ML RESEARCH ENGINEER TECHNICAL ASSESSMENT

PATHWAY AND DISEASE HYPOTHESIS GENERATION

1. OBJECTIVE

At Causaly, we are revolutionizing the way researchers and healthcare professionals access and interpret biomedical information. Our mission is to make biomedical knowledge more accessible and actionable by leveraging state-of-the-art technologies in Generative AI, natural language processing and knowledge graphs. In the ever-expanding field of biomedical research, staying up-to-date with the latest findings and deriving meaningful insights is a daunting task. By automating the extraction of structured knowledge and enabling natural language querying, we can significantly reduce the time and effort required for literature review and data analysis. This assignment aims to explore and implement innovative approaches that align with our mission to empower users with comprehensive, accurate, and easily accessible biomedical information.

This assignment involves creating an LLM-powered agentic workflow that generates hypotheses about the roles of specific genes in diseases, utilizing structured information from KEGG and Gene Ontology (GO).

This is a challenging and fairly open-ended task – therefore we provide indicative steps to help you achieve a working state. Do not focus on optimizing your approach or selecting the best LLM; we are mostly interested in your thinking process.

Background

Pathway analysis plays a crucial role in understanding the complex biological processes underlying diseases. By examining the interactions and pathways that genes and proteins participate in, researchers can uncover the mechanisms that drive disease progression and identify potential targets for therapeutic intervention. This analysis helps in mapping out the cascade of molecular events triggered by genetic and environmental factors, providing a comprehensive view of how diseases develop and progress. By integrating data from various sources, pathway analysis enables the identification of key regulatory nodes and potential biomarkers, thereby facilitating the development of more effective diagnostic and treatment strategies.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Gene Ontology (GO) are two essential resources for pathway analysis in disease biology. KEGG is a comprehensive database that provides information on molecular interaction networks, including pathways related to metabolism, genetic information processing, environmental information processing, and cellular processes. It offers a detailed map of biochemical pathways and their components, helping researchers understand how alterations in these pathways can lead to disease. The Gene Ontology, on the other hand, provides a framework for the representation of gene and gene product attributes across all species. GO categorizes genes and proteins based on their associated biological processes, cellular components, and molecular functions, enabling a standardized and systematic approach to functional annotation. Together, KEGG and GO provide invaluable tools for researchers aiming to decipher the molecular underpinnings of diseases and identify new avenues for intervention.

2. ASSIGNMENT STEPS

A. DATA RETRIEVAL:

Objective: Datasets from KEGG and GO.

Steps:

A1. KEGG Pathways:

You will be provided with the KGML (xml) files of the following specific disease pathways from KEGG:

1. Alzheimer's disease (hsa05010)
2. Parkinson's disease (hsa05012)
3. Type II diabetes mellitus (hsa04930)
4. Colorectal cancer (hsa05210)

(We also include the PNG versions of the aforementioned pathways for illustrative purposes, but they are not required for this assignment.)

A2. Gene Ontology (GO):

1. Go to Gene Ontology Downloads and download the "Gene Association File (GAF)" for Homo sapiens. This file includes gene-to-GO term associations necessary for linking genes to biological processes and diseases.

B. STRUCTURED KNOWLEDGE EXTRACTION:

Objective: Process and structure data from downloaded files for hypothesis generation.

Steps:

B1. Parse KGML files: Extract gene-pathway relationships and associated diseases from the KGML files. The KGML (KEGG Markup Language) files can be parsed using XML parsers available in languages like Python (e.g., `xml.etree.ElementTree` or `lxml`).

B2. Parse GAF files: Extract gene-GO term associations from the GAF files. The GAF files are typically tab-delimited text files that can be processed using pandas or other data processing libraries in Python.

B3. Integrate data: You can either save the extracted relationships in text format or construct a relational database or graph structure (using a graph database like Neo4j or a library like NetworkX) that links genes, pathways, and GO annotations.

C. HYPOTHESIS GENERATION AGENT:

Objective: Develop a system that formulates hypotheses about gene involvement in specific diseases.

Steps:

C1. Develop the agent: Implement an LLM-powered agent that understands the aforementioned data and queries the structured knowledge base to generate hypotheses about the potential involvement of genes in specific pathways and diseases. You are free to use any LLM (open-source or proprietary via API) and agentic framework you like. The agent should process queries like "What diseases is gene X associated with?" and provide informed hypotheses based on the data. The agent can have access to different “tools” that will enable it to plan specific actions and/or reflect on them.

C2. Implement downstream analysis: Enable the agent to predict downstream gene interactions and their possible effects on disease pathways, utilizing network analysis techniques. It is important that the agent can connect information from KEGG to the relevant information and associations from GO to enable basic reasoning across the two databases.

Note: Downstream gene interactions refers to the cascade of events that occur as a result of the activation or repression of a particular gene. When a gene is expressed, it produces a specific RNA transcript, which is then translated into a protein. This protein can act as a transcription factor, enzyme, or structural component, influencing the activity of other genes further along in a signaling pathway or regulatory network.

Optional: Enable the agent to process queries involving multiple genes and offer insights into their combined impact on biological processes and diseases.

3. DELIVERABLES

- **GitHub Repository**: Contains all code, data processing scripts, and query handling mechanisms. A README file that guides setup, operation, and explains the system architecture, and data flow. The code should be reproducible by following the README instructions. Make sure to remove any API keys or other sensitive information from your codebase.
- A brief **report/presentation** (pdf/pptx) with examples, showcasing the system's capabilities via several queries and the corresponding hypotheses generated. You can also include any limitations of your approach.

4. EVALUATION CRITERIA

- **Functionality**: Accurate data retrieval, robust agentic workflow and effective hypothesis generation.
- **Code Quality**: Well-organized, documented, and reproducible code.
- **Innovation**: Creative use of data and agentic techniques to generate hypotheses.

5. TIMELINE

The task should be completed within 7 days.