BIOKG: A KNOWLEDGE GRAPH FOR RELATIONAL LEARNING ON BIOLOGICAL DATA

Brian Walsh, Sameh K. Mohamed, Vít Nováček

[1] https://dl.acm.org/doi/10.1145/3340531.3412776

Content

Motivation

Knowledge graphs have become a popular choice for modelling complex biological systems.

Vast array of open biological data sources available.

There is a lack of open biological knowledge graphs to support relational learning.

A standardised, easily extensible approach to linking entities in different data sources is also missing.

Related Works

Các nguồn dữ liệu sinh học ngày nay đa dạng gồm nhiều thực thể và quy trình sinh học, nhưng lại nhiều định dạng khác nhau, sử dụng lược đồ nhận dạng khác nhau và có thể chứa dữ liệu trùng lặp khiến việc kết hợp các bộ dữ liệu này trở nên khó khăn.

Mỗi database tập trung vào mỗi loại thực thể sinh học khác nhau. Ví dụ, UniProt có Speciaity là: protein, trong khi DrugBank là thuốc. -> Không có một database nào có thể bao quát toàn bộ các khía cạnh của sinh học.

	Properties		Entity coverage							
Database Name	Format	Speciality	Proteins	Drugs	Indications	Diseases	Gene Ontology	Expressions	Pathways	In BioKG ?
UniProt [7]	S/U	PR	✓	✓	✗	✓	✓	✓	✓	✓
REACTOME [8]	S	PA	✓	✗	✗	✗	✓	✗	✓	✓
KEGG [14]	S	PA	✓	✓	✗	✓	✗	✗	✓	✓
DrugBank [15]	S/U	DR	✓	✓	✗	✗	✗	✗	✓	✓
GO [5]	S	GO	✓	✗	✗	✗	✓	✗	✓	✓
CTD [19]	S/U	CH	✓	✓	✗	✗	✓	✗	✓	✓
SIDER [16]	S	DR	✗	✓	✓	✗	✗	✗	✗	✓
HPA [36]	S/U	PR	✓	✗	✗	✗	✓	✓	✗	✓
STRING [33]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
BIOGRID [32]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
IntAct [30]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
InterPro [21]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
PharmaGKB [10]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
TTD [17]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Supertarget [9]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Cellosaurus [2]	S/U	CL	✗	✗	✗	✗	✗	✓	✗	✓
MESH [1]	S/U	CL	✗	✗	✗	✓	✗	✗	✗	✓

Related Works

Các database sử dụng các hệ thống định danh khác nhau cho cùng một thực thể sinh học dẫn đến khó khăn trong việc tích hợp dữ liệu từ nhiều nguồn khác nhau.

Ví dụ, để định danh cho protein:

UniProt sử dụng "UniProt Accessions“

KEGG, CTD sử dụng "Gene Id Numbers“.

	Properties		Entity coverage							
Database Name	Format	Speciality	Proteins	Drugs	Indications	Diseases	Gene Ontology	Expressions	Pathways	In BioKG ?
UniProt [7]	S/U	PR	✓	✓	✗	✓	✓	✓	✓	✓
REACTOME [8]	S	PA	✓	✗	✗	✗	✓	✗	✓	✓
KEGG [14]	S	PA	✓	✓	✗	✓	✗	✗	✓	✓
DrugBank [15]	S/U	DR	✓	✓	✗	✗	✗	✗	✓	✓
GO [5]	S	GO	✓	✗	✗	✗	✓	✗	✓	✓
CTD [19]	S/U	CH	✓	✓	✗	✗	✓	✗	✓	✓
SIDER [16]	S	DR	✗	✓	✓	✗	✗	✗	✗	✓
HPA [36]	S/U	PR	✓	✗	✗	✗	✓	✓	✗	✓
STRING [33]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
BIOGRID [32]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
IntAct [30]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
InterPro [21]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
PharmaGKB [10]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
TTD [17]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Supertarget [9]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Cellosaurus [2]	S/U	CL	✗	✗	✗	✗	✗	✓	✗	✓
MESH [1]	S/U	CL	✗	✗	✗	✓	✗	✗	✗	✓

Related Works

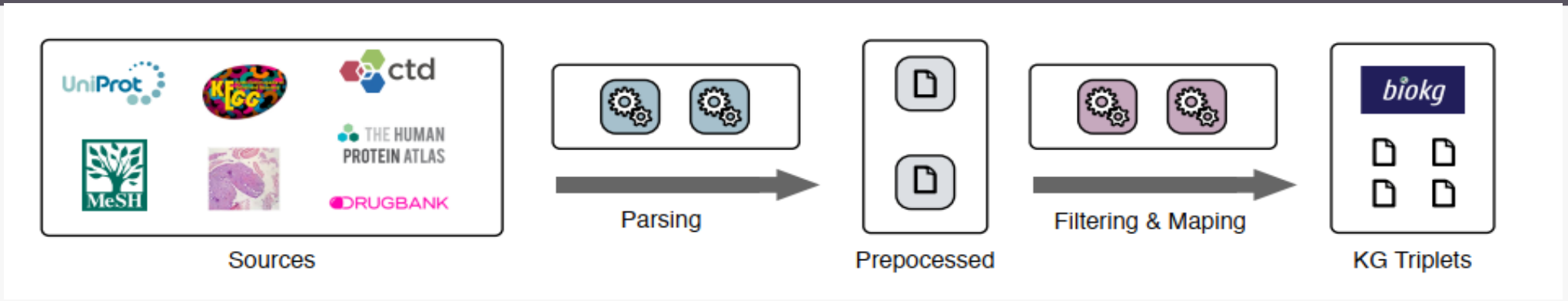
Một số cơ sở dữ liệu có dữ liệu được soạn từ các chuyên gia (như SwissProt trong UniProt), trong khi các cơ sở dữ liệu khác có thể chứa dữ liệu được tạo ra từ các kỹ thuật suy diễn -> Có sự khác biệt về chất lượng.

UniProt và DrugBank được sử dụng nhiều trong nghiên cứu và phát triển thuốc.

Các cơ sở dữ liệu khác có thể ít được biết đến hơn nhưng vẫn cung cấp nhiều thông thông tin cho nghiên cứu.

	Properties		Entity coverage						
Database Name	Format	Speciality	Proteins	Drugs	Indications	Diseases	Gene Ontology	Expressions	Pathways	In BioKG ?
UniProt [7]	S/U	PR	✓	✓	✗	✓	✓	✓	✓	✓
REACTOME [8]	S	PA	✓	✗	✗	✗	✓	✗	✓	✓
KEGG [14]	S	PA	✓	✓	✗	✓	✗	✗	✓	✓
DrugBank [15]	S/U	DR	✓	✓	✗	✗	✗	✗	✓	✓
GO [5]	S	GO	✓	✗	✗	✗	✓	✗	✓	✓
CTD [19]	S/U	CH	✓	✓	✗	✗	✓	✗	✓	✓
SIDER [16]	S	DR	✗	✓	✓	✗	✗	✗	✗	✓
HPA [36]	S/U	PR	✓	✗	✗	✗	✓	✓	✗	✓
STRING [33]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
BIOGRID [32]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
IntAct [30]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
InterPro [21]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
PharmaGKB [10]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
TTD [17]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Supertarget [9]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Cellosaurus [2]	S/U	CL	✗	✗	✗	✗	✗	✓	✗	✓
MESH [1]	S/U	CL	✗	✗	✗	✓	✗	✗	✗	✓

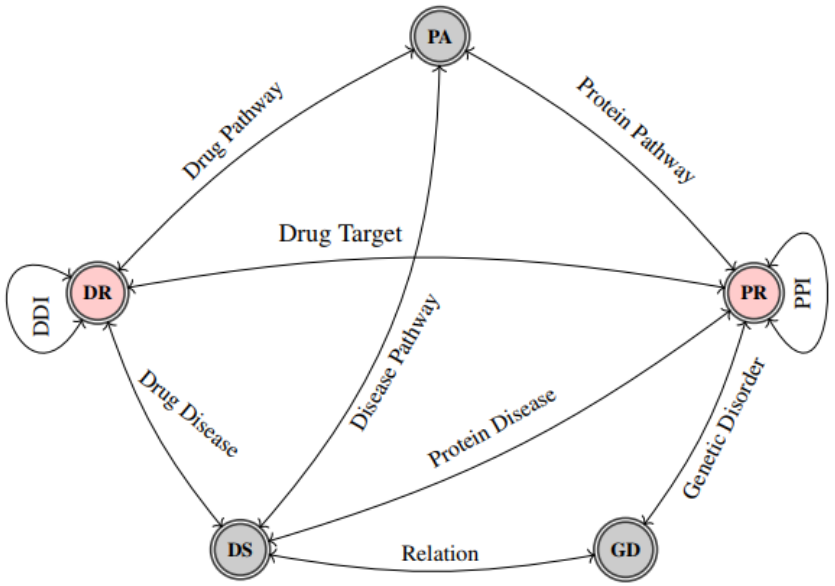
BioKG



Hình 2: Minh họa quy trình quy trình xây dựng dữ liệu của đồ thị tri thức BioKG. Quy trình này giúp dữ liệu được tích hợp một cách có hệ thống gồm 2 giai đoạn:

- Parsing: Thu thập, phân tích và chuyển đổi dữ liệu thành các định dạng cấu trúc trung gian từ các cơ sở dữ liệu sinh học mã nguồn mở.
- Compiling: thực hiện các bước mapping, lọc dữ liệu để xây dựng các Triplet từ dữ liệu bước Parsing, tạo ra một đồ thị tri thức BioKG

BioKG



Hình 1 Mô tả sơ đồ của các thực thể sinh học và các mối quan hệ giữa chúng trong đồ thị tri thức BioKG.

1.Protein (PR): Đại diện cho các protein.

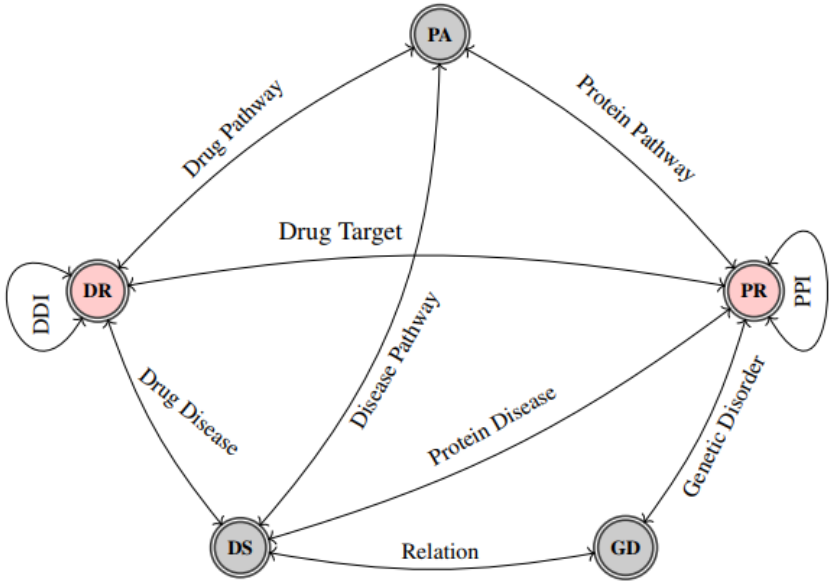
2.Drug (DR): Đại diện cho các loại thuốc, có thể tương tác với các protein và ảnh hưởng đến các quá trình sinh học.

3.Disease (DS): Đại diện cho các bệnh lý mà các protein có thể liên quan.

4.Genetic Disorder (GD): Liên quan đến các rối loạn di truyền có thể ảnh hưởng đến chức năng của protein.

5.Pathway Associations(PA): Đại diện cho các con đường sinh học, là các chuỗi các phản ứng sinh hóa mà protein tham gia.

BioKG



Hình 1 Mô tả sơ đồ của các thực thể sinh học và các mối quan hệ giữa chúng trong đồ thị tri thức BioKG.

Các Mối Quan Hệ

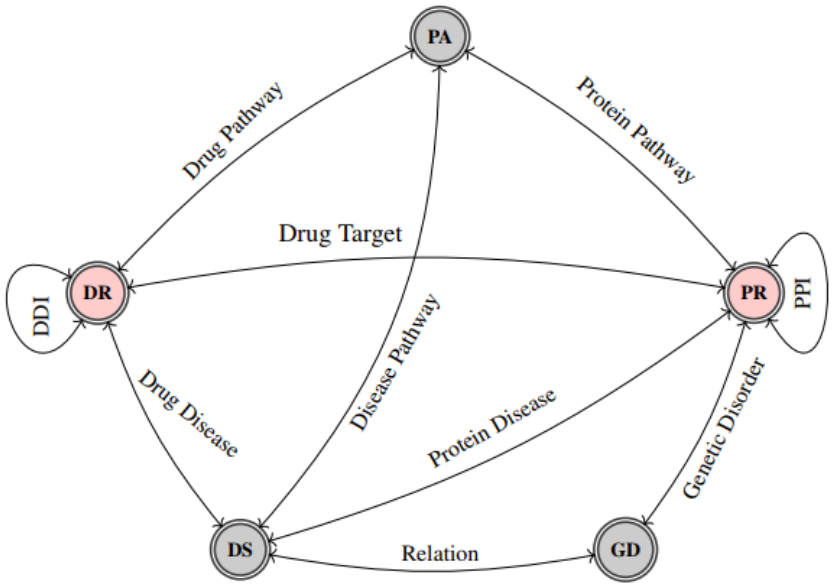
Protein-Protein Interactions (PPI): Mối quan hệ giữa các protein tương tác với nhau.

Drug-Protein Interactions: Mối quan hệ giữa thuốc và protein, cho thấy cách thuốc tác động lên protein.

Drug-Drug Interactions (DDI): Mối quan hệ giữa các loại thuốc tương tác với nhau.

Protein-Disease Relationships: Mối quan hệ giữa protein và bệnh, cho thấy protein nào có thể liên quan đến bệnh nào.

BioKG



Hình 1 Mô tả sơ đồ của các thực thể sinh học và các mối quan hệ giữa chúng trong đồ thị tri thức BioKG.

Các Mối Quan Hệ

Disease-Genetic Disorder Associations: Mối quan hệ giữa bệnh và rối loạn di truyền, cho thấy các bệnh nào có thể liên quan đến các rối loạn di truyền.

Protein-Pathway Associations: Mối quan hệ giữa protein và các con đường sinh học, cho thấy protein nào tham gia vào con đường sinh học nào.

Drug-Pathway Associations: Mối quan hệ giữa thuốc và các con đường sinh học, cho thấy thuốc nào có thể ảnh hưởng đến con đường sinh học nào.

BioKG

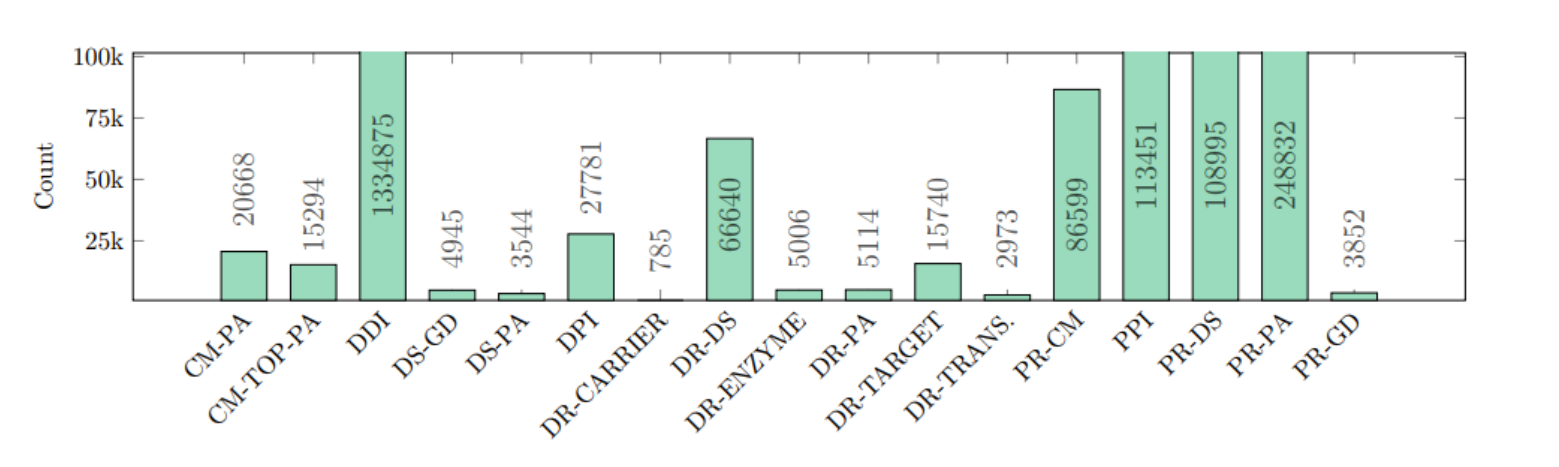
Cấu trúc trọng tâm của đồ thị tri thức BioKG bao gồm:

- **Links**: mô hình hóa các mối quan hệ giữa các thực thể.
- **Properties**: chứa các thông tin liên quan đến các thực thể sinh học và các thuộc tính của chúng. Ví dụ:

  Protein Attributes: Bao gồm các liên kết với Gene Ontology và các chú thích về chuỗi protein.

  Drug Properties: Các thuộc tính của thuốc như tác dụng phụ, chỉ định và mã phân loại ATC.
- **Metadata**: chứa thông tin về tên, loại, từ đồng nghĩa của các thực thể sinh học, giúp tăng cường độ phong phú của thông tin về các thực thể sinh họ.

BioKG



Phần lớn các mối quan hệ tập trung vào các thực thể thuốc và protein dẫn đến sự mất cân bằng trong các mối quan hệ.

Đây là một triệu chứng của sự mất cân bằng trong nghiên cứu Trọng tâm nơi một số sinh học các thực thể liên quan đến các hiện tượng sinh học phổ biến được nghiên cứu kỹ lưỡng dẫn đến các tập dữ liệu và chú thích lớn hơn và phong phú hơn.

BENCHMARKS

Trình bày năm tập dữ liệu chuẩn được cung cấp cùng với BioKG. Những tập dữ liệu này tập trung vào việc phát hiện mục tiêu thuốc và tương tác thuốc-thuốc. Mục đích là cung cấp các tập dữ liệu để huấn luyện và đánh giá các mô hình ML.

DDI-MINERAL

DDI-EFFICACY

DPI-FDA

DPI-FDA-EXP

PPI-PHOSPHO

THANK YOU

Bui Minh Phung

phungbm.work@gmail.com

www.github.com/phungbminh