METHOD ARTICLE

Identifying communities from multiplex biological networks by randomized optimization of modularity [version 1; referees: awaiting peer review]

Gilles Didier¹, Alberto Valdeolivas¹-³, Anaïs Baudot id¹,³

¹Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
²ProGeLife, Marseille, France
³Aix Marseille Univ, Inserm, MMG, Marseille, France

Abstract

The identification of communities, or modules, is a common operation in the analysis of large biological networks. The Disease Module Identification DREAM challenge established a framework to evaluate clustering approaches in a biomedical context, by testing the association of communities with GWAS-derived common trait and disease genes. We implemented here several extensions of the MolTi software that detects communities by optimizing multiplex (and monoplex) network modularity. In particular, MolTi now runs a randomized version of the Louvain algorithm, can consider edge and layer weights, and performs recursive clustering.

On simulated networks, the randomization procedure clearly improves the detection of communities. On the DREAM challenge benchmark, the results strongly depend on the selected GWAS dataset and enrichment p-value threshold. However, the randomization procedure, as well as the consideration of weighted edges and layers generally increases the number of trait and disease community detected.

The new version of MolTi and the scripts used for the DMI DREAM challenge are available at: https://github.com/gilles-didier/MolTi-DREAM.

Keywords

Biological Networks, Multiplex, Multi-layer, Community identification, Clustering, DREAM challenge

This article is included in the DREAM Challenges gateway.
Corresponding authors: Gilles Didier (gilles.didier@univ-amu.fr), Anaïs Baudot (anais.baudot@univ-amu.fr)

Author roles: Didier G: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Valdeolivas A: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; Baudot A: Conceptualization, Funding Acquisition, Investigation, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The project leading to this publication has received funding from the Centre National de la Recherche Scientifique (PEPS BMI IMFMG), the French “Plan Cancer 2009–2013”, and the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French “Investissements d’Avenir” programme.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Didier G et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Didier G, Valdeolivas A and Baudot A. Identifying communities from multiplex biological networks by randomized optimization of modularity [version 1; referees: awaiting peer review] F1000Research 2018, 7:1042 (doi: 10.12688/f1000research.15486.1)

First published: 10 Jul 2018, 7:1042 (doi: 10.12688/f1000research.15486.1)
Introduction

Biological macromolecules do not act isolated in cells, but interact with each other to perform their functions, in signaling or metabolic pathways, molecular complexes, or, more generally, biological processes. Thanks to the development of experimental techniques and to the extraction of knowledge accumulated in the literature, biological networks are nowadays assembled on a large-scale. A common feature of biological networks is their modularity, i.e., their organization around communities - or functional modules - of tightly connected genes/proteins implicated in the same biological processes\(^1\).

The Disease Module Identification (DMI) DREAM challenge aims at investigating different algorithms dedicated to the identification of communities, in a biomedical context\(^1\). The challenge has been divided into two sub-challenges, to identify communities either i) from six biological networks independently, or ii) from all these networks jointly. The clustering approaches proposed by the participants are assessed regarding their capacity to reveal disease communities, defined as communities significantly associated with genes implicated in diseases in GWAS studies\(^2,3\). The challengers proposed various strategies and clustering approaches, including kernel clustering, random walks or modularity optimization. We competed with an enhanced version of MolTi, a modularity-based software that we recently developed\(^4\). MolTi was initially developed to cluster multiplex networks, i.e., networks composed of different layers of interactions. It extended the modularity measure to multiplex networks and adapted the Louvain algorithm to optimize this multiplex-modularity. We have demonstrated that this multiplex approach better identifies the communities than approaches merging the networks, or performing consensus clusterings, both on simulated and real biological datasets\(^5\).

Grounded on these initial results, we here extended and tested our MolTi software, both on simulated data and on the DMI challenge framework. We improved MolTi with the implementation of a randomization procedure, the consideration of edge and layer weights, and a recursive clustering of the classes larger than a given size.

With simulated data, we observed that considering more than one network layer improves the detection of communities, as already noted\(^4\), but also that communities are better detected with the randomization procedure. With the DMI benchmark, we pointed to a great dependence on the GWAS dataset used for the evaluation and on the FDR threshold defined, but, overall, randomizations and edge and layer weights increase the number of detected disease communities.

Methods

**MoTi-DREAM: communities from multiplex networks**

We detected communities with an extended version of MolTi\(^5\), a modularity-based software. Although MolTi was specifically designed for multiplex networks, it deals with monoplex networks by considering them as multiplexes composed of a single layer. All the networks are here considered undirected. The new version of MolTi, MolTi-DREAM, and the scripts used for the DMI DREAM challenge are available at https://github.com/gilles-didier/MoTi-DREAM.

**Modularity.** Network modularity was initially designed to measure the quality of a partition into communities\(^6\), and subsequently used to find such communities. Since finding the partition optimizing the modularity is NP-complete, we applied the meta-heuristic Louvain algorithm\(^6\). Louvain starts from the community structure that separates all vertices. Next, it tries to move each vertex from its community to another, picks the move that increases the most modularity, and iterates until no change increases the modularity anymore. It then replaces the vertices by the detected communities and performs the same operations on the newly obtained graph, until the modularity cannot be increased anymore. In order to handle multiplexes, we use a multiplex-adapted modularity and an adaptation of the Louvain algorithm for optimizing this multiplex-modularity.

**Edge and layer weights** Modularity approaches can deal with weighted networks\(^6\), and we modified MolTi to handle weighted networks. We also added the possibility to weight each layer of the multiplex network: the contribution of each layer in Equation (1) is multiplied by its weight when computing the multiplex modularity.

**Multiplex modularity** The modularity measure to detect communities in a multiplex network \(X^{(g)}\) can be written as

\[
\sum_g \sum_{(i,j)} \frac{w^{(g)}(i,j)}{2m^{(g)}} \left( X^{(g)}_{i,j} - \frac{k_i^{(g)}k_j^{(g)}}{2m^{(g)}} \right) \delta_{c_i,c_j},
\]

where \(X^{(g)}\) denotes the (monoplex) network of the layer \(g\), \(w^{(g)}\) is the user-defined weight associated to the network \(g\), \(m^{(g)}\) is the sum of the weights of all the edges of \(X^{(g)}\), \(k_i^{(g)}\) is the weight of the edge \((i,j)\) in \(X^{(g)}\), \(K^{(g)}\) is the sum of the weights of all the edges involving vertex \(i\) in \(X^{(g)}\), \(\delta_{c_i,c_j}\) is equal to 1 if \(i\) and \(j\) belong to a same community and 0 otherwise.

**Randomization.** We implemented a randomized version of the Louvain algorithm, similar to the one in GenLouvain\(^6\). Rather than updating the current partition by picking the move leading to the greatest increase of the modularity, we randomly pick a move among those leading to an increase of the modularity. Different runs of the randomized Louvain algorithm generally return different partitions, even if the results are often close. MolTi-DREAM runs the randomized Louvain algorithm a user-defined number of times (from one to ten in this work, four by default), and returns the partition with the highest modularity.

**Simulations of Multiplex Networks with a known community structure**

We simulated random multiplex networks with a fixed known community structure, namely 1,000 vertices split into 20 balanced communities, and various topological properties (i.e., dense/sparse/mixed, with/without missing data)\(^7\). Multiplex networks are simulated by drawing each layer according to this...
structure and fixed intra/inter community edge probabilities (0.1/0.01 for sparse layers and 0.5/0.2 for dense ones). We also generated multiplex networks with missing data in which we randomly withdrawn vertices of each layer with probability 0.5. The relevance of a community structure is assessed by computing the adjusted Rand index\textsuperscript{18} between the detected communities and the ones used to simulate the multiplex networks.

The Disease Module Identification challenge benchmark

**Biological Networks.** The DMI challenge provided six biological networks: two protein-protein interactions, one signaling, one co-expression, one network linking genes essential for the same cancer types, and one network connecting evolutionary-related genes. These six networks have various sizes and edge densities (Table 1). All networks have weighted edges, and all networks but the signaling network are undirected. However, we considered the signaling network as undirected.

**Evaluations with GWAS data.** The communities identified by the different challengers were evaluated according to the associations of their member genes with GWAS data, following the PASCAL tool\textsuperscript{1}.

The procedure leverages the SNP-based \textit{p}-value statistics obtained from 180 GWAS datasets, covering common diseases and traits. It is to note that an important parameter is the FDR threshold used to define the significant associations\textsuperscript{3,4}, and to get the number of significant disease communities. We used three datasets: the “Leaderboard” (76 GWAS) and “Final” (104 GWASs), which were used during the challenge, and their union in a “Total” dataset (180 GWASs).

**Obtaining modules in a given size range.** The DMI challenge set up two constraints on the submitted communities: no overlap and a size ranging from 3 to 100 nodes. We tested different pre-filters (pruning leaves), parameters (resolution parameter, recursions, combination of graph weights for multiplexes) and post-filters (density, size, pruning leaves) in each leaderboard round. We took into account both the number of significant communities and the total number of submitted communities to evaluate the pertinence of each combination. All partitions were post-filtered to keep only classes containing from 7 to 100 nodes.

**Table 1. Number of vertices, of (non-zero-weighted) edges and density of the biological networks used in the DMI challenge.**

| Network | Number of nodes | Number of edges | Density |
|---------|-----------------|-----------------|---------|
| 1-ppi   | 17,397          | 2,232,405       | 1.48 x 10^{-2} |
| 2-ppi   | 12,420          | 397,309         | 5.15 x 10^{-3} |
| 3-signal| 5,254           | 21,826          | 1.34 x 10^{-3} |
| 4-coexpr| 12,588          | 1,000,000       | 1.26 x 10^{-2} |
| 5-cancer| 14,679          | 1,000,000       | 9.28 x 10^{-2} |
| 6-homology | 10,405         | 4,223,606       | 7.80 x 10^{-2} |

**Resolution parameter** Modularity-based clustering approaches are often associated to a resolution parameter \(\gamma\) to tune the size of the obtained communities. We tested different values of this parameter (\(\gamma = 1, \gamma = 5, \gamma = 10, \gamma = 110\)), but the leaderboard tests showed clearly better results for the recursive approach. We chose to keep the default \(\gamma = 1\) and focused on this recursive procedure.

**Recursion procedure** We re-clustered all the communities above a certain size (here 100 vertices) by extracting the corresponding subgraphs from the networks and applying recursively the MolTi algorithm. We iterated the process until obtaining only communities with less than 100 vertices, if possible (some communities with more than 100 vertices cannot be split by considering modularity).

**Results**

**Randomization improves community detection on simulated multiplex networks**

To evaluate the accuracy of the community structures detected from the initial MolTi and its improved version that includes the randomization procedure, we simulated random multiplex networks with a fixed, known community structure, and various features\textsuperscript{4}. Considering a greater number of layers always improves the inference of communities, as already observed\textsuperscript{4} (Figure 1). In addition, communities are better detected from sparse multiplexes than from dense ones. We also observed that the randomizations improve the accuracy of the detected communities, in particular for dense multiplex networks, with or without missing data. Increasing the number of randomization runs improves the results, but to a limited extend after more than four runs.

**Finding disease modules with MolTi**

We applied the improved MolTi to the networks provided by the DMI challenge (Methods). We focused on the sub-challenge 2, which was dedicated to the identification of communities from multiple networks. We considered the six DMI biological networks as layers of a multiplex network, and applied the recursion procedure to obtain communities in the required size range. The significant disease communities were selected regarding their enrichments in GWAS-associated genes (Methods). We observed first that the number of detected disease communities varies in a non-trivial way depending on the GWAS dataset and FDR threshold used (Figure 2). However, we can observe that the number of detected significant disease modules slightly increases after randomization, in particular when the FDR threshold is higher (Figure 2).

**Multiplex versus monoplex.** We next evaluated the added value of the multiplex approach as compared to the identification of modules from the individual networks. When analyzing the significant disease modules obtained for a FDR threshold of 0.1, we observed that combining biological networks in a multiplex generally increases the number of significant modules (Figure 3). However, this does not stand for the cancer and/or homology networks, which lower the number of significant modules retrieved when added as layers of the multiplex. We hypothesize that the community structures of these networks (if they exist)
Figure 1. Adjusted Rand indexes between the reference community structure used to generate the random multiplex networks, and the communities detected by standard and randomized MolTi with 1 to 5 randomization runs. Multiplex networks contain from 1 to 9 graph layers. The indexes are averaged over 2,000 random multiplex networks of 1,000 vertices and 20 balanced communities. Each layer of sparse (resp. dense) multiplex networks is simulated with 0.1/0.01 (resp. 0.5/0.2) internal/external edge probabilities. Mixed multiplex networks are simulated by uniformly sampling each layer among these two pairs of edge probabilities. Multiplex networks with missing data (right column) are generated by withdrawing vertices from each layer with probability 0.5.
Figure 2. Number of significant disease modules identified for different GWAS datasets and FDR thresholds. “Leaderboard” and “Final” datasets were used during the training and final evaluation of the challenge, respectively, whereas the “Total” dataset is the union of the two previous ones.

Figure 3. Number of significant disease modules identified for different combinations of multiplex network layers. Ten randomization have been applied, and the FDR threshold is set to 0.1.
are so unrelated that it is pointless to seek for a common structure by integrating them.

These observations are consistent with the DMI challenge observations, in which the top-scoring team in the sub-challenge 2 handled only the two protein-protein interaction networks. Our algorithm also performs well with the two protein-protein interactions networks, but the highest number of disease modules is retrieved by considering network combinations that exclude the cancer and homology network layers (Figure 3).

**Evaluation of the edge and layer weighting.** All the six biological networks used in the DMI challenge have weighted edges. We compared the number of disease modules obtained by considering or not these weights in the MolTi partitioning, for different FDR thresholds (Table 2). We observed that intra-layer edge weights only has a slight effect on the number of identified significant disease modules, except for the very low significance threshold of 0.01, where it seems pertinent to use these weights.

MoTI-DREAM allows assigning weights to each layer of the multiplex network, for instance to emphasize the layers known to contain more relevant biological information. Given the results of the DMI challenge and our first analyses, we decided to test a combination of weights that would lower the importance of the 5-cancer and 6-homology network layers. We observed that this led to detecting more disease modules (Figure 4). Conversely, less disease modules are detected when higher weights are given to these networks (Figure 4).

**Discussion and conclusion**

We applied here the MoTI software and various extensions to identify disease-associated communities following the DMI challenge benchmark. The new version of MoTI, MoTI-DREAM, runs a randomization procedure, takes into account edge and layer weights, and performs a recursive clustering of the classes that are larger than a given size. We finished tied for second in the challenge. However, even if we obtained higher scores than monoplex approaches, the difference was not significant and the organizers of the DREAM challenge declared the sub-challenge 2 vacant.

In the simulations, all the networks are randomly generated from the same community structure. These networks can thereby

**Table 2. Number of significant disease modules detected.**

| FDR | Unweighted | Weighted |
|-----|------------|----------|
| 0.01| 5          | 10       |
| 0.025| 13         | 12       |
| 0.05| 20         | 19       |
| 0.1 | 30         | 32       |

**Figure 4.** Number of significant disease modules identified with FDR thresholds 0.05 and 0.1, and from three different inter-layer weightings: No Weights, i.e., equal weights for all layers, Confidence Weights, i.e., weights proportional to the expected biological relevance: 1-ppi=1, 2-ppi=1, 3-path=1, 4-coexpr=0.5, 5-cancer=0.1, 6-homology=0.1, and Inverse Confidence Weights, i.e., weights inversely proportional to the expected biological relevance: 1-ppi=0.1, 2-ppi=0.1, 3-path=0.1, 4-coexpr=0.5, 5-cancer=1, 6-homology=1.
be seen as different and partial views of the same underlying community structure. Combining their information in a suitable way is thereby expected to recover the original structure more accurately. In contrast, combining networks with unrelated community structures (or no structure at all) is rather likely to blur the signal carried by each network. The DMI biological networks are constructed from different biological sources that might correspond to unrelated community structures.

This may explain the results of the sub-challenge 2, in which the top-performer used only the two protein-protein interaction networks, and the fact that the highest number of modules retrieved by our approach was not obtained from a multiplex containing all the six networks. From a biological perspective, the protein-protein networks and the pathway networks are expected to contain mainly physical or signaling interactions between proteins. It has been shown that interacting proteins tend to be co-expressed\(^1\), which could explain why the co-expression network also provides complementary information. In contrast, both the cancer and the homology networks are determined from processes operating at a very different level.

Evaluating the relevance of the community structure detected from real-life datasets is a very complicated problem since the actual structure is hidden and generally unknown. In this context, the only possibility for assessing the detected communities is to consider indirect evidence provided by some independent biological information. Different teams are thereby developing proxies to evaluate the communities, mainly based on testing the enrichment of genes contained in each community in Pathways or Gene Ontology annotations. The approach followed by the DMI DREAM challenge is based on GWAS data. This GWAS-based evaluation is specific in the sense that it considers \(p\)-value-weighted annotations rather than usual binary ones, i.e., “annotated/not annotated”. This probably contributed to the volatility of the results observed with the DMI DREAM challenge framework.

**Data and software availability**

MolTi-DREAM and the scripts used for the DMI DREAM challenge: https://github.com/gilles-didier/MolTi-DREAM

Archived scripts and source code for MolTi-DREAM as at time of publication: http://doi.org/10.5281/zenodo.1301209\(^12\)

License for MolTi-DREAM: GNU 3

---

**Author information**

GD designed MolTi and its extensions, AB and AV applied MolTi during and after the challenge. AV and AB are currently at Aix*Marseille Univ, Inserm, MMG, France. All authors participated to the design of the study, the interpretation of the results and the writing of the manuscript.

**Competing interests**

No competing interests were disclosed.

**Grant information**

The project leading to this publication has received funding from the Centre National de la Recherche Scientifique (PEPS BMI IMFMG), the French “Plan Cancer 2009–2013”, and the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French “Investissements d’Avenir” programme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

---

**References**

1. Hartwell LH, Hopfield JJ, Leibler S, et al.: From molecular to modular cell biology. Nature. 1999; 402(6761 Suppl): C47–52. Published Abstract | Publisher Full Text
2. 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. Published Abstract | Publisher Full Text | Free Full Text
3. Choobdar S, Ahsen ME, Crawford J, et al.: Open community challenge reveals molecular network modules with key roles in diseases. bioRxiv. 2018. Publisher Full Text
4. 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. Published Abstract | Publisher Full Text | Free Full Text
5. Didier G, Brun C, Baudot A: Identifying Communities from Multiplex Biological Networks. PeerJ. 2015; 3: e1525. Published Abstract | Publisher Full Text | Free Full Text
6. Newman ME, Girvan M: Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004; 69(2 Pt 2): 026113. Published Abstract | Publisher Full Text
7. Blondel VD, Guillaume J-L, Lambiotte R, et al.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008; 2008(10): P10008. Publisher Full Text
8. Newman ME: Analysis of weighted networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004; 70(5 Pt 2): 056131. Published Abstract | Publisher Full Text
9. Mucha PJ, Richardson T, Macon K, et al.: Community structure in time-dependent, multiscale, and multiplex networks. Science. 2010; 328(5980): 876–8. Published Abstract | Publisher Full Text
10. Santos JM, Embrechts M: On the use of the adjusted rand index as a metric for evaluating supervised classification. In: C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, Artificial Neural Networks – ICANN2009, 175–184, Berlin, Heidelberg, Springer Berlin Heidelberg, 2009. Publisher Full Text
11. Rual JF, Venkatesan K, Hao T, et al.: Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437(7062): 1173–1178. Published Abstract | Publisher Full Text
12. Didier G: gilles-didier/MolTi-DREAM: MolTi-DREAM (Version v1.0). Zenodo. 2008. Data Source
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com