Dynamic Complex Protein Detection using Binary Harris Hawks Optimization

Mouna Chellal\(^1\)\(^,*\) and Ilyas Benmessahel\(^2\)

\(^1\) School of Computer Science and Engineering, Central South University, Changsha, China.
\(^2\) College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

\(*\) mounachellal@gmail.com

Abstract. Identifying protein complexes brings new insights in the field of understanding of cellular life and the mechanisms underlying complex diseases. In order to solve the challenge of protein complex prediction, a large number of tools have been developed to create an efficient approach. Most existing methods do not take changing protein-protein interactions with time into consideration. To address these challenges, we propose a new binary technique of Harris Hawks Optimizer (BHHO) to improve the accuracy of protein complex detection. According to the core-attachment structure, BHHO is used to discover the potential cores of protein complexes by simulating the process of cooperative behavior in chasing by Harris’ hawks in nature. We design a new fitness function to detect protein complexes with various densities and modularities. The experimental results show that BHHO achieves an ideal performance PIN in term of the F1 and outperforms of the classical algorithms in yeast in term Jaccard measure.

1. Introduction

Many high-throughput experimental techniques, such as yeast two-hybrid and mass spectrometry, have conducted in the past to detect PPIs for organisms [1]. At present, protein-protein interaction (PPI) is one of the key topics for the development and progress of the modern system’s biology [2]. Protein complexes are molecular aggregations assembled by Multiple PPIs, which is substantial to perform a vast array of biological processes. Predicting protein complexes from large PPI network is of great significance to recognize the structure and function of cellular organization [3].

In order to predict protein complexes, several works have been focused on static PPI networks. The cell cycle is changing over time and has different stages which means that the PPI dynamically constructed and disassembled [4]. In static PPI networks, many important temporal information is lost by which can be understand the protein interactions in the cell. Therefore, the accurate determination of protein complex is crucial for shifting from static PPI networks to dynamic PPI networks (DPI) [5].

Generally, two main directions used to build dynamic PPI networks which are the gene expression data and the high-throughput PPI data. Using gene expression is one major method that is based on calculating the variance of each protein [6]. In general, a protein in an active time point refers that the expression level of the corresponding gene attains the peak point. Based on this assumption, Wang et al. construct dynamic PPI networks using gene expression data and static PPI networks to predict the
protein complexes and the essential proteins [7]. Using co-expression correlations is an alternative direction to compose dynamic PPI networks.

Mathematically, PPI is a group of proteins that interact with each other and can be formed as the undirected graph $G(V, E)$, where $V$ is a list of nodes (proteins) and $E$ is a list of edges (interactions) connecting pairs of nodes [8]. The protein complex in this case is defined as dense proteins that are connected, where the density is defined as the fraction of edges out of all possible vertex pairs [9].

With the continuous development of nature-inspired algorithms, many of them have been successfully applied in the research of protein complex prediction [10]. Harris hawks optimizer (HHO) is a novel swarm intelligent algorithm that mimics the cooperative behavior of the chasing process by Harris hawks for global optimization [11]. HHO is easy to be understood and implemented, which has few parameters to be adjusted. Due to its simplicity and efficiency, HHO showed great success in solving some real-world complex problems. Here, HHO will be used to find the complex proteins.

In this work, we present a new algorithm, called BHHO protein complex detection (BHHO-PCD) in which HHO is merged with topological properties and biological information for complex protein discovery. To the best of my knowledge, most of the methods of complex protein discovery focus on static PPI networks and ignore the intrinsic features of organisms which is dynamism. First, we design the synthetic dynamic PPI networks based on gene expression. The cores are detected using a layer-by-layer scheme and the BHHO-PCD algorithm is used to make Hawks converge towards the prey to form the protein complexes. Next, we utilize filtered out mechanism to refine the predicted protein complexes using an overlap threshold. The experimental results illustrate that BHHO-PCD algorithm outperforms than those competing methods for the identification of protein complexes.

The rest of our paper is arranged as follows. Section 2 clarifies our proposed BHHO-PCD algorithm in great detail. Section 3 shows the results and analysis of the experiments. Finally, Section 4 sums up this study and makes the conclusion.

2. Framework
The framework of our proposed Protein Complex Detection is divided into three modules as depicted in Figure 1, namely, Constructing Dynamic Module, Discovering Seed Core Module and Harris Hawks Optimizer Module.

2.1. Constructing Dynamic Module
In order to build a dynamic model, we use both gene expression variance information and co-expression correlations information to calculate the active time point and active probability of each PPI in dynamic PPI networks.

The different peak points of the gene expression value reflect the dynamic changes of protein activities or conditions.

A global-threshold was used to define whether a gene has a high level which is calculated based on the values of its expression. Generally, it is very complicated to determine the active time point of proteins using a single global threshold, thus many expression value were filtered out [4]. Based on the three-sigma threshold, we construct the Dynamic PIN by calculating the active probability of each protein at various time points [12]. A protein $p$ is considered to be active in a dynamic PPI sub-network only if its gene expression value is greater than or equal to the active threshold $\text{ActiveTh}(p)$:

$$\text{ActiveTh}(p) = \alpha(p) + k \cdot \sigma(p) \cdot \left(1 - \frac{1}{1 + \sigma^2(p)}\right)$$

where $\alpha(p)$ and $\sigma(p)$ are the arithmetic mean and the standard deviation (SD) of the gene expression data $p$, respectively, and $k$ (the times of sigma).

We construct dynamic networks by integrating co-expression correlation and gene expression level shifts in the present work. Calculation of the correlation coefficient requires multiple sequential expression data that cover a period of time. We define a period of time on the original expression data, which involves three successive time stamps. When $t$ is the actual time point, we set period time that covers three-time stamps containing $t-1$, $t$, and $t+1$. 


Figure 1. Framework of the proposed method

Since the dynamic network has 12 times and uses a period time to calculate the average of 3 times stamps together, the final co-expression network at each time point is the average of the three times [13]. Eventually, the entire PPI network was divided into 6 sub-networks, the dynamic PPI network was constructed. Let adjSPI denote the static PPI networks adjacency matrix. Integrating Act, Coe, and adjSPI, we can calculate the dynamic PPI networks adjacency matrix adjDPI at the time point i as follows:

$adjDPI_i = ActiveTh_i * Coe_i * adjSPI$  \hspace{1cm} (2)

2.2. Extracting Cores Module

In the dynamic PPI networks can not only provide the dynamic properties of PPI, which are ignored by static PPI networks but also distinguish the reliability of each protein by the dynamic information of PPI networks. Let DPI denote dynamic PPI networks that include Tk active PPI subnetworks {DPIT1, DPIT2… DPITk}. {adj_DPI T1, adj_DPI T2… adj_DPI Tk} is the adjacency matrices of the DPI at T1, T2… Tk active time points. The cluster score of edge e(u,v) in DP_ITi is defined as follows:

$cluster.\ score.\ Ti = Temp.\ edge.\ weight_{(u,v)} \times \frac{2 \times (|N_u \cap N_v| + 1)}{|N_u| + |N_v|}$  \hspace{1cm} (3)

$Temp.\ edge.\ weight_{(u,v)} = adj.\ DPITi\_{(u,v)} \times ACT_u \times ACT_v$  \hspace{1cm} (4)

Given a subgraphs SG in an active PPI subnetworks DPITi, let VSG and ESG denote the set of proteins and PPIs in SG, respectively. To set subgraphs as seed core [7-14], the density of SG is defined as follows:

$density_{Ti} = \frac{2 \times \sum_{e(u,v) \in E_{SG}} Temp.\ edge.\ weight_{(u,v)}}{|V_{SG}| \times (|V_{SG}| - 1)}$  \hspace{1cm} (5)

2.3. Harris Hawks Optimizer Module

Harris Hawks Optimization (HHO) is a novel method for global optimization, which is inspired by the chasing behavior of Harris hawks. This algorithm shows the tactic strategies of cooperation between hawks in predation the prey through surprise pounce of Harris hawks in nature. In HHO, the exploration phase refers to the spread for discovering the entirely new areas from the search space (perching strategy) and wait to track prey. In perching strategy, Hawk perch based on the positions of...
other family members (to be close enough to them when attacking) [11].

2.3.1. The proposed binary HHO algorithm for complex protein detection. Finding proper complex protein is the objective of our proposed method which can be shaped as combinatorial optimization to use HHO. The most important step in finding complex protein is how to code HHO solution to be compatible with the topology network of PPI. This problem caused by the optimal combination of the seed cores between PPI and HHO solution. Also, another important step, a suitable and reasonable fitness function (cost function) is needed.

The HHO algorithm starts the optimization process by generating a set of solutions and assumes that each individual is a solution in all populations. The size of the solution represents the dimension of problem. In our proposed, each individual represents the number of core extracted from the PPI network. The solution represents the best selected cores in the complete set of $M$ cores computed by:

$$v_i = \begin{cases} 1 & \text{if } M(i) = 1 \\ 0 & \text{if } M(i) = 0 \end{cases}$$  \hspace{1cm} (6)

In our implementation, each individual in BHHO is a vector that includes the values restricted to the binary range $[0,1]$, which the size of each solution represents the complete set of cores found by the previous module. The V-shaped function is used to get a continuous value between $0–1$ as follows:

$$V_{i+1} = \begin{cases} -v(i) & V_{\text{shaped}}(v(i)) \geq \text{rand} \\ v(i) & V_{\text{shaped}}(v(i)) \leq \text{rand} \end{cases}$$  \hspace{1cm} (7)

2.3.2. Objective Function. Fitness function refers to the representation and measures of the represented solution used in the optimization algorithm. In each iteration of the algorithm, every individual is evaluated, fitness function evaluates the efficiency of a single solution in population individuals for a given problem. In this paper, we give the following objective function by combing definition of the density of the cluster and considering the appropriate number of clusters. We also used weight sum of edges to replace the number of edges in clusters. The proposed fitness function is designed as follows:

$$F(C^1 \ldots C^k) = \sum_{i=1}^{k} \frac{C^i_{\text{in}}}{C^i_{\text{out}}}$$  \hspace{1cm} (8)

Where $(C^1 \ldots C^k)$ is the clustering results of best cores, $|E|$ is the number of edges in the cluster $C^i$, $|V|$ is the number of edges in the cluster $C^i$, $W_{kl}$ is the number of edges form $C^i$ to outside cluster.

Then, we augment the seed edge to generate the core structure by adding the suitable neighbor proteins one by one. Finally, the attachment proteins are detected for each core structure based on the AttachScore that is calculated as follow:

$$\text{AttachScore}_{ti} = \frac{\sum_{u \in V_{SG}} C_{T_t}(u,v)}{V_{SG}}$$  \hspace{1cm} (9)

The attachment proteins are added into the core structure to form the candidate protein complex. However, the same or highly overlapping protein complexes are filtered out. For any candidate protein complex Candidate list, we check the overlapped degree between clusters.

3. Computational results

In order to investigate the efficacy of the proposed BHHO optimizer, a well-studied set of diverse PPI networks are selected from the literature. We evaluated the performance of the proposed method using Krogan PPI dataset [15].

The gene expression data of Saccharomyces cerevisiae, with accession number GSE343, was retrieved from which is provided by Gene Expression Omnibus (GEO) [13]. This dataset is used in our experiment and it has an expression profiling that includes 9,335 probes with 12 different time points over three successive metabolic cycles. Therefore, there are 12 active time points (T1, T2... T12) for each gene in a cycle. In our experiments, DPI Krogan is constructed based on gene expression data.
The benchmark protein complex dataset CYC2008 includes 408 manually curated heterogenic protein complexes, which is used to evaluate the protein complexes predicted by our method [15].

Our proposed algorithm was evaluated based on several factors, namely, Precision, Recall and F1. Overlapping Score (OS) is usually used to assess the match score between a predicted protein complex and a known protein complex [16].

For the purpose of performance evaluation, we use the sensitivity (Sn), positive predictive value (PPV) and accuracy (Acc) as an evaluation metrics to evaluate protein complex prediction tools, we applied Sn, PPV and Acc in our method on Krogan PPI dataset [1] [15].

In this experiment, we compare our method with the following established leading protein complex prediction methods: CFinder, ClusterOne, MCL, CORE, RNSC, and MCODE. These methods are used to compare the performance in most of the recent complex prediction studies. To equally compare the performance, we test all comparison methods on the Krogan, and choose the optimal parameters. The highest value in each row was shown in bold. As shown in Table 1, our method achieves the highest F1 of 0.560, recall of 0.538, which significantly outperforms other methods. MCODE achieves the highest precision of 0.75. But the recall of MCODE is only 0.157, which leads to a low F1 of 0.259. RNSC achieves the highest Acc of 0.622 while our method archives 0.547. We also note that our method predicted 649 protein complexes.

Table 1. Performance comparison with other methods in terms mentioned above.

|        | BHBO  | CFinder | MCL   | CORE  | MCODE  | ClusterOne | RNSC |
|--------|-------|---------|-------|-------|--------|------------|------|
| Precision | 0.519 | 0.608   | 0.283 | 0.318 | 0.750  | 0.351      | 0.239 |
| Recall  | 0.560 | 0.201   | 0.550 | 0.444 | 0.157  | 0.508      | 0.551 |
| F1      | 0.538 | 0.302   | 0.374 | 0.371 | 0.259  | 0.415      | 0.363 |
| Sn      | 0.446 | 0.481   | 0.498 | 0.255 | 0.270  | 0.555      | 0.498 |
| PPV     | 0.672 | 0.358   | 0.723 | 0.848 | 0.551  | 0.659      | 0.776 |
| Acc     | 0.547 | 0.415   | 0.600 | 0.465 | 0.386  | 0.605      | 0.622 |
| RANK    | 3.1   | 5.1     | 3.3   | 4.5   | 5.6    | 3          | 3.2   |

Our proposed offers a very promising performance because it achieves the second rank of 3.1 followed by RNSC (3.2) while ClusterOne obtained the first rank (3). Thus, the common parametric techniques (P, R, F1, Sn, PPV, and ACC) only compare the ability to recover an overlap and measures the likelihood of getting an overlap in the overall performance of algorithms. By contrast, Jaccard statistical tests can prove that the results are statistically significant and can perform valid similarity of a predicted to a reference by looking at the size of the intersection relative to the union of the proteins [8]. To avoid bias that may arise from large variations in the size of predicted complexes, we also introduce another Jaccard analysis based on sample intersections. For this analysis, we defined three terms: JaccardS, JaccardI, and Jaccard.

Table 2. Performance comparison with other methods in terms of Jaccard, I, S.

|        | BHBO  | CFinder | MCL   | CORE  | MCODE  | ClusterOne | RNSC |
|--------|-------|---------|-------|-------|--------|------------|------|
| JaccardI | 0.3873 | 0.255   | 0.256 | 0.228 | 0.386  | 0.301      | 0.257 |
| JaccardS | 0.3842 | 0.185   | 0.365 | 0.233 | 0.179  | 0.379      | 0.434 |
| Jaccard | 0.3857 | 0.215   | 0.301 | 0.231 | 0.250  | 0.335      | 0.323 |
| RANK    | 1.3   | 6.3     | 4.3   | 6     | 4.6    | 2.6        | 2.6   |

Tables 2 listed the performance comparison results using CYC2008 as a benchmark in terms of JaccardS, JaccardI and Jaccard. As shown in Table 2, our proposed has the best result and superior performance on Jaccard and Jaccard. However, RNSC presents high result on JaccardS. These last results represents that our detected proposed shows a better matching ratio between detected protein complexes and real protein complexes. In summary, our proposed not only has the best performance
over other comparative algorithms in terms of F1 and Jaccard statistics but also can exactly predict many complexes protein such as RNA polymerase I.

4. Conclusion
Our study relies on the motivation to discover the complex protein by providing new insights through the application of complex protein detection programs combined with stochastic optimization algorithms. We implement the HHO to discover the complex protein from extracted seed cores in DPI. The results show that our proposed method outperforms the current popular and powerful algorithms in the literature. BHHO-PCD benefits from high convergence and can even find solutions that are close to the global optimal of any search space as well. Finally, the results of the dynamic environment confirm that HHO can solve challenging problems with many constraints and unknown search spaces. One of our future insights is to propose the first multi-objective optimization algorithm for complex protein mechanisms including the multi-criteria objective function. No computational details can be found in the literature about Multi-objective optimization approaches for Complex protein detection. Hence, it needs to be explored more in future research.

References
[1] Shen, Xianjun, et al. "Identifying protein complex by integrating characteristic of core-attachment into dynamic PPI network." PloS one 12.10 (2017).
[2] Chowdhury, Archana, et al. "Prediction of Protein Complexes Using an Evolutionary Approach." 2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2018.
[3] Zhao, Jie, and Xiujuan Lei. "Detecting overlapping protein complexes in weighted PPI network based on overlay network chain in quotient space." BMC bioinformatics 20.25 (2019): 1-12.
[4] Zhang, Yijia, et al. "A method for predicting protein complex in dynamic PPI networks." BMC bioinformatics 17.7 (2016): 229.
[5] Zhang, Yijia, et al. "Construction of dynamic probabilistic protein interaction networks for protein complex identification." BMC bioinformatics 17.1 (2016): 186.
[6] Rani, R. Ranjani, D. Ramyachitra, and A. Brindhadevi. "Detection of dynamic protein complexes through Markov Clustering based on Elephant Herd Optimization Approach." Scientific reports 9.1 (2019): 1-18.
[7] Wang J, Peng X, Li M, Pan Y. Construction and application of dynamic protein interaction network based on time course gene expression data. Proteomics. 2013; 13(2):301–12.
[8] Zhang, Wei, et al. "Integrating network topology, gene expression data and GO annotation information for protein complex prediction." Journal of bioinformatics and computational biology 17.01 (2019): 1950001.
[9] Rongquan, et al. "A seed-extended algorithm for detecting protein complexes based on density and modularity with topological structure and GO annotations." BMC genomics 20.1 (2019): 637.
[10] Lei, Xiujuan, et al. "IOPTICS-GSO for identifying protein complexes from dynamic PPI networks." BMC medical genomics 10.5 (2017): 80.
[11] Heidari, Ali Asghar, et al. "Harris hawks optimization: Algorithm and applications." Future generation computer systems 97 (2019): 849-872.
[12] Lei, Xiujuan, Ming Fang, and Hamido Fujita. "Moth–flame optimization-based algorithm with synthetic dynamic PPI networks for discovering protein complexes." Knowledge-Based Systems 172 (2019): 76-85.
[13] LEI, Xiujuan, et al. Topology potential based seed-growth method to identify protein complexes on dynamic PPI data. Information Sciences, 2018, 425: 140-153.
[14] Liang, Jing, et al. "ACO Based Core-Attachment Method to Detect Protein Complexes in Dynamic PPI Networks." International Conference on Swarm Intelligence. Springer, Cham, 2018.
[15] Xu, Bo, et al. "Protein complexes identification based on go attributed network embedding." BMC bioinformatics 19.1 (2018): 535.
[16] Liu, Lizhen, et al. "A method for predicting protein complexes from dynamic weighted protein–protein interaction networks." Journal of Computational Biology 25.6 (2018): 586-605.