METHOD ARTICLE

An unsupervised disease module identification technique in biological networks using novel quality metric based on connectivity, conductance and modularity [version 1; referees: 2 approved with reservations]

Raghvendra Mall[1], Ehsan Ullah[1], Khalid Kunji[1], Michele Ceccarelli[2,3], Halima Bensmail[1]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[2]University of Sannio, Benevento, Italy
[3]Bioinformatics Lab, BIOGEM Istituto di Ricerche Genetiche G. Salvatore, Ariano Irpino, Italy

v1 **First published:** 26 Mar 2018, **7**:378 (doi: 10.12688/f1000research.14258.1)
Latest published: 26 Mar 2018, **7**:378 (doi: 10.12688/f1000research.14258.1)

Abstract

Disease processes are usually driven by several genes interacting in molecular modules or pathways leading to the disease. The identification of such modules in gene or protein networks is the core of computational methods in biomedical research. With this pretext, the Disease Module Identification (DMI) DREAM Challenge was initiated as an effort to systematically assess module identification methods on a panel of 6 diverse genomic networks. In this paper, we propose a generic refinement method based on ideas of merging and splitting the hierarchical tree obtained from any community detection technique for constrained DMI in biological networks. The only constraint was that size of community is in the range [3, 100]. We propose a novel model evaluation metric, called F-score, computed from several unsupervised quality metrics like modularity, conductance and connectivity to determine the quality of a graph partition at given level of hierarchy. We also propose a quality measure, namely Inverse Confidence, which ranks and prune insignificant modules to obtain a curated list of candidate disease modules (DM) for biological network. The predicted modules are evaluated on the basis of the total number of unique candidate modules that are associated with complex traits and diseases from over 200 genome-wide association study (GWAS) datasets. During the competition, we identified 42 modules, ranking 15th at the official false detection rate (FDR) cut-off of 0.05 for identifying statistically significant DM in the 6 benchmark networks. However, for stringent FDR cut-offs 0.025 and 0.01, the proposed method identified 31 (rank 9) and 16 DMIs (rank 10) respectively. From additional analysis, our proposed approach detected a total of 44 DM in the networks in comparison to 60 for the winner of DREAM Challenge. Interestingly, for several individual benchmark networks, our performance was better or competitive with the winner.

Keywords

Disease Module Identification, Biological Networks, Community Detection, GWAS, Modularity, Conductance, PPI, Co-expression

Open Peer Review

Referee Status: ? ?

	Invited Referees 1	 2
version 1 published 26 Mar 2018	? report	? report

1 **Eric E Schadt**, Icahn School of Medicine at Mount Sinai, USA

2 **Yunpeng Liu**, Massachusetts Institute of Technology (MIT), USA

Discuss this article

Comments (0)

This article is included in the DREAM Challenges gateway.

Corresponding author: Raghvendra Mall (rmall@hbku.edu.qa)

Author roles: Mall R: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Ullah E**: Data Curation, Formal Analysis, Investigation, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kunji K**: Data Curation, Formal Analysis, Investigation, Resources, Validation, Writing – Review & Editing; **Ceccarelli M**: Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Bensmail H**: Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Mall R, Ullah E, Kunji K *et al.* **An unsupervised disease module identification technique in biological networks using novel quality metric based on connectivity, conductance and modularity [version 1; referees: 2 approved with reservations]** *F1000Research* 2018, **7**:378 (doi: 10.12688/f1000research.14258.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

First published: 26 Mar 2018, **7**:378 (doi: 10.12688/f1000research.14258.1)

Motivation & background

A variety of genomic data has been used to construct biological networks. Biological networks are scale free by nature[1] and it is well-known that scale-free networks exhibit community like structure[2–5]. Community like structure in networks is equivalent to presence of a high degree of modularity[5]. In biological networks, the modules often comprise of genes or proteins that are involved in the same biological functions. Network module identification methods, commonly known as community detection[4–9] and graph partitioning methods[10–12], attempt to reveal these functional units[2,13,14] which is key to derive biological insights from genomic networks[15–18]. However, the performance of different community detection methods using diverse parameter settings to uncover biologically relevant modules in myriad networks remain poorly understood because there has been no community effort to transparently evaluate module identification methods on common benchmarks and across diverse types of genomic networks. Thus, it is very difficult to objectively compare the strengths and limitations of alternative approaches. Evaluation of module identification methods typically relied either on random graphs[13], which do not allow for assessment of biological relevance of modules, or on pre-annotated functional gene sets[18] (e.g., gene ontology or molecular pathway databases such as KEGG), which are still primarily incomplete and biased towards well-studied pathways.

To address these issues, an open community DREAM challenge enabling comprehensive and rigorous assessment of module identification methods across a broad range of gene and protein networks was initiated. The task in sub-challenge 1 was to identify functional modules in 6 individual benchmark networks s.t. the module size satisfied the constraint: $3 \leq modul\ esize \leq 100$. The predicted modules were evaluated based on data from disease-relevant genome-wide association studies (GWAS). GWAS have successfully identified thousands of genetic loci associated with a broad range of complex traits and diseases. The variants are mapped to genes allowing to ask whether specific network modules are enriched in these genes[19]. The DREAM challenge organizers employed a comprehensive collection of over 200 GWAS datasets, thereby, covering a broad spectrum of functional units, many of which have not been annotated previously.

In this paper, we focus on sub-challenge 1 where the goal is to predict functional modules for individual anonymized networks across a broad range of gene and protein networks. Our proposed pipeline requires a hierarchical tree from any state-of-the-art hierarchical community detection technique as input. The pipeline first identifies the optimal level of hierarchy using an F-score comprising of quality metrics like conductance[13], modularity[2] and connectivity[1]. Then it traverses the hierarchy bottom-up from the optimal level allowing to merge smaller communities based on the weighted connectivity criterion as long as they fit the size constraint. Further, it splits giant connected components ($modulesize > 100$).

For each giant connected component, we re-build the hierarchical tree using a linkage based agglomerative hierarchical technique and identify the optimal cut (number of clusters k) using the proposed F-score criterion. Finally, we propose a metric to indicate the confidence in each module among the final set of detected modules and develop a method to automatically select the right confidence threshold to prune less meaningful modules. Figure 1 depicts the proposed pipeline for the constrained disease module identification problem.

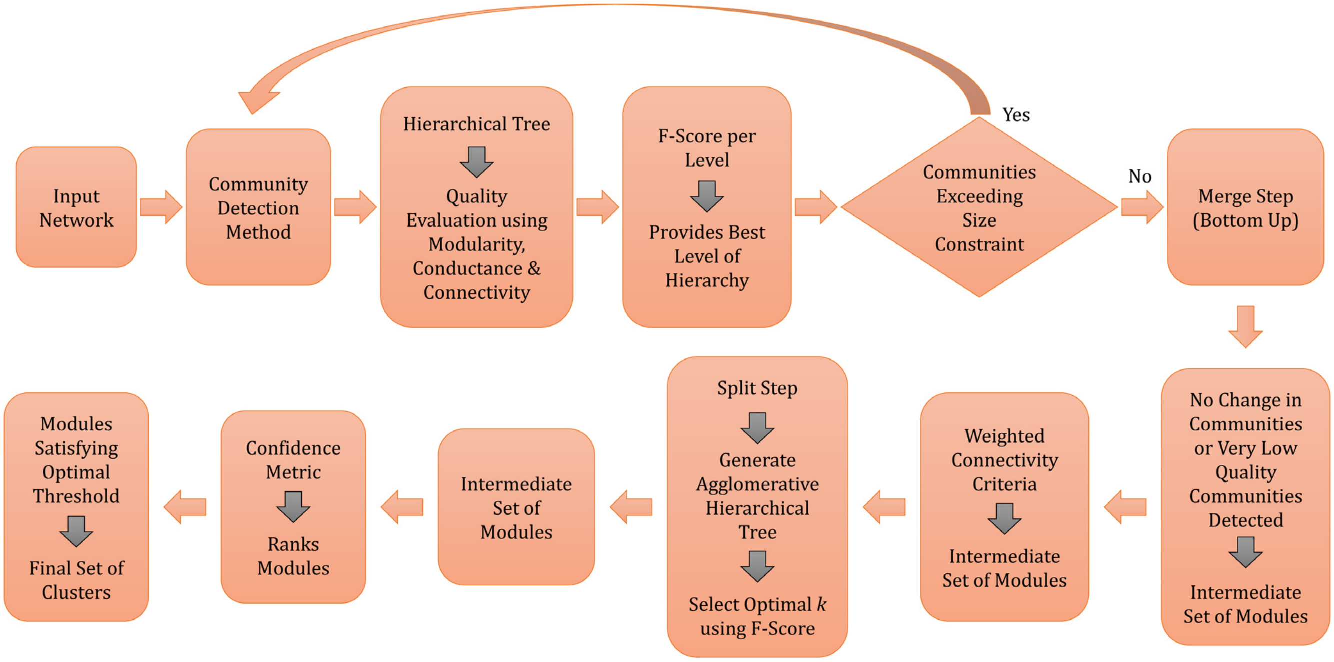
Input Network
Community Detection Method
Hierarchical Tree
Quality Evaluation using Modularity, Conductance & Connectivity
F-Score per Level
Provides Best Level of Hierarchy
Communities Exceeding Size Constraint
Yes
No
Merge Step (Bottom Up)
No Change in Communities or Very Low Quality Communities Detected
Intermediate Set of Modules
Weighted Connectivity Criteria
Intermediate Set of Modules
Split Step
Generate Agglomerative Hierarchical Tree
Select Optimal k using F-Score
Intermediate Set of Modules
Confidence Metric
Ranks Modules
Modules Satisfying Optimal Threshold
Final Set of Clusters

Figure 1. Steps involved in proposed generic constrained disease module identification framework.

Methods

Data

The disease module identification methods were evaluated using 6 benchmark networks. Details of the networks are provided in Table 1.

Preprocessing

There are several preprocessing steps performed before the input network can be processed by the pipeline. The node IDs are mapped to a continuous set of integers starting from 1. If the aforementioned procedure is not performed, the network will end up with several isolated nodes and missing IDs. All the edge-weights in each network are normalized between 0 and 1. The input network are considered weighted and undirected in all our pipeline.

We experimented with removal of edges with a weight lower than a threshold $t = 0.05$ but observed that the corresponding results deteriorated. Hence, we recommend keeping all the edges in the network.

Preliminary experiments

In the initial submission rounds, we ran several out-ofthe-box state-of-the-art community detection techniques including Order Statistics Local Optimization Method (OSLOM)[4], Louvain[5], Multi-level Hieararical Kernel Spectral Clustering (MHKSC)[6,7], Dynamic Tree Cut[20] and METIS[10]. We also tried to use the results obtained from these methods as input to consensus clustering based method PCAgglo[21] and ensemble clustering based method Ensemble-Clue[22] which are evaluated using complex traits and disease modules in 76 European GWAS datasets.

OSLOM is based on the local optimization of a fitness function expressing the statistical significance of communities with regards to random fluctuations, which is estimated with tools of extreme and order Statistics. The Louvain method is a greedy optimization method that attempts to optimize the modularity of a network partition. MHKSC technique uses a kernel spectral clustering formulation to random walk and exploits the structure of the projections in the eigenspace to automatically determine a set of increasing distance thresholds. It then uses these distance thresholds in a test phase to obtain multiple levels of hierarchy using principles of agglomerative hierarchical clustering. Dynamic Tree Cut method implements novel dynamic branch cutting technique for hierarchical clustering where it detects clusters in a dendogram depending on their shape. They are capable of identifying nested clusters, can identify clusters of various shape and are suitable for automation. METIS is a set of serial programs for multilevel recursive partitioning of the graph to produce fill reducing orderings for sparse matrices. PCAgglo performs logistic PCA on the concatenated node membership matrix formed from k different methods and then agglomerative hierarchical clustering is performed on the principal components. For METIS, Dynamic Tree Cut, PCAgglo and Ensemble-Clue, we selected that level of hierarchy for which the average module size was close to the best as per the exploratory data analysis provided by the DREAM Challenge organizers. The results that we obtained by direct application of out-ofthe-box state-of-the-art community detection methods is depicted in Table 2.

Insights gained

The Best of All result were not submitted during the preliminary rounds of the Challenge because the Best of All method depicts the maximum number of enriched modules that can be identified by a simple 'max' combination of these techniques at default settings. However, as per our understanding the goal of the challenge is to develop a method or a generic framework which can optimally identify disease modules from various gene and protein interaction networks at different parameter settings. We gained several insights from these preliminary results including:

- Methods like OSLOM, MHKSC and PCAgglo generated a set of clusters whose cluster size distribution is nearly power law.
- For most of these methods there were several giant connected components which were ignored due to the strict upper bound constraint on the module size.
- For most of these methods nearly half of the nodes in each network were part of giant connected components that were removed due to size constraint.
- METIS generated uniform sized clusters and included most of the nodes in each network, hence can't be optimized further.
- We didn't get more enriched modules from a consensus (PCAgglo) or ensemble (ensemble-clue) based clustering methods.

Table 1. Description of 6 benchmark networks used for evaluation.

ID	Directed	#Nodes	#Edges	Type	Edge-Weight
1_ppi	No	17, 397	2, 232, 405	Protein-protein interaction network	Confidence score
2_ppi	No	12, 420	397, 309	Protein-protein interaction network	Confidence score
3_signal	Yes	5, 254	21, 826	Signaling network	Confidence score
4_coexpr	No	12, 588	1, 000, 000	Co-expression network	Correlation
5_cancer	No	14, 679	1, 000, 000	Connects genes essential for similar tumor types	Correlation
6_homology	No	10, 405	4, 223, 606	Connects genes that are evolutionarily related	Confidence score

Table 2. Preliminary results using several state-of-the-art hierarchical module identification techniques. Comparison of several out-of-the-box community detection methods along with one consensus and one ensemble based clustering method for disease module identification on 6 different biological networks. Here *N* represents total number of candidate disease modules and n_s represents the number of significant/detected disease modules in the 76 genome wide association study (GWAS) datasets. OSLAM - Order Statistics Local Optimization Method, MHKSC - Multi-level Hieararical Kernel Spectral Clustering.

Method	N	n_s	1_ppi	2_ppi	3_signal	4_coexpr	5_cancer	6_homology
OSLOM	842	28	6	1	6	10	4	1
Louvain	833	29	6	5	7	1	5	5
MHKSC	707	31	11	3	3	5	6	3
METIS	1209	30	8	8	5	4	3	2
Dynamic Tree Cut	2118	24	9	4	6	3	0	2
PCAagglo	1803	24	8	3	2	1	4	6
Ensemble-Clue	756	21	9	4	3	3	1	1
Best of All	-	48	11	8	7	10	6	6

Notations

Let $G(V, E)$ be an undirected graph with $n = |V|$ representing number of nodes and $m = |E|$ representing number of edges. Let S be the set of modules (or a partition of the network), where n_s is the number of nodes in a module $s \in S$; m_s be the number of edges in s i.e. $m_s = |(u, v) \in E : u \in s, v \in s|$ and c_s be the number of edges on the boundary of s i.e. $c_s = |(u, v) \in E : u \in s, v \notin s|$ and $d(u)$ is the degree of node u.

Quality metrics

We provide summary of quality metrics used and definition of proposed quality metrics below:

1. **Modularity:** Modularity is a global metric which takes value between −1 and 1. It measures the density of links inside communities compared to links between communities. For a weighted graph, modularity of a network partition is defined as: $Q(S) = \frac{1}{4}\sum_{s\in S}(m_s - E(m_s))$, where $m_s - E(m_s)$ is the difference between m_s, the number of edges between nodes in s and $E(m_s)$, the expected number of such edges in a random graph with identical degree sequence. Modularity value ≤ 0 indicate that the corresponding partition behaves worse than a random partition of the network. Modularity score can only be obtained for graph partitions.

2. **Conductance**: Conductance is a local quality metric which can defined for each individual community in the network and takes values between 0 and 0.5. It is defined as: $CC(s) = \frac{m_s}{2m_s + c_s}$. It measures the fraction of total volume of edges associated with the nodes in a module $s \in S$ pointing outside the cluster. Conductance for a partition S can be calculated by taking an average of the conductance values for all modules $s \in S$.

3. **Connectivity:** Connectivity is a sub-local quality metric which can be defined for each individual node in the network and can be averaged for all the nodes in a module s (considering connectivity to other nodes in the same community) to get local connectivity metric. It can be averaged for all the modules $s \in S$ to obtain the global connectivity $CN(S)$ for a partition S. It was used in [1] to evaluate whether genes perturbed by trait-associated variants are more densely interconnected than expected in complex diseases and generate connectivity enrichment curves. The connectivity matrix K is defined as: $K = (I + \hat{W})^p$; with $\hat{W} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, where K is p-step random walk kernel used to define pairwise connectivity between the nodes, I is an identity matrix, W is the weighted adjacency matrix and D is the weighted diagonal degree matrix. We set $p = 4$ for our biological networks as it allows to capture all meaningful interactions for paths of length ≤ 4. The connectivity of a node i is estimated as $CN(v_i) = \sum_j K_{ij}$, connectivity of module s is $CN(s) = \sum_{i,j\in s}\frac{K_{ij}}{n_s^2}$, and connectivity of partition S as $CN(S) = \sum_{s\in S}\frac{CN(s)}{|S|}$. is Here $|\cdot|$ represents the cardinality function.

4. **F-score:** We need a quality metric which evaluates the quality of a partition using modularity, conductance and connectivity. While modularity captures global information, conductance and connectivity can capture local information. The proposed quality metric is defined as: $F(S) = CC(S)\frac{Q(S) + CN(S)}{2Q(S)CN(S)}$.

Higher value of modularity indicates better quality clusters, lower value of conductance leads to good quality communities and higher value of connectivity indicate better quality of modules. We need to maximize modularity and connectivity while minimizing conductance. Hence, we take harmonic mean of modularity and connectivity in the denominator of F-score metric to give importance to both of the quality metrics. Thus, with conductance in the numerator, the minimum value of F-score corresponds to the partition S with best quality cluster. However, if modularity value is ≤ 0 then we set F-score to a very large value which depicts the poor quality of the partition.

5. **Inverse Confidence**: We need a metric to rank all the modules generated from the proposed framework. We first considered the average connectivity metric $CN(s)$ for a community s. However, the connectivity criterion prefers smaller size modules which tend to be more cliquish than bigger modules. We also considered using the conductance $CC(s)$ of a community s to rank all the modules in partition S. However, conductance value decreases as size of the community increases due to larger volume of the module (which is denominator of $CC(\cdot)$). We propose an inverse confidence metric to rank all the communities in a partition S as: $IC(s) = \frac{CC(s)}{CN(s) \times n_S^2}$. We utilized the Inverse Confidence metric in conjugation with modularity to remove out less meaningful communities as illustrated in Figure 2 and explained within the proposed framework. We finally convert the inverse confidence value of each module into a confidence score as: $\mathrm{Conf}(s) = 1 - \frac{IC(s)}{\arg_S \max(IC(s))}$, where the denominator is used for normalization.

Proposed generic framework

We followed the steps indicated in Figure 1 to build the proposed framework for constrained disease module identification.

1. Given an input network we perform the preprocessing step to create a modified input network where the node IDs are monotonically increasing, edge weights are noramlized, and the network becomes weighted and undirected.

2. Run a state-of-the-art hierarchical community detection technique to generate the hierarchical tree structure.

3. Estimate quality of each level of hierarchy using modularity, conductance and connectivity.

4. Select that level of hierarchy for which the F-score is minimum.

5. For communities of size > 100 go to Step 2 until either the constraint exceeding communities cannot be split further or modularity of resulting cluster memberships becomes very poor.

6. In the merge step, we start with the partition (S) at the best level of hierarchy and traverse the hierarchical tree from that level in a bottom-up fashion. We iteratively merge those communities whose weighted mean connectivity score is less than the connectivity score for a module at next level of hierarchy where the module consists of those previous communities i.e.a. Here p an q are modules at level $h - 1$ and s is community at level h such that $p, q \in s$. This results in an intermediate partition set or a set of modules.

7. We then consider all the communities s s.t. $n_s > 100$. For each such community s, we consider the sub-graph comprising only the nodes from that community. We transform the corresponding weighted adjacency matrix i.e. $\hat{W} = W(v_i,v_j)$, $\forall v_i, v_j \in s$ into a distance matrix $D_{\hat{W}} = 1 - \hat{W}$. We then build the agglomerative hierarchical tree using the linkage clustering with Ward's distance.

8. For each community s ($n_s > 100$), once we obtain the agglomerative hierarchical tree, we cut the tree for different values of k i.e. the number of clusters. We

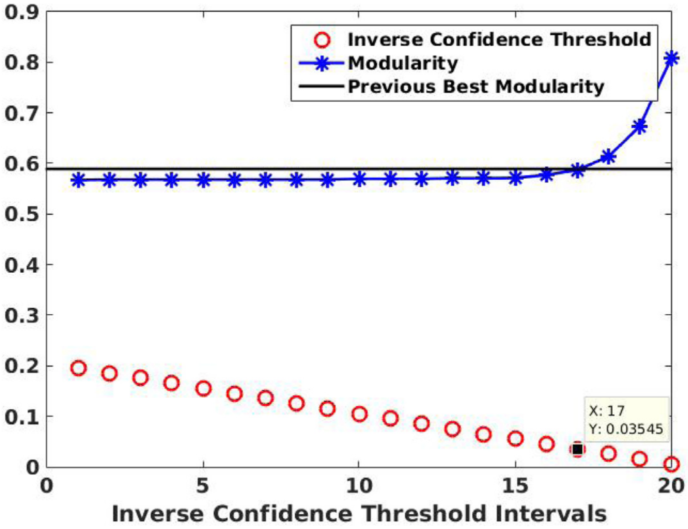

Figure 2. Figure 2 showcases the modularity values for different partitions obtained at various inverse confidence thresholds for network 3_signal. Here we also highlight the optimal inverse confidence threshold value.

evaluate each such partition using the F-score and select that partition which has the minimum positive F-score.

9. Using Steps 6–7 on these bigger modules and the small size communities which satisfy the size constraint, we generate another set of intermediate clusters.

10. We rank this intermediate set of communities using the inverse confidence score i.e. $IC(s)$, $\forall s \in S$. Lower inverse confidence corresponds to higher rank. We now remove all modules whose size exceeds the size constraint i.e. $n_s \leq 3$ and $n_s \geq 100$.

11. In this final step, we propose a mechanism to select the best set of modules for evaluation in an automated fashion independent of the network. We can calculate the maximum and minimum value of inverse confidence (IC) from the inverse confidence (IC) scores of all the communities in the intermediate partition S. We iteratively decrease the inverse confidence threshold from maximum to minimum thereby pruning clusters. At each such threshold, we calculate the modularity of the remaining set of partition using the subgraph corresponding to this partition S' i.e. GS'. We select the threshold where the difference between $Q(S', \theta)$ and Q_{prev} is minimum i.e. $\arg_{\theta} \min|Q(S', \theta) - Q_{\text{prev}}|$. Here $|\cdot|$ represents the absolute value, Q_{prev} is the modularity of the partition obtained at Step 2 and calculated in Step 3. For the final submission, we consider all the modules in the optimum partition i.e. $s \in S'$ obtained by pruning communities whose $IC(s) \geq \theta$.

Results

For our final submission, we utilized the method which is the fastest and most suitable for hierarchical graph partitioning i.e. Louvain method[5] as we were allowed to make only 1 submission. We formulated a recursive version of Louvain method where communities of size greater than 100 were recursively partitioned. We also designed a constraint satisfying version of MHKSC[6,7] and compared its performance with the recursive Louvain method within the proposed generic framework. The evaluation criterion used in the Challenge was the total number of significant modules identified in the 6 benchmark networks on a hold-out set of 104 GWAS datasets at the false discovery rate (FDR) cut-off[23] of 0.05 for multiple testing. We compare the results obtained from proposed generic framework using both the Louvain and MHKSC methods with the winners of the DREAM Challenge in Table 3.

From Table 3, we observe that the winners (Double Spectral Clustering and Resolution Adjusted Clustering) perform far better than Constrained Louvain method on the protein-protein interaction networks (Networks 1 and 2) and homology network (Network 6). However, for the signaling, co-expression and the cancer networks (Networks 3, 4 and 5), proposed Constrained Louvain method has comparable performance with the winners of the challenge. To gain a sense of the robustness of the ranking with respect to the final GWAS data, the challenge organizers sub-sampled the hold-out set by drawing 76 GWASs (same number as during the preliminary phase) out of the 104 GWAS datasets. They created 1, 000 subsamples of the hold-out set. The methods were then scored on each subsample (Sub-sampling was done here without replacement.)

The performance of each competing method t for a given network was compared to the highest scoring method across the sub-samples by the paired Bayes factor B_t i.e. the method with the highest score on this network in the hold-out set (all 104 GWASs) was defined as reference. The score $n_s(t, b)$ of method t in subsample b was thus compared with the score $n_s(ref\ , b)$ of the reference method in the same subsample b. The Bayes factor B_t is defined as the number of times the reference method outperforms method t, divided by the number of times method t outperforms or ties the reference method over all subsamples. Methods with $B_t < 4$ were considered a tie with the reference

Table 3. Final submission results comparing the winners with proposed generic framework. Here the proposed generic frameworks are referred as Constrained Louvain and Constrained Multi-level Hieararical Kernel Spectral Clustering (MHKSC) and we use * to represent the winners of the competition. Here N represents total number of candidate disease modules and n_s represents the total number of significant disease modules identified in the104 genome wide association study (GWAS) datasets. In the final round of the challenge, we submitted the results corresponding to Constrained Louvain method.

Method	FDR Cutoff	N	n_s	1_ppi	2_ppi	3_signal	4_coexpr	5_cancer	6_homology
Double Spectral Clustering*	0.05	2407	60	16	13	9	12	5	5
Resolution Adjusted Clustering*	0.05	2780	60	19	11	5	14	7	4
Constrained Louvain	0.05	1965	42	12	3	7	15	5	0
Constrained MHKSC	0.05	2108	37	5	3	4	18	4	3

method (i.e., method *t* outperforms the reference in more than 1 out of 5 subsamples). For networks 3, 4 and 5, the Bayes factor of proposed Constrained Louvain method was less than 4. This indicates that the proposed generic framework, though not the winner, is useful, generic and robust enough for identification of statistically significant disease modules in biological networks.

With the availability of the de-anonymized version of the networks along with the scoring tools used during the competition, we were able to perform additional experiments for the Constrained Louvain method. After the challenge, we identified an error in labeling the nodes in the significant disease modules that we submitted for the homology network (Network 6) during the competition. After correcting the labeling error, we identified 2 significant disease modules from Network 6.

Moreover, we performed additional analysis using 5 different FDR cut-offs (multiple testing) for each of the 6 benchmark networks to obtain the trends in the number of significant disease modules identified by the proposed generic framework for these cut-offs. This result is depicted in Figure 3. The FDR cut-off used as evaluation criterion during the competition was 0.05.

Discussion

The DREAM Challenge organizers made the GWAS datasets along with de-anonymized networks available to the challenge participants. This allowed us to further analyze our results. For each benchmark network, we identified the proteins or genes that make up the significant disease modules.

We investigated association of identified disease modules with disease/trait of the provided GWAS datasets. We used the official competition FDR cut-off of 0.05 as significance threshold to identify disease modules for each benchmark network. Table 4–Table 9 provides a detailed analysis of the significant modules and their corresponding associated disease (inferred from 104 hold-out GWAS datasets) for Networks 1,2,3,4,5 and 6 respectively. Each module is found to be associated with at least two GWAS datasets of the corresponding disease/trait. Moreover, many modules were found associated with multiple disease/trait

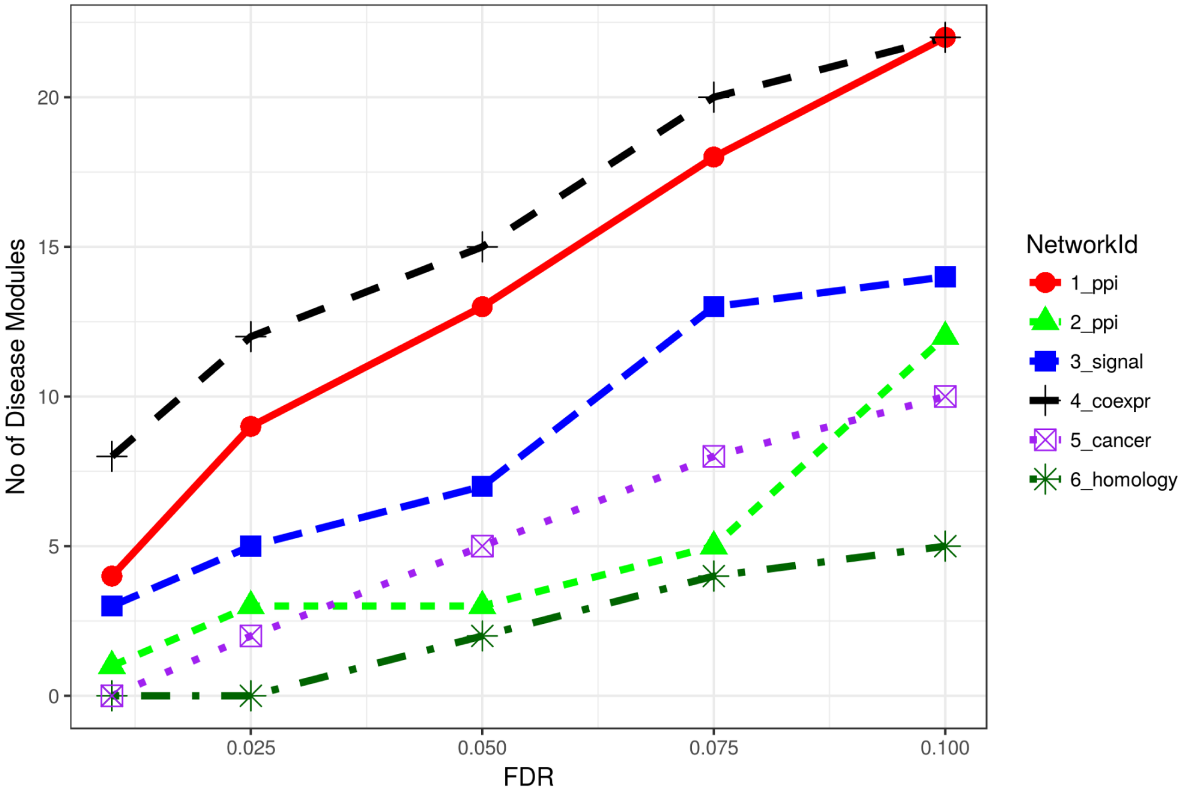

Figure 3. Number of disease modules identified by Constrained Louvain method for different false discovery rate (FDR) cut-offs for 6 benchmark networks.

Table 4. Significant disease/trait modules identified for 1_ppi network by proposed Constrained Louvain method.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
19	Hip Circumference(3), Human Height(4), Waist Circumference(3)	C5orf24, LOX, FBN1, ADAMTSL2, ECM1, FBLN5, MFAP5, EFEMP2, MFAP3, ELN, LTBP2, FBN2, MFAP4, ADAMTSL5, PXDNL, MFAP2, FBLN1, PRSS35, LOXL1, FBLN2, EFEMP1, PXDN
54	Ulcerative Colitis(2)	AMIGO2, AMIGO1, AMIGO3
56	Coronary Artery Disease(2)	C17orf103, ADAMTS5, CST2, ADAMTS7, DSCR8, POFUT2, B3GALTL, ADAMTSL1, ADAMTSL4, CFP
57	Lipid Levels(2)	NCAN, HPSE, APOE, B4GALNT4, IDUA, MSR1, CHST15, APOC4, NDST2, B3GAT3, LRPAP1, CHPF, TMCC3, B3GAT2, HS3ST6, CHSY1, B3GALT6, GPC2, CSPG4, CLEC2L, CSGALNACT1, GXYLT1, HS3ST2, VCAN, PLD5, GPC3, B3GAT1, SDC4, CHST3, APOA4, CHPF2, CSGALNACT2, APOC2, SLAMF7, LRP8, PON3, RBP1, LDLRAP1, KAL1, CHST13, GPR144, SLC35D2, B4GALT7, CHST11, CHST7, HS3ST4, HS3ST3B1, APOB, GPR111, NDST1, CC2D1A, LRP1, BCAN, CSPG5, XYLT2, DSE, LACRT, SDC3, NKG7, SDC1, HS6ST2, GLCE, GPC4, SNX17, TSPAN1, MTTP, HS6ST1, GPC5, ITGB1BP1, HS3ST1, HS3ST5, AGRN, IGSF9, HSPG2, SDC2, GPC1, B4GALNT3, EXT1, APOC3, CHSY3, CHST14, A2ML1, UST, MDK, GPR97, HPSE2, GPC6, HS3ST3A1, XYLT1, LRP2, PTN, TMCC2, LDLR, CHST12, EXT2, TLL2
145	Coronary Artery Disease(2)	HSD3B1, STS, SOAT1, CYP27A1, CYP11A1, CYP11B2, SULT2B1, CYP3A7, DHCR24, LIPA, TM4SF4, CYP11B1, CH25H, CYP46A1, CYP7B1, SUSD4, SOAT2, SLC27A5, CYP1A2, HSD3B2, ALS2CR12, CYP7A1, CYP17A1, FDX1, FDX1L
154	Crohn's Disease(2), Rheumatoid Arthritis(2), Ulcerative Colitis(2)	IL24, IL26, LEPR, IFNAR1, OSMR, IL22RA2, IL23A, RELB, IL28B, IL20RB, CSF2, IFNK, IL7, IL10, CBLC, IL21, IFNGR1, MPO, IL4, IL12RB1, IL19, TBX21, IL15RA, IL5RA, IL9R, IL2RA, SOCS4, SOCS7, CSH2, IL7R, IL3, IL28A, IFNAR2, EPO, EPOR, IL28RA, OSM, IL10RA, IL9, IL3RA, LEP, IL5, IL13, IFNW1, CTF1, IL13RA1, IL11, IFNA13, IL21R, IFNA1, JAK2, GH2, IL15, IL13RA2, IL10RB, CNTFR, IL20, CSF2RB, TSLP, IFNE, CSH1, IL12A, IL12RB2, CNTF, DOK2, IL2RG, CSF3, IL11RA, IFNGR2, CLCF1, LIFR, IL23R, IL6R, IL20RA, CSF2RA, IL2RB, IL29, IL6ST, IL2, CRLF2
157	Ulcerative Colitis(2)	PDCD1LG2, CD3D, CLEC2D, HLA-DRB1, CD274, SLA2, TREML2, PDCD1
174	Lipid Levels(2)	PVRL4, CD226, PVRL1, CADM3, PVRL2, TIGIT, PARD3, PVR, PVRL3, CD96
176	Narcolepsy(2), Rheumatoid Arthritis(2), Ulcerative Colitis(2)	HLA-DQB1, HLA-DQA2, HLA-DMB, HLA-DPA1, HLA-DPB1, HLA-DRA, HLA-DQB2, HLA-DOA, MS4A1, HLA-DMA, HLA-DQA1, HLA-DOB
184	Lipid Levels(5)	TBL1X, NCOA1, HELZ2, TBL1XR1, NR1I3, GPS2, CARM1, POU1F1, G0S2, HMGCS1, GLIPR1, SMARCD3, NCOA6, MED1, PPARA, NCOA2, TGS1, CTGF, CHD9, HMGCR, PEX11A, SULT2A1, GRHL1, NR1I2, NRBF2, HMGCS2, FADS1, SREBF2, DLX2, PLIN2, CPT2, CPT1A, RGL1, APOA5, SLC27A1
211	HbA1C(2)	SP110, RNF112, WDR73, TRIM73, SSRP1, TRIM17, CAPNS2, FN3K, HIST2H4A, ERMAP, PEF1, MLL5, GCA, ZNF618, TRIM69, TRIM60, ATAD2B, KDM5A, TRIM66, TRIM68, SUPT16H, HIST1H4K, DIDO1, TRIM72, SP140, FSD2, RFPL4B, SDR39U1, SLC20A1, TRIM58, HIST1H4I, SH3RF3, H3F3C, HIST2H3D, TRIM44, TRIM31, TRIM49B, TRIM51, WDR76, TRIM39-RPP21, TRIML1, HIST1H4B, TRIM43, KDM5D, TRIM49C, TRIM74, TRIM34, HIST2H3A, SLC20A2, HIST1H3G, TRIM4, HIST1H4G, NHLRC4, HIST1H3F, SP140L, RFPL1, CHD1, RNF39, PYROXD2, TRIM6-TRIM34, HIST1H3H, SRI, TRIM15, C16orf11, TRIM10, HIST2H3C, HIST1H3I, TRIM49, TRIM40, TRIM26, PHF20L1, RNF186, BRD4, TRIM64C, TRIM7, MEFV, TRIM52, HIST1H3B, RNF135, HAT1, SETD7, WDR59, ATAD2, KDM5C, FN3KRP, HIST1H4L, TRIM64B, TRIM48, TRIML2, HIST4H4, TRIM41, RFPL2
251	Psychiatric Disorders(2)	DAPK2, MKNK2, CDKL1, MAPKAPK3, CAMK1G, TSSK4, CAMK1, CDK7, STK32C, STK32A, TSSK1B, NEK3, STK16, TSSK6, MKNK1, PKMYT1, NEK7, CAMK1D, SBK1, MOS, PIM2, CDK10, STK33, NUAK2, ITIH3, MAP2K2, CDK6, PTK6, PSKH1, CDKL2, PIM1, OXSR1, PIM3, STK17A, NEK6, NUAK1, PSKH2, ULK3, CDKL4, PDIK1L, PNCK, SBK2, STK17B, PAK2
252	Lipid Levels(2)	ANKRD61, ANKRD65, ANKRD39, ASB11, ASB9, ASB12, ASB7, ACBD6, ASB1, RFXANK, ASB13, ASB14, ANKRD7, ANKRD49, ANKRD54, ASB4, ANKRD29, ASB3, CDKN2C, ANKRA2, CDKN2B, ANKRD30BL, ANKDD1A, ASB8, ANKRD16, OSTF1, ASB10, FANK1, ANKRD23, ANKRD44, CDKN2D, ANKRD1, ANKK1, ANKRD46, ANKRD22, ANKRD52, ASB5, GABPB1, BCL3, PPP1R27, NFKBID, ANKDD1B, ANKRD2

Table 5. Significant disease/trait modules identified for 2_ppi network by proposed Constrained Louvain method.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
81	Human Height(3), Waist Circumference(2)	NPPB, NPR1, NPR3, NPR2, NPPC, NPPA
109	Myocardial Infarction(2)	SMPD1, GALK2, C12orf4, SLC7A5, GDAP2, MRPS33, RAB23, C7orf43, C14orf50, DPEP2, CARS2, TMEM50A, SRFBP1, PLBD2, LANCL2, C4orf29, MPND
144	Narcolepsy(2), Rheumatoid Arthritis(2)	CD74, TRBV7-9, CD3D, TRDV2, HLA-DMA, HLA-DPB1, HLA-DQB1, HLA-DQA1, HLA-DPA1, TRAV29DV5, CD3E, HLA-DOA, CD3G, TRBV19, HLA-DRB4, HLA-E, TRAV8-4, TRBV12-2, TRAV19, CD247, HLA-DQA2

Table 6. Significant disease/trait modules identified for 3_signal network by proposed Constrained Louvain method.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
1	Coronary Artery Disease(2), Lipid Levels(9), Myocardial Infarction(2)	PCSK9, LDLR, APOB
114	BMI(2), Weight(2)	UBE3C, TLR2, TLR1, TLR10, SFTPA1, PSMD2, CYBB, NEU1, TLR6, DHX36, TRAP1
162	Age-related Macular Degeneration(2)	LTB, LTBR, TNFSF14, TNFRSF14
258	Age-related Macular Degeneration(2)	C3, CD46, C3AR1, CFB, CR1, CFI, CFH
284	BMI(2)	THPO, MPL, ATXN2L
331	Fasting Glucose(7)	GCK, GCKR, DUSP12
337	BMI(2)	BCL2, BCL2L1, CISD2, TMBIM6, TP53AIP1, ITM2B, SPNS1, HRK, BCAP31

Table 7. Significant disease/trait modules identified for 4_coexpr network by proposed Constrained Louvain method.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
3	Hip Circumference(4), Neuroticism(2), Psychiatric Disorders(7), Waist Circumference(2)	HIST1H2BE, HIST1H2BC, HIST1H3G, HIST1H4A, FLJ13224, HIST1H3A, OR2B6, HIST1H2BG, HIST1H2BH, HIST1H1D, HIST1H2BI, HIST1H2BF, HIST1H4H, HIST1H2AG, HIST1H2BB, HIST1H2BL, HIST1H2BN, HIST1H3H, HIST1H4E, HIST1H2AJ, HIST1H2BM, HIST1H2BO, HIST1H2BJ, HIST1H4F, HIST1H2AE, HIST1H2AK, HIST3H2A, HIST1H3J, HIST1H2AI, HIST2H2AA3, HIST1H4D, HIST2H2BE, HIST1H4B
39	Lipid Levels(4)	GLB1, C11orf75, VPS37C, CSF1R, TNFSF13, BLVRB, SH2B3, SCPEP1, NEU1, RNPEP, CFD, BLVRA, MAN2B1, PION, M6PR, KIAA0930, PLEKHB2, NSMAF, FCGRT, LRPAP1, CAPG, SAMHD1, SIDT2, MGAT1, FKBP15
48	Psychiatric Disorders(2)	COCH, B4GALT6, TCEAL4, S100A6, CHL1, TSPO, PIK3R3, CXXC4, CAPN2, PIP5K1B, YES1, GAS2, TRIP6, SLC20A1, TUBB2B, C18orf1, ATP2A3, GLDC, ANXA2P2, HMHB1, ATP8A1, C20orf103, ACAP1, LRRC59, ENPP4, CTNNAL1, ADAM9, CD200, EMID1, GSTM3, VEGFA, LAPTM4B, LIMA1, FXYD2, GGA2, S100A4, DAPK1, S100A11, ITGAV, PARM1, SIDT1, CYFIP1, MGAT3, CEBPB, REXO2, KL, BAG2, IKZF2, WBP5, JAG1, QPRT, VAMP3, PLP2, NCALD, CNIH4, FCGR2B, MXRA7, ASCL1, GNG7, TAX1BP3, PLS3, ARHGAP6, ANXA1, IPCEF1, REC8, BCL2, DMD, EPHB1, MT2A
53	Crohn's Disease(2), Rheumatoid Arthritis(3)	DNAH17, IL2RA, IL3RA, ICOSLG, COL9A2, CIITA, GRAP, IL21R, TEC, TNFRSF4, POU2F2, TCL6, PAOX, IKZF3, STOM, BTK, LILRB1, LILRB4, ADAM28, KMO, SLC2A5, GPR65, SH2D3C, ST6GALNAC4, CD86, SLC15A2, PCMT1, UGT2B17, ABCB4, PTPN7, GATM, PPFIBP2, DOK3, KLK2, VNN2, ADAM29, TLR7, STS, BTN2A2, TNFSF11, HLA-DRB4, SH3D21, LY9, FGD2, GH1, PHACTR1, HSPA1A, HCG26, ALOX5, LOC100505650, HLA-DOA, SCARF1, LTA, TTN, DNASE1L3, CNR1, CXorf21, ZNF318, PDE6G, TNFRSF9, P2RY10, OAT, RASGRF1, IL7, AKAP5, IGHV5-78, CXCR5, SLC9A7, PFN2, IRF5, TRAF3, TNFRSF13B
56	Rheumatoid Arthritis(3)	FLT3LG, TXK, ACSL6, BACH2, ANKRD55, CDKN2D, CCR9, S1PR1, CCR5, HSPB1, ANGPT4, LIME1, CD96, CD28, TNFRSF25, UBASH3A, GPR171, GPA33, CCR4, SIRPG, TNFSF8, XCL1, CD8B, KLRD1, PVRIG, STAT4, TBX21, TRAV8-3, FASLG, TRD@, CD7, RORA, GFI1, CXCR6, SH2D1A, LOC79015, CAMK4, LAG3, IL23A, LRRN3, SPINK2, TRAT1, KLRG1, IFNG, EMR1, SH2D2A, CD3G, CHMP7, KCNA3, CD6, MGST3, GZMM, ICOS, CD5, SLAMF1, PTPN4, CCR8, PDCD1, TRBC2, LOC100507397, RCAN3, TRBV10-2, OCM2, SIT1, PRKCQ
92	Lipid Levels(5)	SNTA1, DEXI, KLHDC3, PHKB, EEF1A1, MID1IP1, SLC25A11, OGDH, VPS4A, GSTM2, MAP7D1, SCN1B, CARM1, KEAP1, USO1, GSTM1, POLDIP2, PHKG1, VPS52, FAM89B, GPS2, TRIP10, SLC2A4RG, FHOD1, CTDNEP1, ARL2, RNF123, UQCC, LRRC20
99	Lipid Levels(4)	C9, F7, CRP, TMPRSS6, TAT, SLC17A2, HMGN2, LECT2, MASP2, C19orf80, TRMT5, LIPC, ABCG5, APOF, SPP2, CFHR5, FGF21
104	Lipid Levels(3)	CYP7A1, GCKR, CLEC4M, PKLR, CRYAA, PRG4, DDO, IGFALS, LPA, FTCD, FN3K, C14orf105, SEC14L4, F13B, MASP1, CLDN16, CPN2, ART4, ADRA1A, FOLH1B, HGFAC, HAAO, FOLH1, MBL2, SLC7A9, DNMT3L, MLXIPL, CA5A, ABCG2, FETUB, LPAL2, CYP3A43, CCL16, F11, GPER, SARDH, HNF4A, GPLD1, CPS1-IT1, NAT8, SLC38A3, APOA4, ONECUT1, EPO, SHBG, HNF1A, SLC26A1, MBNL3, UPB1, NR1I3, ALDOA, RHBG, PON1, CPN1, CCNI, CYP2C19, PROZ, TTPA
107	Lipid Levels(5)	C3orf32, SERPINA6, ADH6, SULT2A1, SERPINA4, C4BPA, RGN, C8A, PLG, UGT2B4, SERPINF2, PGC, SERPINA10, ITIH1, HPR, MTTP, PROC, ANGPTL3, AKR1D1, MAT1A, BHMT

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
109	Psychiatric Disorders(2)	ENDOU, IL37, WNT3, DNASE1L2, KCNK7, KRTAP2-4, KRTAP9-9, KRT83, KRTAP1-3, BPY2, KRT35
126	Leptin(2)	HSPB7, CCDC48, HSPB2, BAALC, CSPG4, SLC16A4, MAP1A, SGCA, CSDC2, DNAJB5, NFASC, FHL5, PLEKHA4, STK32B, DAAM2, TRO, SPEG, ADAMTSL3, TMEM100, CLIP3, CACNA1C, TBX5, GPC4, SLC26A10, GREM2, LTBP4, C8orf84, RRAD, EMILIN1, RAB23, HSPB6, HSPA12A, C7orf58, TACR2, ADAMTS8, ITGA1, CYTL1, SLC2A10, SCN7A, ARHGAP24, GPM6A, PRKG1, RAB40A, NBLA00301, SCRG1, HSPB3, SNAI1, AGTR2, IL17B, BEX1, SGCD, PER1, PKNOX2, CHRM2, FGF7, PDE5A, SMAD9, ENOX1, PGM5, NDNF, HPSE2, ARNT2
135	Waist Circumference(3)	TGFB1I1, FBLN1, TJP1, AXL, CAV2, COL5A1, TIMP3, FBN1, WWTR1, TPM2, UBE4A, LOXL2, OLFML3, FAP, PCOLCE, NUPR1, CTGF, LTBP2, SEPT10, MFAP2, TNC, FN1, PRSS23, PXDN, CDKN1A, CALD1, NID1, TMEM47, LOXL1, MRC2, PPAP2B, FBLN5, PPIC, IL1R1, LARGE, MYO1B, LHFP, MYL9, NID2, LOX, FLRT2, RASL12, C6orf145, OLFML2A, SNAI2, LAMB1, THBS1, PPAP2A, EFEMP1, DSE, ENAH, MAP1B, IGFBP3, DKK3, F2R, ADAMTS1, FERMT2, ARHGAP29, CDH11, MYLK, MYOF, COL1A1, NNMT, COL5A2
138	Narcolepsy(2), Rheumatoid Arthritis(3)	HLA-E, CTSC, KRT19, LAP3, LYZ, HLA-G, HCLS1, LCP1, HLA-DPA1, UCP2, TAPBP, RAC2, HLA-B, GRB10, LYN, SH3BGRL, EIF2C2, LIPA, GRB14, CD74, CNDP2, HLA-F, LAPTM5, MYD88, DLG5, HLA-DRB1, HLA-C, TRIM22, HLA-DPB1, CD53, SRGN, HLA-DMA, LGMN, IFI30
184	Obesity(2)	CINP, NDOR1, FAM158A, ZMAT5, HLCS, SURF2, KCTD2, LIN37, TELO2, C4orf10, ZNF408, CCDC22, COQ6, BAD, C17orf59, RNF25, LIN7B, TBL3, TUG1, RPS6KB2, C21orf2, PIGH, SART1, BRF1, TMEM110, AAGAB, AZI1, SSSCA1, ZNHIT2, NUDT2, PGP, TMEM104, ROM1, ARMC7, MKL1, AKIP1, SUGP1, GTF3C5, E4F1, PPP2R2D, C2CD2L, ETV2, NADSYN1, NUBP2, LOC100129250, C11orf51, WDR25, GPKOW, KCTD17, TMED1, BCL7C, THAP7, NOC4L, TBCD, EXOC3, GNB1L, FAM3A, KLHDC4, NKIRAS2, OPHN1, PIN1, FAU, SNAP29, COMMD9, PUM2, C17orf90, FAM3C, C16orf42, SHARPIN, BNIP1, TXNRD2, PIN1P1, ZNF839, CCDC101, DHRS7B, PANK3, PRMT7, WDR13, DDX49, TMEM11, ASPSCR1, TSR2, ZFPL1
187	Crohn's Disease(2)	ARFGAP1, PCGF3, TAF1C, RTEL1, MUS81, BRD9, CDK10, SH3BP2, INPP5E, C19orf54, ABCC10, SPG7, MAN1B1, DOM3Z, RAD9A, CEP164, NFRKB, MST1, CLASRP, NELF, TJAP1, ASXL1, SLC35C2, TXLNA, PLXNA3, SZT2, SFI1, ATG4B, ASAH1, INPPL1, FAM193B, CUL9, APBA3, RHOT2, SKIV2L, MDC1, RBM14, PAQR6, SLC26A6, FAM193A, PIGO, HTT, MOGS, C9orf86, MFSD10

Table 8. Significant disease/trait modules identified for 5_cancer network by proposed Constrained Louvain method.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
107	Neuroticism(2)	ARSH, ATP2B4, CTSB, MYT1L, MMP21, LIM2, DCSTAMP, KCNT2, ZNF442, T, POC1B-GALNT4, EXOC4, NFATC2, NOD2, SIM1, MYLIP, PGK2, C1orf109, FAAH2, ZNF436, TCF4, NEK1, CLIC2, TMEM206, DIAPH1, CYP4F22, MNDA, CLDN17, GALNT4, HELT, GNPAT, CNGA2, ZGPAT, PPP1R16A, HIVEP1, CSNK2A3, UQCRBP1, RUNX1T1, CYP24A1, ENPP6
181	Hip Circumference(2)	NT5C1B-RDH14, ALDH8A1, ZNF155, GNA12, B3GAT2, SULT1B1, HHIP, AAGAB, TMEM119, RDH14, DNAJA1
211	BMI(2), Obesity(2)	TACC2, BDNF, SLC16A3, AKAP10, PLS3, FAM19A3, PABPC5
656	Neuroticism(2)	LRRC37A4P, INPP5B, LRRC37A, VAMP3, LRRC37A2, LRRC37A3
666	BMI(4), Waist Circumference(2)	SULT1A3, SULT1A2, SULT1A1, SLX1A, SLX1A-SULT1A3, SLX1B, SULT1A4, SLX1B-SULT1A4

Table 9. Significant disease/trait modules identified for 6_homology network by proposed Constrained Louvain method after the challenge.

Module Id	Disease/Trait (Number of GWAS datasets)	Genes/Proteins
105	Coronary Artery Disease(2)	RNASE7, KDM2A, SLC24A1, ESRP2, ASNS, MARCH9, BZW1, EDDM3B, VPS13D, APPBP2, NAA16, THAP4, VWA3B, CYP3A43, FABP2
198	Lipid Levels(4)	SOCS3, NUDT19, AKR1C1, ZNF714, IGFBP6, CAT, ARHGAP32, PITPNC1, NFKBIL1, MAD2L2, EIF1B, LPO, ZNF620, TMEM204, DAND5, ARHGAP25, KRCC1, SP1, FEZF1, LMNTD2, OOSP1, TMED1, HOXA4, SLC36A4, FAM71F2, ASPM, FBXL20, OR5I1, HBG1, SFTPC, APOC4-APOC2, HEXB, ZNF521, TRIM56, CHPT1, IFT20, MLXIP, AJUBA, IDE, GMIP

of similar nature. For example, as shown in Table 4, module 19 in 1_ppi network is found to be associated with anthropometric traits. This indicates that the identified modules correspond to preserved biological functions of genes/proteins.

Data availability

The Challenge datasets for registered participants are available at: https://www.synapse.org/#!Synapse:syn6156761/wiki/400659. Challenge documentation, including the detailed description of the Challenge design, overall results and scoring scripts can be found at: https://www.synapse.org/#!Synapse:syn6156761/wiki/400647.

Source code for the proposed framework is available from: https://github.com/raghvendra5688/DMI/tree/DMI_v1.0

The archived source code for the proposed framework along with a README file can be found at: https://doi.org/10.5281/zenodo.1197424[24].

Competing interests

‘No competing interests were disclosed’.

Grant information

‘The author(s) declared that no grants were involved in supporting this work.’

Acknowledgements

We would like to thank Ms. Kanchan Karnani for helping out with preparing the flowchart (Figure 1).

References

1. Marbach D, Lamparter D, Quon G, *et al.*: **Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases.** *Nat Methods.* 2016; **13**(4): 366–370.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Newman ME: **Modularity and community structure in networks.** *Proc Natl Acad Sci U S A.* 2006; **103**(23): 8577–8582.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Jiang J, Wen S, Yu S, *et al.*: **The structure of communities in scale-free networks.** *Concurr Comp-Pract E.* 2017; **29**(14): e4040.
Publisher Full Text
4. Lancichinetti A, Radicchi F, Ramasco JJ, *et al.*: **Finding statistically significant communities in networks.** *PLoS One.* 2011; **6**(4): e18961.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Blondel VD, Guillaume JL, Lambiotte R, *et al.*: **Fast unfolding of communities in large networks.** *J Stat Mech.* 2008; **2008**(10): P10008.
Publisher Full Text
6. Mall R, Langone R, Suykens JA: **Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks.** *PLoS One.* 2014; **9**(6): e99966.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Mall R, Langone R, Suykens JA: **Furs: Fast and unique representative subset selection retaining large-scale community structure.** *Soc Network Anal Min.* 2013; **3**(4): 1075–1095.
Publisher Full Text
8. Mall R, Langone R, Suykens JA: **Self-tuned kernel spectral clustering for large scale networks.** In *Big Data, 2013 IEEE International Conference on.* IEEE, 2013; 385–393.
Publisher Full Text
9. Mall R, Jumutc V, Langone R, *et al.*: **Representative subsets for big data learning using k-nn graphs.** In *Big Data (Big Data), 2014 IEEE International Conference on.* IEEE, 2014; 37–42.
Publisher Full Text
10. Karypis G, Kumar V: **Metis-serial graph partitioning and fill-reducing matrix ordering.** 2012.
Reference Source
11. Dhillon IS: **Co-clustering documents and words using bipartite spectral graph partitioning.** In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2001; 269–274.
Publisher Full Text
12. Dhillon IS, Guan Y, Kulis B: **Weighted graph cuts without eigenvectors a multilevel approach.** *IEEE Trans Pattern Anal Mach Intell.* 2007; **29**(11): 1944–57.
PubMed Abstract | Publisher Full Text
13. Fortunato S, Hric D: **Community detection in networks: A user guide.** *Phys Rep.* 2016; **659**: 1–44.
Publisher Full Text
14. Parthasarathy S, Tatikonda S, Ucar D: **A survey of graph mining techniques for biological datasets.** In *Managing and mining graph data.* Springer, 2010; 547–580.
Publisher Full Text
15. Barabási AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet.* 2011; **12**(1): 56–68.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Califano A, Butte AJ, Friend S, *et al.*: **Leveraging models of cell regulation and gwas data in integrative network-based association studies.** *Nat Genet.* 2012; **44**(8): 841–7.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Mitra K, Carvunis AR, Ramesh SK, *et al.*: **Integrative approaches for finding modular structure in biological networks.** *Nat Rev Genet.* 2013; **14**(10): 719–32.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics.* 2008; **9**(1): 559.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Lamparter D, Marbach D, Rueedi R, *et al.*: **Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics.** *PLoS Comput Biol.* 2016; **12**(1): e1004714.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.** *Bioinformatics.* 2007; **24**(5): 719–720.
PubMed Abstract | Publisher Full Text
21. Asur S, Ucar D, Parthasarathy S: **An ensemble framework for clustering protein-protein interaction networks.** *Bioinformatics.* 2007; **23**(13): i29–i40.
PubMed Abstract | Publisher Full Text
22. Hornik K: **A clue for cluster ensembles.** *J Stat Softw.* 2005; **14**(12): 1–25.
Publisher Full Text
23. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc B Met.* 1995; **50**(1): 289–300.
Reference Source
24. Mall R: **raghvendra5688/DMI: Disease Module Identification (Version1) (VersionDMI$_v$1.0).** *Zenodo.* 2018.
Data Source

Open Peer Review

Current Referee Status: ? ?

Version 1

Referee Report 08 June 2018

doi:10.5256/f1000research.15518.r34227

? **Yunpeng Liu**
Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

In this paper, the authors describe a new pipeline for identifying disease modules from large-scale biological networks in the DREAM challenge. The pipeline builds upon off-the-shelf hierarchical community detection methods and first generates an initial partitioning of the network using a given community detection algorithm. It then integrates multiple properties of the network including modularity, conductance and connectivity into an F score to benchmark the partitioning at different levels of the hierarchy and selects the best partitioning. Next the pipeline merges modules in a way that increases connectivity, and resulting modules that exceed size threshold are partitioned again using hierarchical clustering and split at a level corresponding to the F score minimum. After a second round of merging similar to previous steps, a final set of modules are generated and ranked using a score termed inverse confidence. Using known disease-gene associations obtained from GWAS datasets, the authors verified the modules identified from their pipeline with multiple community detection algorithms, and compared performance across difference networks with that of the top team in the challenge. The authors conclude that despite the top performing team scoring highest overall, there are several cases where the number of modules identified in this paper are at least comparable to that from the top scoring method. Additionally, the authors claim that their pipeline is a generic framework for identifying statistically significant disease modules from biological networks.

The methodology put forward in this paper seems novel. However, neither the utility of the module identification pipeline nor its generalizability are adequately demonstrated. This is likely due to vagueness in the description of methods and lack of theoretical justification and supporting computational experiments to validate the procedures and scoring metrics devised by the authors. Specific comments are listed in detail below.

1. The description of the module identification framework seems elusive and lacks detail, rendering it unclear whether it is based on sound theoretical foundations. The framework works through a series of merge-split-merge steps that seems to hint at an iterative procedure to refine network partitioning. However, the method stops at the second merging step and discards all modules whose sizes exceed thresholds. The authors need to provide a rationale for this – is this because of empirical results that the modules identified from the pipeline do not change much after these steps and thus do not require further iterations in general? In addition, the paper seems to switch between methods and scoring schemes in different steps of the pipeline, for example the merging step is performed using increasing connectivity as criterion, whereas the splitting step uses a hierarchical clustering with minimum F score as partitioning criterion. At the final step, the authors used minimal change in modularity score as the criterion for setting a cutoff for module significance. These heuristics seem inconsistent with the idea of integrating multiple network properties to assess quality of network partitioning. Without justification for such a design the

framework it would not appear convincing enough to be a fair and generalizable pipeline for module identification.

2. The paper only compares their results with that from the top scoring team in the DREAM challenge and finds that there is a subset of networks where their performance is comparable to that of the top scoring method. This is rather inconclusive in terms of the method of comparison as well as the comprehensiveness of the datasets covered. Is the best scoring team's method (not explained in detail in the paper) applicable to the pipeline that the authors proposed? If so it would be more convincing if the authors could see improvement over the best team's method by refining modules through their pipeline. Is a single comparison adequate? Can the pipeline improve results from other teams that use hierarchical methods for module identification too? Answers to these questions will help further verify the utility of the proposed pipeline.
3. The authors gained some insight by inspecting results from each of the off-the-shelf methods they tested using their module identification pipeline – different methods behave very differently in terms of module number and size distribution. A question that naturally arises from these observations is what goes into a good module identification method? Instead of using real biological networks, the authors may gain a better view of the performance of these algorithms by benchmarking them against synthetic networks where modules are simulated under different generative models. Because these methods, including the novel pipeline in this paper itself, only takes network topology into consideration, it is important to test if they work well with different network models.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Referee Expertise: Transcription regulation, network biology, cancer biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 08 May 2018

doi:10.5256/f1000research.15518.r32972

? **Eric E Schadt**
Department of Genetics & Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

In their paper, "An unsupervised disease module identification technique in biological networks using novel quality metric based on connectivity, conductance and modularity", Mall *et al* present an approach to identifying modules of genes from different types of networks, where their approach uses a novel quality metric to evaluate the quality of the partitions based on a number of network metrics such as modularity, conductance and connectivity. Through a series of steps defining the module detection pipeline employed by the authors, they identify modules for the different types of networks and assess the "truth" of the module using enrichment metrics based on GWAS findings as defined by the DMI Dream competition in which the authors had submitted their approach and results.

The approach detailed by the authors seems reasonable, and the idea of the DMI to test a great variety of methods against one another is excellent. The processing of networks to identify biologically meaningful modules is still an important area of research and competitions such as DMI have helped to assess progress and identify best practices.

However, while the work presented is potentially interesting, as written the general DMI approach and specific results are difficult to understand and so hard to evaluate in light of this (as detailed below in the specific comments). Further, this paper represents a detailing of one of many approaches that were used in the DMI challenge, and shows results only compared to the winner of the challenge. The approach detailed apparently ranked 15th in the competition, and so was beaten by 14 other approaches. There is no discussion around this, no discussion on why a reader should care to know about one approach that ranked 15th compared to, say, the 19 other approaches that ranked in the top 20 (14 of which would have beaten the described method). There is no motivation provided on why knowledge of the authors' approach should be considered in light of 14 other approaches that beat it in the DMI competition. Do the authors believe the DMI competition was the best way in which to assess module identification methods and that the field should adopt the top-scoring methods for this as the state of the art? Do the authors believe that for the type of networks such as coexpression where their approach was comparable to the winners, that their approach has broader utility? Where the other top 13 methods similar with respect to performance across network types?

Specific Comments:

1. The paper is somewhat oddly written in that the methods section contain some methods along with some results and generally a failure to really describe the methods employed by DMI to compare methods.Without this the only way to understand the results in the paper is to invest much time going through the challenge description and the results, and so on. That is a huge burden placed on the reader. The components necessary to understand the results given in the paper should be described in the paper, and if there are references that describe fuller details, then that could be summarized in the methods so that the reader understands what was done and where to go for more details.(Further details on what is missing are given in the following comments.)
2. The authors given preliminary experiments done in the methods section, which ostensibly drove some thinking and refining on the approach they ultimately settled on.The preliminary experiments are not really methods, they are more results. And then there is an "insights gained" section in the methods, which again is not really a method but rather detail learnings from these earlier results.

3. While the authors do detail their own module identification process, the way in which the validity of the modules were assessed is not clearly articulated.What were the criteria set forth by DMI?How were the genes identified given a GWAS finding?There is error associated with identifying the vast majority of genes associated with a GWAS finding, so how was this handled?Was an enrichment score used for genes from GWAS being identified in the module?Was it per disease and combined over all diseases?Did effect sizes come into play? Etc.There should at least be a summary of this so that the reader can understand what it means to be able to count a module in the accuracy score for the competition.But there is nothing on this. The results speak to the paired Bayes factor that was used to compare methods, but you can't really understand the appropriateness of that without have an understanding of the above questions.
4. There is a link to the Synapse platform regarding the challenge, with many scores of pages of material and then a paper posted on biorxiv that provides details on the challenge and the findings.But it is not yet peer reviewed, it does not appear to be published yet, and so all of the missing detail in this present paper simply points to other papers that are not peer reviewed.It's again a pretty tall order to ask a reviewer to sift through endless pages of material to understand the context of a paper they have been asked to review and to then review on top of that papers upon which the paper they were asked to review is based.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: I note that I am co-founder and a board member of Sage Bionetworks, the institution that ran the challenge upon which the results of this paper are based.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.