IMPLEMENTATION OF CROSSBAR WEB SERVICE FOR INTERACTIVE VISUALIZATIONS OF BIOLOGICAL NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET ATAKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2020

Approval of the thesis:

IMPLEMENTATION OF CROSSBAR WEB SERVICE FOR INTERACTIVE VISUALIZATIONS OF BIOLOGICAL NETWORKS

submitted by **AHMET ATAKAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. M. Volkan Atalay
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Prof. Dr. M. Volkan Atalay
Computer Engineering, METU _____

Assoc. Prof. Dr. Tunca Doğan
Computer Engineering, Hacettepe University _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ahmet Atakan

Signature :

ABSTRACT

IMPLEMENTATION OF CROSSBAR WEB SERVICE FOR INTERACTIVE VISUALIZATIONS OF BIOLOGICAL NETWORKS

Atakan, Ahmet
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. M. Volkan Atalay

September 2020, 61 pages

Existing computational tools, resources, and services to assist experimental biomedical research lack data diversity and data connectivity and therefore, they are limited in helping to solve real world problems. Within the framework of CROssBAR project aiming for drug discovery, CROssBAR database (CROssBAR-DB) integrated diverse biomedical resources with the predictions of a comprehensive computing resource developed using machine learning and deep learning-based methods. It is crucial for the users to perform easily search on CROssBAR-DB, and then to visualize the result of the search. For this reason, CrossBAR web service (CROssBAR-WS) is designed, developed and implemented within the scope of this thesis. In CROssBAR-WS, knowledge graphs can be generated by using the open-source graph library Cytoscape.js. CROssBAR-WS is available at crossbar.kansil.org.

Keywords: CROssBAR, graph visualization, protein interactions, cytoscape.js

ÖZ

BİYOLOJİK AĞLARIN İNTERAKTİF GÖRSELLEŞTİRİLMESİ İÇİN CROSSBAR WEB SERVİSİNİN İMPLEMENTASYONU

Atakan, Ahmet
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. M. Volkan Atalay

Eylül 2020 , 61 sayfa

Deneysel biyomedikal araştırmalara yardımcı olmak için mevcut işlemsel araçlar ve hizmetlerin, veri çeşitliliği ve veri bağlantılılığı açılarından eksiklikleri bulunduğundan dolayı, gerçek problemlere çözüm üretme noktasında sınırlı kalmaktadır. Çeşitli biyomedikal kaynaklarla birlikte yapay ve derin öğrenmeye dayalı yöntemlerle geliştirilen kapsamlı bir hesaplama kaynağına dayalı tahminler, ilaç keşfi problemi için oluşturulan CROssBAR projesi kapsamındaki CROssBAR veritabanına (CROssBAR-VT) entegre edilmiştir. Kullanılar açısından CROssBAR-VT'de kolayca arama yapmak ve ardından arama sonuçlarının görsellenmesi büyük önem arz etmektedir. Bundan dolayı CROssBAR ağ servisinin tasarımı, geliştirilmesi ve gerçekleştirilmesi bu tez kapsamında yapılmıştır. CROssBAR ağ servisinde oluşturulan bilgi grafları, açık kaynak kodlu çizge kütüphanesi olan Cytoscape.js kullanırak üretilmiştir. CROssBAR ağ servisine crossbar.kansil.org bağlantısı üzerinden ulaşılabilir.

Anahtar Kelimeler: CROssBAR, graf görselleştirme, protein etkileşimleri, cytoscape.js

to my family

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my supervisor, Prof. Dr. Volkan Atalay, for his guidance, support, and patience during this thesis. I also thank Prof. Dr. Rengül Çetin-Atalay for her priceless advice and critique for the biological aspect of this work. I'm delighted working with them.

I am much obliged to Assoc. Prof. Dr. Tunca Doğan for his fruitful comments, advice, and productive criticism to enhance my study. I would also like to thank Ahmet Rifaioğlu, Gökhan Özsarı, Heval Ataş and Esra Sinoplu for their valuable support to overcome the problems when I am puzzled.

I want to thank my friends Anıl Çetinkaya, Alper Karamanlıoğlu, M.Çağrı Kaya, Özcan Dülger, Alperen Dalkıran, Alperen Eroğlu, Murat Öztürk and Tuğberk İşyapar. It has always been exceptional to spend time with you.

Finally, I would like to thank my family for their continuous support during my study. I'm especially grateful to my mother Nurten Atakan and my father Osman Suat Atakan.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	
CROssBAR	Comprehensive Resource of Biomedical Relations with Deep Learning Applications and Knowledge Graph Representations
KEGG	Kyoto Encyclopedia of Genes and Genomes
OMIM	Online Mendelian Inheritance in Man
EFO	Experimental Factor Ontology
UniProt	The Universal Protein Resource
UniProtKB	The UniProt Knowledgebase
API	An Application Programming Interface
DTI	Drug–Target Interaction
KG	Knowledge Graph
NCBI	The National Center for Biotechnology Information

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

The CROssBAR is a database that offers biologists the relationships between critical biological data for drug discovery researchers. It is an enriched database in terms of finding relationships between biological entities. However, the user must be a database expert to directly use the information from the CROssBAR database (CROssBAR-DB). It is necessary to visualize the data meaningfully and adequately in terms of knowledge graphs to facilitate users' interpretation. Thus, our web service (CROssBAR-WS) allows users to visualize biomedical data without any programming and database knowledge.

1.2 Improvements

In the scope of this study, we developed the CROssBAR web service (CROssBAR-WS) (https://crossbar.kansil.org) that visually presents the data in the CrossBar-DB in an easily interpretable and interactive manner. CROssBAR-WS is a service for researchers with no programming background to investigate data gathered from various publicly available databases. CROssBAR-WS effectively visualizes heterogeneous biological networks that include direct or indirect relations among genes/proteins, drugs, diseases, pathways, and HPO Terms. These intensely-processed heterogeneous biological networks are expected to aid biomedical research, especially to infer mechanisms of diseases concerning biomolecules, systems, and candidate drugs.

The term knowledge graph (KG) defines a specialized data structure, in which a col-

lection of entities (nodes) are linked to each other (edges) in a semantic context. In this study, we chose to represent heterogeneous biomedical data in terms of knowledge graphs. Knowledge graphs (KG) are presented visually on web-browsers as flexible Cytoscape networks. Users can create queries with biomedical terms, individually or in combination, to obtain the relevant graph. The queries obtained as a combination of biological terms are especially critical as they provide the ability to investigate the indirect biological relationships between the terms from both the same and different biomedical components. Since there are billions of ways to query in CROssBAR-WS, it was not feasible to pre-calculate the resulting graphs; therefore, the graphs are constructed on-the-fly, in real-time. Several options are provided to users to customize their queries in a search on CROssBAR-WS. Some sample options are as follows: UniProt databases to be used (UniProtKB/Swiss-Prot or UniProtKB/Swiss-Prot+UniProtKB/TrEMBL), taxonomic identifiers to be included, and the number of terms/nodes to be included from each entity type (selected from enrichment score-based ranked lists). It is also possible to display the graph using various layout options, including our in-house CROssBAR-layout. Saving options let users to store the graph in various formats, including JSON, figure-ready snapshots, and protein-centric delimited data-tables. The interactive visualization also lets users prepare a custom display by relocating the nodes as desired.

1.3 Organization

In CROssBAR knowledge graphs (CROssBAR-KG), biological components/terms (i.e., drugs/compounds, genes/proteins, bio-processes/pathways, phenotypes/diseases) are represented as nodes, and their known or predicted pairwise relationships are annotated as edges (a protein and its coding gene is treated as one merged term/entry/node). These biological components are described in Chapter 2.

CROssBAR-KGs display the direct and indirect relations between all of the terms in the graph. During the construction of graphs, terms (nodes) and their pairwise relations (edges) are directly obtained from the CROssBAR-database (DB). CROssBAR database is a document-oriented database that is one of the most comprehensive and large-scale integrated databases in computational systems biology, as it is explained

in Chapter 3.

At each step of constructing a knowledge graph, an overrepresentation-based enrichment analysis has been performed to select the terms that are significantly associated with the growing graph and discard the rest. This analysis comprises a series of hypergeometric tests, based on the recorded relations in the CROssBAR-DB. Here, we applied a layered construction approach, always taking the genes/proteins at the center of enrichment analysis. Finally, additional relation types are incorporated to the graph as edges between the existing nodes (e.g., drug-disease, disease-pathway and disease-HPO), to further enrich the provided information. All steps of the construction of knowledge graphs and enrichment analysis are detailed in Chapter 4.

CHAPTER 2

BACKGROUND INFORMATION AND RELATED WORK

2.1 Biological Background

In this section, we explain node types used in CROssBAR-WS and their meanings.

2.1.1 Proteins

Proteins are large molecules constituted from amino acid sequences that enable our cells to function correctly. They are structurally and functionally necessary for our body and regulate body cells, tissues, and organs, which cannot happen without them.

2.1.2 Genes

Genes are distinctive units within the chromosomes found in the cell's nucleus and transfer the essential characteristics of living things between generations. Each gene performs a different function in our body. While providing protein synthesis, they are decisive in determining and revealing all our characteristics, from the color of our hair to our finger shape, our blood type, by transferring the information in DNA to RNA.

2.1.3 Diseases

Disease results from protein-based mutations that cause structural and functional impairment in a living thing, mainly producing specific symptoms or affecting a par-

ticular location. These mutations are the leading causes of the disease. They cause biochemically dysfunctional allosteric changes in proteins that affect the binding interface [1].

2.1.4 Compounds

A chemical compound is a chemical substance made of many alike molecules constituted of atoms attached by chemical bonds from more than one element.

2.1.5 Drugs

Drugs are the substances given to humans or animals to treat, prevent, or diagnose a disease. They are used to relieve pain or other disturbing conditions and correct and control abnormal minds and bodies' abnormal circumstances.

The drug development process consists of a few main parts:

- Discovery and research
- Preclinical studies
- Clinical studies
- Treatment approval

We can name drugs as compounds. Drugs are compounds that passed the treatment approval, which is the last stage of the stages mentioned above.

2.1.6 HPOs

Human Phenotype Ontology (HPO) consists of HPO terms that form a standardized vocabulary of phenotypic abnormalities faced for human diseases.

2.1.7 Pathways

A pathway is an ordered sequence of molecular events that create a new molecular product, or a change in a cellular state or process. For example, a pathway could produce a protein, signal the presence of a hormonal signal, or cause a cell to move.

2.2 Employed Technologies

2.2.1 Apache2

Apache is an open-source, free web server software whose official name is Apache HTTP Server developed by the Apache Software Foundation. It moves files back and forth by creating a bridge between the server and website users (Firefox, Chrome, Safari). Apache is a cross-platform software that can run on both Unix and Windows servers. In case of any try to load a page on someone's website, the browser sends a request to that person's server, and Apache sends a response to all requested files (text, images, etc.). The server and client communicate via the HTTP protocol, and Apache is responsible for smooth and secure communication.

2.2.2 Programming environment

The technologies used to build the accessible web service are all open source and are given below:

- PHP
- Javascript (cytoscape.js)
- CSS (BootStrap)
- MySQL (database management system used for local services, data is taken from the CROssBAR database)

2.2.3 Cytoscape.js

Cytoscape.js is a graph theory library developed as open-source in pure JS utilized for graphical analysis and visualization. Cytoscape.js enables users to interact with the graph and connect to the client's user events and facilitates to view and edit rich and interactive graphics for users. Since Cytoscape.js is compatible with both desktop browsers such as Chrome and mobile browsers such as the iPad, it is easily integrable into applications. Cytoscape.js includes all the actions users expect, such as compression for zooming, frame selection, scrolling. Cytoscape.js also includes many useful graph theory functions in its library, as it takes graph analysis into account.

2.3 Related Work

Related studies have different and similar aspects to our approach such as consisting computationally predicted or experimentally validated data as well as attributes and associations among biological entities. Many studies specifically focus on only genes/proteins [2] or drugs/compounds [3] or pathways [4] or diseases, or HPO terms. There are also studies covering more than one of these concepts [5]. Some studies differ in terms of integrated databases such as manually curated/computationally generated and followed methodologies to represent biomedical data. Apart from these studies, services such as [6], [7], [8], which are widely used, also use graph/network-based expressions as service logic and in some respects intersect with the CROssBAR project.

Fabregat et al. presented Reactome Pathway Knowledgebase that provides molecular details of signal transduction, transport, DNA replication, protein synthesis, and intermediate metabolism as an ordered network of molecular transformations [6]. The increasing size and complexity of Reactome have necessitated some performance-enhancing updates. For this purpose, a graph database has been implemented, performance improvements of data analysis tools have been made, and new data structures and strategies have been established. The reason why graph database model is preferred over relational database is that Reactome data contains many relations and therefore needs many join tables. By using Neo4j graphical database, response times

are shortened and significant performance improvement is provided. Also, Enhanced High Level Diagrams that merge an iconography familiar from textbooks and review articles with web functionality were implemented to improve usability.

STRING database [7] aims to collect, score and integrate all publicly available protein-protein interaction information sources and perform computational prediction on them to achieve a comprehensive and objective global network, including physical and functional interactions. STRING offers its users the option of enrichment analysis after version 11.0. This option requires input, especially at the genome scale, where a protein or gene has an associated numerical value for analysis. When users query with a set of proteins, calculation of the functional enrichment analysis is performed in the background, and then the results can be interactively projected into the user's protein network. In addition to high-throughput data mining methods and hierarchical clustering of the association network itself, well-known classification systems such as Gene Ontology and KEGG are also implemented in enrichment analysis calculations.

WikiPathways project [4] was introduced by Kutmon et al. as open access and collaborative web platform. This project is used to capture and disseminate models of pathways for biological data analysis and visualization. In the updated version, extensions for pathway analysis tool PathVisio and the network analysis tool Cytoscape were developed to provide ease of integration of WikiPathways pathways. Apart from that, new features have been added to the study, making it suitable for ease of navigation and consulting. Besides XML, the web service also returns JSON, a format that can be easily parsed into data structures to support developers. WikiPathways' content has been published on the semantic web, so it can be possible to submit semantic queries on the relevant content via specific query languages or APIs.

GeneMANIA [8] is a web platform used for generating hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays, and enables to query genomic, proteomic, and gene function data with its flexible interface. GeneMANIA can work with a single query gene or a longer list of gene queries. Given a single query gene, the closest linked genes can be found among the user-selected networks and attributes, while the query list can be reconstructed by weighting according to the predicted values in longer lists. In order to achieve this, it is aimed

to find functionally similar genes using a large number of genomics and proteomics data. Another capability of the platform is to make network searches using the analysis features of Cytoscape. The current version of the application has a design focusing on interactive network visualization powered by Cytoscape.js. Moreover, GeneMANIA supports different layouts such as concentric bipartite, force-directed and linear bi partite to enable the user to view the network from a different perspective.

BioGraph [9] is a data integration and data mining platform that enables automatic inference of functional hypotheses among biomedical entities by prioritizing putative disease genes. Since any type of concept can be the subject or target of prioritization, the user does not need to define the list of genes that cause the disease, thus the need for domain information is minimized. BioGraph uses 21 publicly curated databases containing biomedical relationships between heterogeneous biomedical entities for data integration. The entities mentioned herein are genes, diseases, compounds, pathways, ontology terms, protein domains, disease and gene families, and microRNAs. The integrated databases included in BioGraph are of three different types: curated databases, curated ontology databases and curated annotation databases. It is possible to extract relationships between concepts from the knowledge resources in common format and annotated semantic relationship types provided by these integrated databases.

Bio4j [10] has been designed and implemented as a high-performance, cloud-enabled platform that integrates semantically rich large-scale biological data using graph models. The main goal of the creation of Bio4j was to integrate with important bioinformatics databases for a modular and truly integrated graph structure, so it was thought that it would be possible to perform genomics, metagenomics and transcriptomics analysis on a wide spectrum from human to virus. Therefore, Bio4j includes protein data, taxonomy data and annotation data in integration with UniProt, NCBI and Gene Ontology, respectively.

Yuan et al. propose an approach [11] to construct biomedical domain-specific knowledge graphs in order to represent complex domain knowledge in a structured format. It was stated that the knowledge graphs were created with minimum supervision since the method developed did not require human annotation until the last stage. In

the last step, semantic relationships need to be defined manually. Biomedical articles were collected from PubMed in an unstructured format and then automatically preprocessed. Other stages of the method, which are fully automated, include steps for text mining such as entity recognition, entity and relation embedding, latent label generation, relation refinement.

CHAPTER 3

CROSSBAR PROJECT

CROssBAR is a comprehensive computational resource that links numerous biomedical and chemical resources and generates predictions for relations via deep learning and machine learning methods. The purpose of the CROssBAR project is to design and develop artificial and deep learning-based prediction methods and models for virtual scanning, to validate predictions obtained from developed models by means of molecular biology experiments and to create visualization tools for further analysis. The CROssBAR project is funded by TÜBİTAK and the British Council and carried out in cooperation with METU and EMBL-EBI,

Multiple sub-projects are implemented in the CROssBAR project in order to provide biomedical researchers a better and quick understanding of relations between biomedical entities.

Sub-projects of CROssBAR are as follows.

- CROssBAR database and its API,
- Deep learning based DTI (drug-target interaction) prediction models,
- Wet-lab Experiments,
- Network-based organization and analysis of large scale biomedical data using knowledge graph representations,
- CROssBAR Web-service.

Figure 3.1 illustrates all parts of the CROssBAR Project.

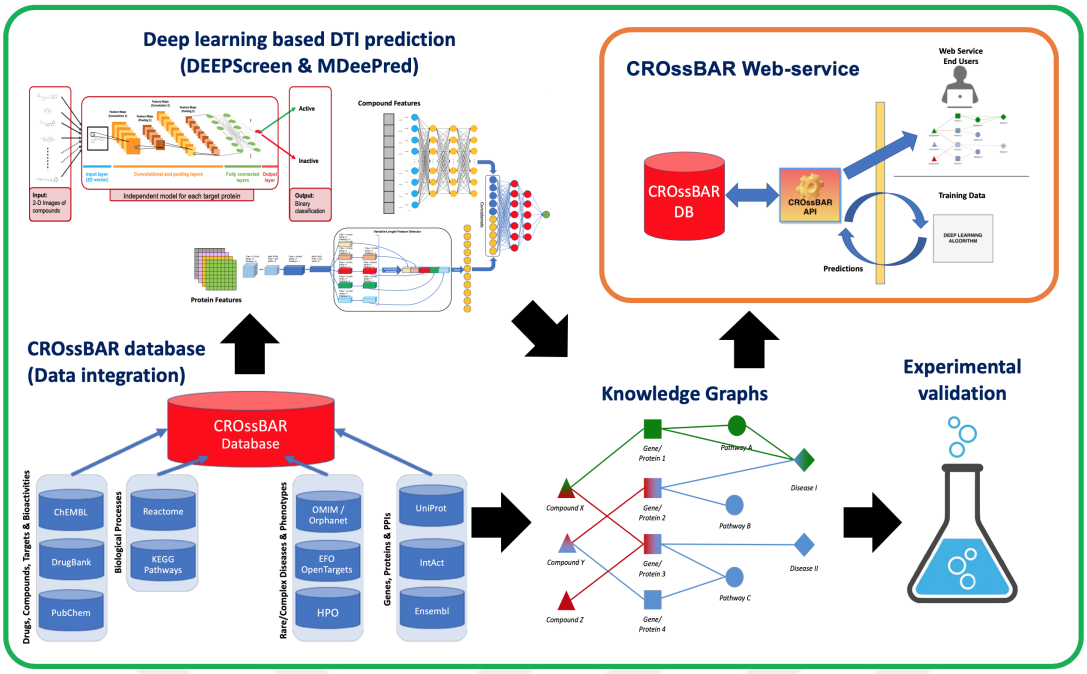

Figure 3.1: General view of CROssBAR project.

3.1 CROssBAR Database and its API

The huge amount of biological and chemical information from the public Resources are unified under the CROssBAR database. Data pipelines have been developed to fetch data from different databases such as UniProt, IntAct, DrugBank, ChEMBL, PubChem, Reactome, KEGG, OMIM, and EFO while maintaining specific data attributes by applying logic rules. Figure 3.2 displays the integrated resources.

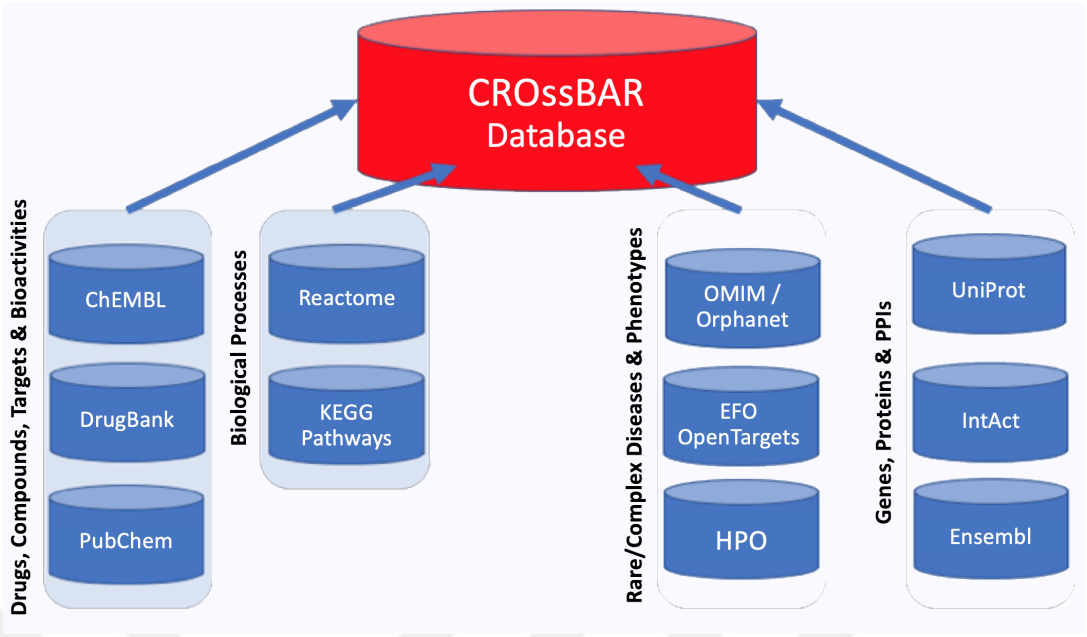

Figure 3.2: Publicly available resources that are integrated into CROssBAR database.

Figure 3.3 displays cross-collection relations of CROssBAR DB. The CROssBAR database of attributes is hosted in self-sufficient, easy to access collections in MongoDB. MongoDB stores data records as documents which are assembled together in collections and the collection term is a grouping of MongoDB documents [12]. Those collections available both to end-users and to the CROssBAR web service through an API at: www.ebi.ac.uk/Tools/crossbar/swagger-ui.html.

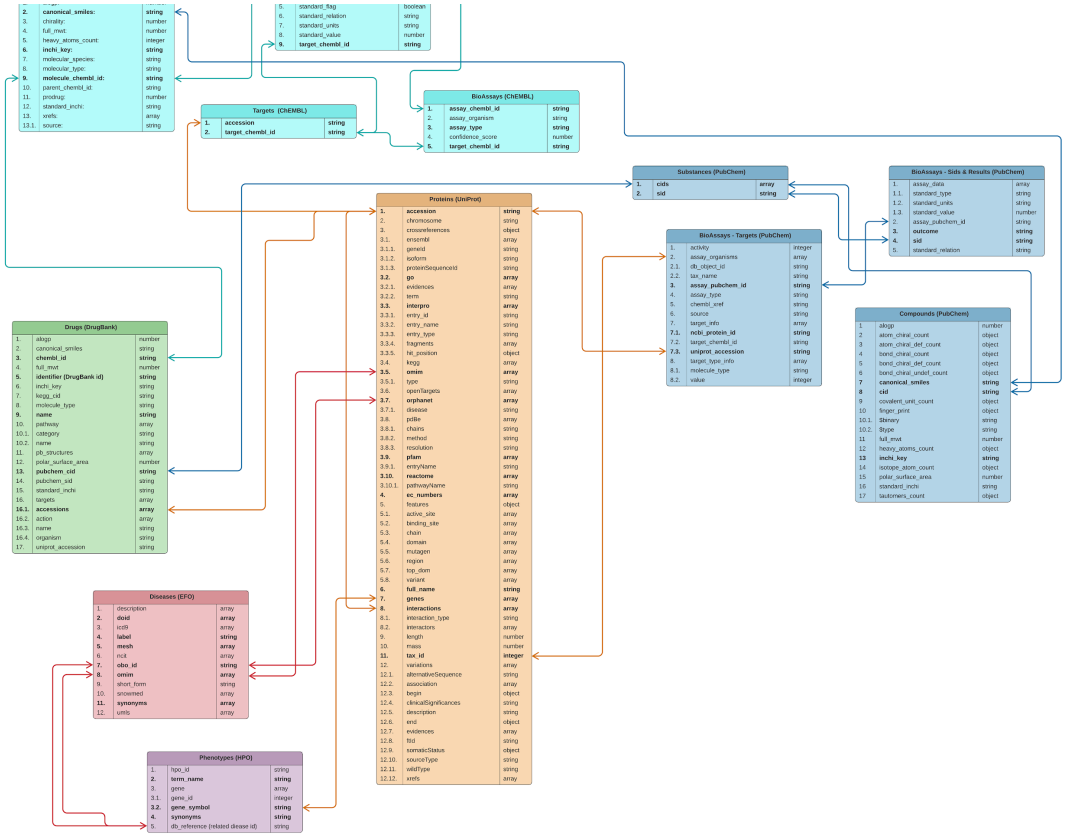

Figure 3.3: Schematic representation of collections of CROssBAR DB.

The number of entries and the size of data in each collection was given in Table 3.1.

Table 3.1: The number of entries and the size of data in collections in the CROssBAR database.

Collection name	Total number of entries	Total file size (MB)
Activities (ChEMBL)	15,207,914	3,000
Assays (ChEMBL)	1,060,283	167.8
Drugs (DrugBank)	11,922	11.4
Targets (ChEMBL)	12,091	15,800
Molecules (ChEMBL)	1,828,820	998.2
Proteins (UniProt)	443,422	6,500
Bioassay Sids (PubChem)	239,151,694	188.5
Bioassay (PubChem)	1,067,568	62,900
Compound (PubChem)	97,351,543	69,000
Substance (PubChem)	252,877,567	670.1
HPO	8,822	22.5
EFO	30,438	15.0

3.1.1 ChEMBL

ChEMBL presents a database of manually curated bioactive molecules with drug-like properties. This database holds chemical, bioactivity, and genomic data to support scientists in interpreting genomic information into effective new drugs [13].

3.1.2 PubChem

PubChem is an extensive collection of freely accessible chemical information. Searching chemicals by name, molecular formula, structure, and other identifiers are presented to researchers' usage. Chemical and physical properties, biological activities, safety, and toxicity information, patents, literature citations, and more are also provided [14].

3.1.3 DrugBank

DrugBank is a comprehensive database of drug-target information that presents a detailed drug resource for bioinformatics and cheminformatics [15].

3.1.4 Reactome

Reactome is an open-source, open-access, manually curated, and peer-reviewed pathway database. The motivation behind the creation of this database is to contribute bioinformatics tools for the visualization, interpretation, and analysis of pathway knowledge to support basic and clinical research, genome analysis, modeling, systems biology, and education.

3.1.5 Kyoto Encyclopedia of Genes and Genomes

KEGG is a database of large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. This database enables us to comprehend high-level functions and utilities of the biological system [16].

3.1.6 Online Mendelian Inheritance in Man

Dr. Victor A. McKusick initiated this database in the early 1960s as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). This database is a comprehensive, authentic collection of human genes and genetic phenotypes that are freely available. OMIM provides information on all known mendelian disorders and over 15,000 genes [17].

3.1.7 Experimental Factor Ontology

The Experimental Factor Ontology (EFO) exhibited by EMBL-EBI provides a well-organized definition of many experimental variables in EBI databases [18].

3.1.8 Human Phenotype Ontology

Human Phenotype Ontology (HPO) consists of HPO terms that form a standardized vocabulary of phenotypic abnormalities faced for human diseases. HPO is developed by using the Orphanet, DECIPHER, and OMIM medical literature. It currently comprises over 13,000 terms and covers 156,000 riders for hereditary diseases. HPO is a part of a project that develops software for phenotype-based differential diagnosis, genome diagnosis, and translation research. Besides, it is the flagship product of the Monarch Initiative, an international NIH-supported consortium dedicated to the semantic integration of biomedical data.

3.1.9 The Universal Protein Resource (UniProt)

UniProt is a database of protein sequences and their annotations. Many inputs have been derived from genome sequencing projects and presented to users as an open-access online platform. The biological functions of proteins have been extracted from the research literature and offered to users. This platform's maintenance is provided by the UniProt Consortium, consisting of several European bioinformatics organizations and a foundation from Washington, DC, United States.
UniProtKB consists of 2 parts: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. Swiss-Prot is a protein sequence database that has annotations with manually reviewed experimental results and scientific results. Swiss-Prot is the updated division of the UniProt knowledgebase. UniProtKB/TrEMBL comprises computationally examined unreviewed datasets that are enhanced with electronic annotations. Manually reviewed and unreviewed entries are stored in the different databases to have high-quality data [19].

3.2 Drug-Target Interaction Models

As a part of the CROssBAR project, two novel deep-learning-based predictive systems developed: DEEPScreen and MDeePred, intending to enrich bioactivity data by identifying unknown interactions between drugs/drug-candidate compounds and

target proteins.

3.2.1 DEEPScreen

DEEPScreen [20] is a high-performance drug–target interaction predictor that utilizes convolutional neural networks and 2-D structural compound representations to predict their activity against intended target proteins. DEEPScreen system comprises 704 target protein-specific prediction models, each independently trained using experimental bioactivity measurements against many drug candidate small molecules, and optimized according to the target proteins' binding properties.

3.2.2 MDeePred

MDeePred is a deep learning method that generates compound-target binding affinity predictions to use for computational drug discovery and repositioning purposes. The method adopts the chemogenomic approach where both compound and target protein properties are used at the input level; this enables to make predictions for proteins that are not sufficiently studied or fully targeted. In MDeePred, multiple types of protein traits, such as sequence, structural, evolutionary, and physicochemical features, are incorporated into multi-channel 2-dimensional vectors, and this is then converted into true-valued compound-target protein interactions.

3.3 Wet-lab Experiments

Wet-lab molecular biology experiments, centered around the topics of mechanisms and the possible treatments of the hepatocellular carcinoma (HCC) disease, have been created and carried out for selected sets of computational predictions to validate the accuracy of the information included in the CROssBAR resource. As a result of the DEEPScreen large-scale prediction run, it has been disclosed that JAK proteins may be Cladribine's new target proteins. Cytotoxicity, cell death, cell cycle analyses, and flow cytometry analysis of STAT-3 phosphorylation and protein immunoblotting

methods were used to demonstrate whether Cladribine drug had an effect on this pathway in hepatocellular carcinoma(HCC) cell lines (Huh7, Mahlavu, and HepG2) [20].

3.4 CROssBAR Knowledge Graphs

The term knowledge graph (KG) defines a specific data representation structure, in which a collection of entities (nodes) are linked to each other (edges) in a semantic context [11]. In the scope of the CROssBAR project, it is chosen to represent heterogeneous biomedical data in KG structures. In CROssBAR knowledge graphs (CROssBAR-KG), biological components (i.e., drugs/compounds, genes/proteins, pathways, diseases) are represented as nodes. Their known or predicted pairwise relations are annotated as edges (a protein and its coding gene is merged as one term). CROssBAR-KGs display the direct and indirect associations among all of the terms in the graph. These intensely-processed heterogeneous biological networks are expected to aid biomedical study, especially to deduce mechanisms of diseases concerning biomolecules, systems, and candidate drugs.

3.5 CROssBAR Web-service

We developed the CROssBAR web-service (CROssBAR-WS) to make the CROssBAR-KGs accessible to the public in a simply interpretable and interactive way. KGs are given visually on web-browsers as flexible Cytoscape [5] networks. Users can perform queries with biomedical terms, separately or in combination, to receive the relevant graph. The combinatory term query is mainly critical as it grants the capability to investigate the indirect biological associations between the terms from both the same and different biomedical components. Considering there are billions of different ways to query CROssBAR, it was not possible to pre-calculate the resulting graphs; consequently, they are set to be created on-the-fly, in real-time. The subject of this thesis implementing the CROssBAR web-service as described in Chapter 4.

CHAPTER 4

METHOD

The knowledge graph construction process starts by getting the input from the user. Figure 4.1 illustrates the flow of a knowledge graph construction process. Users may start the search with one or more component types as his/her choice.

Those component types that are currently available for search are as follows:

- diseases/phenotypes,
- drugs/compounds,
- pathways
- proteins

Once the user chooses component(s) for search, first fetching the gene(s)/protein(s) needs to be brought in which appropriate queries are constructed and sent to CROss-BAR API. We call these proteins as “core proteins” in our system since they are the main components of our knowledge graphs.

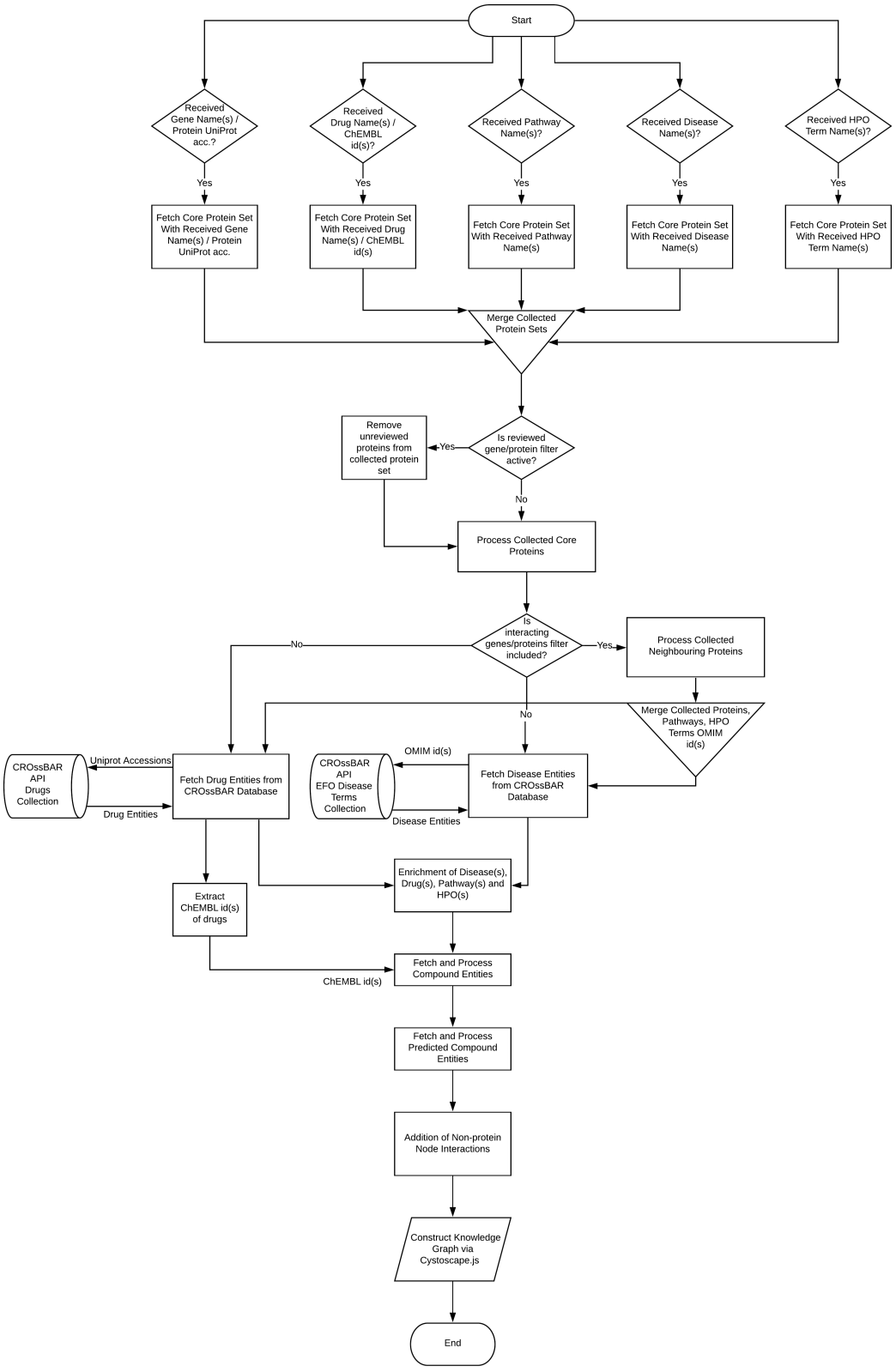

Figure 4.1: Flowchart of Whole Graph Construction Process

While collecting main proteins, we also consider taxonomic classification by adding user-defined taxonomy identifiers to the queries. Since there are nine types of proteins that belong to different species in CROssBAR API, we provide a select box that allows selectable species with their taxonomic identifiers. These taxonomic identifiers written in parentheses.The NCBI contributes unique taxonomic identifiers for all organisms expressed in the international sequence databases. Although an organism is expressed in UniProtKB, the relevant 'taxonomic identifier' may not be visible from the NCBI taxonomic database [21]. Selectable taxonomies are as follows:

- Human (9606),
- Rat (10116),
- Pig (9823),
- Bovine (9913),
- Mouse (10090),
- Baker's yeast (559292),
- Rabbit (9986),
- MYCTU (83332),
- Ecoli (83333)

The user can select more than one species, however, if (s)he does not make any selection, human is chosen by default.

After collecting the core proteins of every type of component, we merge the proteins to avoid duplicates. We pass these core proteins to the next phase. If the user wants the collected proteins to be reviewed proteins by clicking the checkbox "only reviewed gene/protein entries", at this phase, it is checked whether the proteins are reviewed.

When we remove unreviewed protein entries from the collected proteins, we need to process the core proteins. Some types of components give us to complete protein data, but others provide just UniProt accession ids. Before proceeding, we should

fetch all data related to the core proteins from the CROssBAR database and extract the necessary information, as described in Chapter 4.3.

If the user also wants to visualize the interacting proteins, the system extracts the protein-protein interaction data of the core proteins and combine these proteins with the core proteins, as described in Chapter 4.4.

Processing of protein resources contains extracting HPO terms, pathways, and OMIM identifiers. However, their relations with diseases, drugs, and compounds are not accessible directly from CROssBAR Protein Resource. We fetch related disease entities via sending collected OMIM identifiers to CROssBAR EFO Disease Terms Collection. Also, associated drugs are fetched from CROssBAR Drugs Collection in parallel with diseases.

At this point, we have all related diseases, drugs, pathways, and HPO terms, but since there could be hundreds of them and it is hard to understand and visualize such huge data, we apply enrichment analysis as described in Chapter 4.2 to every component to display the most significant ones.

At this phase, first, the experimental components and then the predictive compounds are fetched and processed. Experimental compounds are the ones that have proven useful in the laboratory but have not been put into practice. Predicted compounds, on the other hand, are components whose usefulness is indicated by computational methods. In terms of reliability, experimental compounds and predicted compounds come after drugs. Those processes are detailed in Chapter 4.7.

So far, we have taken a protein-based approach when creating the knowledge graphs. However, some nodes may also have interactions with other non-protein nodes, such as disease-drug. We finalized the knowledge graphs construction by adding such relationships to our network data, described in Chapter 4.8.

At the end of all these steps, we feed with prepared network data to Cytoscape.js to render the corresponding knowledge graph. Every node type has a different shape, as shown in Figure 4.2. Also, a detailed search report which consists of all steps described above provided to users as a downloadable text file. Our web-service is publicly available at https://crossbar.kansil.org/.

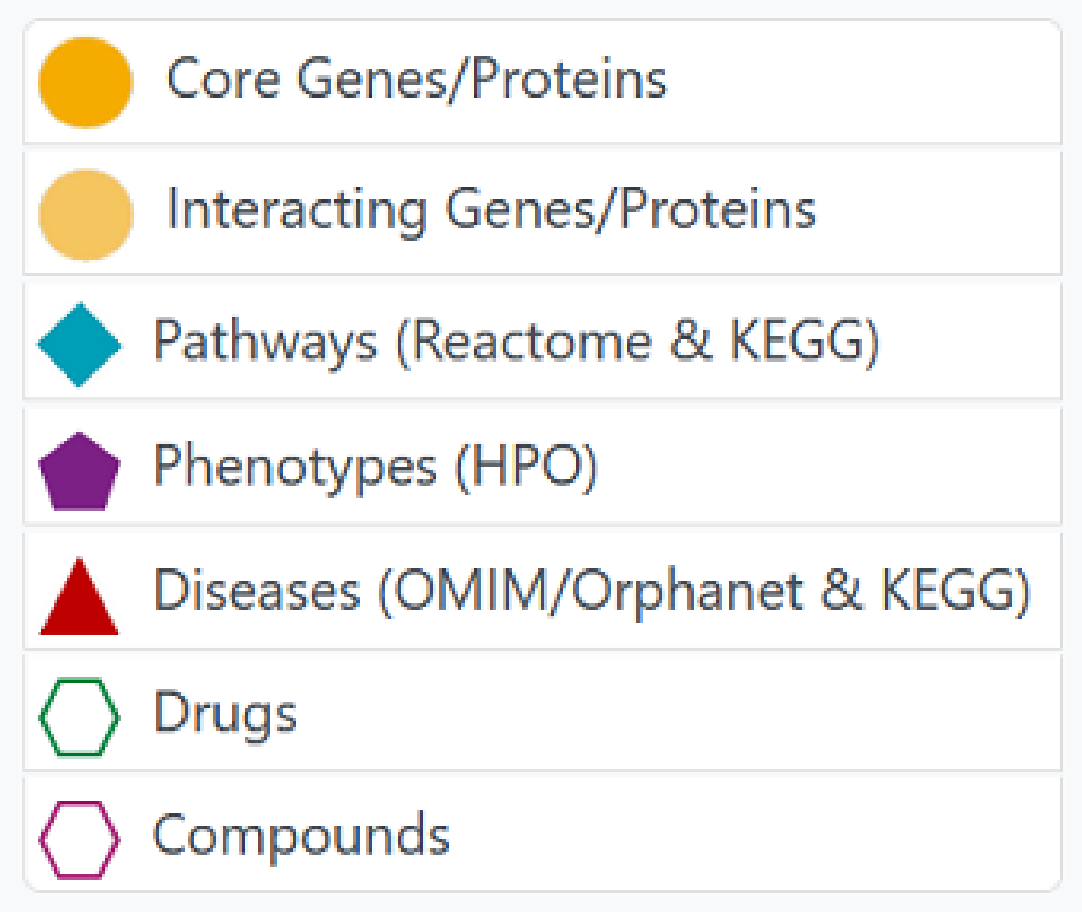

Figure 4.2: Shapes of node types.

4.1 Collecting UniProt Accessions

Knowledge graphs in our system are created based on the proteins. Even if the user searches for a drug or a pathway, we first start to create our knowledge graph by pulling the core proteins associated with them from the database. As explained in other sections, we must follow different ways when collecting proteins from the CROssBAR API. For example, in the pathway search, we fetch core proteins from the protein resource directly. On the other hand, in the disease search, we first fetch the omim ids from the EFO Disease terms collection, and we then use the collected OMIM ids to fetch the protein from the protein resource.

4.1.1 Fetching Proteins via Drugs and Compounds

If we receive data from "Drug name(s) / ChEMBL id(s)" box, we first check if the given input is drug name or not. If it starts with "CHEMBL" prefix, that means it is a compound, and we need to obtain activities data from CROssBAR API. We fetch such data in JSON format via sending an HTTP request to https://www.ebi.ac.uk/Tools /crossbar/activities? **pchemblValue=6** &moleculeChemblId=**CHEMBL_ID** page. Notice that there is a "pchemblValue" parameter. Higher the value of this parameter, more biologically meaningful data we obtain. If we do not get a result with the value of 6, we continue searching with less reliable results by reducing the pchemblValue parameter till 4. After parsing the data that we obtained from the activities collection, we extract the first N "target_chembl_id" by sorting the ChEMBL activities according to pchemblValue to get the ones with the highest experimental values.

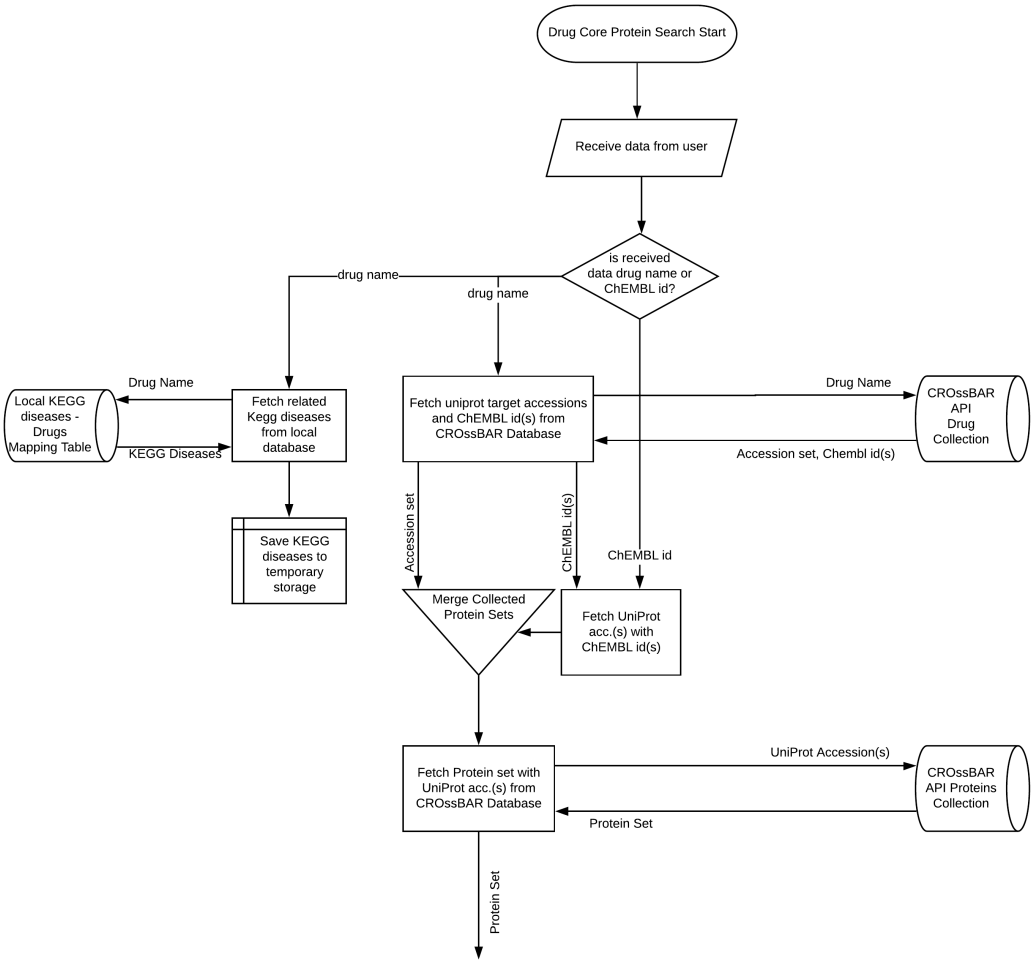

Figure 4.3: Flowchart of collecting drug interacting proteins.

4.1.2 Fetching Proteins via Pathways

In addition to the CROssBAR API, we also make use of the KEGG Pathway - UniProt Accessions mapping table that we keep in local in the pathway core protein collection process. If the searched pathway exists in the KEGG database, we first retrieve the related UniProt accessions from this database. Since Reactome pathways can be accessed from the CROssBAR API Protein collection, we can directly fetch the protein data with the Reactome pathway name. As shown in Figure 4.4, UniProt accessions that we have collected locally with the protein data we obtained with the https://www.ebi.ac.uk/Tools/crossbar/proteins?reactome=**pathway_name** request, we

save them to temporary storage and pursue with the next step.

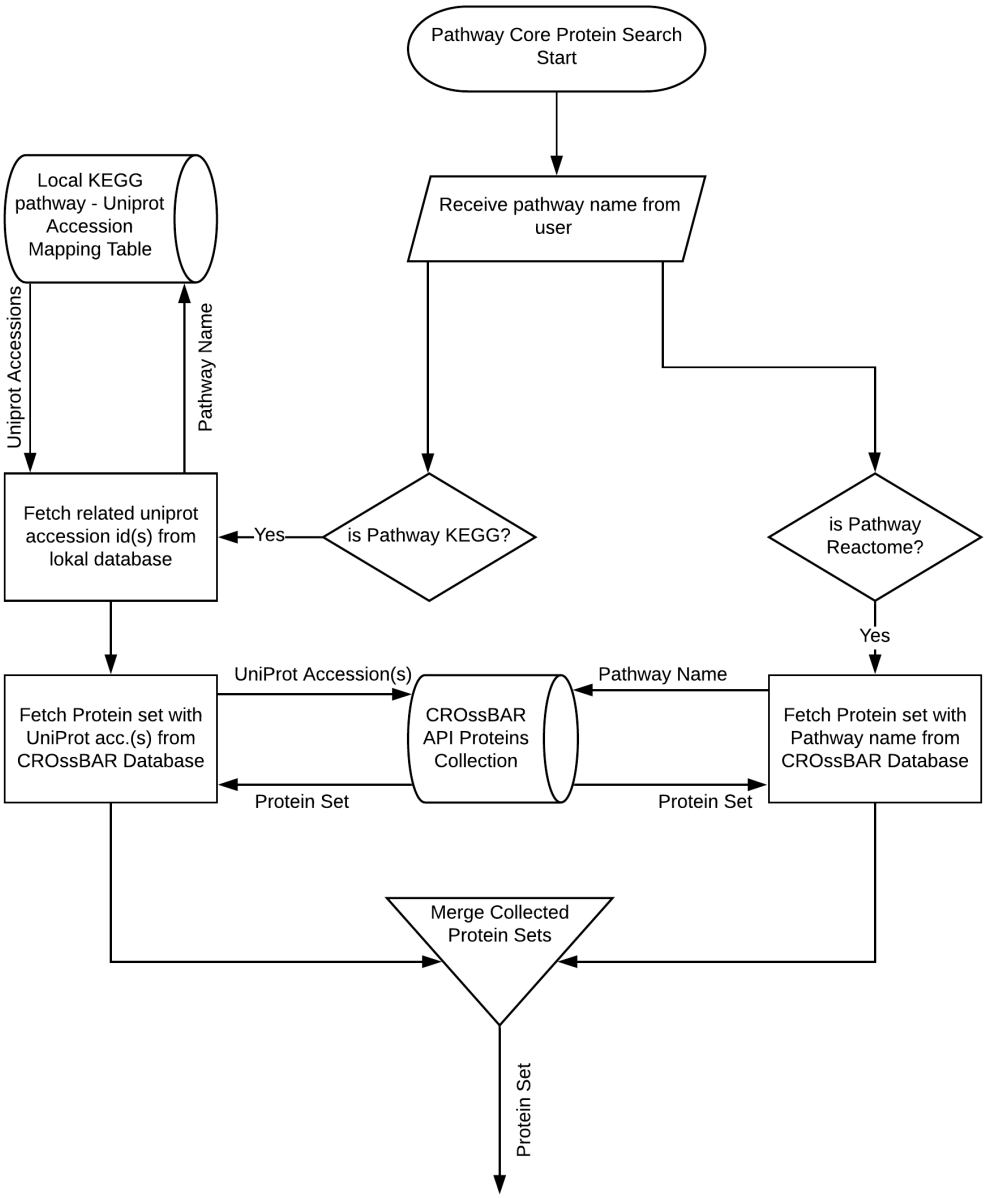

Figure 4.4: Flowchart of collecting pathway interacting proteins.

4.1.3 Fetching Proteins via Diseases

Since the diseases are collected from three different sources in our system, we follow three different paths in the process of collecting the essential proteins of the disease.

If the disease is detected in all three sources, the protein is gathered and merged.

In case the disease is a KEGG disease, we save the proteins we take from the local KEGG Diseases - UniProt accessions mapping table, as well as the corresponding drugs from the local KEGG Diseases - Drugs mapping table to the temporary storage.

If the given disease is an EFO disease, we first extract OMIM IDs from the data we obtain by sending the www.ebi.ac.uk/Tools/crossbar/efo?synonym=**disease_name** request to the CROssBAR API EFO Resource. As a result of the request of www.ebi.ac.uk /Tools/crossbar/proteins?omim=**omim_ids**, we obtain the core proteins of the EFO disease by sending the collected OMIM IDs to the protein collection.

As Orphanet diseases can be accessed directly from the CROssBAR API protein resource, it is enough to send the www.ebi.ac.uk/Tools/crossbar/proteins? orphanet=**disease_name** request.

Figure 4.5 illustrates the process of collecting information from these three different sources.

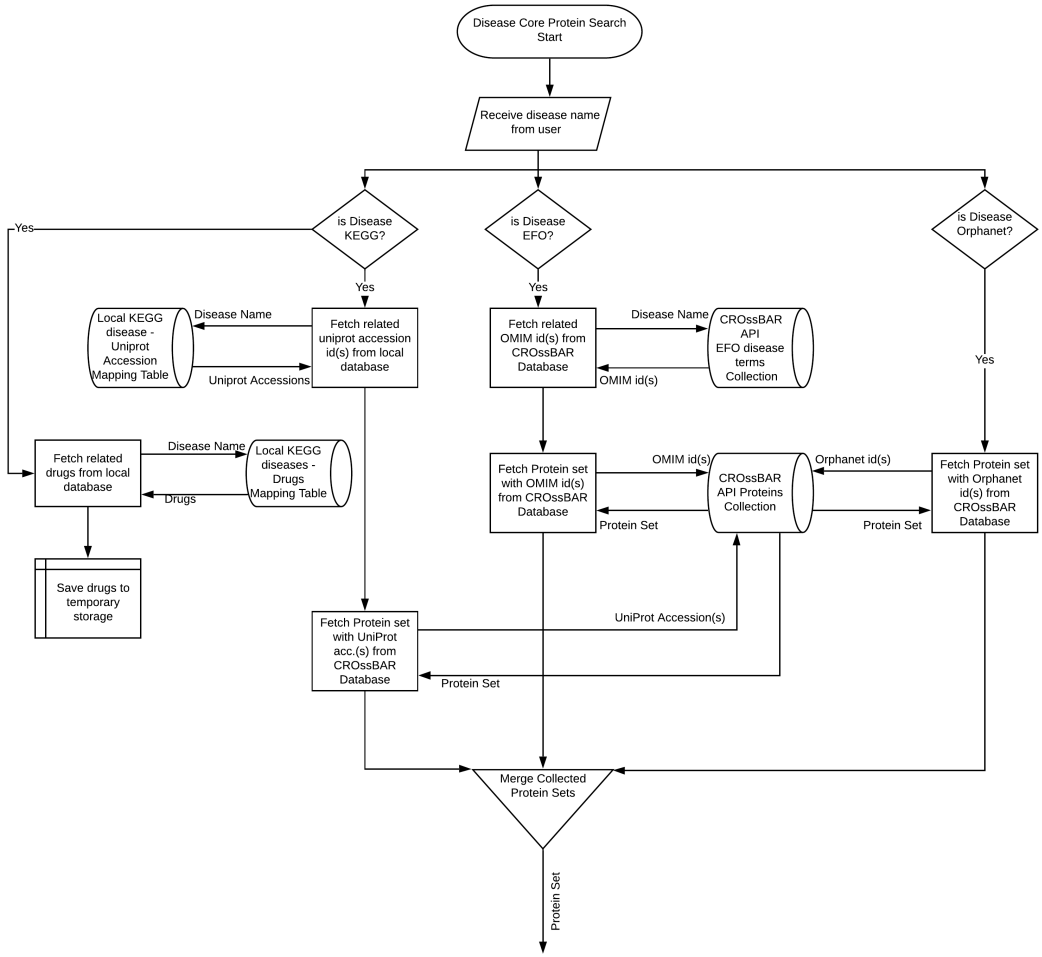

Figure 4.5: Flowchart of collecting disease interacting proteins.

4.1.4 Fetching Proteins via HPO Terms

Since HPO Terms are mapped with gene names in the CROssBAR API, in the HPO Terms Core Proteins collection process, first of all by sending the www.ebi.ac.uk/Tools /crossbar/hpo?hpotermname=**HPO_Term** request we extract the related gene names. We save the data we obtain in temporary storage by sending the www.ebi.ac.uk/Tools /crossbar/proteins?gene=**collected_gene_names** request to the protein collection with the collected gene names.

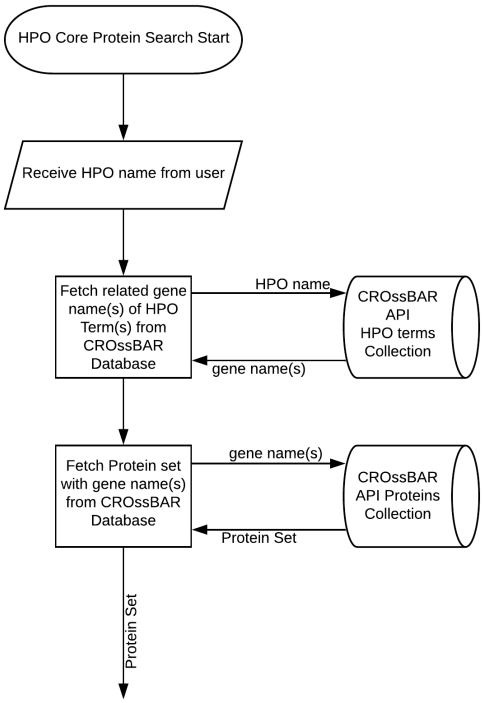

Figure 4.6: Flowchart of Collecting of HPO Terms Interacting Proteins

4.2 Enrichment Analysis

Due to a large number of records in CROssBAR DB, the knowledge graphs become very crowded. Therefore, we use the enrichment analysis on the knowledge graph. The enrichment analysis enables us to create biologically meaningful knowledge graphs in a reasonable size and only obtain the most relevant nodes and edges.

The variables and their definitions used in the calculation of the enrichment score of a component Q are as follows:

- x = number of proteins in the network that Q is associated with
- y = total number of proteins in the network
- X = total number of proteins that Q is associated with (not only the proteins in the network!)

- Y = total number of proteins in the CROssBAR DB

$$enrichment\ \ score = \frac{x^2/y}{X/Y}$$

While x and y are calculated on the fly during the knowledge graph construction process, we fetch X and Y from the enrichment tables that we keep in our local database.

Enrichment scores are applied to each node type. We then extract the first N items with the highest scores to be represented in the knowledge graph. We accept the N value as the value that the user enters in the "# of nodes" input box.

4.3 Processing Core Proteins

Processing of proteins starts with filtering the collected protein set. First of all, we remove the proteins that do not match with the taxonomy identifiers input by the user. Suppose the user has included "interacting genes/proteins." In this case, we put the UniProt accessions that we obtained by following the proteins-> interactions path to enrichment analysis. After taking the first N, we save them to temporary storage for processing, as described in Chapter 4.4.

From the accumulated protein data, we extract OMIM identifiers via following the path of proteins->crossreferences->omim, and Reactome pathways via following the path of proteins->crossreferences->Reactome, and we save them to temporary storage.

Since HPO Terms cannot be obtained directly from the collected protein data, we save the obtained HPO Terms to temporary storage by sending www.ebi.ac.uk/Tools/crossbar /hpo?genesymbol=**collected_gene_names** query with collected gene names.

Figure 4.7 illustrates the processing of core proteins.

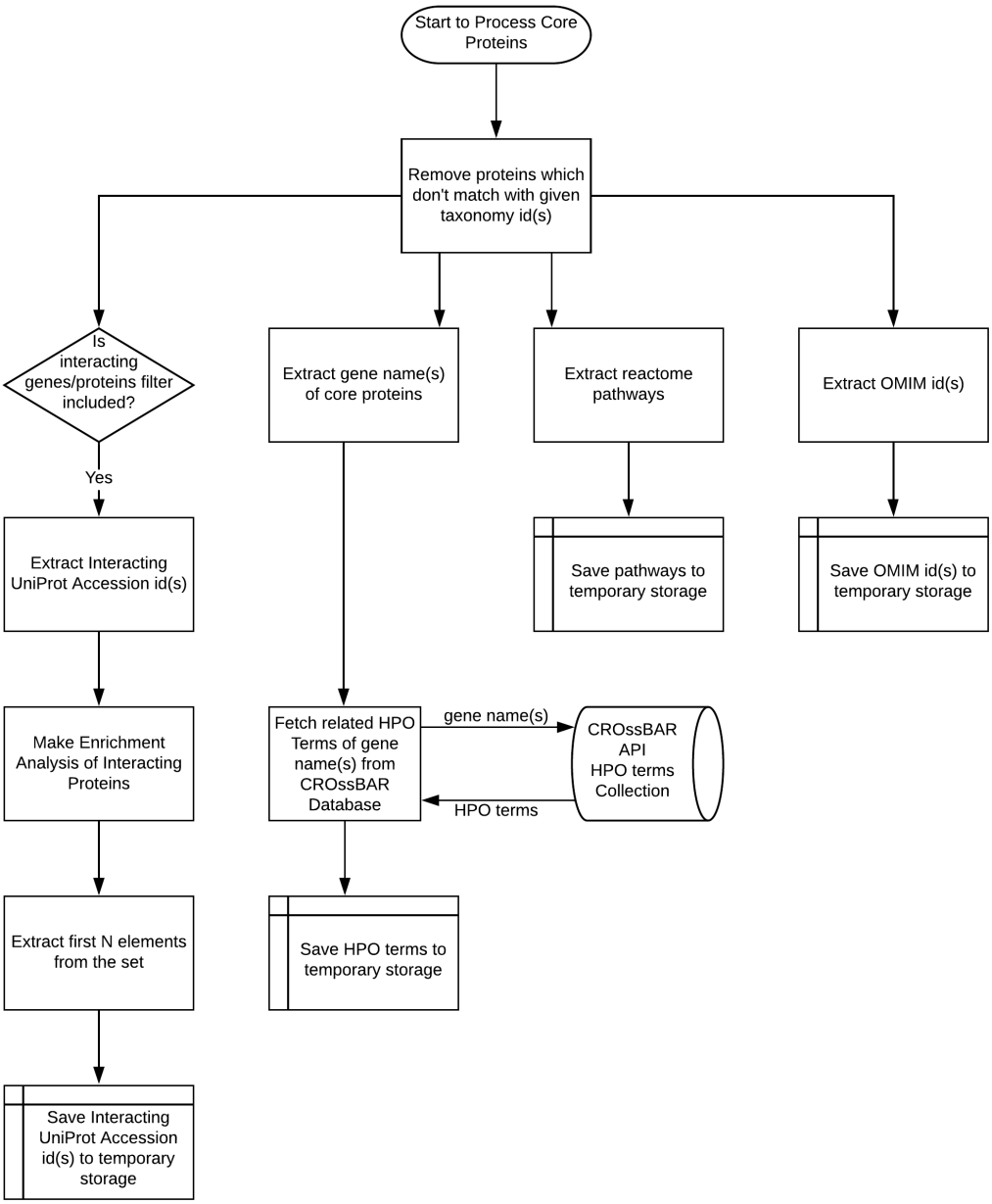

Figure 4.7: Flowchart of processing core proteins.

4.4 Processing Neighbouring Proteins

In the process of neighbor proteins, we follow the proteins-> crossreferences -> omim path, as in the same core proteins, and extract the Reactome pathways by following the pathways of OMIM ids, proteins -> crossreferences -> Reactome and save them

to temporary storage. Similarly, we send the gene names extracted from proteins-> genes field to CROssBAR API HPO Resource and keep the HPO Terms in temporary storage. Since we represent only the first neighbor proteins of core proteins in our system, the proteins obtained from the proteins-> interactions field of the first neighbors are ignored.

Figure 4.8 illustrates the processing of neighboring proteins.

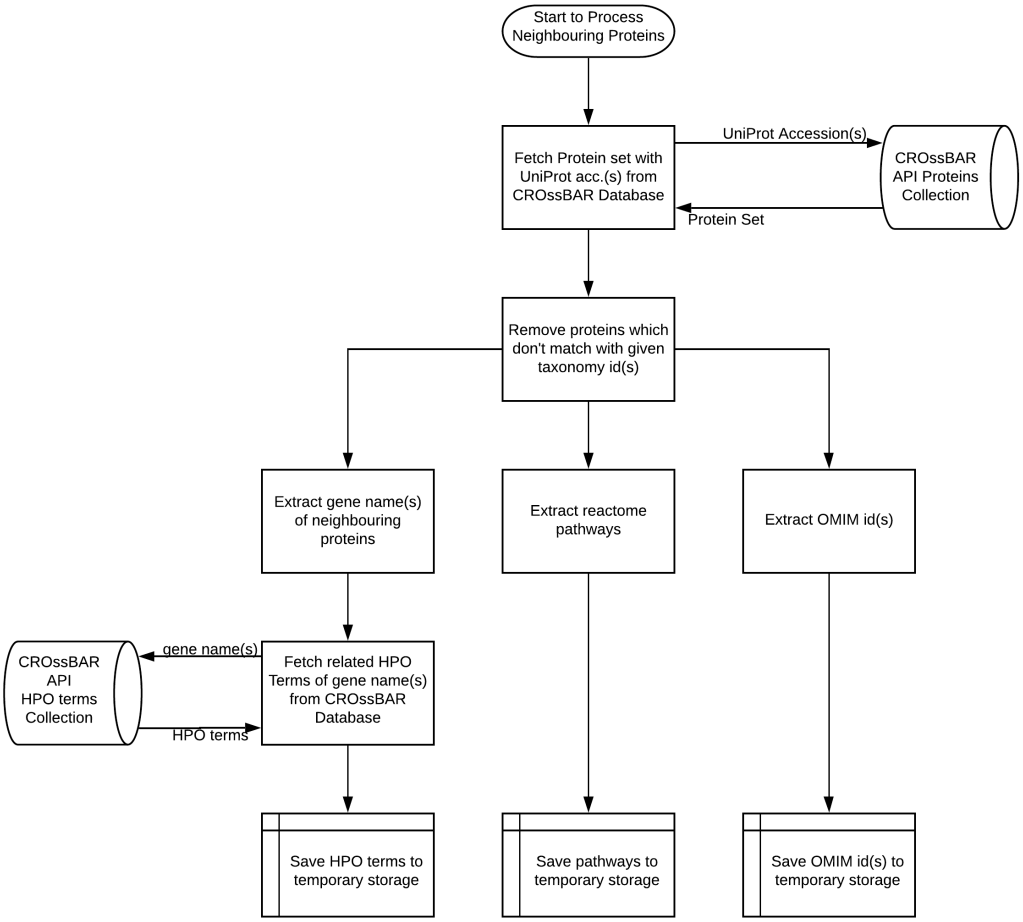

Figure 4.8: Flowchart of processing neighbouring proteins.

4.5 Fetching Drug Entities

We obtain the drug data by sending the www.ebi.ac.uk/Tools/crossbar/drugs? accession=**uniprot_accessions** request to the CROssBAR API Drugs collection with the

collected UniProt accession identifiers. After enrichment analysis, the first N drug nodes with the highest score are kept in the system. Besides, we store the "chembl_id" field of these N drugs in temporary storage. We use these "chembl_ids" in the creation of compound nodes, as explained in Chapter 4.7.

4.6 Fetching Disease Entities

To fetch the disease entities, OMIM identifiers that we collect during the processing of core proteins and neighboring proteins are sent to the CROssBAR API EFO Disease Terms collection via www.ebi.ac.uk/Tools/crossbar/efo?omimId=**collected_omim_ids** request. After we perform enrichment analysis on the diseases, we obtain by parsing the JSON data delivered from this request and we store the first N as the knowledge graph disease nodes in temporary storage.

4.7 Fetching Experimental and Predicted Compounds

To fetch Experimental Compounds, we first send the www.ebi.ac.uk/Tools/crossbar /targets?accession=**uniprot_accessions** request to CROssBAR API Chembl Targets Resource. We fetch ChEMBL activities by adding the "target_chembl_ids" extracted from this data to the www.ebi.ac.uk/Tools/crossbar/activities? targetChemblId=**target_chembl_ids** request that we send to the activities collection. Then we apply enrichment analysis to the experimental compounds by extracting the activities->molecule_chembl_id field from the data we have obtained. After listing according to their enrichment scores, these compounds are processed one by one, starting with the one with the highest enrichment score.

While each compound is being processed, as described in Section 4.5, when collecting drugs, the interactions of the matching compounds are added to the related drug as a blue edge by comparing with the ChEMBL identifiers we have registered in the temporary storage. Compounds that already exist in the system as drug are not added to the system again.

We check whether the compound we process is a drug in the CROssBAR API Drugs

collection by sending the request www.ebi.ac.uk/Tools/crossbar/drugs? chemblId= **molecule_chembl_id**. If we obtain a drug data from this request, we define protein interactions by converting the compound to a drug. At this stage, we define the green edge with the proteins from the Drugs collection, and the blue edge with the proteins that enable the compound to be fetched. These processes continue until they reach N compounds. Figure 4.9 illustrates compound fetching process.

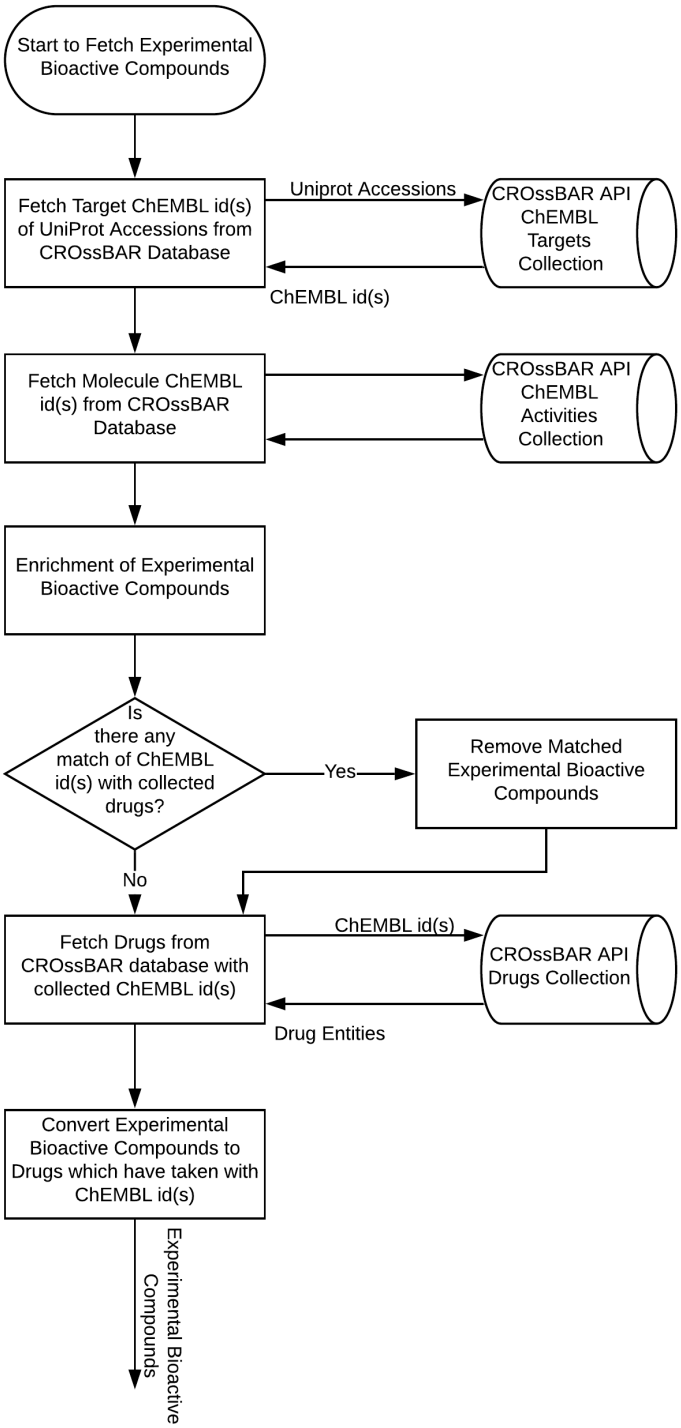

Figure 4.9: Flowchart of fetching compounds.

As described in Chapter 3, we keep the mapping table between the compounds and

proteins predicted by DEEPScreen developed within the CROssBAR project's scope in our local database. The predicted compounds from this table and the collected proteins are processed as it is done for experimental compounds. After the enrichment analysis, these compounds are sorted based on their enrichment scores, and edge merging operations are performed according to the ChEMBL ids of the drugs collected. These compounds and their equivalents in the CROssBAR API Drugs collection are checked.

Since experimental compounds have higher reliability than predicted compounds, the collected predicted compounds are checked whether their experimental versions are available in the CHEMBL database. If it is so, these predicted compounds are converted into experimental compounds, and then protein interactions are identified as blue edges. This process continues until N predicted compounds are obtained.

4.8 Additional Non-protein Interactions

Since KEGG database is not added to CROssBAR API, we keep this data in our local MySQL database. In this section, we add KEGG pathways and diseases to our KG, which we fetched from KEGG Diseases - UniProt accessions and KEGG Pathways - UniProt accessions mapping tables after making enrichment analysis.

Finally, we end the knowledge graph construction process after defining non-protein interactions that we obtain from KEGG Diseases - Drugs and KEGG Diseases - Pathways mapping tables.

4.9 Edge Types

The edges of the graph express relationships between different types of biological entities. Edge types vary according to the defined relationships.

The edge labels for a relation between:

- two proteins: "interacts_with",

- a gene/protein and a disease: “related_to”,
- a drug/compound and a protein: “targets”,
- a gene/protein and a pathway: “involved_in”,
- a gene/protein and a phenotype term: “associated_with”,
- a drug and a disease: “indicates”, and
- a disease and a pathway: “modulates”.

4.10 Layout Algorithm

We developed our proprietary CROssBAR layout plugin, because the layout algorithms built-in in the Cytoscape.js library are not very effective with our knowledge graphs.

We will continue to develop the CROssBAR layout in our future work, and for now, we kept the user experience at the forefront by keeping it simple. Our layout algorithm has 2 options: four layers or seven layers. The layers mentioned here represent nested circles. In the four-layer version, there are similar node types in each layer, while in the seven-layer version, all node types are taken into separate layers and different node types are gathered together.

Seven layer CROssBAR layout demonstration is used in all of the sample searches given in Chapter 5.

CHAPTER 5

RESULTS

We can summarize our contribution to the literature in two items;

- Being able to observe indirectly related biological terms altogether and thus generating hypotheses through their relations,
- Generating hypotheses pursuing the system biology approach by observing in the same graph all of the relevant biological items.

The search process of each component type is summarized in the examples given in this chapter. In Section 5.5, an example of a search for all component types together is given. The most important feature that distinguishes our study from the other studies is illustrated in this example. As the last use-case example, an example of misuse is given in Section 5.6. The meaningless results that may be encountered if the components searched together are unrelated are represented.

The biological interpretation of these examples is left to the users and biologists.

5.1 Protein Search Case: ACE2

A gene/protein search begins with directly fetching the proteins from CROssBAR API Protein Resource. ACE2 gene used in our example is known as the gene most affected by the new generation COVID-19 virus.

In this example, we process gene search by writing to the input box 'ACE2' where "Gene names/Protein Uniprot acc." was indicated. Leaving the other parts blank, only HUMAN (9606) of the taxonomy filter, and the number of N (the number of

biological terms to be converted into nodes after enrichment analysis), were provided to get default values as 10.

When we examine data for ten proteins returned as a result of https://www.ebi.ac.uk /Tools/crossbar/proteins?gene=ACE2 request that we sent to Protein Resource, only one protein (Q9BYF1) has the "tax_id" field with a value of 9606. Thus, the other proteinss are deleted, and core protein processing starts only on Q9BYF1 (ACE2).

Since neighboring proteins are also included by default, UniProt accessions P01019, P01042, P30990, Q695T7, O15393, P59594, P0DTC2, Q6Q1S2 that are extracted from the proteins -> interactions field are saved to the temporary storage. Since these collected neighboring proteins can not be more than 10, which is the current N value, enrichment analysis is not necessary at this point.

When JSON data is parsed, there is no Reactome pathway and OMIM identifier. Besides, as there is no result from the request https://www.ebi.ac.uk/Tools /crossbar/hpo?genesymbol=ACE2 that is sent to the HPO Terms collection, we can say that there is not a directly related disease, Reactome pathway, or HPO term with the searched ACE2.

After the core protein Q9BYF1, neighboring proteins are processed. When the data returned from the request https://www.ebi.ac.uk/Tools/crossbar/proteins? accession= P01019, P01042, P30990, Q695T7, O15393, P59594, P0DTC2, Q6Q1S2 is examined, tax_id of P30990, P01042, P01019, Q695T7, O15393 proteins comes out to be 9606 (HUMAN). The data for these five proteins is then parsed and Pathways, HPO Terms, and OMIM identifiers are extracted. This information is combined with those coming from the core protein, but in our example, no such data comes from the core protein. After extracting the necessary information, these proteins are combined with the core protein and the process proceeds with six proteins.

The request https://www.ebi.ac.uk/Tools/crossbar/drugs? accession=Q9BYF1, P30990, P01042, P01019, Q695T7, O15393 is sent to Drug collection to fetch the Drugs with the collected proteins. ChEMBL identifiers of these drugs are saved in temporary storage. These drugs are considered in the next step by adding experimental and predicted compounds. Since only ten drugs are imported from the drug collection,

enrichment analysis is not required. However, enrichment analysis is still provided in the search report.

When we send OMIM identifiers that are collected during the processing of core proteins and neighboring proteins to EFO disease terms collection, four diseases are obtained. Even though these diseases are lower than N which is 10, enrichment analysis is again provided in the search report.

Collected 100 HPO Terms and 13 Pathways are incorporated into enrichment analysis, and the most essential 10 are saved in temporary storage to be added to the network.

Remark that the compound with the id 'CHEMBL429844' is included in the DrugBank as a drug (ORE-1001). However, we prefer to leave it with a blue edge, since we retrieved the interaction information from ChEMBL database.

There are no predicted compounds in the network, as there is no result with the proteins that are collected from the "predicted compounds - UniProt accessions mapping table".

As described in Chapter 4, we implemented KEGG database integration with the tables that we keep locally. As the last step of our knowledge graph, we add KEGG Diseases and KEGG Pathways, which are fetched from the local KEGG tables by incorporating them into enrichment analysis. As non-protein interactions, the edges between the 'Renal tubular dysgenesis of genetic origin' EFO disease with many HPO Terms and the 'Hartnup disorder' KEGG disease with KEGG Pathways have also been obtained from the local KEGG tables.

The network created as a result of the ACE2 search is given in Figure 5.1. In addition to the detailed search report, can be accessed through the link https://crossbar.kansil.org/job.php? id=1599317614.

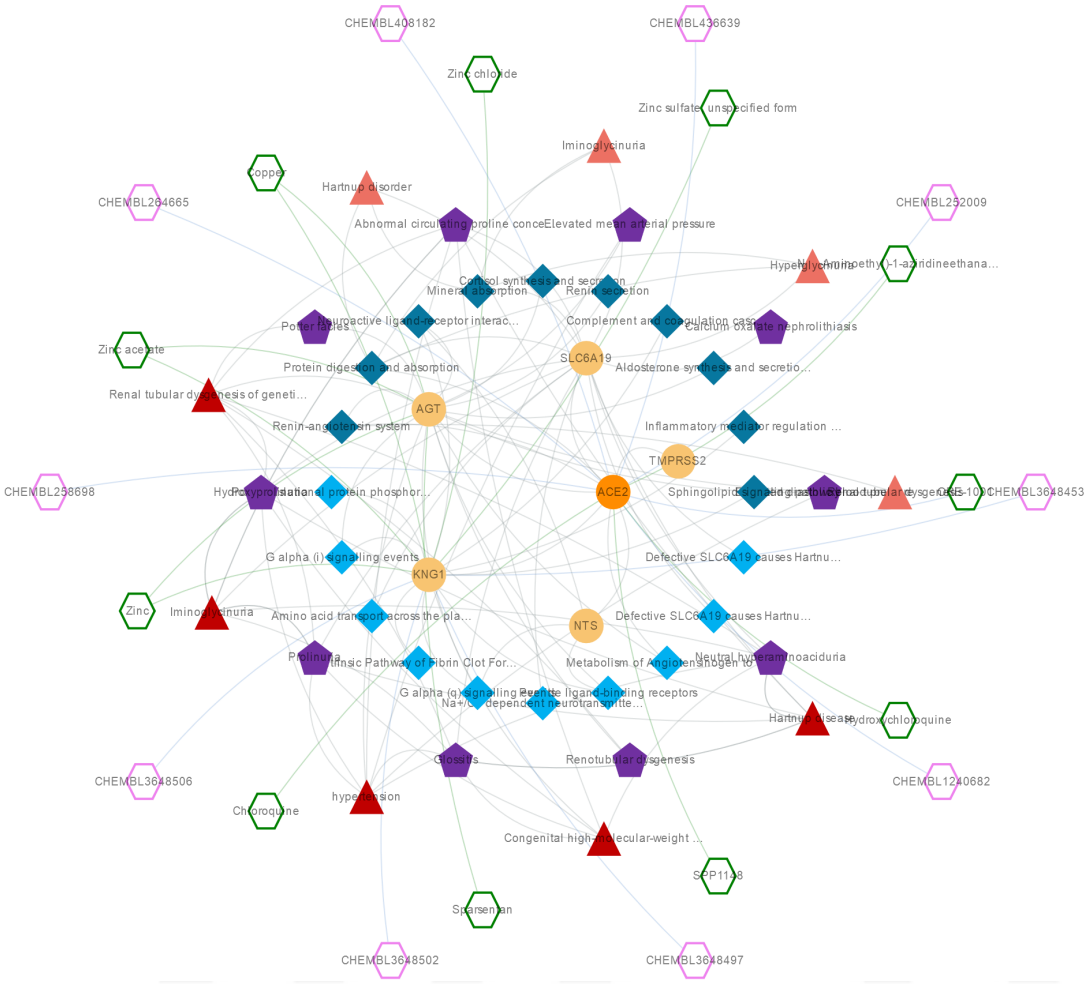

Figure 5.1: Network created for ACE2 protein search case.

5.2 Disease Search Case: Hepatocellular Carcinoma

Hepatocellular Carcinoma (HCC) search starts with default values; core proteins are fetched from both KEGG database and CROSSBAR API EFO disease terms collection as these two sources contain HCC.

Unreviewed proteins being removed, 20 Uniprot accessions are fetched from the KEGG database. With these accessions, the following request (https://www.ebi.ac.uk/Tools/crossbar/proteins?accession=P08581, P60484, P04637, P42771, P01106, P06400, P35222, P37173, P01135, P42336, P08069, O14746, O75084, P01344, O15169, Q14145, Q16236, P14210, O14497, Q68CP9) is sent to the protein resource and

the returned data is saved to temporary storage.

The following request https://www.ebi.ac.uk/Tools/crossbar/efo?label=hepatocellular carcinoma is sent to EFO Resource collection to collect core proteins from the CROssBAR API. We get six more core proteins by sending the OMIM identifier (114550) obtained from this request to the CROssBAR API with the request https://www.ebi.ac.uk /Tools/crossbar/proteins?omim=114550.

These proteins collected from KEGG and EFO disease terms collection are merged and processed, as described in Section 4.3. A total of 1658 HPO terms, 190 Pathways, and 50 OMIM identifiers are collected for these proteins. Since interacting genes/proteins are included by default, enrichment analysis is applied on the neighboring proteins. From CROssBAR API Protein Resource, we fetch data for N (its default value is 10) proteins that have the highest enrichment scores. We combine the HPO Terms, Pathways, and OMIM identifiers extracted from this data with the ones that we fetch for core protein. Disease entities are fetched with the request https://www.ebi.ac.uk/Tools /crossbar/efo?omimId=OMIM:114550 sent to EFO disease terms collection over the combined OMIM identifiers. Drug entities are fetched to be sent to the Drugs Collection via the request https://www.ebi.ac.uk/Tools/crossbar /drugs?accession=, as described in Section 4.5.

Enrichment analysis is applied to disease, drug, HPO, and pathway entities that are fetched and the first 10 entities are registered in temporary storage to be represented in the knowledge graph.

Experimental and predicted compound collection process is also implemented with all collected proteins, as described in Section 4.7. At this stage, as a result of the request from https://www.ebi.ac.uk/Tools/crossbar/drugs?chemblId=CHEMBL3785909, the drug version of compound 'CHEMBL3785909' is named as 'AMG-337'. By turning this compound into a drug, its interaction with the P08581 (MET) protein is defined as a blue edge.

Non-protein interactions, KEGG Diseases and KEGG Pathways are retrieved by means of the queries performed on KEGG tables. Figure 5.2 illustrates the final generated network.

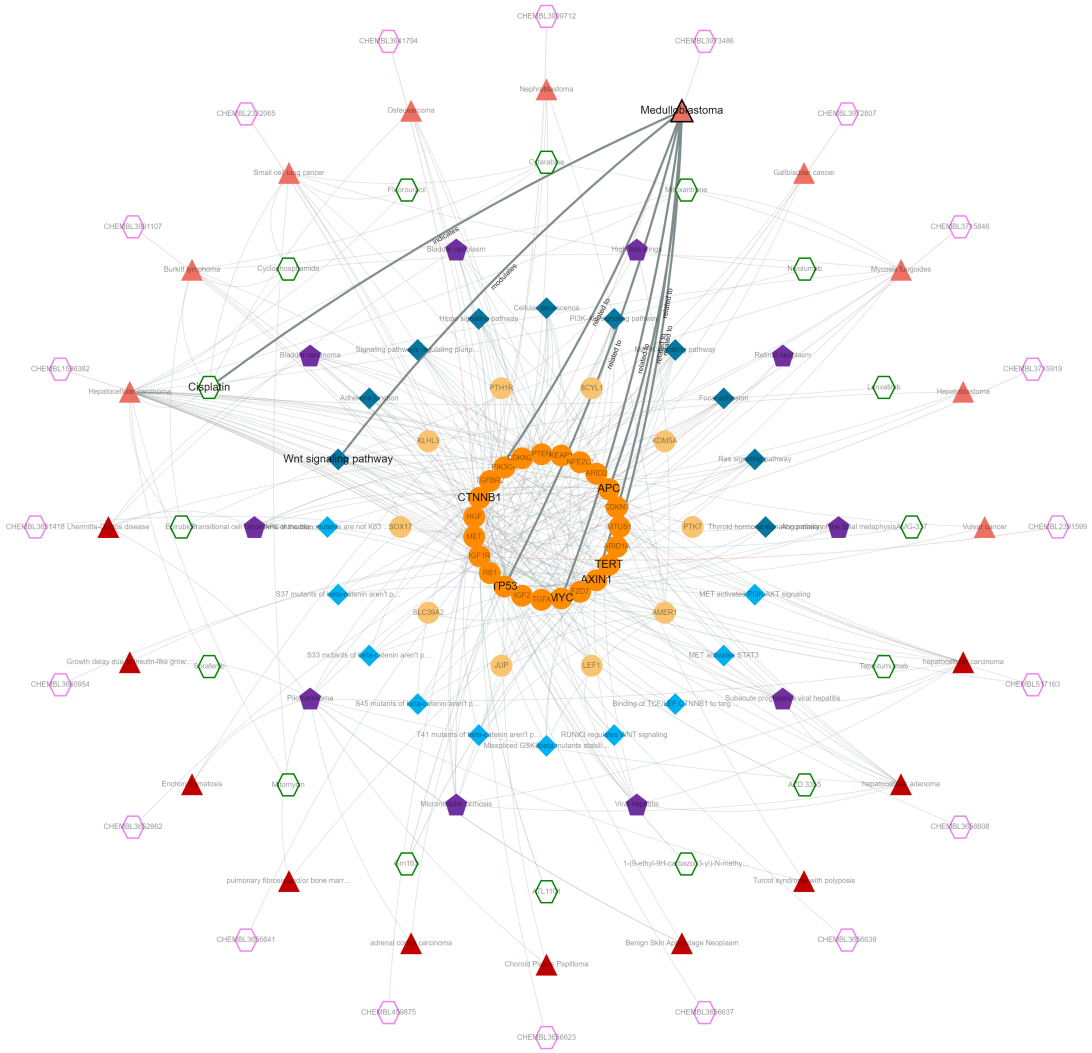

Figure 5.2: Network created for Hepatocellular Carcinome disease search case.

5.3 Drug Search Case: Sorafenib

'Sorafenib' Drug search starts by sending the drug name to CROssBAR API with the following request https://www.ebi.ac.uk/Tools/crossbar/drugs?name=Sorafenib. The process continues by collecting proteins with Sorafenib's ChEMBL identifier and then by saving the proteins to temporary storage (P15056, P04049, P35916, P35968, P36888, P09619, P10721, P11362, P07949, P17948) extracted from the drug data. As described in Section 4.1.1, the request https://www.ebi.ac.uk/Tools/crossbar /activities?pchemblValue=6&moleculeChemblId=CHEMBL1336 with the identifier 'CHEMBL1336' is sent to the activities collection. The proteins P00533, Q08345,

Q16832, Q8NE63, and Q9NYL2 are retrieved in response. These collected proteins are merged and form the core proteins of the Sorafenib search.

Knowledge graph is initially constructed with the obtained JSON file by sending the request https://www.ebi.ac.uk/Tools/crossbar/proteins?accession= P00533, Q08345, Q16832, Q8NE63, Q9NYL2, P15056, P04049, P35916, P35968, P36888, P09619, P11362, P07949, P17948 to the protein resource.

The processes (processing core and neighboring proteins, fetching drug and diseases, enrichment analysis of diseases, drugs, pathways, and HPO terms, fetching experimental and predicted compounds entities, defining non-protein node interactions KEGG integration and processes, respectively) are executed as described in Chapter 4.

Figure 5.3 illustrates the knowledge graph generated as a result of the Sorafenib search. The detailed report for the search is available at https://crossbar.kansil.org /job.php?id=1599485555.

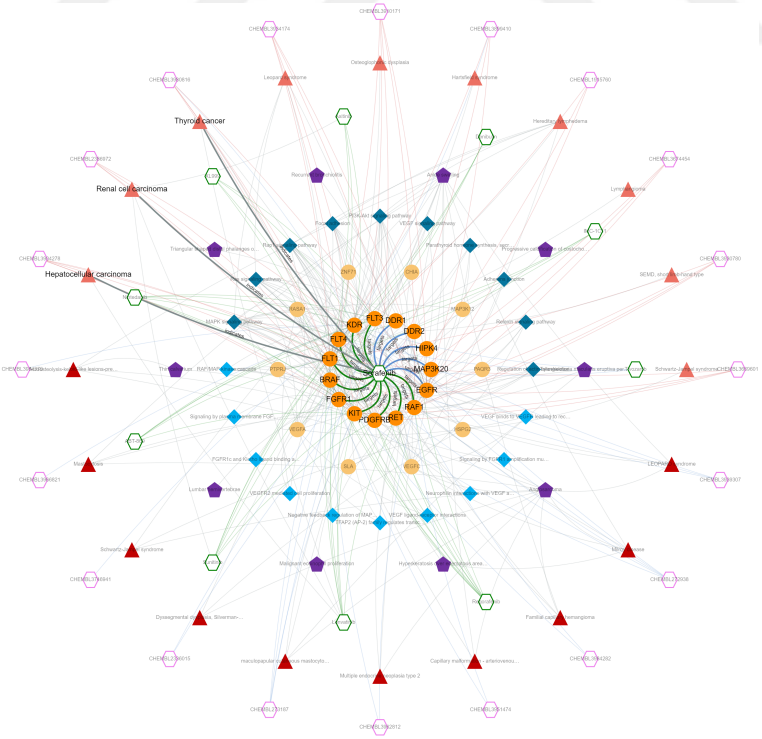

Figure 5.3: Network generated as the result of Sorafenib drug search case.

5.4 Pathway Search Case: Regulation of gene expression by Hypoxia-inducible Factor

Data can be fetched directly from the CROssBAR API Protein Resource in Pathway searches. In this example, since the "Regulation of gene expression by Hypoxia-Inducible Factor" pathway is not in the KEGG database, the core protein collection process is implemented solely through the CROssBAR API Protein Resource. After removing unreviewed proteins from 38 proteins obtained from the request "https://www .ebi.ac.uk /Tools/crossbar/proteins?reactome=Regulation of gene expression by Hypoxia-inducible Factor", we have 11 proteins (P01588, P15692, P27540, Q09472, Q16665, Q16790, Q92793, Q99814, Q99967, Q9Y241, Q9Y2N7). Complete data of these proteins are obtained by sending the request https://www.ebi.ac.uk/Tools/crossbar/proteins?accession=P01588, P15692, P27540, Q09472, Q16665, Q16790, Q92793, Q99814, Q99967, Q9Y241, Q9Y2N7 to protein resource.

Other steps required to create the knowledge graph are the same as the examples given above, and the generated knowledge graph is presented in Figure 5.4 and it can be accessed at https://crossbar. kansil.org/job.php?id=1599484011.

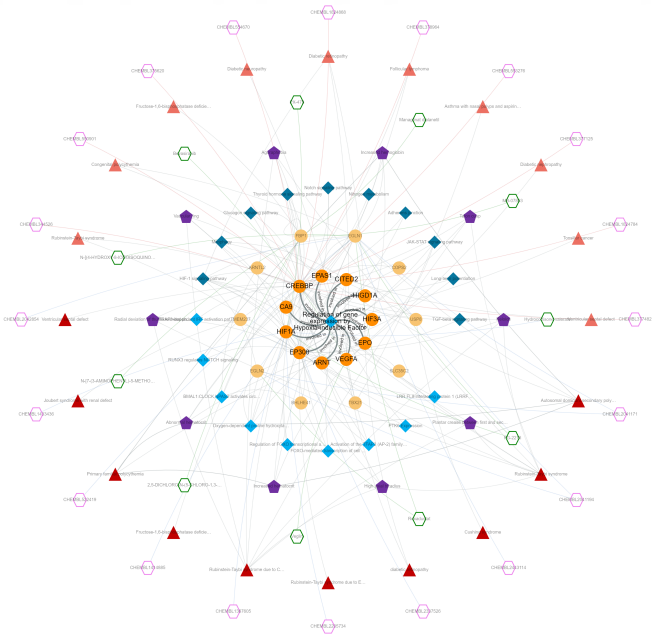

Figure 5.4: Knowledge graph of "Regulation of gene expression by Hypoxia-inducible Factor"

5.5 Mixed Search Case

In this example, a knowledge graph is generated by adding a search term to each node type. Searched terms are as follows:

- Gene: CXCL8
- Drug: Sorafenib
- Pathway: PI3K-Akt pathway
- Disease: Hepatocellular carcinoma
- HPO Term: Abnormal liver philosophy

The relationship between the CXCL8 gene and Hepatocellular Carcinoma was first shown by Kahraman et al. [22]. As illustrated in Figure 5.5, this relationship (between CXCL8 and Hepatocellular Carcinoma) through 'Cellular senescence' is confirmed by the result of this search.

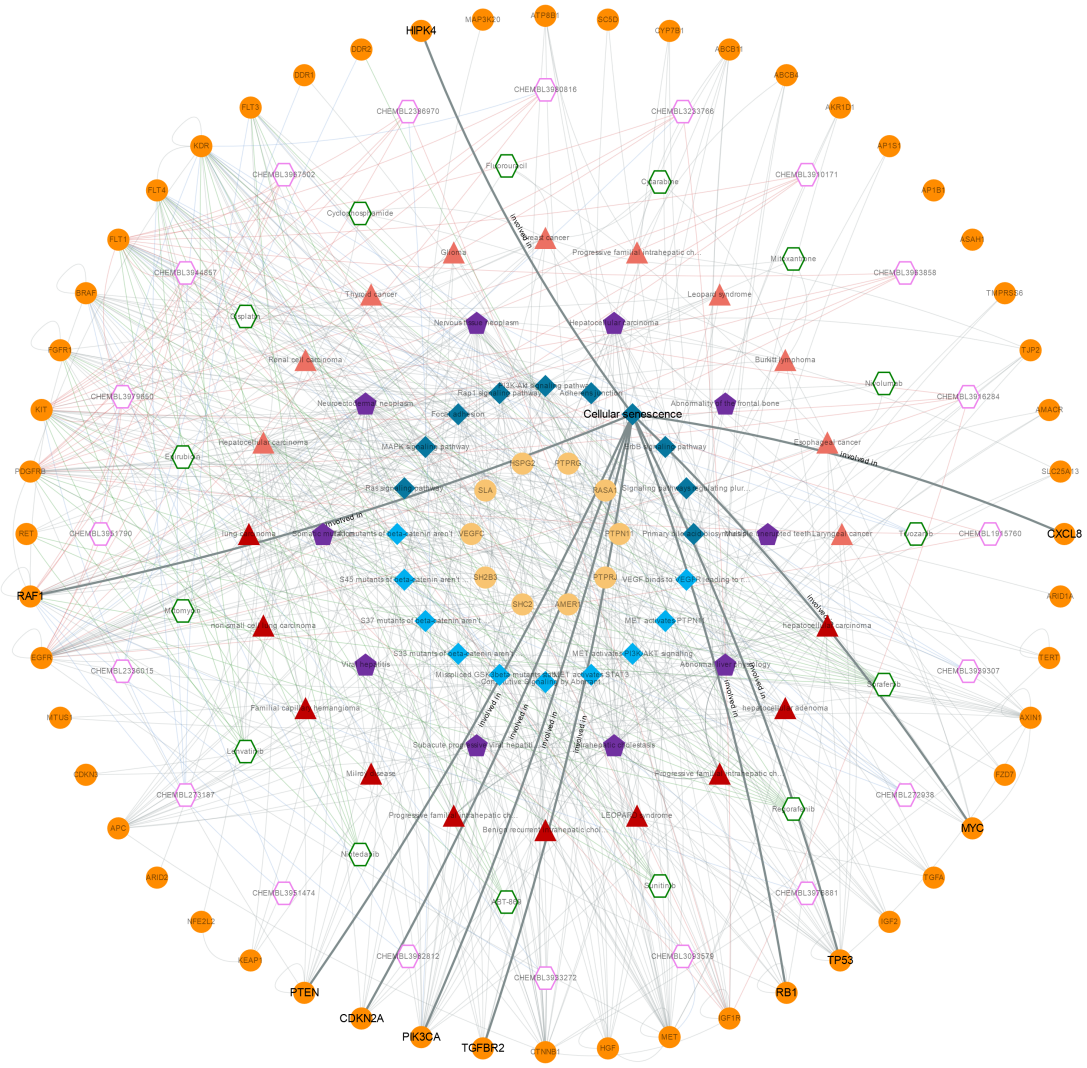

Figure 5.5: Knowledge graph generated for mixed search case.

To the best of our knowledge, there is no web-service in the literature where all component types (diseases, drugs, proteins, pathways, and HPO Terms) can be searched all together, which is an important indicator of the originality of our study. This sample search could be performed again on our web-service, and the same results can be obtained, as well as a detailed report review (https://crossbar.kansil .org/job.php?id=1599484256).

5.6 Unrelated Search Example

The bottleneck of our knowledge graph construction process is that if users choose unrelated components for search, after the construction of the knowledge graph, nodes without any connections may appear and this is not meaningful. There could be two separate clusters that seem to relate on their own, but since most of the nodes are eliminated after enrichment analysis, about half of the information is lost. Example of unrelated search with gene MYC (oncogene, Transcription Factor) and gene ICAM1 (Intercellular Adhesion Molecule 1) is given in Figure 5.6. The detailed report of this search can be accessed at https://crossbar.kansil.org/ job.php?id=1599497736.

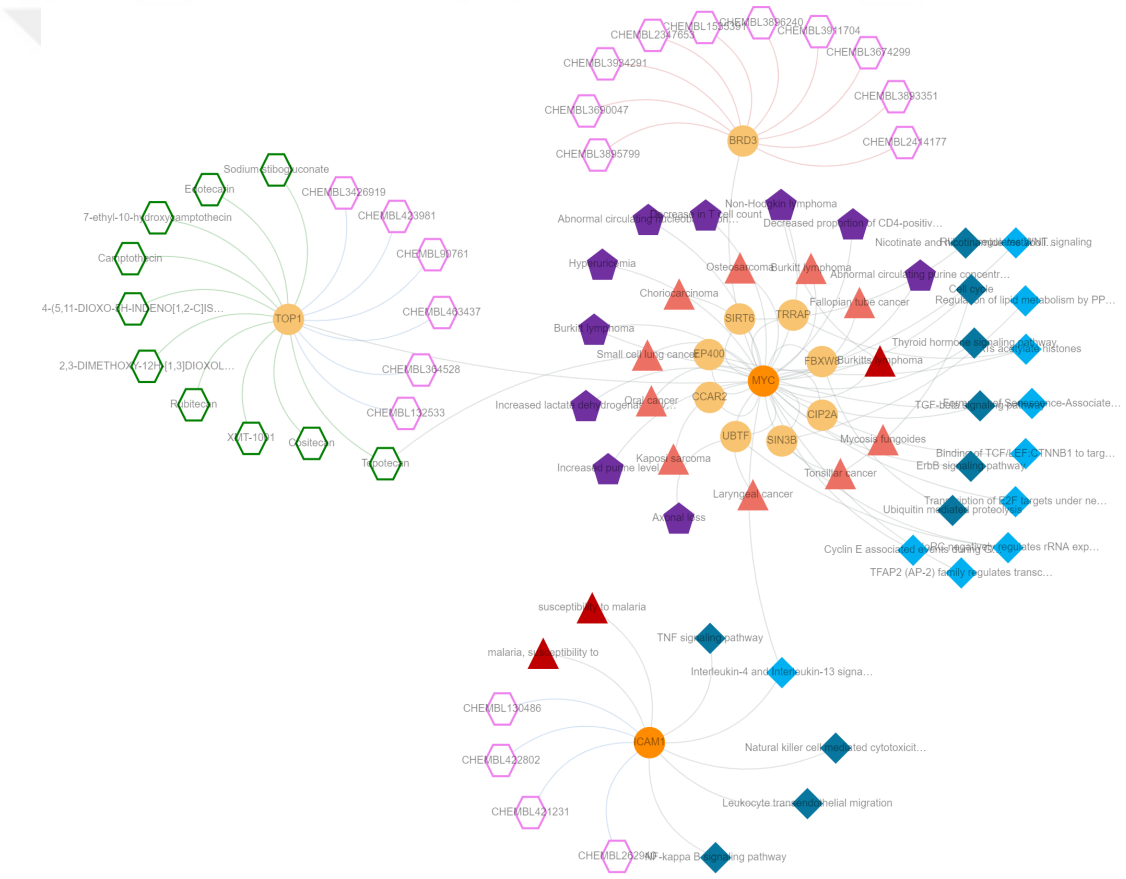

Figure 5.6: Network constructed by unrelated search (MYC & ICAM1).

To avoid that kind of construction knowledge graph, one should search the components separately, as given in Figure 5.7 (MYC) and Figure 5.8 (ICAM1).

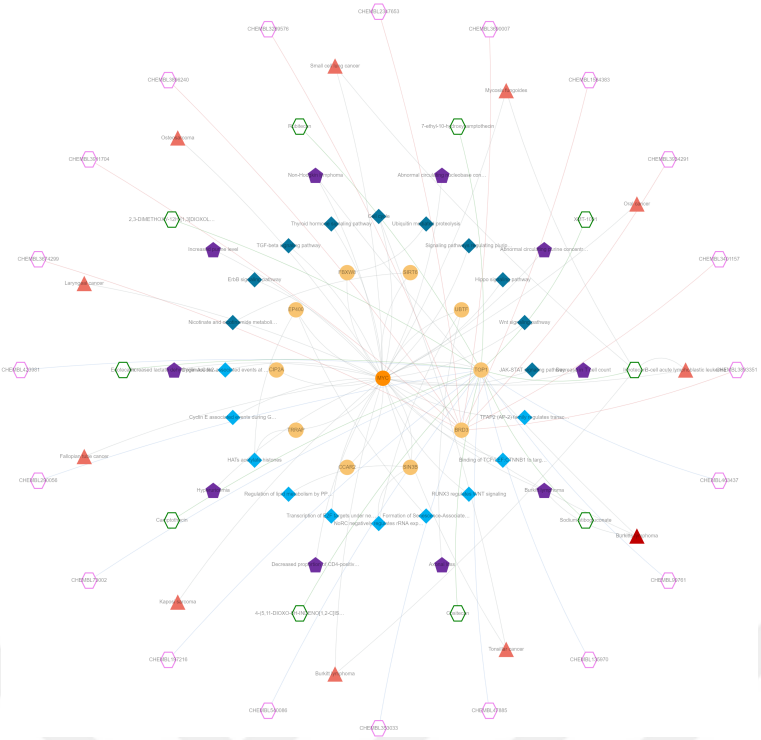

Figure 5.7: Network of MYC gene search.

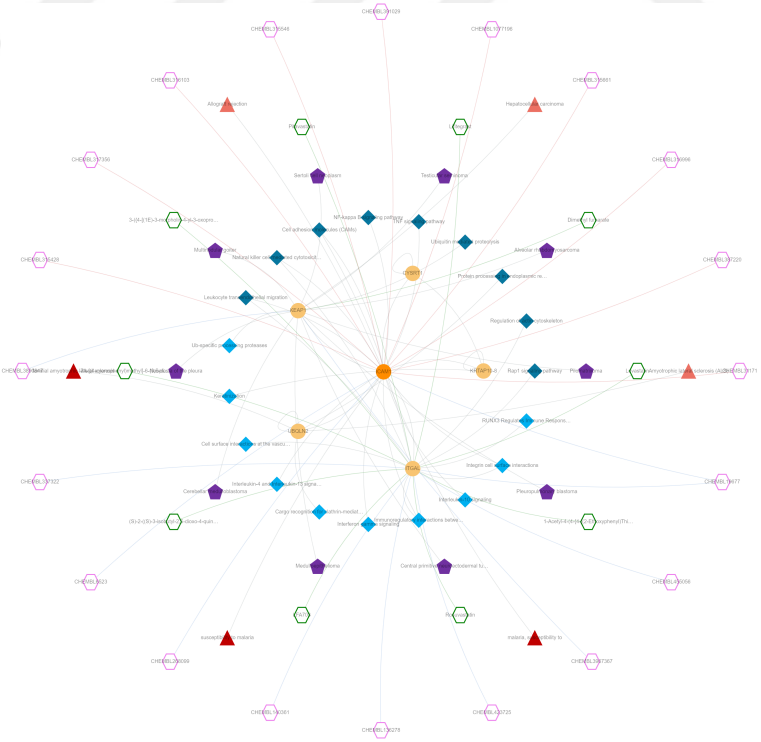

Figure 5.8: Netrok by ICAM1 gene search.

Detailed reports of these searches can be found at https://crossbar.kansil.org/job.php?id=1599498965 and https://crossbar.kansil.org/job.php?id=1599498965. As can be seen in the reports, much of the information obtained when searched alone could not be displayed in MYC & ICAM1 search as a result of the enrichment analysis we implemented to prevent the formation of crowded graphs and to create graphs with only the most significant nodes.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The CROssBAR database was created within the framework of the CROssBAR project. This database comprises various biomedical resources used for the drug discovery process. Additionally the predictions of deep learning-based models are integrated to the CROssBAR-DB.

The ability to access easily all biomedical resources collected within the CROssBAR database framework under a single framework provides a crucial service to researchers, mostly biologists. CROssBAR web service provides the visualization of these combined resources in terms of knowledge graphs. The visualization of knowledge graphs are of great importance in terms of ease of perception and interpretation.

To the best of our knowledge, no study provides the search of the following five different components altogether:

- Genes/Proteins,
- Drugs/Compounds,
- Pathways,
- Diseases,
- HPO Terms.

To fill this literature gap, we present the open-access web service via crossbar.kansil.org

to researchers' use. Thanks to our project's data integration contributors, our study contributes to the literature in two ways:

- Being able to observe indirectly related biological terms altogether and thus generating hypotheses through their relations,
- Generating hypotheses pursuing the system biology approach by observing in the same graph all of the relevant biological items.

We apply enrichment analysis to prevent unnecessary crowds in CROssBAR-WS. We implemented the visualization process of the knowledge graphs on our web service by using Cytoscape, an open-source graph library, to allow free access to information. Cytoscape hosts many interactive events created with the contributions of developers and the core team for visualization problems. CROssBAR-WS has a user-friendly interface based on Cytoscape provided services such as searching the node on the graph, highlighting the selected nodes, and adjusting the node sizes. Also, CROssBAR-WS allows users to customize the knowledge graphs. Multi-layered circular CROssBAR-layout is also added which is developed to make the nodes on the interpretable knowledge graphs look better, as an extension to CROssBAR-WS. We have made possible to apply the built-in layout algorithms provided by Cytoscape to the graphs created together with the CROssBAR-layout.

CROssBAR-WS provideS search report, exporting network as image and downloading data as JSON features via CROSSBAR-WS so that users can trace the details of the research and use them as a source for their own work.

6.2 Future Work

In addition to the CROssBAR-layout algorithm that we have developed by placing the same type of nodes at the points of the circle over a certain radius, we plan to develop it in a way that minimizes edge crossing. We also intend to contribute novel predictions to the database by developing new deep learning models that support the developed models.

REFERENCES

[1] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein interactions and disease," Oct 2016.

[2] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data," *Nucleic acids research*, vol. 44, no. D1, pp. D380–D384, 2016.

[3] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *Elife*, vol. 6, p. e26726, 2017.

[4] M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, J. Mélius, A. Waagmeester, S. R. Sinha, R. Miller, *et al.*, "Wikipathways: capturing the full diversity of pathway knowledge," *Nucleic acids research*, vol. 44, no. D1, pp. D488–D494, 2016.

[5] Q. Cong, Z. Feng, F. Li, L. Zhang, G. Rao, and C. Tao, "Constructing biomedical knowledge graph based on semmeddb and linked open data," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1628–1631, IEEE, 2018.

[6] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 46, no. D1, pp. D649–D655, 2018.

[7] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[8] M. Franz, H. Rodriguez, C. Lopes, K. Zuberi, J. Montojo, G. D. Bader, and Q. Morris, "Genemania update 2018," *Nucleic acids research*, vol. 46, no. W1, pp. W60–W64, 2018.

[9] A. M. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk, and J. Del-Favero, "Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation," *Genome biology*, vol. 12, no. 6, p. R57, 2011.

[10] P. Pareja-Tobes, R. Tobes, M. Manrique, E. Pareja, and E. Pareja-Tobes, "Bio4j: a high-performance cloud-enabled graph-based data platform," *BioRxiv*, p. 016758, 2015.

[11] J. Yuan, Z. Jin, H. Guo, H. Jin, X. Zhang, T. Smith, and J. Luo, "Constructing biomedical domain-specific knowledge graph with minimum supervision," *Knowledge and Information Systems*, vol. 62, no. 1, pp. 317–336, 2020.

[12] MongoDB, "The most popular database for modern apps," 2020.

[13] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. Magariños, J. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. Radoux, A. Segura-Cabrera, A. Hersey, and A. Leach, "ChEMBL: towards direct deposition of bioassay data," *Nucleic Acids Research*, vol. 47, pp. D930–D940, 11 2018.

[14] NIH, "Pubchem," 2020.

[15] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, pp. D1074–D1082, 11 2017.

[16] K. Laboratories, "Kyoto encyclopedia of genes and genomes," 2020.

[17] M.-N. I. of Genetic Medicine, "Online mendelian inheritance in man (omim)," 2020.

[18] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an Experimental Factor Ontology," *Bioinformatics*, vol. 26, pp. 1112–1118, 03 2010.

[19] EBI, "Uniprotkb/swiss-prot and uniprotkb/trembl," 2018.

[20] A. S. Rifaioglu, E. Nalbat, V. Atalay, M. J. Martin, R. Cetin-Atalay, and T. Doğan, "Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations," *Chem. Sci.*, vol. 11, pp. 2531–2557, 2020.

[21] EBI, "Taxonomic identifier," 2018.

[22] D. C. Kahraman, T. Kahraman, and R. Cetin-Atalay, "Targeting pi3k/akt/mtor pathway identifies differential expression and functional role of il8 in liver cancer stem cell enrichment," *Molecular Cancer Therapeutics*, vol. 18, no. 11, pp. 2146–2157, 2019.