ISSN: 0067-2904

AHeuristic Strategy for Improving the Performance of Evolutionary Based Complex Detection in Protein-Protein Interaction Networks

Qusay Z. Abdullah*, Bara'a Ali Attea
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq.

Abstract

One of the most interested problems that recently attracts many research investigations in Protein-protein interactions (PPI) networks is complex detection problem. Detecting natural divisions in such complex networks is proved to be extremely NP-hard problem wherein, recently, the field of Evolutionary Algorithms (EAs) reveals positive results. The contribution of this work is to introduce a heuristic operator, called *protein-complex attraction and repulsion*, which is especially tailored for the complex detection problem and to enable the EA to improve its detection ability. The proposed heuristic operator is designed to fine-grain the structure of a complex by dividing it into two more complexes, each being distinguished with a core protein. Then, it is possible for each of the remaining proteins associated with the original coarse-grained complex to repulse from one of the new generated complexes while attracted by the core protein of the second complex. The topology-based complex detection models presented in the literature are adopted to inter-play with the proposed heuristic operator inside the EA general framework. To assess the performance of the EA when coupled with the proposed heuristic operator, the well known Saccaromycaes Cerevisiae yeast PPI network and one reference set of benchmark complexes created from MIPS are used in the experiments. The results prove the positive impact of the heuristic operator to harness the strength of almost all adopted EA models.

Keywords: Complex detection, evolutionary algorithm, heuristic operator, PPI networks.

أستراتيجية أرشادية لتحسين كشف المركبات في الشبكات البروتينية التفاعلية والمعتمد على الخوارزمية التطورية

قصي زهير عبدالله*، براء علي عطية
قسم الحاسوب ، كلية العلوم ، جامعة بغداد ، بغداد ، العراق

الخلاصة

واحدة من أهم المشاكل والتي جذبت حديثا العديد من الابحاث في مجال الشبكات البروتينية التفاعلية (PPI) هي مشكلة كشف المركبات. ثبت هذه المشكلة بأنها صعبة للغاية، وحديثا تم أثبات بأن مجال الخوارزميات التطورية (EAs) له نتائج ايجابية. في هذا البحث تم أستحداث عامل ارشادي، يدعى *تجاذب وتنافر البروتين الى المركب البروتيني* وقد صممت خصيصا لمشكلة اكتشاف المركبات البروتينية ولأجل

*Email: qusay.zuhair.abdullah@gmail.com

تمكين خوارزمية ال EA لتحسين قدرته الاكتشافية. صمم العامل الارشادي المقترح لغرض تصفية أو صقل هيكلية المركب البروتيني وذلك بمحاولة شطره الى اثنين من المركبات البروتينية، يميز كل واحد منها عن طريق بروتين جوهري. وعلى هذا الأساس يتم أعادة توزيع بقية بروتينات المركب الأصلي غير المصقول، كل بروتين حسب تجاذبه مع أحدى من البروتينات الجوهرية المستخلصة وتنافره من الآخر. تم في هذا البحث أيضا أعتماد النماذج الرياضية للخوارزمية التطورية والخاصة باكتشاف المركبات البروتينية والموجودة في الادبيات وتوضيف التعاون المتبادل بينها وبين العامل الارشادي المقترح بداخل الاطار العام التابع الى (EA). وعلى هذا الأساس تم تقييم اداء الخوارزمية التطورية عندما أرتباطها بالعامل الارشادي المقترح، مع استخدام شبكة البروتين التفاعلية (Saccaromycaes Cerevisiae yeast) ومصدر واحد للمركبات تم اناشاؤه من قبل (MIPS) في التجارب. النتائج اثبتت التاثير الايجابي للعامل الارشادي لاضهار قوة أغلب النماذج الرياضية للخوارزمية التطورية.

1. Introduction

Protein-protein interaction (PPI) networks have received much attention in the past few years. For example, a large volume of experimented data is determined to reflect the proteins different structures and their mutual interactions in protein-protein interaction (PPI) networks [1]. Figure-1 depicts an example of a PPI network being represented as a graph where proteins act as nodes and interactions as links. The prediction of protein complexes (or functional modules) is crucial and an important problem in biological network analysis (BNA), giving a valuable guide in understanding the behavior of the cell. This has triggered a race for new high performance clustering algorithms for discovering and characterizing different complexes of PPI networks.

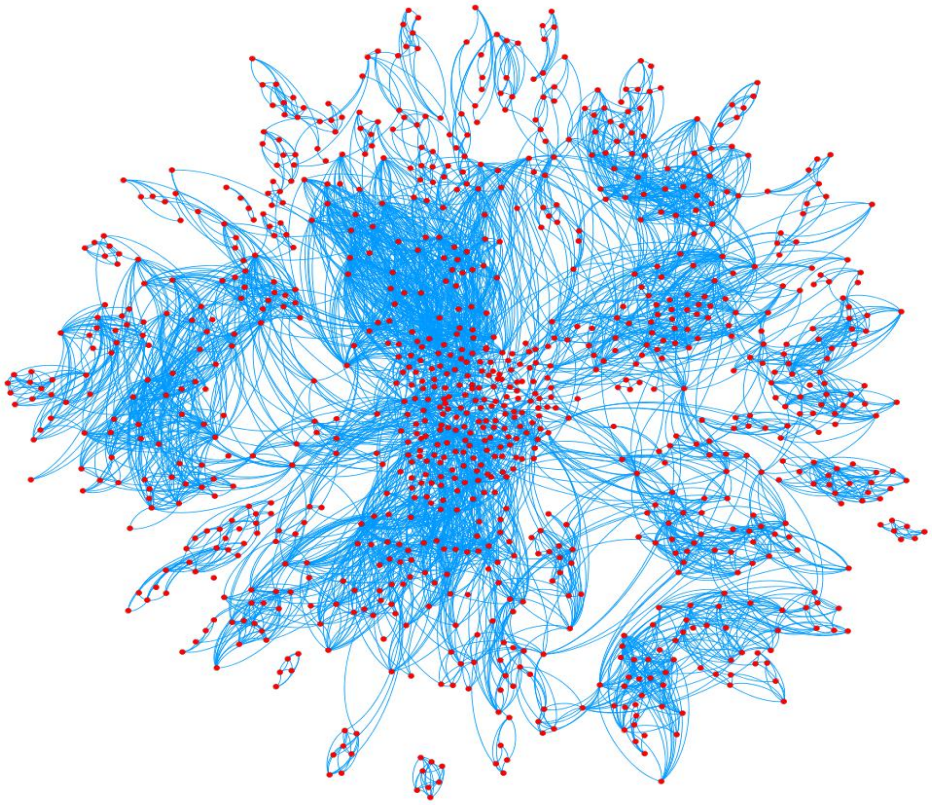

Figure 1- An example of a PPI network

Many of the research studies proposed bottom-up strategies based on some cost optimization function to find dense subgraphs from the whole PPI network. For example, Bader and Hogue [2] proposed Molecular Complex Detection (MCODE) algorithm to detect densely connected regions as molecular complexes in large PPI networks. MCODE consists of two main steps: network weighting and complex detection. In network weighting, all vertices are assigned weights based on their local network density This is followed by outward traversal from a locally dense seed protein to isolate the dense regions.

King et al. [3] proposed restricted neighborhood search clustering algorithm (RNSC) that partition the nodes of network throw clusters, based on low-cost clustering function (called homogeneity P value). It starts with an initial random clusters and then randomly moving a protein from one cluster to another satisfying a minimum deleterious of the cost function.

Altaf-Ul-Amin et al. [4] suggested a deterministic algorithm to select initial cluster as the seed highest weighted node or highest degree node. The cluster then gradually grows by adding neighbor nodes to the cluster one by one depending on neighbor priority. A cluster then continues to expand until cluster density and/or cluster property violating the initial constraint, at which a new cluster starts to born from the remaining nodes of the original graph.

Adamcsek et al. [5] proposed CFinder, an independent platform for locating overlapping group of interconnecting nodes. This strategy merges up nodes into clusters. It uses the Clique Percolation Method (CPM) of Palla et al. [6] to form the $\mathrm{k}-\mathrm{clique}$ (for $4 \leq \mathrm{k} \leq 6$) percolation clusters of the network.

Pizzuti and Rombo [7 – 9] proposed three local search co-clustering strategies (RANCoC, PINCoC, and $\mathrm{MF}-\mathrm{PINCoC}$). The basic concept behind these co-clustering methods is to search for dense sub-matrices in the adjacency matrix. The quality of sub-matrices differs from one strategy to another depending on the contribution of the proteins to improve the quality function. Additionally, the same authors, i.e. Pizzuti and Rombo, in 2014 [10] stated the complex detection problem in PPI networks as a single-objective optimization problem and devised the methodology of evolutionary algorithms (EAs) to solve the formulated problem. They formulated different topology-based quality functions include community score (CS), conductance (CO), normalized cut (NC), Internal Density (ID), Expansion (EX), and Cut Ratio (CR) as fitness models. Their investigations showed that EA has more detection ability than the traditional complex detection algorithms.

Although Pizzuti and Rombo, in 2014 showed that EA has advantageous over other complex detection algorithms, but they presented EA with its more general form. The main contribution of this paper is to introduce a heuristic operator to be injected into the general framework of the EA to improve its detection ability The remaining of this paper is organized into the following sections. Next section presents essential background related to the topic. Section 3 presents the general characteristic components of the proposed EA for tackling complex detection problem in PPI networks and the proposed heuristic methodology. The followed section, then, presents the experimental results to evaluate the performance of EA. The evaluation is reported with respect to different evaluation metrics. Final section presents conclusion of the current work.

2. Background

Mathematically, a network is a graph of nodes and edges. A PPI network $\mathcal{N}$ can be modeled as undirected graph $G = (P, E)$. The set of n proteins in $\mathcal{N}$ is noted as the set of nodes or vertices $P(G) = \{p_1, p_2, \dots, p_n\}$ while the mutual interaction between any pair of proteins in $\mathcal{N}$ is noted as edges (p_i, p_j). Normally, an undirected graph G can be represented by a symmetric $n \times n$ matrix called adjacency or connection matrix A. Rows and columns of A are labeled with the proteins of P with either 1 or 0 in entry (i, j) if protein p_i has mutual interaction with protein p_j, i.e. if $(p_i, p_j) \in E$. In list notation, matrix A can be represented by a set of n adjacency lists $L = \{l_1, l_2, \dots, l_n\}$, one list l_i for each protein $p_i \in P$aggregating all 1 entries in row i. Thus, $|l_i| = \sum_{j=1}^{n}(i, j)$ and $|L| = \sum_{i=1}^{n}|l_i|$. Mathematically, n is said to be the cardinality of G, $|l_i|$ is the degree of vertex p_i, while $|L|$ denotes the volume of G.

Graph co-clustering problem is a fundamental problem in computer science that is proved to be NP-hard [11] Consider a data set matrix A consisting of n objects, each being characterized by n features, i.e. $A = [a_{i,j}],\ i, j = 1, \dots, n$. Any clustering algorithm tries to partition the space of A into a

partition set $\mathcal{C}$ of K clusters, i.e. $\mathcal{C} = \{C_k\}_{k=1}^{K}$, according to the correlation among n objects. On the other hand, co-clustering means simultaneous clustering of both objects and features of A into sub-matrices, each of which consists of locally correlated objects under a subset of their features.

Given a graph $G = (P, E)$, the main problem in graph co-clustering is to find the set of sub-graphs $G_i = (P_i, E_i) \subset G$ such that the number of inter-edges connecting vertices from two different sub-graphs, usually known as *cut size*, is minimum [12]. Let $G_1 = (P_1, E_1)$ and $G_2 = (P_2, E_2)$ be two sub-graphs of G, the cut set and cut size of G_1 and G_2 can be expressed in Eq. 1 and 2, respectively.

$$cut(G_1, G_2) = \{(p_i, p_j) \in E \big| p_i \in P_1 \wedge p_j \in P_2\} \quad (1)$$

$$|cut(G_1, G_2)| = \sum_{p_i \in P_1 \wedge p_j \in P_2} A(i, j) \quad (2)$$

The second issue that should be carefully addressed in graph co-clustering problem is to group individual nodes of the graph into disjoint sets of dense clusters. Each cluster should have *intra*-contributions among its nodes as more as possible than its *inter*-contributions with other clusters. In context of social networks, Radicchi et al. [13] semantically define a sub-graph $G_i = (P_i, E_i) \subset G$ as a community in a strong sense if for every node p belongs to G_i, the intra-edge connections are larger than inter-connections, i.e.

$$\forall p \in G_i \Rightarrow \sum_{w \in G_i}(p, w) > \sum_{w \notin G_i}(p, w) \quad (3)$$

However, if this intra-connections versus inter-connections relation only holds over the aggregation of all G_i's nodes (see Eq. 4), then G_i is said to be a community in a weak sense.

$$\sum_{p \in G_i} \sum_{w \in G_i}(p, w) > \sum_{p \in G_i} \sum_{w \notin G_i}(p, w) \quad (4)$$

3. Improving EA based complex detection models

Evolutionary algorithms (EAs) are heuristic search and optimization techniques that simulate the process of natural evolution. The main idea of EAs is to evolve a population of candidate solutions towards better and better solutions. A typical EA has three main operators (selection, crossover and mutation) which are used collaboratively to improve the initial solutions set. Selection strategy selects sub-set of best solutions depending on fitness value. Crossover strategy creates new solutions from the existing solutions available in the mating pool after applying selection operator. This strategy exchanges the gene information between the solutions in the mating pool. Mutation is the occasional introduction of new features into the solution to maintain diversity in the population.

3.1 EA for complex detection problem

In this section, the characteristic components of the EA are presented and expressed in such a way to handle complex detection problem in PPI networks. The first component to express is how to define an individual solution in EA, i.e. chromosome. Here, the chromosome, I of the population $\mathbb{P}$ is defined as a collection of protein-protein interaction genes. A single gene in I is defined by its locus and its allele. Thus, in n loci chromosome, locus i identifies protein i in the PPI network, while its allele value j corresponds to protein j that has an actual interaction with protein i in the PPI network. Formally speaking, $I: PPI \rightarrow (S_i)^n$, where S_i is the set of all interacting proteins with protein i in the network PPI. The decoding function $\delta(I)$: $\mathcal{C} = \{C\}_{k=1}^{K}$ of individual I will outline different complexes of the network.

Given that EA is population-based optimization algorithm, then a population P is a set of N solutions and can be represented as: $\mathbb{P} = \{I_1, I_2, \dots, I_N\}$.

The iterative structure of the adopted EA can be defined as $\Psi(\mathbb{P}_t) = \mathbb{P}_{t+1}$, where $\mathbb{P}_t$ and $\mathbb{P}_{t+1}$ are the population of chromosome solutions at generation t and $t + 1$, respectively. The population starts with an initial random population $\mathbb{P}_0$ and continues until a maximum number of iterations max_t has been reached.

Uniform crossover and mutation operators are used with probability p_c and p_m, respectively. Consider two chromosomes I_1 and I_2 to be the two participating parents in the crossover. Under p_c control, a child I' can be generated by uniformly mixing allele values of I_1 and I_2 together. This is formally defined by, $\forall i, 1 \leq i \leq n$:

$$I'_i = \begin{cases} I_{1,i} & if\ r \leq 0.5 \\ I_{2,i} & otherwise \end{cases} \quad (5)$$

where $r \sim [0,1]$ is a uniform random number. For the mutation operator, the allele of the mutated gene I_i can be altered to any value j providing that protein I_i and j has an interaction in the PPI (i.e. $A(I_i, j) = 1$).

3.2 EA based complex detection models

Some of the recent and successful efforts for tackling complex detection problem in PPI networks are based on evolutionary algorithms (EAs). In [10], Pizzuti and Rombo addressed the problem as single-objective optimization functions. They projected different quality, i.e. fitness, functions used to solve community detection in complex networks as fitness functions. These include $modularity$, $community\ score$ (CS), $conductance$ (CO), $normalized\ cut$ (NC), $Internal\ Density$ (ID), $Expansion$ (EX), and $Cut\ Ratio$ (CR). Before formulating these models, let us express some mathematical notations [10]. Consider a network $\mathcal{N}$ of n individuals being modeled by $G = (V, E)$. Let $\mathcal{C} = \{C_1, \dots, C_K\}$ be a candidate partitioning of $\mathcal{N}$ with K complexes. Le fort $1 \leq k \leq K$, n_k and $m_k = \sum_{v,w \in C_k}(v, w)$ be the degree and volume of C_k, respectively. Moreover, let $in_k(v) = \sum_{w \in C_k}(v, w)$ and $out_k(v) = \sum_{w \notin C_k}(v, w)$ be, respectively, the number of intra-connections and inter-connections of node v which belongs to cluster C_k (i.e. $|l_v| = in_k(v) + out_k(v)$).

$Modularity$, Q, awards a clustering solution $\mathcal{C} = \{C_1, \dots, C_K\}$ according to the fraction of intra-connections inside $\{C_1, \dots, C_K\}$ (see Eq. 6). In Eq. 6, two contradictory objectives are handled. The first term in Eq. 6 biases towards a solution $\mathcal{C}$ with a densely intra-connected modules. On the other hand, the second term expresses that the expected value of the same edge density in $\mathcal{C}$ with the same community structure $\{C_1, \dots, C_K\}$ but fall at random between the vertices should be small. Q will approach its minimum at 0 if the number of within-community edges is no better than random. On the other hand, values approaching $Q = 1$, which is the maximum, indicate strong community structure.

$$max\ Q(\mathcal{C}) = \sum_{k=1}^{K} \left[\frac{m_i}{|L|} - \left(\frac{\sum_{v \in C_k} |l_v|}{2|L|} \right)^2 \right] \quad (6)$$

$Community\ score$, CS in Eq. 7 is considered as a global quality measure of $\mathcal{C} = \{C_1, \dots, C_K\}$, allowing the detection of the maximal and dense partitions. $Conductance$, CO and $Normalized\ cut$ in Eq. 8 and Eq. 9, respectively, measure the fraction of inter-connections of a clustering solution $\mathcal{C} = \{C_1, \dots, C_K\}$. $Internal\ density$, ID, qualifies a partitioning solution according to the internal edge density, while $Expansion$, EX, and $Cut\ ratio$, CR, qualify the solution based on the number of inter-edges per node (refer to Eq. 10, Eq. 11, and Eq. 12, respectively).

$$max\ CS(\mathcal{C}) = \sum_{k=1}^{K} \sum_{v \in C_k} \left(\frac{\sum_{v \in C_k} in_k(v)}{n_k} \right)^2 \times \frac{2m_k}{n_k} \quad (7)$$

$$min\ CO(\mathcal{C}) = \sum_{k=1}^{K} \frac{\sum_{v \in C_i} out_k(v)}{2m_k + \sum_{v \in C_k} out_k(v)} \quad (8)$$

$$min\ NC(\mathcal{C}) = \sum_{k=1}^{K} \frac{\sum_{v \in C_k} out_i(v)}{2m_k} + \frac{\sum_{v \in C_k} out_i(v)}{2(|L| - m_k) + \sum_{v \in C_k} out_k(v)} \quad (9)$$

$$min\ ID(\mathcal{C}) = \sum_{k=1}^{K} 1 - \frac{m_k}{n_k(n_k - 1)/2} \quad (10)$$

$$min\ EX(\mathcal{C}) = \sum_{k=1}^{K} \frac{\sum_{v \in C_k} out_k(v)}{n_k} \quad (11)$$

$$min\ CR(\mathcal{C}) = \sum_{k=1}^{K} \frac{\sum_{v \in C_k} out_k(v)}{n_k(n - n_k)} \quad (12)$$

These studies showed that EA based methods deserve the credits as a powerful and competitive computational technique to cope with complex detection problem. However despite their success on this problem, the characteristic components of the adopted methods are still in their more or less traditional forms. They provide single-objective community detection being modeled with the very general form of EA. In other words, they didn't exploit any possible heuristic to harness the strength of the adopted models. In the next section a heuristic operator is introduced to improve the quality of the solutions provided by the state-of-the-art EA based models.

3.3 The proposed protein-complex attraction and repulsion strategy

To improve the performance of any evolutionary algorithm, one should design some problem-specific operators. In this paper, we propose a heuristic perturbation $h: I \rightarrow I'$ operator that is tailored for the complex detection problem in PPI networks. The proposed h, called protein-complex attraction and repulsion operator, is designed to fulfill the topological properties at the complex level.

The main guidelines on the design of h operator is to fine-grain the structure of a complex by releasing its sparse interactions and delimiting its boundary as close as possible towards a core protein (e.g. many complexes in nature are very small composed of only two or three proteins). How to fix the boundary of a complex depends on our perspective towards utilizing topological features of proteins complexes. In each complex C_k, two distinct proteins are identified, these are, core protein (Eq. 13) and odd protein (Eq. 14). Core protein $core_k$ is protein v which belong to C_k and satisfies, in terms of distance closeness centrality (CC), the closest protein with respect to all other proteins of C_k. Odd protein odd_k, on the other hand, represents the furthest away protein with respect to all other proteins of C_k.

$$core_k = \underset{v \in C_k}{argmin} \frac{\sum_{w \in C_k} dis(v,w)}{n_k} \quad (13)$$

$$odd_k = \underset{v \in C_k}{argmax} \frac{\sum_{w \in C_k} dis(v,w)}{n_k} \quad (14)$$

Note that the distance between two proteins v and w in C_k, $dis(v,w)$, is computed by considering Dijkstra shortest path between v and w.

Based on the above considerations, complex C_k, then, is allowed to be divided into two more complexes, one complex is structured to attract proteins close to $core_k$ while the second complex is centered around odd_k. Here, we simply claim that there should be a strong relation among proteins, so that proteins with small distance to either core or odd protein should form together unique function. Thus, one relaxed or coarse-grained complex can be divided further into two more compact complexes. Let us assume that $C_{k,core}$ and $C_{k,odd}$ are the two complexes derived from C_k after identifying $core_k$ and odd_k, then, the association of the remaining proteins of C_k to $C_{k,core}$ and $C_{k,odd}$ can be specified by attraction operation following Eq. 15 and by repulsion operator following Eq. 16, respectively.

$\forall v \in C_k$:

$$C_{k,core} = \{v | dis(v, core_k) < dis(v, odd_k)\} \quad (15)$$

$$C_{k,odd} = \{v | dis(v, odd_k) < dis(v, core_k)\} \quad (16)$$

Moreover, we can repeatedly fine-grained over-sized complexes until a certain minimum size is reached. Note that very coarse-grained condition is satisfied if the size of the complex exceeds the maximum size of a true or reference complex taken from the golden ttrue reference set. We suggest to apply complex division at every generation t, however, enlarged, or coarse-grained, complexes are considered for further divisions after every t_{gap} of generations. Algorithm 1 outlines the main steps of the proposed protein-complex attraction and repulsion operator.

4. Experimental results

In the performance evaluations, yeast *Saccharomyces cerevisiae* PPI network is used. Yeast is proven to be highly effective PPI network of model organism for mammalian biological functions and diseases. PPI_D1 network was prepared by Gavin et al. [14] and filtered by Zaki et al. [15]. The filtered version of this network contains 990 proteins with 4687 interactions.

Algorithm 1: Protein-complex attraction and repulsion operator (I_i; I_i')

Input: $I_i | \ i \in \{1,2,\dots,N\}$

Output: $I_i' | \ i \in \{1,2,\dots,N\}$

1: Decode I_i to a partition set $\mathcal{C} = \{C_1, \dots, C_K\}$

2: **for** $k \leftarrow 1$ **to** K

3: $n_k \leftarrow |C_k|$ // degree of C_k

4: // find out core and odd proteins in C_k

// 1. find closeness centrality CC for each protein

$\forall v \in C_k$: $CC_v \leftarrow \frac{\sum_{w \in C_k} dis(v,w)}{n_k}$

// 2. find core protein having closest centrality

$$core_k \leftarrow \underset{v \in C_k}{argmin} CC_v$$

// 3. find odd protein having furthest centrality

$$odd_k \leftarrow \underset{v \in C_k}{argmax} CC_v$$

5: // divide C_k into two complexes centered at $core_k$ and odd_k

//1. Initialize complexes to empty

$C_{k,core} = \emptyset$

$C_{k,odd} = \emptyset$

//2. Re-assign proteins of C_k to $C_{k,core}$ or $C_{k,odd}$

$\forall v \in C_k$: **if** $(dis(v, core_k) < dis(v, odd_k))$

$C_{k,core}$=$C_{k,core} \cup v$

Else

$C_{k,odd}$=$C_{k,odd} \cup v$

End

6: **end**

To validate the quality of the predicted complexes, a reference set, denoted as $Cmplx\_D1$, drawn from the Munich Information Center for Protein Sequence (MIPS) catalog is used in the experiments [16]. $Cmplx\_D1$ contains 81 complexes of sizes ranges from 6 and 38. Mathematically, these can be expressed as $S^* = \{S_1, S_2, \dots, S_{K_S}\}$, $|S^*| = K_S = 81$ and $\forall_{C_i \in S^*}\ 6 \leq n_i \leq 38$.

A predicted cluster C_i is said to match a gold standard complex S_j if their proteins are overlapped or intersected with overlapping score, OS equals or larger than a specific threshold σ_{OS} [17].

$$OS(C_i, S_j) = \frac{|C_i \cap S_j|^2}{|C_i||S_j|} \quad (17)$$

$$match(C_i, S_j) = \begin{cases} 1 \ if \ OS(C_i, S_j) \geq \sigma_{OS} \\ 0 \ otherwise \end{cases} \quad (18)$$

where $|\cdot|$ is the number of proteins common to both a predicted cluster and a gold standard complex. Based on matching expressed in Eq. 18, the notions of $recall$, $precision$, and cumulative F score are

defined. Recall refers to the fraction of gold standard complexes that are matched to any predicted cluster. As $match$ function returns either 0 or 1, then $max_{C_j \in \mathcal{C}} match(S_i, C_j)$ in the numerator of Eq. 19 will be computed to 1 if a given gold standard complex S_i has one or more matches with the predicted clusters. Precision, on the other hand, refers to the fraction of predicted clusters that are matched to any gold standard complex (Eq. 20). A harmonic mean of both recall and precision is reflected by F score (Eq. 21).

$$recall = \frac{|S_i | S_i \in S^* \wedge \exists C_j \in \mathcal{C} \rightarrow match(S_i, C_j)|}{K_S} \quad (19)$$

$$precision = \frac{|C_i | C_i \in \mathcal{C} \wedge \exists S_j \in S^* \rightarrow match(C_i, S_j)|}{K_{\mathcal{C}}} \quad (20)$$

$$F = \frac{2 * recall * precision}{recall + precision} \quad (21)$$

As $recall$ and $precision$ evaluate the cumulative quality of the prediction at the complex level, $recall_N$ and $precision_N$ (Eq. 22 and Eq. 23, respectively) can similarly estimate the accuracy of the prediction, however, at the protein level [15]. For these two measurements, we formulate an F_N score (Eq. 24) to imitate F score, but at the protein level.

$$recall_N = \frac{\sum_{i=1}^{K_S} |m_i|}{\sum_{i=1}^{K_S} |S_i|} \quad (22)$$

where $|m_i| = max_{C_j \in \mathcal{C}} |match(S_i, C_j)$.

$$precision_N = \frac{\sum_{i=1}^{K_{\mathcal{C}}} |m_i|}{\sum_{i=1}^{K_{\mathcal{C}}} |C_i|} \quad (23)$$

where $|m_i| = ax_{S_j \in S^*} |match(C_i, S_j)$.

$$F_N = \frac{2 * recall_N * precision_N}{recall_N + precision_N} \quad (24)$$

In the experiments, the seven single EA based models presented in Eq. 6 – Eq. 12 are annotated as CO, Q, EX, CR, NC, ID, and CS. In the simulation runs, population size, $PopSize$, and maximum number of generations, $iter_{max}$, are set to 100, probabilities of crossover and mutation are set to 0.8 and 0.2, respectively, and $t_{gap} = 10$. The objective of the experiments is to test the impact of the proposed heuristic operator on the final prediction power of all EA models. The results present the performance of the tested models operating with no heuristic and when there exists collaboration between the model and the proposed heuristic operator. Here, performance evaluation is reported in terms of F and F_N scores evaluation metrics presented in Eq. 21 and Eq. 24. Moreover, the threshold σ_{OS} is assumed to vary from 0.1 to 0.5 in steps of 0.05. Figures-2, 3, 4, 5, 6, 7 and 8 report the performance evaluations in terms of F and F_N scores for the seven tested complex detection models.

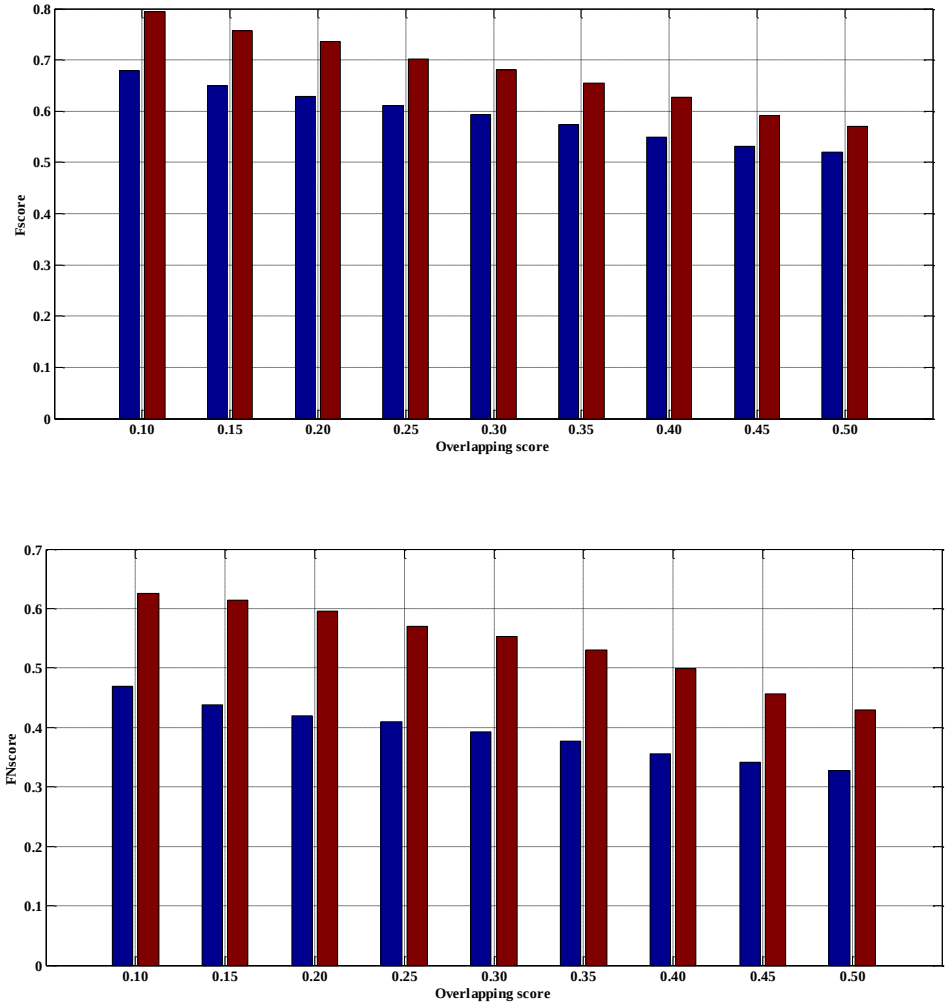

Figure 2- Performance evaluation of conductance model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

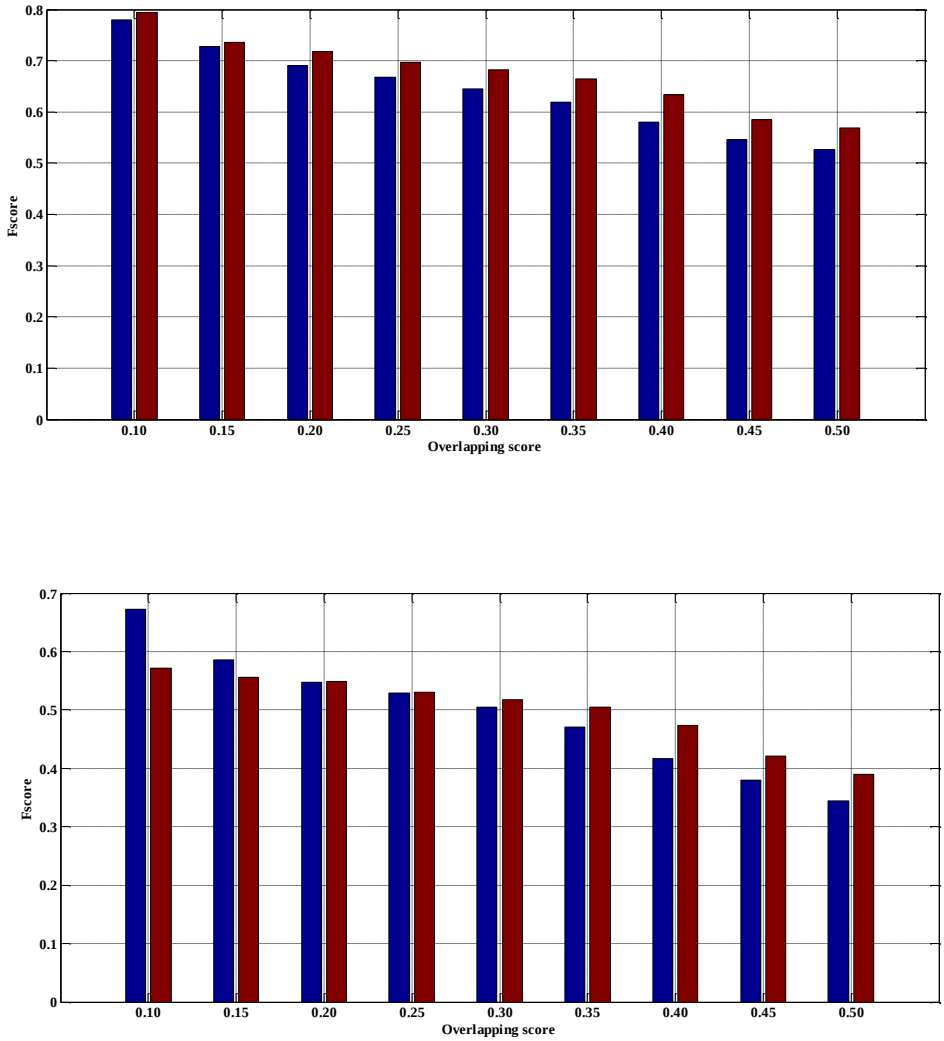

Figure 3- Performance evaluation of Q model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

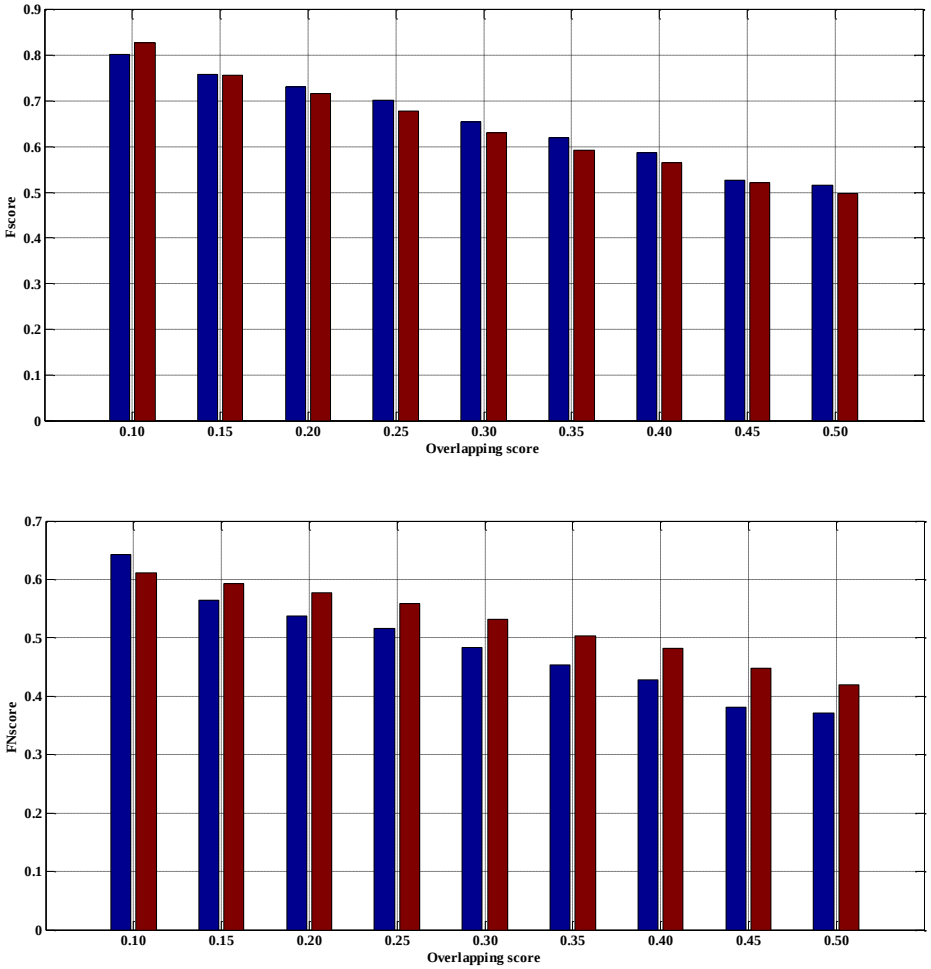

Figure 4 - Performance evaluation of expansion model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

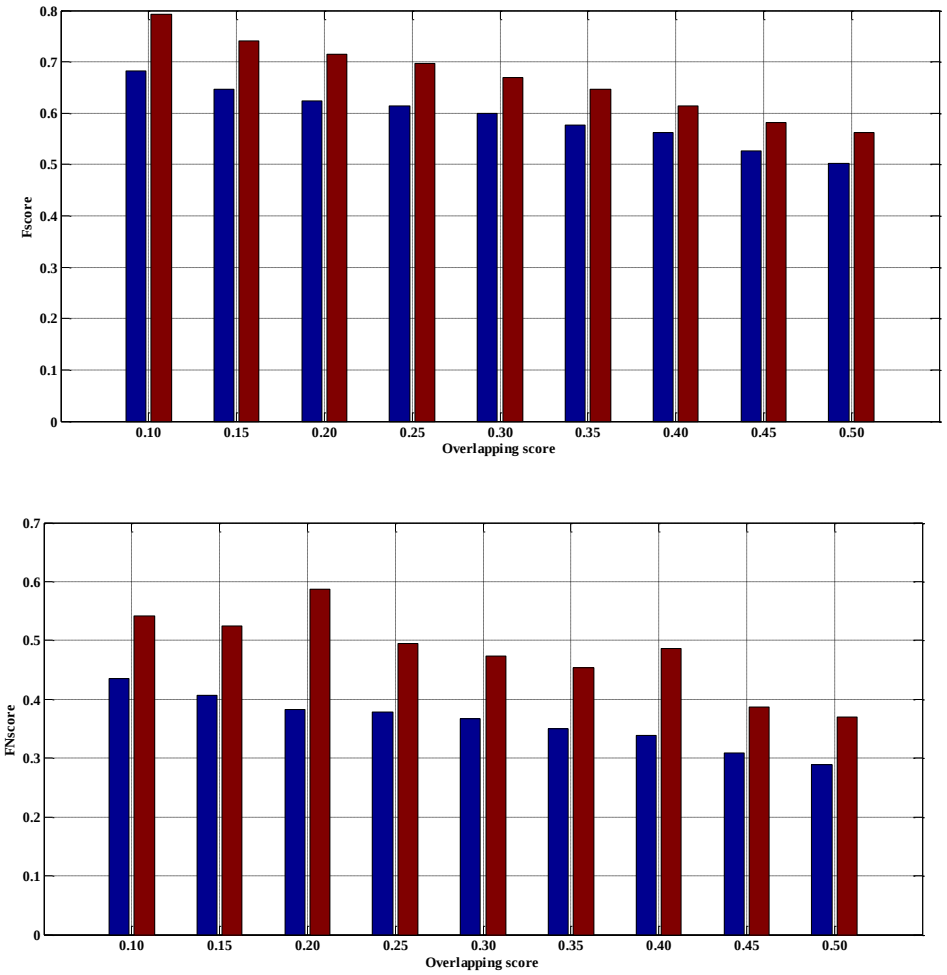

Figure 5 - Performance evaluation of cut ratio model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

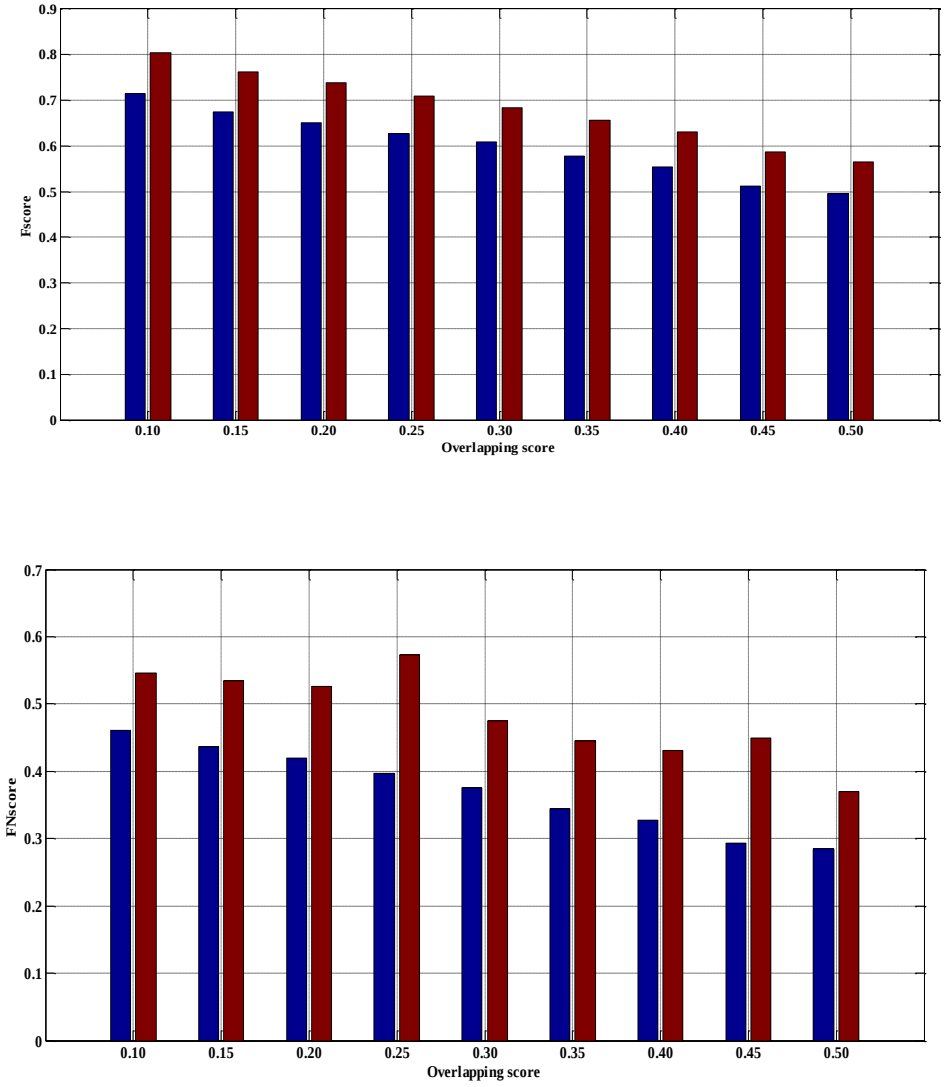

Figure 6 - Performance evaluation of normalized cut model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

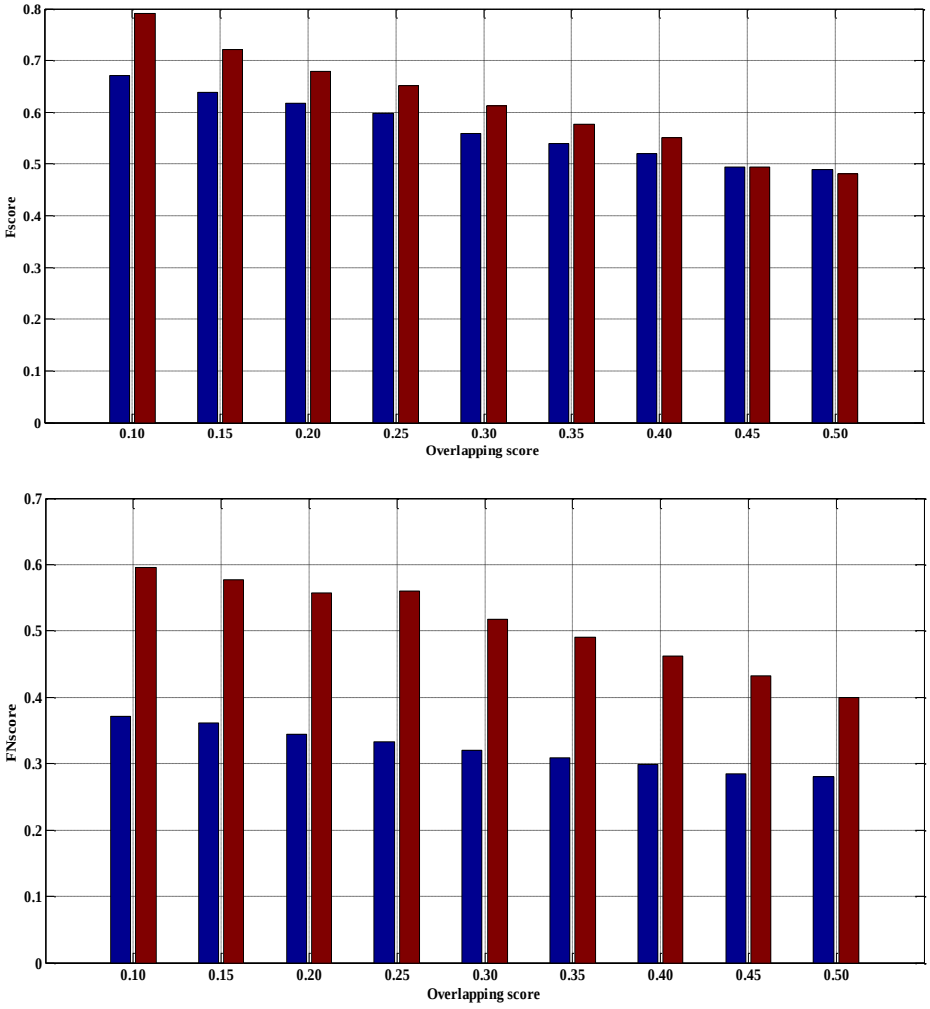

Figure 7 - Performance evaluation of internal density model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

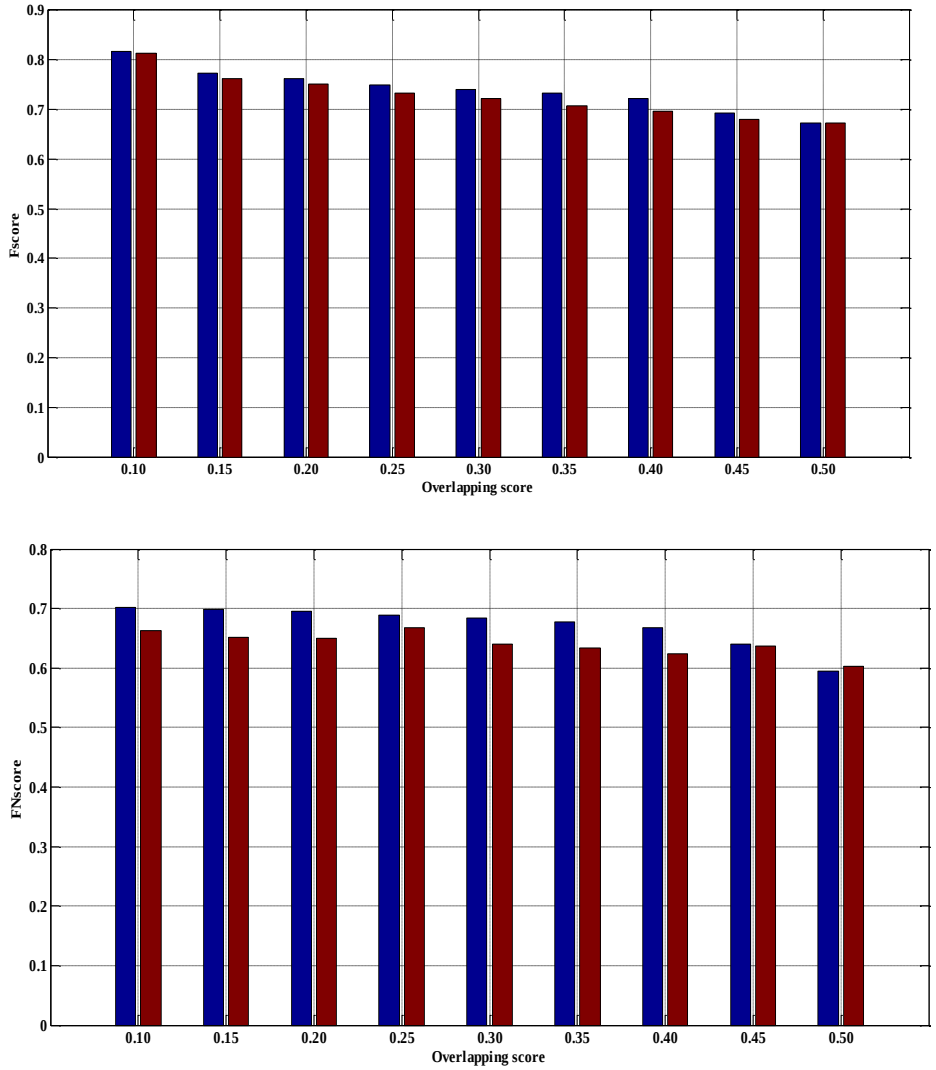

Figure 8 - Performance evaluation of community score model (F (top) and F_N(bottom) with no heuristic (left bar), and in heuristic version (right bar).

The introduction of the heuristic operator is supposed (due to its mechanism) to improve the prediction power of the EA based complex detection model. The results reported in Figures- 2 and 8 reveal that conductance Figure-2, cut ratio Figure-5, normalized cut Figure-6, and internal density Figure-7 models have positive collaboration with the proposed heuristic operator to improve their detection ratio in terms of F and F_N scores. For modularity, one can see that the proposed heuristic operator is beneficial in terms of F and in almost all results of F_N score. The performance of community score model, however, is generally bewildered by the introduction of the heuristic operator (compare left bars against right bars in Figure-8).

References

1. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. **2002**. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, *417*(6887), 399-403.
2. Bader, G. D., & Hogue, C. W. **2003**. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, *4*(1), 1.

3. King, A. D., Pržulj, N., and Jurisica, I. **2004**. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17), 3013–3020.

4. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., & Kanaya, S. **2006**. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC bioinformatics*, *7*(1), 207.

5. Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., & Vicsek, T. **2006**. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8), 1021-1023.

6. Palla, G., Derényi, I., Farkas, I., & Vicsek, T. **2005**. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814-818.

7. Pizzuti, C., & Rombo, S. E. **2007**. Pincoc: a co-clustering based approach to analyze protein-protein interaction networks. IDEAL'07 Proceedings of the 8th international conference on Intelligent Data Engineering and Automated Learning ,pp. 821-830, Springer Berlin Heidelberg.

8. Pizzuti, C., & Rombo, S. E. **2008**. Discovering meaningful protein-protein interaction modules by a co-clustering based approach. In *SEBD*, pp. 294-301.

9. Pizzuti, C., & Rombo, S. E. **2012**. A coclustering approach for mining large protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *9*(3), 717-730.

10. Pizzuti, C., & Rombo, S. E. **2014**. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, *30*(10), 1343-1352.

11. Garey M.R. & D. S. Johnson, D.S., **1979**. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company.

12. Noack, A. **2007**. Energy Models for Graph Clustering. *J. Graph Algorithms Appl.*, *11*(2), 453-480.

13. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. **2004** Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U S A*,101(9), pp: 2658-2663.

14. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., & Edelmann, A. **2006**. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, *440*(7084), 631-636.

15. Zaki, N., Berengueres, J., & Efimov, D. **2012**. Detection of protein complexes using a protein ranking algorithm. *Proteins: Structure, Function, and Bioinformatics*, *80*(10), 2459-2468.

16. Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., & Weil, B. **2002**. MIPS: a database for genomes and protein sequences. *Nucleic acids research*, *30*(1), 31-34.

17. Brohée, S. and van Helden, J. **2006**. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, *7*(1), pp: 1-19.