A Multiobjective GO based Approach to Protein Complex Detection

Sumanta Ray\textsuperscript{a}, Moumita De\textsuperscript{b}, Anirban Mukhopadhyay\textsuperscript{c}

\textsuperscript{a,b,c}Department of Computer Science and Engineering, University of Kalyani, Kalyani-741235, West Bengal, India

Emails: \textsuperscript{a}sumanta_ray86@rediffmail.com, \textsuperscript{b}moumita.de2013@gmail.com, \textsuperscript{c}anirban@klyuniv.ac.in

Abstract

Considering the large number of proteins and the high complexity of the protein interaction network, decomposing it into smaller and manageable modules is a valuable step toward the analysis of biological processes and pathways. There are different types of graph clustering approaches available for modularization of PPI networks. But most of them seek to identify groups or clusters of proteins on the basis of the local density of the interaction graph in the network. So they find the denser region in the interaction graph, as the denser regions are expected to have the high chance to form a protein complex. In this article we present a novel Multiobjective Gene Ontology based Clustering (MGOC) in PPI network- to find protein complexes. We consider both graphical properties of PPI network as well as biological properties based on GO semantic similarity as objective functions. The proposed technique is demonstrated on yeast PPI data and the results are compared with that of some existing algorithms.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

Keywords: Protein-protein interaction, protein complex, multiobjective genetic algorithm, gene ontology.

1. Introduction

A PPI network can be described as a complex system of proteins linked by interactions. The simplest representation takes the form of a mathematical graph consisting of nodes and edges [1]. Proteins are represented as nodes and the interaction of two proteins are represented as adjacent nodes connected by an edge. The component proteins within protein-protein interaction (PPI) networks are associated in two types of groupings: protein complexes and functional modules. Protein complexes are assemblages of proteins that interact with each other at a given time and place, forming a single multimolecular machine. Functional modules consist of proteins that participate in a particular cellular process while binding to each other at various times and places. For the sake of simplicity, in this paper we use protein complexes and functional modules as a same concept. The main objective of PPI network clustering is finding dense subgraphs that represent functional modules and protein complexes. Identification of functional modules in protein interaction networks is the first step in understanding the organization and dynamics of cell function.

Previous PPI network clustering methods detect near-clique subgraphs in a network [2, 3, 4]. Clique is a complete graph where there is an edge between each pair of its vertices. A molecular complex detection
(MCODE) algorithm is proposed in [5] that is an efficient approach for detecting densely-connected regions in the PPI networks.

In this article the problem of finding protein complexes in a Protein-Protein Interaction Network (PPIN) is modeled as a Multiobjective Optimization problem. The searching is performed over a number of objectives and the final solution set contains a number of Pareto-optimal solutions (each solution represents a protein complex). Here we used the information content based semantic similarity measure between GO terms of each pair of proteins as an objective. Some of the graphical properties of protein interaction network are used as other objectives. The results were evaluated by comparing to other well-known MCODE [5] and ClusterONE [6] approaches.

2. Multi-objective Optimization using Genetic Algorithms

The multi-objective optimization can be formally stated as follows [7]: Find the vector \( \mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T \) of the decision variables which will satisfy the \( m \) inequality constraints: \( g_i(\mathbf{x}) \geq 0, \quad i = 1, 2, \ldots, m \), the \( p \) equality constraints \( h_i(\mathbf{x}) = 0, \quad i = 1, 2, \ldots, p \), and optimizes the vector function \( \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})]^T \). The constraints define the feasible region \( \mathcal{F} \) which contains all the admissible solutions. The vector \( \mathbf{x}^* \) denotes an optimal solution in \( \mathcal{F} \). The concept of Pareto optimality comes handy in the domain of multi-objective optimization. A formal definition of Pareto optimality from the viewpoint of minimization problem may be given as follows: A decision vector \( \mathbf{x}^* \) is called Pareto optimal if and only if there is no \( \mathbf{x} \) that dominates \( \mathbf{x}^* \), i.e., there is no \( \mathbf{x} \) such that \( \forall i \in \{1, 2, \ldots, k\}, f_i(\mathbf{x}) \leq f_i(\mathbf{x}^*) \) and \( \exists i \in \{1, 2, \ldots, k\}, f_i(\mathbf{x}) < f_i(\mathbf{x}^*) \). In words, \( \mathbf{x}^* \) is Pareto optimal if there exists no feasible vector \( \mathbf{x} \) which causes a reduction on some criterion without a simultaneous increase in at least one other. In general, Pareto optimum usually admits a set of solutions called non-dominated solutions. There are different approaches to solving multi-objective optimization problems [8] e.g., aggregating, population based non-Pareto and Pareto-based techniques. Here we use NSGA-II [9] as underlying multi-objective algorithm for finding protein complex.

3. GO-based Semantic Similarity Measure

The Gene Ontology (GO) project [10] is a collaborative effort to provide consistent description of genes and gene products. GO provides a collection of well-defined biological terms, which are called GO terms. The GO terms are shared across different organisms. They comprise three categories as the most general concepts: biological processes, molecular functions and cellular components. The GO terms are structured by the relationships to each other, such as is-a and part-of. The measurement of semantic similarity between two concepts can be easily extended to measure the degree of similarity between terms in the GO structures. In Information Content based method the semantic similarity between two concepts can be measured based on the commonality of their information contents. The information content of the concept \( C \) is formulated by the negative log likelihood – log \( p(c) \), where \( p(c) \) is the probability of encountering a concept. According to Lin [11] every object is considered to belong to a class of taxonomy. Similarity between two objects is the similarity between those two classes.

In this article we use three different taxonomies i.e Biological Process, Molecular Function, Cellular Component of the Lin [11] measure and use this measures as objective function for computing three different results.

4. Proposed Method

In this section the multiobjective GO based clustering (MGOC) algorithm in PPI network is described.

4.1. Chromosome Representation

A protein complex is nothing but a subgraph of the whole protein protein interaction graph. Here a protein complex is encoded as a chromosome. So in the resulting population a chromosome of the type: \( n_1 \ n_2 \ n_3 \ \ldots \ n_p \) represent a protein complex consisting of \( p \) number nodes or proteins. All nodes in the chromosome are not necessarily connected.
4.2. Population Initialization
Initially the whole network is broken in several numbers of biclusters. Biclustering is done by applying K-means clustering from both the dimensions of PPI matrix and taking intersections of the clusters formed in these two dimensions. Each bicluster represents a densely connected region in the network. We sort these biclusters on the basis of density and pick up first 50 biclusters and use these in the initial population. The next and subsequent populations are created using the genetic operator of multiobjective GA.

4.3. Representation of Objective Functions
Here we use two types of objective functions: one is totally depended on the typical graph properties of protein interaction network and another is based on Gene Ontological annotations of proteins.

4.3.1. Graph Based Objective
All graph theoretic approaches for finding protein complexes seek to identify dense subgraph by maximizing the density of each subgraph on the basis of local network topology. The density of a graph is the ratio of the number of edges present in a graph to the possible number of edges in a complete graph of the same size. As there are large number of interaction (or edges) between proteins (or nodes) in a protein complex (or subgraph) so the density of each complex (or subgraph) is generally very high. So using density as an objective function and maximizing it for individual subgraphs will yield much better and denser complexes. For choosing next objective we count the number of interconnecting nodes that are not present in the current cluster. For example in Fig 1 the chromosome is represented as black nodes and the interconnecting nodes of this chromosome (which are not present in the current chromosome) is shown in yellow colored nodes. This may be written as: \( N(C) = \bigcup_{i \in C} n_i \), where \( C \) is any cluster in \( G \) and \( n_i \) is the number of nodes which are connected with node \( i \) in \( C \), and are not present in \( C \). Minimizing this will result clusters, which have lesser number of outward interaction partners and we get compact clusters.

4.3.2. Semantic Similarity Based Objective
The semantic similarity measured between two GO terms can be directly converted to a measurement of the similarity between two proteins. Since a protein is annotated to multiple GO terms, several researchers [12] have defined the similarity between two proteins as the average similarity of the GO term cross pairs, which are associated with both interacting proteins. The package csbl.go is used for calculating similarity matrix \( S \). Here we use the similarity measure proposed by Lin [11] to compute the similarity matrix. Now for calculating the fitness of a chromosome in our proposed multiobjective approach the average similarity of each pair of protein comprising the chromosome is computed. For example to compute the fitness of the chromosome: \( \{n_1, n_2, \ldots, n_p\} \), we compute a submatrix \( s \) with rows and columns comprising these nodes from the similarity matrix \( S \). Now the fitness function may be written as: \( \text{sim}(s) = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} s(i,j)}{p} \). We can group the functionally similar proteins in same cluster, by maximizing this function.

4.4. Selection and Mutation
The popularly used genetic operations are selection, crossover, and mutation. General crossover operation between two chromosomes results many disconnected subgraphs and produces a large number of...
isolated nodes. So crossover is not performed here and instead mutation is performed with high probability (mutation probability =0.9). The selection operation used here is the crowded binary tournament selection used in NSGA-II. If a chromosome is selected to be mutated then addition or deletion of nodes in the chromosome is performed in the following way: for a chromosome a random node is selected and either of the two tasks is performed with equal probability: Delete the node or add the nodes which, are direct neighbor of node $n_i$ and, are not included in the parent chromosome. The whole operation is performed five times to create new diversified chromosome from the parent chromosome.

5. Experimental Results

We run the proposed algorithm MGOC on the PPI network of Saccharomyces Cerevisiae (yeast) dataset downloaded from the DIP [13]. Out of 5000 S.Cerevisiae proteins we use 4669 proteins due to the availability of their annotation data. Table 1 summarizes the PPI network. For comparison of the result of MGOC with other algorithms MCODE and ClusterONE the following parameters are used: For MCODE the basic parameters used are: Node Score Cuttoff=0.2, Maximum Depth=100, K-Score=2. For ClusterONE the basic parameters used are: Minimum Cluster Size=10, Minimum Density=0.3, Overlap Threshold=0.8. For MGOC the basic parameters are Population Size=50, Number of Generations=5, Mutation Probability=0.9.

We compute the proposed MGOC algorithm based on each of the three orthogonal taxonomies or aspects that hold terms describing the Molecular Function (mf), Biological Process (bp) and Cellular Component (cc) for a gene product. The results are verified by biological and non-biological criteria [14, 15]. One of the non-biological or topological criteria of comparison is network modularity proposed by Garvin and Newman [16]. The network modularity is a metric for assessment of partitioning a network into clusters defined as, $Q = \sum_{i} (e_{ii} - a_i^2)$, where $i$ is the index of cluster, $e_{ii}$ is the number of edges or interactions which have both ends located in the $i$ th cluster, and $a_i$ is the number of interactions or edges which have exactly one end located in the $i$ th cluster. In Table 2 we compare MCODE and ClusterONE with our proposed algorithm MGOC depending on the value of $Q$. It shows that network modularity of MGOC is higher than that of the other methods.

For biological validation we match our clustering result with the known protein complexes consist of 491 complexes, downloaded from the site http://yeast-complexes.russelllab.org/. We build a Contingency Table with rows as protein complexes and columns as resulting clusters. So, the contingency table $T$ is a $n \times m$ matrix having $n$ complexes and $m$ resulting clusters, where row $i$ corresponds to the $i$-th annotated complex, and column $j$ to the $j$-th cluster. The value of a cell $T_{i,j}$ indicates the number of proteins found in common between complex $i$ and cluster $j$. Some proteins belong to several complexes, and some proteins may be assigned to multiple clusters or not assigned to any cluster.

Sensitivity, positive predictive value (PPV), and accuracy are classically used to measure the correspondence between the result of a classification and a reference. Now we compare the results with respect to these measures.

5.1. Sensitivity

Sensitivity is the fraction of proteins of complex $i$ found in predicted cluster $j$: $S_{n_{ij}} = \frac{T_{ij}}{N_i}$, where $N_i$ is the number of proteins belonging to complex $i$. A complex-wise sensitivity $S_{n_{e0}}$ may be defined as:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#Proteins</th>
<th>#Interactions</th>
<th>Min. Degree</th>
<th>Avg. Degree</th>
<th>Max. Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIP</td>
<td>4669</td>
<td>21217</td>
<td>0</td>
<td>9.2305</td>
<td>241</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MCODE</th>
<th>ClusterONE</th>
<th>MGOC_bp</th>
<th>MGOC_cc</th>
<th>MGOC_mf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q Value</td>
<td>-16572751</td>
<td>-10742041</td>
<td>-819555</td>
<td>-124669</td>
</tr>
</tbody>
</table>

Table 1. Summary of the PPI network data set used here

Table 2. Network modularity comparison between the MCODE, ClusterONE and MGOC approaches
Table 3. Summery of clustering result with respect to Senesitivity, PPV and Accuracy

<table>
<thead>
<tr>
<th></th>
<th>MCODE</th>
<th>ClusterONE</th>
<th>MGOC-mf</th>
<th>MGOC-bp</th>
<th>MGOC_cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Sensitivity</td>
<td>0.1168</td>
<td>0.2135</td>
<td>0.2490</td>
<td>0.2726</td>
<td>0.2321</td>
</tr>
<tr>
<td>General PPV</td>
<td>0.4922</td>
<td>0.4078</td>
<td>0.8186</td>
<td>0.8184</td>
<td>0.6891</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.2397</td>
<td>0.2951</td>
<td>0.4515</td>
<td>0.4711</td>
<td>0.4000</td>
</tr>
</tbody>
</table>

Table 4. The number of complexes covered by different algorithm

<table>
<thead>
<tr>
<th></th>
<th>MCODE</th>
<th>ClusterONE</th>
<th>MGOC-mf</th>
<th>MGOC-bp</th>
<th>MGOC_cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Of Cluster</td>
<td>53</td>
<td>64</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Matched Complex</td>
<td>197</td>
<td>163</td>
<td>301</td>
<td>308</td>
<td>266</td>
</tr>
<tr>
<td>Matched Cluster</td>
<td>42</td>
<td>61</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ S_{n_{i,j}} = \max_{j=1}^m S_{n_{i,j}} \]

The General Sensitivity \((S_n)\) is the weighted average of \(S_{n_{i,j}}\) over all complexes and defined as:
\[
S_n = \frac{\sum_{i=1}^n N_i S_{n_{i,j}}}{\sum_{i=1}^n N_i}.
\]

5.2. Positive Predictive Value

The positive predictive value is the proportion of members of predicted cluster \(j\) which belong to complex \(i\), relative to the total number of members of this cluster assigned to all complexes:
\[
PPV_{i,j} = \frac{T_{i,j}}{T_j} = \frac{t_{i,j}}{T_j}
\]

where \(T_j\) is the marginal sum of a column \(j\). Cluster-wise positive predictive value \(PPV_{cl,j}\), represents the maximal fraction of proteins of cluster \(j\) found in the same complex:
\[
PPV_{cl,j} = \max_{i=1}^m PPV_{i,j}.
\]

The General PPV \((PPV)\) of a clustering result is the weighted average of clustering-wise-PPV \((PPV_{cl,j})\) over all predicted clusters:
\[
PPV = \frac{\sum_{j=1}^m T_j PPV_{cl,j}}{\sum_{j=1}^m T_j}.
\]

5.3. Accuracy

The Geometric Accuracy \((Acc)\) represents a tradeoff between sensitivity and the positive predictive value and it is defined as:
\[
Acc = \sqrt{S_n * PPV}.
\]

It is the geometrical mean of the \(S_n\) and the \(PPV\). The advantage of taking the geometric is that it yields a low score when either the \(S_n\) or the \(PPV\) metric is low. High accuracy values thus require a high performance for both criteria.

Table 3 shows the general sensitivity, PPV and accuracy of different algorithms including our proposed technique. From this table it is clear that our proposed algorithm MGOC shows much better result with respect to sensitivity and positive predictive value.

Table 4 shows the number of protein complexes covered by different algorithms. First row represents the number of clusters which are identified by the corresponding methods or algorithm in the whole network. The second row represents the number of complexes covered by the corresponding algorithm. Here we say that a complex is matched against any one of the cluster, when at least one of the core protein in that complex is included in any one of the clusters. In Table 4 we see that MGOC identifies over 300 of such protein complexes which is much better than the MCODE or ClusterONE algorithm. The third row represent the number of clusters in which at least one of the core proteins belongs. In MCODE 11 clusters and in ClusterONE 3 clusters have no core proteins, but in MGOC we see that all clusters have at least one of the core proteins within it.

6. Conclusions

In this article we present a Multiobjective GO based Genetic Algorithm for finding protein complexes in the protein-protein interaction network. Here some graphical properties of PPIN is used as objective functions. To incorporate the functional similarity we use the semantic similarity measure of GO terms between protein pairs as another objective function. The information content based semantic similarity measure proposed by Lin [11] is used. As a future work other semantic similarity measures can be used as objective functions.
References