Incorporating Semantic Similarity into Clustering Process for Identifying Protein Complexes from Affinity Purification/Mass Spectrometry Data

BINGJING CAI
School of Computing and Mathematics
University of Ulster
N. Ireland, UK
cai-b@email.ulster.ac.uk

HAIYING WANG
School of Computing and Mathematics
University of Ulster
N. Ireland, UK
hy.wang@ulster.ac.uk

HUIRU ZHENG
School of Computing and Mathematics
University of Ulster
N. Ireland, UK
h.zheng@ulster.ac.uk

HUI WANG
School of Computing and Mathematics
University of Ulster
N. Ireland, UK
h.wang@ulster.ac.uk

Abstract — **This paper presents a framework for incorporating semantic similarities in the detection of protein complexes from Affinity Purification/Mass Spectrometry (AP-MS) data. AP-MS data is modeled as a bipartite network, where one set of nodes consist of bait proteins and the other set are prey proteins. Pair-wise similarities of bait proteins are computed by combining similarities based on topological features and functional semantic similarities. A hierarchical clustering algorithm is then applied to obtain 'seed clusters' consisting of bait proteins. Starting from these 'seed' clusters, an expansion process is developed to recruit prey proteins which are significantly associated with bait proteins, to produce final sets of identified protein complexes. In the application to real AP-MS datasets, we validate biological significance of predicted protein complexes by using curated protein complexes. Six statistical metrics have been applied. Results show that by integrating semantic similarities into the clustering process, the accuracy of identifying complexes has been greatly improved. Meanwhile, clustering results obtained by the proposed framework are better than those from several existent clustering methods.**

Keywords - Protein-protein interactions; Affinity purification/mass spectrometry (AP-MS); Protein compelxes; Gene Ontology; Semantic Similarity

I. INTRODUCTION

With development of experimental and computational approaches, huge amount of protein-protein interactions (PPIs) have been detected [1-5]. How to accurately identify protein complexes from such large-scale networks of PPIs becomes a challenge. PPIs have been characterized as two categories based on which experimental approach was used: yeast-two-hybrid (Y2H) identifies physically pair-wise PPI; while affinity-purification/mass-spectrometry (AP-MS) detects co-membership of complexes by purifying proteins (called prey) that are associated with tagged proteins (called bait) [3-5].

In decades, most researches have been carried out to detect protein complexes by treating PPIs as binary. In 2003, Bader and Hogue [6] proposed the 'Spoke' model and the 'Matrix' model for PPI networks, and developed the Molecular Complex Detection (MCODE) algorithm for deriving protein complexes using these models. MCODE Markov Clustering Algorithm (MCL) [7] is a well-known clustering method, simulating random walks on the graph. The process converges towards a partition of the graph, with a set of regions of high density. CFinder [8] exploits local topological structure of nodes, exploring clusters consisting of numbers of k-cliques where two adjacent k-cliques share k-1 nodes, and then yields overlapping clusters.

However, since the AP-MS technique detects co-membership between bait proteins and prey proteins instead of directly pair-wise interactions, AP-MS data have non-binary nature. CODEC [9], proposed in 2011, takes advantage of its non-binary nature by constructing AP-MS data as a bipartite graph, aimed to detect complexes as dense bipartite sub-graphs.

Besides physically interacting relationship between two proteins, semantic similarity describes another type of between-protein relationship by estimates of ontology-based functional similarity [10-11]. Semantic similarity measures closeness in meaning between two proteins, given two ontology terms or two sets of terms annotating these two proteins. The Gene Ontology (GO) [12] has become the *de facto* standard for annotating gene products [10]. Results from a study [13] demonstrated that there is a significant correlation between semantic similarity between pair-wise proteins and their co-complex membership. Therefore, we assume that incorporating semantic similarity into clustering process will improve the accuracy of identifying protein complexes.

This paper presents a framework for detecting protein complexes from AP-MS data, based on a previously proposed method [14]. The algorithm intends to detect groups of prey proteins that are significantly co-associated with bait proteins. We first construct AP-MS data as a bipartite graph, where one set of nodes consist of bait proteins and the other set are prey proteins. We calculate similarities between pair-wise bait proteins by combining topology-based similarity and semantic similarity. Then we apply hierarchical clustering to group bait proteins based on their similarities, to produce a set of 'seed' clusters composed of bait proteins. Starting from these 'seed' clusters, an expansion process in a greedy fashion is developed to recruit prey proteins which are significantly associated. Then, final sets of identified protein complexes are output.

The organization of the paper is shown below. Section II introduces the methodology of our proposed framework followed by a description of datasets under study in section III. Statistical metrics used to assess the quality of clustering are introduced in section IV. In section V, experimental results are presented and discussed. We validate biological

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

significance of predicted protein complexes by using curated complexes. Finally, the conclusion and future work is presented in Section VI.

II. Methodology

Our proposed framework is developed from previously proposed method [14]. The method lies on the assumption that, as AP-MS experiment directly detects complex membership by purifying prey proteins which are co-associated with tagged bait proteins [3-4], thus, institutively, a protein complex is composed of a set of bait proteins along with a set of prey proteins that are significantly associated with the same set of bait proteins. Our previously proposed method [14] calculates pair-wise similarities between bait proteins based on number of common neighbors that two bait proteins share. In this paper, we incorporate GO-driven semantic similarities with the topology-based similarity.

A. Calculate pair-wise similarities between bait proteins

In our proposed framework, we calculate pair-wise similarities between bait proteins by combining topology-based similarity with semantic similarity.

a) Topology-based similarity

We calculate topological similarity between pairs of bait proteins based on number of common neighbors they share. Let b_1 and b_2 be the two baits, $N(b_1)$ and $N(b_2)$ denote the set of neighbours of b_1 and b_2, respectively. Then the similarity $t\_sim(b_1, b_2)$ is calculated by Equation (1):

$$t\_sim(b_1, b_2) = \frac{|N(b_1) \cap N(b_2)|}{|N(b_1) \cup N(b_2)|} \tag{1}$$

The similarity is actually generalized from the notion of Jaccard Similarity Coefficient [15].

b) Semantic similarity

The GO [12] has three ontologies, *Molecular Function* (*MF*), *Biological Process* (*BP*), and *Cellular Component* (*CC*). MF refers to information on what a gene product does. BP is related to a biological objective to which a gene product contributes. CC refers to the cellular location of the gene product, including cellular structures and complexes. The GO terms under each ontology are organized hierarchically, and their relationships are represented by *Directed acyclic graphs* (DAGs). The reader can refer to [15] for more details.

The basic idea for calculating similarity between gene products is to calculate similarities between all terms that are used to annotate gene products. We employ a tool, called *seGOsa* [16], which is a user-friendly cross-platform system to support large-scale assessment of GO-driven similarity among gene products. It implements three semantic similarity measures to measure between-term similarity within each of the GO hierarchies (MF, *BP* and *CC*). In this paper, we use BP semantic similarity as the first instance. The semantic similarity, *s_sim(b₁,b₂),* has two numeric values, that is:

$$s\_sim(b_1, b_2) = \begin{cases} -1, & if\ s\_sim(b_1, b_2)\ not\ \ identified \\ & simValue, \quad otehrwise \end{cases} \tag{2}$$

Here, -1 indicates that one or both proteins have not been annotated using GO terms. *simValue* falls between [0,1], representing the closeness between the pair of proteins based on the information derived from GO annotations. IEA annotations were excluded in the calculation due to their lack of reliability.

c) Combination of two similarities

We incorporate the topology-based similarity and the semantic similarity to generate new pair-wise similarity measures for bait proteins. We adopt a simple way to combine the two different similarities as first trial by calculating the arithmetic average of topology-based similarity and semantic similarity.

$$sim(b_1, b_2) = \begin{cases} \frac{t\_sim(b_1, b_2) + s\_sim(b_1, b_2)}{2}, If\ sim(b_1, b_2)! = -1 \\ t\_sim(b_1, b_2), \text{If}\ s\_sim(b_1, b_2) = -1 \end{cases} \tag{3}$$

After calculating the combined similarities in (3), a network composed of similarities between pair-wise bait proteins could be obtained accordingly.

B. Proposed framework

The proposed framework in the paper is based on our previously proposed method [14]. The clustering process consists of the following five steps.

a) *Model the input AP-MS network as a bipartite graph;*

b) *Calculate pair-wise similarities between bait proteins;*

c) *Cluster bait proteins to obtain preliminary seed clusters;*

d) *Expand process to form complete clusters;*

e) *Filter clusters and output final set of clusters.*

Using the similarities calculated above as a metric, we apply the Agglomerative Hierarchical Cluster algorithm to cluster the bait proteins and obtain a set of "seed" clusters. Complete clusters are formed from expansion from these seed clusters and the expansion process is in a greedy fashion. We calculate the overlap rate between two clusters, that is, $|C_1 \cap C_2|/(|C_1 \cup C_2|)$, where $|C|$ is the size of the cluster. If the overlapping rate is above a given threshold, we merge the two clusters. In our algorithm, we use 0.2 as the threshold value which is obtained from empirical studies. For the detail description, please refer to [14].

III. Datasets Under Study

We used two well-studied AP-MS datasets for empirical study of the proposed framework in the paper: a) the dataset obtained by Gavin et al. [4] in 2006, which contains 2671 proteins in total with 1993 tagged as bait proteins; b) the dataset obtained by Krogan et al. [17], which has totally 5215 proteins with 2231 bait proteins. Note, the network downloaded from the study [17] contains 94 prey proteins which were suspected as non-specific contaminants. Thus, these 94 proteins have been excluded from the raw dataset.

We downloaded a set of hand-curated complexes derived from the Wodak lab CYC2008 catalog [18], which consists of 408 complexes.

IV. Quality Assessment

We use six statistical metrics in this paper to measure the quality of a clustering. The six metrics have been suggested by Brohée and Helden [19], which are *sensitivity*, *predictive precision value (PPV)*, *accuracy*, *cluster-wise separation(Cl-Sep)* , *complex-wise separation (Co-Sep)* and *separation*.

These assessment metrics measure the overlap degree between predicted clusters and benchmark complexes. The value of each measure falls into the interval between 0 and 1. The higher the value, the better the result is. Please refer to [19] for more details. We adopt the same strategy to pre-process the set of predicted clusters and benchmark complexes as in [14].

V. Experimental Results and Discussion

In order to gauge the effect after incorporating the semantic similarity in clustering process, we firstly compare the proposed framework against the method without incorporate any semantic similarity. Then, we evaluate the performance of the proposed framework against several existing clustering methods.

A. Parameter selection

We choose un-weighted average linkage criteria in the hierarchical clustering in the following experiments. According to experimental results, the clustering is better when selecting cut-off value as 0.3 and 0.25 on Gavin_2006 and Krogan_2006 networks, respectively.

We choose the set of parameters of MCL and MCODE recommended by Broheé and Helden [19]. Specifically, we use the inflation rate of 1.8 for MCL. For MCODE, we use the following parameter settings: the depth equals to 100, the node score percentage is 0, Haircut is set to TRUE, Fluff is set to FALSE and the percentage for complex fluffing is 0.2. As for the CODEC algorithm, there are two schemes, CODEC-w0 and CODEC-w1. The result of CFinder achieves better accuracy when *k* value is 5. We compare our method to both schemes of CODEC.

B. The effect of incorporating semantic similarity on detecting protein complexes

Table I shows evaluation results of predicted protein complexes on the two real PPI networks comparing with the set of benchmark complexes, CYC_2008. The column under "Topology-based similarity" shows evaluation figures of clustering results from the method using topological similarity only, and the column under "Combination similarity" presents results for the proposed framework in this paper.

TABLE I. Performance Comparison On Gavin_2006 And Krogan_2006 With Cyc-2008

Evaluation metrics	Gavin_2006		Krogan_2006	
	Topology-based similarity	***Combination similarity***	***Topology-based similarity***	***Combination similarity***
Sensitivity	0.461	**0.469**	0.300	**0.419**
PPV	0.711	**0.727**	0.550	**0.595**
Accuracy	0.573	**0.584**	0.406	**0.499**
Co-Sep	0.307	**0.334**	0.134	**0.221**
Cl-Sep	0.819	**0.871**	0.745	**0.846**
Separation	0.502	**0.540**	0.316	**0.432**

On Gavin_2006 network, the proposed framework achieves marginally higher accuracy and homogeneity value. On the Krogan_2006 network, the proposed framework performs significantly better compared to the method without semantic similarity with an increase in accuracy and homogeneity of 23% and 36.7%, respectively. Hence, The consistently better performance demonstrates that, incorporating semantic similarity helps improving the quality of predicting complexes.

C. Comparison to other clustering methods

We compare the performance of the proposed framework against that of four other clustering methods, including MCODE [6], MCL [7], CFinder [8], and CODEC [9], on these two PPI networks.

Sensitivity shows the average maximal fraction of proteins in complexes found by best-matching clusters. According to Tables II and III, we observe that, MCL has higher sensitivity than other algorithms do, which indicates that a higher fraction of proteins in the benchmark complexes are revealed by best-matching clusters yielded by MCL. The proposed framework achieves significantly better PPV than other methods, which reflects the proposed framework is more reliable in predicting proteins belonging to the "right" complexes.

Accuracy reflects the trade-off between sensitivity and PPV, thus it indicates the general quality of a clustering. As seen from Tables II and III, the proposed framework achieves considerably high accuracy against other methods, which reflects better quality of predicted clusters generated by the proposed framework.

TABLE II. Performance Comparison Among Different Clustering Methods On Gavin_2006 With Cyc-2008

Evaluation metrics	MCL	MCODE	CFINDER	CODEC-w0	CODEC-w1	Proposed framework
Sensitivity	**0.721**	0.338	0.390	0.584	0.582	0.469
PPV	0.296	0.342	0.365	0.511	0.546	**0.727**
Accuracy	0.462	0.340	0.377	0.546	0.564	**0.584**
Co-Sep	0.156	0.123	0.087	0.234	0.272	**0.334**
Cl-Sep	**0.875**	0.748	0.613	0.086	0.107	0.871
Separation	0.369	0.303	0.231	0.141	0.171	**0.540**

TABLE III. Performance Comparison Among Different Clustering Methods On Krogan_2006 With Cyc-2008

Evaluation metrics	MCL	MCODE	CFINDER	CODEC-w0	CODEC-w1	Proposed Framework
Sensitivity	**0.659**	0.275	0.346	0.595	0.562	0.419
PPV	0.140	0.135	0.389	0.399	0.422	**0.595**
Accuracy	0.304	0.193	0.366	0.487	0.487	0.**499**

Co-Sep	0.052	0.036	0.063	**0.232**	0.218	0.221
Cl-Sep	0.537	0.474	0.566	0.024	0.048	**0.846**
Separation	0.166	0.131	0.189	0.075	0.102	**0.432**

The proposed framework also achieves better complex-wise separation (Co-Sep) and cluster-wise separation (Cl-Sep). High Co-Sep value indicates that a benchmark complex corresponds very well to its best-matching cluster or split into two or several more clusters, meanwhile, most members in the cluster or each of these clusters are also members in the complex. Similarity, high Cl-Sep value explains that most members in a cluster is from one or several complexes. Highest Separation value indicates the proposed framework achieves best degree of the bidirectional correspondence between the clustering and the set of benchmark complexes among these methods, that is, maximal fraction of known proteins in the clustering are correctly identified.

According to different treatment on bait nodes, CODEC [9] has two schemes, w0 and w1 (CODEC-w0 and CODEC-w1), which yield two clustering results.

CODEC has the higher sensitivity on both PPI networks than our proposed framework, but our proposed framework achieves better PPV and consistently higher accuracy. CODEC and our proposed framework obtain comparative Co-Sep value. However, our proposed framework achieves significantly higher Cl-Sep value than CODEC. Cl-Sep value penalizes strong overlapping clusters since it is calculated by using marginal sums rather than the cluster size. In other words, low Cl-Sep value of CODEC suffers from its high internal overlap rate. Take CODEC-w1 for example, in the applications on Gavin_2006 and Krogan_2006 networks, the average internal overlap rate between predicted clusters is 52% and 50%, respectively.

Considerably higher homogeneity (Ho) reveals our proposed framework performs better than CODEC does on both networks.

VI. CONCLUSION

In this paper, we propose a new framework for combining topology features of graph and semantic similarities between proteins to extract protein complexes in AP-MS PPI networks. It has been tested on two real AP-MS PPI networks. The results indicated that, by integrating semantic similarities into the clustering process, greater accuracy and better homogeneity have been achieved in comparison with those produced by state-of-art clustering algorithms. The main feature of our method is that it detects protein complexes by taking co-complex relations into account from AP-MS data, as well as incorporating prior knowledge, semantic similarity, into clustering to help detect protein complexes. The proposed framework is also able to detect overlapping modules in PPI networks. In addition, it could also be easily to extend to detect functional modules on bipartite graph from other application domain.

In our future work, different topology features, such as, shortest paths, as well as more mathematical models for integrating these two types of information, worth to be further investigated and assessed, in order to find out the better way which could facilitate the identification of protein complexes in AP-MS data.

ACKNOWLEDGMENT

Bingjing Cai is supported by the Vice Chancellor's Research Scholarships, University of Ulster, UK.

REFERENCES

[1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", Proceedings of the National Academy of Science of the United States of America, Vol 98, pp. 4569-4574, 2001.

[2] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J. R. Knight, et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae", Nature, Vol 403, pp.623-627, 2000

[3] A.C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes", Nature, Vol 415(6868), pp.141-7, 2002.

[4] A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, et al., "Proteome survey reveals modularity of the yeast cell machinery", Nature, Vol 440, pp.631-636, 2006.

[5] J. Yu and F. Fotouhi, "Computational approaches for predicting protein-protein interactions: A survey", Journal of Medical System, Vol 30, pp.39-44, 2006.

[6] G.D. Bader and C. WV Hogue, "An automated method for finding molecular complexes in large protein interaction networks". BMC Bioinformatics, Vol 4(1), pp.2, 2003.

[7] AJ Enright, S Dongen, CA Ouzounis, "An efficient algorithm for largescale detection of protein families", Nucleic Acids Research Vol.30(7), pp.1575-1584, 2002.

[8] B. Adamcsek, G. Palla, I. Farkas, I. Derényi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks". Bioinformatics, Vol. 22(8), pp. 1021-1023, 2006.

[9] G. Geva and R. Sharan, "Identification of protein complexes from co-immunoprecipitation data". Bioinformatics, Vol. 27(1), pp.111-117, 2011.

[10] F. Azuaje, H.Y. Wang, H. Zheng, O. Bodenreider, A. Chesneau, "Predictive integration of gene ontology-driven similarity and functional interactions", In Proceeding of the 6th IEEE International Conference on Data Mining. pp. 114-119, 2006.

[11] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, F.M. Couto, "Semantic similarity in biomedical ontologies", PLos Comput. Biol. Vol. 5, e1000443, 2009.

[12] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, et al., "Gene ontology: tool for the unification of biology", Nat. Genet. Vol.25(1), pp.25-9, 2000.

[13] F. Azuaje and O. Bodenreider, "Incorporating Ontology-Driven Similarity Knowledge into Functional Genomics: An Exploratory Study", Proceeding of the IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE-2004), pp. 317-324, 2004.

[14] B. Cai, H.Y. Wang, H. Zheng, H. Wang, " Detection of protein complexes from Affinity Purification/Mass Spectrometry data", BMC Systems Biology Supplement for ICIBM 2012, accepted.

[15] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Société Vaudoise des Sciences Naturelles, Vol.37, pp. 547–579, 1901.

[16] H. Zheng, F. Azuaje and H.Y. Wang, "seGOsa: Software Environent for Gene Ontology-driven Similarity Assessment". 2010 IEEE Internation Conference on Bioinformatics and Biomedicine, pp. 539-542, 2010.

[17] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, et al., "Global landscape of protein complexes in the yeast saccharomyces cerevisiae". Nature, Vol. 440, pp.637-643, 2006.

[18] S. Pu, J. Wong, B. Turner, E. Cho, S. Wodak, "Up-to-date catalogues of yeast protein complexes". Nucleic Acids Res., Vol. 37, pp.825-831, 2009.

[19] S. Brohée and J.V. Helden, "Evaluation of clustering algorithms for protein-protein interaction networks". BMC Bioinformatics, Vol. 7, pp. 488, 2006.