Data Optimization from a Multiple Species Network Using Modified SFLA(Shuffled Frog Leaping Algorithm)

N.Kannaiya Raja, M.E., (P.hd) .,A.P[1], Dr. K.Arulanandam, Prof & Head[2], S.K. Sugunedham, (M.E)[3]

1CSE Dept. Arulmigu Meenakshi Amman College of Engg
Thiruvannamalai Dt, Near Kanchipuram ,India

[2]CSE Department Ganadipathy Tulsi's Jain Engineering
College, Vellore , India

[3]CSE Dept.Arulmigu Meenakshi Amman College of Engg ,
Thiruvannamalai Dt, Near Kanchipuram ,India

Abstract--In this work we present a novel approach that uses interspecies sequences homology to connect the networks of multi species and possible more species and possible more species together with gene ontology dependencies in order to improve protein classification for research work. Proteins are involved in many for all biological process such energy metabolism, signal transduction and translation initiation. Even though for a large portion of proteins and their biological function are unknown or incomplete, therefore constructing efficient and reliable models for predicting the protein function has to be used in research work. Our method readily extends to multi species food and produce the improvements similar to them multi species. In the presence of multi interacting networks are using data mining for integration of a data from various sources and contributing increased accuracy of the function prediction of the multiple species for research work. We further enhance our model to account for the gene ontology dependencies by linking multiple related ontology categories such as, we have selected the food items from various countries such as from America the famous food items of yoghurt and Australia food items of oats and Indian food items of soya bean. The data sets are highly desirable for this use from various countries using logical networks from center for bioinformatics research institute (Chennai) and stored in the mining.SFLA aims to set a generic paradism of the efficient mining that acquire the data set of proteins for these food items and promotes predictions of protein functions with gene ontology for research work.

Keywords— Biology and genetics, machine learning, bioinformatics (genome or protein) databases.

1. Introduction

Protein-protein interactions (PPIs) can reveal insights on biological regulatory path ways and metabolic processes. a complete and reliable protein interaction map provides us with an opportunity to understand the basic biological processes within a cell. Every node have protein-protein networks, which represents protein between nodes should represent a protein and edges between nodes represents different types of function has been assigned, such as protein-protein interactions, sequence similarity, co-expression patterns and others. in biological research, various experiments have been developed to map the interactions among the proteins, such as mass spectrometry, yeast two-hybrid tandem affinity purification and co-immuno prediction and chain graph probabilistic approach [1][2][3]. The computational methods for protein classification depend on nearest neighbors in a protein-protein network, can share related functions. These methods are appointed the functions to a protein of interest based on the annotation of its neighbors. Those approaches have accomplishment in cases where protein have multiple, [4][5][6]mostly annotated neighbors. On the other hand the method shows much less success on protein with inadequate neighborhoods those proteins isolated in their own network or the ones surrounded by weakly annotated neighbors. In this paper, we propose a new approach to protein function prediction which trounce the drawbacks of these method and illustrates interspecies sequences homology to connect the networks of multi species and possible more species together with multifunctional gene ontology (GO) dependencies. The primary conceptual improvement of our method is to connect protein-protein networks of two or more different food items,

but related multiple species, into a single-computational model. During the edges of high homology, proteins can increase their learning neighborhood and acquire extra functional information from their neighbors-homolog's of a different food items networks.

Our new and promising technique, the modified shuffled frog leaping algorithm for classifying the protein functions from various sources with the use of interspecies sequence similarity and integrate multiple sources of information. In connecting networks we rely on the fact that proteins of different food items from various countries, which share high sequence similarity, are likely to share similar protein classification. Thus, high similarity of sequences between species is likely to lead to shared functions. Even though the resulting shuffled frog leaping algorithm used the data mining with leave-one-out cross validation is used to evaluate classification accuracy when applied to the datasets, compared all multi species protein functions. We demonstrate that the combined models often lead to efficient implementations and significant improvements in predictive accuracy not experimental in isolated networks or other competing approaches.
Additionally we propose an efficient evolutionary approach for selecting protein subsets from stored mining data's, that effectively achieves higher accuracy for classification problems. The purpose of classification is to build an efficient model for predicting protein functions [6]. We embedded a genetic algorithm in a shuffled frog leaping algorithm and to serve as a local optimizer for each generation in implementing feature selection. To evaluate the SFLA and GA results showed that our proposal achieves superior classification accuracy when applied to the data sets. The SFLA is derived from a virtual population of protein functions in which individual proteins are equivalent to the GA chromosomes, and represent a set of solutions. Each protein is distributed to a different subset of the whole population called a memeplex. An independent local search is conducted for each protein memplex, in what is called memeplex evolution. After a defined number of memetic evolutionary steps, protein functions are shuffled among memeplexes [10][11], enabling proteins to interchange messages among different memplexes and ensure that they move to an optimal position. Local search and shuffling continue until defined convergence criteria are met. SFLA have demonstrated effectiveness in a number of global optimization problems difficult to solve using other methods.

The remaining portion of the paper as follows: in section 2, we first present a general view of closely related network approaches to protein function prediction. then we goes in section 3, the modified shuffled frog leaping algorithm that combines both go structure and the information from protein-protein networks of multiple species. section4 illustrates the effectiveness of the proposed approach when implemented to yoghurt, oats and soya beans networks, at different granularities of the go(gene ontology). We discuss the new results and analysis in section5. Finally in section 6 conclusion and future enhancement.

2. Related Work

Proteins and are used in many biological process such as energy gains and DNA-rna metabolism, signal transduction and translation, initialization. However the large portion of protein and their biological function cannot be used in maximum in biological research work. Therefore we have proposed constructing efficient and reliable model for predicting protein function which has the task of our model. For this purpose we have selected three food items for prediction of protein functions[1][2]. The computational biology described a enormous methods for understanding the importance of protein-protein interactions. In these methods, supervised learning is a leading approach. The state-of-the-art supervised learning methods include k-nearest neighbor, support vector machines (SVMS),[4] Random forest and so on. Using learning classifier positive examples of truly interacting protein pairs and negative examples of non-interacting protein pairs, to predict a not noticed relationship between two proteins. Each protein pair is fixed as a feature vector in the data. This attempt has been tried on developing informative and effective feature representation methods for PPI prediction. Feature vectors may be extracted based on protein sequences directly or may involve indirect evidences, including domain compositions, motif pairs and related mRNA expression. Compositions of amino acids and physiochemical descriptors based SVM method is used by bock and gouge. Urquiza et al. extracted 26 genomic or proteomic features of yeast from various databases for each pair, such as information of protein domains, domain-domain interactions in proteins whose 3D-structures are known and high quality annotation of gene ontology. Espadaleret al. considered protein structural similarities among domains found in the databases of interacting proteins, combining maintenance of pairs of sequence patches based on the observation that structural evidence has shown that usually interacting pairs of close homolog's physically interact in the same way. Numerous methods are facilitate to concluded protein interactions based on the conservation of gene neighborhood, conservation of gene order, gene fusion events, and the co-evolution of interacting protein pair sequences. The valuable facts of protein interaction networks across organisms. A sensitive idea is to utilize the interaction map of one organism as a pattern to predict interactions in another . The 'interaction domain profile pairs' method (IDPP) applied to protein interaction datasets of Escherichia coli and Helicobater pylori by Wojcik and Schachter. The IDPP method required a high quality protein interaction map to be real world, protein interaction datasets are frequently sparse,

incomplete and deafening. The researchers in machine learning develop a matrix factorization based methods for reassign learning. An important application is recommendation systems, where collaborative filtering is modeled using matrix factorization, moreover countenance the problem of network sparsity. We are interested in predicting links between nodes, which corresponds to binary connectivity, where the proteins have different semantics from users and products. The family probabilistic graphical models utilized by computational biology for protein function prediction, such as hopeful networks, to conclude the functions datasets of incomplete annotated proteins. With the use of partial knowledge of functional observation, probabilistic assumption is to determine other proteins unknown functions of momentary on and accumulating undecided information on a huge sets of related proteins. The preference of functional association between proteins of the network models performance impacts by a critical factor. The most traditional methods are based on sequence similarly using blast. The large set of method relies on the truth that similar proteins are likely to share common functions, sub cellular location, or protein- protein interactions. The same kind of similarity based methods take account of sequence homology similarity in short signaling motifs, amino acid composition and expression data. The PPI data to established protein function within a own network. For example Markov random filed method is to define the complete set of proteins involved in PPI[17][18]. These methods are based on the concept that interacting neighbors in networks might also share function. The approach of incorporating gene ontology structure into probabilistic graphical models has gave the possible results for predicting protein functions. The approach considers multiple functional discipline in the gene ontology (GO) simultaneously. In this model, each protein is represented by its own annotation space - the go structure. The information is distributed within the ontology structure as well as between neighboring proteins, leading to an added ability of the model to explain potentially tentative single term predictions. Markov random field (MRF)[12] model for protein function prediction using protein-protein interaction data. Two main features differentiate the MRF based methods from other guilt-by-association methods. One is that the MRF model uses global information on the whole interaction network as a replacement for of local interaction network. the method was applied to predict protein function based on "cellular role" using protein functions defined in yeast proteome database (YPD)[14][15]. The results showed that the MRF based method outperforms the two guilt-by-association based methods. Features of individual proteins have long been used for protein function prediction. A feature here refers to an surveillance about a protein. It can be the presence or absence of a motif signal, its isoelectric point, its absolute mRNA expression level, or mutant phenotypes from experiments about the sensitivity or resistance of disruption mutants under various growth conditions. features were used for protein function prediction as pattern recognition problems. Drawid and Gerstein developed a general Bayesian approach to predict protein localization based on a large number of features of individual protein. we broaden the MRF[10][11] based method to an integrated approach including other protein pair wise relationship such as correlations of gene expression patterns, genetic interactions, and features of individual proteins such as domain information. The model is bendable that other protein pair wise relationship information such as pair wise protein sequence similarities and features of individual proteins can be easily incorporated. The integrated approach to predict functions of yeast proteins based on MIPS protein functions and the interaction networks based on MIPS physical and genetic interactions, gene expression profiles, tandem affinity purification (tap) protein complex data, and protein domain information. We study the sensitivity and specificity of the integrated approach using different sources of information by the leave-one-out approach. Compared to using MIPS physical interactions only, the integrated approach combining all the information increases the sensitivity from 57% to 87% when the specificity is set at 57%, an increase of 30%. it should also be noted that by enlarging the interaction network, the number of proteins whose functions can be predicted is also greatly increased. Many learning approaches rely on information available from neighbors in a protein network. however, two largest protein networks of yeast and fly as well as joint yeast-fly network. predictive performance[2][3][4] of network models is evaluated in a five-cross-validation setting. the experiment set consists of a 20 percent same proportion of negatively and positively annotated proteins, as the left behind 80 percent of the data used for training the model. For each arbitrarily chosen test protein, all of its annotations are left out—the gene ontology structure remains in place but the functions at all terms now listed as unknown. in the case of a joint fly-yeast network, we eradicate annotations of 20 percent of annotated proteins from each network. in the testing phase, upon junction of the message-passing process, predictions at terms whose annotations were left out are tested against the known eliminated annotations. We conducted a 10 experimental rounds using the random splitting process for each tested network, and compared results of runs on single networks (without joining) to that of the joint network. individual and joint networks are skilled and evaluated on the same training/testing data. normalized BLAST scores measures the intra and interspecies similarity, and its divided by self score of query (i.e., blast score of the homolog divided by the blast score of the protein against itself), ranging from 0 to 1.From the saccharomyces Genome database for yeast and fly base for fly we got sequence and annotation data. Same as protein-protein interaction data were obtained from biogrid database. This evidence resulted in a combined se of fly and yeast proteins that were used to construct the joint hopeful networks.

Gene ontology structure downloaded from gene ontology databases. while analysis gene ontology annotation, we consider two basic go assumptions: GO hierarchy is expanded up for positively annotated proteins (if a protein is positively annotated to a term, then it is also positively annotated to all of its parents/ancestors) and is expanded down for negatively annotated proteins (if a protein is negatively annotated to a term, then it is negatively annotated to all of its children/ descendants). We construct a negative set relying on co-annotation (co-occurence) statistics of GO annotations in the data (further maintaining two fundamental go assumptions). In exacting, a protein is considered negatively annotated to a definite GO term if this term has never been observed to co-occur with a known function for this protein in the training data. This method can be applied to the entire gene ontology, at the cost of time and space complexity. Though, in practice, biologists and clinicians are interested in precise, relatively small, subontologies, targeted in our study.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining technique is one of a number of analytical tools for analyzing data. it allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. While large-scale information technology has been evolving separate[14] transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

Data mining consists of five major elements:
1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

TECHNIQUES OF DATA MINING

Data mining involves following four techniques:

Clustering - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.[14][15]

Classification - is the task of generalizing known structure to apply to new data. for example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naïve bayesian classification,neural networks and support vector machines.

Regression - attempts to find a function which models the data with the least error.

Association rule learning - searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket Analysis

3. Method

Multi species method is to join networks of two (or more) species by edges of high sequence similarity into one computational model, as shown in fig. 1. In exacting, an edge is introduced between homologous proteins in two species if their normalized blast score is above 0.5 (the similarity is high). On the other hand, interspecies edges are not introduced [1][2]when the score is below 0.5 (the similarity is low), since dissimilar proteins may or may not be involved in the same biological process. Furthermore, most of the protein pairs would share some low similarity, which would obscure the network with potentially irrelevant low similarity edges. We take a three-species setting, and we classify a similarity measure between protein i in yoghurt network and protein y in oats network, and protein z in soya bean network at term c, as $s_{i,y,c}^{between}$ $s_{i,z,c}^{between}$ a normalized pair wise blast score. Consequently, the potential function for homolog's between different species is defined as if the relationships between children and parents are directional, the protein function is positively annotated to a child as well as positively annotated to a parent. The reverse relationships does not hold but the parent have negatively annotated protein function, it will be negatively annotated to all children terms. By the definition clearly known where the presence of multiple parents in negative state,it immediately yields a negative state of child.

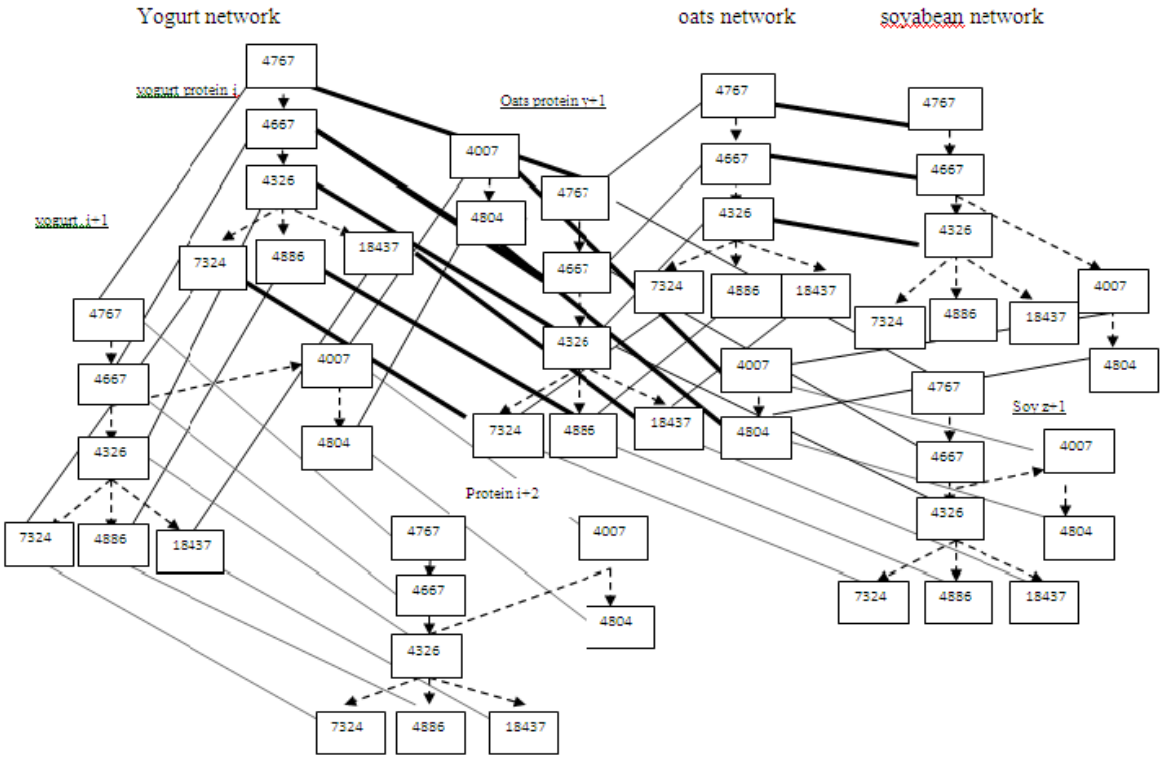
$$\omega(+,+) = \omega(-,-) = s_{i,y,c}^{between}$$

$$\omega(+,-) = \omega(-,+) = 1 - s_{i,y,c}^{between}$$

$$\omega(+,+) = \omega(-,-) = s_{i,z,c}^{between}$$

$$\omega(+,-) = \omega(-,+) = 1 - s_{i,z,c}^{between}$$

bond angles ω, the creation of these structures neutralizes the polar groups on each amino acid. the secondary structures are tightly packed in the protein core in a hydrophobic environment. each amino acid side group has a some degree of volume to occupy and a limited number of possible interactions with other near- by side chains, a situation that must be taken into account in molecular modeling and alignments. While this assumption maybe open to debate, it is shown to lead to improved annotation performance. Considering heterogeneous values of similarity $s_{i,y}^{between}$ at each term c may lead to additional improvements, at a cost of a more complex and demanding parameter estimation process. The combined model for joint yoghurt-oats-soya bean (referred to as species 1 and 2) network now defines a joint Gibbs distribution of functional term annotations over a set of all proteins in the shuffled frog leaping, detailed in (2). Here, z is the normalizing constant, ω_{within} is similarity measure within one species network, $\omega_{between}$ is a similarity measure between multiple species network. After the joint network is built, the confidence broadcast is used to compose predictions at all ontology terms in both species. We consider a state of convergence and decision thresholds. Adding interspecies homology information into the learning model has exclusive advantages and shows noteworthy Improvements in protein function prediction.The model is specifically beneficial for proteins isolated in their own networks (having no interaction neighbors) or for proteins which are surrounded by poorly annotated neighbors. in a multispecies setting, the neighborhood of such proteins is expanded so that they can learn their functional annotations from their homologs in the different species



Basic SFLA

The SFLA has four main stages: generating initial population, partition, local search and shuffling.

Generating initial population

Generate p frogs as initial population arbitrarily. Each Frog (correctly, the position of a frog) represents a sufficient solution of the problem, Frog i is denoted as $x_{i=}x_{i1}, x_{i2}, \ldots\ldots x_{is}$ where S is the space dimension of a solution. later than that, calculate the fitness for each frog.

Partition

The Frogs are sorted in descending order according to their fitness. then partition the entire population into m memeplexes (subsets), each one containing n frogs. in this scheme, put the first memeplex, the mth frog into the mth memeplex, and the (m +1) th frog back into the first memeplex, etc. finally, all the frogs are put into m different memeplexes and p (m×n) . Here, the different memeplexes symbolize different cultures of the specified frogs.

Local search

The local search process improves the average fitness of the memeplexes by serving the frogs with the worst fitness in the subsets communicate with the frogs by way of local, even global best fitness. the upgrading process is described as follows. to a given memeplex, find out the frogs with the best and the worst fitness and categorize them as x_b and x_w respectively. Also find out the frog with the global best fitness within the whole population and identify it as x_g . Then apply a process similar to particle swarm optimization to improve only the frog with the worst fitness (i.e. x_w) in each cycle. Accordingly, the position of the frog with the worst fitness is adjusted as follows:

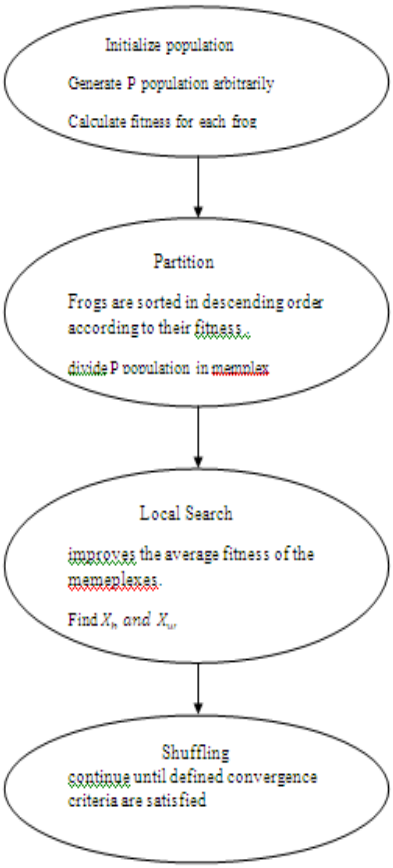


Figure2 Stages in SFLA

where rand() is a random number between 0 and 1; d_{max} is the maximum allowed change in a frog's position. If the fitness of new x_w is better than that of current x_w , then replace current x_w with new x_w . Otherwise, replace x_b with x_g and repeat the calculation according to (1) and (2). If no improvement is made in this case yet, then randomly generate a new frog to replace the current x_w . After the memetic change for the current frog in the memeplex, sort the frogs of the memeplex in order of decreasing performance value and repeat local iteration shuffling.

The next step of memetic evolution steps within each memeplex, ideas are agreed among memeplexes in a shuffling process which helps to increase the quality of the memes after being infected by frogs from different cultures[6]. The partition, the local search and the shuffling processes continue until defined convergence criteria are satisfied. Usually, the convergence criteria can be defined as follows:

The qualified change in the fitness of the best frog within a number of successive shuffling iterations is less than a pre-specified tolerance. The main parameters of SFLA are: number of total frogs, number of memeplexes, number of generation for each memeplex before shuffling, number of iteration, and maximum step size.

procedure msfl

Randomly initialize population of *p* frogs;
for *g*=1 to *gshuff*
begin
calculate the fitness of all frogs;
sort all frogs in order of descending fitness and do the memeplex
division according to (1);
for k=1 to m do
begin
for *j*=1 to *gmeme* do
begin
set *xbk* ,*xwk* and *xg*;
do local evolution according to (4)(5);
if the evolution can't produce a better solution,
(4)(5) are repeated
but with *xbk* replaced by *xg* ;
else if still no improvement, randomly generate a new solution;
resort the *k*-th memeplex in order of decreasing fitness;
end;
end;
re-shuffling the entire frogs;
end

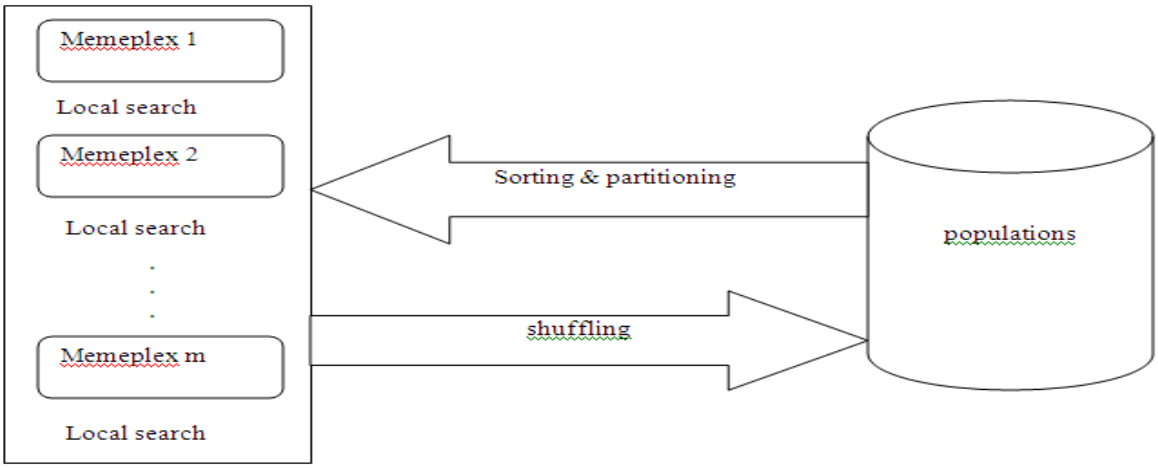


Figure 3 Construction of MSFLA

4. Experiments and Results

As a new and hopeful technique, the Modified shuffled frog leaping algorithm for multi species networks . We apply our method to three largest joint protein networks of yogurt, oats and soya bean , for optimizing the data from multispecies networks. Initially we discussed the two search techniques: the local search of the "particle swarm optimization" technique and the competitiveness mixing of information of the "shuffled complex evolution" technique . with each other. So the frogs stand for the protein functions here. Frog i can be written as P_i =($p_{i1}, p_{i2}, \ldots\ldots\ldots p_{iS}$) where s is total number of food items. The decision variable F_{ij} (j=1,2,…….s) is the number equivalent to a food items. Here we took three food items so, s=3 then calculating the fitness function ,it evaluate the protein's position and returns a single numerical value, and higher return value, the better the protein The SFLA is a heuristic search algorithm.Experiments was carried out by using different numbers of food items from various countries from America the food item of yogurt. Australia food item of oats and Indian food item of soya bean. In the joint yogurt-oats-soya bean networks the best solution of protein function is to be selected for protein through interaction together function.

Now a protein function starts to find the nearest neighbor for making strong link between each network. According to this simple near neighborhood rule, we can generate S food items as part of intial population, in which the first resolution variable is 1,or 2 or S. The second variable is the number which represents the closest protein function to the first protein. The third variable is the number which represents the nearest protein function among the protein networks has not been connected and next to second protein node.

The protein functions are sorted in descending order according to their fitness and then partitioned into subsets called as memeplexes (m). Within each memeplex, the frog with worst and best fitness are identified as x_b , x_w protein function with global best fitnees is identified as x_g the protein function with worst fitness is improved according to above equation. We want to calculate intra- and interspecies similarity, used BLAST scores, and defined as a BLAST score divided by character score of protein functions, its score of the protein against itself ranging from 0 to 1. We obtained sequence and annotation data from NCBI genome database [3] for yogurt, oats and soya-bean. NCBI provides various levels of computation, analysis, and Curation as needed per organism. Blast results, as well as the sequence features, are readily displayed on NCBI'S map viewer. Protein-Protein interaction data were obtained from string [7] database. Network visualization of the protein interactome where each point represents a protein function and each line between them is an interaction. This resulted in a combined set of 13,200 yogurt,5,004 oats and 6,008 soyabean proteins that were used to construct our joint belief networks. To generate the database to mine we formed a single deductive database of genes and their known functional assignments.

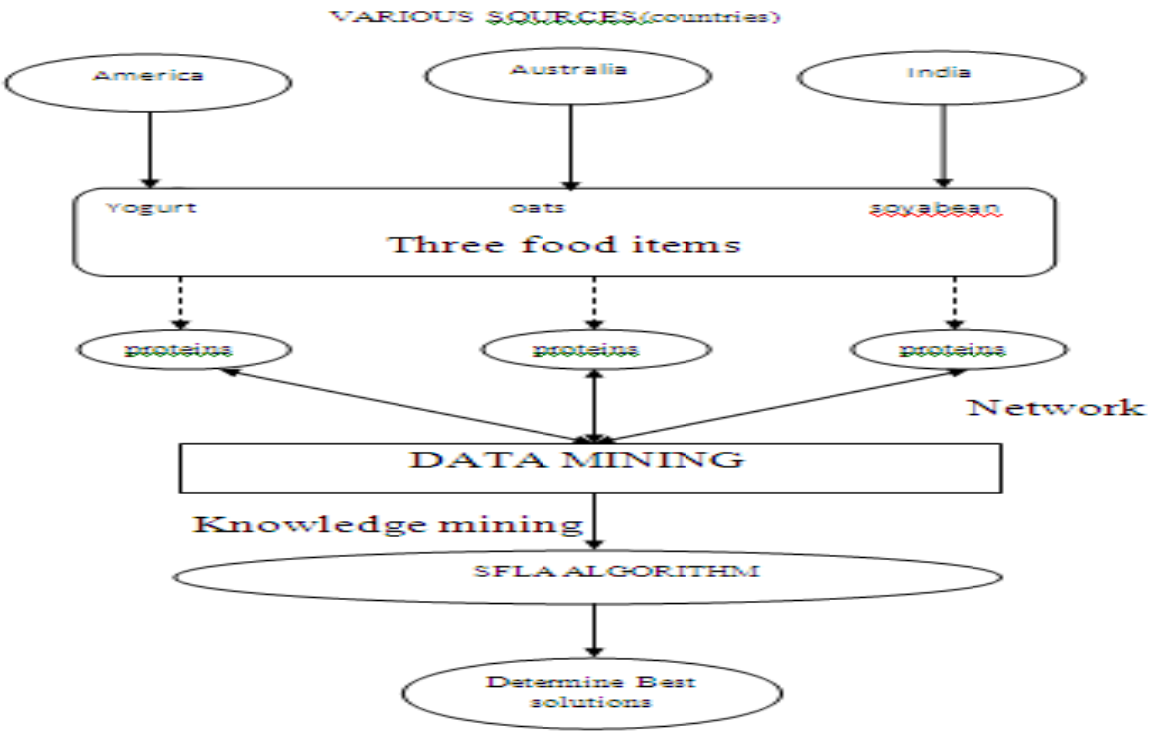


Figure 4 Architecture Diagram

We then processed this data to form sequence descriptions of the genes. the most commonly used technique to gain information about a sequence is to run a sequence similarity search, and this was used as the starting point in forming descriptions. The most commonly used technique to gain information about a sequence is to run a sequence similarity search, and this was used as the starting point in forming descriptions. Such rules provide a way of predicting function in the absence of recognizable sequence homology. The data mining approach described is extendable to analysis of other forms of bio informatics data, such as expression profiles, pathway analysis,structural studies.

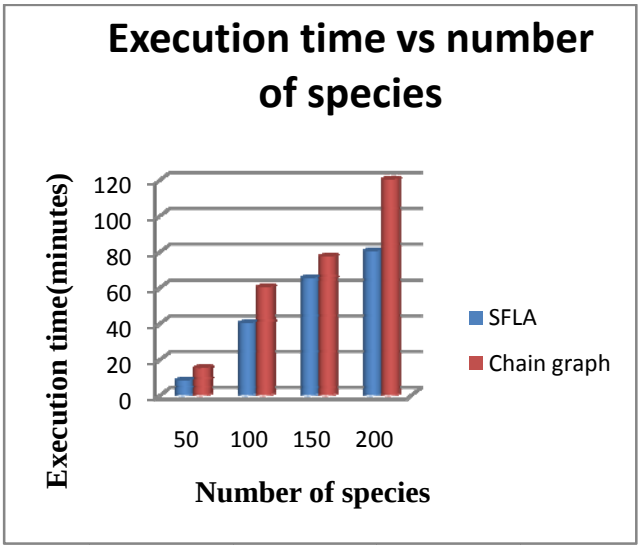


Figure5 Execution time vs Number of species

Table1 The number of protein function found are those selected on the validation set. A rule predictsmore than one homology class if there is more than one sequence similarity. A rule predicts a new homology class if there is a sequence similarity cluster in the test predictions that has no members in the training data. average test accuracy is the accuracy of the predictions on the test proteins of assigned function (if conflicts occur, the prediction with the highest *a priori* probability is chosen). Default test accuracy is the accuracy that could be achieved by always selecting the most populous class.

Networks	Yogurt	oats	soya-bean
number of protein functions found	27	32	21
rules predictiong more than one homology class	19	18	8
rules predicting a new homology class	14	16	2
average test accuracy	66%	65%	73%
default test accurach	55%	20%	6%

Table2.Number of species Execution time comparison

Number of Species	Execution Time of Chaingraph	Execution Time of SFLA
50	12.22 min	1.34 min
100	25.34 min	3.02 min
150	52.77 min	7.33min
200	1.65 hour	12.2 min
above 200	infinite	Finite

Figure6 Total cost with respect to space occupy

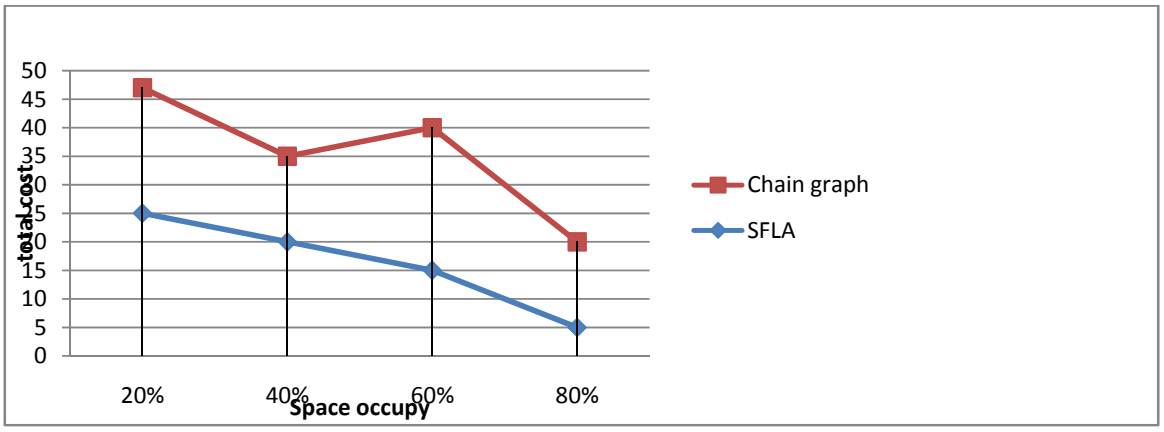


Table3 classified protein function values in SFLA and

Function	Dimension	Best		Mean Best		Std. Dev	
		SFLA	MSFLA	SFLA	MSFLA	SFLA	MSFLA
f_1	10	8.24E-18	9.68E-24	3.44E-16	8.65E-23	3.91E-16	6.31E-23
	20	2.74E-10	1.28E-20	2.41E-09	6.20E-20	1.70E-09	3.19E-20
	30	7.54E-07	4.45E-19	3.33E-06	1.22E-18	2.16E-06	5.78E-19
f_2	10	0	0	1.01	0	9.65E+01	0
	20	3.98	0	7.90	0	2.08	0
	30	7.96	0	1.60E+01	0	4.25	0
f_3	10	7.27E-10	0	4.0E-09	0	2.31E-09	0
	20	2.63E-06	0	1.10E-05	0	5.33E-06	0
	30	1.91E-04	0	4.18E-04	0	2.07E-04	0
f_4	10	1.99E-02	0	7.30E-02	1.02E-02	3.47E-02	5.17E-02
	20	1.47E-08	0	2.74E-02	0	2.59E-02	0
	30	5.70E-06	0	9.32E-03	0	1.20E-02	0
f_5	2	1.60E-16	0	9.09E-16	0	5.82E-16	0

5. Conclusion

In this paper, we proposed a modified SFLA techniques to classified the protein functions with joining networks of three different species. Experimental results on the standard test data sets show that the proposed algorithm is an effective and efficiency algorithm for optimizing data from multispecies networks. The use of the Gene Ontology enables synchronized deliberation of multiple but related functional categories,forfurther improvements to the model's predictive ability.

Our method readily extends to multiple, but not related species setting . The presence of multiple interacting networks may further enable integration of additional sources of evidence, thus contributing to increased accuracy of functional predictions.

References

[1] http://www.geneontology.org/, 2010.

[2] Antonina Mitrofanova, Vladimir Pavlovic, And Bud Mishra"Prediction Of Protein Functions With Gene Ontology And Interspecies Protein Homology Data" IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 8, No. 3, May/June 2011.

[3] U. Karaoz et al., "Whole-Genome Annotation by Using Evidence Integration in Functional-Linkage Networks," Proc. Nat'l Academy of Sciences USA, vol. 101, pp. 2888-2893, 2004.

[4] A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K.-H. Glatting, and S. Suhai, "Applying Support Vector Machines for Gene Ontology Based Gene Function Prediction," BMC Bioinformatics, vol. 5, p. 116, 2004.

[5] J. Liu and B. Rost, "Comparing Function and Structure between Entire Proteomes," Protein Science, vol. 10, pp. 1970-1979, 2001.

[6] Alireza R.V., Ali Hossein M, "Solving a bi-criteria permutation flowshop problem using shuffled frog-leaping algorithm," Soft Computing, vol. 12(5), pp. 435-452, 2008.

[7] Eusuff M. M, Lansey K.E, "Shuffled frog-leaping algorithm: A memetic meta-heuristic for discrete optimization," Engineering Optimization, vol. 38 (2), pp. 129-154, 2006.

[8] N. Nariai, E. Kolaczyk, and S. Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS ONE, 2(3), 2007.

[9] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90 (1999).

[10] Altschul,S.F., Gish,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410

[11] Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein functionprediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6): 697–700, 2003

[12] von Mering, R. Krause, B. Snel, M. Cornell, S. G. Olivier, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

[13] G. Yona, N. Linial, and M. Linial. Protomap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Structure, Function, and Genetics*, 37:360–678, 1999.

[14] Han, J., Kamber M. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, (2006).

[15] Jensen, L., et al. "Prediction of Human Protein Function from Post - Translational Modifications and Localization Features", Journal of Molecular Biology, vol. 319(5), (2002) pp 1257-65.

[16] E.Elbeltagi, T.Hegazy, And D.Grierson, "comparison among five evolutionary-based optimization algorithms," advanced engineering informatics,vol.19,pp.43-53, january 2005.

[17] A. R. Rahimi-Vahed And A. H. Mirzaei, " solving a bi-criteria permutation flowshop problem using shuffled frog-leaping algorithm," soft computing, vol.12(5), pp. 435–452, december 2008.

[1]Mr.N.Kannaiya Raja received degree MCA from Alagappa University and ME from Anna University Chennai in 2007 joined assistant professor in various engineering colleges in Tamil Nadu affiliated to Anna University and has eight years teaching experience. His research work in deep packet inspection. He has been session chair in major conference and workshops in computer vision on algorithm, network, mobile communication, image processing papers and pattern recognition. His current primary areas of research are packet inspection and network. He is interested to conduct guest lecturer in various engineering in Tamil Nadu.

[2]Dr.Mr..KArulanandam received PhD doctorate degree in 2010 from Vinayaka Missions University. He has twelve years teaching experience in various engineering colleges in Tamil Nadu which are affiliated to Anna University and his research experience network, mobile communication works, image processing papers and algorithm papers. Currently working in Ganadipathy Tulasi's Jain Engineering College Vellore .

[3]S.K.Sugunedham received degree B.Tech Computer Science Engineering from Pondicherry University Puducherry in 2010. Now pursuing ME Computer Science and Engineering in Arulmigu Meenakshi Amman College of Engineering Kanchipuram affiliated to Anna University Chennai.