Computational Molecular Biology and Bioinformatics

iKraph

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

October, 2025

Knowledge graph

A knowledge graph (KG) is a structured representation of information that captures entities (things, people, places, concepts) and the relationships between them in a graph format.

The KGs enable efficient information retrieval and automated knowledge discovery. However, transforming unstructured scientific literature into KGs remains a significant challenge.

Biological knowledge graph

A biological knowledge graph is a structured network that represents biological entities and their relationships in a graph format. It may connect data from diverse biological sources – genes, proteins, diseases, drugs, pathways, phenotypes, and more – into an integrated, machine-readable framework.

Such knowledge graphs comprises the following components:

- Nodes (entities): Biological objects such as genes, proteins, drugs, diseases, pathways, organisms, phenotypes, etc.
- Edges (relations): Interactions or associations between entities, such as "gene encodes protein", "drug treats disease", "protein interacts with protein", "gene involved in pathway", etc.

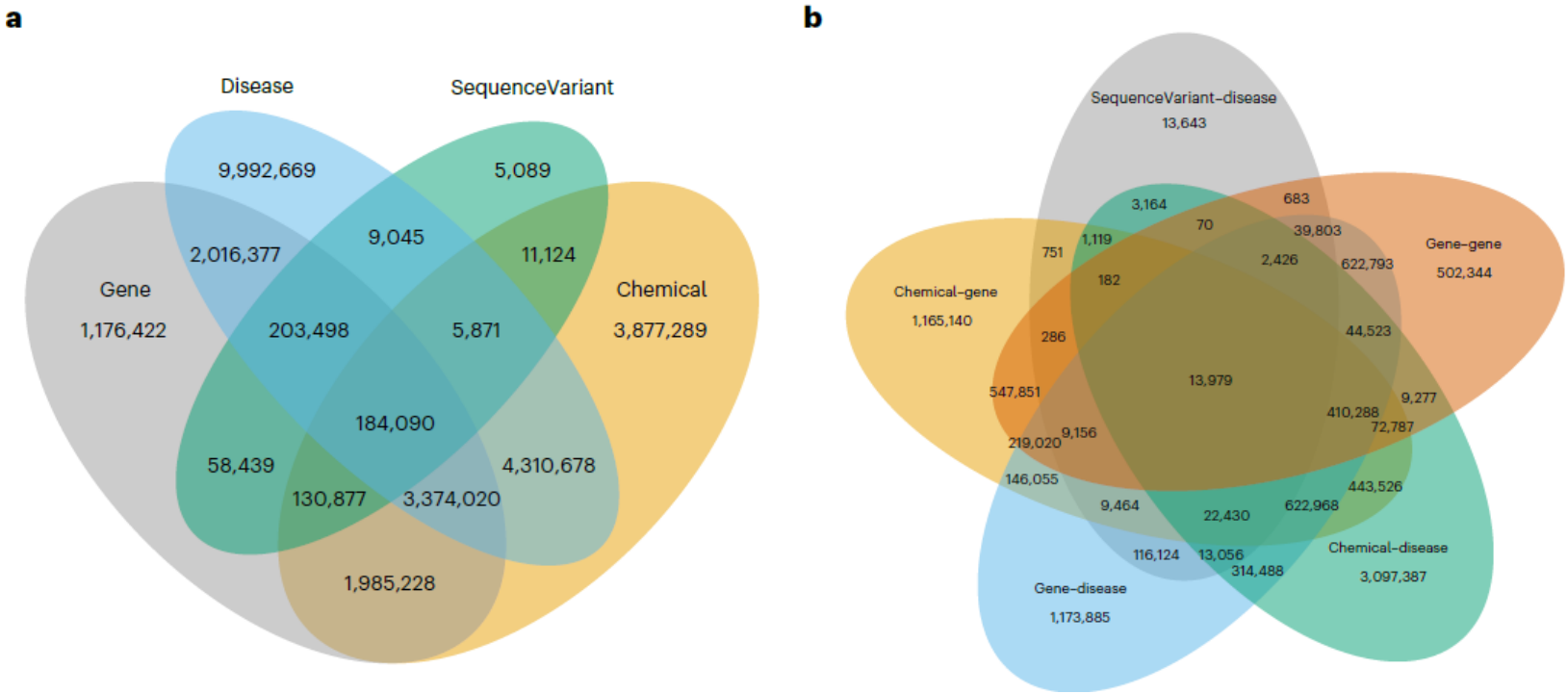
The iKraph

iKraph is a large-scale KG covering all the scientific abstracts available on PubMed, which contains more than 34 million abstracts, resulting in 10,686,927 unique entities and 30,758,640 unique relations [1]. It was constructed using an information extraction pipeline that won first place in the LitCoin Natural Language Processing Challenge (2022). To enhance the

comprehensiveness of the KG, the relation data from 40 public databases and the relation information inferred from high-throughput genomic data were integrated into iKraph. A cloud-based platform (`https://biokde.insilicom.com`) was also developed for academic users to access this rich structured data and associated tools.

PubMed coverage

The number of PubMed articles included in iKraph that consist of certain types of (a) entities and (b) relations are shown below.



The construction of iKraph

The construction of iKraph involves three primary stages: named entity recognition (NER), relation extraction and novelty classification.

Constructing a causal KG

To infer causal relations, causal direction for 4,572 relations in the LitCoin dataset were annotated. Among them, 2,009 cases have direction from the first entity to the second, 1,611 cases have direction from the second entity to the first and 952 cases have no direction.

This annotation allowed to train a model for predicting the directions for relations, which achieved an F1 score of 0.924 in a 5-fold cross-validation test on the LitCoin dataset. Using a causal KG, we can infer indirect causal relations more effectively for entities not directly connected in the KG.

Probabilistic inference

For inferring the indirect relation from A to C using direct relations from A to B and the relation from B to C, we first extract the two direct relations. Notably, relation A to B and B to C will probably occur many times in different PubMed abstracts. So, we calculate the overall probability of whether two entities have a particular relation using the following formula

$$P_{A \to B} = 1 - \prod_{i=1}^{N} (1 - p^{i}_{A \to B}),$$

where $P_{A \to B}$ is the overall probability of A-B entity pair having a particular relation and $p^{i}_{A \to B}$ is the probability of being true for the i^{th} occurrence of these two entities in a PubMed abstract.

Outcome from probabilistic inference

It also introduces an interpretable and probabilistic inference method to identify indirect causal relations and applied it to real-time COVID-19 drug repurposing between March 2020 and May 2023.

Around 1,200 candidate drugs were identified in the first 4 months, with one-third of those discovered in the first 2 months later supported by clinical trials or PubMed publications.

References

1. Zhang, Y., Sui, X., Pan, F., Yu, K., Li, K., Tian, S., Erdengasileng, A., Han, Q., Wang, W., Wang, J. and Wang, J., A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. Nature Machine Intelligence, 7:602-614, 2025.