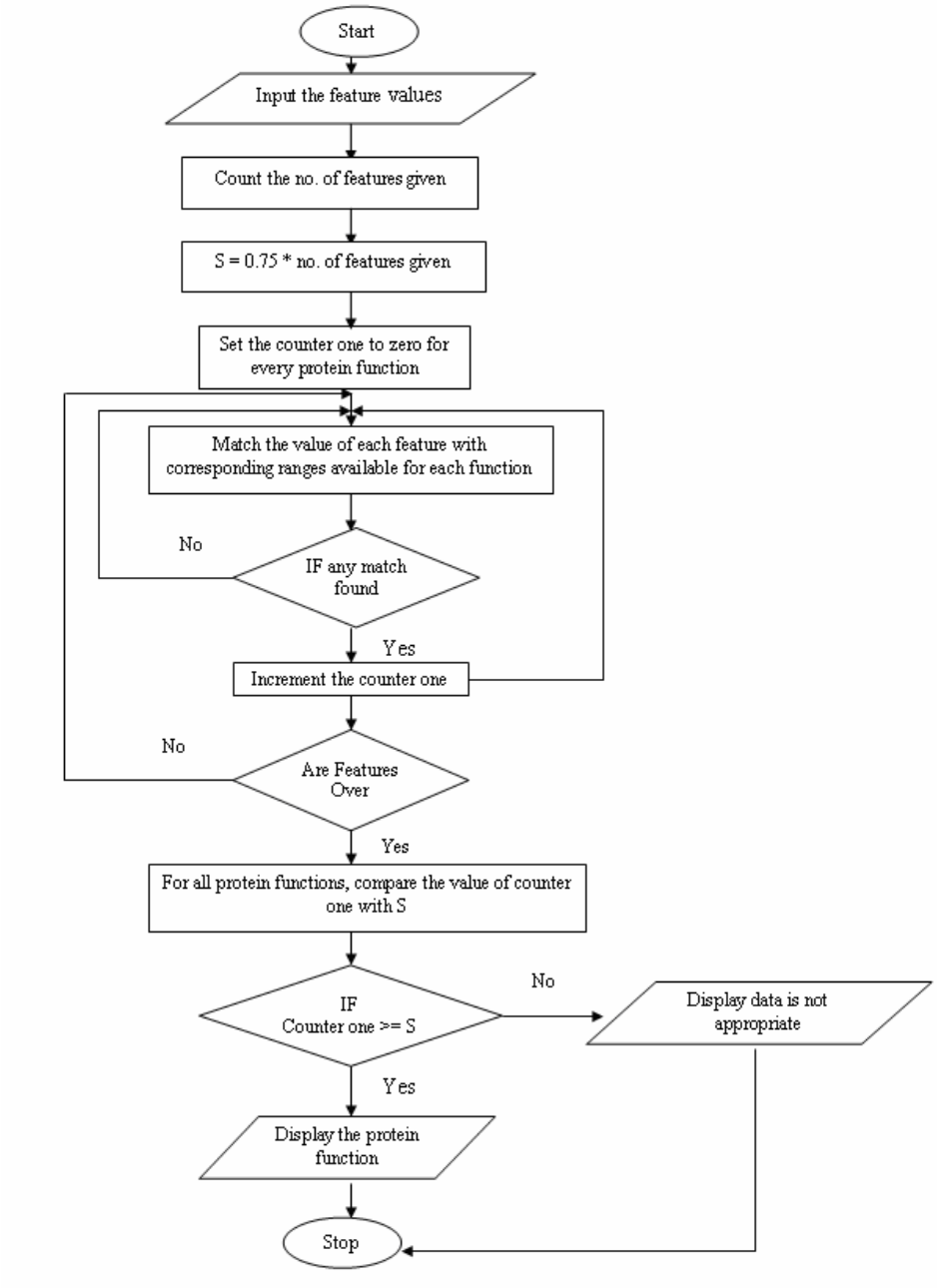

Fig. 2 Working Model for HPF predictor