MetaPath2Vec in LUCIA: Lung Cancer Risk Factor Discovery through Heterogeneous Graphs

As a contribution to the LUCIA project, researchers at the Center for Biomedical Technology (CTB) at Universidad Politécnica de Madrid (UPM) developed a heterogeneous knowledge graph that combines clinical and biological data. The biological data comes from multiple public sources including DISNET, WikiPathways, Mitelman, and DisGeNET [1], [2], [3]. The resulting graph connects seven entity types: patients, diseases, proteins, pathways, drugs, clinical measurements and chromosomal rearrangements. The disease layer includes a wide range of conditions, among them lung cancer and its different subtypes.

To discover lung cancer risk factors within this graph, MetaPath2Vec [4] was used. MetaPath2Vec learns vector representations (embeddings) on heterogeneous graphs by guiding random walks through **metapaths:** predefined sequences of node types that define semantically meaningful routes. By restricting each step of the walk to a specific node type, the algorithm ensures that the resulting embeddings capture biologically coherent co-occurrence patterns rather than arbitrary structural proximity. The metapaths selected for this analysis are:

- **PERSON-DISEASE-PERSON-DISEASE-PROTEIN** for protein discovery,
- **PERSON-DISEASE-PERSON-DISEASE-PROTEIN-PATHWAY** for pathway discovery, and
- **PERSON-DISEASE-PERSON-DISEASE-PROTEIN-CHROMOSOMAL REARRANGEMENT** for chromosomal rearrangement discovery.

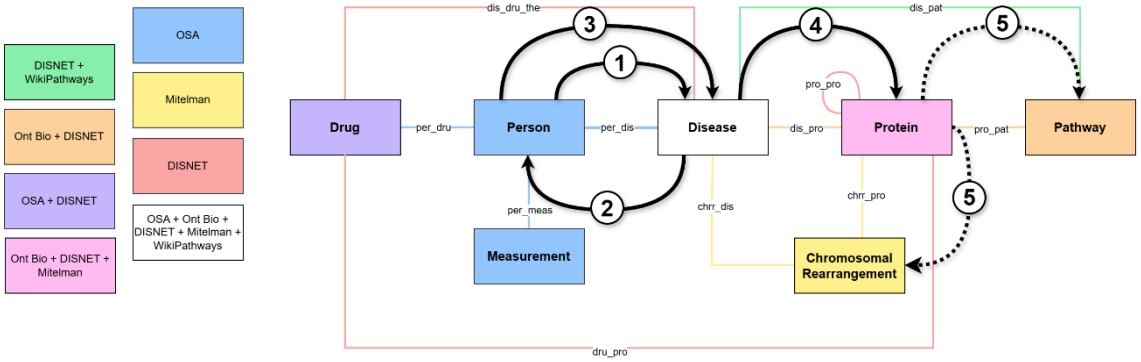

Figure 1. Schema of the heterogeneous knowledge graph developed. Nodes represent the seven entity types integrated in the graph: person, disease, protein, pathway, chromosomal Rearrangement, drug and measurement. Edges represent the semantic relations connecting them, including *per_dis*, *dis_pro*, *chrr_pro*, *pro_pat*, and *per_meas*, among others. The colored boxes on the left reflect the multi-source nature of the biological data. The numbered arrows indicate the sequential steps of the selected metapaths: starting from a Person node, step (1) traverses to a Disease, step (2) returns to a second Person who shares that disease, step (3) moves to another Disease of that second person, step (4) reaches a Protein associated with that disease, and step (5) arrives at either a Pathway or a Chromosomal Rearrangement connected to that protein, depending on the target entity type.

These routes share a common structural logic: starting from a patient, they traverse to a disease, then to a second patient who shares that disease, and from there to another disease before arriving at the target biological entity. This two-hop patient bridge captures second-order patient similarity, meaning the model identifies biological entities that characterize patients who are clinically related through shared disease context, rather than relying solely on direct database associations. For pathways and chromosomal rearrangements, the additional protein intermediate grounds the association in molecular mechanism, ensuring that the final

embedding reflects not just disease-level links but the underlying protein biology connecting patients to these entities.

The resulting embeddings are evaluated via cosine similarity against three complementary methods. Method 1 measures the similarity of each candidate factor to the Lung Cancer disease node embedding, answering the question of which entities are structurally closest to lung cancer in the learned latent space. Method 2 computes the mean embedding of all lung cancer patients and measures proximity to that centroid, capturing which factors are most associated with the clinical profile of the patient population. Method 3 computes the mean embedding of a curated set of known oncogenic driver proteins (including EGFR, KRAS, ERBB2, BRAF, ALK, MET, ROS1, TP53, RET, STK11, RB1) and measures which factors cluster near that biological reference. Each method produces an independent ranked list, and the three rankings are then fused via Reciprocal Rank Fusion (RRF), a technique that combines rankings based solely on position rather than raw scores, making the final result robust to scale differences between methods.

This approach combines different sources of evidence to identify risk factor candidates that are consistently relevant across the network. The resulting ranked lists of proteins, pathways, and chromosomal rearrangements provide a data-driven starting point for investigating and validating potential lung cancer risk factors.

References

[1] *A. Agrawal et al., "WikiPathways 2024: next generation pathway database," Nucleic Acids Res., vol. 52, no. D1, pp. D679–D689, Jan. 2024, doi: 10.1093/nar/gkad960.*

[2] *F. Mitelman, B. Johansson, and F. Mertens, "Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer." [Online]. Available: https://mitelmandatabase.isb-cgc.org/*

[3] *J. Piñero et al., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," Nucleic Acids Res., vol. 45, no. D1, pp. D833–D839, Jan. 2017, doi: 10.1093/nar/gkw943.*

[4] *Y. Dong, N. V Chawla, and A. Swami, "metapath2vec: Scalable Representation Learning for Heterogeneous Netwo rks," in KDD'17, ACM, 2017, pp. 135–144. doi: 10.1145/3097983.3098036.*

The LUCIA project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101096473. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.