Bioinformatics, 36(6), 2020, 1881–1888
doi: 10.1093/bioinformatics/btz857
Advance Access Publication Date: 18 November 2019
Original Paper

Data and text mining

Discovery of disease- and drug-specific pathways through community structures of a literature network

Minh Pham , Stephen Wilson, Harikumar Govindarajan, Chih-Hsu Lin and Olivier Lichtarge*

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 2, 2019; revised on October 29, 2019; editorial decision on November 11, 2019; accepted on November 15, 2019

Abstract

Motivation: In light of the massive growth of the scientific literature, text mining is increasingly used to extract biological pathways. Though multiple tools explore individual connections between genes, diseases and drugs, few extensively synthesize pathways for specific diseases and drugs.

Results: Through community detection of a literature network, we extracted 3444 functional gene groups that represented biological pathways for specific diseases and drugs. The network linked Medical Subject Headings (MeSH) terms of genes, diseases and drugs that co-occurred in publications. The resulting communities detected highly associated genes, diseases and drugs. These significantly matched current knowledge of biological pathways and predicted future ones in time-stamped experiments. Likewise, disease- and drug-specific communities also recapitulated known pathways for those given diseases and drugs. Moreover, diseases sharing communities had high comorbidity with each other and drugs sharing communities had many common side effects, consistent with related mechanisms. Indeed, the communities robustly recovered mutual targets for drugs [area under Receiver Operating Characteristic curve (AUROC)=0.75] and shared pathogenic genes for diseases (AUROC=0.82). These data show that literature communities inform not only just known biological processes but also suggest novel disease- and drug-specific mechanisms that may guide disease gene discovery and drug repurposing.

Availability and implementation: Application tools are available at http://meteor.lichtargelab.org.

Contact: lichtarge@bcm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Even though pathway information is routinely used to explore hidden biological processes in large omics data, current pathway information is incomplete. Not only knowledgebases do not cover interactions of all human protein-encoding genome (Fabregat *et al.*, 2018), they also focus on mainly general pathways, with few of them related to specific diseases and drugs. Manual curation to complete pathway information is the gold standard but hard-pressed to keep up with a biomedical literature of more than 29 million papers, with 900 000 added each year (Citations Added to MEDLINE® by Fiscal Year: https://www.nlm.nih.gov/bsd/stats/cit_added.html). Professional curators could review only a minuscule portion of the literature body: about 42 000 articles in 6 years (Davis *et al.*, 2013, 2019). The difficulty of curating pathway information from the literature slows knowledge discovery.

Text mining can automate extracting information from the literature, supporting pathway curation (Krallinger *et al.*, 2005). Multiple tools explore biological associations between genes, diseases and drugs from text. First, relevant biological entities have to be identified, through pre-defined rules and dictionaries (Narayanaswamy *et al.*, 2003) or by learned text data features of machine learning (ML) (Habibi *et al.*, 2017). These approaches often yield high precision but low recall due to limited training data. These tools are also not generalized enough to apply to other corpora and are computationally intensive. Medical Subject Headings (MeSH) offers a solution. MeSH terms reliably capture key entities of all articles in MEDLINE because they are manually curated by biocurators to index these articles. MeSH terms include over 28 000 MeSH main heading descriptors and 240 000 Supplementary Concept Records (SCRs) (Fact Sheet Medical Subject Headings: https://www.nlm.nih.gov/pubs/factsheets/mesh.html), which are updated daily to annotate new chemicals and rare diseases that are not covered by MeSH descriptors yet. MeSH terms can also be mapped to IDs in other databases, supporting integration of literature-mined information to knowledgebases and experimental networks.

Biological associations for recognized bio-entities can be extracted from text through co-occurrence and natural language processing (NLP). Co-occurrence approaches assume that co-mentioned bio-entities in text are biologically related (Alako *et al.*, 2005). The confidence for biological relatedness is based on the number of articles with co-mentions. NLP methods rely on prior knowledge or apply ML to learn how biological events are mentioned in text (Li *et al.*, 2017). NLP approaches are often more accurate than co-occurrence methods and can reveal nature of the extracted associations. However, NLP methods are restricted to pre-defined/learned relationships and computationally intensive. Co-occurrence methods are more robust to detect novel biological associations and can be easily scaled up. Co-occurrence associations also cover various biological aspects, such as disease candidate genes (Hristovski *et al.*, 2005), PPI (Tsuruoka *et al.*, 2008) and chemical–gene associations (Rebholz-Schuhmann *et al.*, 2007). Therefore, co-occurrence associations can efficiently supplement pathway information.

To construct pathways, individual text-mined associations can be manually integrated, which is a tedious task due to a great number of associations. Efforts to automate this process have been done, most notably through BioNLP 2013 Pathway Curation task (Pyysalo *et al.*, 2015), which aimed to aggregate relevant entities and their genetic/molecular interactions together to curate pathways. Participating methods applied ML and achieved competitive performance, yet they were restricted by training data and had trouble scaling up. This urges a new approach to efficiently assemble relevant bio-entities and their associations in order to synthesize pathways.

Graph theory can model the vast number of co-occurrence associations of bio-entities, facilitating pathway synthesis. Co-occurrence links can be constructed into networks, in which nodes are biological entities and edges are their co-occurrence relations. MeSH co-occurrence networks robustly recapitulate pathway information in knowledgebases and discover new associative patterns (Wilson *et al.*, 2018). These networks also exhibit small world properties with dense local clusters: entities within clusters are highly interconnected while links among clusters are sparser (Kastrin *et al.*, 2014). When entities have similar neighbors in the networks, they are more likely to link together or be involved in similar processes. In experimental PPI networks, small-world properties help pinpoint most relevant associations for certain biological processes and pathways (Chen *et al.*, 2015; Voevodski *et al.*, 2009). Therefore, we hypothesized that clusters in MeSH co-occurrence networks may also represent functional biological pathways.

Network clusters or *communities* can be constructed through community detection algorithms using topological properties of the networks. Since crosstalk between biological pathways is observed, we are particularly interested in methods that detect overlapping and hierarchically nested communities, reflecting the intricated nature of biology. Previously, Clauset–Newman–Moore (Clauset *et al.*, 2004), Louvain (Blondel *et al.*, 2008), BIGCLAM (Yang and Leskovec, 2013), and Recursive Louvain (RL) (Wilson *et al.*, 2017) methods detected meaningful clusters that represented functional pathways and disease processes in STRING 9.1 experimental PPI network (Franceschini *et al.*, 2012). These tools performed well and fast on the massive network. They searched for densely overlapping and hierarchically nested communities as well as non-overlapped clusters to resemble actual biological pathways. RL, which iteratively breaks down further Louvain-detected communities into smaller groups, detected the similar number of communities with the most similar size distribution to those of annotated pathways than the aforementioned clustering tools. Genes in RL communities were significantly overlapped with control reference pathways, biased to pathogenic genes and co-expressed in breast cancer (Wilson *et al.*, 2017). These data show that RL detected biologically functional gene groups in the PPI network.

Because community structure of the PPI network revealed functional pathways, we proposed generating communities in a co-occurrence network of both MeSH main heading descriptors and SCRs in order to efficiently synthesize pathways from literature-mined associations. Specifically, we now hypothesized that the detected communities, which were groups of highly associated genes, diseases and drugs/chemicals in the network, captured biological pathways and mechanisms of diseases and drugs. Our data validated the biological relevance of these literature communities, which were enriched for curated biological pathways. We further demonstrated that the communities captured curated pathogenic gene sets for diseases and chemical-perturbed gene expression profiles. Clinical relationships between disease–disease and between chemical–chemical in same communities were also validated. Finally, we showed that community structures in the MeSH network helped propose novel disease and drug mechanisms. Overall, the detected literature communities complement pathway curation from text by automating pathway synthesis, and support disease gene discovery and drug repurposing through novel predictions of disease- and drug-specific mechanisms.

2 Materials and methods

2.1 Detecting literature communities

We extracted functional gene sets for specific diseases and chemicals by applying community detection RL algorithm (Wilson *et al.*, 2017) to Mesh Term Objective Reasoning (MeTeOR) network (Wilson *et al.*, 2018) (Fig. 1A). MeTeOR was selected because it robustly captures biological knowledge by comprehensively aggregating co-occurrences of both MeSH main heading descriptors and SCRs, from 22 million MEDLINE publications up to year 2017 (Wilson *et al.*, 2018). MeTeOR also extracts more high-quality associations from the literature than other text-mining methods (Wilson *et al.*, 2018), i.e. STRING Literature (Szklarczyk *et al.*, 2015), STITCH Literature (Szklarczyk *et al.*, 2016), EVEX (Van Landeghem *et al.*, 2013) and BeFree (Bravo *et al.*, 2015). The network consists of 1.07×10^5 nodes of genes (12%), chemicals (83%) and diseases (5%). Human genes are mapped to EntrezID using NCBI's annotations and chemicals to PubchemCID, facilitating our validation against curated databases (mapping details are described in Wilson *et al.*, 2018). The network data were downloaded from http://meteor.lichtargelab.org. RL is the selected community detection algorithm due to its usability in detecting biologically meaningful clusters in a PPI network (Wilson *et al.*, 2017). After running Louvain algorithm, RL makes a subgraph of communities with more than 10 nodes and reruns Louvain detection. The process is done iteratively until all communities are broken down to communities with at most 10 nodes or a node is detected in more than 3 communities. Finally, communities that were highly overlapped, i.e. Jaccard similarity score >0.9, were collapsed, reducing clustering redundancy.

$$Jaccard\ Similarity,\quad J(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{1}$$

where a set of communities $C_{i,j} \in C$.

RL detected communities that contained genes, diseases and chemicals and were proposed to represent functional pathways. We excluded communities with fewer than three genes since they would not represent any reasonable pathways. In addition to communities constructed based on the whole literature that we have up to year 2017, we also retrieved communities based on past literature, i.e. up to year 2005 and up to year 2013, for retrospective studies.

2.2 Evaluating the communities against curated pathways

We evaluated whether the literature communities captured knowledge from curated pathway databases (Fig. 1B). Selected databases include Molecular Signature Database (MSigDB) (Liberzon *et al.*, 2011, 2015), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), WikiPathways (Kelder *et al.*, 2012), Reactome (Fabregat *et al.*, 2018), Gene Ontology Annotation

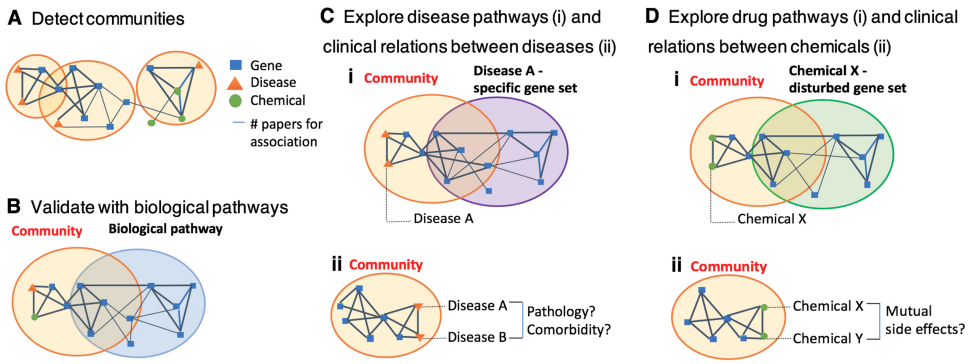

Fig. 1. Overview of discovering biologically meaningful groups of genes, diseases and chemicals. (**A**) Communities of genes, diseases and chemicals were detected in the MeTeOR network. Through enrichment analyses, we evaluated whether the communities could capture biological pathways (**B**), disease pathogenic genes (**C.i**) and drug-perturbed gene expression (**D.i**). Clinical associations between disease pairs (**C.ii**) and drug pairs (**D.ii**) in same communities were also explored

(GOA) for aspects of Cellular Component, Biological Process and Molecular Function (Huntley *et al.*, 2015; The Gene Ontology Consortium, 2017). These databases have different year versions from year 2005 to year 2017. All gene sets of MSigDB were used and downloaded from http://software.broadinstitute.org/gsea/msigdb. The other references were downloaded from http://amp.pharm.mssm.edu/Enrichr/ (Kuleshov *et al.*, 2016). We performed hypergeometric tests for each pair of a literature community and a curated gene set from a database. The hypergeometric *P*-value is calculated by the following equation:

$$Hypergeometric\ \ Test, P(X \geq |C_i \cap R_i|) = 1 - \sum_{j=0}^{|C_i \cap Ri|-1} \frac{\binom{|R_i|}{j}\binom{M-|R_i|}{|C_i|-j}}{\binom{M}{|C_i|}} \quad (2)$$

where M is the number of genes in both the reference and MeTeOR network; C_i is the genes in a given community i; R_i is the genes in a given pathway in a referenced control.

We hypothesized that enrichment *P*-values of our communities for curated pathways were more significant than those of random sets, measured by Kolmogorov–Smirnov (KS) test. Random gene sets were constructed with similar sizes with the detected communities.

In addition, we compared the performance of MeTeOR communities with communities that were constructed from other text-mining networks. Specifically, we aggregated predictions of gene–gene (GG), gene–disease (GD) and gene–chemical (GC) associations from other literature-mining methods into a combined network. The other selected literature-mining methods are STRING Literature v10.5 (Szklarczyk *et al.*, 2015), STITCH Literature v5.0 (Szklarczyk *et al.*, 2016), EVEX (Van Landeghem *et al.*, 2013) and DisGeNET's BeFree v5.0 (Bravo *et al.*, 2015) (network statistics summarized in Supplementary Table S1). To make the comparison fair, we also built communities from associations of GG, GD and GC from MeTeOR predictions. We hypothesized that the communities built from MeTeOR gene-specific edges only are more enriched for pathway information than communities built from other methods.

We also performed time-stamped experiments to validate our predictive power for future curated pathways. We conducted enrichment analyses of communities of past literature, i.e. up to year 2005 or 2013, against pathway databases dated in later years. False discovery rate (FDR) was applied and a community significantly captured a pathway when q-value $\leq$0.1. We evaluated whether a *novel* community, which did not capture any curated pathway at the year it was constructed, could later capture newly added pathways or be *confirmed*.

2.3 Assessing disease knowledge of the communities

2.3.1 Evaluating how well the communities captured curated disease pathways

We hypothesized that genes in communities that contained disease entities explained etiology of those specific diseases (Fig. 1C.i). We gathered curated disease pathways from Online Mendelian Inheritance in Man (OMIM: https://www.omim.org) (Hamosh, 2004), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) (Landrum *et al.*, 2014) and CTD (http://ctdbase.org) (Davis *et al.*, 2019). We also took union of pathways of all sources for each disease to form a 'Total' set. Disease pathways with fewer than three genes were excluded. For communities that contained any diseases with pathway data, we performed hypergeometric tests for disease pathways against the communities. We measured area under Receiver Operating Characteristic curve (AUROC) and area under Precision-Recall curve (AUPRC) to evaluate how well a disease-specific community captured genetic causes of that specific disease. The ranking was the inverse of enrichment *P*-values for each pair of a disease pathway and a community. The truth table indicated whether that disease was in that community.

2.3.2 Evaluating clinical and genetic disease–disease relationships in the communities

We observed cases of multiple diseases in same communities, motivating us to understand their clinical relationships, i.e. shared pathology and comorbidity (Fig. 1C.ii). We obtained hierarchies of disease pathology classes from Disease Ontology (http://disease-ontology.org) (Schriml *et al.*, 2012). We explored whether diseases in same communities fell under similar classes. We also retrieved odds ratio for diseases to occur in same patients from a study (Blair *et al.*, 2013), which examined 8 clinical cohorts with more than 123 million unique patients. We explored whether diseases co-appearing in more communities have higher comorbidity. In both relationship types, we compared distributions of these values in disease pairs with 0, 1 and 2 (or more) communities. We performed KS and Fischer's Exact tests to evaluate whether these distributions were significantly different from each other.

2.4 Assessing drug information of the communities

2.4.1 Evaluating how well the communities captured drug-specific gene expression

We assessed whether drug-related communities explained drug-perturbed gene expression profiles (Fig. 1D.i). We obtained LINCS L1000 Connectivity Map—mRNA expression profiles for cell lines following small molecule perturbation (Subramanian *et al.*, 2017) from https://amp.pharm.mssm.edu/Enrichr. For each drug in LINCS that were mapped to the communities, we performed hypergeometric tests for the communities and drug-perturbed expressed gene

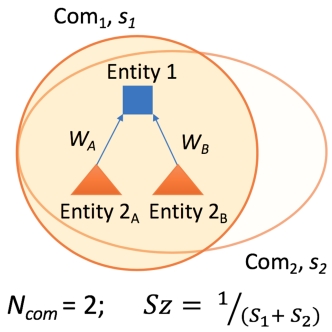

Fig. 2. Predicting multi-relations for genes, diseases and chemicals. For example, a prediction of a mutual pathogenic gene (*Entity 1*) for a disease pair (*Entities* 2_A and 2_B) entailed *W* and *Com-Sz* [calculated by Equations (3)–(5)]. The calculations took into account of edge weights among entities (W_A and W_B), number of the communities that these entities co-occur (N_{com}) and sizes of those communities (*Sz*)

sets. We used FDR *q*-value ≤0.1 as a cutoff for a significant pair of a community and gene expression set. We also performed similar enrichment analyses using 100 random gene sets with similar genes and drugs in the communities. We compared the number of drugs whose expression profiles were explained by the communities compared with random sets.

2.4.2 Evaluating drug–drug relationships in communities

We investigated whether drugs in same communities shared side effects (Fig. 1D.ii). SIDER database v4.1 (http://sideeffects.embl.de) (Kuhn *et al.*, 2016) curates side effects for drugs. We hypothesized that the more communities a pair of drugs share, the more similar side effects that they have. We compared the distribution of number of shared side effects of drug pairs in same communities with that of drug pairs in different communities through KS test.

2.5 Predicting multi-relations for genes, diseases and drugs

In the communities, some of their genes, diseases and chemicals had no MeTeOR associations. We hypothesized that these entities were biologically connected, even though no paper had documented their associations through MeSH terms. We were particularly interested in predicting the following three-entity relations: mutual pathogenic genes for disease pairs, common diseases for gene pairs, shared gene targets for drug pairs and common drugs for gene pairs. In making predictions, we considered number of supporting papers for predicted associations (*W*), and number of shared communities of the three entities (N_{com}), adjusting with sizes of shared communities (*Sz*). Figure 2 summarizes how we made the multi-relation predictions. For example, a prediction of mutual pathogenic gene (*Entity 1*) for a pair of diseases (*Entities* 2_A and 2_B) entailed *W* and *Com-Sz*, which is a combination of N_{com} and *Sz*. If there is no edge linking *Entity 1* to either *Entity* 2_A or *Entity* 2_B, *W* is 0 as it would be impossible to use only the number of supporting papers to predict the 'mutual' pathogenic gene *Entity 1* because there was no paper linking that gene to either of the diseases.

$$W = \frac{W_A + W_B}{\max(W_A + W_B)} \quad \text{if } W_A \neq 0, \quad W_B \neq 0; \text{else } W = 0 \tag{3}$$

$$Sz = \frac{1}{\sum_{i=1}^{N_{com}} S_i} \tag{4}$$

$$Com - Sz = \frac{N_{com} + Sz}{\max(N_{com} + Sz)} \tag{5}$$

where W_A is the edge weight between *Entity 1* and *Entity 2A* (i.e. the number of articles in which *Entity 1* and *Entity 2A* co-occur); W_B is the edge weight between *Entities 1* and *2B*; and S_i is the number of entities in a community index *i* in a set of N_{com} shared communities.

Each multi-relation prediction was ranked by three measures: just *W*, just *Com-Sz* and their sum. Validation was made against curated gene-disease associations in databases OMIM, CTD and DisGeNET v5.0 (Piñero *et al.*, 2017) and against gene-chemical associations in databases CTD and STITCH v5.0 (http://stitch.embl.de) (Szklarczyk *et al.*, 2016). The AUROC was implemented to evaluate the performances of different measures. Note that since we were interested in pathway-relevant communities, we limited our communities to have equal and fewer than 100 genes in this analysis.

Table 1. Summary for the detected communities in the MeTeOR network

Entity type	No. entities	No. communities
Gene	11 127	3444
Disease	4773	1159
Chemical	67 385	2288

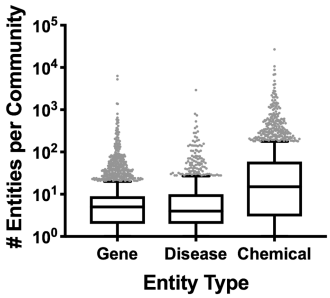

Fig. 3. Distribution of the number of entities per communities

3 Results

3.1 Overview of literature communities

When using the whole MeTeOR network information, the RL algorithm detected 3444 communities comprising a mix of gene, disease and chemical entities (Table 1; Fig. 3). All were defined to contain at least three genes in order to capture biological pathway context. A total of 958 communities consisted of all 3 entity types, potentially reflecting intricate relationships among the contained genes, diseases, and drugs. Of the rest, 201 communities contained just genes and diseases while 1330 communities had only genes and chemicals. This likely reflects that the number of chemical entities dominated the MeTeOR network. These communities should allow us to explore disease and drug processes.

3.2 Communities captured current and predicted future biological pathways

To investigate whether the literature communities were biologically meaningful, a positive control first evaluated whether the communities captured knowledge from already curated pathways or functional gene sets. This was measured by the significance of any overlap between MSigDB pathways and the literature communities (Fig. 1B) compared with that with random genes. Figure 4A shows that the literature communities were systematically enriched for curated pathways (KS test: $P < 0.0001$). Specifically, nearly all of the overlap *P*-values of the communities were <0.01 unlike those of random genes centered above that threshold. Furthermore, MeTeOR communities, either with gene-specific edges only or all edges, captured MSigDB pathway information more than communities using other text-mining methods (Supplementary Fig. S1). These enrichment patterns were not special to MSigDB but held true

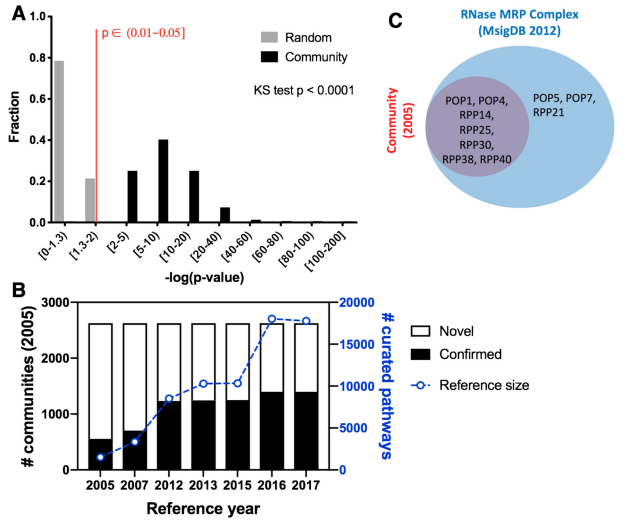

Fig. 4. The communities captured knowledge from the curated pathways of MSigDB. (**A**) The *community P*-values for enrichment of MSigDB pathways were skewed to significance compared with *random*. (**B**) Novel communities in year 2005 predicted pathways added to MSigDB after year 2005. (**C**) RNase MRP complex gene set, which was curated in 2012, was majorly annotated in a community in year 2005 (Jaccard similarity=0.70)

against all the other databases we tested: Reactome, KEGG, WikiPathways, GO Biological Process, GO Cellular Component, GO Molecular Function (Supplementary Figs S1 and S2). Furthermore, the communities robustly captured the majority of gene sets in these databases, from 63.5 to 95.6% (Supplementary Table S2). These data go beyond prior tests (Wilson *et al.*, 2018) to show that MeTeOR literature communities are enriched for high-quality meaningful, curated pathway associations.

Based on the capture of known pathways, we hypothesized that MeTeOR communities could suggest new pathways as well. This was tested through time-stamped experiments. We constructed MeTeOR communities based on the literature up to a past date and then assessed their overlap against pathways added to the database after that date. This is possible because pathway databases, such as MSigDB, have sequential dated versions (from year 2005 and beyond), allowing us to keep track of newly added pathways. We showed that at a given time point, i.e. 2005, many 'novel' communities which were not enriched with any pathway then, eventually developed a significant overlap with the newly added pathways (Fig. 4B). For example, in 2012, MSigDB annotated the RNase MRP complex, 70% of which was already annotated as a MeTeOR community in 2005 (Fig. 4C). The information added to MSigDB fluctuates over time (Fig. 4B), with little in 2016–2017 so that few of the novel communities could be further validated. Over 12 years from 2005 to 2017, however, more than half of the communities were confirmed. These data show that literature communities can discover new pathways, and communities that are still not validated remain potential candidates for unknown biological processes and drug/disease pathways. In Sections 3.3 and 3.4, we specifically explored these novel communities for diseases and drugs, respectively.

There are two additional layers of complexity to this retrospective analysis. First, communities from 2005 could become enriched for pathways or stay novel depending on the reference used (Supplementary Fig. S3A). This is because each reference may hold a different type and source of data. We chose to focus on MSigDB because it is a high-quality dataset often used for pathway analysis, but our conclusions hold across many different sources (Supplementary Fig. S3). Second, the year 2005 is not unique in offering insights into pathways, and we demonstrated that communities created on 2013 data also predicted groups of genes about to be annotated (Supplementary Fig. S3B). Additionally, there were more communities obtained based on literature up to 2013 than 2005, suggesting that with more literature information, more groupings of biological entities can be generated that are of potential value. Interestingly, for similar reference versions, there were more communities in year 2013 validated than those constructed in year 2005, indicating that updating the literature information allowed more accurate recapitulation of biological pathways.

In conclusion, the communities built from modularity of the MeSH co-occurrence network summarized current pathway knowledge and predicted future pathway information worthy of experimental assessment.

Table 2. Area under ROC curve (AUROC) and area under precision-recall (AUPRC) for community predictions of disease pathways

Reference	AUROC	AUPRC
OMIM	0.80	0.71
ClinVar	0.68	0.44
CTD	0.64	0.20
Total	0.64	0.20

3.3 Communities captured disease-specific pathways

Multiple communities contained diseases in addition to genes. We hypothesized that these communities suggested meaningful relationships among diseases and genes, such as disease pathogenic genes and disease comorbidity. We first investigated whether genes in same communities of diseases were causative for these diseases. We gathered curated disease pathways from OMIM, ClinVar and CTD, and used them to validate the gene sets for each disease-contained community. We proposed that when a community contained a specific disease, that community was highly enriched in pathogenic genes for the disease (Fig. 1C.i). Indeed, when predicting diseases that each community contained, we achieved an overall AUROC of 0.64 and up to 0.80 for OMIM (Table 2). AUPRC was worst for CTD even though AUROC for CTD was comparable with that for ClinVar, suggesting that predictions for CTD were less precise. Predictive performances for OMIM and ClinVar were better than that for CTD because CTD is more general in scope while OMIM and ClinVar are more clinically focused. This is best demonstrated at the level of annotation, where clinicians and physicians annotate OMIM and ClinVar while non-clinical specialists went through the literature to curate CTD to include a much greater number of general disease–gene associations (Supplementary Table S3). The majority of disease pathway information of the *total* or combined set was from CTD (Supplementary Table S3), leading to similar predictive performances between *total* and CTD. These data show that generally communities captured gene sets for diseases but represented clinically significant genes especially well.

We further unraveled clinical associations between disease–disease pairs within communities. Our hypothesis was that diseases in the same communities shared similar pathology (Fig. 1C.ii). We first looked at Disease Ontology, in which 2397 diseases were mapped to the MeTeOR communities. We investigated whether diseases detected in the same communities fell under similar disease classifications. We found that most diseases that shared communities were in the same classes, suggesting that they shared similar pathologies (Fischer's Exact test: $P < 0.0001$, odds ratio = 6.3) (Supplementary Table S4). This finding highlights that communities depicted pathophysiological relationships for the diseases.

Another disease–disease relationship type that we explored in communities was comorbidity (Fig. 1C.ii). Previous studies show that there is a strong correlation between disease comorbidity and their genetic and molecular risk interactions (Blair *et al.*, 2013; Lee *et al.*, 2008). Likewise, we saw that the more communities a pair of diseases were detected together, the higher odds ratio that they co-occur in same patients (Fig. 5; Supplementary Table S5). This result further suggests that diseases in the same communities share pathophysiology. Combining with the previous finding that disease-specific communities recapitulated well curated information of disease driver genes, the data demonstrate that the common underlying

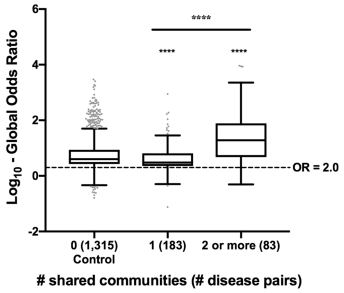

Fig. 5. Diseases co-occurring in same communities had higher comorbidity, suggesting that they shared mechanisms. $^{****}P \leq 0.0001$ (KS test)

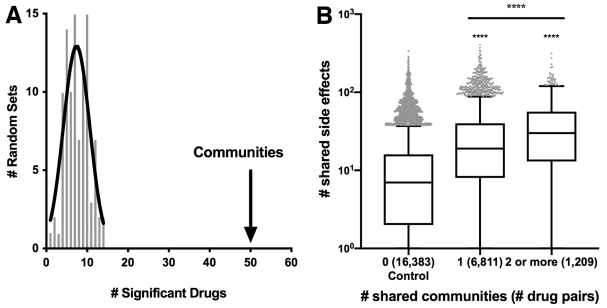

Fig. 6. The communities captured genetic and clinical information of drugs, and chemicals. (**A**) The communities significantly captured perturbed gene expression profiles for 50 drugs, much more than the number of drugs captured by random sets ($N = 100$) (gray bars) (z-score = 15). (**B**) Drugs in same communities shared great numbers of side effects, suggesting that they shared mechanisms. $^{****}P \leq 0.0001$ (KS test)

mechanisms for diseases in same communities can be explained by genes in same communities.

3.4 Communities captured drug-specific pathways

We explored whether genes in the communities reflected drug pathways (Fig. 1D). Specifically, we examined whether genes in drug-specific communities recapitulated drug-perturbed gene expression. We performed hypergeometric enrichment analyses for drug-containing communities and experimental gene expression profiles for specific drugs from LINCS database (Fig. 1D.i). We also compared the enrichment results with those when randomizing genes in the drug-specific communities. We observed that the number of drugs whose expression profiles were significantly captured by the communities was much higher than that by random sets (z-score = 15; Fig. 6A). This result demonstrates that genes in drug-specific communities were enriched for genes up/down-regulated by the corresponding drugs.

Communities that contained a drug usually contained many drugs, motivating us to explore their clinical associations. We examined whether drugs in same communities shared side effects (Fig. 1D.ii). SIDER annotates side effects for 792 drugs mapped to the MeTeOR communities. Figure 6B illustrates that drugs co-occurring in more communities shared similar side effects, suggesting that they acted through similar mechanisms. We noted that multiple drug pairs shared up to 20 common side effects (e.g. itchy skin and headache), but drug pairs that co-occurred in 2 or more communities shared more side effects, up to 409. Two drugs that shared the highest number of side effects were *Aripiprazole* and *Escitalopram*, both treating depression (Nelson *et al.*, 2008). The community that they co-occurred in was enriched for genes relating to neuroactive ligand–receptor interactions, explaining 409 side effects that they shared. Our findings demonstrate that drugs in the same communities often interacted or acted on genes in the same communities, inducing similar side effects.

3.5 The literature communities proposed plausible disease and drug mechanisms

Since the communities captured clinical and biological associations between genes, diseases and chemicals, we proposed that they could generate plausible disease and drug mechanisms. We were particularly interested in detecting the following multi-relations: mutual pathogenic genes for disease pairs, common diseases for gene pairs, shared gene targets for drug pairs and common drugs for gene pairs. Identifying mutual genes for disease pairs could explain the common underlying mechanisms for diseases with similar pathology and/or high comorbidity. Detecting common diseases for gene pairs shed light on disease pathways. Predicting shared gene targets for drug pairs and common drugs for gene pairs facilitates drug repurposing. The predictions could suggest new drugs for the same gene targets or repurposed drugs for other genes and diseases.

We started with detection of shared pathogenic genes for a pair of diseases (Fig. 2). Naively, we prioritized genes annotated on multiple papers together to indicate that these genes were linked to both diseases. For our literature network, the number of supporting papers for associations (W) is the sum of edge weights between genes and diseases. When compared against a combined gold standard of curated gene-disease associations, communities predicted shared genes for diseases well (AUROC = 0.79) (Table 3), suggesting that genes that are highly co-mentioned with diseases in publications may explain shared mechanisms of these diseases. Next, we explored whether the number of shared communities that a gene and a pair of diseases co-occur (*Com-Sz*) recovered underlying mechanisms for disease pairs. *Com-Sz* achieved a high AUROC of 0.75 (Table 3), indicating that community structure defined meaningful clusters of genes and diseases that explain common genetics of diseases.

Table 3. Area under ROC for multi-relation predictions

Entity 1 type	Entity 2 type	W	*Com-Sz*	W + *Com-Sz*
Gene	Disease	0.79	0.75	0.82
Disease	Gene	0.83	0.79	0.86
Gene	Chemical	0.58	0.77	0.80
Chemical	Gene	0.63	0.70	0.75

The edge weight (W) is more intuitive than the community measure (*Com-Sz*) because W prioritizes already known associations. However, *Com-Sz* unravels unannotated associations in the literature network that were grouped together due to connections with other entities in same communities. This goes back to our major hypothesis in which entities that share neighbors may be as well biologically related. Combining both edge weight and community structure can fully utilize their benefits. Indeed, summation of both measures improved detection of shared pathogenic genes for disease pairs (AUROC = 0.82) (Table 3).

We observed similar patterns when predicting common diseases for gene pairs, shared gene targets for drug pairs and common drugs for gene pairs (Table 3). In detection of shared diseases for gene pairs, W outperformed *Com-Sz*. In the case of identifying shared gene targets for drug pairs and common drugs for gene pairs, *Com-Sz* had higher predictive power than W. An explanation for this observation could be that the network information on relationships between genes and diseases was denser than relationships between genes and drugs. In all cases, combining both measures yielded the best performance, on average of 0.81 for AUROC (Table 3). These data show that community structure filled in the missing gap of annotating information and complemented with known knowledge.

We also searched for examples of drug repurposing through the use of the literature communities in a time-stamped experiment. Using the literature up to year 2013, we looked for disease–chemical pairs in communities that had no co-occurrence association. We proposed that the fact that these disease–chemical pairs co-occurred in multiple communities indicated that they had some clinical associations that would be validated later. Indeed, for disease–chemical

Table 4. Novel disease–chemical associations suggested by the top number of shared communities in 2013 (seven communities) were validated by publications in later year

Disease	Chemical	Drug function	PMID (year)
Neovascularization, pathologic	DCC-2701	Therapeutic	26285778 (2015)
Vitreous detachment	perfluorooctane	Used in surgery	24800216 (2014)
Exfoliation syndrome	AR-12286	Therapeutic	27552517 (2016)
Exfoliation syndrome	K-115	Therapeutic	28349329 (2017)
Gastrointestinal neoplasms	dotatate gallium ga-68	Diagnostic	29159606 (2018)
Anhedonia	fluvoxamine	Therapeutic	27987210 (2017)
Adenomyosis	gestrinone	Palliative	25510683 (2014)
Heavy metal poisoning	sodium arsenite	Causative	26091798 (2015)

pairs that co-occurred in multiple communities, even though many of them were not found associated together by MeSH co-occurrence in year 2013, several drugs were successfully repurposed for the corresponding diseases in later years (Table 4). For example, drug *DC-2701* was found in seven communities with *Neovascularization* and its known treatments because *DC-2701* and the treatments were highly connected together due to their similar mechanisms of action. Even though in year 2013, *DC-2701* was not studied for the disease yet, the fact that they co-occurred in many communities suggests that *DC-2701* may treat *Neovascularization*, which in fact, was confirmed in year 2015. These data suggest that the communities mimicked human hypothesis generation, efficiently grouped clinically relevant chemicals and diseases and thus, reliably synthesized plausible drug repurposing hypotheses.

4 Discussion

Individual associations between bio-entities are extensively explored, yet it is challenging to integrate them into useful pathway information. Here, we proposed a new approach that utilized modularity of an MeSH co-occurrence network in order to efficiently mine new pathway information from the literature. Specifically, we generated functional groups or *communities* of genes, diseases and chemicals that reliably summarized knowledge from biological processes. Most communities captured significant portions of the curated pathways or included pathways as their subsets, thus highlighting core drivers of curated pathways and expanding current curated information. The communities that were not significantly overlapped with any curated pathway, on the other hand, proposed yet unknown functional processes or disease- and drug-specific mechanisms, which were not curated in the tested pathway databases. The proposed processes can be confirmed in the future (Fig. 4B) or may have already been experimentally validated but not curated yet. Overall, the communities provided meaningful biological information to supplement current pathway knowledgebases.

Furthermore, the communities provided known and novel genetic and clinical information for diseases and drugs. The communities detected diseases and drugs with similar clinical manifestations (e.g. disease comorbidity) and captured disease- and drug-specific mechanisms of actions through the genes in the same communities. The communities also robustly recovered multi-relations among genes, diseases and drugs (e.g. mutual drug targets), many of which were not already curated. Finally, the communities proposed promising hypotheses for disease gene discovery and drug repurposing with many successful cases as shown in a retrospective study (Table 4).

Overall, the literature communities automate integrating related associations to synthesize functional pathways and provide genetic and clinical contexts for genes, diseases and drugs of interest. Furthermore, the communities can imitate human hypothesis generation and reliably propose ideas for disease gene discovery and drug repurposing that are worthy of experimental assessment. Instead of scientists going through individual associations to formulate hypotheses, the communities efficiently aggregate relevant biological interactions to propose plausible mechanisms. Even if entities of novel associations are distant from each other in the network, the fact that they share multiple neighbors, as detected by the communities, strongly suggests that they are biologically relevant, and that their associations are valid. For predictions of novel associations, the communities also inform their functional and clinical nature in the context of other genes, diseases and drugs co-occurring in same communities and easily point out specific publications that were used to construct communities and thus, are relevant to novel associations.

There are aspects that we would like to improve with the communities. Currently, knowledge of the communities is limited to MeSH terms. MeSH terms were selected because they are reliably curated by biocurators and their IDs can be mapped easily to other databases, supporting knowledge integration. Yet, they do not cover all keywords, leading to insufficient information in the communities. We attempted to complete gene information of the communities by supplementing the MeTeOR network with more than 8000 additional human genes extracted by PubTator (Wei *et al.*, 2019), a named entity recognition method for PubMed-indexed citations. The communities after supplementing with PubTator-mined genes were significantly less enriched with curated pathway information than the original MeTeOR communities. This suggests that the genes extracted by PubTator were not necessary key entities of articles, adding spurious information (Supplementary Fig. S4). Only when we restricted to use PubTator-mined genes that appeared more than 50 citation mentions in order to reduce the redundant information, the newly constructed communities improved their pathway information enrichments and was even comparable to the original communities for some pathway references (Supplementary Fig. S4). For a future study, we plan to combine other curated network data in order to improve completeness and quality of the community information.

In addition, even though the communities propose meaningful associations, they lack directionality, e.g. the communities could not differentiate whether chemicals treat or cause diseases (Table 4). Currently, the nature of novel associations can be inferred from the literature information of relevant associations in same communities. To automate this process, we plan to integrate the literature network with directional biological networks to deduce directionality of proposed associations. We can also apply NLP to specific publications relevant to the novel associations to validate and improve our annotations.

To support utilization of the communities, we provide data and tools on http://meteor.lichtargelab.org. Users can explore the detected literature communities for discovery of novel mechanisms of diseases and drugs. The website can also extract significantly overlapped communities for any given groups of genes, diseases and/or chemicals of interest. For example, the website can detect communities that are significantly enriched for users' genes of interest and simultaneously highlight diseases and chemicals that co-occur in these communities and thus, are functionally related with the genes. This function is particularly helpful for users to identify novel diseases and drugs linked to genes of interest and to investigate unknown biological processes for omics data.

Funding

This work has been supported by the National Institutes of Health [GM079656, GM066099, AG061105].

Conflict of Interest: none declared.

References

Alako,B.T. *et al*. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.

Blair,D.R. *et al*. (2013) A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell*, **155**, 70–80.

Blondel,V.D. *et al*. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.

Bravo,À. *et al*. (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, **16**, 55.

Chen,C.-Y. *et al*. (2015) Dissecting the human protein-protein interaction network via phylogenetic decomposition. *Sci. Rep*., **4**, 7153.

Clauset,A. *et al*. (2004) Finding community structure in very large networks. *Phys. Rev*. E, **70**, 066111.

Davis,A.P. *et al*. (2013) A CTD-Pfizer collaboration: manual curation of 88, 000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database (Oxford)*, **2013**, bat080.

Davis,A.P. *et al*. (2019) The comparative toxicogenomics database: update 2019. *Nucleic Acids Res*., **47**, D948–D954.

Fabregat,A. *et al*. (2018) The reactome pathway knowledgebase. *Nucleic Acids Res*., **46**, D649–D655.

Franceschini,A. *et al*. (2012) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., **41**, D808–D815.

Habibi,M. *et al*. (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, **33**, i37–i48.

Hamosh,A. (2004) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.

Hristovski,D. *et al*. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform*., **74**, 289–298.

Huntley,R.P. *et al*. (2015) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res*., **43**, D1057–D1063.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Kastrin,A. *et al*. (2014) Large-scale structure of a network of co-occurring MeSH terms: statistical analysis of macroscopic properties. *PLoS One*, **9**, e102188.

Kelder,T. *et al*. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res*., **40**, D1301–D1307.

Krallinger,M. *et al*. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today*, **10**, 439–445.

Kuhn,M. *et al*. (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res*., **44**, D1075–D1079.

Kuleshov,M.V. *et al*. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*., **44**, W90–W97.

Van Landeghem,S. *et al*. (2013) Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, **8**, e55814.

Landrum,M.J. *et al*. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*., **42**, D980–D985.

Lee,D.-S. *et al*. (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA*, **105**, 9880–9885.

Li,F. *et al*. (2017) A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, **18**, 198.

Liberzon,A. *et al*. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Liberzon,A. *et al*. (2015) The molecular signatures database Hallmark gene set collection. *Cell Syst*., **1**, 417–425.

Narayanaswamy,M. *et al*. (2003) A biological named entity recognizer. *Pac. Symp. Biocomput*., **2003**, 427–438.

Nelson,J.C. *et al*. (2008) Augmentation treatment in major depressive disorder: focus on aripiprazole. *Neuropsychiatr. Dis. Treat*, **4**, 937–948.

Piñero,J. *et al*. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Pyysalo,S. *et al*. (2015) Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. *BMC Bioinformatics*, **16**(**Suppl. 10**), S2.

Rebholz-Schuhmann,D. *et al*. (2007) EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

Schriml,L.M. *et al*. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*., **40**, D940–D946.

Subramanian,A. *et al*. (2017) A next generation connectivity map: l 1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.e17.

Szklarczyk,D. *et al*. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., **43**, D447–D452.

Szklarczyk,D. *et al*. (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*., **44**, D380–D384.

The Gene Ontology Consortium. (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.

Tsuruoka,Y. *et al*. (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.

Voevodski,K. *et al*. (2009) Finding local communities in protein networks. *BMC Bioinformatics*, **10**, 297.

Wei,C.-H. *et al*. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res*., **47**, W587–W593.

Wilson,S.J. *et al*. (2017) Discovery of functional and disease pathways by community detection in protein-protein interaction networks. *Pac. Symp. Biocomput*., **22**, 336–347.

Wilson,S. *et al*. (2018) Automated literature mining and hypothesis generation through a network of Medical Subject Headings. *bioRxiv*, 403667. doi: 10.1101/403667.

Yang,J. and Leskovec,J. (2013) Overlapping community detection at scale. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data mining - WSDM'13*. ACM Press, New York, New York, USA, p. 587.