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

This manuscript (permalink) was automatically generated from greenelab/connectivity-search-manuscript@0622e9d on June 8, 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 [**???**]. 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 [4], SemMedDB [5,6,7], SPOKE, and DRKG.

Rephetio

Unsupervised connectivity search

Results

Hetmatpy Package

DWPC null distribution

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

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. **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

5. **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

6. **Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data**
   Qing Cong, Zhiyong Feng, Fang Li, Li Zhang, Guozheng Rao, Cui Tao
   Institute of Electrical and Electronics Engineers (IEEE) (2018-12) https://doi.org/ggzb26
   DOI: 10.1109/bibm.2018.8621568

7. **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