A Multi-objective Evolutionary Approach 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
Department of Math and Computer Science,
Liverpool Hope University,
Liverpool, U.K
nagara@hope.ac.uk

***Abstract*—Protein-Protein Interactions (PPIs) play an important role in various cellular processes. This paper attempts to solve the PPI prediction problem in a multi-objective optimization framework. The scoring functions for the trial solution deal with simultaneous minimization of functional dissimilarity, intra- as well as inter-molecular energy and the difference in phylogenetic profiles of interacting proteins. The above optimization problem is solved using Firefly Algorithm with Non-dominated Sorting. The proposed technique outperforms existing methods, including gene-ontology based Relative Specific Similarity, Fuzzy SVM, phylogenetic profile and evolutionary/swarm algorithm based approaches, with respect to sensitivity, specificity and F1 score.**

Keywords—protein–protein interaction networks; gene ontology; CHARMM energy; phylogenetic profiles; firefly algorithm; non-dominated sorting.

I. INTRODUCTION

Proteins are groups of amino acids linked together by peptide bonds. They play a vital role in organisms and participate in many processes within cells. Proteins interact with other proteins to form PPI. In nature it is rare to find proteins that perform biological actions without the assistance of other proteins [1]. Thus Protein–Protein Interaction (PPI) is important. There are several in vivo methods and in vitro methods for identifying PPI [2]. There are many experiments that provide physical interactions among proteins while few experiments provide functional associations among proteins.

The high throughput methods may lead to many erroneous results in the form of false positive and false negative. Computational methods, however, can overcome the limitations of high throughput methods and thus are gaining importance in bio-informatics research. Various computational approaches have been developed in the past to predict PPIs utilizing different characteristic features of the existing PPIs [3-5].

In the present context of predicting possible interaction between two proteins, the objectives for the scoring functions are based on minimizing 1) dissimilarity of their functions, 2) the intra- and inter-molecular energy of their complex, and 3) the difference in their phylogenetic profiles. The first criterion of the objective function is based on Gene Ontology (GO) annotation of proteins. GO annotation, referring to a unified representation of genes and gene product across all species, has been identified as one of the strongest predictors for protein interaction [7]. GO annotation-driven interaction inference is based on the observation that proteins localized in the same cellular compartment are more likely to interact than proteins that reside in spatially distant compartments [8]. Similarly, proteins that share a common biological process or molecular function have been found to be predictive for PPI [1]. The second criterion considers the energy of the stable protein-protein complex. There are several challenges for the correct description of the energetics at protein interfaces. Several models have been successful in identifying energetically important interactions in protein–protein interfaces [9], [10]. These models state that readily computable descriptions of electrostatic and hydrogen bonding interactions are important for suitable modeling of energetics at protein–protein interfaces. The final criterion is based on phylogenetic profile of proteins. The phylogenetic profile represents the presence or absence of proteins in various genomes. It has been observed that interacting proteins often display coordinated evolution, so that proteins with similar phylogenetic profiles are more likely to interact with each other [11], [12].

Apparently it may seem that the energy of a protein-protein complex will be minimized simultaneously along with the minimization of the difference in the phylogenetic profile and difference in the functions assigned to the proteins. However, there are evidences supporting that mere evaluation of the presence or absence of proteins over various genomes (using phylogenetic analysis) may not always confirm the desired structural and chemical properties of the residues in the active sites of the interacting proteins (for stabilizing the protein-protein complex). Similarly, different proteins may be found to possess identical phylogenetic profile but with rare functional similarity required to validate the real world interaction. Thus, it can be concluded that these three properties are mutually independent and hence need to be optimized simultaneously to validate the predicted PPIs.

To accomplish this, PPI problem has been formulated in a Multi-Objective Optimization (MOO) framework where the objectives are jointly minimized. In this paper, we study the scope of our proposed Firefly Algorithm with Non-dominated Sorting (FANS) for solving PPI prediction problem. It draws inspiration from the collective behavior and biochemical properties of fireflies. FANS is an evolutionary MOO strategy that utilizes the advantages of Firefly Algorithm (FA) [6] with the mechanisms of Pareto based ranking and crowding distance sorting [27]. A novel approach is also recommended for

978-1-4799-7492-4/15/$31.00©2015 IEEE

improving the performance of FANS by controlling the step size for random movement of each firefly according to its relative supremacy in the population. The strategy is devised with an aim to drive an inferior firefly by explorative force while a relatively better firefly is confined in its local neighborhood.

This paper is an approach to significantly improve the work proposed in [13]. In **Error! Reference source not found.**[13], the authors used Chaotic Local Search based Bat Algorithm (CLSBA), whereas the present version examines the scope of FANS. The inclusion of the two new objectives(functional similarity and force field of proteins) in the formulation and replacement of the BAT algorithm by FANS algorithm results in significant improvement in performance as indicated by three useful metrics, namely *Specificity*, *Sensitivity* and *F*1 *score*.

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. In section III, FANS algorithm is proposed. Section IV presents the experimental settings and the results. Section V concludes the paper.

II. Formulation of Protein-Protein Interaction Identification Problem

This section attempts to formulate the PPI prediction as an MOO problem. The characteristic features used for the PPI prediction are outlined to develop the objective functions, which when jointly optimized returns the desired network.

A. Predicting Protein-Protein Interactions using Functional Characteristic

The functional annotation of the proteins is very important. Function assignment of proteins is proteome-wide and is determined by the global connectivity pattern of the protein network. Consequently it is a challenging task to design a PPI network based on the functional assignments of the proteins. The assessment of the functional similarity of interacting proteins in [14] has motivated us to consider protein functions as an influential feature of PPI prediction.

The development of function annotation schemes, such as Gene Ontology (GO), have made it possible to combine the semantic information of proteins with the protein functional category . To be more specific, the protein functions are annotated using GO terms. GO is represented as a directed acyclic graph. A GO term may have multiple parents GO terms, often called ancestors. Let *annotate*(*f*) signifies the set of GO terms annotating a protein function *f*. It is apparent that if a function *f* is annotated by a GO term *g*, then *f* is also annotated with ancestor GO terms of *g*. With this representation scheme, the dissimilarity between any two protein functions, *f* and *f*′, of the PPI network can be captured by the following formulation:

$$d(f,f') = 1 - \frac{|annotate(f) \cap annotate(f')|}{|annotate(f) \cup annotate(f')|} \quad (1)$$

Here $|S|$ represents the cardinality of a set *S*. It is evident from (1) that $d(f, f')=0$ if *f* and f' have identical set of annotating GO terms. On the contrary, if there is no common GO term between *f* and f', $d(f, f')=1$. Thus, minimization of the functional dissimilarity between any two predicted interacting proteins p_i and p_j in the network can be given as follows:

$$J_1 = \sum_{i=1}^{N} \sum_{\forall p_j \in Set_i} \left(\frac{1}{\max\left(\left|F(p_i)\right|, \left|F(p_j)\right|\right)} \sum_{\forall f \in F(p_i)} \sum_{\forall f' \in F(p_j)} d(f,f') \right) \quad (2)$$

Here, *N* is the total number of proteins in the network and $F(p)$ represents the set of functions associated with protein *p*. Set_i symbolizes the set of proteins which are predicted to be interacting with protein p_i for $i=[1, N]$.

B. CHARMM (Chemistry at HARvard Macromolecular Mechanism) Force Field

In PPI identification problem, the second objective considered is to minimize the energy of a protein-protein complex for its stability. The energy comprises intra-molecular and inter-molecular energy. The intra-molecular energy is given by (3) and is estimated from the interaction energy among different functional groups inside individual proteins corresponding to bond stretching, angle bending and torsion.

$$E_{bond} = \sum_{bond} E_{bond} + \sum_{angle} E_{angle} + \sum_{dihedral} E_{dihedral} + \sum_{improper} E_{improper} + \sum_{UB} E_{Urey-Bradley} \quad (3)$$

The inter-molecular energy value, contrarily, is influenced by the chemical properties and the interaction energy between the residues present in the active sites of the interacting proteins. The inter-molecular non bonding interaction energy is calculated in terms of the Van der Waals energy and the electrostatic energy and is given by (4). Van der Waals interactions between two atoms within the active site of two proteins are approximated with a Lennard-Jones potential.

$$E_{non-bond} = \sum_{non-bond} \left[E_{Lennard-Jones} + E_{electrostatic} \right] \quad (4)$$

Hence, there is a need of energy functions, commonly known as force fields, for qualitative analysis of the PPI network conformation in large space. In this work the CHARMM force fields [15] are used for evaluating the cost of the PPI conformations.The force field which includes both inter-molecular and intra-molecular energy is given by (5)

$$V_{CHARMM} = E_{bond} + E_{non-bond}$$

$$V_{CHARMM} = \sum_{bond} K_b (b-b_0)^2 + \sum_{angle} K_\theta (\theta-\theta_0)^2 + \sum_{dihedral} K\chi \left(1+\cos(n\chi-\delta)\right) + \sum_{improper} K_\psi (\psi-\psi_0)^2 + \sum_{UB} K_{UB} (S-S_0)^2 + \sum_{non-bond} \left[\varepsilon_{i,j} \left[\left(\frac{R_{\min,i,j}}{r} \right)^{12} \right] - 2 \left[\left(\frac{R_{\min,i,j}}{r} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon r} \right] \quad (5)$$

where K_b is a constant that depends on the identity of the two atoms sharing the bond in a protein, *b* is the length of the bond and b_0 is the unstrained bond length in equilibrium. K_θ is a constant that depends on the three atoms defining the angle θ between the atoms within a protein and θ_0 is the unstrained angle in equilibrium. $K\chi$ and δ are constants that depend on the adjacent atoms, *n* is an integer that depends on the number of bonds made by atoms and χ is the value of the dihedral angle.

$K\psi$ is a constant, ψ_0 is the equilibrium improper angle and ψ is the improper angle that depends on the coordinates of the atoms. K_{UB} is the Urey-Bradley force constant, S is the distance between two atoms separated by two covalent bonds (1, 3 distance) and S_0 is the equilibrium distance. $\varepsilon_{i,j}$ is the Lennard-Jones well depth, r is the distance between atoms i and j, $R_{\min,i,j}$ is the minimum interaction radius. q_i and q_j are the charges of the two atoms and ε is the dielectric constant of the surrounding medium.

The anlysis of conformations of a PPI network is represented as follows:

$$\begin{aligned} J_2 &= \sum_{i=1}^{N} \sum_{\forall p_j \in Set_i} E_{CHARMM,i,j} \\ &= \sum_{i=1}^{N} \sum_{\forall p_j \in Set_i} \left(E_{bond,i} + E_{bond,\mathrm{j}} + E_{non-bond,i,j} \right) \end{aligned} \tag{6}$$

Here $E_{bond,i}$ represents intra-molecular energy of protein p_i and $E_{non\text{-}bond,i,j}$ represents the inter-molecular energy between proteins p_i and p_j for i, j=[1, N] but $i \bullet j$. Set_i represents the proteins predicted to be interacting with protein p_i for i=[1, N]. A few constraints incorporated in the algorithm includes the distances between the interacting residues in the active sites of the proteins, predicted to be interacting, to be between 5Å (range for the force to be applicable) and 0.65Å (to avoid steric hindrance).

C. Predicting Protein-Protein Interactions using Phylogenetic Analysis

The specific pattern of presence or absence of a protein p in a set of organisms (genomes) **G** is referred to as *phylogenetic profiles* [16], denoted by $s(p)$. It has been observed that proteins with similar phylogenetic profiles are most likely to interact with each other to perform some biological process [29]. In effect there should be an evolutionary pressure on a group of proteins to stay together which are involved in a function beneficial for an organism.

The possibility of interaction between a pair of proteins can be predicted by measuring the dissimilarity in their respective phylogenetic profiles. The above mentioned fact can be captured by the following formulation:

$$J_3 = \frac{1}{L} \sum_{i=1}^{N} \sum_{\forall p_j \in Set_i} \sum_{\forall g \in \mathbf{G}} (s_g(p_i) - s_g(p_j))^2 \tag{7}$$

Here, N is the total number of proteins in the network, L is the length of phylogenetic profile of protein p_i, $s_g(p_i)$ represents the presence or absence of protein p_i in genome $g \in \mathbf{G}$ and Set_i symbolizes the set of proteins predicted to be interacting with protein p_i for i=[1, N].

The above discussion has motivated us to recast the mathematical problem of PPI prediction problem in an MOO framework for simultaneous minimization of three objectives,

i) J_1 for assuring functional similarity between interacting protein pairs,
ii) J_2 for ensuring stability of a protein-protein complex, and
iii) J_3 for affirming similarity in phylogenetic profiles of interacting protein pairs.

D. Formation of a Protein-Protein Interaction Network

The PPI network of N proteins can be represented by a two dimensional binary matrix $\mathbf{M}$= $[M_{i,j}]$, of dimension $N \times N$. The binary elements {0, 1} of **M** signify the possible interaction or non interaction of proteins. Hence

$$M_{i,j} = \begin{cases} 1 & \text{if proteins } p_i \text{ and } p_j \text{ are predicted to interact} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

For example, the solution matrix **M** for the PPI network with N=4 can be represented by Fig. 1 and equation (9).

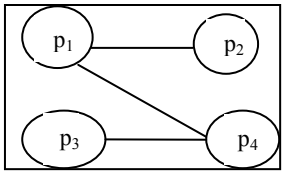

Fig. 1. Example of a PPI network with 4 proteins

This implies p_1 is connected to p_2 and p_4 while p_3 is connected to p_4. It is also evident from Fig. 1 that Set_1= {p_2, p_4}, Set_2= {p_1}, Set_3= {p_4} and Set_4= {p_1, p_3}. Thus from the solution matrix the values of different objective function values are evaluated for interacting and non-interacting proteins using (2), (6) and (7).

$$\mathbf{M} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & P_3 & P_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array} \tag{9}$$

III. FIREFLY ALGORITHM WITH NON-DOMINATED SORTING (FANS)

In Firefly Algorithm [6] with Non-dominated Sorting [17] (FANS), the position of a firefly represents a possible solution of the optimization problem and the light intensity at the position of the firefly corresponds to the fitness of the associated solution. An overview of the main steps of the FANS algorithm for jointly minimizing the N objectives are as follows:

1. Initialization: FANS involves a population P_G of NP, D-dimensional firefly positions $\vec{Z}_i(G) = \{z_{i,1}(G), z_{i,2}(G), ..., z_{i,D}(G)\}$ at the current generation G = 0.Firefly positions are randomly initialized in the range $[\vec{Z}^{\min}, \vec{Z}^{\max}]$ where $\vec{Z}^{\min} = \{z_1^{\min}, z_2^{\min}, ..., z_D^{\min}\}$ and $\vec{Z}^{\max} = \{z_1^{\max}, z_2^{\max}, ..., z_D^{\max}\}$. The j-th element of the i-th firefly position is thus initialized by

$$z_{i,j}(0) = z_j^{\min}(0) + \text{rand}(0,1) \times (z_j^{\max}(0) - z_j^{\min}(0)) \tag{10}$$

for j=[1, D]. Here rand(0,1) is a uniformly distributed random number in the range (0, 1). The k-th objective function $J_k(\vec{Z}_i(0))$ of $\vec{Z}_i(0)$ is evaluated for i= [1, NP] and k= [1, N].

2. Identification of Dominating Sets: Corresponding to each firefly $\vec{Z}_i(G)$, two sets of candidates are recognized from the current generation population P_G. The first set, denoted by $Set_i^1(G)$, comprises the position vectors of the fireflies

dominating $\vec{Z}_i(G)$. To be more specific, a member $\vec{Z}_j(G) \in Set_i^1(G)$, for j=[1, NP] but $i \neq j$, should satisfy the following conditions:

(a) $J_k(\vec{Z}_j(G)) \leq J_k(\vec{Z}_i(G))$ for k =[1, N].

(b) $J_l(\vec{Z}_j(G)) < J_l(\vec{Z}_i(G))$ for at least one $l \in$ [1, N].

Similarly, the second set, $Set_i^2(G)$, is constructed by including the position vectors of fireflies which are dominated by $\vec{Z}_i(G)$. This categorization procedure is repeated for all fireflies with i=[1, N].

3. Attraction to Brighter Fireflies: Firefly $\vec{Z}_i(G)$ is attracted towards the positions of the brighter or dominating fireflies, $\vec{Z}_j(G) \in Set_i^1(G)$ such that $\vec{Z}_j(G) \prec \vec{Z}_i(G)$. The attractiveness $\beta_{i,j}$ of $\vec{Z}_i(G)$ towards $\vec{Z}_j(G)$ is proportional to the light intensity seen by adjacent fireflies. However attractiveness $\beta_{i,j}$ decreases exponentially with the distance between them, denoted by $r_{i,j}$ as given in (11).

$$\beta_{i,j} = \beta_o \exp(-\gamma \times r_{i,j}^m), \quad m \geq 1 \tag{11}$$

where β_0 denotes the maximum attractiveness experienced by the i-th firefly at its own position (i.e., at $r_{i,j} = r_{i,i} = 0$) and γ is the light absorption coefficient, which controls the variation of $\beta_{i,j}$ with $r_{i,j}$. This parameter is responsible for the convergence speed of FA [18]. A setting of γ=0 leads to constant attractiveness while γ approaching infinity is equivalent to complete random search [18]. In (11) m is a positive constant representing a non-linear modulation index. The distance between $\vec{Z}_i(G)$ and $\vec{Z}_j(G)$ is computed using Euclidean norm as follows.

$$r_{i,j} = \| \vec{Z}_i(G) - \vec{Z}_j(G) \| \tag{12}$$

This step is repeated for i= [1, NP].

4. Movement of Fireflies: The firefly at position $\vec{Z}_i(G)$ stores its current position in its memory, symbolized by $\vec{Z}_i^{cur}(G)$ and then moves towards a more attractive position $\vec{Z}_j(G) \in Set_i^1(G)$ occupied by a brighter firefly j by following the dynamics given in (13).

$$\vec{Z}_i^{next}(G) = \vec{Z}_i^{cur}(G) + \beta_{i,j} \times (\vec{Z}_j(G) - \vec{Z}_i(G)) + \alpha \times (\vec{r} - 0.5) \tag{13.1}$$

$$\vec{Z}_i^{cur}(G) \leftarrow \vec{Z}_i^{next}(G) \tag{13.2}$$

The movement of the i-th firefly, governed by (13), is continued for j=[1, $|Set_i^1(G)|$]. The first term in the position updating formula (13.1) represents the firefly's position after the last movement. The second term in (13.1) denotes the change in the position of the firefly at $\vec{Z}_i(G)$ due to the attraction towards the brighter firefly at $\vec{Z}_j(G) \in Set_i^1(G)$.Hence it is apparent that the brightest firefly with no more attractive firefly in the current population P_G will have no motion due to the second term and may get stuck at the local optima. To circumvent this problem, the last term is introduced in (13.1).It is used for the random movement of the fireflies with a maximum step size of $\alpha \in$ (0, 1). Here $\vec{r}$ is a D-dimensional vector with its j-th component r_j being a random number uniformly distributed in the range (0, 1). After completing the motion towards brighter fireflies, the updated position of the i-th firefly is represented by $\vec{Z}_i^{next}(G)$ for i= [1, NP]. It is to be noted that the i-th firefly memorizes both the positions, the one before starting its motion and the one after completing its journey, given by $\vec{Z}_i(G)$ and $\vec{Z}_i^{next}(G)$ respectively. This step is repeated for i= [1, NP].

It is noteworthy that the random movement of a firefly (or a colony) with step size α in (13.1) in traditional FA [6] helps the population individuals to avoid local optima by their expedition proficiency. Particularly, the convergence of fireflies towards global optima greatly relies on the step size (α) profile. However in traditional FA, α is taken to be constant for all fireflies in the current population, irrespective of their fitness. Consequently, fireflies in vicinity of the global optima may deviate away (with α value greater than the requirement) and may get trapped at local optima. Contrarily, fireflies far away from the global optima in the fitness landscapes (with α smaller than necessity), may not be given any opportunity to be attracted towards the global optimum. To overcome this problem, α, used for random movement of a firefly, needs to be modulated with its relative merit over other members of the population P_G. It is realized here by setting

$$\alpha_i = 1 - \frac{\left| Set_i^2(G) \right|}{NP} \quad \text{for } i=[1, NP]. \tag{14}$$

It is evident from (14) that greater (or smaller) is the size of $|Set_i^2(G)|$ reresenting the set of members of P_G being dominated by $\vec{Z}_i(G)$, less (or more) is its corresponding step size value. It in turn ensures that the better quality fireflies (which dominates a large fraction of the population) should search in the local neighborhood with a small step size to prevent the exclusion of the global optima whereas a poor performing member (which is most frequently dominated by its competitors) should participate in the global search to explore promising regions. The k-th objective function $J_k(\vec{Z}_i^{next}(G))$ of $\vec{Z}_i^{next}(G)$ is evaluated for i= [1, NP] and k= [1, N].

4. Selection: If $\vec{Z}_i^{next}(G)$ dominates $\vec{Z}_i(G)$, $\vec{Z}_i^{next}(G)$ replaces $\vec{Z}_i(G)$ in the memory of the i-th firefly. However if $\vec{Z}_i(G)$ and $\vec{Z}_i^{next}(G)$ are non-dominated, both positions are kept in her memory P_G. This step is reiterated for i= [1, NP] and hence, a population of firefly positions is achieved with size $|P_G| \in [NP, 2NP]$.

5. Non-dominated Sorting: The population P_G, thus obtained, is sorted to Pareto fronts following the non-domination principle. All the non-dominated firefly positions of the current population are included in the optimal Pareto front, *Front_Set*(1). The second front *Front_Set*(2) is formed by the non-dominated firefly positions of the set {P_G–*Front_Set*(1)}. This process is continued and eventually identifies all the non-dominated sets.

6. Truncation of Extended Population: The non-dominated set of firefly positions are filtered from P_G (of size $NP < |P_G| < 2NP$) so as to pass NP firefly positions to the next generation population P_{G+1}, according to the ascending order of their

Pareto ranking. Let *Front_Set*(l) be the set such that by adding *Front_Set*(l) to P_{G+1}, $|P_{G+1}|$ exceeds *NP*. Then the firefly positions in *Front_Set*(l) are sorted in descending order of their crowding distance *CD*, revealing the perimeter of a hypercube formed by their nearest neighbors at the vertices in the fitness landscapes. To ensure diversity in population, the firefly positions in *Front_Set*(l) with the highest crowding distances are given priority for being included in P_{G+1} until $|P_{G+1}|$ becomes *NP*.

After each evolution, we repeat from step 2 until termination condition for convergence is satisfied. The pseudo-code for the proposed FANS algorithm with *N* objectives is given below.

Procedure FANS

Begin

1. Initialize a population P_G of *NP*, *D*-dimensional firefly position vectors $\vec{Z}_i(G)$ at generation G=0 using (10) for i= [1, *NP*].
2. Evaluate $J_k(\vec{Z}_i(G))$ for i= [1, *NP*] and k= [1, *N*].
3. **While** termination condition is not reached **do**

   Begin

   3.1. Identify the sets $Set_i^1(G)$ and $Set_i^2(G)$ corresponding to $\vec{Z}_i(G)$ for i= [1, *NP*].

   3.2. Store $\vec{Z}_i(G)$ in $\vec{Z}_i^{cur}(G)$ and perform its movement towards all $\vec{Z}_j(G) \in Set_i^1(G)$ to generate a new position $\vec{Z}_i^{next}(G)$ following the dynamics in (13) and (14) for i= [1, *NP*].

   3.3. Evaluate $J_k(\vec{Z}_i^{next}(G))$ for i= [1, *NP*] and k= [1, *N*].

   3.4. **If** $\vec{Z}_i^{next}(G) \prec \vec{Z}_i(G)$ **Then** replace $\vec{Z}_i(G)$ with $\vec{Z}_i^{next}(G)$;

   Else If $\vec{Z}_i(G) \equiv \vec{Z}_i^{next}(G)$ (i.e., non-dominated)

   Then $P_G \leftarrow P_G \bigcup \vec{Z}_i^{next}(G)$;

   End If

   Repeat the step for i= [1, *NP*].

   3.5. Sort P_G into subsequent Pareto fronts *Front_Set* using non-dominated sorting principle.

   3.6. Include the firefly positions from the Pareto fronts *Front_Set* of P_G into P_{G+1} starting from *Front_Set*(1) until *Front_Set*(l) is found such that $|P_{G+1}| + |Front\_Set(l)| > NP$. Sort the position vectors in *Front_Set*(l) in descending order of crowding distance and set $P_{G+1} \leftarrow P_{G+1} \bigcup$ top $(NP - |P_{G+1}|)$ firefly position vectors of *Front_Set*(l).

   3.7. $G \leftarrow G+1$.

   End While

End

IV. Experiments and Results

A. Competitor Algorithms and Parameter Settings

We have compared our proposed method with other computational methods including Relative Specific Similarity (RSS) method (using GO terms) [24], Fuzzy Support Vector Machines (SVM) based classifier [25], Phylogenetic Profile (PP) [26] and Chaotic Local Search based Bat Algorithm (CLSBA) [13]. We have also compared the proposed FANS based PPI prediction approach with two well known evolutionary MOO algorithms namely Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [27] and Multi-Objective Particle Swarm Optimization (MOPSO) [28]. RSS method utilizes the similarity in gene annotation of proteins to predict whether the proteins interact or not. The fuzzy support vector machine utilizes the domain-based protein interaction prediction methods and assumes that conservation of domains, over the course of evolution may contribute to the interaction of proteins. The phylogenetic profiles of proteins are compared to find interacting proteins in PP method. The same objectives given in (2), (6) and (7) are considered for the NSGA-II and MOPSO based simulations of PPI prediction problem.Best parameter settings for all competitor algorithms are employed in this paper as stated in the respective sources.

B. Performance Metrics

The PPI network obtained by the proposed method is compared with the standard PPI network obtained through experiments. Four different classes of interaction can be observed namely True Positive (*TP*), True Negative (*TN*), False Positive (*FP*) and False Negative (*FN*). Based on these four interconnection classes, we have considered well known metrics enlisted in Table-I to compare the relative performance of our proposed PPI prediction algorithm with its competitors.

TABLE I. PERFORMANCE METRICS FOR PPI PREDICTION

Performance	Expressions
Sensitivity (*Recall*)	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Positive Likelihood Ratio	$Sensitivity/(1 - Specificity)$
Negative Likelihood Ratio	$(1 - Sensitivity)/Specificity$
Precision/*Positive Predicted Value*	$\frac{TP}{TP + FP}$
Negative Predicted Value	$\frac{TN}{TN + FN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
*F*1_*score*	$\frac{2\times TP}{2\times TP+FP+FN}$
Mathews Correlation	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Receiver Operating Curve	Plot of *Sensitivity* against (1−*Specificity*)
Area Under Curve (*AUC*)	Area under *ROC* Curve

C. Database Used

In our simulation, protein interaction data of *Saccharomyces cerevisiae* (*SC*) is acquired from BioGrid [19](July 2013). It consists of 6391 proteins and 326967 interactions. The total possible interactions with 6391 proteins will be 20419245, out of which according to the dataset it contain 220066 positive unique interactions. Since the dataset only provides information about positive interacting proteins, we have considered the difference between the total possible interactions and positive interaction as non-interacting protein pairs. The Cartesian coordinates of the proteins in *SC* are acquired from Protein Data Bank [20]. The GO terms of each protein for evaluating

functional similarity is obtained from Saccharomyces Genome Database [21]. CHARMM Force field for pair of proteins in the solution network is evaluated using CHARMM GUI [22]. The phylogenetic profiles of the proteins are obtained from PhyloPat [23]. The phylogenetic profile of the proteins in *Saccharomyces cerevisiae* is generated with nine other species namely *Caenorhabdities elegans*, *Anopheles gambiae*, *Aedens aegypti*, *Drosophila melanogaster*, *Ciona savignyi*, *Ciona intestinalis*, *Tetraodon nigroviridis*, *Takifugu rubripes* and *Oryziaslatipes*.

D. Results and Performance Analysis

According to the discussion in section II, the PPI prediction problem, boils down to a MOO problem to optimally determine the existence of plausible interaction between any two proteins by jointly satisfying three individual objectives as given in (2), (6) and (7). The optimized PPI network can be obtained by decoding the best firefly position from the approximate Pareto front A (*Front_Set*(1) of FANS). It is, however, noteworthy that all position vectors in A are equally good. To select the best one among many possible candidates, the following composite measure is considered for each firefly position vector $\vec{Z}_i \in A$.

$$J(\vec{Z}_i) = J_1^*(\vec{Z}_i) \times J_2^*(\vec{Z}_i) \times J_3^*(\vec{Z}_i) \quad \text{for } i = [1, |A|] \qquad (15)$$

where $|A|$ is the number of non-dominated solutions in A and (16) represents the normalized estimate of $J_k(\vec{Z}_i) \in (0,1)$ for k=[1, 3].

$$J_k^*(\vec{Z}_i) = J_k(\vec{Z}_i) \Big/ \sum_{l=1}^{|A|} J_l(\vec{Z}_k) \qquad (16)$$

The effective non-dominated firefly position $\vec{Z} \in A$ having the smallest $J(\vec{Z}_i)$ for i= [1, $|A|$] is now identified for decoding the optimal PPI network obtained by FANS.

The Receiver Operating Curves (*ROC*s) for different PPI prediction is plotted in Fig. 2. It is a useful technique for examining the efficacy of a prediction algorithm in inferring true 'interacting' and 'non-interacting' pairs of proteins. The curve plots *Sensitivity* against (1− *Specificity*). The *ROC*s for the evolutionary methods (FANS, MOSPSO, NSGA-II and CLSBA) are drawn for function evaluations ranging from 10×10^4 to 50×10^4 and for the rest of the classification methods , they are drawn for varying thresholds (ranging from 0.3 to 0.8).

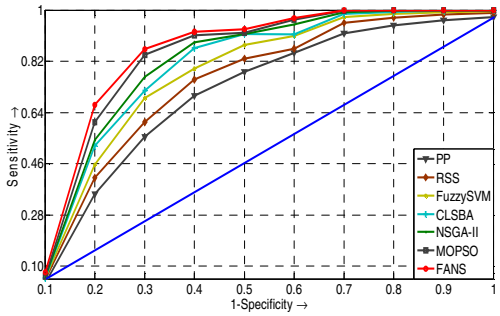

Fig. 2. ROC plot for different PPI prediction algorithms

The relative positions of the ROC curves indicate the relative efficiency of the algorithms to predict truly interacting protein-pairs. It is evident from Fig. 2 that FANS exhibit highest efficiency. A quantitative measure of the *ROC* induced efficiency of a PPI prediction algorithm can be captured by its respective Area under Curve (*AUC*). It is apparent from Fig. 2 and Table-II that *AUC* for FANS, MOPSO, NSGA-II and CLSBA employing evolutionary optimization methods have attained higher values than other competitor classification method.

TABLE II. AREA UNDER CURVE OBTAINED FROM FIG.2

FANS	MOSPO	NSGA-II	CLSBA	Fuzzy SVM	RSS	PP
0.925 (0.09)	0.908 (0.11)	0.892 (0.17)	0.888 (0.19)	0.848 (0.23)	0.7438 (0.39)	0.602 (0.92)

A plot of *Precision* versus *Recall* is given in Fig. 3. In the task of predicting PPIs, we put more emphasis on *Precision* since low reliability is one of the main weaknesses of the experimental methods; at the same time, we aim to obtain a reasonable value of *Sensitivity* (*Recall*). To assess the relative merits of the algorithms, a straight line is drawn making an angle of 45^0 with the *Recall* axis such that it passes through all the curves corresponding to all contender algorithms. In our analysis we have taken the distance from the origin to the intersecting point between the straight line and each *PROC* curve as a measure of the performance of respective algorithm. The higher the measure, the better is the performance. Symbol "≥" is used to represent the relative performance of any two algorithms. Using this convention, the ranking of the algorithms can be depicted as: FANS ≥ MOPSO ≥ NSGA-II ≥ CLSBA ≥ FuzzySVM ≥ RSS ≥ PP. Fig. 3 indicates that the proposed PPI prediction algorithm offers good level of *Precision* and *Recall*.

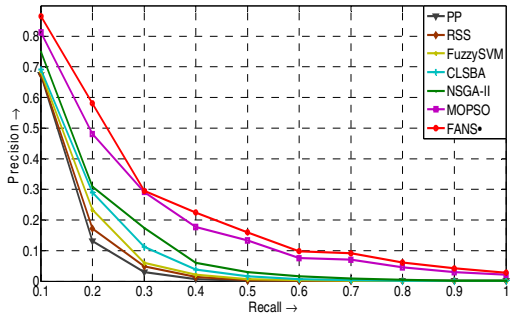

Fig. 3. PROC plot for different PPI prediction algorithms

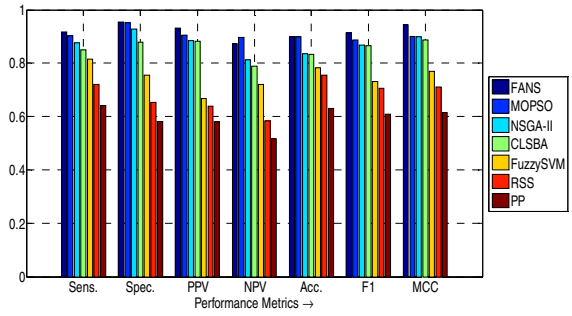

Fig. 4. Plot of performance metrics for different PPI prediction algorithms

Table-III is used to report the mean and standard deviation of best-of-run values of the performance metrics, for 25 independent runs of each PPI prediction algorithm. The

standard deviation is given in parenthesis below its respective mean value. To compare the mean of each performance metric obtained by the best and the second best algorithm, paired two-tailed t-test has been used. The statistical significance of the difference of mean of two best algorithms is provided in the last row of Table-III. Note that here "+" indicates that the t value of 24 degrees of freedom is significant at a 0.05 level of significance, whereas "–" means that the difference of means is not statistically significant, and "NA" stands for not applicable, covering cases for which two or more algorithms achieve the best metric values. The best metric value achieved by an algorithm is marked in bold.

It is interesting to see that out of 9 performance metrics, in 7 cases, FANS outperforms its nearest competitor in a statistically significant manner. In case of *Accuracy*, FANS has obtained statistically equivalent result to that of MOPSO. The mean values of the performance metrics over 25 runs of each PPI prediction algorithm are plotted in Fig. 4. A close inspection of Fig. 4 and Table-III indicates that the performance of the proposed FANS-based PPI prediction algorithm has remained consistently superior to that of the other contender approaches.

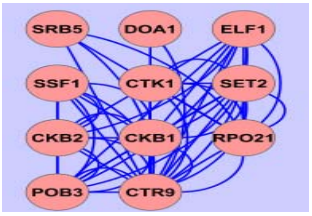

Fig. 5. Original sub-network of PPI network in yeast

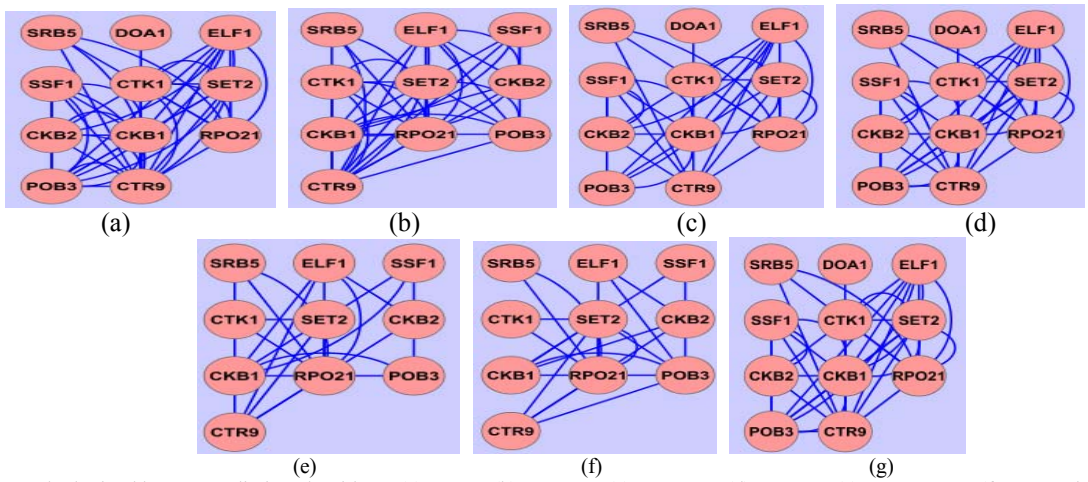

Fig. 6. Sub-network obtained by PPI prediction algorithms: (a) FANS, (b) MOPSO, (c) NSGA-II, (d) CLSBA, (e) Fuzzy SVM, (f) RSS and (g) PP

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

Algorithms	*Sensitivity*	*Specificity*	*PLR*	*NLR*	*Precision*	*NPV*	*Accuracy*	*F1_score*	*MCC*
FANS	**0.9173 (0.071)**	**0.9542 (0.111)**	**20.028 (0.107)**	**0.0867 (0.103)**	**0.9323 (0.004)**	0.8714 (0.107)	**0.8992 (0.060)**	**0.9131 (0.053)**	**0.9428 (0.002)**
MOPSO	0.9033 (0.132)	0.9511 (0.166)	18.472 (0.152)	0.1017 (0.145)	0.9049 (0.007)	**0.8974 (0.085)**	**0.8992 (0.129)**	0.8866 (0.069)	0.8999 (0.003)
NSGA-II	0.8765 (0.157)	0.9280 (0.188)	12.173 (0.164)	0.1331 (0.161)	0.8848 (0.009)	0.8125 (0.116)	0.8354 (0.138)	0.8690 (0.072)	0.8986 (0.003)
CLSBA	0.8503 (0.196)	0.8782 (0.285)	6.9811 (0.178)	0.1705 (0.186)	0.8833 (0.011)	0.7877 (0.167)	0.8324 (0.162)	0.8665 (0.078)	0.8866 (0.004)
FuzzySVM	0.8148 (0.251)	0.7551 (0.319)	3.3271 (0.429)	0.2453 (0.374)	0.6669 (0.498)	0.7217 (0.263)	0.7828 (0.521)	0.7309 (0.319)	0.7692 (0.415)
RSS	0.7214 (0.443)	0.6537 (0.587)	2.0832 (0.553)	0.4262 (0.658)	0.6399 (0.573)	0.5842 (0.289)	0.7551 (0.542)	0.7071 (0.508)	0.7105 (0.529)
PP	0.6410 (0.629)	0.5815 (0.617)	1.5317 (0.651)	0.6174 (0.792)	0.5804 (0.734)	0.5160 (0.509)	0.6313 (0.694)	0.6092 (0.896)	0.6151 (0.561)
Statistical Significance	+	+	+	+	+	–	NA	+	+

In order to visualize the performance of different algorithms for predicting PPI, we have considered a sub-network of *Saccharomyces cerevisiae* in Fig. 5 comprising of 11 proteins, namely POB3, SET2, CTR9, RPO21, CKB1, CKB2, CTK1, SSF1, ELF1, DOA1 and SRB5.The three objectives were calculated for one interacting protein pair, POB3–SET2,and also for one non-interacting pair, POB3–CLN1. It is observed that value of functional dissimilarity objective is lower for the interacting protein pair than non-interacting proteins indicating that reduction in functional dissimilarity is a good indicator of interaction. A lower value of CHARMM energy for the interacting pair signifies stable protein-protein interaction. The values of difference in phylogenetic profile are observed to be same for both interacting and non-interacting protein pairs substantiating that phylogenetic profile alone is not a good characteristic feature for identifying protein interaction. The predicted PPIs for the same sub-network obtained using seven competitor algorithms is pictorially represented in Fig. 6.

Comparing Fig. 6 with Fig. 5, it is apparent that FANS based method outperforms other competitors in predicting correct PPIs.

V. Conclusion

In this paper the PPI problem is solved with an MOO approach. The proposed FANS algorithm is used to predict PPI network. Here we have analyzed the effect of three essential characteristic features of proteins on competently predicting PPI network, including i) functional similarity, ii) minimal intra- and inter- molecular energy and iii) similarity in phylogenetic profiles of interacting protein pairs. The proposed method performs well on unbalanced data. Experiments undertaken reveal the superiority of the proposed method over its state-of-art contenders in predicting PPIs in a statistically significant manner.

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.

References

[1] Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O., Protein function in the post-genome era, Nature 405 (2000) 823–826.

[2] Qi, Y., Bar-Joseph, Z., & Klein-Seetharaman, J., (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics, Volume 63, Issue 3, 490-500.

[3] 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.

[4] 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.

[5] 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.

[6] Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In Stochastic algorithms: foundations and applications (pp. 169-178). Springer Berlin Heidelberg.

[7] Patil, A., & Nakamura, H. (2005). Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC bioinformatics,6(1), 100.

[8] Shin, Chang J., Simon Wong, Melissa J. Davis, and Mark A. Ragan. "Protein-protein interaction as a predictor of subcellular location." BMC systems biology3, no. 1 (2009): 28.

[9] Kortemme, Tanja, and David Baker. "A simple physical model for binding energy hot spots in protein–protein complexes." Proceedings of the National Academy of Sciences 99, no. 22 (2002): 14116-14121.

[10] Guerois, Raphael, Jens Erik Nielsen, and Luis Serrano. "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." Journal of molecular biology 320, no. 2 (2002): 369-387.

[11] Goh, Chern-Sing, and Fred E. Cohen. "Co-evolutionary analysis reveals insights into protein–protein interactions." Journal of molecular biology 324, no. 1 (2002): 177-192.

[12] Ramani, Arun K., and Edward M. Marcotte. "Exploiting the co-evolution of interacting proteins to discover interaction specificity." Journal of molecular biology 327, no. 1 (2003): 273-284.

[13] Chowdhury, A., Rakshit, P., Konar, A., & Nagar, A. K. (2014, July). A modified bat algorithm to predict Protein-Protein Interaction network.In Evolutionary Computation (CEC), 2014 IEEE Congress on (pp. 1046-1053).IEEE.

[14] Hou, Jingyu, and Xiaoxiao Chi. "Predicting protein functions from PPI networks using functional aggregation." Mathematical Biosciences 240, no. 1 (2012): 63-69.

[15] Rakshit, Pratyusha, Amit Konar, Archana Chowdhury, Eunjin Kim, and Atulya K. Nagar. "Muti-objective evolutionary approach of ligand design for protein-ligand docking problem."In Evolutionary Computation (CEC), 2013 IEEE Congress on, pp. 237-244.IEEE, 2013.

[16] Rakshit, P., Das, P., Chowdhury, A., Konar, A., & Janarthanan, R. (2012, July). Evolutionary approach for designing protein-protein interaction network using artificial bee colony optimization. In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on(pp. 1-8). IEEE

[17] Srinivas, N., & Deb, K. (1994). Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, *2*(3), 221-248.

[18] Roy, A. G., Rakshit, P., Konar, A., Bhattacharya, S., Kim, E., & Nagar, A. K. (2013, June). Adaptive Firefly Algorithm for nonholonomic motion planning of car-like system. In *Evolutionary Computation (CEC), 2013 IEEE Congress on* (pp. 2162-2169). IEEE.

[19] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., &Tyers, M. (2006).BioGRID: a general repository for interaction datasets. Nucleic acids research, 34(suppl 1), D535-D539.

[20] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., ... & Tasumi, M. (1978). The Protein Data Bank: a computer-based archival file for macromolecular structures. Archives of biochemistry and biophysics, 185(2), 584-591.

[21] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., ...& Wong, E. D. (2011). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research, gkr1029.

[22] Jo, S., Kim, T., Iyer, V. G., &Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. Journal of computational chemistry,29(11), 1859-1865.

[23] Hulsen, T., Groenen, P. M., de Vlieg, J., &Alkema, W. (2009).PhyloPat: an updated version of the phylogenetic pattern database contains gene neighborhood. Nucleic acids research, 37(suppl 1), D731-D737.

[24] Wu, X., Zhu, L., Guo, J., Zhang, D. Y., & Lin, K. (2006). Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. Nucleic acids research, 34(7), 2137-2150.

[25] Chiang, J. H., & Lee, T. L. (2008). In Silico Prediction of Human Protein Interactions Using Fuzzy–SVM Mixture Models and Its Application to Cancer Research. Fuzzy Systems, IEEE Transactions on, 16(4), 1087-1095.

[26] Sprinzak, Einat, and Hanah Margalit."Correlated sequence-signatures as markers of protein-protein interaction." Journal of molecular biology 311, no. 4 (2001): 681-692.

[27] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on*, *6*(2), 182-197.

[28] Coello C. C. A., and Lechuga M., "MOPSO: A Proposal for Multiple Objective Particle Swarm Optimization," in *Proceedings of IEEE Congress of Evolutionary Computation*, vol. 2, May, 2002, pp. 1051–1056.

[29] Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.