International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No. 4,October 2012 – www.ijais.org

Cluster Algorithm using Distributed Processing for Human Protein Function Prediction

Manpreet Singh
Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, Punjab, INDIA

Gurvinder Singh, PhD.
Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar, Punjab, INDIA

Karanjeet Singh Kahlon, PhD.
Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar, Punjab, INDIA

ABSTRACT
For the pharmaceutical industry, the discovery of a new drug presents an enormous scientific challenge, and consists essentially in the identification of the target responsible for the disease. Once the therapeutic target is identified, scientists then find one or more leads that interact with the therapeutic target. Usually leads are searched by employing a long and costly process of trial and error. But if the protein class of the target would have been known it will become very easy to find the complementary lead for the responsible molecule. The 50 protein sequences related to 10 different molecular classes are obtained from Human Protein Reference Database (HPRD). Then the Sequence Derived Features (SDFs) for each of the available sequence are obtained using the different online tools. For the whole SDF database, the variation in the values obtained is analyzed and priorities are assigned accordingly. In the present work, priority based cluster algorithm is used for human protein function prediction. Then the distributed processing using four MATLAB workers is applied for different iterations in the algorithm. Two different methods for distributing the code are applied and the cpu times are computed for these methods.

General Terms
Protien Function Prediction, Distributed Processing, Algorithms.

Keywords
Human Protein Function Prediction, Cluster Algorithm, Sequence Derived Features.

1. DISTRIBUTED PROCESSING
Distributed computing is a science which solves a large problem by giving small parts of the problem to many computers to solve and then combining the solutions for the parts into a solution for the problem.

Distributed system is a collection of individual computing devices that can communicate with each other. Distributed computing is often used to refer to the implementation of applications on the distributed memory architectures. In the biomedical research field, it is the most widely used form of parallel processing. Parallel processing describes a computing environment where multiple processors cooperate to solve a given computational problem. Distributed processing implies that the processing will occur on more than one processor in order for a transaction to be completed. In other words, processing is distributed across two or more machines and the processes are most likely not running at the same time, i.e. each process performs part of an application in a sequence. Often the data used in a distributed processing environment is also distributed across platforms. A problem is broken into discrete parts that can be solved concurrently. Each part is further broken down to a series of instructions. Instructions from each part execute simultaneously on different CPUs. In the present work, the algorithm runs concurrently in four lab space. Parallel computing toolbox in MATLAB is used.

2. PROTEINS
Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function.

In an organism, the biological function of a protein cannot be predicted from its chemical function. This is due to the fact that proteins do not function alone. The proteins interact with other proteins and form complexes [1,2].

2.1 Protein Classification
As proteins are responsible for many different functions in the living cell, it is possible to classify proteins on the basis of their functions [3,4,5] as given below:

2.1.1 Enzymes
Enzymes are the proteins that catalyze chemical and biochemical reactions within living cell and outside.

2.1.2 Hormones
Hormones are the proteins those are responsible for the regulation of many processes in organisms.

2.1.3 Transport Proteins
These proteins are used for transporting or storing some other chemical compounds and ions.

2.1.4 Antibodies
Proteins that involved into immune response of the organism to neutralize large foreign molecules, which can be a part of an infection.

2.1.5 Structural Proteins

These proteins are responsible to maintain structures of other biological components, like cells and tissues.

2.1.6 Motor Proteins

These proteins can convert chemical energy into mechanical energy.

2.1.7 Receptors

These proteins are responsible for signal detection and translation into other type of signal.

2.1.8 Signaling Proteins

This group of proteins is involved into signaling translation process.

2.1.9 Storage Proteins

These proteins contain energy, which can be released during metabolism processes in the organism.

2.2 Importance of Protein Classification

The proteins can be divided into various classes on the basis of their functionality. Their importance lies in the process of drug development. An understanding of the classes of proteins is an important component of drug development because proteins are the most common drug targets. Drug development has two major components [2]: Discovery and Testing.

The testing process involves preclinical and clinical trials. The computational methods are not generally subjected to produce significant enhancement in testing processes of drugs.

But in the discovery process the computational methods are very helpful. The drug discovery process is labour intensive and expensive and has provided a fertile ground for bioinformatics research. Bioinformatics promises to reduce the labour associated with this process, allowing drugs to be developed faster and at a lower cost. The drug discovery process itself can be broken into several components as discussed below:

2.2.1 Target Identification

It involves the identification of the target on which the drug acts.

2.2.2 Lead Discovery and Optimization

It involves the docking algorithms i.e. the algorithms that help in determining the lead compounds. A lead compound is the compound in the drug which will bind to the target. The optimization includes the use of the database indexing techniques in the docking algorithm to reduce the number of lead compounds by ruling out those that are highly unlikely to bind to the target.

2.2.3 Toxicology

It involves the study of all the biochemical reactions that will take place when drug is taken.

2.2.4 Pharmacokinetics

It includes the study of the kinetics of the biochemical reactions.

Thus, protein class prediction is necessary for the drug discovery process. The drug discovery process is time consuming and expensive. The process of drug discovery, involves the prediction of protein class based upon existing facts. Sophisticated mining models are needed for protein class prediction. Bioinformatics promises to reduce the labour, time as well as cost associated with this process [6].

2.3 Protein Function Prediction Techniques

The computational techniques for predicting the structure and functions of unknown proteins are as follows:

2.3.1 QM/MM Scheme

Quantum Mechanical/Molecular Mechanical scheme is used by software named GAMESS (General Atomic and Molecular Electronic Structure System) to predict an unknown protein. It requires a large computer memory to perform mathematical calculations and it runs on Linux operating system.

2.3.2 SWISS Model

Software named as SWISS-Model is available for automated building of the theoretical structural models of a given protein (amino-acid sequence) based on the known proteins' structures.

2.3.3 Classifiers

Classifiers, for example, neural networks, decision trees etc. learn classification rules from the given training data which are used to predict functions of unknown proteins [7].

The support vector machines and neural network are considered as black box models as their working is invisible to the researcher [8, 9]. On the other hands, decision trees and rule sets are white box models [10-13].

3. PROBLEM STATEMENT AND DATA COLLECTION

Place Class prediction of a protein facilitates to enhance the process of drug discovery. In drug discovery, it is very problematic to find out the complementary protein for each protein individually. But if the class of the protein will be known for which the drug is to be discovered then it will become very smooth to find the complementary protein sequence which can be attached to the active site of the protein to stop it to expand.

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.

Sequence derived features are utilized to predict the class of protein. The drug discoverer provides unknown protein sequence based on which he wants to predict the class of protein. SDFs can be derived by using web-based tools. These SDFs are then processed into a suitable format that can be utilized to predict class [6]. The model will be tested based on the available classes and existing protein sequences. The data related to human protein is accessed from Human Protein Reference Database (HPRD). From HPRD 10 protein functions are considered and 5 amino acid sequences are extracted for each protein function. These amino acid sequences are then given as input to various web based bioinformatics tools which further provide sequence derived

features as output [14]. It includes approximately 163 classes of protein functions. From HPRD, the sequences related to ten molecular classes are obtained. These classes are: Defensin (Def), Heat Shock Protein (HSP), Voltage Gated Channel (VGC), Cell Surface Receptor (CSR), DNA Repair Protein (DRP), Amino peptidase (Ami), Decarboxylase (Dec), G-Protein (GP), RNA Binding Protein (RBP) and Transport/Cargo Protein (T/CP). For each of the molecular class five amino acid sequences are obtained.

The various bioinformatics tools and their respective SDFs are shown in table 1[7, 15].

Table 1. SDFs' obtained from web based tools

Tool used	SDFs' Obtained
ExPASy ProtParam	Extinction Coefficients
	No. of negatively charged residues
	No. of positively charged residues
	Instability Index
	Aliphatic Index
	GRAVY
PSORT	Protein localization sites
NetOGlyc	O-Glycosylation sites
NetNGlyc	N-Glycosylation
SignalP	Signal Peptide
TMHMM	Transmembrane Helices

Values of the SDFs for each of the protein class are stored for each class [7]. The algorithm for Protein Class Predictor using Distributed Environment exemplifies the prediction of protein class by setting number of sequence, number of SDF, credit of sequence, score of SDF and range of SDF with the help of counters in distributed environment. For allocating score to each of the SDF, the variation in the values of each of the SDFs is studied for each functional class. High priority features are the large variation of the value of particular SDF from class to class. Low priority features are the less variation of the value of SDF from class to class. On the basis of priority, score is allocated to each SDF. This score is utilized to find the highest score sequence to predict the class of the entered sequence.

4. ALGORITHM

The algorithm for protein class predictor is implemented for seventeen sequence derived features and ten protein classes. Five amino acid sequences are considered in each class. The computation is enhanced by identifying the different areas in the code where the distributed processing can be implemented. Then the distributed processing is embedded in those areas having four worker labs. The algorithm for Protein Class Predictor in Distributed Environment is written here:

Begin Algorithm (Number of sequences M, Number of SDFs N, Credit of Sequence C, Score of SDF S, Priority of SDF P, Priority of previous cluster Pprev, Current maximum credit Maxc, Previous maximum credit Maxp, Maximum possible credit Mp, Range of SDF R, Counters i, j and k)

Step 1: Start
Step 2: Enter features Fj
Step 3: Set j =1, Ci =0
Step 4: Set i = 1
Step 5: Check the value of Aij
Step 6: If Fj+Rj/2>=Aij>=Fj-Rj/2 then set Ci=Ci+Sj otherwise check the value of i.
Step 7: If ith sequence is in any previous cluster, then check If Pprev>Pj and repeat step 8 to step 15, otherwise enter ith sequence in jth cluster
Step 8: If Pprev>Pj, then check the value of i, otherwise Shift ith sequence from previous to jth cluster and check the value of i.
(*Distributed Processing is applied in this loop by distributing the loop computation in four labs*)
Step 9: If i=M, then check If j = N? And repeat step 10 to step 15, otherwise set i=i+1.
(*Distributed Processing is applied in this loop by distributing the loop computation in four labs*)
Step 10: If j = N?, then Set K = N (Highest Priority Cluster) and Maxc=0, Maxp=0, otherwise j=j+1.
(*Distributed Processing is applied in this loop by distributing the loop computation in four labs*)
Step 11: Set Maxc = MaxK (Maximum credit in kth cluster).
Step 12: If Maxp>Maxc, then Maxc=Maxp, otherwise check whether Maxc=MpK? And repeat step 13 to step 15.
(*Distributed Processing is applied in this loop by distributing the loop computation in four labs*)
Step 13: If Maxc=MpK, then display the Maxc sequence class, otherwise check the value of K.
(*Distributed Processing is applied in this loop by distributing the loop computation in four labs*)
Step 14: If K=1 (Lowest priority cluster), then display the Maxc sequence class, otherwise set K = K-1 and Maxp=Maxc.
Step 15: End

For the given values of all input 17 features, this technique will match the value each of the entered feature with the respective value of all the sequences in the database and if the value of a sequence will be in the specified range of the entered value then that sequence will be included in the cluster of that feature. Every time a sequence enters a cluster, its credit will be incremented by the score of that feature. So in this manner clusters for all the SDFs will be generated.

Starting with the highest priority cluster, sequence with the maximum credit will be determined from the current cluster. If the maximum credit obtained from the current cluster will be greater than the previous cluster then the new sequence will become the maximum credit sequence otherwise previous sequence will remain the highest credit sequence. While traveling from higher priority clusters to lower priority, in each of the cluster if the maximum score of the current credit will be greater than all the previous clusters and also equal to the maximum possible credit of that cluster then the sequence with this credit will be the maximum credit sequence in the whole database and need not to travel further. Then the class of the sequence with maximum credit will be included in the prediction result. After implementing the algorithm, its complexity is studied and the points are located, where the distributed processing has to be applied. Then the distributed processing is embellished in target areas of the data mining algorithm. For transmutation of the algorithm in parallel mode, the algorithm needs to be divided up in labs. Two methods called parfor and batch are used to distribute the cluster algorithm under consideration. In parfor, the loop iterations are distributed to different systems and in batch, the distribution is done automatically locating the different distribution points.

5. RESULTS AND DISCUSSIONS

The input values of sequence derived features are listed in table 2 [7] and protein class obtained for this data is Heat Shock Protein. The computation is distributed in four matlab workers and the performance is measured the help of profiler

which displays the cpu time for parfor and batch methods as shown in figure 1 and figure 2 respectively. It is evident that parfor takes lesser cpu as compared to batch and hence gives the better result.

Table 2. Input SDF values to class predictor

S. No.	Features Name	Values Given
1	Nneg	*28*
2	Npos	*37*
3	Exc1	*50545*
4	Exc2	*50420*
5	Instability Index	*47.19*
6	Aliphatic Index	*73.55*
7	GRAVY	*-0.488*
8	S	*1*
9	T	*2*
10	Ser	*15*
11	Thr	*4*
12	Tyr	*2*
13	Mean S	*0.078*
14	D	*0.067*
15	Probability	*0.000*
16	ExpAA	*0.03*
17	PredHel	*0*

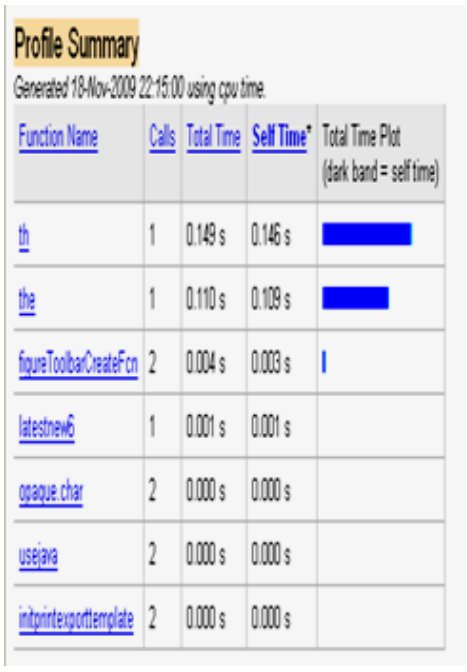

Fig. 1: CPU time for parfor method

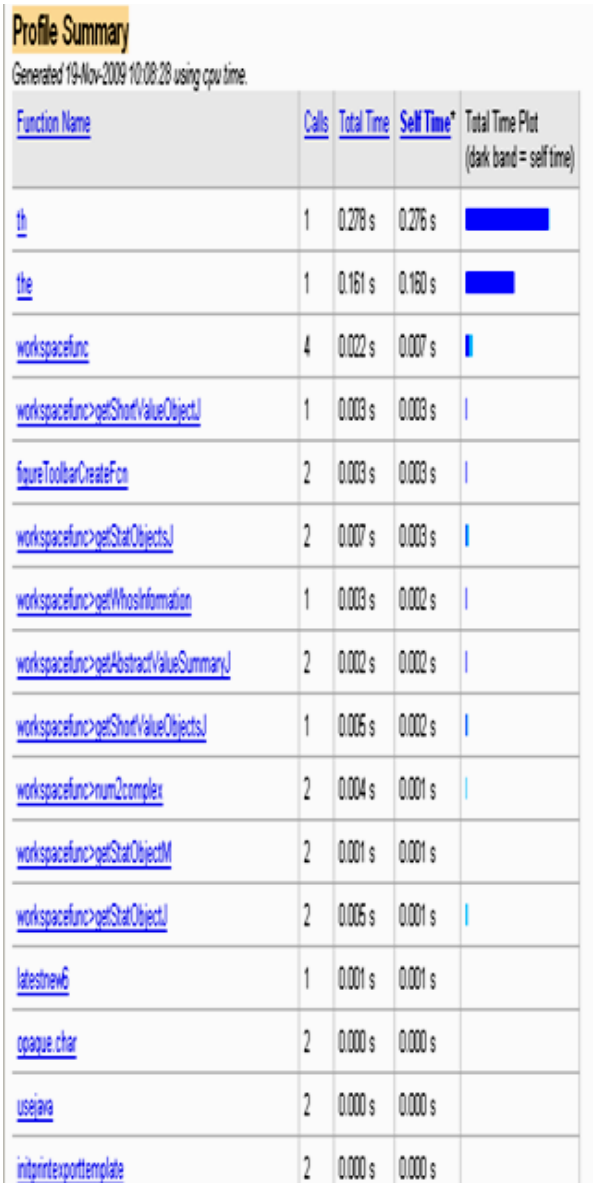

Fig. 2: CPU time for batch method

6. CONCLUSION AND FUTURE SCOPE

As the database for the known protein sequence is very large, the sequence derived features will also form a large database when more number of protein classes are included. Thus for a large database, which ensures accurate prediction, the distributed processing will also ensure minimum time. Thus the SDF database can be increased to achieve the better results. The distributed processing can be also applied to different other classifiers available for HPF prediction.

7. REFERENCES

[1] I. Friedberg, "Automated Protein Function Prediction-the genomic challenge Briefings in Bioinformatics", vol. 7, No. 3, January 2006, pp. 225-242.

[2] Krane, D. and Raymer, M. 2006 Fundamental Concepts of Bioinformatics. Pearson Education Publishers.

[3] N. Narai, E.D. Kolaczyk, S. Kasif, "Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data", PLoS ONE 2(3), issue 3, 2007, pp. 1-7.

[4] Rastogi, S.C., Mendiratta, M. and Rastogi, P. 2005 Bioinformatics Methods and Applications, 3rd edition. PHI publication.

[5] http://proteincrystallography.org/protein/

[6] M. Singh, P. Singh, P.K. Wadhwa, "Human Protein Function Prediction using Decision Tree Induction", International Journal of Computer Science and Network Security, vol. 7, No. 4, 2007, pp. 92-98.

[7] M. Singh, G. Singh "Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction", International Journal of Computer Applications, vol. 20, no.3, 2011, pp. 22-27.

[8] L.J. Jensen, R. Gupta, H.H. Staerfeldt, and S. Brunak, "Prediction of Human Protein Function According to Gene Ontology Categories", Bioinformatics, vol. 19, no. 5, 2003, pp. 635-642.

[9] W.R. Weinert and H.S. Lopes, "Neural Networks for Protein Classification", Applied Bioinformatics, vol. 3, no. 1, 2004, pp. 41-48.

[10] A. Clare, A. Karwath, H. Ougham, and R.D. King, "Functional Bioinformatics for Arabidopsis thaliana", Bioinformatics, vol. 22, no. 9, pp. 1130-1136, 2006.

[11] J. He, H.-J. Hu, R. Harrison, P.C. Tai, and Y. Pan, "Transmembrane Segments Prediction and Understanding Using Support Vector Machine and Decision Tree", Expert Systems with Applications, vol. 30, 2006, pp. 64-72.

[12] G.L. Pappa, A.J. Baines, and A.A. Freitas, "Predicting Post-Synaptic Activity in Proteins with Data Mining", Bioinformatics, vol. 21, no. Suppl. 2, 2005, pp. ii19-ii25.

[13] Singh, M., Sandhu, P.S. Singh, H. 2006. Decision Tree Classifier for Human Protein Function Prediction. In Proceedings of International Conference on Advanced Computing and Communications, ADCOM 2006, 20-23 Dec., 2006, pp. 564-568.

[14] Jensen, L. 2002. Prediction of Protein Function from Sequence Derived Protein Features. Ph.D. thesis. Technical University of Denmark.

[15] M. Singh, G. Singh, K. S. Kahlon, "Classifier for Human Protein Function Class Prediction", International Journal of Engineering & Information Technology, vol. 1, No. 1, 2009, pp. 1-4.