Supplementary Material

Identifying protein complexes based on an edge weighted algorithm and core-attachment structure

Rongquan Wang, Guixia Liu and Caixia Wang

# Testing EWCA against other algorithms

## 2.1 Data sets

We concentrate our study on yeast since it is a well studied model for mammalian. Two experimental yeast PPI data sets are used to test the performance. We refer to DIP [2] and BioGRID [3] data sets. As BioGRID provide weights for only a low proportion of the interactions, we treat it as unweighted in a similar manner to that of [1]. For the DIP network, self-interactions, redundant interactions and interactions involving proteins of which the systematic names are not available are filtered out. For simplicity, we just analyze its the largest connected component. Some statistics of the two data sets are presented in Table 1.

We use the CYC2008 [4] and NewMIPS [5-7] benchmarks as the gold standards of yeast protein complexes. The first set is the CYC2008 catalogue of manually curated protein complexes from Wodak’s lab [4]. The second set used in [8,9] (denoted as “NewMIPS”) was derived from three sources: MIPS [5], Aloy et al. [6] and the Gene Ontology (GO) annotations in the SGD database [7]. Complexes smaller than 3 proteins were filtered out from both benchmarks. After this step, there are 236 complexes left in the CYC2008 and 328 complexes in NewMIPS. For the CYC2008 benchmark, the largest complex is the cytoplasmic ribosomal large subuit with 81 proteins and the average size of the complexes is 6.68 proteins. For detail of the construction of the benchmark, please refer to [10]. The general properties of the reference sets are listed in Table 2.

## 2.2 Evaluation methods

To assess the performance of a considered approach, we need a quantitative criterion to evaluate how a set of predicted complexes matches with a set of reference complexes. Due to the fact that the gold standard complexes (and the predicted complexes if the used algorithm handle overlaps) overlap with each other, a gold standard complex can have a (partial) match with more than one predicted complex and vice versa [1]. It is therefore difficult to find a universal evaluation metric that can work well on this task. In this paper, we use five kinds of measures to evaluate the predicted complexes by comparing them with the reference complexes: recall [11-13], precision [11-13], F-measure [14,15], maximum matching ratio (MMR) [1] and coverage rate (CR) [17,18]. The nine metrics assess the performance from different perspectives and have complementary strength.

The first measure are recall, precision and F-measure, which generally assess not overlapping clustering algorithms. Recall is the fraction of the known complexes that match at least one one predicted complex among all known complexes. Precision is the fraction of the number of the predicted complexes that match at least one known complex over the total number of all predicted complexes. The F-measure is the harmonic mean of recall and precision and shows the overall performance of a predicted result.

The second measure is the MMR, which is strongly recommended by the authors, offers a natural way to compare a set of predicted complexes and a set of reference complexes [1]. As discussed by the authors, it penalizes an approach which tends to split a reference complex into more than one part in the predicted complexes. Owing to focusing mainly on test how well the gold standard is matched by the predicted complexes, it does not take the false positive predictions into account. As a result, other methods which quantify the functional homogeneity of predicted complexes are advised to complement the maximum matching ratio[1].

The last measure is coverage rate which assesses how many proteins in the real complexes can be covered by the predicted complexes [17,18].

## 2.3 General considerations

Evaluating the performance of clustering algorithms on graphs is a trick problem. Therefore, careful counter-measures can and should be taken to avoid the typical biases in the evaluation of a novel method, in particular the over-optimization of algorithm parameters to a given dataset or a given quality score. In short, we have decided on the following:

We have tested each of the algorithms on two different datasets: DIP [2] and BioGRID [3].

1. We have used five quality score to assess the performance of each algorithm from different perspectives and have complementary strength..
2. We have also used two different gold standards: NewMIPS [5-7] and CYC2008 [4]. Note that since the two gold standards are not entirely consistent with respect to the membership of some proteins in some complexes, we decided to test these two gold standards separately.
3. For each algorithm, the final results were obtained after having optimized the algorithm parameters to yield the best possible results as measured by F-measure on the gold standard that was being used in the benchmark (either NewMIPS or CYC2008).

## 2.4 Common settings for all the algorithms

We compared the performance of EWCA to a representative set of other approaches: MCL [26], CFinder [19], Core [27], IPCA [31], COACH [24], SPICi [15], ClusterONE [1], PEWCC [32], GMFTP [33], DPC-NADPIN[34], Prorank+ [35] and CMC [36]. Some of these algorithms supported the use of edge weights (MCL, ClusterONE, CMC), and some could handle overlapping clusterings (IPCA, CFinder, ClusterONE, PEWCC , GMFTP, COACH, DPC-NADPIN, Prorank+ ).

Predicted complexes containing less than three proteins were excluded from the results unless the authors of the original algorithm suggested different size limits. In such cases, the new size limits are always mentioned explicitly in the upcoming sections.

# Parameter settings of compared algorithms

In this paper, in order to evaluate the performance of CP in detecting protein complexes, we compare it with twelve competitive methods: MCL [26], CFinder [19], Core [27], IPCA [31], COACH [24], SPICi [15], ClusterONE [1], PEWCC [32], GMFTP [33], DPC-NADPIN[34], Prorank+ [35] and CMC [36]. Table S0 lists the websites where we download the softwares of these algorithms, the version numbers of these softwares and several indications about whether these algorithms could be applied to weighted PPI networks or handle overlaps. Before describing the parameter settings for each algorithm, we declare several general consideration first. Since the performance of each algorithm depends on the choice of its inherent parameters and the data set under consideration, for all the considered algorithms, we optimize the parameters that yield the best results in a similar way to that of [1].

**Table S0**: Characteristics of the compared algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Downloading website | Version | Weights supported | Overlap supported |
| MCL | http://micans.org/mcl/ | 14-137 | yes | no |
| CFinder | http://cfinder.org/ | 2.0.6 | no | yes |
| Core | http://alse.cs.hku.hk/complexes/ | 0.10 | no | yes |
| IPCA | http://  netlab.csu.edu.cn/bioinformatics/limin/IPCA | 2018.11.8 | no | yes |
| COACH | http://www1.i2r.a-star.edu.sg/~xlli/coach.zip | 2018-6-12 | no | yes |
| SPICi | http://compbio.cs.princeton.edu/spici/ | 2018-5-24 | yes | no |
| ClusterONE | http://apps.cytoscape.org/apps/clusterone | 0.94 | yes | yes |
| PEWCC | http://faculty.uaeu.ac.ae/nzaki/Research.htm | 2018-5-27 | yes | yes |
| GMFTP | http://mail.sysu.edu.cn/home/stsddq@mail.sysu.edu.cn/dai/others/GMFTP.zip | 1.0 | no | yes |
| DPC-NADPIN | https://github.com/ccnuyili/DPC-NADPIN | 1.0 | no | yes |
| ProRank+ | http://faculty.uaeu.ac.ae/nzaki/Research.htm | 2017-5-27 | yes | yes |
| CMC | http://www.comp.nus.edu.sg/~wongls  /projects/complexprediction/CMC-26may09/ | 2.0 | yes | yes |

**3.1 MCL**

Markov Clustering Algorithm (MCL) [26] is a competing protein complex detection algorithm and has been implemented in different languages, such as JAVA, R and C. The key parameter of MCL is inflation, which tunes the granularity of clustering. Here, we try different values of inflation, ranges from 1.2 to 5.0 with 0.2 increment. MCL can handle weighted networks, thus we also list its results on the weighted PPI networks constructed. The optimal value of inflation for each PPI network is shown in Table S1.

**Table S1**: The value of inflation selected for MCL.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | Inflation | 2.0 | 4.0 |
| CYC2008 | Inflation | 2.0 | 3.2 |

**3.2 CFinder**

CFinder [19] is a fast program that implements the CPM (Clique Percolation Method) for graph clustering. The algorithm first finds all k-cliques in the input network, then associates two k-cliques if they share a common (k −1)-clique. CPM detects overlapping clusters by finding k-clique percolation communities. Therefore, a key parameter of CFinder is the size of k-clique. In this paper, for each PPI network, we test CFinder with k-clique size from 3 to 10, step size by 1. Note that the produced distinct clusters may share nodes, however the intersection of two clusters may contain (k −2)-cliques but never a (k −1)-clique. For BioGRID network, since CFinder did not give any result within 48 hours, we set an optional time limit (10 seconds) for the time to spend on each node of the network, such that it can analyze the network efficiently. The original CPM method can only handle unweighted networks. Even though Farkas et al have proposed a weighted extension of CFinder [20], the computational cost of the new variant is more prohibitive and can not analyze BioGRID network in 48 hours. Therefore we just list the results on the unweighted PPI networks.

Table S2 lists the optimal values of parameter k for each PPI network.

**Table S2**: Parameters selected for CFinder.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | k | 3 | 3 |
| CYC2008 | k | 3 | 3 |

**3.3 Core**

Core [27] addresses the problem of predicting protein complexes from protein-protein interaction (PPI) network of one species using a computational approach. Most of the previous methods rely on the assumption that proteins within the same complex would have relatively more interactions. This translates into dense subgraphs in the PPI network. However, all existing software tools have limited success. Recently, Gavin et al [29] provided a detailed study on the organization of protein complexes and suggested that a complex consists of two parts: a core and an attachment. Based on this core-attachment concept, we developed a novel approach to identify complexes from PPI network by identifying their cores and attachments separately.

The procedure takes in input the PPIN file, and has no other parameters.

## 3.4 DPClus

The DPClus algorithm [11] is a approach for detecting protein complexes based on topological structure,which is a combination of subgraph cluster property and subgraph density. It consists of three stages: input initialization, Termination check, Seed selection, Cluster formation and Output and update. Among its inherent parameters, the -Minimun cluster size parameter sets, the density of cluster parameter and the cluster property. In this paper, we set the minimun size of protein complexes identification for all algorithms is 3 to keep consistent. As suggested by the authors, we do not tune the parameters for a particular network. Thus, we use the suggestion parameters in the DPClus\_manual.doc.

The value of all parameters for each PPI network is shown in Table S3.

**Table S3**: Parameters selected for DPClus.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | -Minimun cluster [size] | 3 | 3 |
| -CPin[value] | 0.5 | 0.5 |
| -Density[value] | 0.6 | 0.6 |
| CYC2008 | -Minimun cluster [size] | 3 | 3 |
| -CPin[value] | 0.5 | 0.5 |
| -Density[value] | 0.6 | 0.6 |

**3.5 COACH**

COACH [24] is an algorithm whose main strategy is a greedy local refinement of the core graph of a graph G = (V,E). Formally a core graph CG(G) is defined as follows. Take any vertex v  V and a vertex in its 1-neighborhood u  N1 (v). If the degree of u in the induced subgraph G[N1 [v]] is larger then the average degree in G[N1[v]], then u is labeled as a core vertex. The core graph CG(G) is the subgraph induced in G by the collection of all the core vertices of G. Even if the name is similar, the notion of a core vertex in ([24]) is quite different form the notion of core number/core decomposition of [25] which is based instead on the minimum degree of an induced subgraph.

As suggested by the authors, we do not tune the parameters for a particular network. Thus, we use the default settings of parameters in the software.

**3.6 SPICi**

SPICi [15] is a computationally efficient local network clustering algorithm for large biological networks, which can be applied on PPI networks for complex detection. SPICi can handle weighted networks, so we apply it on the weighted and unweighted PPI networks to test its performance. SPICi has two parameters: the density threshold and the support threshold. Here, we try different values of density threshold, ranges from 0.1 to 1.0 with 0.1 increment. For support threshold, the optimal preference value is determined by trying different values (ranges from 0.1 to 1.0 with 0.1 increment) and setting on the preference value that results in the best quality score.

Table S4 lists the optimal value of density parameter for each PPI networks.

**Table S4**: Parameters selected for SPICi.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | Minimum cluster density | 0.4 | 0.7 |
| Minimum support threshold | 0.5 | 0.2 |
| Minimum cluster size | 3 | 3 |
| Graph mode  0: sparse graph  1:dense graph  2:large sparse graph | 0 | 0 |
| CYC2008 | Minimum cluster density | 0.4 | 0.7 |
| Minimum support threshold | 0.5 | 0.2 |
| Minimum cluster size | 3 | 3 |
| Graph mode  0: sparse graph  1:dense graph  2:large sparse graph | 0 | 0 |

**3.7 ClusterONE**

ClusterONE is recently proposed by Nepusz et al. [21] to detect overlapping protein complexes in PPI networks based on overlapping neighborhood expansion. ClusterONE can deal with weighted and unweighted networks, therefore, besides the original network, we also test its performance on the weighted networks which are constructed by assigning weights to interactions according to the GO annotations.

All parameters are set as suggested by author, and we therefore used the default parameters of ClusterONE implementation. These are: density threshold set to auto for weighted networks. The merging threshold is set to 0.8 and the penalty value of each node is 2. The parameters of ClusterONE for each PPI network are listed in TableS5.

**Table S5**: Parameters selected for ClusterONE.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | Minimum size | 3 | 3 |
| Minimum density | Auto | Auto |
| Edge weights | yes | yes |
| Node penalty | 2 | 2 |
| Overlap threshold | 0.8 | 0.8 |
| CYC2008 | Minimum size | 3 | 3 |
| Minimum density | Auto | Auto |
| Edge weights | yes | yes |
| Node penalty | 2 | 2 |
| Overlap threshold | 0.8 | 0.8 |

**3.8 PEWCC**

The PEWCC algorithm [32] has three main tunable parameters. One of them is that “-r” parameter which is used to remove edges relaibility score less than this value, and according to author’s suggestion, it should be set to default is 0.1. The other two are the “Join parameter” and the “overlap threshold”. We tried all the possible combinations of the following parameter values, and the one resulting having the best F-measure was kept. The optimal values of the parameters of PEWCC for each PPI network are listed in TableS6.

**Table S6**: Parameters selected for PEWCC.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | Join parameter | 0.5 | 0.3 |
| Overlap threshold | 0.8 | 0.8 |
| CYC2008 | Join parameter | 0.5 | 0.3 |
| Overlap threshold | 0.8 | 0.8 |

**3.9 GMFTP**

GMFTP [33] develops a generative model with Functional and Topological properties to simulate the generative processes of topological and biological information and clusters that maximize the likelihood of generating the given PIN are considered as protein complexes. The model provides a mechanism for capturing PPI network and the functional patterns of proteins. By combining the functional and topological properties, they transform the problem of identifying protein complexes as that of mining in the functional profile. This model could deal with overlapping among complexes. For a detailed explanation of the algorithm, please see Ref [33].

As suggested by the authors, we do not tune the parameters for a particular network. Thus, we use the default settings of parameters in the software.

**3.10 DPC-NADPIN**

DPC-NADPIN (Discovering Protein Complexes based on Neighbor Affinity and Dynamic Protein Interaction Network) [34] is a novel algorithm to identify temporal protein complexes from the time course protein interaction networks. Its main idea is that the tighter a protein’s neighbors inside a module connect, the greater the possibility that the protein belongs to the module, DPC-NADPIN first chooses each of proteins with high clustering coefficient and its neighbors to consolidate into an initial cluster, and then the initial cluster becomes a protein complex by appending its neighbors according to the relationship between the affinity among neighbors inside the cluster and that outside the cluster. It has better performance on discovering protein complexes that some state-of-the-art algorithms. Meanwhile, it obtains many protein complexes with strong biological significance, which provide helpful biological knowledge to the related researchers.

Because all compared methods are tested based on static PPI networks, thus, for a fair. In this paper, we only test DPC-NADPIN in the whole static PPI network and not to identify in different time point subnetworks.The procedure takes in input the PPIN file, and has no other parameters. Thus, we don’t tune any parameters for a particular network, we use the default settings of parameters in the algorithm.

**3.11 ProRank+**

Prorank+ [35] use several edge pruning and filtering steps before applying a ranking step inspired by the well known Page-Rank algorithm in order to highlight candidate protein seeds. Afterwards, a complex is identified starting from a seed by using a spoke model. Finally highly overlapping complexes are removed in alternated duplicate-detection and cohesiveness-based-merge phases.

The procedure takes in input the PPIN file, and has no other parameters.

**3.12 CMC**

The CMC algorithm [36] is based on an iterative scoring algorithm that assesses the probability of whether two given proteins are in the same complex, followed by a maximal clique finding process. Highly overlapping cliques are then merged in order to achieve the final set of complexes. The algorithm is primarily governed by the overlap threshold which determines when should two cliques be considered highly overlapping, and the merge threshold, which determines what to do with two highly overlapping cliques: they will be merged if the part of the network between the two complexes is denser than the merge threshold, otherwise the smaller clique will be discarded.

The range of both parameters is between zero and one, although low overlap thresholds do not make sense as they would result in only a few giant complexes. Similarly, high overlap thresholds would result in a very large number of redundant complexes, as almost none of them would be allowed to merge with others. Therefore, the tested range of the overlap threshold was limited to real values between 0.1 and 1.0, sampled with a step size of 0.1. The merge threshold was tested on uniformly sampled real values between 0.1 and 1.0 with a step size of 0.1.

In our benchmarks, we used the original implementation of the CMC software (version 2), downloaded from http://www1.comp.nus.edu.sg/ ~ wongls/projects/complexprediction/CMC-26may09/. According to the suggestions of the authors of the algorithm [22]. Here, the size of size limit is 3. N/A for the BioGRID dataset indicates that the algorithm produced a prohibitively large number of clusters (more than 6000) for all parameter settings we have tried. The optimal value of extend thres for each PPI network is shown in Table S7.

**Table S7**: Parameters selected for CMC.

|  |  |  |  |
| --- | --- | --- | --- |
| Gold Standard | Parameter | DIP | BioGRID |
| NewMIPS | Overlap threshold | 0.2 | 0.2 |
| Merge threshold | 0.1 | 0.7 |
| CYC2008 | Overlap threshold | 0.3 | 0.2 |
| Merge threshold | 0.1 | 0.8 |

References

1. Nepusz,T.Y.et al (2007) Detecting overlapping protein complexes in protein-protein interaction networks, Nature Methods, 9, 471-472.
2. Xenarios,I.S.et al (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Research, 30,303.
3. Stark,C.B.etal(2005)BioGRID:a general repository for interaction datasets, Nucleic Acids Research, 34(Suppl. 1), 535-539.
4. Pu,S.Wetal(2009)Up-to-date catalogues of yeast protein complexes.,Nucleic Acids Research, 37, 825-31.
5. Mewes,H.A et al (2006) MIPS: analysis and annotation of proteins from whole genomes., Nucleic Acids Research, 34, 169-72.
6. Aloy,P.B et al (2004) Structure-based assembly of protein complexes in yeast., Science, 303, 2026.
7. Dwight,S.H et al (2002) Structure-based assembly of protein complexes in yeast., Nucleic Acids Research, 30, 69-72.
8. Friedel CC, Krumsiek J, Zimmer R: Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB) 2008, 3-16.
9. Wu M, Li XL, Kwoh CK: Algorithms for Detecting Protein Complexes in PPI Networks: An Evaluation Study. Proceedings of Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008), Australia, Oct, 15-17, 2008 2008, 135-146.
10. Wong L K, Li X L, Wu M, et al. PLW: Probabilistic Local Walks for detecting protein complexes from protein interaction networks[J]. Bmc Genomics, 2013, 14(5):1-15.
11. Altaf-Ul-Amin M, Shinbo Y, Mihara K, et al. Development and implementation of an algorithm for detection of protein complexes in large interaction networks[J]. BMC bioinformatics, 2006, 7(1): 207.
12. Li,X.L. et al. (2007), Computational Systems Bioinformatics, Vol. 6, pp. 157-168.
13. Wu,M.Li.et al (2009) A core-attachment based method to detect protein complexes in PPI networks, BMC Bioinformatics, 10, 1-16.
14. Cho,H.R.etal(2007) Semantic integration to identify overlapping functional modules in protein interaction networks, BMC Bioinformatics, 8, 1-13.
15. Li,X.W.etal (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey, BMC Genomics, 11(Suppl. 1), 1-19.
16. Ji,J.Z.et al (2007) Survey: Functional Module Detection from Protein-Protein Interaction Networks, IEEE Transactions on Knowledge Data Engineering, 26,261-277.
17. Brohee,S.V.et al (2006) Evaluation of clustering algorithms for protein-protein interaction networks., BMC Bioinformatics, 7, 488.
18. Friedel,C.C. (2008) Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast., Research in Computational Molecular Biology, Springer Berlin Heidelberg, Berlin, Heidelberg, pp.3–16.
19. Adamcsek,B., et al (2006) CFinder:locating cliques and overlapping modules in biological networks, Bioinformatics, 22, 1021-1023.
20. Farkas, I.,´Abel, D., Palla, G. & Vicsek, T. Weighted network modules. New J. Phys.9, 180 (2007).
21. King,A.D., et al (2004) Protein complex prediction via cost-based clustering.,Bioinformatics, 20, 3013-3020.
22. Liu,G., et al (2009) Complex discovery from weighted PPI networks.,Bioinformatics, 25, 1891-1897.
23. Hanna, E. M. and Zaki, N. (2014) Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure., BMC Bioinformatics, 15, 204.
24. Wu,M.Li.et al (2009) A core-attachment based method to detect protein complexes in PPI networks, BMC Bioinformatics, 10, 1-16.
25. Seidman, S.B.: Network structure and minimum degree. Social Networks 5(3), 269–287 (1983).doi:10.1016/0378-8733(83)90028-X
26. VanDongen, S(2000)Graph Clustering by Flow Simulation.PhD Thesis, University of Utrecht, The Netherlands.
27. Leung, H. C., et al (2009) Predicting protein complexes from PPI data: a core-attachment approach, Journal of Computational Biology A Journal of Computational Molecular Cell Biology, 16, 133-144.
28. Becker, E., et al (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network, Bioinformatics, 28, 84-90.
29. Gavin,A.A., et al (2006) Proteome survey reveals modularity of the yeast cell machinery., Nature, 440, 631-6.
30. Shu,Tadaka and Kinoshita,Kengo(2016) NCMine:Core-peripheral based functional module detection using near-clique mining, Bioinformatics, 32, 3454-3460.
31. Min Li, Jianer Chen, Jianxin Wang, et al. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures[J]. BMC Bioinformatics, 2008, 9(1):398-398.
32. Zaki, Nazar, Berengueres, et al. Protein complex detection using interaction reliability assessment and;weighted clustering coefficient[J]. BMC Bioinformatics,14,1(2013-05-20), 2013, 14(1):163-163.
33. Zhang X F, Dai D Q, Ouyang L, et al. Detecting overlapping protein complexes based on a generative model with functional and topological properties.[J]. Bmc Bioinformatics, 2014, 15(1):186.
34. Shen X, Yi L, Jiang X, et al. Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network[J]. Methods, 2016, 110:90-96.
35. Hanna, E. M. and Zaki, N. (2014) Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure., BMC Bioinformatics, 15, 204.
36. Liu,G., et al (2009) Complex discovery from weighted PPI networks.,Bioinformatics, 25, 1891-1897.