Hetnet connectivity search provides rapid insights into how two biomedical entities are related

This manuscript (permalink) was automatically generated from greenelab/connectivity-search-manuscript@9f0f132 on June 21, 2020.

Authors

Manuscript in preparation
The authorship information below is incomplete and preliminary. This notice will be updated once all contributors meeting authorship criteria have added themselves to `metadata.yaml`.

- **Daniel S. Himmelstein**
  0000-0002-3012-7446 · dhimmel · dhimmel
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America · Funded by GBMF4552

Abstract

Hetnets, short for “heterogeneous networks”, contain multiple node and relationship types and offer a way to encode biomedical knowledge. For example, Hetionet connects 11 types of nodes — including genes, diseases, drugs, pathways, and anatomical structures — with over 2 million edges of 24 types. Previously, we trained a classifier to repurpose drugs using features extracted from Hetionet. The model identified types of paths between a drug and disease that occurred more frequently between known treatments.

For many applications however, a training set of known relationships does not exist; Yet researchers would still like to know how two nodes are meaningfully connected. For example, users may want to know not only how metformin is related to breast cancer, but also how the GJA1 gene might be involved in insomnia. Therefore, we developed hetnet connectivity search to propose the most important paths between any two nodes.

The algorithm behind connectivity search identifies types of paths that occur more frequently than would be expected by chance (based on node degree alone). We implemented the method on Hetionet and provide an online interface at https://het.io/search. Several optimizations were required to precompute significant instances of node connectivity at scale. We provide an open source implementation of these methods in our new Python package named hetmatpy.

To validate the method, we show that it identifies much of the same evidence for specific instances of drug repurposing as the previous supervised approach, but without requiring a training set.

Introduction

A *network* (also known as a graph) is a conceptual representation of a group of entities — called *nodes* — and the relationships between them — called *edges*. Typically, a network has only one type of node and one type of edge. But in many cases, it is necessary to be able to distinguish between different types of entities and relationships.

Hetnets

A *hetnet* (short for **het**erogeneous information **net**work [1]) is a network where nodes and edges have type. The ability to differentiate between different types of entities and relationships allows a hetnet to accurately describe more complex data. Hetnets are particularly useful in biomedicine, where it is important to capture the conceptual distinctions between various concepts, such as genes and diseases, or upregulation and binding.

The types of nodes and edges in a hetnet are defined by a schema, referred to as a metagraph. The metagraph consists of metanodes (types of nodes) and metaedges (types of edges). Note that the prefix *meta* is used to refer to type (e.g. compound), as opposed to a specific node/edge/path itself (e.g. acetaminophen).

Hetionet

Hetionet is a knowledge graph of human biology, disease, and medicine, integrating information from millions of studies and decades of research. Hetionet v1.0 combines information from 29 public databases. The network contains 47,031 nodes of 11 types (Table 1) and 2,250,197 edges of 24 types.

Table 1: Node types in Hetionet The abbreviation, number of nodes, and description for each of the 11 metanodes in Hetionet v1.0.

Metanode	Abbr	Nodes	Description
Anatomy	A	402	Anatomical structures, excluding structures that are known not to be found in humans. From Uberon.
Biological Process	BP	11381	Larger processes or biological programs accomplished by multiple molecular activities. From Gene Ontology.
Cellular Component	CC	1391	The locations relative to cellular structures in which a gene product performs a function. From Gene Ontology.
Compound	C	1552	Approved small molecule compounds with documented chemical structures. From DrugBank.
Disease	D	137	Complex diseases, selected to be distinct and specific enough to be clinically relevant yet general enough to be well annotated. From Disease Ontology.

Metanode	Abbr	Nodes	Description
Gene	G	20945	Protein-coding human genes. From Entrez Gene.
Molecular Function	MF	2884	Activities that occur at the molecular level, such as “catalysis” or “transport”. From Gene Ontology.
Pathway	PW	1822	A series of actions among molecules in a cell that leads to a certain product or change in the cell. From WikiPathways, Reactome, and Pathway Interaction Database.
Pharmacologic Class	PC	345	“Chemical/Ingredient”, “Mechanism of Action”, and “Physiologic Effect” FDA class types. From DrugCentral.
Side Effect	SE	5734	Adverse drug reactions. From SIDER/UMLS.
Symptom	S	438	Signs and Symptoms (i.e. clinical abnormalities that can indicate a medical condition). From the MeSH ontology.

Hetionet provides a foundation for building hetnet applications. It unifies data from several different, disparate sources into a single, comprehensive, accessible, common-format network. The database is publicly accessible without login at https://neo4j.het.io. The Neo4j graph database enables querying Hetionet using the Cypher language, which was designed to interact with networks where nodes and edges have both types and properties.

One limitation that restricts the applicability of Hetionet is incompleteness. In many cases, Hetionet v1.0 includes only a subset of the nodes from a given resource. For example, the Disease Ontology contains over 9,000 diseases [2], while Hetionet includes only 137 diseases [3]. Nodes were excluded to avoid redundant or overly specific nodes, while ensuring a minimum level of connectivity for compounds and diseases. See the Project Rephetio methods for more details [4]. Nonetheless, Hetionet v1.0 remains one of the most comprehensive and integrative networks that consolidates biomedical knowledge into a manageable number of node and edge types. Other integrative resources, some still under development, include Wikidata [5], SemMedDB [6,7,8], SPOKE, and DRKG.

Rephetio

Project Rephetio is the name of the study that created Hetionet and applied it repurpose drugs [4]. This project predicted the probability of drug efficacy for 209,168 compound–disease pairs. The approach learned which types of paths occur more or less frequently between known treatments than non-treatments (Figure 1). To train the model, Rephetio created PharmacotherapyDB, a physician-curated catalog of 755 disease-modifying treatments [9].

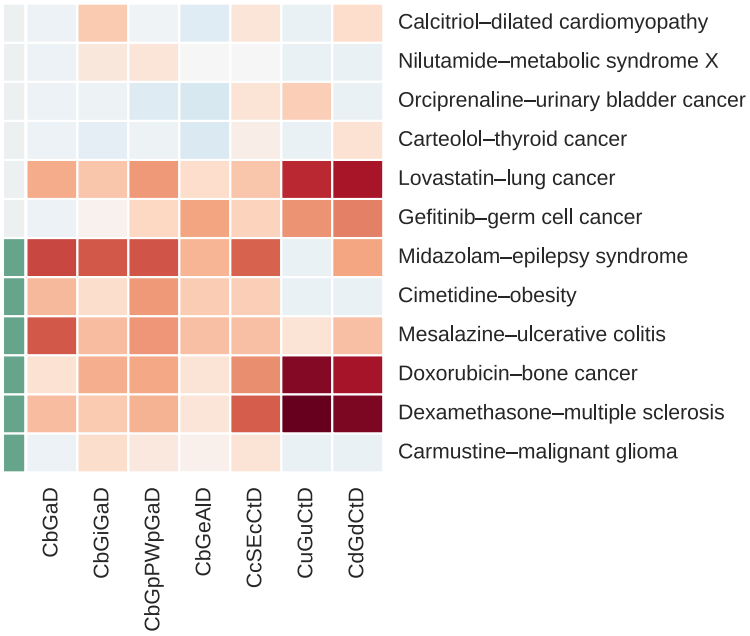

Figure 1: Supervised machine learning approach from Project Rephetio. This figure visualizes the feature matrix used by Project Rephetio to make supervised predictions. Each row represents a compound–disease pair. The top half of rows correspond to known treatments (i.e. positives), while the bottom half correspond to non-treatments (i.e. negatives, not known to be treatments in PharmacotherapyDB). Here, an equal number of treatments and non-treatments are shown, but in reality the problem is heavily imbalanced. Project Rephetio scaled models to assume a positive prevalence of 0.36% [4,10]. Each column represents a metapath, labeled with its abbreviation.

Feature values are DWPCs (transformed and standardized), which assess the connectivity along the specified metapath between the specific compound and disease. Maroon colored values indicate above-average connectivity, whereas blue values indicate below average connectivity. In general, positives have greater connectivity for the selected metapaths than negatives. Rephetio used a logistic regression model to learn the effect of each type of connectivity (feature) on the likelihood that a compound treats a disease. The model predicts whether a compound–disease pair is a treatment based on its features, but requires supervision in the form of known treatments.

TODO: Other works

https://github.com/greenelab/hetmech/issues/56

Network embeddings edge2vec [11], metapath2vec [12], HINE [13].

14 training node pairs to important metapaths (Forward Stagewise Path Generation). MetaExp [15] user selects two sets of nodes. MetaExp detects metapaths and interacts with the user to progressively refine metapaths.

Unsupervised connectivity search

Results

Hetmatpy Package

We created the hetmatpy Python package, available on GitHub and PyPI under the permissive BSD-2-Clause Plus Patent License. This package provides a matrix-based utilities for hetnets.

DWPC null distribution

To assess connectivity between a source and target node, we use the DWPC (degree-weighted path count) metric. The DWPC is similar to path count (number of paths between the source and target node along a given metapath), except that it downweights paths through high degree nodes. Rather than using the raw DWPC for a source-metapath-target combination, we transform the DWPC across all source-target node pairs for a metapath to yield a distribution that is more compact and amenable to modeling [16].

Previously, we had no technique for detecting whether a DWPC value was exceptional. One possibility is to evaluate the DWPCs for all pairs of nodes and select the top scores (e.g. the top 5% of DWPCs). Another possibility is to pick a transformed DWPC score as a cutoff. The shortcomings of these methods are twofold. First, neither the percentile nor absolute value of a DWPC has inherent meaning. To select transformed DWPCs greater than 6, or alternatively the top 1% of DWPCs, is arbitrary. Second, comparing DWPCs between node pairs fails to account for the situation where high-degree node pairs are likely to score higher, solely on account of their degree (TODO: figure).

To address these shortcomings, we developed a method to compute the right-tail *p*-value of a DWPC. *p*-values have a broadly understood interpretation — in our case, the probability that a DWPC equal to or greater than the observed DWPC could occur under a null model. By tailoring the null distribution for a DWPC to the degree of its source and target node, we account for degree effects when determining the significance of a DWPC.

Enriched metapaths

Enriched paths

Comparison to Rephetio

Detecting Mechanisms of Action for Indications

Assess ability to predict paths in https://github.com/SuLab/DrugMechDB

Connectivity Search Webapp

Use cases

Discussion

Methods

The HetMat awakens

At the core of the hetmatpy package is the HetMat data structure for storing and accessing the network. HetMats are stored on disk as a directory, which by convention uses a `.hetmat` extension. A HetMat directory stores a single heterogeneous network, whose data resides in the following files.

1. A `metagraph.json` file stores the schema, defining which types of nodes and edges comprise the hetnet. This format is defined by the hetnetpy Python package. Hetnetpy was originally developed with the name hetio during prior studies [4,17], but we renamed it to het**net**py for better disambiguation from het**mat**py.
2. A `nodes` directory containing one file per node type (metanode) that defines each node. Currently, `.tsv` files where each row represents a node are supported.
3. An `edges` directory containing one file per edge type (metadata) that encodes the adjacency matrix. The matrix can be serialized using either the Numpy dense format (`.npy`) or SciPy sparse format (`.sparse.npz`).

For node and edge files, compression is supported as detected from `.gz`, `.bz2`, `.zip`, and `.xz` extensions. This structure of storing a hetnet supports selectively reading nodes and edges into memory. For example, a certain computation may only require access to a subset of the node and edge types. By only loading the required node and edge types, we reduce memory usage and read times.

Additional subdirectories, such as `path-counts` and `permutations`, store data generated from the HetMat. By using consistent paths for generated data, we avoid recomputing data that already exists on disk. A HetMat directory can be zipped for archiving and transfer. Users can selectively include generated data in archives. Since the primary application of HetMats is to generate computationally demanding measurements on hetnets, the ability to share HetMats with precomputed data is paramount.

The `HetMat` class implements the above logic. A `hetmat_from_graph` function creates a HetMat object and directory on disk from the pre-existing `hetnetpy.hetnet.Graph` format.

We converted Hetionet v1.0 to HetMat format and uploaded the `hetionet-v1.0.hetmat.zip` archive to the Hetionet data repository.

Computing DWPCs with matrix multiplication

Permuted hetnets

Degree-grouping of node pairs

Gamma-hurdle distribution

Prioritizing metapaths for database storage

Rest API & backend

Webapp & Frontend

Realtime open science

Software & data availability

References

1. **Renaming “heterogeneous networks” to a more concise and catchy term**
   Daniel Himmelstein, Casey Greene, Sergio Baranzini
   ThinkLab (2015-08-16) https://doi.org/f3mn4v
   DOI: 10.15363/thinklab.d104

2. **Human Disease Ontology 2018 update: classification, content and workflow expansion**
   Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, … Carol Greene
   Nucleic Acids Research (2019-01-08) https://doi.org/ggx9wp
   DOI: 10.1093/nar/gky1032 · PMID: 30407550 · PMCID: PMC6323977

3. **Unifying disease vocabularies**
   Daniel Himmelstein, Tong Shu Li
   ThinkLab (2015-03-30) https://doi.org/f3mqv5
   DOI: 10.15363/thinklab.d44

4. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
   Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
   eLife (2017-09-22) https://doi.org/cdfk
   DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

5. **Wikidata as a knowledge graph for the life sciences**
   Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, … Andrew I Su
   eLife (2020-03-17) https://doi.org/ggqqc6
   DOI: 10.7554/elife.52614 · PMID: 32180547 · PMCID: PMC7077981

6. **SemMedDB: a PubMed-scale repository of biomedical semantic predications**
   H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch
   Bioinformatics (2012-10-08) https://doi.org/f4hp3x
   DOI: 10.1093/bioinformatics/bts591 · PMID: 23044550 · PMCID: PMC3509487

7. **Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data**
   Qing Cong, Zhiyong Feng, Fang Li, Li Zhang, Guozheng Rao, Cui Tao
   2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2018-12)
   https://doi.org/ggzb26
   DOI: 10.1109/bibm.2018.8621568

8. **Time-resolved evaluation of compound repositioning predictions on a text-mined knowledge network**
   Michael Mayers, Tong Shu Li, Núria Queralt-Rosinach, Andrew I. Su
   BMC Bioinformatics (2019-12-11) https://doi.org/ggpcsr
   DOI: 10.1186/s12859-019-3297-0 · PMID: 31829175 · PMCID: PMC6907279

9. **Announcing PharmacotherapyDB: the Open Catalog of Drug Therapies for Disease**
   Daniel Himmelstein
   ThinkLab (2016-03-15) https://doi.org/f3mqtv
   DOI: 10.15363/thinklab.d182

10. **Our hetnet edge prediction methodology: the modeling framework for Project Rephetio**
Daniel Himmelstein
ThinkLab (2016-05-04) https://doi.org/f3qbmj
DOI: 10.15363/thinklab.d210

11. **edge2vec: Representation learning using edge semantics for biomedical knowledge discovery**
Zheng Gao, Gang Fu, Chunping Ouyang, Satoshi Tsutsui, Xiaozhong Liu, Jeremy Yang, Christopher Gessner, Brian Foote, David Wild, Ying Ding, Qi Yu
BMC Bioinformatics (2019-06-10) https://doi.org/ggpcsq
DOI: 10.1186/s12859-019-2914-2 · PMID: 31238875 · PMCID: PMC6593489

12. **metapath2vec: Scalable Representation Learning for Heterogeneous Networks**
Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami
KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017-08) https://doi.org/gfsqzn
DOI: 10.1145/3097983.3098036

13. **HINE: Heterogeneous Information Network Embedding**
Yuxin Chen, Chenguang Wang
Lecture Notes in Computer Science (2017) https://doi.org/gg2c7t
DOI: 10.1007/978-3-319-55753-3_12

14. **Discovering Meta-Paths in Large Heterogeneous Information Networks**
Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, Wangda Zhang
Association for Computing Machinery (ACM) (2015) https://doi.org/gg2c7v
DOI: 10.1145/2736277.2741123

15. **MetaExp: Interactive Explanation and Exploration of Large Knowledge Graphs**
Freya Behrens, Fatemeh Aghaei, Emmanuel Müller, Martin Preusse, Nikola Müller, Michael Hunger, Sebastian Bischoff, Pius Ladenburger, Julius Rückin, Laurenz Seidel, … Davide Mottin
WWW '18: Companion Proceedings of the The Web Conference 2018 (2018-04)
https://doi.org/gg2c7w
DOI: 10.1145/3184558.3186978

16. **Transforming DWPCs for hetnet edge prediction**
Daniel Himmelstein, Pouya Khankhanian, Antoine Lizee
ThinkLab (2016-04-01) https://doi.org/f3qbmd
DOI: 10.15363/thinklab.d193

17. **Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes**
Daniel S. Himmelstein, Sergio E. Baranzini
PLOS Computational Biology (2015-07-09) https://doi.org/98q
DOI: 10.1371/journal.pcbi.1004259 · PMID: 26158728 · PMCID: PMC4497619