Wireless Personal Communications
https://doi.org/10.1007/s11277-021-08851-z

Improved Firefly Optimization for Pairwise Network Alignment with its Biological Significance of Predicting GO Functions and KEGG Pathways

R. Ranjani Rani[1] · **D. Ramyachitra**[2]

Accepted: 6 August 2021

Abstract
Global pairwise biological network alignment is a pervasive technique in bioinformatics and computational biology. Even now, the computation of network alignment is a challenging effort for delivering an efficient and statistically significant results. Thus, the optimization algorithms have been used to get the precise results of protein network alignment. In this work, an Improved Firefly Optimization Algorithm method was used to align the biological protein networks in a pairwise technique which resulted in an optimal solution. By utilizing the final outcome of network alignment, the function of proteins in a network and KEGG pathways was also obtained and found that the aligned proteins have more functions that are common in nature.

Keywords Protein network alignment · IFOA method · Synthetic networks · Real-world networks · Symmetric substructure score · Gene-ontology Precision

Abbreviations

PPI	Protein–protein interaction
GRAAL	GRAph Aligner
S^3	Symmetric substructure score
IFOA	Improved Firefly Optimization Algorithm
GoP	Gene-ontology precision
SANA	Simulated annealing network aligner

1 Introduction

Bioinformatics field works with a large amount of data and it is very difficult to discover new biological knowledge using generic algorithms. Due to the drastic development in the large-scale computation, many metabolic networks, signalling networks, protein–protein

✉ D. Ramyachitra
jaichitra1@yahoo.co.in

[1] Department of Computer Science, PSG College of Arts and Science, Coimbatore, India

[2] Department of Computer Science, Bharathiar University, Coimbatore, India

interaction (PPI) networks and many other networks are built and deposited in several databases. However, the biological function annotations of these networks are not much identified [1].

Aligning the PPI networks of any unique organism is vital. It has a widespread application in detecting orthologous proteins, protein complex detections, predicting gene functions, shared pathways, conserved functional modules and evolutionary common ancestor and many more. In concise, the network alignment goal is to identify the analogy and differences between various biological networks.

Primarily, the approaches of network alignment can be categorized into two techniques. One is a local network alignment (LNA), which compares a part of the protein network (sub-network) to detect the topological similarity. But here the aligned sub-networks are overlying which gives rise to many-to-many indefinite relationship mappings [2, 3]. To overcome this drawback, another alignment technique, global network alignment (GNA) that compares the entire protein network, which results in a one-to-one relationship mapping of proteins [4] is used. Also, the other two major approaches for aligning protein networks are classified into two types. One is pairwise network alignment, which aligns with only pairs of networks. Another approach is the multiple network alignment, which aligns with more than two protein networks simultaneously to infer the biological functions of diverse organisms.

Generally, protein network alignment is used to detect how precisely the similarity should be identified and matched between different protein networks. The similarity detected can be either topological or biological function similarity. The similarity of either structure of a protein network, topology or edges around the proteins in the PPI network is called topological similarity and the similarity is calculated by using protein functions is called functional similarity [5–7].

Most of the network aligners are classified into two main categories. One is the most popular two-stage alignment method. In the first stage, the similarity matrix was calculated to show the similarity between two networks, and in the second stage, an identical pair of nodes was aligned primarily by giving the input of similarity matrix to 'seed-and-extend' method [8].

Some of the aligners are IsoRank [9], GRAAL (GRAph ALigner) which has the family of algorithms that use the graphlet counts to calculate mathematically hard topological node similarity scores. C-GRAAL (common neighbors based global biological network alignment) [10], L-GRAAL (Lagrangian based global biological network alignment) [11], MI-GRAAL (Matching based integrative global protein interaction network alignment [12], GHOST [13], ModuleAlign [14], PROPER (PROtein–protein interaction network alignment based on PERcolatin) [15] and Ulign [16].

The second most popular alignment category is search-based alignment techniques. They practice metaheuristic algorithms to improve a population of alignments. Some of the example aligners are MAGNA [17], MAGNA++ [18], Artificial Bee Colony optimization (NABEECO) [19], Ant Colony Optimization (ACOGNA) [4], Optnetalign [8], SUMONA [20], Particle Swarm Optimization (PSONA) [1], PISwap [21], SANA (Simulated annealing network aligner) [22], and WAVE (Weighted Alignment VotEr) [23].

Evolving an accurate network alignment of two or more protein networks obtained from different organisms is a hard-computational task which is considered as an NP-complete optimization problem. It is difficult to explore the tradeoff among the biological functions and topological similarity. When dealing with large networks the computational complexity is a big threat. Thus, this paper deals with a stochastic optimization technique for the global network alignment to distinguish the tradeoff between the topological and biological

function similarity with better alignment results. The remaining segments of the paper are ordered as follows: Sect. 2 defines the materials and methods of proposed IFOA method for global network alignment. Section 3 shows the analyses of the experiments accomplished on different datasets, comparison of results with existing methods and finally, Sect. 4 discusses the implementation and its outcome of the proposed approach and also deliberates the conclusion of the paper and recommendations for future enhancement.

2 Materials and Methods

2.1 Firefly Optimization Algorithm

The projected IFOA (Improved Firefly Optimization Algorithm) method was employed to find, the optimal solution for the pairwise global network alignment. The firefly optimization algorithm is a swarm intelligence and a novel meta-heuristic optimization algorithm stimulated by the population of fireflies twinkling behaviour. It was proposed by Yang [24] and it has been evidenced to be an effective optimization algorithm to explore the global optimum solution.

The communication among the fireflies is directed by the following rules:

1. Every single firefly is unisex; therefore, they are fascinated with alternative fireflies irrespective of their sex.
2. The distance and brightness between the fireflies direct the attractiveness. For any pair of firefly, the attractiveness is directly corresponding to the brightness of a firefly. Hence, the less bright firefly will transfer towards the brighter firefly. As the distance raises between these pair of fireflies, then both the attractiveness and brightness of the firefly will be decreased as they are directly proportional. If none of the fireflies is brighter than a distinct firefly, then it will travel arbitrarily.
3. The firefly's brightness is denoted as light intensity, (I) and it is affected by the significance score of the fitness function.

2.1.1 Major steps involved in Firefly Optimization Algorithm (FOA) method:

(a) **Unaligned biological network:** This work portrays the pairwise global network alignment which is the distinctive alignment among two biological networks, by exploring the maximal sequence, topological and biological function similarity among them. In this work, the two biological PPI networks, namely $N_1\ (Vt_1, Ed_1)$ and $N_2\ (Vt_2, Ed_2)$ with Vt_1, Ed_1 and Vt_2, Ed_2 as the group of nodes which represents the proteins and group of edges that represents the interactions between the proteins respectively. Let $m = |Vt_1|$ and $n = |Vt_2|$ represents the number of protein elements in Vt.

(b) **Initial population of networks:** Initially the number of protein network population is formed by constructing an alignment between two protein networks randomly. Always, the network will be aligned when $|Vt_1| \leq |Vt_2|$ where all the proteins in Vt_1 can be mapped with Vt_2. Here, the alignment of two protein networks N_1 and N_2 are defined as an injective function $f: Vt_1 \rightarrow Vt_2$, where every single protein of Vt_1 is mapped exclusively with a protein of Vt_2 [17]. The example illustration of global pairwise network alignment used in this paper is depicted in Fig. 1.

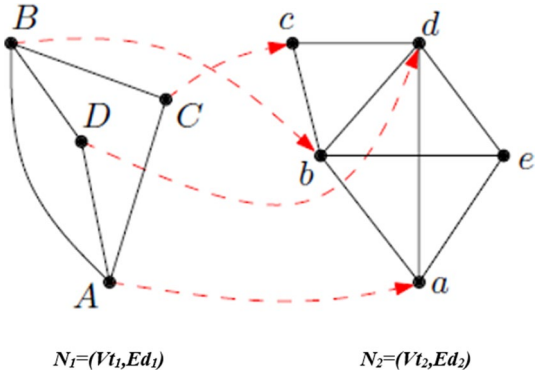

Fig. 1 Illustration of Global Pairwise Network Alignment

In Fig 1, the global pairwise network alignment between two protein networks are N_1 and N_2. The N_1 protein network has Vt_1 vertices and Ed_1 edges also the N_2 protein network has Vt_2 vertices and Ed_2 edges. Here every single protein of Vt_1 is connected exclusively with a protein of Vt_2 such as A→a, B→b, C→c, D→d.

Here, each firefly signifies a protein network alignment between two networks that gives a candidate solution.

(c) **Fitness calculation:** Subsequently, once the network alignment is built, the fitness (brightness of the firefly) of the particular network alignment is calculated. In this paper, the symmetric substructure score (S^3) and Gene-ontology Precision (GoP) measures were the criteria to design the fitness function.

- **Brightness:** For a maximization problem, the brightness of a firefly directly corresponds to the significance score of the objective function (fitness function) which is denoted by: $LI(x) = f(x)$, where $f(x)$ is the objective function to calculate the performance measures such as S^3 and GoP. The formula for calculatindg S^3 and GoP is given in Eqs. (10) and (11).

(d) **Find the best individual:** Once the evaluation of fitness was accomplished, the objective function scores are arranged in an ascending order and the best individual (the most attractive) of protein network alignment that has the maximum objective function is identified among all other individuals.

- **Attractiveness:** Here, the movement of one firefly (protein network alignment) towards other fireflies is based on the attractiveness of other fireflies. Here the movement represent the changing of the alignment. The attractiveness of the fireflies varies with the brightness which in turn is related to the objective function. The light intensity is determined as the quantity of brightness transferred and it differs with the distance between proteins. It is known to differ inversely with the square of increasing distance and is given by

$$LI = LI_0 e^{-\gamma d} \tag{1}$$

where LI_0 is the opening light intensity and γ is the coefficient of light absorption which regulates the decline of light intensity and d is the distance between two fireflies.

Certainly, larger the distance d ab (distance between the firefly a and b), the dim light of the fireflies can perceive from each other. The better alignment cannot be obtained when the distance is larger between the fireflies.

The main form of firefly attractiveness is denoted by $\beta(d)$,

$$\beta(d) = \beta_0 e^{-\gamma d^2} \tag{2}$$

where d is the space gap between two fireflies a and b, β_0 is the firefly attractiveness for $d = 0$ and γ is a coefficient constant of light absorption.

- **Distance:** The Euclidean distance is used in this paper to calculate the space gap between two fireflies a and b at x_a and x_b and it is denoted by

$$d_{a,b} = ||x_a - x_b|| = \sqrt{\sum_{k=1}^{r} \left(x_{a,k} - x_{b,k}\right)^2} \tag{3}$$

where $x_{a,k}$ stands for the kth component node of the spatial coordinate of aligned graph x_a of ath firefly and $x_{b,k}$ stands for the kth component node of the spatial coordinate of aligned graph x_b of bth firefly.

(e) **Realignment of network towards its attractiveness:** The network alignment of all other remaining individuals is realigned (movement) towards the best individual to get better objective function scores.

- **Movement:** The migration of a firefly a is fascinated to another more attractive (brighter) firefly b is defined by

$$x_a^{t+1} = x_a^t + \beta\_0 e^{-\gamma d_{ab}^2}\left(x_b^t - x_a^t\right) + \alpha(rand - 1/2) \tag{4}$$

The above equation is based on the attraction and the randomization parameter where α is termed as random weight parameter or size of the step which lies in the interval [0–1] and the *rand* is a random value initiator with uniform distribution in [0,1] and t represents the current iteration.

(f) **Check the condition:** The above process is accomplished till the algorithm reaches the maximum number of generations or the objective function scores get diverged or the objective function score is identical for a few consecutive generations.

(g) **Stop the process:** Once the condition becomes true, the process has been terminated.

2.2 Improved Firefly Optimization Algorithm (IFOA)

The goal of any population-based algorithm is to ensure the balance between exploitation and exploration to identify the optimal solution. So, to accommodate the above specifications, it is essential that population should initially diverge and explore the entire search space instead of probing around the best firefly. Then in future steps, population starts to converge towards the best firefly and try to identify the optimum solution around it. In addition, the early convergence around the local minima is the common issue for all population-based algorithms [25].

2.2.1 Balance Between Exploitation and Exploration

The traditional FOA method suffers from the appropriate balance between exploitation and exploration. To rectify the disadvantages of traditional FOA, a novel improved firefly optimization algorithm (IFOA) is proposed by presuming that fireflies are arbitrarily generated and dispersed in search space and initially every firefly updates its location according to Eq. (8) and its corresponding fitness function is computed. Later, the updated firefly's fitness function is compared to remaining fireflies and updates its location in further process. Fundamentally, initial location of firefly is modified and employed for subsequent calculations.

In this method the exploitation and exploration of an algorithm is controlled by including a new parameter termed '*E*'. The value of '*E*' chooses whether firefly travel towards or away from the best firefly (b_f). Moreover, the '*E*' depends on the value of *e* index which lies in the range of [-3e,3e] where *e* is the linear decreasing function as given in the Eq. (6) and it modifies its value depending on the current iteration and maximum number of iterations. In the initial stage, when $E < 1$ the firefly deviate and move away from the best firefly (b_f) in desire of getting improved best firefly (b_f) in the search space. During second stage, when $E > 1$, fireflies updates its location towards the best firefly (b_f) that directs to more convergence. On that account, the proper balance between the exploitation and exploration is sustained by adaptively varying the parameter '*E*'.

$$E = 3 * e * r_1 - e \tag{5}$$

$$e = 2 - M * \left((2)/M_{Generation}\right) \tag{6}$$

$$alpha = abs(b_f - x_a^t) \tag{7}$$

$$x_a^t = best - E * alpha \tag{8}$$

where r_1 is a random value in the range [0,1], M represents the current iteration and $M_{Generation}$ represents the maximum number of iterations.

2.2.2 Enhanced Exploration Ability

At the beginning, fireflies are arbitrarily produced and separated by larger distance by employing a traditional FOA method which leads to increases in diversity. But, after some iteration, the diversity gets decreased, because the fireflies arrive closer to each other and try to identify the optimal solution around best solution and consequently increase the probabilities of falling in local optima. The traditional FOA's search ability gets reduced when premature convergence occurs. So to recover this limitation, the firefly equation is altered by including one extra term that comprises of the difference between two random fireflies in search space. This will lead to more random walk in search space when firefly updates its location towards brighter firefly. The altered equation is given in Eq. (9).

$$x_a^{t+1} = x_a^t + \beta_{0e^{-\gamma d_{ab}^2}}(x_b^t - x_a^t) + \alpha(\text{rand} - 1/2) + \text{rand} * (x_p^t - x_q^t) \tag{9}$$

where x_p^t and x_q^t are randomly chosen fireflies in search space.

Table 1 Various parameters used for the proposed IFOA method

Parameter	Notation in Algorithm
Brightness (*LI*)	Objective function of a problem
Alpha (α)	Random weight parameter
Beta (β)	Attractiveness of a firefly
Gamma (γ)	Absorption coefficient of a firefly
Maximum generations ($M_{Generation}$)	Number of Iterations
Number of fireflies	Number of Population
Dimension (r)	Problem dimension space
Distance (d)	Exact distance length from the light source

The parameter selection used for the IFOA was displayed in Table 1 with its explanation:

The general pseudocode for IFOA method has been depicted in Fig. 2.

The Schematic representation of the proposed IFOA method for pairwise biological network alignment is depicted in Fig. 3.

From Fig. 3, it is inferred that initially an unaligned two biological protein networks have been taken as an input. The pairwise network alignment between these unaligned protein networks has been done using the IFOA optimization method. As a result of IFOA method, an aligned protein network has obtained where every single protein of Vt_1 is mapped exclusively with a protein of Vt_2. Consecutively, once the pairwise protein network alignment is built, the similarity of the particular network alignment is calculated using the symmetric substructure score (S^3) and Gene-ontology Precision (*GoP*) measures. Finally, with these similarity measures the biological significance of attained network alignments

```
Begin:
Load the algorithm parameters f(x), x = (x1, x2 ...,  xs )^M
Randomly create a preliminary population of fireflies xa (a = 1, 2,..., m)
Compute the fitness value of each firefly (x);
Initialize the parameters like β, γ, α, max generation;
While (M < MGeneration)
   For a=1:n
      Update solution using Equation (7) and (8);
      Evaluate the fitness LI(x)
    For b=1:n
     If (Ib > Ia)
       Move firefly Ia towards Ib using Equation (9)
     Else
       Compute fitness of new solution;
     End if
    End for b
   End for a
Rank all the fireflies and update the final best
End while
Start the next process for the best results achieved
```

Fig. 2 Pseudocode for Proposed IFOA method

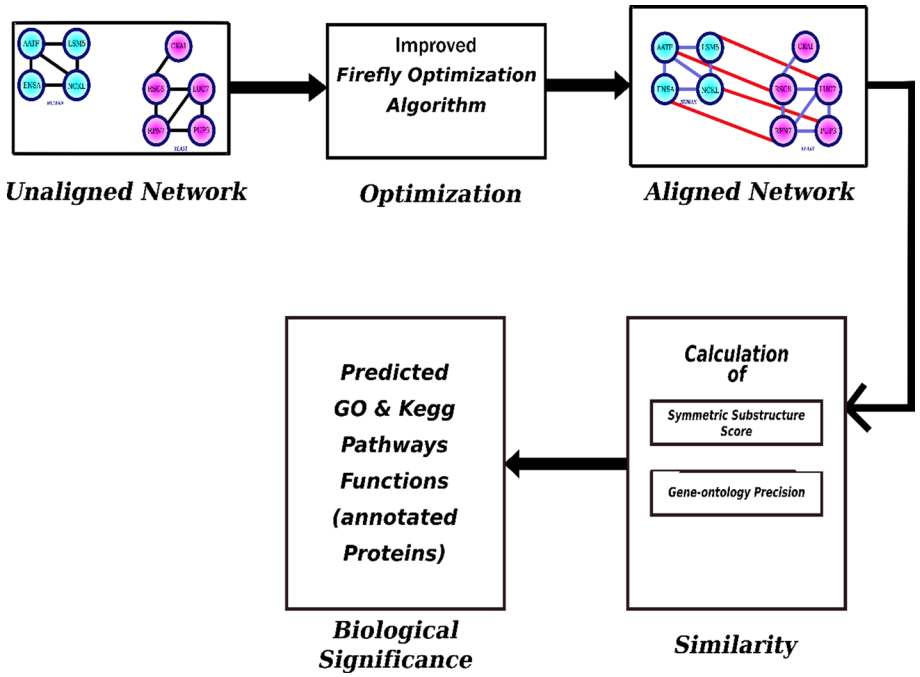

Fig. 3 Schematic Representation of the Proposed IFOA method

identifies the gene ontology function and KEGG pathways in a pair of aligned protein network.

The schematic representation of the major steps involved in the proposed IFOA method is shown in Fig. 4.

In Fig. 4, the key steps comprised in the proposed IFOA method has been depicted clearly. In the First step, an input of two unaligned biological networks have been taken. Secondly, the number of protein network population is formed by constructing an alignment between two protein networks randomly. Consecutively, once the network alignment is built, the fitness (brightness of the firefly) of the particular network alignment is calculated using S^3 and *GoP* measures.

Once the evaluation of fitness was accomplished, the best individual was found using the intensity based on its attractiveness. The above process is accomplished by checking the condition that algorithm reaches maximum number of iterations or brightness gets converged. Once the condition becomes true, the process has been terminated. Else, the network alignment of all other remaining individuals is realigned (movement) towards its attractiveness.

2.3 Performance Measures

In this work, the IFOA method was used for biological global pairwise network alignment. The maximization of alignment quality measures such as Symmetric substructure score *(S^3)* and Gene-ontology Precision (*GoP*) leads to explore the optimal solution.

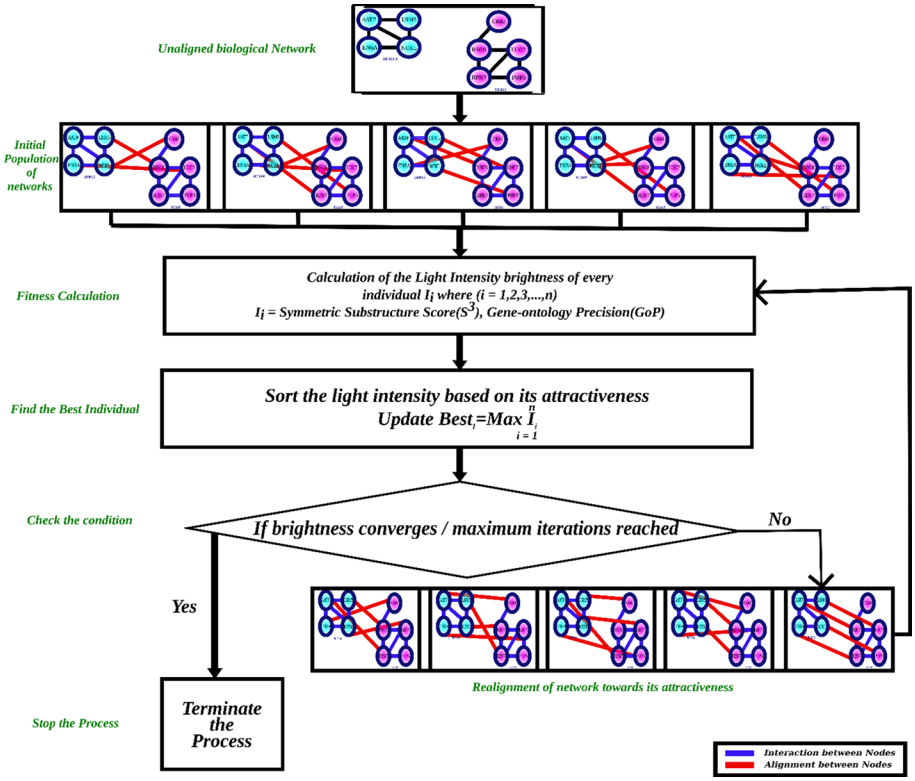

Fig. 4 Schematic representation of the major steps involved in the proposed IFOA method

Let f: $Vt_1 \rightarrow Vt_2$ be an alignment among two networks N_1 (Vt_1, Ed_1) and N_2 (Vt_2, Ed_2). If $P \subseteq N_2$, let N_2 $[P]$ be the induced subnetwork of N_2 with a set of node P. Similarly, if Q is a subnetwork of N_2, let $Ed(Q)$be its set of edges. Let $f(Ed_1)=\{(f(a), f(b)) \epsilon\ Ed_2 : (a,b) \epsilon\ Ed_1\}$, and let $f(Vt_1)=\{f(b) \epsilon\ Vt_2 : b \epsilon\ Vt_1\}$.

2.3.1 Symmetric Substructure Score (S^3)

Symmetric substructure score (S^3) is defined with reference to both source and target network. The $\left|f\left(Ed_1\right)\right|$ is described as the ratio of the total quantity of conserved edges with function f. The $\left|Ed_1\right|$ is defined as the total quantity of edges in the source network [26] and $\left|Ed\left(N_2\left[f\left(Vt_1\right)\right]\right)\right|$ is described as the total quantity of edges present in the sub-network of N_2 which are aligned to the nodes in N_1. Automatically, if N_1 and N_2 $[f(Vt_1)]$ are overlapped into a complex graph, then the denominator of S^3 is the total quantity of unique edges in this complex graph. The percentage of the S^3 value of the function f is 100% only when the function is faultless [17].

$$S^3(f) = \frac{\left|f\left(Ed_1\right)\right|}{\left|Ed_1\right| + \left|Ed\left(N_2\left[f\left(Vt_1\right)\right]\right)\right| - \left|f\left(Ed_1\right)\right|} \tag{10}$$

2.3.2 Gene-ontology Precision (*GoP*)

Gene-ontology Precision (*GoP*) is used to evaluate the accuracy of the aligned network with biological relevance [15]. It is defined as the capability of an alignment that involves similar functional gene ontology functions such as biological process, molecular function and cellular component and it is expressed as:

$$GoP(f) = \frac{|GO(p) \cap GO(q)|}{|GO(p) \cup GO(q)|'} \tag{11}$$

where $p \in N_1$ and $q \in N_2$ for aligning pairs of nodes in two protein networks. And *GO(p)* and *GO(q)* denotes gene ontology annotations related to the protein p and q.

Now the summation of Gene-ontology Precision (*GoP*) of all aligned proteins in π is defined as

$$GoP(n) = \sum_{p \in N1} GoP(p, n(p)) \tag{12}$$

3 Results

The proposed IFOA method was examined with benchmark datasets that used to address the protein biological network alignment. It is a high eminence standard protein network dataset to find the weak and strong facts of numerous network alignment techniques in pursuance to analyze the efficiency of the algorithms. The datasets were collected from various benchmark databases like STRING [27], INTACT [28], BIOGRID [29] and Synthetic protein interaction network data [17] to observe the proposed algorithm against the other existing methods.

This paper employed two prevalent sets of network data such as "synthetic networks" in which the node mapping is known in prior and "real-world networks" in which the node mapping is unknown. The benchmark synthetic networks dataset comprises of high-confidence yeast protein–protein interaction network [17] which has 1,004 nodes and 8,323 protein interactions, and also 5 noisy networks have been built by including a proportion of low-confidence protein interactions to the high-confidence network from same data set. The noise percentage has been varied from 5 to 25% in intervals of 5%. The primary high-confidence network has been aligned to each of 5 noisy networks, which results in a five network aligned pairs. The real-world networks dataset comprises 21 different pairs of benchmark protein networks. All the 21 protein networks of different organisms are classified into three categories. The first category is for bacteria, pathogen and virus, the second category is for worms and fly, and the final third category is for mammals and fungus. The protein networks used in this research are described in Table 2. Further, the efficacy of the projected IFOA method has been evidenced to be better by comparing with different network alignment methods, namely, IsoRank, MI-GRAAL, GHOST, MAGNA, PROPER, NABEECO, ACOGNA and PSONA. The network alignment performance measures, namely, S^3 and *GoP* were considered for an optimal solution.

The algorithm's initial members of a population are various random alignments of the pair of protein networks. The proposed IFOA method was tested with different population

Table 2 Common benchmark protein networks from various databases

S.NO	Datasets (Real-world datasets)	Category
1	*Campylobacter jejuni (C. jejuni)—Escherichia coli (E. coli)*	Bacteria, Pathogen and Virus
2	*Mesorhizobium loti (Meso)—Synechocystis (Syne)*	
3	*Francisella tularensis (Fratt)—Bacillus anthracis (Bacan)*	
4	*Bacillus anthracis (Bacan)—Yersinia pestis (Yerpe)*	
5	*Herpes Simplex Viruses (HSV)—Epstein-Barr Virus (EBV)*	
6	*Kaposi's sarcoma-associated herpesvirus (KSHV)—Varicella zoster virus (VZV)*	
7	*Kaposi's sarcoma-associated herpesvirus (KSH.**V)—Maize Chlorotic Mottle Virus (MCMV)*	
8	*Caenorhabditis Elegans (CE)—Mus Musculus (MM)*	Worms and Fly
9	*Caenorhabditis Elegans (CE)—Drosophila melanogaster (DM)*	
10	*Caenorhabditis Elegans (CE)—Homo Sapiens (HS)*	
11	*Caenorhabditis Elegans (CE)—Saccharomyces Cerevisiae (SC)*	
12	*Drosophila melanogaster (DM)—Homo Sapiens (HS)*	
13	*Drosophila melanogaster (DM)—Saccharomyces Cerevisiae (SC)*	
14	*Drosophila melanogaster (DM)—Mus Musculus (MM)*	
15	*Yeast-Human*	Mammals and Fungus
16	*Saccharomyces Cerevisiae (SC)—Rattus Norvegicus (RAT)*	
17	*Saccharomyces Cerevisiae (SC)—Drosophila melanogaster (DM)*	
18	*Homo Sapiens (HS)—Mus Musculus (MM)*	
19	*ULITSKY-HPRD*	
20	*Human-HPRD*	
21	*Homo Sapiens (HS)—Rattus Norvegicus (RAT)*	
Synthetic datasets		
22	Krogan_2007 Yeast dataset with noise levels	Fungus

sizes such as 200, 500, 1000, 2000, 5000, 10,000, 20,000, 25,000, 30,000. The fitness function of the algorithm is the performance measure of network alignment quality, namely, S^3 and *GoP*. The termination criterion is fixed, if the best solution formed in each generation remains alike for 100 successive generations or the maximum quantity of generations is reached. The proposed IFOA method has been executed for several generations, from 0 to 3000 in the increment interval of 200.

The grouping of the initial population, the size of the population, the number of generations and the fitness function of the alignment results in one ultimate alignment.

Consequence of the initial population: The initial population of alignments is constructed randomly.

Consequence of the size of the population: The largest population size is always preferred to gain more alignments. Here, the largest population size is 30000.

Consequence of the number of generations: A large number of generations are always preferred to increase more variety of alignments. This may result in better alignments and in the improvement of the results.

Consequence of fitness function: The random weight parameter α has the range of [0,1]. Because of the existence of the α parameter in the fitness function, the actual alignment and the similarity values get varied across different α values and also across different runs of the algorithm of same α value. To know the specific random value that gives some better results, it requires a trial and error technique on the manipulator side and the complete recalculation of alignment is done when the α gets changed. Once the value is found, it can be used for other population sizes and generations. In this work, the highest topological and biological function similarity values have been achieved when the value of α is 0.8.

The results of network alignment with the diverse values of the population and iterations were given in supplementary results. The best appropriate alignment from the previous generations is conveyed as the final alignment. The comparison of S^3 scores of all three categories of real-world benchmark protein network pairs of proposed and existing algorithms is shown in Figs. 5, 6 and 7. When the value of S^3 is higher, it is inferred that the better alignment of a network is predicted based on the both, dense to the sparse network and vice versa.

From Figs. 5, 6 and 7, it is inferred that the proposed IFOA method has been performed better and produced the better S^3 scores when compared to all other popular existing methods in all three categories of the real-world protein networks.

The comparison of *GoP* scores of all three categories of real-world benchmark protein network pairs of proposed and existing algorithms is shown in Figs. 8, 9, and 10.

From Figs. 8, 9, and 10, it is clearly inferred that the proposed IFOA method has good *GoP* scores when compared to other methods and also it is inferred that although the IsoRank method has less S^3 scores when compared to others, it has good average *GoP* scores when compared to MI-GRAAL, GHOST and MAGNA.

The robustness of the proposed IFOA method has been evaluated by adding noise to the benchmark high-confidence yeast protein interaction network called synthetic networks. The comparison of S^3 scores for all 5–25% of noise levels of yeast protein interaction networks have been depicted in Fig. 11.

As predicted, a higher noise in a network has a more undesirable effect on the alignment results of most of the techniques. However, the proposed IFOA method still reliably outperforms all other existing techniques beyond different noise levels.

The proposed IFOA method was assessed using the non-parametric test, namely Wilcoxon Matched-signed Rank test between each pair of techniques in order to produce statistical significance. The transformation between various existing methods and the proposed

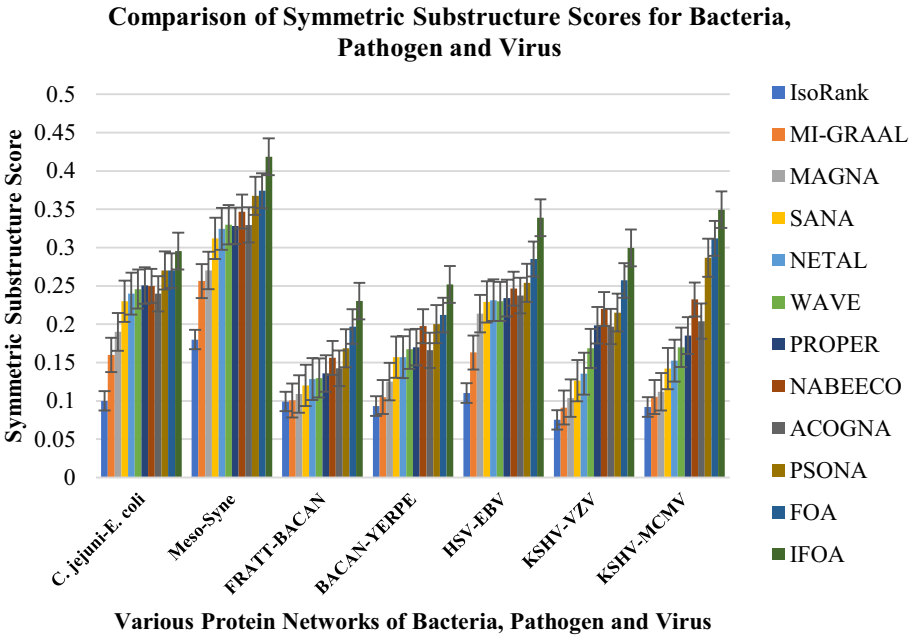

Fig. 5 Comparisons of average S^3 scores for bacteria, pathogen and virus with proposed and existing algorithms

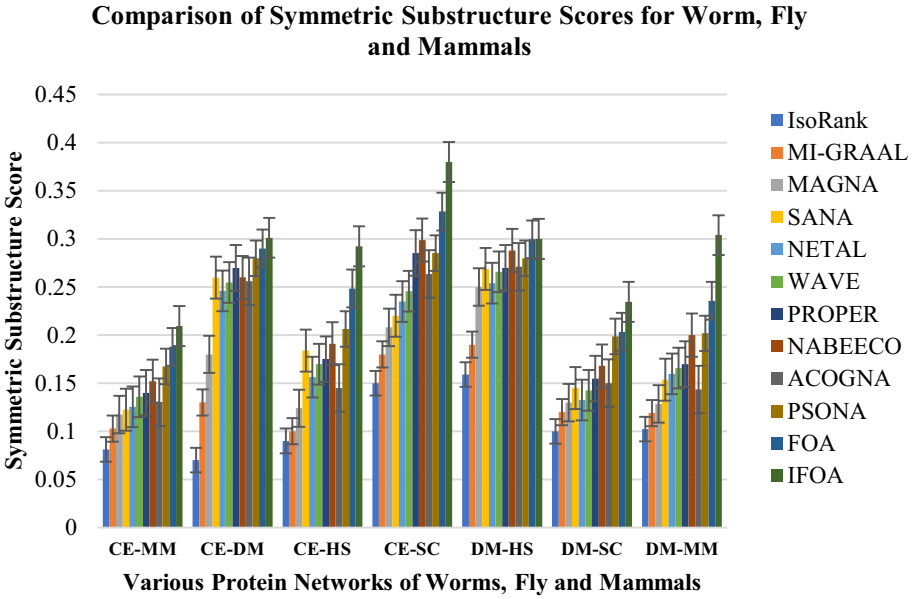

Fig. 6 Comparisons of average S^3 scores for worms and fly with proposed and existing algorithms

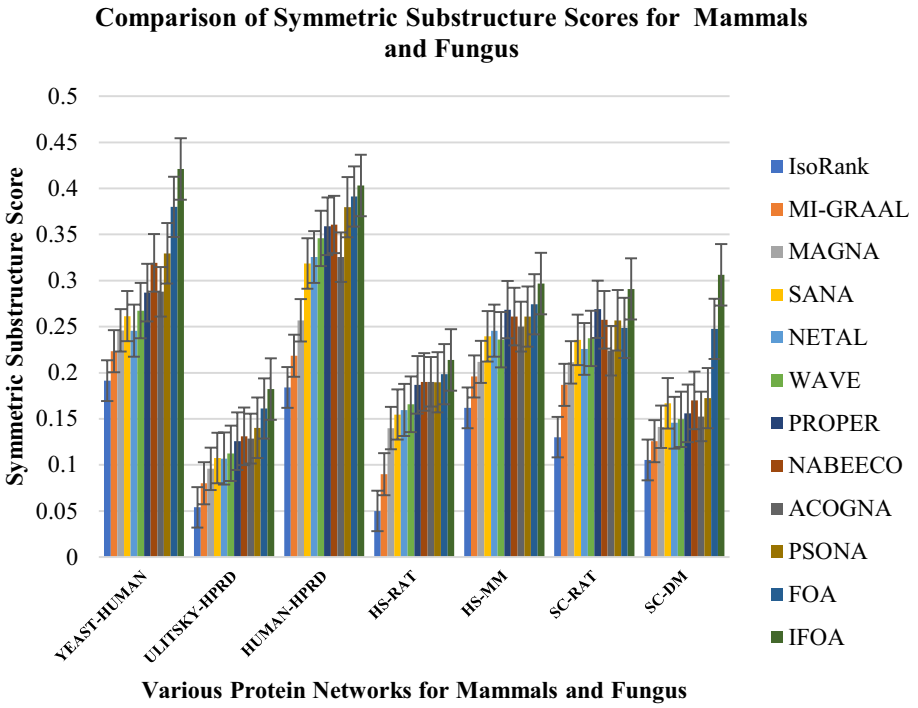

Fig. 7 Comparisons of average S^3 scores for mammals and fungus with proposed and existing algorithms

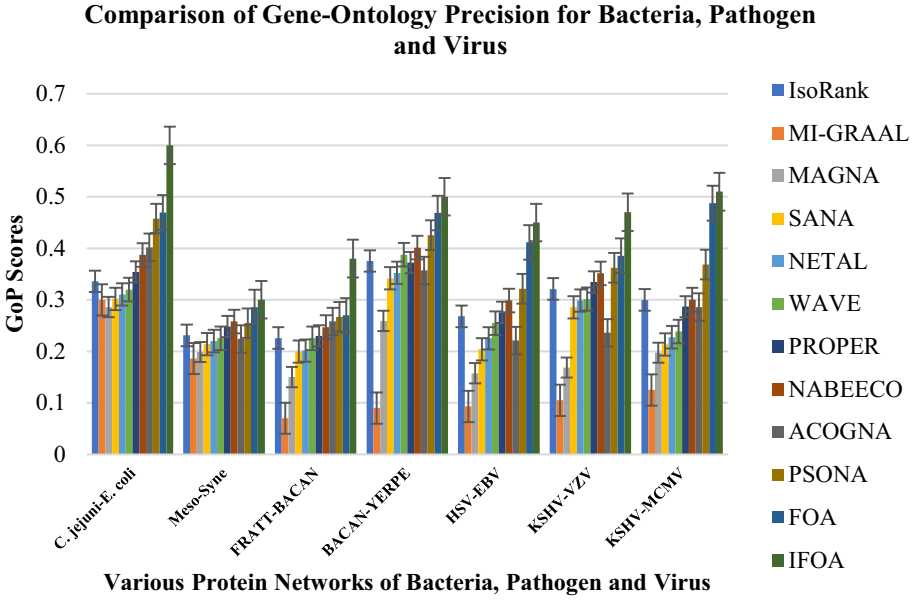

Fig. 8 Comparisons of average *GoP* scores for bacteria, pathogen and virus with proposed and existing algorithms

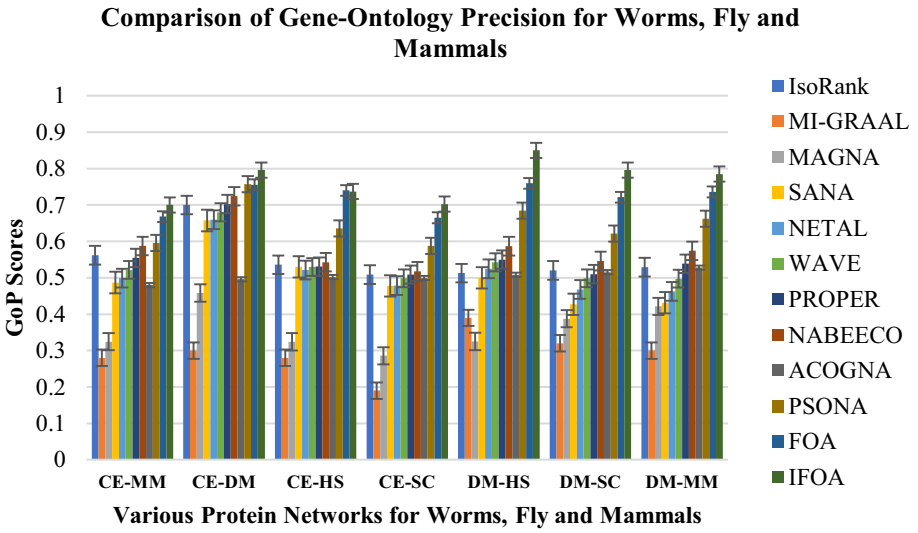

Fig. 9 Comparisons of average *GoP* scores for worms and fly with proposed and existing algorithms

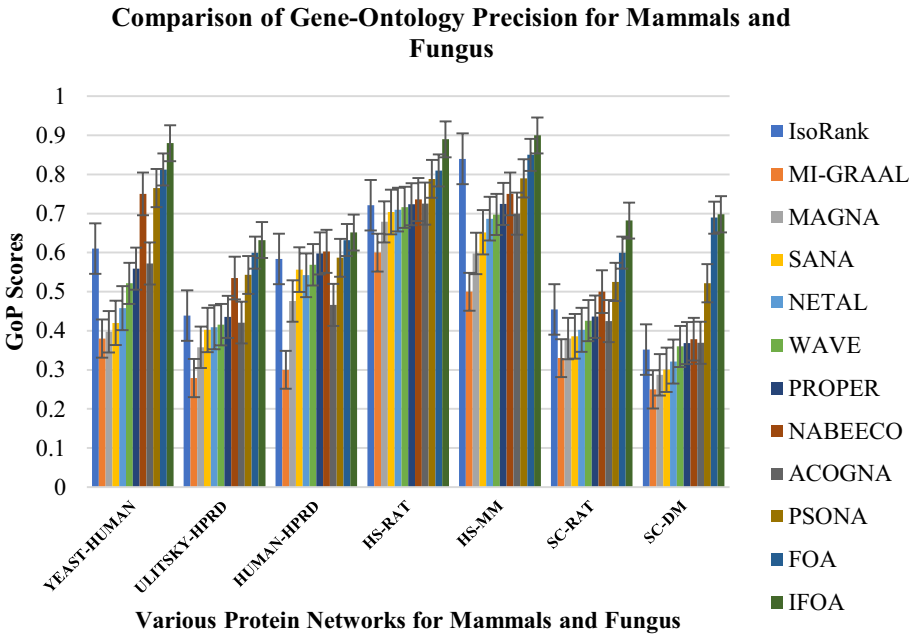

Fig. 10 Comparisons of average *GoP* scores for mammals and fungus with proposed and existing algorithms

method using S^3 and *GoP* scores was verified using the significant confidence level of 5% (P-value < 0.05). The p-values less than 0.05 were described as highly significant and the values superior than 0.05 were described as insignificant values.

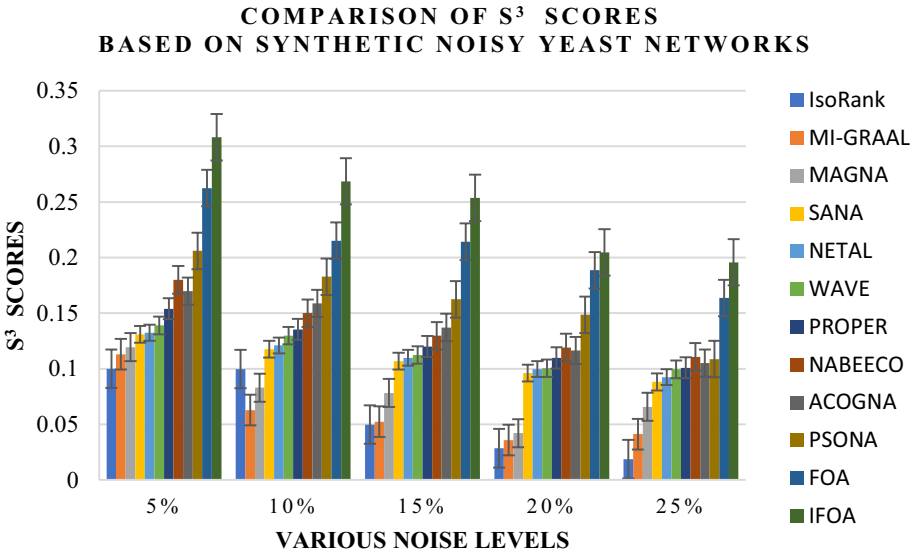

Fig. 11 Comparisons of average S^3 scores for synthetic noisy Yeast networks with proposed and existing algorithms

In Table 3, the upper right corner of the matrix is attained from S^3 values and the lower left corner of the matrix is attained from *GoP* values by using the Wilcoxon Matched-signed Rank test as a non-parametric statistical analysis. It is diagnosed that the proposed IFOA method achieves statistically better results over all other methods.

After aligning the pairwise protein networks using the IFOA method, the aligned proteins have more common gene ontology functions and KEGG pathways. One of the protein network pair namely, Yeast-Human (TAF6_YEAST- TAF10_HUMAN) have shared functions and they are highlighted and given below in Table 4. Remaining pairs of protein networks and their common functions are given in the supplementary material.

Table 3 Comparison of S^3 and *GoP* values of the proposed algorithm against other existing algorithms using Wilcoxon Signed Rank Test

	IsoRank	MI-GRAAL	MAGNA	PROPER	NABEECO	ACOGNA	PSONA	FOA	IFOA
IsoRank	0	0.086	0.071	0.062	0.072	0.052	0.050	0.042	0.040
MI-GRAAL	0.078	0	0.062	0.059	0.060	0.057	0.055	0.049	0.042
MAGNA	0.070	0.075	0	0.067	0.065	0.065	0.069	0.059	0.052
PROPER	0.052	0.050	0.048	0	0.040	0.042	0.045	0.048	0.039
NABEECO	0.050	0.048	0.042	0.048	0	0.040	0.040	0.042	0.046
ACOGNA	0.061	0.050	0.048	0.046	0.039	0	0.030	0.038	0.035
PSONA	0.060	0.054	0.047	0.042	0.038	0.035	0	0.032	0.030
FOA	0.045	0.049	0.050	0.042	0.040	0.035	0.031	0	0.029
IFOA	0.039	0.036	0.039	0.034	0.035	0.032	0.027	0.025	0

Table 4 The common gene ontology functions and KEGG pathways for protein TAF6 and TAF10 in Yeast-Human organism protein networks

	Protein network 1	Protein network 2	KEGG Pathways for Network 1	KEGG Pathways for Network 2
Yeast-Human	TAF6_YEAST	TAF10_HUMAN		
Molecular Function	chromatin binding identical protein binding protein complex scaffold activity RNA polymerase II activating transcription factor binding transcription factor complex TFIID	DNA binding enzyme binding estrogen receptor binding RNA polymerase binding RNA polymerase II activating transcription factor binding transcription factor (TFIID) complex	Basal transcription factors	Basal transcription factors Herpes simplex infection
Biological Process	chromatin organization histone acetylation regulation of transcription, DNA-templated RNA polymerase II transcriptional preinitiation complex assembly transcription from RNA polymerase II promoter	apoptotic process DNA-templated transcription, initiation hepatocyte differentiation histone acetylation histone H3 acetylation protein deubiquitination regulation of DNA binding regulation of transcription, DNA-templated transcription from RNA polymerase II promoter transcription initiation from RNA polymerase II promoter		
Cellular Component	Nucleus Cytosol	Nucleus Cytosol		

3.1 Implementation and Discussion

Computing the global pairwise biological protein network alignment with high topological and biological function accuracy is still a difficult task. This article presented an IFOA method to solve the biological network alignment problem. It has established a significant improvement in the alignment accuracy over all other leading network aligners. The network performance measures, namely S^3 and *GoP* were considered for an optimal solution.

The computational execution time of a network alignment using the proposed IFOA method is $O(n^2p)$ where n is the number of population and p is the number of iterations. The construction of realignment will take time. The growing size of the biological network may lead to increased time complexity. The proposed IFOA method could be effortlessly executed as a parallel to work with huge PPI networks and to reduce the running time.

Subsequently, not all aligners were available under the same platform, the execution of many of the aligners was done on virtual machines, which prohibited us from

accomplishing an exact comparison of their relative execution time. Thus, here the comparison of execution time between different population size and generations has been given for the proposed method. The proposed technique consumes several hours to obtain better alignment results. Since our technique is a meta-heuristic in nature, it can conceivably keep trying to improve its alignment indefinitely. The execution time for a proposed IFOA method with the various sizes of the population and a different number of generations are depicted in Fig. 12.

From Fig. 12, it has been inferred that when the size of the population and the generations of the population increased, simultaneously the execution time of the algorithm has also been increased, they are directly proportional to each other.

In favour of obtaining the biological significance of resulted protein network alignments the gene ontology function prediction and KEGG pathways were identified. From this, it is inferred that, the network alignment using the IFOA method, has more shared protein functions in nature.

The highest protein sequence similarity habitually correlates the highest similarity of protein function. Although the protein pairs which are considered in this work had less sequence similarity in nature, they exhibit the more common biological function by aligning their protein network using IFOA approach. The sequence similarity between two proteins in a network is evaluated using EMBOSS Needle pairwise sequence alignment technique [30]. The minimum sequence similarity was 2.3% among the protein pairs and maximum sequence similarity was 64.4% between the protein pairs and the average of all protein pairs in the dataset was 16.62% of the sequence similarity.

For example, though the protein pair of *Yeast-Human* namely TAF6- TAF10 has a sequence similarity of 6.0%, it has two molecular functions, three biological processes and two cellular components in common and the names of these common functions are given in Table 4. The KEGG pathway for the above-aligned protein pair has one common pathway namely "basal transcription factors". More than aligning the Human protein with other organisms like Mouse, Rat and others, it is better to align with Yeast which has balanced data. The protein pair STH1- REG3A has a sequence similarity of

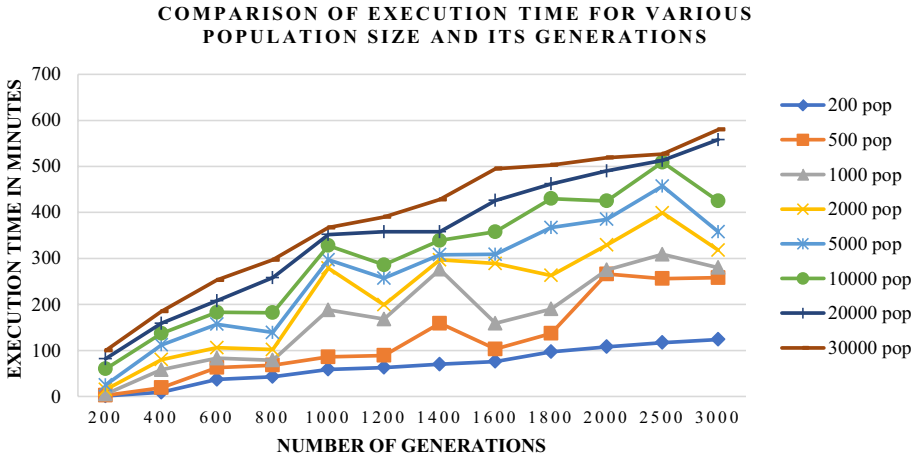

Fig. 12 Comparison of execution time for various population size and its generations

4.2% and it shares two molecular functions and cellular components. The detailed discussion of remaining protein pairs is described in supplementary material.

In order to identify some useful conclusion with a certain level of statistical significance, this implementation was run on a 2GHZ Intel CPU with 1 GB of memory, running on windows 8.1. When increasing the number of population and iterations the values of performance measures increases simultaneously.

Also, the performance of the alignment algorithm depends on the nature of the protein network to be aligned which has the diverse interactions between proteins. For example, the pair of Yeast and Human organisms has more similarity between them and the accuracy also is higher than other pairs of biological networks. All the performance measures had fluctuated during the first 100 iterations of the experiment and in the later run, consistency was observed. The algorithm gets dismissed only if the total number of iterations reached or the best solution produced in each generation remains identical for 100 consecutive iterations.

4 Conclusion

The network alignment problem is an open question to researchers. The alignment techniques employed to solve this problem should be boosted habitually as they play a vital role in the analysis of massive data delivered by next-generation sequencing and high-throughput experiments. The goal of this experiment is to attain the purpose of assessing evolutionary algorithms and discovering ways to further progress their performance to accomplish the optimal solution. This research work proposed the IFOA method, to perform global pairwise protein biological network alignment and leads the result in the direction of the optimal solution. The performance measures, namely symmetric substructure score (S^3) and gene-ontology precision (GoP) was used as a network alignment quality.

The statistical significance was calculated to compare the significance of the proposed approach with existing methods. As the stochastic optimization techniques were employed to align the network alignment, few execution concerns had raised. The major concerns in execution were the computational limitations and execution time as it is based on successive iterations, parameter selection, the number of populations, etc. In the future, this method can be achieved on a multicore/ more powerful CPU. It can also be extended or combined with any other evolutionary algorithms for optimal results. Different network quality measures may be introduced to attain the best results of network alignment and to find the phylogeny based on network similarity between species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11277-021-08851-z.

Declarations

Conflict of interest There are no conflicts of interest.

Competing interests The authors declare no competing financial interests.

References

1. Huang, J., Gong, M., & Ma, L. (2016). A global network alignment method using discrete particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 99*, 1. https://doi.org/10.1109/TCBB.2016.2618380
2. Ciriello, G., Mina, M., Guzzi, P. H., Cannataro, M., & Guerra, C. (2012). AlignNemo: A local network alignment method to integrate homology and topology. *PLoS ONE, 7*(6), e38107. https://doi.org/10.1371/journal.pone.0038107
3. Mina, M., & Guzzi, P. H. (2014). Improving the robustness of local network alignment: Design and extensive assessment of a Markov Clustering-based approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics., 11*, 561–572. https://doi.org/10.1109/TCBB.2014.2318707
4. Ngoc, H. T., & Xuan, H. H. (2016). ACOGNA: An efficient method for protein-protein interaction network alignment. In: *Proceedings of IEEE eighth international conference on knowledge and systems engineering*. https://doi.org/10.1109/KSE.2016.7758021
5. Elmsallati, A., Clark, C., & Kalita, J. (2015). Global alignment of protein-protein interaction networks: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 13*, 689–705. https://doi.org/10.1109/TCBB.2015.2474391
6. Yerneni, S., Khan, I., Wei, Q., & Kihara, D. (2018). IAS: Interaction specific GO term associations for predicting protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2015.2476809
7. Wei, Q., Khan, I. K., Ding, Z., Yerneni, S., & Kihara, D. (2017). NaviGO: An interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinformatics, 18*, 177. https://doi.org/10.1186/s12859-017-1600-5
8. Clark, C., & Kalita, J. (2015). A multiobjective memetic algorithm for PPI network alignment. *Bioinformatics, 31*(12), 1988–1998. https://doi.org/10.1093/bioinformatics/btv063
9. Singh, R., Xu, J., & Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceeding of the National Academy of Sciences of the United States of America., 105*, 12763–12768. https://doi.org/10.1073/pnas.0806627105
10. Memisevica, V., & Przulj, N. (2012). C-GRAAL: Common-neighbors-based global GRAph ALignment of biological networks. *Integrated Biology., 7*, 734–743. https://doi.org/10.1039/c2ib00140c
11. Malod-Dognin, N., & Przulj, N. (2015). L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics, 31*, 2182–2189. https://doi.org/10.1093/bioinformatics/btv130
12. Kuchaiev, O., & Przulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics, 27*, 1390–1396. https://doi.org/10.1093/bioinformatics/btr127
13. Patro, R., & Kingsford, C. (2012). Global network alignment using multiscale spectral signatures. *Bioinformatics, 28*, 3105–3114. https://doi.org/10.1093/bioinformatics/bts592
14. Hashemifar, S., Ma, J., Naveed, H., Canzar, S., & Xu, J. (2016). ModuleAlign: Module-based global alignment of protein-protein interaction networks. *Bioinformatics, 32*, i658–i664. https://doi.org/10.1093/bioinformatics/btw447
15. Kazemi, E., Hassani, H., Grossglauser, M., & Modarres, H. P. (2016). PROPER: Global protein interaction network alignment through percolation matching. *BMC Bioinformatics, 17*, 527. https://doi.org/10.1186/s12859-016-1395-9
16. Dognin, N. M., Ban, K., & Pruzlj, N. (2017). Unified alignment of protein-protein interaction networks. *Scientific Reports., 7*, 953. https://doi.org/10.1038/s41598-017-01085-9
17. Saraph, V., & Milenkovic, T. (2014). MAGNA: Maximizing accuracy in global network alignment. *Bioinformatics, 30*, 2931–2940. https://doi.org/10.1093/bioinformatics/btu409
18. Vijayan, V., Saraph, V., & Milenkovic, T. (2015). MAGNA++: Maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics, 31*, 2409–2411. https://doi.org/10.1093/bioinformatics/btv161
19. Ibragimov, R., Martens, J., Guo, J., & Baumbach, J. (2013). NABEECO: Biological network alignment with bee colony optimization algorithm. In: *Proceeding of 15th annual conference companion on genetic and evolutionary computation* (pp. 43–44). https://doi.org/10.1145/2464576.2464600

20. Tuncay, E. G., & Can, T. (2016). SUMONA: A supervised method for optimizing network alignment. *Computational Biology and Chemistry., 63*, 41–61. https://doi.org/10.1016/j.compbiolchem.2016.03.003
21. Chindelevitch, L., Ma, C. Y., Liao, C. S., & Berger, B. (2013). Optimizing a global alignment of protein interaction networks. *Bioinformatics, 29*, 2765–2773. https://doi.org/10.1093/bioinformatics/btt486
22. Mamano, N., & Hayes, W. B. (2017). SANA: Simulated Annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics, 33*, 2156–2164. https://doi.org/10.1093/bioinformatics/btx090
23. Sun, Y., Crawford, J., Tang, J., & Milenkovic, T. (2014). Simultaneous optimization of both node and edge conservation in network alignment via WAVE. In M. Pop, sH. Touzet (Eds.), *Algorithms in bioinformatics*. WABI 2015. *LNCS* (p. 9289). https://doi.org/10.1007/978-3-662-48221-6_2.
24. Yang, X. (2009). Firefly algorithms for multimodal optimization. stochastic algorithms: Foundations and applications SAGA 2009. *LNCS* (p. 5792). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-04944-6_14
25. Kaur, K., Salgotra, R., & Singh, U. (2017). An improved firefly algorithm for numerical optimization. *Proceedings of International Conference on Innovations in Information, Embedded and Communication Systems*. https://doi.org/10.1109/ICIIECS.2017.8275914
26. Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., & Przulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *Journal of Royal Society Interface*. https://doi.org/10.1098/rsif.2010.0063
27. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Cepas, J. H., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. V. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research., 47*, D607–D613. https://doi.org/10.1093/nar/gky1131
28. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The intact molecular interaction database in 2012. *Nucleic Acids Research, 40*, D841–D846. https://doi.org/10.1093/nar/gkr1088
29. Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., et al. (2013). The biogrid interaction database: 2013 update. *Nucleic Acids Research., 41*, D816–D823. https://doi.org/10.1093/nar/gks1158
30. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology., 48*, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

R. Ranjani Rani is an Assistant Professor in the Department of Computer Science, PSG College of Arts and Science in Coimbatore. India. She has published various articles in reputed SCI/SCIE indexed journals and attended various National and International conferences in and outside India. Her area of research is Data Mining and Bioinformatics.

D. Ramyachitra is an Assistant Professor in the Department of Computer Science at Bharathiar University in Coimbatore, India. She has 16 years of teaching experience and 9 years of Research experience. She has published numerous articles in various reputed SCI/ SCIE indexed journals and attended various National and International conferences in and outside India. She has delivered lectures on various seminars and conferences. Her area of research is Data Mining and Bioinformatics.