Detecting Protein Complexes from Noisy Protein Interaction Data

Dmitry Efimov
Fac. of Mechanics and Math.
Moscow State University,
Moscow
diefimov@gmail.com

Nazar Zaki†
Intelligent Systems
Faculty of Info. Technology
UAEU, Al Ain 17551, UAE
nzaki@uaeu.ac.ae

Jose Berengueres‡
Intelligent Systems
Faculty of Info Technology
UAEU, Al Ain 17551, UAE
jose@uaeu.ac.ae

ABSTRACT
High-throughput experimental techniques have made available large datasets of experimentally detected protein-protein interactions. However, experimentally determined protein complexes datasets are not exhaustive nor reliable. A protein complex plays a key role in disease development. Therefore, the identification and characterization of protein complexes involved is crucial to the understanding of the molecular events under normal and abnormal physiological conditions. In this paper, we propose a novel graph mining algorithm to identify protein complexes. The algorithm first checks the quality of the interaction data, then predicts protein complexes based on the concept of weighted clustering coefficient. To demonstrate the effectiveness of our proposed method, we present experimental results on yeast protein interaction data. The level of accuracy achieved is a strong argument in favor of the proposed method. Novel protein complexes were also predicted to assist biologists in their search for protein complexes. The datasets and programs are freely available from http://faculty.uaeu.ac.ae/nzaki/PE-WCC.htm.

Categories and Subject Descriptors
G.2.2 [GRAPH THEORY]: Network problems

†Dr. Nazar Zaki (corresponding author) is the Director of the Bioinformatics Lab and Coordinator of the Intelligent Systems program at the Faculty of Information Technology, United Arab Emirates University, Al Ain, P. O. Box 17551, UAE, Email: nzaki@uaeu.ac.ae
‡Dr. Jose Berengueres is the Director of the Media Lab and Assistant Professor of Intelligent Systems, Faculty of Information Technology, United Arab Emirates University, Al Ain, P. O. Box 17551, UAE, Email: jose@uaeu.ac.ae

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2012 ACM 978-1-4503-1552-4...S15.00.

1. INTRODUCTION
A protein complex is a group of associated polypeptide chains which play crucial role in disease development [1, 2]. A complex can have various functions inside the cell such as: cellular machines, rigid structures, dynamic signaling and post-translational modification systems. Therefore, the identification of protein complexes is one of the steps necessary to understand the molecular events of the cell.

Several excellent computational approaches have been developed to predict protein complexes from a protein-protein interaction (PPI) network. In this case, PPI is modeled as a graph \( G = (V, E) \) where \( V \) is a set of nodes and \( E \) is a set of edges connecting pairs of nodes. Thus in the PPI networks, the nodes represent proteins and the edges represent interactions. The protein complexes are modeled as dense subgraph of proteins. The density of a graph is the fraction of edges it has out of all possible vertex pairs. One of the most commonly used algorithms for predicting protein complexes via dense protein subnetworks model is molecular complex detection (MCode) algorithm [3]. Other clustering methods for inferring protein complexes include Markov clustering (MCL) [4], restricted neighborhood search clustering (RNSC) [5,6], super paramagnetic clustering (SPC) and CFinder [7]. Leung et al. [8] utilized the core-attachment idea to develop an algorithm called Core to identify complexes from PPI network. However, the accuracy of the protein interaction data produced by high-throughput experiments is often questioned. It is believed to be noisy and incomplete, which makes it difficult to predict complexes accurately. For example, for Y2H screens, it is thought that the false positive rate could be as high as 64%, and the false negative rate may range from 43% to 71% [9]. Sprinzak et al. [10] showed that the reliability of high-throughput yeast two-hybrid assays is about 50%, and that the size of the yeast interactome is estimated to be 10,000 to 16,600 interactions. Therefore, several methods have been proposed to assess the reliability of the high-throughput protein interaction data [11–17]. Other authors were able to improve several clustering algorithms by proposing weighting schemes.
based on the number of common neighbors such as CDdistance [18] and FSWeight [19].

Liu et al. [20] has recently developed an algorithm called Clustering-based on Maximal Cliques (CMC) to discover protein complexes from weighted PPI networks. A clique in a graph is a fully connected subgraph, that is, a subgraph in which every two nodes are connected by an edge. They used an iterative scoring method called AdjustCD to assign weights to protein pairs. The AdjustCD weight in this case indicates the reliability of the interaction between protein pairs. AdjustCD iterative algorithm [18, 21, 22] is mainly based on the number of common neighbors of protein pairs in the PPI network. If two neighbor proteins are denoted as u and v then the CD-distance [21] between these proteins is defined as:

\[
CD(u, v) = 1 - \frac{2|N_u \cap N_v|}{|N_u| + |N_v|} \quad (1)
\]

where \(N_u\) and \(N_v\) are the numbers of neighbors of proteins \(u\) and \(v\), respectively. Equation 1 was further modified by Chua et al. [22] to decrease the CD-distance for proteins with insufficient number of interactions:

\[
\text{AdjustCD}(u, v) = \frac{2|N_u \cap N_v|}{\max(|N_u|, N_{avg}) + \max(|N_v|, N_{avg})} \quad (2)
\]

where \(N_{avg} = \sum_{x \in V} |N_x|\) is the average number of neighbors in the network, \(N\) is the total number of nodes in the network.

Equations 1 and 2, show how many 3-cliques are based on interactions between proteins \(u\) and \(v\), but do not take into account the 3-cliques based on other outgoing interactions from proteins \(u\) and \(v\). To solve this problem, Chua et al. [22] suggested an iterative method which considers all 3-cliques based on interactions from all neighbor proteins \(u\) and \(v\):

\[
w^k(u, v) = \sum_{x,y \in V} \frac{|N_x \cap N_y|}{|N_x| + |N_y|} \quad (3)
\]

where \(w^0(x, u) = 1\), if \(x\) and \(u\) interact, \(w^0(x, u) = 0\), otherwise; \(w_{avg} = \sum_{x \in V} \sum_{y \in V} w^{k-1}(x, y)\) is the average number of weights on \((k-1)^{th}\) step; \(w^1(x, u) = \text{AdjustCD}(x, u)\) and eventually \(w^k(u, v)\) will determine the reliability of interactions between proteins \(u\) and \(v\). CMC showed that the iterative scoring method can significantly improve the performance of CMC and other well known protein complex detection methods. However, CMC works accurately in clean protein interaction data (no false or missing interactions). It is quite difficult to identify unreliable edges or to find maximal cliques when the data is noisy. This weakness is demonstrated in Figure 1. The calculation of the reliability weight of the edge \(e1\) using AdjustCD depends on the outgoing edges \(e6, e7, ..., e10\). In a case of noisy network there is a possibility that many of the outgoing edges \(e6, e7, ..., e10\) may not be reliable. Moreover, the reliability of the edge \(e1\) should not be influenced by the number of outgoing edges.

In this paper, we propose a simple yet effective method for protein complex detection. We are aware that, beside the improvement in graph mining techniques, it is necessary to have high quality benchmarks by assessing the protein interactions reliability. We propose a novel method for assessing the reliability of the interaction data and detecting protein complexes. Unlike CMC, our method works to find the near maximum cliques (maximal cliques without unreliable interactions). We employ the concept of weighted clustering coefficient as a measure for subgraphs that defines which subgraph is closer to the maximal clique. The clustering coefficient of a vertex in this case is the density of its neighborhood [23].

2. METHOD

Computational approaches for detecting protein complexes from PPI data are useful complements to the limited experimental methods such as Tandem Affinity Purification (TAP) [24]. Beside the improvement in graph mining techniques, the success of accurate detection of a protein complex depends on the availability of high quality benchmarks. The bottleneck of different computational methods remains to be the noise associated with the protein interaction data. Therefore, a rigorous assessment of protein interactions reliability is crucial. In this section, we introduce a novel method "PE-WCC" which has two main steps: first, assess the reliability of the protein interaction data using PE-measure. Second, detect protein complexes using weighted clustering coefficient [23]. In the subsequent sections, we describe the two steps in details.

2.1 Assessing the reliability of protein interactions

In this section we introduce "PE-measure", a new measure for protein pairs. PE-measure enables us to reduce the level of noise associated with PPI networks and it is defined as follows:

Given a PPI network with \(N\) proteins. We represent the PPI network by an undirected graph \(G = (V, E)\), where the vertex set \(V\) represents the proteins, edge set \(E\) represents the set of interactions between pairs of proteins. The elements \((p_0)_{ij}\) of the initial \((N \times N)\) reliability matrix \(P_0\) is equal to 0.5 (given that \(i\) interacts with \(j\)). We then calculate the elements \((p_k)_{ij}\) of the matrix \(P_k\) in \(k\) iterations as:

\[
(p_k)_{ij} = 1 - \prod_{v_l} (1 - (p_{k-1})_{ij} \cdot (p_{k-1})_{jl}) \quad (4)
\]

where we take the product by all \(v_l : (v_l, v_i) \in E, (v_l, v_j) \in E\).
To illustrate the weighting scheme consider the hypothetical network shown in Figure 2

![Figure 2: A simple hypothetical network of 5 proteins and 6 interactions to illustrate how the weight of the edge $e_1$ is determined.](image)

Suppose we would like to determine the weight of the edge $e_1$ (between protein 1 and protein 2). According to Equation 4, the probability that protein 3 and protein 4 do not "support" the edge $e_1$ is $(1-p_{1,3} \cdot p_{2,3})$ and $(1-p_{1,4} \cdot p_{2,4})$, respectively. Thus, the probability that protein 3 and 4 do not "support" the edge $e_1$ is $(1-p_{1,3} \cdot p_{2,3}) \cdot (1-p_{1,4} \cdot p_{2,4})$. Therefore, the probability that protein 1 and protein 2 interact (and "supported" by protein 3 and protein 4) is the complementary probability $1 - [(1-p_{1,3} \cdot p_{2,3}) \cdot (1-p_{1,4} \cdot p_{2,4})]$.

Starting with the initial probability matrix $P_0$ (where $p_{1,1} = 1$, $p_{2,2} = 1$, $p_{3,3} = 1$, $p_{4,4} = 1$, and $p_{5,5} = 1$), the algorithm proceeds as follows:

1. Calculate the degree of each node: $N_i$.
2. Sort the nodes in decreasing order of degree.
3. For each node $i$ with degree $N_i$, calculate the probability that protein 1 and protein 2 interact with protein $i$ using Equation 4.
4. If the probability is greater than the threshold, remove the edge $e_i$.

The algorithm stops when no further edges can be removed. The resulting subgraph is a valid core protein complex.

3. RESULTS AND DISCUSSIONS

The effectiveness of our method is evaluated using two different yeast PPI datasets. The first dataset (PPI-D1) was prepared by Gavin et al. [25] and it contains 869 proteins. The second dataset (PPI-D2) is generated by six different experiments, including interactions characterized by mass spectrometry technique [25–28], and interactions produced using two-hybrid techniques [29,30]. PPI-D2 contains 23,399 interaction pairs and 3,869 proteins.

Three reference sets of protein complexes are used in these experiments. The first set of complexes (Cimplx-D1) is created from MIPS [31] and only complexes that were manually
annotated from DIP interaction data are considered. Following Leung et al. [8], complexes of sizes less than 5 proteins are excluded and therefore, 81 complexes are considered. The second set of complexes (Cmplx-D2) comprises of 162 hand-curated complexes (size no less than 4 proteins) from MIPS [32]. The third dataset (Cmplx-D3) which contains 63 complexes is generated by Aloy et al. [33]. Both datasets Cmplx-D2 and Cmplx-D3 were used by Liu et al. [20] to evaluate the performance of the CMC method. In this case we carry only those complexes with sizes no less than 4.

In the first experimental work, we tried to find the optimal number of iterations $k$ which will lead to the best performance of PE-measure. In Table 1, we listed the effect of varying the number of iterations $k$ and the corresponding complex detection accuracy measure in terms of the number of matched complexes, $Rec_c$ and $Pre_c$. The results show that the performance of PE-measure based on the reference dataset Cmplx-D2 remains approximately the same when $k$ is greater or equal to 12.

Table 1: Optimizing the number of integrations $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>Detected Cmplx</th>
<th>$Rec_c$</th>
<th>$Pre_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>64</td>
<td>0.388</td>
<td>0.197</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>0.400</td>
<td>0.209</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>0.394</td>
<td>0.201</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>0.394</td>
<td>0.194</td>
</tr>
<tr>
<td>10</td>
<td>66</td>
<td>0.400</td>
<td>0.194</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>0.406</td>
<td>0.194</td>
</tr>
<tr>
<td>14</td>
<td>67</td>
<td>0.406</td>
<td>0.194</td>
</tr>
<tr>
<td>16</td>
<td>67</td>
<td>0.406</td>
<td>0.194</td>
</tr>
<tr>
<td>18</td>
<td>67</td>
<td>0.406</td>
<td>0.193</td>
</tr>
<tr>
<td>20</td>
<td>67</td>
<td>0.406</td>
<td>0.193</td>
</tr>
</tbody>
</table>

To analyze the performance of PE-measure and AdjustCD in a noisy interaction dataset, we added different random sets of interaction pairs to Cmplx-D1 and the complex detection accuracy, $Rec_c$ and $Pre_c$ are calculated. In Figures 4, 5 and 6 we show the number of matched complexes detected, $Rec_c$ and $Pre_c$ using PE-measure and AdjustCD in the presence of different random sets of interaction pairs (1,000 to 10,000). The results presented in Figure 4 and 5 show that PE-measure is able to significantly detect more complexes with higher $Rec_c$ than AdjustCD. AdjustCD however, performed better in terms of $Pre_c$ as shown in Figure 6.

Table 2 shows the performance of PE-measure compared to AdjustCD method. In the case of AdjustCD, the iterative scoring parameter $k$ is set 2 as it was shown by Liu et al. [20] that the iterative scoring method reaches the best performance when $k = 2$. On the other hand $k$ is set to 12 when PE-measure is applied.

The results in this case show that PE-measure is able to
Table 2: Comparing the performances of the PE-measure to AdjustCD in terms of the number of matched complexes, $Rec_c$ and $Prec_c$, where $Acc(K, P) \geq 0.5$.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>AdjustCD Matched Cmplx</th>
<th>AdjustCD $Rec_c$</th>
<th>AdjustCD $Prec_c$</th>
<th>PE-Measure Matched Cmplx</th>
<th>PE-Measure $Rec_c$</th>
<th>PE-Measure $Prec_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cmplx-D1</td>
<td>51</td>
<td>0.629</td>
<td>0.371</td>
<td>60</td>
<td>0.741</td>
<td>0.384</td>
</tr>
<tr>
<td>Cmplx-D2</td>
<td>62</td>
<td>0.376</td>
<td>0.213</td>
<td>67</td>
<td>0.406</td>
<td>0.194</td>
</tr>
<tr>
<td>Cmplx-D3</td>
<td>57</td>
<td>0.905</td>
<td>0.263</td>
<td>55</td>
<td>0.873</td>
<td>0.248</td>
</tr>
</tbody>
</table>

Figure 5: $Prec_c$ using PE-measure and AdjustCD in the presence of different random sets of interaction pairs.

Figure 6: $Rec_c$ using PE-measure and AdjustCD in the presence of different random sets of interaction pairs.

detect significant number of matched complexes when applied on Cmplx-D1 and Cmplx-D2 datasets with better $Rec_c$. Note that it is possible to increase the accuracy of the $Prec_c$ by adopting more flexible merging procedures. However, our data analysis (e.g. Cmplx-D2) shows that the number of proteins involved in known complexes is approximately 31% and therefore, many real complexes could still remain unknown. The predicted complexes will allow biologist to investigate and identify more novel complexes. The performance of AdjustCD on Cmplx-D3 is slightly better than PE-measure and the reason is that, Cmplx-D3 is a considerably clean data.

It is appealing to note that the performance of AdjustCD on Cmplx-D2 in conjunction with our WCC has significantly increased Table 3. This is strong evidence in favor of our method.

3.1 Comparing PE-WCC to the existing methods for detecting protein complex

Table 4 compares the performance of our method to CMC, CFinder, MCL and MCode based on different reference datasets. For the CMC, the iterative scoring parameter $k$ is set to 2. For MCL, inflation is set to 1.8. For MCode, the depth is set to 100, node score percentage to 0, and percentage for complex fluffing to 0.2 (as suggested by [34]). For CFinder, we set $k$-clique size to 4. The rest of the parameters were set to their default values.

As shown in Table 4, the proposed method is able to detect more matched complexes (67 matching complexes) than any of the state-of-the-art methods with higher recall.

For generality, the proposed method is also compared to Core, MCL, MCode and CFinder ($k$-clique size of 3 and 4). Table 5 shows the potential of the proposed method to detect significantly more complexes than any of the other four methods.

4. CONCLUSION

In this paper, we have provided a novel method which we call it PE-WCC for detecting protein complexes from PPI network of yeast. We have shown that our approach, which first assesses the quality of the interaction data and then detect the protein complex based on the concept of weighted clustering coefficient is superior to most of the well known methods. All complex predictions made by PE-WCC can be found at http://faculty.uaeu.ac.ae/nzaki/PE-WCC.htm.

An interesting open challenge is to study the incorporation of additional biological knowledge of protein complexes. To this end, a probabilistic calculation of the affinity score between two proteins [35], measuring the structural similarity between interacting proteins [36] and incorporating co-immunoprecipitation data to determine sets of preys that significantly co-associate with the same set of baits [37] could further improve the performance of proposed method.

Acknowledgment

The authors would like to acknowledge the assistance provided by the Emirates Foundation (EF Grant Ref. No. 2010/116), the National Research Foundation (NRF Grant Ref. No. 21T021) and the Research Support and Sponsored Projects Office and the Faculty of Information Technology at the United Arab Emirates University (UAEU).
Table 3: Comparing the performances of the PE-measure to AdjustCD in terms of $Rec$, $Prec$ and number of matched complexes, where $Acc(K, P) \geq 0.5$.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>AdjustCD + CMC</th>
<th>AdjustCD + WCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Matched Cmplx</td>
<td>$Rec$</td>
</tr>
<tr>
<td>Cmplx-D2</td>
<td>56</td>
<td>0.346</td>
</tr>
<tr>
<td>Cmplx-D3</td>
<td>56</td>
<td>0.889</td>
</tr>
</tbody>
</table>

Table 4: Compare PE-WCC to CMC [20], CFinder [7], MCL [4] and MCode [3], where $Acc(K, P) \geq 0.5$.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Detected Cmplx</th>
<th>$Rec$</th>
<th>$Prec$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE-WCC</td>
<td>67</td>
<td>0.406</td>
<td>0.194</td>
</tr>
<tr>
<td>CMC</td>
<td>56</td>
<td>0.346</td>
<td>0.297</td>
</tr>
<tr>
<td>CFinder</td>
<td>46</td>
<td>0.284</td>
<td>0.379</td>
</tr>
<tr>
<td>MCL</td>
<td>51</td>
<td>0.315</td>
<td>0.353</td>
</tr>
<tr>
<td>MCode</td>
<td>39</td>
<td>0.241</td>
<td>0.330</td>
</tr>
</tbody>
</table>

Table 5: Comparison of PE-WCC to Core [8], MCL [4], MCode [3] and CFinder ($k$-clique size = 3 and 4) [7] based on PPI-D1 and Cmplx-D1, where $Acc(K, P) \geq 0.6$.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Detected Complexes</th>
<th>$Rec$</th>
<th>$Prec$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE-WCC</td>
<td>43</td>
<td>0.740</td>
<td>0.384</td>
</tr>
<tr>
<td>Core</td>
<td>35</td>
<td>0.432</td>
<td>0.151</td>
</tr>
<tr>
<td>MCL</td>
<td>32</td>
<td>0.395</td>
<td>0.138</td>
</tr>
<tr>
<td>MCode</td>
<td>23</td>
<td>0.284</td>
<td>0.219</td>
</tr>
<tr>
<td>CFinder ($k$-clique size = 3)</td>
<td>22</td>
<td>0.272</td>
<td>0.225</td>
</tr>
<tr>
<td>CFinder ($k$-clique size = 4)</td>
<td>25</td>
<td>0.309</td>
<td>0.352</td>
</tr>
</tbody>
</table>

5. REFERENCES


