[Frontiers in Bioscience E4, 2150-2161, January 1, 2012]

Comparison and evaluation of network clustering algorithms applied to genetic interaction networks

Lin Hou[1,2,3], Lin Wang[2], Arthur Berg[4], Minping Qian[1,2], Yunping Zhu[3], Fangting Li[5], Minghua Deng[1,2,6]

[1]LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China, [2]Center for Theoretical Biology, Peking University, Beijing 100871, China, [3]State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China, [4]Center for Statistical Genetics, Pennsylvania State University, Hershey, Pennsylvania, USA, [5]School of Physics, Peking University, Beijing 100871, China, [6]Center for Statistical Science, Peking University, Beijing 100871, China

TABLE OF CONTENTS

1. ABSTRACT

The goal of network clustering algorithms detect dense clusters in a network, and provide a first step towards the understanding of large scale biological networks. With numerous recent advances in biotechnologies, large-scale genetic interactions are widely available, but there is a limited understanding of which clustering algorithms may be most effective. In order to address this problem, we conducted a systematic study to compare and evaluate six clustering algorithms in analyzing genetic interaction networks, and investigated influencing factors in choosing algorithms. The algorithms considered in this comparison include hierarchical clustering, topological overlap matrix, bi-clustering, Markov clustering, Bayesian discriminant analysis based community detection, and variational Bayes approach to modularity. Both experimentally identified and synthetically constructed networks were used in this comparison. The accuracy of the algorithms is measured by the Jaccard index in comparing predicted gene modules with benchmark gene sets. The results suggest that the choice differs according to the network topology and evaluation criteria. Hierarchical clustering showed to be best at predicting protein complexes, Bayesian discriminant analysis based community detection proved best under epistatic miniarray profile (EMAP) datasets, the variational Bayes approach to modularity was noticeably better than the other algorithms in the genome-scale networks.

2. INTRODUCTION

A genetic interaction arises when the phenotype expressed by a double mutation deviates from the combined phenotype of single mutation(1). Large-scale genetic interaction networks are available in model organisms, such as *S. cerevisiae* (2-7), *S. pombe* (8-10), and *E. coli* (11). These datasets have been effective in revealing cellular functions of proteins and understanding the organizational principles of the living cell on a systems level.

There are several high-throughput experimental techniques used to measure genetic interactions in *S. cerevisiae*. The synthetic genetic array technique (SGA) (12-13) generates double mutation strains by crossing a query strain against a library of genome-wide single mutation test strains. After mating is performed, special drugs are used to select out the double mutation strains. The double mutation strains are then grown in a rich media for a defined time period after which the sizes of the colonies are measured. A lethal or sickly interaction occurs when double mutation strain dies or is sicker (less abundant) than expected. Epistatic Miniarray Profiles (EMAP) is modified from SGA (7), which is more quantitative and can measure positive (healthy) interactions as well as negative (sickly) interactions. The interaction network derived can be represented as a matrix, with rows

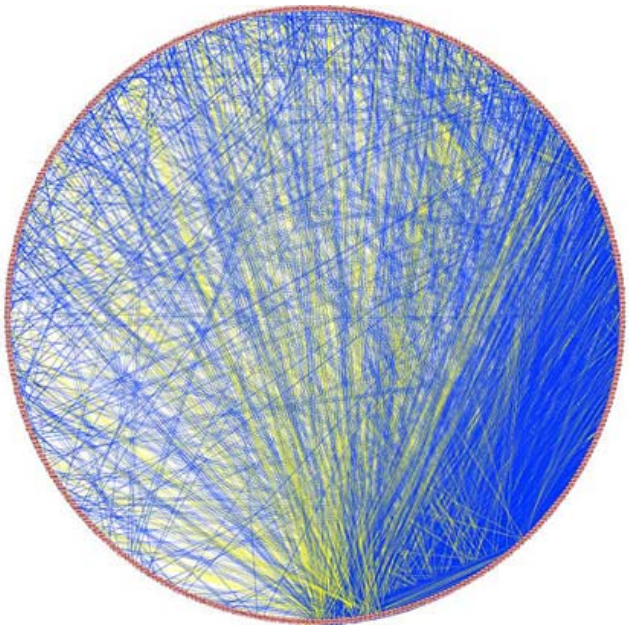

Figure 1. Network view of the early secretory pathway EMAP datasets. Blue edges: negative interactions; yellow edges: positive interactions.

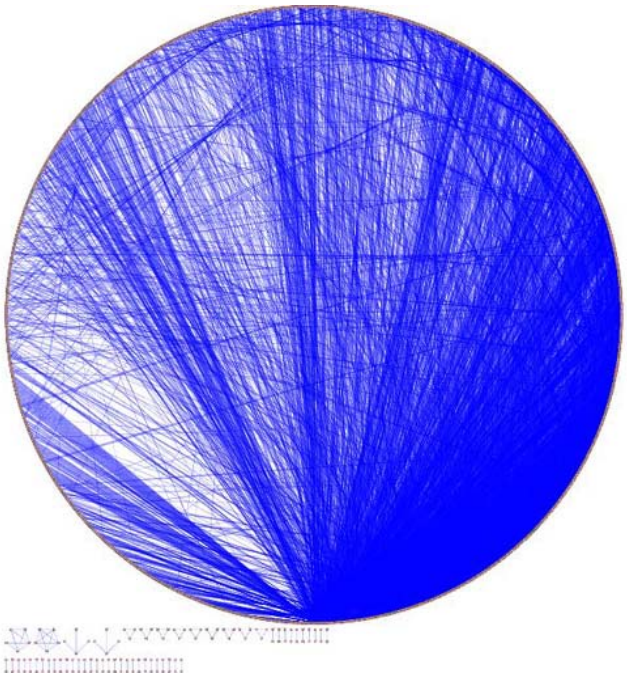

Figure 2. Network view of the synthetic lethal network. Blue edges: negative interactions; yellow edges: positive interactions.

and columns corresponding to query and test genes respectively. In EMAP datasets, query genes and test genes are essentially the same set of genes, resulting in a symmetric matrix.

Large-scale genetic interaction networks are informative, and they can provide answers to many biological questions (1, 3). However, due to the large amounts of data and the substantial noise present in the data, it is difficult to analyze and interpret such networks (Figure 1-2). Network clustering algorithms can overcome some of these difficulties by predicting modules within the network, potentially yielding novel biological hypotheses. Although there are many different network clustering algorithms (14-16), the question as to which algorithms are best at predicting modules in genetic interaction networks

Table 1. Parameter setups for synthetic datasets

	N		p_0	m_0	m_1
A1	500	20	0.3	0.05	0.2
A2	500	20	0.3	0.05	0.3
A3	500	20	0.3	0.05	0.4
A4	500	20	0.3	0.05	0.5
B1	2000	100	0.1	0.005	0.15
B2	2000	100	0.1	0.005	0.2
B3	2000	100	0.1	0.005	0.25
B4	2000	100	0.1	0.005	0.3

has not been previously researched in the literature. Actually, hierarchical clustering is predominately used in the field of genetic interaction networks (2-3, 17), but its accuracy has not been evaluated and compared with other algorithms.

In this study, we seek to answer the following questions: (1) Are there algorithms that significantly outperform the dominantly applied hierarchical clustering algorithm? (2) Which network clustering algorithms have favorable qualities (and which should be avoided)? (3)What factors are important in selecting an optimal algorithm? To answer these questions, a wide array of network clustering algorithms is implemented over several different experimentally derived and synthetically constructed biological networks. Among the six algorithms investigated, some have been applied in genetic interaction networks, like hierarchical clustering (HC) (7), bi-clustering (18), and Bayesian discriminant analysis based community detection (BDA) (19), while others have not been introduced before, including topological overlap matrix (TOM), Markov clustering (MCL), and variational Bayes approach to modularity (VBM). Markov clustering algorithm (20) is shown to be superior in predicting protein complexes within physical protein interaction networks. In gene co-expression networks, hierarchical clustering (21), topological overlap matrix (22), and bi-clustering (23) are effective algorithms in predicting co-regulated gene sets. The performances of these algorithms are compared and evaluated based on the similarity between predicted modules and benchmark gene sets using the Jaccard index. Based on the results on experimental genetic interaction networks, we conclude that in EMAP studies, hierarchical clustering works best for predicting protein complexes, and the performance of these algorithms are comparable to each other in recovering GO co-functional gene sets, with BDA slightly outperforms others. In genome-scale genetic interaction networks, VBM achieves the best accuracy. Plus, we studied how the accuracy of network clustering algorithms varies with the signal to noise ratio in the synthetic network. The accuracy of VBM increases rapidly with the increase of the signal to noise ratio. As the genetic interaction network are wide spread and extensively investigated, a more complete network can be expected, which indicates the potential use of VBM in the future.

3. MATERIALS AND METHODS

3.1. Notation

N: number of nodes in the network.

M: number of edges in the network.

K: number of modules in the network.

A: adjacency matrix of the network; $A_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes i and j:} \\ 0, & \text{otherwise.} \end{cases}$

p_0: probability that a node does not belong to any of the K modules.

$p_i, i = 1, 2, \cdots, K$: prior probability that a node belongs to module i.

s_j: module membership of node j; $s_j = 0$ if node j does not belong to any module.

m_1: probability that a "within" module edge exists ($m_1 = P(e_{ij} = 1 | s_i = s_j)$).

m_0: probability that an edge exists randomly ($m_1 = P(e_{ij} = 1 | s_i \neq s_j)$)

3.2. Experimental genetic interaction networks

Three experimental genetic interaction networks in S. cerevisiae are used here, including an EMAP dataset studying the early secretory pathway (ESP) (7), an EMAP dataset studying the chromosome biology (CHR) (6), and a synthetic lethal genetic interaction network from the BioGrid database (SLN) (24). The ESP-EMAP, CHR-EMAP, and SLN are datasets consist of 424, 743, and 2828 genes respectively. The EMAP datasets were downloaded from http://interactome-cmp.ucsf.edu/ and the SLN data was downloaded from BioGrid (release 3.1.71). It is noted that the EMAP datasets consists of continuous-valued measurements, whereas the SLN consists of binary data.

3.3. Synthetic data

Two sets of synthetic data are simulated. Scenarios A1-A4 are designed to reflect the signal to noise ratio in EMAP and enhanced EMAP datasets. Similarly, scenarios B1-B4 are designed to reflect the signal to noise ratio in SLN and the enhanced SLN. The parameter setups are listed in Supplementary Table 1.

In order to simulate the binary networks, the parameter values need to be appropriately chosen. Previously reported genetic interaction networks guide selection of these parameters. First, parameters determining the number of nodes and interactions are described. In EMAP datasets, the study is focused on a general biological process, such as early secretory pathway (7), chromosome biology (6), unfolded protein response pathway (25), and phosphorylation network (5). The number of query genes ranges from four hundred to seven hundred, and the chance that two genes have a genetic interaction is greater than what is expected by random chance since the genes are co-regulated. Thus, in scenarios A1-A4, the number of nodes is set to 500 and the null probability (m_0) that two genes have an interaction is set to 0.05. The choice of m_0 is consistent with what is observed in the experimental datasets (see Supplementary Table 2). In genome-scale networks like SLN, the interaction network is a

Table 2. Statistics for experimental genetic interaction networks

	N	m_1	m_0	#modules (PC)
ESP	424	0.15	0.04	34
CHR	743	0.17	0.05	125
SLN	2828	0.12	0.003	267

combination of several independent studies, which makes the resulting network larger and sparser. In scenarios B1-B4, the number of nodes is set to 2000. The null probability m_0 is set to 0.005, which also reflects the parameters estimated from the SLN experimental datasets (see Supplementary Table 2).

Next we describe how parameters determining the number and sizes of the modules are chosen. In EMAP studies, the query genes are distributed to several related signaling pathways or functions. While in genome-scale datasets, genes can be classified into distinct functional categories, and represent many aspects of a biological system. We set K to 20 for scenarios A1-A4 and set K to 100 for scenarios B1-B4, as suggested by the experimental data (see Supplementary Table 2). The size of each module is determined by a random sampling. The probability that a node is in module i is randomly sampled from the uniform (0, 1) distribution. An additional parameter, p_0, is introduced to reflect the fact that some portion of nodes cannot be assigned to any of the modules based on the available data; this parameter is arbitrarily set to 0.3 and 0.1. The module membership s_i is sampled from the multi-nominal distribution $\text{Multi-norminal}(N, p_0, p_1, \cdots, p_K)$.

Finally, the adjacency matrix is sampled from Bernoulli distributions. Interactions (e_{ij}) is generated by random sampling from $\text{Bernoulli}(m_1)$ if $s_i = s_j$, and otherwise from$\text{Bernoulli}(m_0)$. We chose m_1 according to the experimental datasets in scenario A1 and B1 (see Supplementary Table 2). In scenarios A2-A4 and B2-B4, we increased m_1 to represent a more complete network.

3.4. Benchmark functional gene sets

GO co-functional gene sets and protein complexes are used as benchmark functional gene sets to compare and evaluate the network clustering algorithms (26-27). Biological Process, Molecular Function, and Cellular Component ontologies are analyzed separately. The mapping of yeast gene products to GO-Slim terms and the information of protein complexes is downloaded from SGD (27)

(http://downloads.yeastgenome.org/literature_curation, updated 26-Feb-2011).

3.5. Network clustering algorithms

We compared six network clustering algorithms—hierarchical clustering (21), topological overlap matrix (22), bi-clustering (23), Markov clustering (20), Bayesian discriminant analysis based community detection, and variational Bayes approach to modularity (28). Hierarchical clustering is widely used in analyzing EMAP datasets, and it is useful for illuminating gene functions (17). Using each gene's genetic interactions in the genome as an interaction profile, the Pearson correlation coefficient between two gene profiles can be used as a similarity measure. Based on this similarity measure, hierarchical clustering arranges the network into a dendrogram. The topological overlap matrix yields another similarity measure, which reflects the commonality of the nodes they connect to (22).

$$TOM_{ij} = \frac{\sum_{k \neq i,j} A_{ik} \times A_{jk} + A_{ij}}{min\{\sum_{k \neq i} A_{ik}, \sum_{k \neq j} A_{jk}\} + 1 - A_{ij}}$$

It is demonstrated as an appropriate measure of similarity when applied to metabolic networks (29) and gene co-expression network (30). Genes can then be clustered through hierarchical clustering, with topological overlap matrix as the similarity matrix. Bi-clustering is an unsupervised machine learning technique, which is useful in gene expression datasets (23). When applied to genetic interaction networks, the two indicies correspond to query and test genes, which are usually the same in EMAP studies. Bi-clustering finds a subset of query genes and test genes, within which interactions are denser compared to background. Unlike hierarchical clustering, which clusters genes with globally coherent genetic interaction profiles, bi-clustering detects genes with both global and local coherent profiles. Bi-clustering is demonstrated as a valuable tool for mapping genes into biological pathways (18). Markov clustering algorithm is based on a simulation of stochastic flow in the network, and it has been successfully applied in protein-protein interaction networks (20). In each iteration step, the stochastic flow is re-distributed by an expansion operation and an inflation operation, so that strong flows are inflated and weak flows are weakened. This procedure is simply implemented as the power of the adjacency matrix of the network, followed by a normalization of each row, which will converge to a partition of the network. The Bayesian discriminant analysis based community detection is an algorithm for analyzing EMAP data we developed previously, which identifies modules in a probabilistic genetic interaction network. It calculates the likelihood ratio of a set of genes by contrasting "module" model with "random set" model, assuming genes in biological significant modules are densely interacting, while interactions in random gene sets are sparse and random. The variational Bayes approach to modularity is an efficient implementation of a Bayesian framework, which poses the community detection problem as an inference of latent variables in a probabilistic model. It infers the number of modules, model parameter, and module assignment simultaneously through optimizing a likelihood function (28, 31).

These algorithms are applied to experimental and synthetic genetic interaction networks. Hierarchical clustering is performed by the hclust() function in the R program (http://www.r-project.org/). The prediction results of ESP-EMAP and CHR-EMAP by bi-clustering were downloaded from the supplementary materials of Pu *et al.* (18). The other algorithms have been implemented in-house according to the methodological descriptions in the respective published sources. Bi-clustering and BDA are specifically designed for EMAP data and their results are

Table 3. Topological properties of experimental genetic interaction networks

	N[1]	Ave. Deg[2]	Ave. NCC[3]
ESP*	424	16	0.417
CHR*	743	39	0.424
SLN	2828	8	0.304

*: the EMAP datasets are transformed into binary network by thresholding at 2.5.

1: N: number of nodes

2: average degree, $\text{Ave. Deg} = \frac{\sum_i |N_i|}{N}$

3: average network clustering coefficient;

$\text{NCC}_i = \frac{\sum_{j \in N_i, k \in N_i} A_{jk}}{|N_i| \times (|N_i|-1)}$, $\text{Ave. NCC} = \frac{\text{NCC}_i}{N}$

limited to this type of data. Among the six algorithms investigated, HC, bi-clustering, and BDA are designed to work with continuous data, while the others are designed to work with binary data. For the algorithms requiring binary data, the EMAP datasets are transformed to binary network by thresholding the values at 2.5; this thesholding approximately corresponds to a 5% loss of information (7). The prediction results of HC and TOM vary according to the depth at which the dendrogram is cut. Also, the results of MCL are dependent on the choice of inflation parameter. To account for the sensitivity of these parameters, several thresholds and inflation parameters were considered, and the results with the best performance were selected for comparison with other algorithms. Only clusters containing between 3 and 50 genes were used in the comparison.

3.6. Jaccard index: evaluation measure of the predicted modules

We used the Jaccard index to determine how well the predicted modules correspond to benchmark ("theoretical") gene sets (16). The Jaccard index between two sets M_i and B_j is defined as

$$\frac{\#\{M_i \cap B_j\}}{\#\{M_i \cup B_j\}}.$$

For module M_i, the Jaccard index between M_i and each gene set B_j in the benchmark is computed, and the Jaccard index of M_i and the benchmark gene sets is defined as the maximum of Jaccard index between M_i and any gene set in the benchmark:

$$\text{Jaccard Index}(M_i, B) = max_j\{\text{Jaccard Index}(M_i, B_j)\}$$

.

Thus, the average Jaccard index of the predicted modules and the benchmark gene sets can be computed:

$$\text{Jaccard Index}(M, B) = \frac{\sum_{i \in 1,\cdots,k} \text{Jaccard Index}(M_i, B)}{k}$$

The accuracy of network clustering algorithms is evaluated by the average Jaccard index of the predicted modules and benchmark gene sets. In the ideal situation where the predicted modules perfectly match the benchmark gene sets, the Jaccard index is 1. The larger the Jaccard index, the better the predictions are. Furthermore, the significance of difference between different network clustering algorithm X and Y is tested by the Wilcoxon rank sum test between $\{\text{Jaccard Index}(M_i, B), i \in \{1, \cdots, k_X\}$ and $\{\text{Jaccard Index}(M_i, B), i \in \{1, \cdots, k_Y\}$.

4. RESULTS

4.1. Comparisons with the experimental data

A major difficulty in comparing and evaluating network clustering algorithms is the lack of established criteria. Here, we propose to assess a network clustering algorithm with the biological significance of the modules it predicted, which we refer to here as prediction accuracy. Accuracy is measured by the Jaccard index between predicted modules and benchmark gene sets. In order to systematically compare the network clustering algorithms, three genetic interaction networks are investigated here. ESP-EMAP and CHR-EMAP are EMAP datasets, which focuses on a particular method of measuring genetic interactions in yeast. SLN is a synthetic lethal genetic interaction network from BioGrid, which is a literature-curated interaction database. The two kinds of networks differ in three aspects: 1) networks from EMAP studies contain fewer genes than literature-curated networks; 2) the proportion of genetic interactions in EMAP studies is larger, since it targets at a potentially co-functional set of proteins; 3) the EMAP experiments yield a continuous measure of interaction whereas the SLN data is binary. Because of the structural differences in the EMAP and SLN datasets (see Supplementary Table 3), we proceed in providing different preferences for the network clustering algorithms for these two types of datasets.

We applied the network clustering algorithms to these networks, and the accuracy of each algorithm is evaluated by comparing the predicted modules to benchmark gene sets with the Jaccard index. Additionally, the size of the predicted modules and the number of benchmark gene sets that can be recovered are also investigated. Comparisons are provided across the different algorithms, networks, and benchmark gene sets.

In EMAP datasets, the algorithms are comparable to each other in predicting GO functional gene sets, though HC and BDA are slightly more accurate (Figure 3AB). The results are consistent in ESP-EMAP and CHR-EMAP. Moreover, in ESP-EMAP more GO functional gene sets are recovered by BDA compared to HC (Supplementary Table 5). VBM did not predict any module of appropriate size (consisting of 3 to 50 genes) in both datasets. When we use GO protein complexes as the benchmark gene set, HC achieves a noticeably higher Jaccard index than the other algorithms. Particularly, in CHR-EMAP, the Jaccard index of HC is significantly higher than others (Figure 4, supplementary Table 4). This is because subunits of a protein complex form a functional unit with highly similar genetic interaction profiles (17), which is captured by hierarchical clustering directly. However, a stringent cutoff needs to be applied for HC to achieve a high Jaccard index. In that case, fewer modules

Table 4. Significance of difference between different network clustering algorithms

Algorithm[&]/p-value*	BP	MF	CC	PC
ESP-EMAP	BDA/0.3588	BDA/0.06681	BDA/0.06664	HC/0.1003
CHR-EMAP	BDA/0.1319	HC/0.7809	BDA/0/0174	HC/0.02821
SLN	VBM/4.94E-05	VB/0.00073	VBM/0.01431	HC/0.809

&: the algorithm which achieves the highest average Jaccard index in the corresponding dataset and benchmark set is given .*: the p-values are calculated by the Wilcoxon rank sum test between Jaccard index of two algorithms, the algorithm shown and the one with the second highest average Jaccard index.

Table 5. Summary statistics of predicted modules in ESP-EMAP

PC*	Jaccard Index	#module	Ave. size	#enriched benchmark set@
HC$	0.3389	6	4	5
TOM$	0.1318	30	4	5
Bi-Clustering	0.1602	391	6	9
MCL$	0.0746	29	5	1
BDA	0.1184	8	18	6
BP	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.1231	6	4	3
TOM	0.1136	11	4	6
Bi-Clust	0.1229	391	6	9
MCL	0.0853	24	4	0
BDA	0.1288	8	18	7
MF	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0714	33	6	2
TOM	0.0635	52	5	1
Bi-Clust	0.0592	391	6	1
MCL	0.0689	23	6	0
BDA	0.0846	8	18	1
CC	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0711	16	6	2
TOM	0.0626	52	5	0
Bi-Clust	0.598	391	6	4
MCL	0.0655	34	5	0
BDA	0.0935	8	18	4

*: the benchmark dataset. $: the clustering result of HC, TOM, and MCL is dependent on cutoffs or parameters. We tried different cutoffs or parameters, the one with the highest Jaccard index in corresponding dataset and benchmark set is presented.@: hyper-geometric test applied to test the enrichment of gene sets. Significance level: FDR<=0.05.

are predicted, and the average size of predicted modules are small (around 5), leaving a large proportion of genes as singletons (see Supplementary Table 5-6).

In the genome-scale synthetic lethal network, VBM is significantly better than the other algorithms in predicting GO functional gene sets (Figure 3C). The other three algorithms are comparable to each other, but measurably less accurate than the VBM. All these algorithms are comparable in predicting protein complexes in SLN (Figure 4). A comparison of the results in EMAP datasets and SLN highlights that differences of the Jaccard index between different datasets is greater than the differences among the different algorithms in the same network.

Based on these observations, we concluded that the results in ESP and CHR are similar. BDA and HC are the best choice for predicting co-functional modules, and HC is best for predicting protein complex memberships. As for SLN, the VBM is best regardless of the biological focus. Generally, it is more difficult to identify modules in SLN because the network is less complete compared to the EMAP datasets. The best Jaccard index obtained in the SLN data is much smaller than any of the Jaccard indices calculated in the EMAP datasets. Our results also demonstrate that the accuracy of network clustering algorithms is affected by network topologies, as reported previously (16).

4.2. Comparisons with the synthetic data

In addition to the experimental datasets, we applied the algorithms to simulated binary networks. There are two advantages to this approach. First, in real datasets, a true module is frequently regarded as false positive because of limited biological knowledge. As a result, the false positive rate is usually over-estimated, and the true positive rate is under-estimated. On the other hand, in simulated datasets, we always know what the true modules are, thus the algorithms can be more accurately evaluated. Second, in the synthetic data parameters can be manipulated to study how the parameter values affect the results.

After an inspection of the experimental genetic interaction networks, eight sets of parameters are employed. The different parameter setups represent different types of genetic interaction networks. Scenarios A1-A4 mimics the EMAP datasets with the percentage of within module interactions varying from 20% to 50%. In synthetic network A1, similar Jaccard index levels were observed as those in the experimental datasets and the performance of the algorithms are comparable to each other. VBM, as in the experimental data, failed to predict any module of size between 3 and 50. When we improve the signal-to-noise ratio, naturally the performance of all the algorithms demonstrated improvement (Figure 5A). The most striking change was with VBM. In the A4

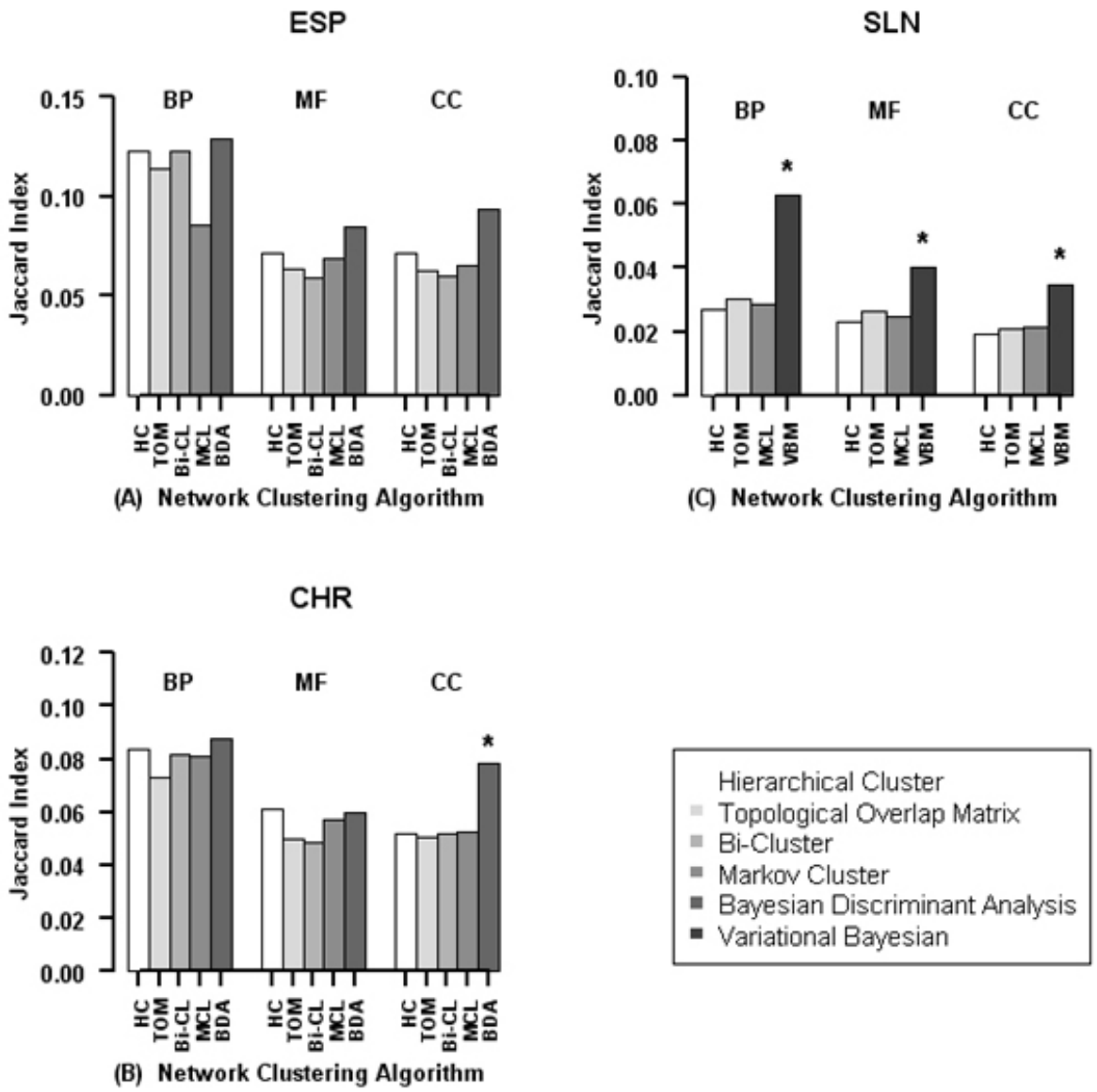

Figure 3. Comparison of the Jaccard index between predicted modules and the GO functional gene sets on the early secretory pathway EMAP (A), chromosome biology EMAP (B), and the genome-scale synthetic lethal network (C). The algorithms compared are hierarchical clustering (HC), topological overlap matrix (TOM), Bi-clustering (Bi-CL), Markov clustering (MCL), Bayesian discriminant analysis based community detection (BDA), and variational Bayes approach to modularity (VBM). The biological process (BP), molecular function (MF), and cellular component (CC) ontologies are compared separately. The larger the Jaccard index, the better the result. The * indicates the algorithm with the highest Jaccard index and the algorithm with the second highest Jaccard index are statistically significant ($p<0.05$).

scenario, it achieved a Jaccard index around 0.8, whereas the second best algorithm, Markov clustering, only yielded a Jaccard index of 0.58. In scenarios B1-B4, the number of nodes is much larger, but the network is sparser. VBM consistently outperformed the others in average Jaccard index (Figure 5B). The simulation studies show the hierarchical clustering based methods (HC and TOM) are robust across the different networks. However, their prediction accuracy can be considerable inferior to the other methods. On the other hand, the VBM is very sensitive to the signal-to-noise ratio, but when it works, it can considerably outperform the other methods.

In addition to the Jaccard index, we also compared the number of benchmark sets that can be predicted by each algorithm (Supplementary Table 8). In synthetic EMAP dataset A2-A4, VBM not only recovered the most number of

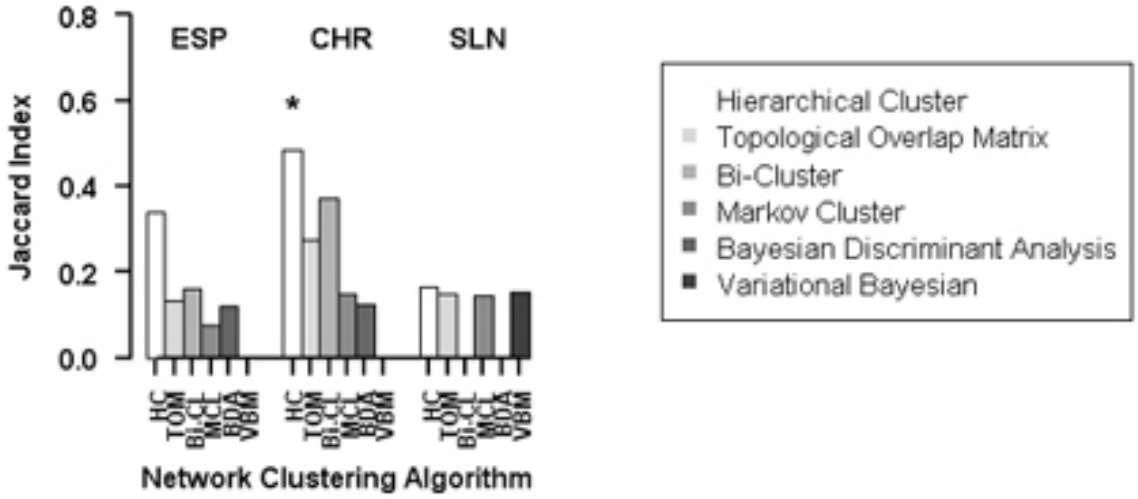

Figure 4. Comparison of the Jaccard index between predicted modules and the GO protein complexes on the early secretory pathway EMAP, chromosome biology EMAP, and the genome-scale synthetic lethal network. The algorithms compared are hierarchical clustering (HC), topological overlap matrix (TOM), Bi-clustering (Bi-CL), Markov clustering (MCL), Bayesian discriminant analysis based community detection (BDA), and variational Bayes approach to modularity (VBM). The biological process (BP), molecular function (MF), and cellular component (CC) ontologies are compared separately. The larger the Jaccard index is, the better the result is. The * indicates the algorithm with the highest Jaccard index and the algorithm with the second highest Jaccard index are statistically significant ($p<0.05$).

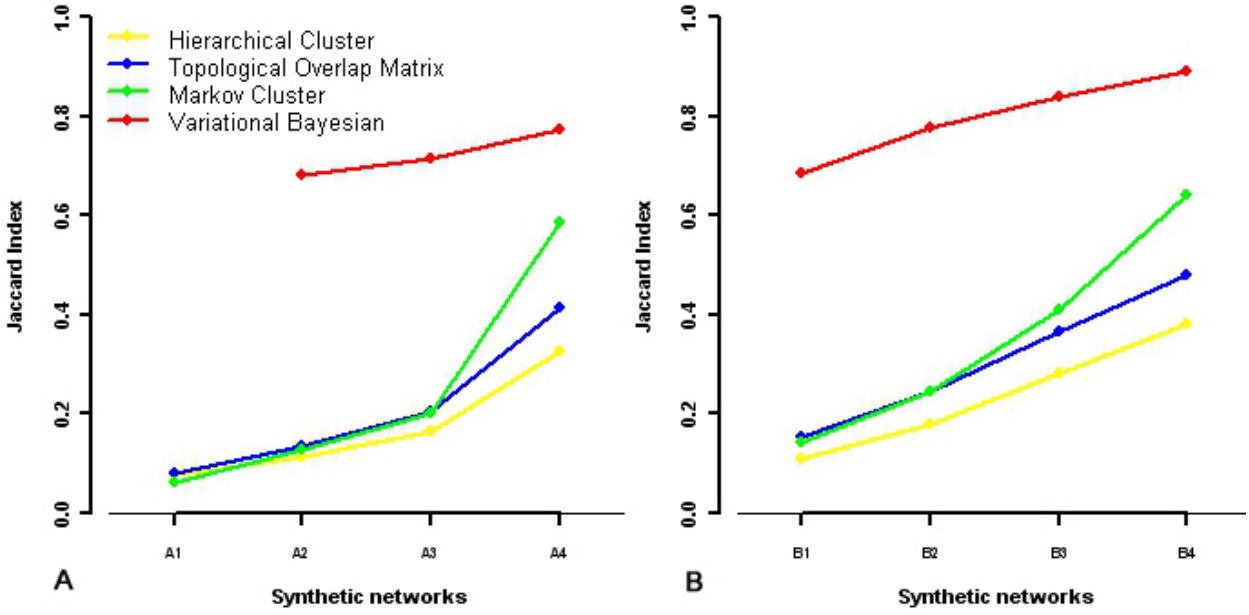

Figure 5. Comparison of the Jaccard index on the synthetic datasets A1-A4 (left) and B1-B4 (right). The algorithms compared are hierarchical clustering (HC, yellow), topological overlap matrix (TOM, blue), Markov clustering (MCL, green), and variational Bayes approach to modularity (VBM, red).

modules, but also achieved the best accuracy at the largest average module size. In genome-scale datasets, VBM can achieve a Jaccard index of 0.9 and recover 60% of true modules when 30% of interactions can be observed (B4). We also tested the computation time and memory consumption of these algorithms on synthetic datasets (see Supplementary Table 9). Overall, for most algorithms it needs less than 50M and 3 minutes for A1-A4, and less than 300M and 15 minutes for B1-B4. VBM took around 80 minutes on datasets B1-B4, however, this time can be cut down by decreasing iteration steps.

Table 6. Summary statistics of predicted modules in CHR-EMAP

PC	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.4864	21	4	28
TOM	0.2734	24	4	20
Bi-Clust	0.3698	268	9	52
MCL	0.1482	39	7	4
BDA	0.125	24	18	21
BP	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0839	62	6	18
TOM	0.0729	83	5	6
Bi-Clust	0.0818	268	9	17
MCL	0.0809	26	4	0
BDA	0.0875	24	18	4
MF	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0609	62	6	0
TOM	0.0496	83	5	1
Bi-Clust	0.0484	268	9	1
MCL	0.0575	26	4	0
BDA	0.06	24	18	0
CC	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0519	11	4	3
TOM	0.0506	7	3	2
Bi-Clust	0.0519	268	9	6
MCL	0.0523	33	4	0
BDA	0.0781	24	18	5

Table 7. Summary statistics of predicted modules in SLN

PC	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.1644	187	6	42
TOM	0.1483	354	5	59
MCL	0.1424	335	6	44
VBM	0.1535	22	16	24
BP	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0269	258	7	6
TOM	0.0305	280	7	1
MCL	0.0287	298	7	1
VB	0.0627	22	16	2
MF	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0233	258	7	0
TOM	0.0262	280	7	0
MCL	0.025	298	7	1
VB	0.04	22	16	0
CC	**Jaccard Index**	**#module**	**Ave. size**	**#enriched benchmark set**
HC	0.0191	258	7	0
TOM	0.0208	280	7	0
MCL	0.0212	298	7	4
VB	0.0349	22	16	1

5. DISCUSSION

Large-scale genetic interaction network data are becoming increasingly available bringing insight into the cellular organization of the cell. To interpret the network and make biological inference, network clustering algorithms are needed to identify modules from hundreds or thousands of interactions. Various network clustering algorithms have been developed and applied to biological and social networks. In this study, we compared and evaluated different network clustering algorithms applied to genetic interaction networks so as to provide a guide in selecting an accurate and effective algorithm. No single algorithm universally outperformed the rest. In clustering EMAP datasets, both hierarchical clustering and Bayesian discriminant analysis based community detection are recommended. For genome-scale networks as SLN, the varational Bayes approach to modularity works consistently well for identifying protein complexes and GO co-functional gene sets.

In addition to the comparison and evaluation study in experimental genetic interaction networks, we also applied the algorithms to synthetic data. These studies shed light on what we can achieve with fixed network topologies, and what we should do to improve the understanding of genetic interaction networks. Comparison of network clustering algorithms between different synthetic networks suggests that the accuracy of module identification is highly dependent on the completeness of the network. Thus, great effort should be made to increase the overall coverage of the genetic interaction networks, which can be achieved by both experimental and computational strategies.

Experimentally, almost every pair of genes in *S. cerervisiae* has been tested for their genetic interaction (2). In spite of this advancement, we believe large amounts of false negatives exist in current networks for two reasons. First, some pathways are not active in normal laboratory conditions. As a result, cells usually do not behave differently after perturbations on these pathways, and the corresponding interactions cannot be identified in normal conditions. To remove these false negatives, it is helpful to re-test the genetic interaction network under stress or with

Table 8. Summary statistics of predicted modules in synthetic networks

	A1	A2	A3	A4	B1
HC	102/(5)/2*	93/(5)/11	55/(9)/11	37/(9)/15	346/(5)/42
TOM	80/(6)/3	64/(8)/11	46/(11)/11	24/(11)/14	263/(8)/59
MCL	4/(4)/0	7/(6)/4	11/(13)/8	9/(18)/9	229/(7)/58
VBM		14/(23)/12	15/(25)/12	19/(19)/17	24/(31)/23

*The numbers shown are (1) number of modules predicted; (2) average size of modules; (3) number of benchmark sets recovered (FDR <=0.05).

Table 9. Computation time and memory consumption for synthetic datasets

Time(S)/Memory(M)*	HC	TOM	MCL	VB
A1	2/33	15/30	10/41	143/34
A2	2/33	15/30	10/41	143/34
A3	2/33	15/30	9/41	143/34
A4	2/33	15/30	9/38	143/34
B1	132/280	740/206	922/295	4709/191
B2	132/280	741/206	923/295	4728/191
B3	132/280	741/206	931/295	4738/191
B4	132/280	741/206	1005/295	4735/191
Time complexity	$O(n^2 \times \log n)$	$O(n^2 \times \log n)$	$O(n \times K^2)$	$O(MK)$

*: time complexity of BDA, $O(n \times M)$; bi-clustering, $O(l \times k \times n^2)$; n, number of nodes; K, number of modules; M, number of edges.

drug treatment. A recent work on the DNA damage pathway (32) highlights the additional insights gained by comparing the network in normal conditions and under DNA-damage agent treatment. Second, current studies choose growth rate as the phenotype to measure genetic interactions. While growth rate is easy to measure in a quantitative and high-throughput fashion, it may not respond sensitively to a particular biological process of interest. Both experimental studies (25) and theoretical studies (33) proved that introducing more phenotypes indeed increases the coverage of genetic interaction networks. Therefore it is helpful to measure more phenotypes when experimentally applicable. Separate from the experimental methods, the coverage of genetic interaction networks can be improved through computational approaches. Recently, several studies tried to predict genetic interactions based on diverse biological data, including physical protein interaction network, gene co-expression data, and functional annotation data (34-36).There is also imputation method to deal with the missing data in EMAP studies (37). Although these methods are effective in predicting genetic interactions, it remains unclear to what extent the predicted networks can improve the accuracy of module identification.

In conclusion, in order to choose the best algorithm, the network topology and biological questions of interest should be considered. In EMAP datasets, hierarchical clustering and Bayesian discriminant analysis based community detection is recommended to predict protein complexes and identify co-functional gene sets. In genome-scale datasets, the problem becomes more difficult. Our suggestion is that the variational Bayes approach to modularity should be implemented.

6. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No.10871009 to M.D., No. 10721403 to M.D. and No. 10774009to F.L.); National High Technology Research and Development of China (No. 2008AA02Z306 to M.D.); National Key Basic Research Project of China (No.2009CB918503 to M.D., No.2006CB910706 to F.L.); The Fundamental Research Funds for the Central Universities in China to F.L. Support from the Center for Statistical Science at Peking University to A.B.

7. REFERENCES

1. S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews and C. Boone: Systematic mapping of genetic interaction networks. *Annu Rev Genet*, 43, 601-25 (2009)

2. M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z. Y. Lin, W. Liang, M. Marback, J. Paw, B. J. San Luis, E. Shuteriqi, A. H. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pal, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A. C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews and C. Boone: The genetic landscape of a cell. *Science*, 327(5964), 425-31 (2010)

3. P. Beltrao, G. Cagney and N. J. Krogan: Quantitative genetic interactions reveal biological modularity. *Cell*, 141(5), 739-45 (2010)

4. J. Zheng, J. J. Benschop, M. Shales, P. Kemmeren, J. Greenblatt, G. Cagney, F. Holstege, H. Li and N. J. Krogan: Epistatic relationships reveal the functional organization of yeast transcription factors. *Mol Syst Biol*, 6, 420 (2010)

5. D. Fiedler, H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, A. C. Silva, M. Shales, S. R. Collins, S. van

Wageningen, P. Kemmeren, F. C. Holstege, J. S. Weissman, M. C. Keogh, D. Koller, K. M. Shokat and N. J. Krogan: Functional organization of the S. cerevisiae phosphorylation network. *Cell*, 136(5), 952-63 (2009)

6. S. R. Collins, K. M. Miller, N. L. Maas, A. Roguev, J. Fillingham, C. S. Chu, M. Schuldiner, M. Gebbia, J. Recht, M. Shales, H. Ding, H. Xu, J. Han, K. Ingvarsdottir, B. Cheng, B. Andrews, C. Boone, S. L. Berger, P. Hieter, Z. Zhang, G. W. Brown, C. J. Ingles, A. Emili, C. D. Allis, D. P. Toczyski, J. S. Weissman, J. F. Greenblatt and N. J. Krogan: Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446(7137), 806-10 (2007)

7. M. Schuldiner, S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, J. S. Weissman and N. J. Krogan: Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3), 507-19 (2005)

8. A. Roguev, S. Bandyopadhyay, M. Zofall, K. Zhang, T. Fischer, S. R. Collins, H. Qu, M. Shales, H. O. Park, J. Hayles, K. L. Hoe, D. U. Kim, T. Ideker, S. I. Grewal, J. S. Weissman and N. J. Krogan: Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322(5900), 405-10 (2008)

9. A. Roguev, M. Wiren, J. S. Weissman and N. J. Krogan: High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe. *Nat Methods*, 4(10), 861-6 (2007)

10. S. J. Dixon, Y. Fedyshyn, J. L. Koh, T. S. Prasad, C. Chahwan, G. Chua, K. Toufighi, A. Baryshnikova, J. Hayles, K. L. Hoe, D. U. Kim, H. O. Park, C. L. Myers, A. Pandey, D. Durocher, B. J. Andrews and C. Boone: Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*, 105(43), 16653-8 (2008)

11. G. Butland, M. Babu, J. J. Diaz-Mejia, F. Bohdana, S. Phanse, B. Gold, W. Yang, J. Li, A. G. Gagarinova, O. Pogoutse, H. Mori, B. L. Wanner, H. Lo, J. Wasniewski, C. Christopolous, M. Ali, P. Venn, A. Safavi-Naini, N. Sourour, S. Caron, J. Y. Choi, L. Laigle, A. Nazarians-Armavil, A. Deshpande, S. Joe, K. A. Datsenko, N. Yamamoto, B. J. Andrews, C. Boone, H. Ding, B. Sheikh, G. Moreno-Hagelseib, J. F. Greenblatt and A. Emili: eSGA: E. coli synthetic genetic array analysis. *Nat Methods*, 5(9), 789-95 (2008)

12. A. H. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers and C. Boone: Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550), 2364-8 (2001)

13. A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A. M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey and C. Boone: Global mapping of the yeast genetic interaction network. *Science*, 303(5659), 808-13 (2004)

14. A. L. Barabasi and Z. N. Oltvai: Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2), 101-13 (2004)

15. T. Aittokallio and B. Schwikowski: Graph-based methods for analysing networks in cell biology. *Brief Bioinform*, 7(3), 243-55 (2006)

16. J. Song and M. Singh: How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25(23), 3143-50 (2009)

17. S. R. Collins, M. Schuldiner, N. J. Krogan and J. S. Weissman: A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol*, 7(7), R63 (2006)

18. S. Pu, K. Ronen, J. Vlasblom, J. Greenblatt and S. J. Wodak: Local coherence in genetic interaction patterns reveals prevalent functional versatility. *Bioinformatics*, 24(20), 2376-83 (2008)

19. L. Hou, L. Wang, M. Qian, D. Li, C. Tang, Y. Zhu, M. Deng and F. Li: Modular analysis of the probabilistic genetic interaction network. *Bioinformatics*, 27(6), 853-9 (2011)

20. S. Brohee and J. van Helden: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7, 488 (2006)

21. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25), 14863-8 (1998)

22. A. M. Yip and S. Horvath: Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8, 22 (2007)

23. A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler: A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9), 1122-9 (2006)

24. C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski and M. Tyers: The BioGRID Interaction

Database: 2011 update. *Nucleic Acids Res*, 39(Database issue), D698-704 (2011)

25. M. C. Jonikas, S. R. Collins, V. Denic, E. Oh, E. M. Quan, V. Schmid, J. Weibezahn, B. Schwappach, P. Walter, J. S. Weissman and M. Schuldiner: Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, 323(5922), 1693-7 (2009)

26. C. Gene Ontology: The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38(Database issue), D331-5 (2010)

27. J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer and D. Botstein: Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl), 67-73 (1997)

28. J. M. Hofman and C. H. Wiggins: Bayesian approach to network modularity. *Phys Rev Lett*, 100(25), 258701 (2008)

29. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi: Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551-5 (2002)

30. B. Zhang and S. Horvath: A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4, Article17 (2005)

31. A. E. Teschendorff, Y. Wang, N. L. Barbosa-Morais, J. D. Brenton and C. Caldas: A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21(13), 3025-33 (2005)

32. S. Bandyopadhyay, M. Mehta, D. Kuo, M. K. Sung, R. Chuang, E. J. Jaehnig, B. Bodenmiller, K. Licon, W. Copeland, M. Shales, D. Fiedler, J. Dutkowski, A. Guenole, H. van Attikum, K. M. Shokat, R. D. Kolodner, W. K. Huh, R. Aebersold, M. C. Keogh, N. J. Krogan and T. Ideker: Rewiring of genetic networks in response to DNA damage. *Science*, 330(6009), 1385-9 (2010)

33. E. S. Snitkin and D. Segre: Epistatic interaction maps relative to multiple metabolic phenotypes. *PLoS Genet*, 7(2), e1001294 (2011)

34. C. Y. Park, D. C. Hess, C. Huttenhower and O. G. Troyanskaya: Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *PLoS Comput Biol*, 6(11), e1001009 (2010)

35. S. L. Wong, L. V. Zhang, A. H. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, C. Boone and F. P. Roth: Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A*, 101(44), 15682-7 (2004)

36. G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, V. Kumar and E. E. Schadt: An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol*, 6(9) (2010)

37. I. Ulitsky, N. J. Krogan and R. Shamir: Towards accurate imputation of quantitative genetic interactions. *Genome Biol*, 10(12), R140 (2009)

Abbrevations: EMAP: epistatic miniarray profiles; SGA: synthetic genetic array technique; HC: hierarchical clustering; BDA: Bayesian discriminant analysis based community detection; TOM: topological overlap matrix; MCL: Markov clustering; VBM: variational Bayes approach to modularity; GO: gene ontology; ESP: early secretory pathway; CHR: chromosome biology; SLN: synthetic lethal genetic interaction network

Key Words: Genetic interaction, Network, Clustering algorithm, Jaccard index, Comparison, Epistatic miniarray profiles

Send correspondence to: Minghua Deng, LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China, Tel: 86-10-62767562, Fax: 86-10-62751801, E-mail: dengmh@math.pku.edu.cn

http://www.bioscience.org/current/vol4E.htm