2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

A Modified Bat Algorithm to Predict Protein-Protein Interaction Network

Archana Chowdhury[1], Pratyusha Rakshit[2], Amit Konar[3]
Department of Electronics and Telecommunication Engineering
Jadavpur University
Kolkata, India
[1]chowdhuryarchana@gmail.com,[2]pratyushar1@gmail.com,
[3]konaramit@yahoo.co.in

Atulya K. Nagar[4]
Department of Math and Computer Science,
Liverpool Hope University,
Liverpool, U.K
nagara@hope.ac.uk

***Abstract*—This paper provides a novel approach to predict the Protein-Protein Interaction (PPI) network using a modified version of the Bat Algorithm. The attractive trait of the proposed approach is that it attempts to analyze the impact of physicochemical properties, structural features and evolutionary relationship of proteins, to predict the PPI network. Computer simulations reveal that our proposed method effectively predicts the PPI of Saccharomyces Cerevisiae with a sensitivity of (0.85) and specificity of (0.87) and outperforms other state-of-art methodologies.**

Keywords—protein–protein interaction networks; domains; phylogenetic profiles; accessible solvent area; bat algorithm.

I. INTRODUCTION

Proteins regulate every process in the cell. Proteins do not function in isolation. Proteins interact with each other or ligands to arbitrate biological processes. Protein-Protein Interactions (PPIs) play a vital role in understanding the action mechanism of protein. A number of high throughput methods have been proposed for detecting PPIs. The high throughput methods including yeast two-hybrid (Y2H) and tandem affinity purification (TAP) provide interactions for several organisms such as S. cerevisiae [1], C. Elegans [2], D. Melanogaster [3] and H. Sapiens [4]. However there are evidences of PPIs not satisfying the presumed conditions of high throughput screening, for inferring PPIs[5]. Naturally, a failure to catch transient interactions often leads to false negative PPI prediction. The experiments are also very tiresome and laborious and it is very difficult to verify individual interaction as the number of interactions within a cell is very high. This motivated the use of computational methods for PPI prediction.

A number of computational approaches to PPI prediction have been developed over the past decade utilizing different characteristic features of existing PPIs [6-10]. The most important aspect of the paper is that the PPI prediction problem has been formulated in an evolutionary optimization framework, optimizing the objective responsible for the configuration of a PPI.

In the present context, optimal prediction of PPI is inspired by maximization of a fitness function satisfying three criteria– 1) similarity in phylogenetic sequences [11], 2) deviation of the accessible solvent area (ASA) [12] of the PPI with respect to that of individual proteins, and 3) similarity in domain-domain interaction profiles of two predicted interacting proteins in the PPI. The first criterion of the fitness function is based on the evolutionary relationship of proteins. The design philosophy adopted here relies on the fact that proteins with similar phylogenetic profiles are more likely to interact with each other [13]. PPI prediction remains incomplete if the energy of the stable protein-protein complex is not considered. In [14], solvation energy of proteins are used to describe the energetics at protein interfaces. In [12], it is shown that the free energy of protein solvation is linearly related to ASA in a continuum approach. It has motivated us to introduce ASA of the predicted protein-protein complex as a second fitness measure to ensure stable connectivity between the proteins in the PPI. The third criterion of the fitness function is based on the underlying premise that proteins interact with one another through some interacting domains, and it has become a common approach for predicting protein-protein interaction by identifying these domains [15].

In this paper, we study the scope of proposed Chaotic Local Search-based Bat Algorithm (CLSBA) to judiciously predict PPIs. The choice of Bat Algorithm (BA) [16] in the present context is inspired partly heuristically because of the background of the algorithm in the topic, and partly because of its established performance in the literature [16]. Bat Algorithm (BA) is selected for its fewer control parameters, good run-time accuracy and faster speed of convergence. CLSBA is an evolutionary strategy that utilizes the composite benefit of global exploration of BA [16] and the chaotic local search capability realized with logistic map [17] and Rechenberg's 1/5 mutation rule [18].

The rest of the paper is divided into four sections: Section II gives a brief idea about the formulation of the PPI identification problem and explains the criteria used. Section III describes the traditional BA. The proposed CLSBA is presented in section IV. Section V presents the discussion of results. Section VI concludes the paper.

II. FORMULATION OF PROTEIN-PROTEIN INTERACTION IDENTIFICATION PROBLEM

In this paper, combination of three important characteristics of PPI is considered as primary objective for its prediction.

A. Predicting Protein-Protein Interactions using Phylogenetic Analysis

The presence or absence of N proteins in a collection of K completely sequenced genomes G from different organisms is represented in a specific pattern called the phylogenetic profiles [11]. For each protein p_i, a phylogenetic profile is represented as a K-length binary string $s = s_1 s_2 \cdots s_k$ where $s_j = 1$ if protein p_i is present in genome g_j and $s_j = 0$ if protein p_i is absent in genome g_j. The clustering of proteins based on the similarity of their phylogenetic profiles can provide crucial information regarding the protein networks. It is observed that proteins present in the same cluster are functionally related. The logic underlying this reasoning is that proteins with similar phylogenetic profiles are likely to interact in performing some biological process.

To meet this issue, we evaluate the accuracy of the produced PPI network by comparing the phylogenetic profiles of two predicted interacting proteins in the network with the hope that if the two proteins interact with each other in reality then the cosine similarity (dot product) between their corresponding phylogenetic profiles will be more. Let, N be the total number of proteins in the network and K be the length of phylogenetic profile of each protein. $s_{k,i}$ represents the presence or absence of proteins p_i in genome g_k and Set_i symbolizes the set of proteins predicted to be interacting with protein p_i. Then the similarity between the phylogenetic profiles of interacting proteins in the PPI can be measued by (1).

$$C_1 = \frac{1}{K}\sum_{i=1}^{N}\sum_{\forall j \in Set_i} \frac{\sum_{k=1}^{K}(s_{i,k} \times s_{j,k})}{\sqrt{\sum_{k=1}^{K}(s_{i,k})^2} \times \sqrt{\sum_{k=1}^{K}(s_{j,k})^2}} \quad (1)$$

B. Accessible Solvent Area

The hydrophobic effect is usually defined as the reduction of the unfavorable interactions in PPI occurring between water and non-polar atoms, such as hydrophobic residues in protein which are incapable of forming hydrogen bonds in aqueous solution [19]. Hence upon binding, the binding sites of two interacting proteins must be desolvated. Once bound, the side-chain or main-chain non-polar functional groups in the binding sites of bound protein-pair become (partially or completely) immobilized and now they construct intermolecular interaction. These non-polar molecules (functional groups) stay together to minimize water-exposed Accessible Solvent Area (ASA). As a consequence, a strong binding between two proteins p_i and p_j can be ensured by the extent of reduction in the ASA of the protein complex, $ASA(p_i\_p_j)$ with respect to their individuals ASAs, i. e., $ASA(p_i)+ASA(p_j)$ by maximizing (2). Here Set_i represents the set of proteins predicted to be interacting with protein p_i.

$$\begin{aligned} C_2 &= ASA\,of\ individual\,proteins - ASA\,of\ protein\,complex \\ &= \sum_{i=1}^{N}\sum_{\forall j \in Set_i}\left(ASA(p_i) + ASA(p_j) - ASA(p_{i\_j})\right) \end{aligned} \quad (2)$$

C. Protein-Protein Interaction Prediction based on Domain-Domain Interaction

Proteins interact with each other through their small substructures, known as domains. Such domain architecture governs the protein-protein complex formation, offering a framework for prediction model of PPI [20]. In the domain-based structural quantification approach, the knowledge about the strength of interaction between domain d_i in protein p_1 and domain d_j in protein p_2 is used to predict whether proteins p_1 and p_2 interact.

Let,

M be the number of domain-domain interactions of yeast,

$dom(p)$ be the set of domains present in protein p,

$ddi(p)$ be the set of domains interacting with domain $d_k \in dom(p)$, for $k = [1, |dom(p)|]$, where $|dom(p)|$ is the number of unique domains present in protein p.

The DOMINE database [25] is used to extract domain-domain interaction data of yeast. Given the domain-domain interaction data we can measure the similarity between predicted interacting protein-pair, p_i and p_j based on their domain-domain interaction information using Pearson coefficient as follows.

$$r(p_i, p_j) = \frac{M \times \left|ddi(p_i) \cap ddi(p_j)\right| - \left|ddi(p_i)\right| \times \left|ddi(p_j)\right|}{\sqrt{\left(M \times \left|ddi(p_i)\right| - \left|ddi(p_i)\right|^2\right) \times \left(M \times \left|ddi(p_j)\right| - \left|ddi(p_j)\right|^2\right)}} \quad (3)$$

Here $|ddi(p)|$ represents the number of domains present in $ddi(p)$ for any arbitrary protein p. It is apparent from (3) that if all the domains present in $dom(p_i)$ and $dom(p_j)$ are same, i.e., $1 \le |ddi(p_i)| = |ddi(p_j)| = |ddi(p_i) \cap ddi(p_j)| \le M$, $r(p_i, p_j) = 1$. It in turn represents a high possibility of interaction between these two proteins with high structural similarity. On the other hand, $r(p_i, p_j) < 0$ if there is no common domain-domain interaction of proteins p_i and p_j i.e., $ddi(p_i) \cap ddi(p_j) = \varphi$, indicating a rare chance of interaction.

Hence the value of $r(p_i, p_j)$ diminishes with increase in the dissimilarity between domain-domain interaction data of proteins p_i and p_j (a reduction in the first term of numerator of (3)). Consequently, the accuracy of prediction of interaction between any two proteins, mediated by a great variety of interacting domains, can be improved by maximizing the third objective as given in (4).

$$C_3 = \sum_{i=1}^{N}\sum_{\forall j \in Set_i} r(p_i, p_j) \quad (4)$$

We now construct a fitness function, the maximization of which yields a possible solution to the PPI identification problem. The expressions to be considered are (1), (2) and (4). Hence the overall fitness function to be maximized is given by

$$fit = C_4 = w_1 \times C_1 + w_2 \times C_2 + w_3 \times C_3 \quad (5)$$

where w_1, w_2, and w_3 (>0) are scale factors. These parameters are set in a manner to have all the terms on the right hand side of (5) in the same order of magnitude. The larger the value of the function *fit* the better is the performance of PPI identification.

D. Formation of a Protein-Protein Interaction Network

In the proposed method for N proteins of interest, each with K dimensional phylogenetic sequence, a solution is represented by a two dimensional binary matrix $\boldsymbol{Z}= [z_{j,k}]$, $\forall j$, $k \in [1, N]$ of dimension $N \times N$. It describes the presence or absence of an interaction between two proteins. Hence

$$z_{j,k} = \begin{cases} 1 & \text{if proteins } p_j \text{ and } p_k \text{ interact with each other} \\ 0 & \text{if there is no interaction between } p_j \text{ and } p_k \end{cases} \quad (6)$$

III. BAT ALGORITHM (BA)

An overview of the Bat Algorithm (BA) is given below.

1. Initialization: The position $\vec{X}_i(t) = \{x_{i,1}(t), x_{i,2}(t), ..., x_{i,D}(t)\}$, velocity $\vec{V}_i(t) = \{v_1(t), v_2(t), ..., v_D(t)\}$, loudness $A_i(t)$ and the pulse emission rate $r_i(t)$ of the i-th bat at generation t=0 is selected randomly in the range $[\vec{X}_{\min}, \vec{X}_{\max}]$, $[\vec{V}_{\min}, \vec{V}_{\max}]$, $[A_{\min}, A_{\max}]$ and $[r_{\min}, r_{\max}]$ respectively for i= [1, NP] where $\vec{X}^{\min} = \{x_1^{\min}, x_2^{\min}, ..., x_D^{\min}\}$, $\vec{X}^{\max} = \{x_1^{\max}, x_2^{\max}, ..., x_D^{\max}\}$ and $\vec{V}^{\min} = \{v_1^{\min}, v_2^{\min}, ..., v_D^{\min}\}$, $\vec{V}^{\max} = \{v_1^{\max}, v_2^{\max}, ..., v_D^{\max}\}$ respectively.

2. Evaluating the Global Best Position: The fitness $fit(\vec{X}_i(t))$ is evaluated for i= [1, NP]. The position of a bat with highest fitness is selected as the global best position $\vec{X}^{best}(t)$ at generation t.

3. Frequency Selection: The frequency f_i of the emitted pulse by the i-th bat is determined as follows for i= [1, NP].

$$f_i = f_{\min} + \beta \times (f_{\max} - f_{\min}) \quad (7)$$

Here β is a random number within (0, 1), $f_{\min}$ and $f_{\max}$ are minimum and maximum frequencies respectively.

4. Velocity Update: The velocity of the i-th bat for i=[1, NP] is updated as follows:

$$\vec{V}_i(t+1) = \vec{V}_i(t) + f_i \times (\vec{X}_i(t) - \vec{X}_{best}(t)) \quad (8)$$

5. Position Update: The position of the i-th bat for i=[1, NP] is updated as follows.

$$\vec{X}_i(t+1) = \vec{X}_i(t) + \vec{V}_i(t+1) \quad (9)$$

6. Generating Local Position : A new position $\vec{X}'_i(t) = \{x'_{i,1}(t), x'_{i,2}(t), ..., x'_{i,D}(t)\}$ is discovered by the i-th bat around $\vec{X}^{best}(t) = \{x_1^{best}(t), x_2^{best}(t), ..., x_D^{best}(t)\}$ for i=[1, NP] with a probability $(1-r_i)$ following (10) for j= [1, D].

$$x'_{i,j}(t) = x_j^{best}(t) + \varepsilon \times A_{avg}(t) \quad (10)$$

Here ε is slected randomly from [–1, 1]. $A_{avg}(t)$ is the average loudness of all NP bats in the current population.

$$A_{avg}(t) = \sum_{i=1}^{NP} A_i(t) \Big/ NP \quad (11)$$

7. Selection: $\vec{X}^{best}(t)$ is replaced with new $\vec{X}'_i(t)$ with a probability A_i provided that $fit(\vec{X}'_i(t)) > fit(\vec{X}^{best}(t))$. This is repeated for for i=[1, NP].

8. Update Loudness and Pulse Emission Rate: If $\vec{X}_{best}(t)$ is successfully replaced by $\vec{X}'_i(t)$, the loudness is reduced and pulse emission rate is increased by following (12) and (13) respectively for i=[1, NP].

$$A_i(t+1) \leftarrow \alpha \times A_i(t) \quad (12)$$

$$r_i(t+1) \leftarrow r_i(t) \times (1 - \exp(-\gamma \times t)) \quad (13)$$

This process is iterated from step 2 till termination condition has been reached.

IV. PROPOSED CHAOTIC LOCAL SEARCH BAT ALGORITHM (CLSBA)

A proper tuning of the frequency of pulse emission f_i in (8) plays a significant role in the generation of new promising position of the bats. This paper proposes a Chaotic Local Search-based Bat Algorithm (CLSBA) combining standard BA with chaotic sequences generated by the logistic map [17]. The use of chaos makes the frequency adaptive and more random in nature to balance the trade-off between global exploration and local exploitation. The chaotic behavior in f_i follows the non-linear dynamics of logistic map.

$$f_i(t+1) = \mu \times f_i(t) \times (1 - f_i(t)) \quad (14)$$

Here μ is a control parameter. When we set μ=4 and $f_i(0) \neq \{0, 0.25, 0.5, 0.75, 1\}$ then the value of f_i(t) distributes with proper randomness and irregularity. Moreover, the local search capability of the traditional BA being a crucial deterministic factor of its performance has been further improved here by adapting the step-size parameter ε. The value of ε lies within the range [–1, 1] in the traditional BA while it varies within the range [–SF(t), SF(t)] in the proposed CLSBA with magnitude of the perturbation in (10) being guided by the scaling factor (SF). A lower value of SF(t) influences the local exploitation while a larger value of SF(t) accelerates the global search. Here, automatic tuning of SF(t) is performed using Rechenberg's 1/5 mutation rule [18]. It

adapts $SF(t)$ based on $\varphi(m)$, the ratio of the number of successful replacements of the global best position with the newly discovered neighborhood position to the total number of local search carried out around the best position in m cycles of the algorithm. The adaptation rule of $SF(t)$ is given below.

$$SF(t+1) = \begin{cases} SF(t) \times 0.85 & if\ \varphi(m) < 1/5 \\ SF(t)/0.85 & if\ \varphi(m) > 1/5 \\ SF(t) & if\ \varphi(m) = 1/5 \end{cases} \quad (15)$$

V. EXPERIMENTS AND RESULTS

A. Simulation Results

The performance of the proposed CLSBA is examined here with respect to minimizing 25 CEC-2005 recommended benchmark functions [21] of 50 dimensions each. Here, we compare CLSBA with traditional Bat Algorithm, Global Best Particle Swarm Optimization (g-best PSO) [22] and Harmony Search (HS) [23] algorithms. For each algorithm, the population size is kept at 50 and the maximum function evaluations (FEs) is set as 500000. We employ the best parametric set-up for all these four competitor algorithms as prescribed in their respective sources. For the proposed CLSBA algorithm, we have selected $A_i(0)= 1$, $r_i(0)= 0.5$, $f_{\min}= 1$, $f_{\max}= 2$, m=50 and $SF(0)$=1.

The mean and standard deviation (within parenthesis) of the best-of-run values of 50 independent runs for each of the four algorithms are presented in Table-I. In the sixth column of Table-I we present the statistical significance level of the difference of the mean of the best two algorithms using t-test of 25 samples. Note that here "+" indicates that the t value of 49 degrees of freedom is significant at a 0.05 level of significance by two-tailed test, whereas "–" means the difference of mean is not statistically significant, and "NA" stands for not applicable, covering cases for which two or more algorithms achieve the best accuracy results. The best algorithm is marked in bold. A close scrutiny of the simulation results in Table-I reveals that CLSBA remains consistently superior to its competitors with respect to the quality of solutions outperforming its competitors over 19 cases out of 25 benchmark instances in a statistically significant manner. In two cases (f06 and f25), BA, which remains the second best algorithm outperforms CLSBA.

B. Performance of CLSBA in PPI Prediction

B.1. Experimental Set-up

In each generation of CLSBA, the position of the bat is decoded to obtain the corresponding PPI network. In order to identify the PPI network, we need to maximize the expression in (5) which determines the best position of the bat. The raw data set consists of 118,363 interactions involving 6593 *Saccharomyces Cerevisiae* proteins, of which 75,748 interactions are unique. The dataset is pruned by removing unannotated protein, self-interactions and repeated interactions to obtain the final dataset which consists of 69,331 interaction pairs involving 5386 annotated proteins. For experiments, the Cartesian coordinates of the proteins in *Saccharomyces Cerevisiae* are obtained from Protein Data Bank [http://www.rcsb.org/pdb/home/home.do]. The phylogenetic profile of these proteins is generated with respect to ten species namely *Saccharomyces Cerevisiae*, *Caenorhabdities Elegans*, *Aedens Aegypti*, *Anopheles Gambiae*, *Drosophila Melanogaster*, *Ciona Intestinalis, CionaSavignyi*, *Tetraodon Nigroviridis*, *Takifugu Rubripes* and *Oryziaslatipes* using PhyloPat: Phylogenetic Patterns [http://www.cmbi.ru.nl/cdd/phylopat/52/]. The ASA is calculated using GETAREA [http://curie.utmb.edu/getarea.html]. The protein domain information is gathered from Pfam [24], which is a protein domain family database that contains multiple sequence alignments of common domain families. In total, there are 4293 Pfam domains defined by the set of proteins in *Saccharomyces Cerevisiae*. The domain-domain interaction is obtained from DOMINE [25].

B.2. Competitor Algorithms and Parameter Settings

We have compared our proposed method with other computational methods including Support Vector Machines (SVM) [26], Random Forests (RF) [27], Artificial Neural Networks (ANN) [28], Bayesian Classifiers (BC) [29], Maximum Likelihood Estimate (MLE) [15] and Phylogenetic Profile (PP) [30]. We have also compared the proposed CLSBA-based PPI prediction approach with other swarm/evolutionary algorithm based methods including traditional BA [16], PSO [22] and HS [23] algorithms for the same application. All the competitor algorithm-based simulations use same solution representation scheme (6) and fitness function (5) as in case of CLSBA-based approach.

B.3. Performance Metrics

True Positive (*TP*): It is the number of interactions that are predicted as interactions and are indeed true interactions.
False Positive (*FP*): It is the number of predicted interactions that are in fact not real interactions.
False Negative (*FN*): It is the number of protein pairs that are reported as not interacting but are indeed true interactions.
True Negative (*TN*): It is the number of protein pairs that are correctly predicted not to interact.

Based on the above four interconnection states, the performance of PPI prediction can be analyzed using following metrics:
Sensitivity or Recall: It measures actual proportion of the positive interactions (*TP*) that is predicted correctly.

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \quad (16)$$

Specificity: It measures actual proportion of the negative interactions (*TN*) that is predicted correctly.

$$Specificit\ y = \frac{TN}{FP + TN} \quad (17)$$

Positive Likelihood Ratio (*PLR*): It is the probability of *TP* predictions in the PPI network with respect to the probability of the incorrectly predicted positive interactions i.e., *FP*.

$$PLR = \frac{Sensitivity}{1 - Specificity} \tag{18}$$

Negative Likelihood Ratio (*NLR*): It is the probability of the incorrectly predicted positive interactions i.e., *FN* predictions in the PPI network with respect to the probability of *TN* predictions.

$$NLR = \frac{1 - Sensitivity}{Specificity} \tag{19}$$

Precision or Positive Predicted Value (*PPV*): It measures the percentage of true positive interactions among all of the predicted positive interactions.

$$Precision\,(PPV) = \frac{TP}{TP + FP} \tag{20}$$

Negative Predicted Value (*NPV*): It measures the percentage of true negative interactions among all of the predicted negative interactions.

$$NPV = \frac{TN}{TN + FN} \tag{21}$$

Accuracy: It is the overall correctness of the predictive model and is calculated as the sum of correct predictions divided by the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

F1_score: An F_1 score reaches its best value at 1 and worst score at 0.

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{23}$$

Mathews Correlation Coefficient (*MCC*): It is in essence a correlation coefficient between the observed and predicted binary classes (positive and negative interaction in case of PPI); it returns a value between −1 and +1.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{24}$$

Receiver Operating Characteristic (*ROC*) curve and Area Under Curve (*AUC*): The accuracy of an algorithm to distinguish between TP and FP prediction can be analyzed using *ROC*. It plots the *Sensitivity* against (1-*Specificity*) over a range of varying control parameter of an algorithm. The *AUC* is the measurement of the area under the *ROC* curve.

B.4. Results and Performance Analysis

To evaluate our proposed method we plot the ROC curves for different PPI prediction algorithms in Fig. 1. The *ROC* for the evolutionary methods (CLSBA, BA, PSO, HS) is drawn for various iterations (ranging from 100000 to 500000) and for classification methods (SVM, RF, ANN, BC, MLE, PP), they are drawn for varying thresholds (ranging from 0.3 to 0.8). The closer the curve is to the upper left hand corner (i.e., the larger the area under curve), the better the predictive algorithm achieving a high value for both sensitivity and specificity. From Fig. 1 and Table-II, we note that the *AUC* for CLSBA method attains highest value than other competitor classification algorithms.

A relative analysis of PPI prediction performance of ten algorithms based on *Precision* and *Recall*, can be obtained from Fig. 2. Here we consider a straight line passing through the origin making an angle of 45^0 with the *Recall* axis and intersecting all the curves. The distance of the intersection points on the curves (corresponding to different predictive algorithms) from the origin is used as a measure of the performance of the algorithm. The higher the measure, the better is the performance. We use ">=" symbol to represent the relative performance of any two algorithms and the ranking thus obtained is depicted as: CLSBA>=BA>=PSO>=HS>=SVM>=RF>=ANN>=BC>=MLE>=PP. The mean values of the performance metrics over 50 runs of each PPI prediction algorithm are plotted in Fig. 3 The mean and the standard deviation (within parenthesis) of the best-of-run values of the performance metrics for 25 independent runs of each of the algorithm are presented in Table-III. We also use paired t-test to compare the mean of each performance metric produced by the best and the second best algorithm. In the last row of Table-III the statistical significance level of the difference of the mean of the best two algorithms is provided. A close inspection of Fig. 3 and Table-III indicates that the performance of the proposed CLSBA-based PPI prediction algorithm has remained consistently superior to that of the other competitor methods.

In order to visualize the influence of different objectives (C_1 to C_3) for predicting PPI, we have considered a sub-network of the PPI dataset of *Saccharomyces Cerevisiae* (Fig. 4) comprising 11 proteins, namely CDC73, CTR9, LEO1, RTF1, SPT16, PAF1, POB3, CKA2, CKA1, HTZ1 and GAL11. The three objectives (C_1 to C_3) are measured for one interacting protein-pair, CDC73–LEO1 and also for one non-interacting protein-pair, LEO1–GAL11 in Table-IV. In Table-IV-A, we see that C_1 is highest for the interacting protein-pair. In Table-IV-B a higher C_2 value for the interacting pair (3283.68) signifies greater reduction in the ASA of the complex resulting in a strong binding between them. The consideration of domain similarity for PPI prediction is confirmed by a higher positive value of C_3 (0.30757) for interacting proteins (Table-IV-C). The predicted PPI for the same sub-network obtained using ten competitor algorithms are also pictorially

represented in Fig. 5. Comparing Fig. 5 with Fig. 4, it is apparent that CLSBA-based method outperforms other competitors in predicting correct PPIs.

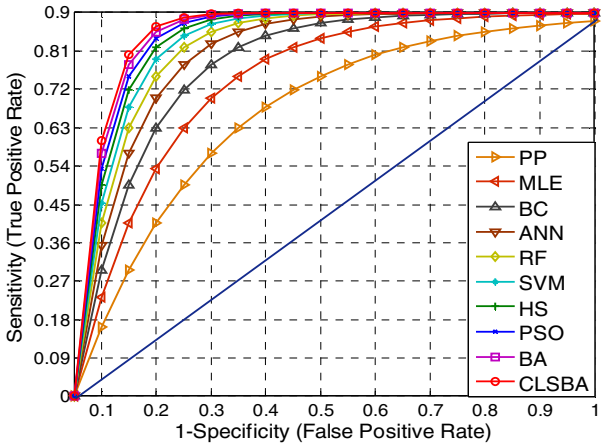

Fig. 1. ROC plot for different PPI prediction algorithms

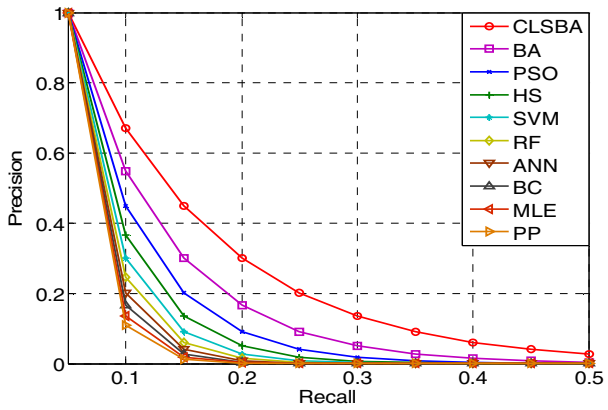

Fig. 2. PROC plot for different PPI prediction algorithms

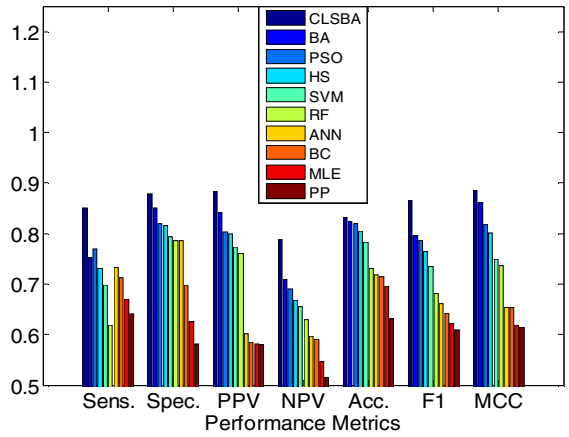

Fig. 3. Plot of performance metrics for different PPI prediction algorithms for 50 runs

TABLE I. COMPARISON OF MEAN FITNESS FUNCTION VALUES OF CLSBA WITH OTHER COMPETITORS OVER 50 INDEPENDENT RUNS

Function No.	HS	PSO	BA	CLSBA	Statistical Significance
f 01	1.469e-010 (2.05e-011)	5.684e-014 (0.00e+000)	**0.000e+000 (4.18e-014)**	**0.000e+000 (0.00e+000)**	NA
f 02	4.124e-005 (5.60e+002)	1.070e-006 (1.63e-006)	**0.000e+000 (9.91e-014)**	**0.000e+000 (0.00e+000)**	NA
f 03	1.608e+007 (2.96e+006)	8.320e+005 (4.30e+005)	1.047e+005 (7.58e+004)	**0.000e+000 (0.00e+000)**	+
f 04	1.714e+004 (2.32e+003)	9.705e+000 (8.10e+000)	5.895e-009 (1.33e-008)	**0.000e+000 (0.00e+000)**	+
f 05	4.363e+003 (4.68e+002)	1.385e+003 (8.44e+002)	6.656e+002 (1.68e+002)	**2.273e-012 (1.31e-012)**	+
f 06	4.050e+001 (1.80e+001)	2.179e+001 (2.00e+001)	**1.352e+000 (1.88e+000)**	4.783e-001 (1.32e+000)	–
f07	1.712e-002 (8.00e-005)	8.965e-003 (6.88e-003)	4.078e-001 (6.69e-002)	**1.833e-001 (5.16e-002)**	+
f08	6.298e+001 (1.85e+002)	2.094e+001 (5.21e-002)	**2.036e+001 (7.14e-002)**	**2.036e+001 (7.14e-002)**	NA
f09	2.778e+002 (1.59e+007)	1.089e+001 (2.74e+001)	1.985e+000 (6.11e-001)	**0.000e+000 (2.32e-014)**	+
f10	2.275e+002 (1.15e+003)	1.209e+002 (1.97e+001)	1.854e+001 (5.83e+000)	**1.365e+001 (1.50e+000)**	+
f11	3.841e+001 (6.25e+002)	2.867e+001 (1.16e+000)	6.427e+000 (4.89e-001)	**5.735e+000 (6.30e-001)**	+
f12	8.870e+005 (1.18e+005)	1.034e+003 (1.43e+002)	1.330e+002 (2.34e+002)	**9.116e+001 (2.79e+002)**	+
f 13	4.961e-001 (1.38e-001)	9.948e-003 (7.12e-003)	8.965e-003 (6.88e-003)	**6.062e-000 (5.15e-002)**	+
f14	1.409e+001 (1.12e-001)	1.413e+001 (7.26e-002)	1.387e+001 (1.13e-001)	**1.011e+001 (2.23e-001)**	+
f 15	2.094e+001 (5.04e-002)	2.112e+001 (4.27e-002)	2.094e+001 (5.21e-002)	**2.018e+001 (7.18e-002)**	+
f 16	9.045e+001 (1.15e+000)	1.416e+001 (8.33e+001)	1.089e+001 (2.74e+001)	**7.404e+000 (2.28e+000)**	+
f 17	1.595e+002 (1.24e+001)	1.828e+002 (7.32e+001)	1.209e+002 (1.97e+001)	**6.938e+000 (1.06e+000)**	+
f 18	7.181e+001 (1.51+000)	2.828e+001 (1.66e+000)	2.867e+001 (1.16e+000)	**6.670e+000 (2.09e+000)**	+
f 19	1.157e+005 (1.35e+004)	4.856e+006 (5.22e+005)	8.870e+005 (1.18e+005)	**5.393e+003 (4.25e+003)**	+
f 20	2.443e+001 (2.38e+000)	1.166e+001 (1.83e+000)	4.090e+000 (2.91e-001)	**1.180e+000 (2.70e-001)**	+
f 21	1.304e+001 (1.42e-001)	2.303e+001 (2.09e-001)	1.311e+001 (2.23e-001)	**2.880e+000 (1.38e-001)**	+
f 22	1.219e+002 (1.88e+001)	9.490e+001 (2.77e+001)	5.008e+001 (7.30e+001)	**4.292e+001 (1.12e+001)**	–
f 23	2.012e+002 (1.46e+001)	1.512e+002 (5.87e+001)	1.457e+002 (1.93e+001)	**1.128e+002 (1.31e+001)**	+
f 24	2.733e+002 (6.79e+001)	2.672e+002 (2.53e+001)	2.178e+002 (8.04e+001)	**1.340e+002 (1.84e+001)**	+
f 25	8.478e+002 (2.07e+000)	8.460e+002 (2.12e+001)	**7.163e+002 (2.46e-001)**	7.633e+002 (1.77e+002)	–

TABLE II: AREA UNDER CURVE OBTAINED FROM FIG. 1

CLSBA	BA	PSO	HS	SVM	RF	ANN	BC	MLE	PP
0.888 (0.19)	0.885 (0.24)	0.871 (0.25)	0.868 (0.25)	0.838 (0.34)	0.778 (0.47)	0.750 (0.61)	0.661 (0.81)	0.611 (0.84)	0.602 (0.92)

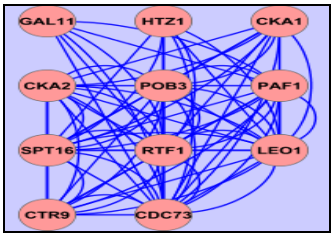

Fig. 4. Original sub-network in yeast PPI

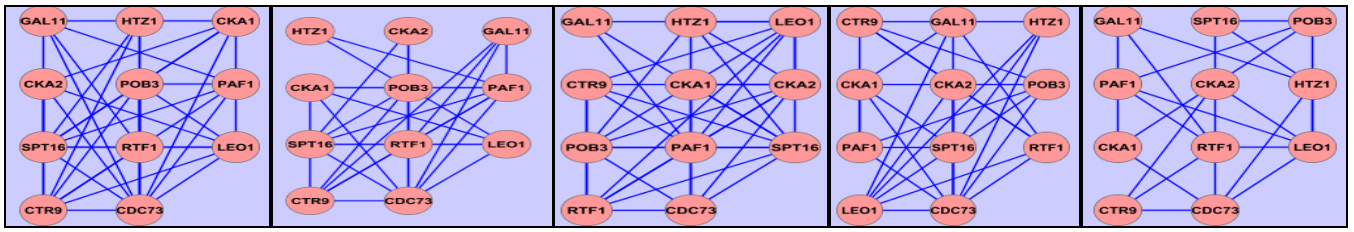

Fig. 5. Sub-network obtained by PPI prediction algorithms: (a) CLASBA (b) BA (c) PSO (d) HS (e) SVM (f) RF (g) ANN (h) BC (i) MLE and (j) PP

TABLE III. COMPARISON OF DIFFERENT PPI PREDICTION ALGORITHMS FOR 25 RUNS

Algorithms	Sensitivity	Specificity	PLR	NLR	Precision	NPV	Accuracy	F1_score	MCC
CLSBA	**0.8503 (0.196)**	**0.8782 (0.285)**	**5.6800 (0.178)**	**0.1761 (0.186)**	**0.8833 (0.011)**	**0.7877 (0.167)**	**0.8324 (0.162)**	**0.8665 (0.078)**	**0.8866 (0.004)**
BA	0.7532 (0.243)	0.8503 (0.351)	3.0519 (0.193)	0.3277 (0.382)	0.8415 (0.053)	0.7085 (0.327)	**0.8324** (0.165)	0.7949 (0.083)	0.8630 (0.084)
PSO	0.7701 (0.251)	0.8204 (0.380)	3.3497 (0.259)	0.2985 (0.438)	0.8026 (0.075)	0.6904 (0.438)	0.8198 (0.262)	0.7860 (0.152)	0.8172 (0.106)
HS	0.7306 (0.254)	0.8159 (0.549)	2.7120 (0.389)	0.3687 (0.445)	0.7998 (0.129)	0.6691 (0.514)	0.8047 (0.311)	0.7636 (0.228)	0.8005 (0.259)
SVM	0.6983 (0.349)	0.7934 (0.567)	2.3146 (0.495)	0.4320 (0.489)	0.7742 (0.337)	0.6550 (0.571)	0.7835 (0.528)	0.7343 (0.442)	0.7501 (0.399)
RF	0.6182 (0.473)	0.7863 (0.585)	1.6192 (0.561)	0.6176 (0.646)	0.7615 (0.469)	0.6299 (0.668)	0.7312 (0.601)	0.6824 (0.450)	0.7366 (0.774)
ANN	0.7340 (0.616)	0.7866 (0.753)	2.7594 (0.577)	0.3624 (0.709)	0.6027 (0.530)	0.5967 (0.733)	0.7178 (0.654)	0.6619 (0.538)	0.6540 (0.800)
BC	0.7136 (0.814)	0.6990 (0.757)	2.4916 (0.632)	0.4013 (0.754)	0.5858 (0.568)	0.5911 (0.762)	0.7146 (0.689)	0.6434 (0.825)	0.6542 (0.817)
MLE	0.6713 (0.840)	0.6269 (0.830)	2.0423 (0.644)	0.4896 (0.765)	0.5817 (0.779)	0.5473 (0.849)	0.6955 (0.748)	0.6233 (0.913)	0.6191 (0.868)
PP	0.6410 (0.929)	0.5815 (0.917)	1.7855 (0.651)	0.5601 (0.792)	0.5804 (0.934)	0.5160 (0.909)	0.6313 (0.794)	0.6092 (0.996)	0.6151 (0.961)
Statistical Significance	+	+	+	+	+	+	NA	+	+

VI. CONCLUSION

In this paper, a novel method based on CLSBA is used to predict PPI Network. An important aspect of our approach is the way we combine various criteria of PPI to formulate the objective function. We evaluate our result and compare them with several existing methods. The results reveal that the proposed method outperforms its competitors in predicting PPIs with respect to ten performance metrics.

ACKNOWLEDGMENT

Funding by Council of Scientific and Industrial Research (CSIR) (for awarding Senior Research Fellowship to the second author) and UGC (for UPE-II program) are gratefully acknowledged for the present work.

TABLE IV. CASE STUDY ON FITNESS FUNCTION VALUES FOR INTERACTING AND NON-INTERACTING PROTEIN PAIRS

TABLE IV-A: SIMILARITY OF PHYLOGENETIC PROFILES

Phylogenetic Profiles Of Interacting Proteins										
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
CDC73	1	1	1	1	1	1	0	1	1	0
LEO1	1	1	1	1	1	1	1	1	1	1
C_1= 0.89442										
Phylogenetic Profiles Of Non-interacting Proteins										
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
LEO1	1	1	1	1	1	1	1	1	1	1
GAL11	1	0	0	0	0	0	0	0	0	0
C_1= 0.31623										

TABLE IV-B: REDUCTION IN ACCESSIBLE SOLVENT AREA

Accessible Solvent Area of Interacting Protein Complex	
CDC73	9576.88
LEO1	9751.94
CDC73-LEO1	15046.36
C_2= 4282.46	
Accessible Solvent Area of Non-interacting Protein Complex	
LEO1	9056.16
GAL11	9576.88
LEO1-GAL11	15349.36
C_2= 3283.68	

TABLE IV-C: SIMILARITY OF DOMAIN INTERACTION PROFILES

Domain-Domain Interaction of Interacting Proteins		
Proteins	Unique Domains	Interacting Domains
CDC73	PF05179	PF00069, PF00923, PF01214, PF03985, PF04004
LEO1	PF04004	PF00999, PF03985, PF00324, PF04037, PF04046, PF05179
C_3= 0.30757		
Domain-Domain Interaction of Non-interacting Proteins		
Proteins	Unique Domains	Interacting Domains
LEO1	PF04004	PF00999, PF03985, PF00324, PF04037, PF04046, PF05179
GAL11	PF05397	PF00125, PF02002, PF02186, PF03902, PF04934, PF05001, PF05669, PF05983, PF07544, PF08601, PF08633
C_3= -0.00214		

REFERENCES

[1]. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, "A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae." *Nature* 403, no. 6770 (2000): 623-627.

[2]. S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, "A map of the interactome network of the metazoan C. elegans." *Science* 303, no. 5657 (2004): 540-543.

[3]. L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, "A protein interaction map of Drosophila melanogaster."*Science* 302, no. 5651 (2003): 1727-1736.

[4]. T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, "A physical and functional map of the human TNF-α/NF-κB signal transduction pathway." *Nature cell biology* 6, no. 2 (2004): 97-105.

[5]. Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M: Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 2005,23:839-844

[6]. R. Kini, R. Manjunatha, and H. J. Evans, "Prediction of potential protein-protein interaction sites from amino acid sequence: Identification of a fibrin polymerization site." *FEBS letters* 385, no. 1 (1996): 81-86.

[7]. S. Jones and J. M. Thornton. "Prediction of protein-protein interaction sites using patch analysis." *Journal of molecular biology* 272, no. 1 (1997): 133-143.

[8]. F. Pazos, M. H. Citterich, G. Ausiello, and A. Valencia, "Correlated mutations contain information about protein-protein interaction." *Journal of molecular biology* 271, no. 4 (1997): 511-523.

[9]. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events." *Nature* 402, no. 6757 (1999): 86-90.

[10]. C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. "Co-evolution of proteins with their interaction partners." *Journal of molecular biology* 299, no. 2 (2000): 283-293.

[11]. D. Eisenberg, and A. D. McLachlan. "Solvation energy in protein folding and binding." (1986): 199-203.

[12]. T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, "Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides." *Proceedings of the National Academy of Sciences* 84, no. 10 (1987): 3086-3090.

[13]. F. Pazos, and A. Valencia, "Similarity of phylogenetic trees as indicator of protein–protein interaction." *Protein engineering* 14, no. 9 (2001): 609-614.

[14]. B. V. Freyberg, and W. Braun. "Efficient search for all low energy conformations of polypeptides by Monte Carlo methods." *Journal of computational chemistry* 12, no. 9 (1991): 1065-1076.

[15]. M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain–domain interactions from protein–protein interactions." *Genome research* 12, no. 10 (2002): 1540-1548.

[16]. X. S. Yang, "A new metaheuristic bat-inspired algorithm." In *Nature inspired cooperative strategies for optimization (NICSO 2010)*, pp. 65-74. Springer Berlin Heidelberg, 2010.

[17]. R. M. May, "Simple mathematical models with very complicated dynamics." *Nature* 261, no. 5560 (1976): 459-467.

[18]. T. Bäck, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Vol. 996. Oxford: Oxford university press, 1996.

[19]. S. Nilapwar, "Characterization and exploitation of protein ligand interactions for structure based drug design." PhD diss., UCL (University College London), 2009.

[20]. J. Reimand, S. Hui, S. Jain, B. Law, and G. D. Bader, "Domain-mediated protein interaction prediction: From genome to network."FEBS letters 586, no. 17 (2012): 2751-2763.

[21]. P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. -P. Chen, A. Auger, and S. Tiwari, "Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization", in *Technical Report. 2005. Nanyang Technological University*, Singapore, May 2005 AND KanGAL Report #2005005, IIT Kanpur, India.

[22]. J. Kennedy, "Particle swarm optimization." In *Encyclopedia of Machine Learning*, pp. 760-766. Springer US, 2010.

[23]. Z.W. Geem, J. H. Kim, and G. V. Loganathan, "A new heuristic optimization algorithm: Harmony search," Simulation, vol. 76, no. 2, pp. 60–68, Feb. 2001.

[24]. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. G. Jones, A. Khanna, "The Pfam protein families database."*Nucleic acids research* 32, no. suppl 1 (2004): D138-D141.

[25]. S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, and R. Jothi. "DOMINE: a comprehensive collection of known and predicted domain-domain interactions." *Nucleic acids research* 39, no. suppl 1 (2011): D730-D735.

[26]. C. Z. Cai, W. L. Wang, L. Z. Sun, and Y. Z. Chen, "Protein function classification via support vector machine approach." *Mathematical biosciences*185, no. 2 (2003): 111-122.

[27]. X. W. Chen, and M. Liu, "Prediction of protein–protein interactions using random decision forest framework." *Bioinformatics* 21, no. 24 (2005): 4394-4400.

[28]. P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein–protein interaction sites in heterocomplexes with neural networks." *European Journal of Biochemistry* 269, no. 5 (2002): 1356-1361.

[29]. R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data." *Science* 302, no. 5644 (2003): 449-453.

[30]. S. V. Date, and E. M. Marcotte, "Protein function prediction using the Protein Link EXplorer (PLEX)." *Bioinformatics* 21, no. 10 (2005): 2558-2559.