



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Doctor of Philosophy

Integration and Visualization of
Pharmacogenomic Data of Human Cancer Cell
Lines and Tissues

The Graduate School
of the University of Ulsan
Department of Medicine

Muhammad Shoaib

Integration and Visualization of
Pharmacogenomic Data of Human
Cancer Cell Lines and Tissues

Supervised by: Professor Suhwan Chang

A Dissertation

Submitted to
the Graduate School of the University of Ulsan
In partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Muhammad Shoaib

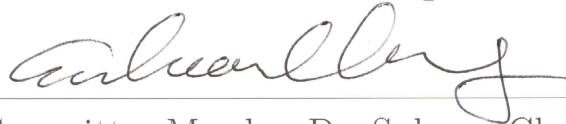
Department of Medicine
Ulsan, Korea
June 2019

Integration and Visualization of Pharmacogenomic Data of Human Cancer Cell Lines and Tissues

This certifies that the dissertation of Muhammad Shoaib is
approved



Committee Chair Dr. Chang Ohk Sung



Committee Member Dr. Suhwan Chang



Committee Member Dr. Sung Min Ahn



Committee Member Dr. Inki Kim



Committee Member Dr. Deokhoon Kim

Department of Biomedical Engineering
University of Ulsan, Republic of Korea
June 2019

dedicated to

Abū Alī al-Husayn ibn Abdillāh ibn al-Hasan ibn Alī ibn Sīnā

and

Abu Ja'far Muhammad ibn Mūsā al-Khwārizmī

Acknowledgments

All praises and glory belong to Almighty Allah – the absolute lord of the entire Universe – Who has blessed me with the power, health, knowledge, intellect, and abilities to conduct research during the course of my doctoral studies. This work would not have been possible without His blessings. I ask Allah to send peace and blessings upon his prophets Abraham, Mousses, and Jesses and upon his last messenger and prophet Muhammad and upon his pure household and noble companions.

I would like to extend my thanks to my advisors Prof Suhwan Chang and Prof. Sung-Min Ahn for their kind support, guidance and valuable advice during my studies at Asan Medical Center. I am very grateful to Prof Ahn for giving me a chance of working in his lab as a Doctoral Student. His mentoring and guidance really helped me to improve my philosophical thinking and scientific writing. I also would like to extend my thanks to committee members Prof. Chang Ohk Sung, Prof Inki Kim and Dr. Deokhoon Kim for their value able suggestions, which helped me a lot in order to improve this dissertation. I also express gratitude to Mr. Young-Sang Park and Ms. Ji-Young Lee from the college of medicine, and Mr. Sean Park at International Student Office, University of Ulsan for their kind help and support during my studies.

I would like to offer humble gratitude to my mother who not only put her all effort

in her capacity for my future but she always encouraged me during my studies and research. Without her prayers, support, and love any of my achievement would not have been possible. I also would like to present my humble gratitude to my father, who has been a great source of inspiration for me, and who taught me patience, honesty, courage, and importance of knowledge. Furthermore, I would like to thank my wife for her support and love during my research. I also thank my brothers Ghasan, Raheeq, Talha, and Awais for their love.

I would like to extend my thanks to my friends Dr. Adnan Ahmad Ansari, Dr. Farhaq Haq and Rayan Muwafiq Alanazi for their kind help, support, healthy discussions, and valuable suggestion. I would like to say thanks to Mrs. Eun-Jun Yu for her kind help, support and assistance during the admission process and during my early days stay at Asan Medical Center. I would like to say thanks to all my labmates Amir Khan, Bilal Mustafa, Ji-Youn Shin, and Sun-Yoeng Lee for their help. It was a really great time with them. I would like to thank all of my friends from Pakistan Community Korea and Pakistan Student Association of Korea. Thank you all.

Lastly, I dedicate this work to Muhammad ibn Musa al-Khwarizmi who was a Persian mathematician, astronomer, astrologer geographer and a scholar of 7th century in the House of Wisdom in Baghdad and Ibn-e-Sina also is also known as Avicenna, author of the famous book *canon of medicine*, a physician and an influential metaphysical philosopher of the pre-modern era.

Muhammad Shoaib

2019.06.10

Abstract

Technical advancement in biology has led to the generation of an enormous amount of multi-omics datasets. Integration of these multi-omics datasets with biological databases is essential for biologists since it allows them to uncover hidden connections between biological entities. However, the process of integrating these datasets is challenging because of their diverse and heterogeneous nature. Since each biological database and omics dataset is developed and generated independently to cover specific biological and omics domain, therefore, their structure – how data is organized – differs from each other. Because of their heterogeneous nature, integration of omics databases has been one of the challenging tasks for omics data scientists.

Resource Description Framework (RDF) is de-facto standard that enables linking heterogeneous resources by providing a unified mechanism to publish data in the form of triples. Databases containing triples is known as a triple store. EBI-RDF platform enabled interpretable and integrated access to six independent biological databases by publishing their triple stores using RDF technology. However, querying these triple stores requires in-depth knowledge about their schema and SPARQL query language. To overcome this limitation in the first part of this dissertation presents cMapper, a gene-centric platform to visualize integrated biological databases in biologist-friendly

fashion. cMapper allows biologists to query six biological databases – (1) UniProt, (2) Expression Atlas, (3) REACTOME, (4) ChEMBL, (5) BioModels and (6) Biosamples – in an integrated fashion without technical knowledge of RDF and SPARQL query language.

The second part of the dissertation presents IPCT – an extended version of cMapper –, a framework that integrates pharmacogenomics data with other biological databases. IPCT integrates genomic aberrations of cancer cell lines from CCLE, drug response data from CTRP, genomic aberrations of cancer tissues from cBioPortal experimental conditions of differentially expressed genes from Expression Atlas, and biological pathways from REACTOME. IPCT allows biologists to search for genomic aberrations of cancer cell lines sensitive to a drug of interest. Conversely, they can search for drugs sensitive to cell lines of interest. Furthermore, IPCT allows users to compare genomic aberrations in cancer cell lines and tissues by integrating

cMapper and IPCT allow users to apply filters on entities of interest. If users enter more than one genes, small molecule or cell lines, they can select options to find common biological objects connected with input. Furthermore, both platforms allow users to visualize their graph on screen or download them in as PNG or GraphML format. IPCT additionally also allows users to download data in CSV and JSON format to perform further analysis. Conclusively the research done in this dissertation addresses the problem of data integration in biology and demonstrates how modern-day data computational methods can be used to present integrated biological data in biologists' friendly way so that biologists can use them to uncover to build their hypothesis by identifying potential hidden relationships between biological entities.

Contents

1	Introduction	2
1.1	Genomics in Cancer Research	2
1.2	Pharmacogenomics data in Target Identification	4
1.3	Role of Cancer Cell Lines in Cancer Drug Discovery	5
1.4	Data Integration to identify cancer Bio-markers	7
1.5	Data Integration to identify anticancer drug targets	10
1.6	Methods for Omics Data Integration	13
1.7	Research Questions	15
1.8	Contribution	16
1.9	Structure of Desertion	17
2	Semantic Web based Data Integration in Life Sciences	18
2.1	Data Integration in Life Sciences	19
2.2	Semantic Web a tool for Data Integration	20
2.3	Semantic Web based tools for Life Sciences	23

2.3.1	RDF based Bioinformatics Knowledge System (BIO2RDF)	23
2.3.2	RDF-based access to NCBI databases (NCBI2RDF)	25
2.3.3	EBI-RDF Platform for Life Science Data Integration	26
2.3.4	The Semantic Enrichment of the Scientific Literature (SESL)	28
2.3.5	Text mining for Disease-Gene associations	29
2.4	Limitations and Challenges	31
3	cMapper: gene-centric connectivity mapper for EBI-RDF platform	32
3.1	Abstract	32
3.2	Introduction	34
3.3	cMapper Overview	37
3.3.1	cMapper Connectivity Tables	38
3.3.2	cMapper Data Tables	40
3.3.3	Web Portal	42
3.4	Results	42
3.4.1	Browsing, Searching and Filtering	42
3.4.2	Finding shared connections	45
3.4.3	Downloading GraphML	48
3.5	cMapper Updater	49
3.6	Case Study	51
3.7	Discussion	54

4	Integrated Pharmacogenomic Platform of Human Cancer Cell Lines and Tissues	56
4.1	Abstract	56
4.2	Introduction	58
4.3	Materials and Methods	59
4.4	Results	64
4.4.1	Data Exploration	65
4.4.2	Comparison Between Cell Lines and Real Tissues	66
4.4.3	Filtering Genes	73
4.4.4	Finding Shared Connections	73
4.5	Download Graph	75
4.6	Case Study	78
4.7	Discussion and Conclusions	79
5	Discussion	93
	Bibliography	97

List of Tables

3.1	Basic information in cMapper data tables	41
3.2	Database filters	43
3.3	User Input query for Figure 3.3	45
3.4	User Input query for Figure 3.3	47
3.5	User Input query for Figure 3.4	49
4.1	Database filters that can be applied to searches in the IPCT	67
4.2	User Input query for Figure 4.3	70
4.3	User Input query for Figure 4.7	75
4.4	User Input query for Figure 4.8	77
4.5	Genes associated with Lapatinib. These genes have genomie aberrations in more than 20% cell lines sensitive to Lapatinib	83
4.6	Genes associated with Afatinib. These genes have genomie aberrations in more than 20% cell lines sensitive to Afatinib	84
4.7	Genes association score with Lapatinib and Afatinib	85

4.8	Pathways associated with genes having genomic changes in cell lines sensitive to Lapatinib and Afatinib	86
4.9	Genes associated with Dabrafenib. These genes have genomic aberrations in more than 20% cell lines sensitive to Dabrafenib	87
4.10	Genes associated with Tramentinib. These genes have genomic aberrations in more than 20% cell lines sensitive to Tramentinib	88
4.11	Genes association score with Dabrafenib and Tramentinib	89
4.12	Pathways associated with genes having genomic changes in cell lines sensitive to Dabrafenib and Tramentinib	90

List of Figures

1.1	Methods developed to study omics data during past 50 years	3
1.2	Drug Protein intersection Network, an example to demonstrate relationship between drugs and genomic entities	9
2.1	Architecture of Bio2RDF	25
2.2	NCBI2RDF System Architecture.	27
2.3	SESL architecture for integration of scientific literature.	29
3.1	Graph illustrating how data points are linked in EBI-RDF platform using ontologies and shared vocabularies	36
3.2	cMapper output for the user query FGF19, CTNNB1, and STAT3 with the shared connection filter disabled	44
3.3	cMapper output for the user query FGF19, CTNNB1, and STAT3 with the filter for shared connections enabled	46

3.4	cMapper output for the user query FGF19, CTNNB1, and STAT3 with the filter for shared connections and the associated genes database filter enabled. In this map, connected data entities represent genes connected with multiple inputted genes through metabolic pathways	48
3.5	cMapper showing data entities connected with FGF19, and CD274, without applying any filter	51
3.6	cMapper showing all FGF19, and CD274 pathways	52
3.7	cMapper output showing common associated genes between PDL1 and FGF19 using shared connection filter.	53
3.8	cMapper showing common pathways and associated genes between PDL1 and FGF19 and PIK3CA.	53
4.1	Overall Architecture of IPCT Database	61
4.2	Entities connected in the IPCT Database	62
4.3	IPCT output for small molecule user query lapatinib with shared pathways	68
4.4	IPCT output for small molecule user query lapatinib, with all pathways	69
4.5	IPCT output for small molecule user query Lapatinib, Sorafenib, Gefitinib and Sunitinib to identify commonly mutated genes in cell lines sensitive to input small molecules.	71
4.6	FAT4's genetic profile in real tumors extracted from cBioPortal. (A) FAT4's mutation and alteration frequency in different cancer studies (B) FAT4's Differential expression in different cancer studies.	72

4.7	IPCT output for small molecule user query Lapatinib, Sorafenib, Gefitinib and Sunitinib after disabling REACTOME and Expression Atlas Databases and enabling Cell Lines and Mutated Genes only. (A) illustrates the results with gene filter = cancer genes, and (B) illustrates the value with gene filter = exclude common mutations.	74
4.8	IPCT output for small molecule user query LAPATINIB, SORAFENIB, GEFITINIB and SUNITINIB with shared connection filter enabled. . .	76
4.9	IPCT output for small molecule user query lapatinib and afatinib. The graph shows all data points connected with lapatinib and afatinib. . . .	80
4.10	IPCT output for small molecule user query lapatinib and afatinib with the shared connection filter enabled.	81
4.11	IPCT output for small molecule user query lapatinib and afatinib with the shared connection filter and the relationship filter enabled.	82
4.12	IPCT output for small molecule user query dabrafenib and trametinib with the shared connection filter and the relationship filter enabled. . .	91

“Data is a precious thing and will last longer than the systems themselves”

Sir. Tim Berners-Lee

“You can have data without information, but you cannot have information without data”

Daniel Keys Moran

Chapter 1

Introduction

1.1 Genomics in Cancer Research

Omics is derived from Greek suffix *ome* that means collection or body. The term omics, in biological science, is used to study the molecular activities at the different levels in a biological system such as genes, proteins, transcripts, and their functional interactions in an integrated fashion³⁹⁾. This integrated analysis facilitates researchers in understanding connections between multiple complex biological systems. The recent development in technology has transformed the way of data collection that has resulted in the availability of massive omics data. This data has served as a building block for the development of large biomedical data repositories.

The term genomics was coined by Thomas H. Roderick in 1986 by melting prefix of word genetics and suffix of omics that merged as a branch of omics technologies which studies genomes of different organism⁸⁴⁾. In cancer, genomics studies are undertaken to identify unique biomarkers such as discovering novel oncogenes, tumor suppressor genes and driver mutations⁷⁰⁾ and identifying deregulated pathways to understand tumori-

genic mechanisms in tumor cells¹⁸⁾. Once these biomarkers are identified, researchers then try to find their clinical relevance to identify target candidates of anticancer drugs. Cancer drugs can be classified into two different categories, (1) cytotoxic drugs and (2) targeted therapeutic drugs. The main tasks of cytotoxic drugs are to control cancer cell proliferation and prevent their replication and growth, whereas targeted therapeutic drugs are used to block specific pathway by controlling the activities of particular gene(s) products^{76,28)}. Research has shown that cancer cells require specific kinases such as HER2, mTOR for proliferation and cell growth²⁸⁾ and targeting these specific kinases can result in promising outcomes to control cancer cells proliferation.

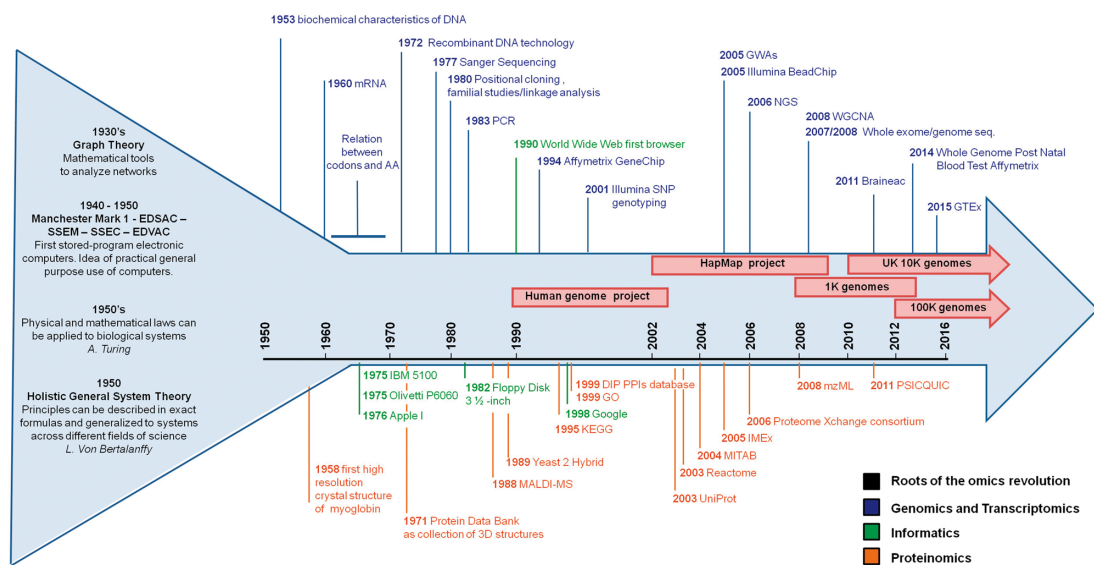


Figure 1.1: Figure adapted from Manzoni et al.⁴⁸⁾ represents methods developed to study omics data during past 50 years. This paradigm shift has revolutionized the way that biologists used to study science to understand biology of human body and make scientific discoveries related to human diseases.

The human genome project⁷⁹⁾ had made the paradigm shift in Genomics research and has revisited the methods of understanding complex diseases like cancer²⁷⁾. Prior to Human Genes Project, researchers used to study individual genes or a small set of genes

and their functions that are assumed to have a role in cancer or other genetic diseases. Figure 1.1 presents the landscape of cancer research for the past five decades. However, after the Human Genome Project, thousands of cancer genomes have been sequenced to study causes of different types of cancers by identifying genomic level changes like Single Nucleotide Polymorphism (single Point mutations) and Copy Number Variants. These efforts have provided an enormous amount of genomic data to study the role of genomic variations in different cancer types in the context of developing new targeted therapeutic drugs.

1.2 Pharmacogenomics data in Target Identification

Pharmacogenomics is an interdisciplinary domain which studies the association between genomic aberrations of an individuals and drug response. Studies have shown that some drugs response was better to the patients with specific genomic aberrations than patients without those genomic aberrations. Similarly, individuals having particular genomic aberrations can respond better to some drugs than others. For example, researchers have found a relationship between patients with HER2 overexpression and their response to the Lapatinib⁶⁴). This is also true for many other drugs. Therefore, researchers in the domain of Pharmacogenomics are trying to find novel associations between drugs and genomic aberrations to aid clinicians in finding appropriate individual treatment to their patients based on their genotype. Conversely, Pharmacogenomics is also being used to identify new drug targets by performing integrated analysis on drug response and genomic datasets.

As mentioned above, cancer is a genetic disease and is studied using genomic aberra-

tions which includes gene expression, somatic mutations, and copy number alterations. Most cancer treatments rely on cytotoxic agents that disrupt the process of cell proliferation in cancer tissues or sites and do not take genomic changes into account. However, researchers are now trying to develop targeted therapeutic drugs since they have improved the survival rate in the past. Trastuzumab, for example, a therapeutic agent that is used to treat metastatic HER2 overexpressed breast cancer patients have demonstrated promising results⁵⁴⁾. Pharmacogenomics databases in cancer research are used to identify clinically relevant new subclasses and associate them targeted therapies by identifying a correlation between genomic profiles and drug response. These databases provide aid to researchers in developing new targeted agents and repositioning or repurposing existing drugs against new targets.

1.3 Role of Cancer Cell Lines in Cancer Drug Discovery

The role of cell lines to understand diseases and drug mechanism can be tracked backed to 1950s⁶⁹⁾. Since then, researchers have been using cell lines to understand the reasons for complex diseases, and the mechanism of action for multiple drugs. In cancer research, BT-20, a breast cancer cell line was the first Cultivated in-vivo model that was established in 1958⁴¹⁾ and has led cancer research towards a new orientation. Since then, researchers and research groups started using cell lines cultivated from cancer tissues to study genomic aberrations and molecular activities of diseases and treatments. These cell lines are now being used to understand the molecular mechanism of different complex diseases like cancer and Alzheimer. This resulted in the development of cell line panels containing cell lines cultivated from different tumor and organs to study tumor heterogeneity and homogeneity along with the mechanism of action for a

particular drug or set of drugs in different organs and tumors.

In the 1990s, for the first time, researchers at the National Cancer Institute derived cell lines from 59 human tumor samples known as NCI60 panel that provided opportunities to researchers to study genomic characteristics and molecular properties of tumors. Experiments performed by employing NCI60 have resulted in large-scale datasets that have been used to identify driver mutations and design gene signatures. Moreover, NCI60 has been used to test the response of more than seventy thousands of compounds²⁰).

Besides the NCI60 panel, cancer-specific cell lines panels are being widely used by researchers to study the biology of specific cancer, and discover clinically relevant sub-types based on genomic aberrations. Liu et al. analyzed a panel 56 colorectal cancer (CRC) cell line to uncover the impact of TP53 mutation and on its expression⁴⁴) in Colorectal cancer. Ovarian cancer cell line panel (OCCP) is another cell lines panel which profiles 39 Ovarian cancer cell lines cultivated at European collection of cell cultures, University of Innsbruck, and Utrecht University. OCCP covers four morphological (molecular) subtypes of Ovarian cancer⁴).

These studies have been presented as an example, dozens of other cell line panels have been profiled to study a different type of cancers and other diseases. However, all these studies were limited in term of breadth and depth, therefore, In order to overcome this problem Cancer Cell Encyclopedia was developed which provides genomic characteristics of 947 Cancer cell lines of 17 organs, that had been tested against 24 anticancer drugs. Because of its unique nature, from its development, CCLE has widely being used in cancer research since it can present near to all types of known cancers. Cancer Therapeutic Response Portal (CTRP) developed by Broad Institute of MIT

profiled 841 drugs response against 860 CCLE cancer cell lines. CCLE and CTRP provide substantial information to design in-silico and in-vivo experiments in the domain of cancer research and drug development. The objective to develop CCLE and CTRP databases was to enable researchers to characterized genomic features against drug sensitivity and resistance. For example, by integrating data from CCLE and CTRP one can explain how cell lines resistant to a certain drug differ from those that are sensitive to that drug at the genomic level.

For the first time, the relevance between in-Vetro models of cancers and cancer cell line was questioned by Kummar et.al. in 1970s⁴⁰⁾. Since then this discussion is ongoing and it is believed by a group of scientific communality that with the passage of time cell lines start differing from the original tissues because of missing micro-environment from which they were cultivated⁸³⁾. Lack of the tumor microenvironment allows cell lines to lose the tumor heterogeneity which was originally present in the tissue. Since then researchers are studying coherence between cancer cell lines and tissues using different methodologies. This dissertation presents a data-centric method to compare cancer cell lines and tissues in the perspective of drug repurposing and repositioning.

1.4 Data Integration to identify cancer Bio-markers

Data integration in biology has always been an essential challenge in biomedical research. Data integration is essential because it allows biologists to examine available information in multiple contexts. However, it possesses an equal challenge for data scientists because of its diversity and heterogeneity. Since data is being generated independently and without considering other data generation systems into account

therefore, each dataset is unique in its nature. Common vocabulary plays a key role in data integration, therefore, researchers have been trying to develop common vocabularies. Chapter 2 of this dissertation provides a detailed overview of efforts taken by researchers to develop common vocabularies using ontologies.

In cancer research, biomarkers are quantitative characteristics of tumors that are measured using different techniques. These biomarkers are used in disease prognosis and drug response prediction. Genomic aberrations are one of those quantitative characteristics that can be measured in the form of genes expression, single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) using different techniques. With the development of NGS technologies, measuring these genomic aberrations has become very easy and an effective way of identifying cancer biomarkers. Therefore research in cancer drug discovery is about trying to associate genomic aberrations with drug response. Figure 1.2 presents an example of drug-protein interaction network and demonstrates a relationship between drugs and genomic entities.

Developments in the domain of genomics have resulted in a paradigm shift in cancer research and treatment. Technology has allowed the generation of an enormous amount of data at a very low cost. A very simple example of this advancement and its effect on science and social welfare is the cost of sequencing a human genome which is approximately 1000 USD today. This cost is far lower than the cost of the first human genomic project and technologists are aiming to reduce it to 100 USD in the near future. This advancement has led us to generate an enormous amount of data for genetic diseases like cancer. An example of this is The Cancer Genome Atlas (TCGA) that provides genomic data for nearly 11,000 patients from 36 different cancer types.

On the other hand, advancement in computational technology has allowed the de-

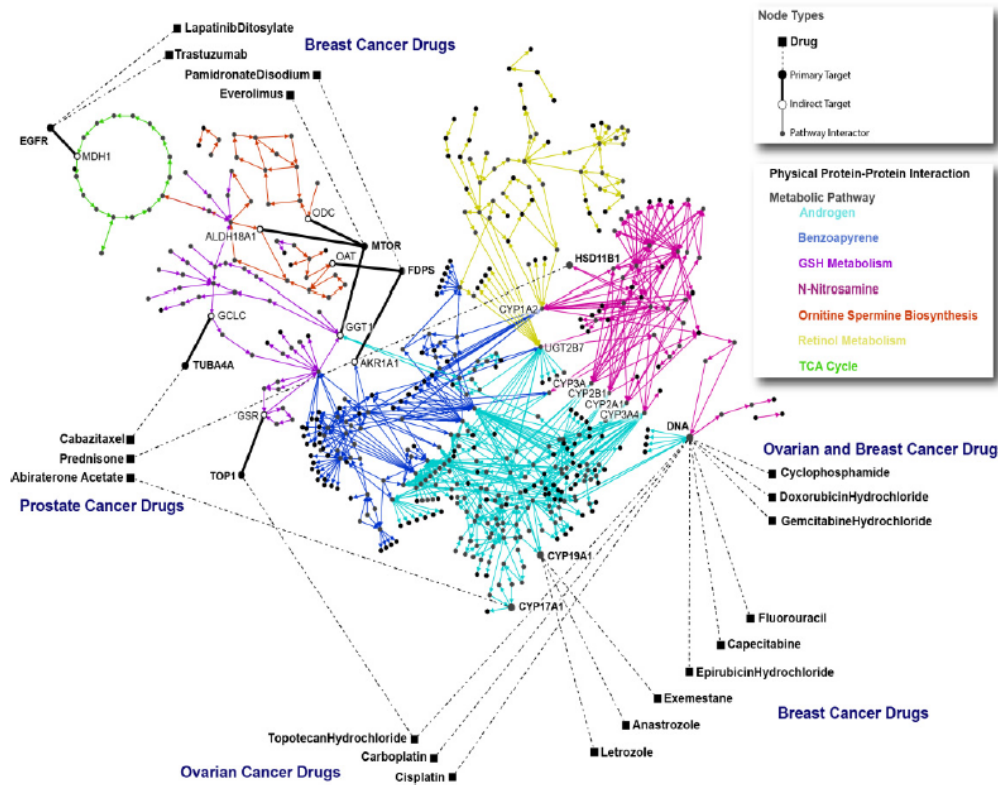


Figure 1.2: Drug Protein intersection Network, an example to demonstrate relationship between drugs and genomic entities

velopment of large scale biological databases to store an enormous amount of biological data, analyze it to understand cancer mechanism and drug actions. Researches have proven that integrated data analysis is an effective tool for understanding cancer mechanism, classifying it into subtypes and designing targeted treatments based on patients' profiles⁵¹). For example, Michaut et.al has used integrative analysis for subtyping of invasive lobular breast cancer⁵³). Similarly, Verhaak at.al. used integrative genomic analysis to study the association between of Glioblastoma multiform subtypes and different neural lineages⁸⁰). Another study conducted by Chitale et.al. has demonstrated the association between expression of DUSP4 and EGFR mutations in lung cancer and showed that DUSP4 has lost functions in EGFR mutated samples in lung cancer¹⁰).

Rampal et al. at Memorial Sloan Kettering Cancer Center, New York used integrative analysis to identify the role of JAK-STAT pathway in myeloproliferative neoplasm pathogenesis⁶³). These are not only four examples from different cancer types, but a number of examples can be found that have used integrated analysis to find cancer biomarkers and their clinical associations with different drugs.

Beside all efforts made by biomedical and omics data scientists in the domain integration, it is still one of the major limitations in the domain of biological data analysis because it is an essential component of large scale analysis.

1.5 Data Integration to identify anticancer drug targets

Approvals of new drugs by Federal Drug Administration (FDA) has decreased by 50% in last decade despite the fact that pharmaceutical industry is investing more in research and development (R&D) to develop new drugs⁶⁰). Failure of a new drug is not lost of money but also time since average drug development duration span over 9-12 years. These challenges have led the concept of drug repositioning, finding new targets for existing drugs or that are already in the development or production pipeline. Governments are encouraging the concept of drug repositioning by making curated databases of approved drugs publicly available. Drug repositioning and repurposing have been of prominent interest in cancer research where researchers are trying to identify new targets for existing anticancer drugs and developing novel therapies using existing anticancer drug^{42,34}).

Data integration is an integral component of drug repositioning and repurposing processes. Recalling from the genetic era, researchers have been using integrated data

analysis for drug profiling based on the relationship between their chemical structures and mode of action in a particular molecular context³⁰). These efforts rocketed after advancement in the domain of genomics technologies that have resulted in the generation of an enormous amount of genomic and transcriptomic data for cancer cell lines and tissues. The Cancer Genome Atlas (TCGA) project, for example, provides genomic and transcriptomic data for about 11,000 cancer tissues. Similarly, Cancer Therapeutic Resource Portal (CTRP) provides with drug response of 450 drugs tested on 800 cancer cell lines of 17 organs. Integration of genomic and transcriptomic databases with drug response databases provides a unique opportunity to understand disease mechanism, mode of actions and identify new use of existing drugs using computational approaches like machine learning, network-based analysis, and text mining^{42,85}). Therefore, applying data integration in drug targets identification to integrate homogeneous genomic and transcriptomic databases with drug response data is very common.

Integrating genomic, transcriptomic, and pathway databases with drug response allows researchers to identify prospective drug targets and their mode of actions for any particular drugs³²). Therefore, researchers have been trying to develop databases by integrating genomic aberrations with drug response. Mutations and Drug Portal (MDP) developed by Cristian et al integrates mutation and pharmacological information from CCLE and NCI60 databases to find pharmacogenomics associations between cancer cell lines and drugs⁷⁴). Liu et al developed a database by integrating microRNA and mRNA data with drug activities for NCBI cell lines to study the correlation between transcriptomic and drug activities⁴³). Pharmacogenomic Knowledge Base PharmGKB is a curated database containing associations between 5000 variants, 900 genes and 600 drugs extracted by mining biomedical literature⁷⁷). The PharmacoGenomic Mutation

Database (PGMD) is a curated database of 117,000 Pharmacogenomic variants extracted manually from biomedical literature along with their study design, statistical significance, and disease contexts. Data contained in PGMD covers 24 diseases, 1400 drugs and variants from 2800 genes³⁷). Developed in 2013, DGIdb is another curated database of 14,144 drug-genes interactions covering 6,307 drugs and 2,611 known target genes and 7,668 potential targets genes. Drugs in DGIdb are classified into two categories; drugs with targeted genes or proteins have been characterized and drugs whose targeted proteins are not characterized¹³). Takarabe et.al. also developed a tool based on data integration approach to identify drug-targets and integrated Drug-target introspection and chemical structure of compounds from KEGG Drugs with adverse event keywords database for drug-target prediction⁷⁵).

These are limited examples of integrating genomic data to find drug targets. However, all these databases to integrate genomics and pharmacology datasets to identify new drug targets limits in term of breadth and depth. Moreover, most of these databases are based on biomedical literature and does not utilize biomedical data. Furthermore, these databases also do not address the problem of integration impotent component of analyzing cell lines and tissues together. In addition, these techniques are also limited to either genomic or transcriptomic approaches while finding potential drug targets whereas combining genomic and transcriptomic data provides a piece of stronger evidence than using only one them. These limitations can be addressed by providing an integrated platform of genomic, transcriptomic and pathway databases of cancer cell lines and tissues. This dissertation has tried to address the challenge of integrating genomic and transcriptomic data of cancer cell lines, and tissues with anti-cancer drug response datasets in a biologist's friendly fashion to facilitate them in

generating new hypothesis based on available data.

1.6 Methods for Omics Data Integration

Previous sections of this chapter laid down the foundation of the fundamental need for omics data integration in biomarker discovery and drug-target indentations. Examples presented in previous sections demonstrate the importance and need for data integration for biologists in general and omics scientist in particular. Conclusively, data integration is an integral component in omics data science because it allows biologists and omics scientists to put information in multiple contexts while generating new hypothesis⁶⁸). This has encouraged omics data scientists and researchers to design and develop methods for omics data integration. This section presents the state of the art methods and techniques that have been used for the purpose of omics data integration.

Data integration in omics is a challenging task because of heterogeneous nature. Data warehouses, databases constructed using mediated schema and federated databases are three well-known database architectures that are being used for omics data integration⁴⁶). Data warehouses store all data in a single large-scale database. Querying data from a Data warehouse is fast however it is impossible to create a global Data warehouse because of individual requirements and needs of each research group or consortium. These constraints make it a good choice for curated data where that database schema is already agreed upon⁵⁵). On the other hand, Federated Databases enable users to query over multiple databases that are created and maintained separately using common schema(s). Querying these databases is similar to querying web using Google or other search engines where each database represents a website and is

queried individually with the ability of aggregating outputs from all databases⁸²). To eliminate the problem of common schema, researchers developed the concept of the mediated schema(s) that acts as middleware between different databases. In general, the mediated schema is a graph containing nodes that represent objects in all database and edges representing the relationship between nodes. Each database implements an additional layer to translate entities and relationship of the mediated schema into executable queries. Mediated schema(s) enables users to ask query from the Federation of databases without understanding the type of each database and their local schema⁴⁶).

Techniques used to integrate omics data coming from multiple sources mainly rely on a common vocabulary. Therefore, researchers in the domain of omics data science have tried to build common vocabularies. Development of common vocabulary is also a major limitation of mediated schemas based federated databases as mediated schema(s) should be developed using in a standard way to enable access to the data in integrated fashion⁴⁶). Thanks to Semantic Web Technologies that has proliferated the role of ontologies in the process of data integration because of their ability to represent and share complex domain information in a systematic way and to provide methods for formal representation of domain knowledge. In addition to these, ontologies are also being used to design and develop information models that enable storing data in an organized structure⁶⁵). Resource Description Framework (RDF) has provided a framework to represent ontologies in a machine-readable format and to allow users to query multiple data sources in a standardized method. Ontologies developed using RDF not only be used for data integration but they also enable users to query data in an integrated fashion using SPARQL query language⁸).

Like other domains, ontologies have been successfully employed in the domain of

biomedical informatics to build common vocabularies. Gene Ontology, for example, is one of the old biomedical Ontologies that were developed to the conceptualization of knowledge about genes and their products¹²⁾. Experimental factor Ontology is another biomedical ontology that provides an information model to enables them to publish experimental data in a structured fashion⁴⁷⁾. These efforts have also played a key role in the development of integrated biomedical databases because they have been used as a core component in the process of data standardization. However, the major limitation of ontology-driven data integration is that it does not allow biologist-friendly access to data as users need to have in-depth knowledge of ontologies and SPARQL query language to get their query answered. To overcome the problem of biologist-friendly data integration, this dissertation has proposed networked based data integration that deals with the limitation of ontology-based data integration by converting ontological vocabulary to networks and graphs. In addition to this, the dissertation has also demonstrated how domain knowledge can be used with experimental data in an integrated fashion while developing a new hypothesis.

1.7 Research Questions

based on the limitations and challenges explained in section 1.6, work presented in this dissertation was carried out to address the following research questions

1. How integrated biological data can be presented in a biologist-friendly way
2. How integrated pharmacogenomics data can be used to identify hidden links between drugs and genomic aberrations
3. How cancer cell lines and tissues data can be integrated to carry out comparative

investigations

1.8 Contribution

The main objective of this dissertation was to study that how data science methods and techniques can be applied to put biological data in a biologist-friendly way. The research presented in this dissertation was undertaken to address the challenge of investigating cancer pharmacogenomics data in an integrated fashion by integrating genomic data of cancer cell lines and tissues with drug response data. Research in this dissertation has been carried out in two phases. The first phase studied the need for genomic data integration, methods developed for applied and adapted integration of genomic data and their limitations to biologists. This research resulted as a framework cMapper, which integrated six different biological databases and presents their contents to users in a biologist-friendly fashion. The second phase studied the methods to integrate pharmacogenomics data of cancer cell lines and tissues in a biologist-friendly way. cMapper framework developed as the result of research performed in the first phase was adopted in the second phase. Frameworks presented in this research allow biologics to (1) search connected objects with genes or small molecules (2) filter objects based on multiple filtering objects (3) observe hidden connections between genes and objects and (4) search for objects that are connected with multiple genes or small molecules.

This dissertation is partially based on the following academic publications

- **Muhammad Shoaib**, Adnan Ahmad Ansari, Farhan Haq, and Sung Min Ahn. "IPCT: Integrated Pharmacogenomic Platform of Human Cancer Cell Lines and Tissues." *Genes* 10, no. 2 (2019): 171.

- **Muhammad Shoaib**, Adnan Ahmad Ansari, and Sung-Min Ahn. "cMapper: gene-centric connectivity mapper for EBI-RDF platform." *Bioinformatics* 33, no. 2 (2016): 266-271.

1.9 Structure of Dissertation

This dissertation is organized as following:

- **Chapter 2:** presents literature review about data integration in biology, introduces semantic web technologies and presents a summarized version of their application in the domain of biology and genomics
- **Chapter 3:** presents the first phase of research in the form of cMapper framework. It concludes the challenges in the domain of biomedical data integration
- **Chapter 4:** presents the second phase of research in the form of IPCT database
- **Chapter 5:** concludes this dissertation by presenting an overview of the dissertation along with possible future extension and directions that could be taken in the near future

Chapter 2

Semantic Web based Data

Integration in Life Sciences

Advances in computational technology have enabled humans to records the massive amount of data and preserve it whereas the computational algorithms are providing methods of new frameworks for analyzing this recorded data⁵²⁾. Data-intensive computing has converted biological science into a quantitative science. The emergence of data-driven approaches in biological science is changing the way we thought about diseases and their treatments. Medical records help us in understanding the nature, type, frequency, and patterns of any particular diseases⁵⁷⁾²¹⁾. Proteins relationship help us in understanding the way any particular protein affect any specific organism of the human body. The truth is that while doing the treatment we do not know the original relationship between disease and treatment. This is why the most treatments and neediness does not work for patients and we do not know which treatment is working fine and which is failing in depth with real causes. Collecting real-time data from

treatment, integrating them with the literature and making some analytical work can help us in solving this problem of uncertainty about treatments. Furthermore, this can help clinicians and physicians in making scientific reasoning upon their decisions, treatments, and prescriptions.

2.1 Data Integration in Life Sciences

The focus of life science research is to discover and identify the components, represent them as data objects (data points) that make life, understand their function and know the relationship among those data points with that they interact with each other to form a biological system.²⁵⁾ The collection of these data objects can be done using biological data however knowing their intersection with each other requires integration between different databases. Therefore the process of understanding biological systems using data-driven techniques can be explained in two different steps (1) collection of biological data objects and their properties and (2) identifying the relationship among these data points (biological system's components).

Fortunately, the biomedical community is full of data what the missing is creating values from the available data and extract more meaningful relations from it. From modern instruments that record the data to literature, biomedical community has published the enormous amount of data during the previous decade in form of open databases, text files, reports, and web pages to help the medical practitioners, scientists, professional, clinicians and researchers in accessing the recent findings. This data can be for discovering the relationship between different medical entities, For example, common reactions of two genes, experiments performed on two or more than

two specific genes. Finding this kind of relationships are not only helpful for researchers but also helps clinicians and medical professionals to understand the in-depths of their decisions.

Although data integration provides a way to look inside data from many different angles, However, the task of biological data integration is not straightforward and creates many challenges for the data scientists³³⁾ The major problems in the life science data integration include heterogeneity and inconsistency. Efforts have been made in integrating life sciences data however these are limited to integration of three datasets.

The issue of data integration in life sciences and biomedical informatics and its challenges has been discussed among data scientists and biologist from a quiet time²⁴⁾. NCBI2RDF¹⁾, BIO2RDF⁷⁾ and EBI-RDF platform³⁵⁾ are examples of well known efforts that have been taken to develop frameworks for data integration and sharing. However, the focus of these frameworks was to provide a baseline for Semantic Data sharing for the biological community. Among this issue of data integration, creation of association among genes, their platform and diseases remain very common.

2.2 Semantic Web a tool for Data Integration

Semantic vision is to provide conceptual organization to available data. The core aim of Semantic Web also known as Web 3.0 is (1) to provide formal semantics to describe different entities (2) providing frameworks for publishing data on the web in uniform format and (3) integrating heterogeneous data sources with each other and (4) provide a global overview of information by logically connecting data points present in those data sources. To fulfill objectives the concept of Ontology development has gained

huge attention among the research community. Ontologies in any domain provide a way to add formal semantics and describe things using them, provide a unified way to identify the objects and allow integration of heterogeneous data points by adding unified references to their identifiers which help linking objects of different domains with each. Resource Description Framework (RDF and) Web Ontology Language (OWL) is being widely used for the development of Ontologies.

From the emergence of Semantic Web, open linked data projects gained huge attention not only in computing society but in other societies to link information present on different sources with each other using the Uniform Resource Identifiers (URIs). DBPedia project and freebase project and DBLP project is some well-known examples of publishing data in the open linked format.

Semantic Web Ontologies have been accepted widely by the scientific community. The scientific experiment Ontology SEO⁷³⁾ was developed to provide a formal description for scientific experiments by describing them using their characteristics and features. Like other domains, Semantic Web technologies have also affected the research of biomedical informatics⁵⁹⁾. Efforts have been made from a decade to create biological ontologies to provide clear semantics and representations for biological entities. One of the very first projects that were undertaken by the biomedical community for publishing biological data using Semantic web was Gene Ontology (GO)¹²⁾. Gene Ontology has now been accepted as common vocabulary for working genetic datasets and is helping in integrating functional genomes data. It provides annotations about molecule functions, biological processes and cellular components in the graphical structure. GO has been used in many integration and data publishing projects including EBI-RDF platform and integration of molecule data network²³⁾. Another ontology for biomedical

investigation (OBI) was developed to describe biological experiments, their experimental processes, and their components required for biological investigations⁷²⁾. It was developed to address the vocabulary challenge in the cross-disciplinary investigation by providing the descriptions about biological and chemical investigations. Additionally, it also defines roles and functions that are used in different biological investigations. A use case of OBI has been presented in⁶⁾ where authors have explains how biological experiments can be modeled using OBI. OBI has also been used in EBI-RDF platform as well. Applications like SPARQLGraph⁶⁷⁾ was also developed to enable semantic and graphical queries over the semantically enriched biological databases. These applications help biologists to search for resources with semantic annotations without having deep technical knowledge. A detailed literature review on the usage of semantic web tool in biological data integration can be found in³¹⁾.

Efforts have also been made in publishing the biological data in the form of open-linked data on the web to make it accessible for other researchers. EMBL-EBI project is one of the major project undertaken by the European Biomedical Celebratory to publish biological data on the web. To the best of our knowledge, EBI has made the more datasets in open access format then any other organization.

Since Ontologies provide an excellent way of linking entities with each other even in different databases by recognizing the (1) common identifiers and (2) semantics of the entities, they have been widely used in data integration⁸⁾. Although Ontologies have been widely in practice for integration of structural and non-structural data how-ever integrating semantically enriched data and its aggregation in particular in biomedical domains have not been studied yet. One possible reason is the lack of semantically enriched data in the biomedical domain. However, with the development of EBI-RDF

platforms, RDF dumps provide a sufficient of the amount of data for integrating and aggregating semantically enriched heterogeneous data.

2.3 Semantic Web based tools for Life Sciences

Semantic Web Tools have gained huge attention in the biomedical research community. The well known and most widely used ontology tool protege was also developed in the biomedical community. One reason is that SW Technologies allows modeling of complex information in an easy well-defined way and allow the representation of captured domain knowledge as well as in a standardized way.

As briefed in Introduction that enormous amount of data is being published in the biomedical domain. Manual access to all this data is not possible for humans. Therefore efforts have been made in to provide access to all information in a uniform way. Here we in-depth review the work is done to publish biomedical data to linked-open data cloud.

Following subsections presents a brief overview of well-known semantic web based data integration tools developed for making biological data part of linked open data.

2.3.1 RDF based Bioinformatics Knowledge System (BIO2RDF)

*BIO2RDF*⁵⁾ is an open source project that addresses the problem of biological data integration by published life science data into semantic web compliant linked open data format. Two major contributions of BIO2RDF initial version was the creation of stranded URIs for biological objects and make them accessible using REST and converting non-structured biological data into structured (RDF) format. Its focus was

on conversion of documents published life science domain and available on the NCBI website in RDF format. BIO2RDF framework consists of two modules (1) ontology generation and (2) RDFizer program. The process of ontology generation was manual and was accomplished using protege – a widely used tool for ontology creation – whereas RDFizer is a computer program written in Java to convert already existed non RDF documents into RDF format. Rdfizer consists of two different components (1) XML to RDF that convert XML documents of NCBI into RDF using XPath and (2) SQL to RDF for Ensemble databases that fetches the relational data required and translate it into RDF documents and (3) Text to RDF that convert textual documents like proteins information in the text format using regular expression. The three steps process of BIO2RDF framework includes (1) creation and normalization of URIs, (2) data cleansing and pre-processing to make it suitable for RDF representation and (3) development of RDFizer to convert structured and non-structured data into RDF format. Finally, the access to the information was given through Use a REST like an interface. REpresentational State Transfer (REST) URI System and SerQL – a previous version of SPARQL – query language. REST-Like URI architecture allows accessing of objects using HTTP in a normalized and standardized way using rewrite rules.

BIO2RDF group made several remarkable changes to their framework from URI standardization to SPARQL queries originally released in 2008⁷). According to the latest reporting in 2013. Now BIO2RDF framework is compliant with SIO ontology and uses stranded URI <http://www.bio2rdf.org/> and for the recourses. Furthermore, additional new format tools have been added to the framework to convert TSV, CSV and semi-structured data into RDF format. Additionally, federated SPARQL queries have also been allowed on different data sources.

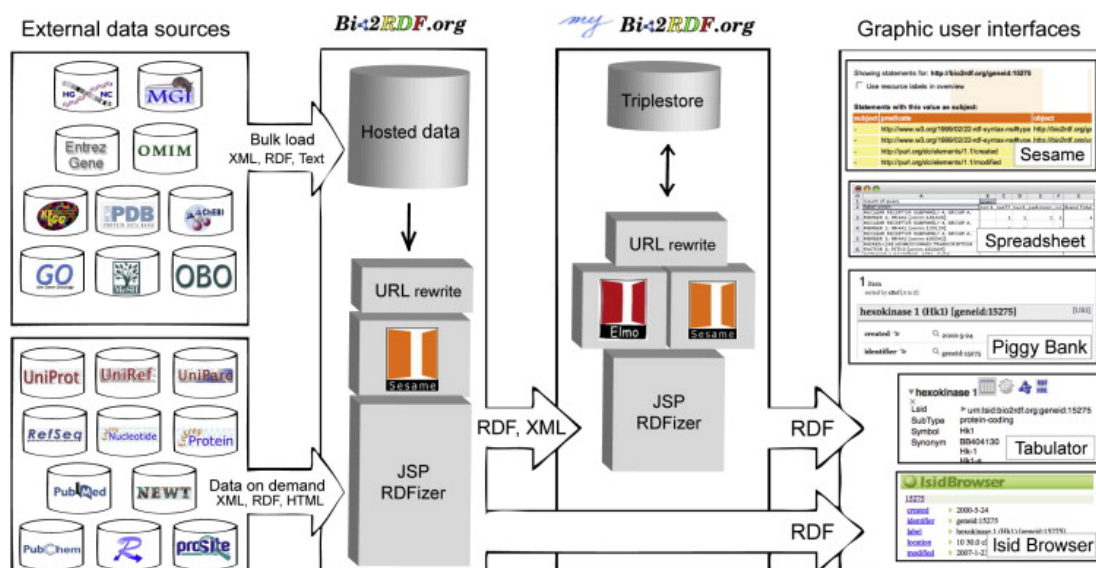


Figure 2.1: Overall Architecture of Bio2RDF Platform. Figure demonstrates the system level flow of information. It shows which data point is connected from which database. Figure adapted from ⁵⁾

2.3.2 RDF-based access to NCBI databases (NCBI2RDF)

NCBI2RDF¹⁾ was a project undertaken to integrate various NCBI resources using RDF and provide SPARQL based access to NCBI resources. It provides access to entire NCBI databases using SPARQL queries that can be run using their developer API. Meta-data Generation and Query Resolution are two basic functions of NCBI2RDF API that it performs. Meta-data Generation module keeps collecting metadata from NCBI about its databases using her E-utilities services that provides databases schema in XML format. This meta-data explains the fields in databases their attributes and their relationships with other databases. Once the meta-data files are fetched RDF is created using these files to allow correct generation of SPARQL queries by the users and to allow mapping of SPARQL variables to databases fields. Each database is defined as a class and its fields are defined as object or data type properties depending

upon the type of field. If it contains only literal value and is not linked with other field NCBI2RDF API marks it as a data type property otherwise it marks it as an object property. In the beginning, this work was done manually however because of frequent updates in NCBI databases the process was automated. Once the meta-data is generated NCBI2RDF API is ready to translate the SPARQL queries into an equivalent set of NCBI service requests that are accepted by the NCBI service endpoints. Queries with the one database are translated into services straightforwardly however queries with the join between databases are translated into more than one service requests that are pipelined then and executed sequentially. Results gathered at the end of each service request are used to create the next request. This makes the process query execution slow. Once all the translated service requests are executed the API combines the results into SPARQL compliant result-set and return back to end the process cycle.

The major limitation of NCBI2RDF database is that in order to utilize the API users must have good knowledge of SPARQL and NCBI2RDF RDF schema and queries must be aligned with it. This may be an easy task for semantic web programmer however it is not for biologist even with a competent knowledge of programming.

2.3.3 EBI-RDF Platform for Life Science Data Integration

*EBI-RDF*³⁵⁾ platform is one of the recent development in bringing biological data to linked open data cloud. The core focus of the project was identifying the points of integration between diverse datasets to answer the question required integration of various datasets. The rationale behind using the Graph-based approach was to build a road map for the development of tools that can easily identify the relationship between

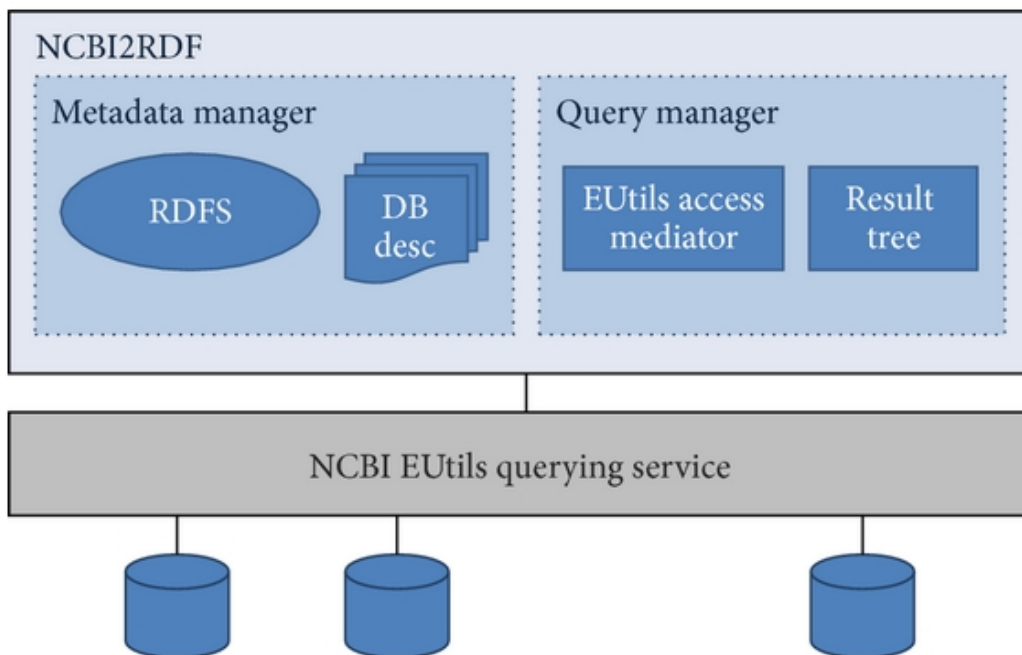


Figure 2.2: Overall Architecture of NCBI2RDF Platform. Figure demonstrates the system level flow of information. Figure adapted from¹⁾

cross-database data points. Instead of creating new URI schemas like NCBI2RDF and BIO2RDF, EBI-RDF platform has made extensive use of already existing ontologies for creation of identifiers of resources and adapted existing vocabularies. Six most widely used databases (1) UniProt, (2) Expression Atlas, (3) Bio-Samples (4) Bio models (5) Reactome and (6) ChEMBL has been made available to the linked open data cloud and are accessible through a SPARQL endpoint. Each database has its own SPARQL endpoint that can be used to query the database. These SPARQL endpoints have the capability of executing simple to complex queries. The key advantage of EBI-RDF datasets is that these datasets use one synchronized URI that has not been ensured in NCBI2RDF and BIO2RDF frameworks. Common URI scheme and ontology-based semantic annotation of data are helpful in data integration to allow the

creation of explicit links between different databases. (we will discuss this aspect of EBI-RDF while discussing our framework in detail in section 3) However, the limitation of EBI-RDF platform is it has not designed the phase of federated queries. EBI-RDF dumps can be downloaded from EBI's website free of cost.

One must have good knowledge of SPARQL, RDF and EBI-RDF schema to compose the SPARQL queries for EBI-RDF platform. this is the same problem that NCBI2RDF and BIO2RDF platforms are suffering from. An effort has been reported in [] to overcome the issue of SPARQL query generation by providing a GUI tool for creation of SPARQL queries automatically. One can make an output graph and the tool will automatically create the SPARQL query that can be passed to EBI-RDF platform. This tool has also implemented the option of federated queries over EBI-RDF platform.

2.3.4 The Semantic Enrichment of the Scientific Literature (SESL)

The Semantic Enrichment of the Scientific Literature *SESL's*²⁹⁾ pilot project aimed to use Semantic Web Technology for adding a meaningful annotation to scientific literature and integrating them using linked open technologies. The major difference between previously explained frameworks and SESL is that SESL deals has used the unstructured and strutted data collectively to create gene-disease relationships and provide the created relationship semantic annotations using already developed biomedical ontologies. In the first phase gene and diseases, identification was performed using LexEBI and UMLS repositories. These repositories acted as the baseline in the identification process. Once the basic identification process was completed sentences having gene-disease pair were identified and the relationship between gene-diseases was marked. Finally, the pair along with the reference information were loaded into the triple store. Once

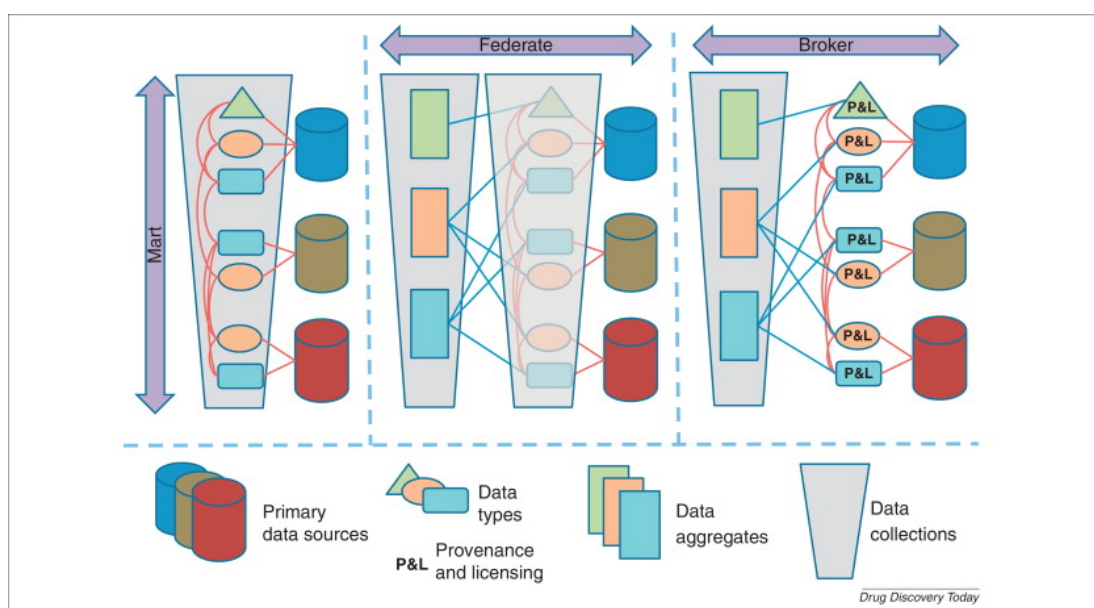


Figure 2.3: SESL architecture for integration of scientific literature. Figure demonstrates the system level flow of information and process of federated query execution at database federation of complex schemes. Figure adapted from²⁹⁾

the pairs are identified and annotated they are linked with UniProtKB and Gene Expression. For this purpose 20,272 proteins, 100,723 functional annotations and 13,897 protein interactions were used in the pilot release of SESL.

2.3.5 Text mining for Disease-Gene associations

Text mining and data integration of disease–gene associations⁶²⁾ is another project that was initiated for the creation of links between diseases and their respective genes from biological literature. The core focus of this project was the relationship discovery between genes and disease using text mining. Four step process of gene-disease annotation includes, (1) construction disease dictionary using Ontology, (2) identification of genes and diseases names from text, (3) Extraction of association between genes and diseases and (4) the integration of computationally extracted data with humanly created liter-

ature. Natural Language based approach for extracting gene-disease relationship with genes from biological literature has been used whereas genes and diseases are extracted using named entity-relationship (NER) and are marked as a paired entity using based on confidence value. The focus of this work was on textual mining, understanding of medical language using NLP. In the first step, DISEASES extracts the list of diseases and their synonyms from disease Ontology and creates a dictionary. In the next step, it identifies the variants of the disease terms which were used interchangeably in the literature. As the next step from textual documents names of genes and diseases are extracted and mapped against the document in which they were found. Data created as the result of gene-disease extraction are then used to create an association between diseases and proteins using co-occurrence based scoring system. To compute the score of co-occurrence authors have used a well-known cosine similarity measure. UniProtKB was used to extract the proteins of the given genes and vice versa. Finally, diseases were linked with the data available on Genetics Home Reference (GHR)⁵⁶⁾. For this purpose, the authors used a crawler to download information about each disease on its web page. Using this way authors were able to map 390 diseases from diseases they extracted using text mining and disease Ontology to the with the GHR database.

Other Life science Data Integration projects include⁵⁰⁾ that explains the ontology-based approach for integrating clinical data. In⁴⁹⁾ authors have presented different case studies of the projects that published there data to open linked data cloud using RDF technologies. This project includes publishing (1) CheEMBL, (2) neurosciences micro-array experiment results and (3) DrugBank as RDF triple store. Notice that these projects were undertaken by different groups and were developed independently, therefore, one can not extract relationships between objects using these three RDF

datasets.

2.4 Limitations and Challenges

The scope of all existing data integration and open linked data publishing tool uses simple Semantic Web Technologies (RDF/OWL) for integration purposes. This is not reasonable for bioscientists, engineers and physicians to ask federated questions using SPARQL queries. It is because of the reason that many few of biological-scientists understand SPARQL, and know how to query using SPARQL endpoint. Because of this technology gap, the fruits of the efforts have not given to the real users.

Chapter 3 of this dissertation has tried to overcome the issue and minimizing the gap by providing a graphical tool based on EBI-RDF datasets to the users. Properties of the developed tools are (1) it allows users to search all entities related to a specific gene along with their relationship with the given gene (2) Limit the output by reducing the number of databases, level of information details, (3) it has capability of finding relationships between two are more than two genes.

Chapter 3

cMapper: gene-centric connectivity mapper for EBI-RDF platform

3.1 Abstract

Motivation: Motivation: In this era of biological big data, data integration has become a common task and a challenge for biologists. The Resource Description Framework (RDF) was developed to enable interoperability of heterogeneous datasets. The EBI-RDF platform enables an efficient data integration of six independent biological databases using RDF technologies and shared ontologies. However, to take advantage of this platform, biologists need to be familiar with RDF technologies and SPARQL query language. To overcome this practical limitation of the EBI-RDF platform, we developed cMapper, a web-based tool that enables biologists to search the EBI-RDF databases in

a gene-centric manner without a thorough knowledge of RDF and SPARQL.

Results: cMapper allows biologists to search data entities in the EBI-RDF platform that are connected to genes or small molecules of interest in multiple biological contexts. The input to cMapper consists of a set of genes or small molecules, and the output are data entities in six independent EBI-RDF databases connected with the given genes or small molecules in the user's query. cMapper provides output to users in the form of a graph in which nodes represent data entities and the edges represent connections between data entities and inputted set of genes or small molecules. Furthermore, users can apply filters based on database, taxonomy, organ and pathways in order to focus on a core connectivity graph of their interest. Data entities from multiple databases are differentiated based on background colors. cMapper also enables users to investigate shared connections between genes or small molecules of interest. Users can view the output graph on a web browser or download it in either GraphML or JSON formats.

Availability and Implementations: cMapper is available as a web application with an integrated MySQL database. The web application was developed using Java and deployed on Tomcat server. We developed the user interface using HTML5, JQuery and the Cytoscape Graph API. cMapper can be accessed at <http://cmapper.ewostech.net> Readers can download the development manual from the website

3.2 Introduction

Data integration has become both a common task and a challenge in biological research. High throughput profiling technologies have led to large-scale data-rich biological research, facilitating the development of various-omics: genomics, epigenomics, transcriptomics, lipidomics, metabolomics, etc.²⁵⁾. In this era of biological big data, it is both an opportunity and a challenge to analyze the vast and various collections of data to discover the underlying biology²⁶⁾.

For biologists, data integration is essential because of the need to put biological questions in various contexts to find unknown connections or new hypotheses. As biologists are the main users of biological data and databases, the traditional approach of data integration has been biologist-friendly. This approach does not require much knowledge or experience in information technology. ENSEMBL is a good example, as it is an integrated platform of multiple genomic databases with a variety of bioinformatics pipelines for data analysis¹⁵⁾

The problem of data integration is not intrinsic to biological big data. This issue has been more thoroughly investigated through studies of the Web Science⁶⁶⁾. Each website is a database containing heterogeneous, but potentially related, data entries. Searching the Web for a piece of information is similar to searching a set of biological big data for a useful biological connection. A data integration approach, such as ENSEMBL, will not work for Web searchers, as it is both large-scale and heterogeneous¹⁶⁾. In other words, it is impossible to create a super-website integrating all data entities from all websites. Instead of creating integrated data repositories, Web scientists have developed a framework called Resource Description Framework (RDF) that enables

linking resources (data entities) from multiple websites based on their Uniform Resource Identifiers (URIs)¹⁶⁾. RDF provides a unified common mechanism to create data models for describing information in the form of subjects, predicates, and objects which are collectively called triples. Each RDF document contains multiple triples. Databases that store RDF documents are called triple stores⁵⁸⁾. Each resource in a RDF database has a URI. These URIs are used to find identical resources (identical data entities) in multiple databases and connect resources from databases with different data models³⁸⁾. Beyond providing a standard framework for publishing data and data models, RDF is flexible, extendable, adaptable, evolvable, and incremental. In summary, RDF provides an easy method of data integration without creating an integrated platform, such as ENSEMBL.

Given the advantages of integrating large-scale heterogeneous datasets, researchers have tried to adopt RDF for biological databases. For example, NCBI2RDF was created to provide integrated access to NCBI databases using the SPARQL query language (Anguita et al., 2013). BIO2RDF was developed to provide integrated access to publicly available biomedical databases such as KEGG, PDB, MGI, HGNC, and a few NCBI databases (Callahan et al., 2013). The EBI-RDF platform is the most recent and systematic effort to apply RDF to biological databases to allow users to ask complex biological questions using SPARQL³⁶⁾.

Technically, the EBI-RDF platform as illustrated in 3.1 is a common RDF triple store of six independent RDF triple stores: 1) Expression Atlas, 2) BioModels, 3) BioSamples, 4) ChEMBL, 5) REACTOME, and 6) UniProt. These triple stores are interconnected through common URIs and shared ontologies. The EBI-RDF platform enables users to ask questions that require integrating data from multiple triple stores

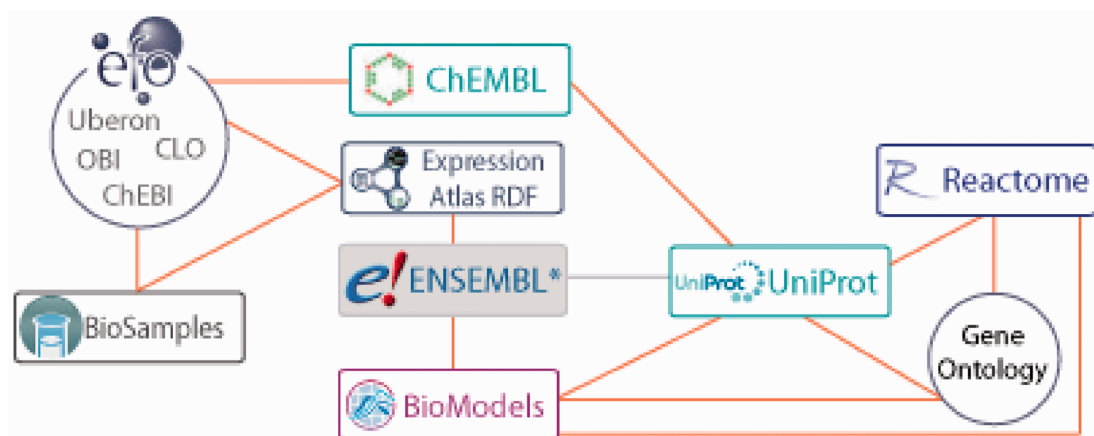


Figure 3.1: Graph illustrating how data points are linked in EBI-RDF platform using ontologies and shared vocabularies. Figure taken from³⁶⁾

using a single SPARQL query. A practical limitation of the EBI-RDF platform is the need for advanced knowledge of SPARQL and data models of six RDF triple stores in order to ask queries across different biological databases.

We have developed cMapper, a gene-centric connectivity mapper for the EBI-RDF platform, starting with the assumption that biologists are most likely to search the EBI-RDF platform in a gene-centric manner. cMapper displays data entities connected with the given genes or small molecules of interest and visualizes them using a graphical interface. Users can easily identify whether their gene of interest has a connected entity in six biological databases in the EBI-RDF platform. Furthermore, they can identify a connected entity and its relationship with the original query. In addition, users can filter output based on (1) shared pathways (i.e., shared data entities between inputted genes or small molecules), (2) taxonomy, (3) organs, (4) metabolic pathways, and (5) signaling pathways. In summary, cMapper enables biologists with limited knowledge of RDF and SPARQL to utilize the EBI-RDF platform in a biologist-friendly way.

3.3 cMapper Overview

The EBI-RDF platform provides a new way of exploiting data in different contexts, such as understanding gene expression in the context of pathways or small molecules³⁶). Our tool allows for investigating connectivity between data entities in a graphical manner. Biologists can start an investigation with a small number of genes or small molecules and then expand the context of the genes or small molecules in order to generate new hypotheses. The distinctive features of cMapper are:

1. Users can construct a graph of data entities in the EBI-RDF platform that are connected to the input (a set of genes or small molecules).
2. Users can select databases. For example, a user may select the databases Expression Atlas and ChEMBL to investigate connectivity of data entities to genes or small molecules of interest.
3. Users can input up to five genes or small molecules to find potential connectivity with other data entities.
4. Users can view the output directly at the cMapper site. Alternatively, users can download and view the file using Cytoscape or other graphing tools.
5. Users can download output graph in GraphML, JSON and Image formats.
6. Upon entering a small molecule, users can obtain a list of genes affected by the small molecule and its relationship with the genes.

cMapper was implemented in two phases: 1) creation of the connectivity database using the EBI-RDF platform containing data from six independent databases; 2) development of the web application with an interactive interface which enables users to

investigate the connectivity map of genes or small molecules of interest in six different databases.

The cMapper database was developed using the MySQL database management system. We used MySQL RDBMS since RDF stores are not suitable for storing and fast processing of large datasets. In other words, on-the-fly queries on RDF stores require a lot of computational resource and time, whereas RDBMSs can efficiently handle queries over large-scale databases. The web application of cMapper was developed using a Java servlet and deployed using the Tomcat web application server. The cMapper database contains two types of tables: 1) connectivity tables that store information about connections between different data entities, and 2) data tables that store basic information about data entities.

3.3.1 cMapper Connectivity Tables

We created connectivity tables by identifying identical data entities across databases using string matching and the owl:sameas property in the EBI-RDF platform. In RDF, the owl:sameas property is used to connect two data entities that are semantically the same but have different URIs. String matching was used to identify data entities with identical URIs across databases, while the property owl:sameas was used to identify similar data entities.

To create connectivity tables, we first extracted gene-protein relationships from UniProt. Second, we connected proteins to REACTOME using UniProt accessions that are common in both databases. Third, we connected the gene-protein-REACTOME connectivity map to Expression Atlas. Expression Atlas contains connections with both genes and proteins. The connections between proteins were directly mapped from RDF

triples using dbXRef property, and connections between genes were mapped by extracting gene names from rdfs:label property. Gene names were identified from rdfs:label objects using Named Entity Resolution (NER). Fourth, we connected ChEMBL to the gene-protein-REACTOME-expression connectivity map. Small molecules in ChEMBL were connected to proteins in the gene-protein-REACTOME-expression connectivity map based on their therapeutic targets. We identified therapeutic targets of small molecules using the combination of RDF properties hasAssay and hasTarget. Identified targets were linked with the proteins in the connectivity map using the RDF property targetCmptXref.

Next, we connected BioSamples and BioModels databases to the gene-protein-REACTOME-expression-chemical connectivity map. Data entities in these two databases do not have direct connections with either genes or proteins. In other words, we were not able to connect them to genes or proteins using the RDF property. To solve this problem, we developed an algorithm that identifies connections between data entities of BioSamples and BioModels and those of UniProt and Expression Atlas. The algorithm (1) filters RDF triples with objects identical to data entities in UniProt or Expression Atlas, (2) marks the predicates of the filtered triples obtained in the first step as bridge properties, and (3) adds triples with bridged properties to mapping tables.

We used pav:derivedFrom properties in BioSamples as a bridging property. BioSamples have two types of sample data: samples used in Expression Atlas experiments, and samples derived from PubMed documents. Samples used in Expression Atlas experiments were linked to the connectivity map using the pav:derivedFrom property. Samples derived from PubMed documents were linked using gene names in the sample description. We extracted genes names from sample descriptions using NER. Finally,

we connected BioModels to the gene-protein-REACTOME-expression-chemical-sample connectivity map using the model’s annotations. First, we identified the necessary annotations using the RDF property `sbmlrdf:versionOf`, a sub property of the annotation property `sbmlrdf:sbmlAnnotation`. Next, we connected UniProt data entities using objects of those triples that have the predicate `sbmlrdf:versionOf` and objects identical to the database references of UniProt data entities. Models having UniProt accessions in data references were connected to the connectivity map directly using the UniProt accessions. However, not all models in BioModels have UniProt accessions; some models have accession IDs from other databases, such as ENSEMBL and Interpro. In these cases, we converted their IDs to UniProt accessions using UniProt DR annotations and NER.

3.3.2 cMapper Data Tables

cMapper data tables contain basic information about data entities in the EBI-RDF platform, summarized in Table 3.1. Since our objective is not to duplicate the information in the EBI-RDF platform but to enable user-friendly, gene-centric, quick access to the EBI-RDF platform to users who do not have essential knowledge of RDF and SPARQL, we have added only basic information about data entities in the data tables. The remaining information can be retrieved using URIs. For example, biological processes and functions can be retrieved from UniProt, while experimental conditions and details, as well as chemical formulae of small molecules; can be retrieved from Expression Atlas and ChEMBL.

Table 3.1: Basic information in cMapper data tables

Database	Basic information about data entities in each database
UniProt	UniProt accessions for protein identification, gene IDs and names for gene identification, organism, and data entities references for cross database connections.
Expression Atlas	Probes, gene names, UniProt accessions, experiment identifiers, assays, organ, up- or down-regulation, organism, assays' short description, and P-value.
ChEMBL	Small molecule registration identifier, assays and respective UniProt accessions to identify connections between protein and small molecules.
REACTOME	Pathway identifiers, upstream and downstream genes, and pathway hierarchy.
BioModels	Experimental assays used to link BioModels with Expression Atlas, gene names, and organism.
BioSamples	Sample title, brief summary, sample identifier, species, gene names, and sample group.

3.3.3 Web Portal

Our web portal provides users with a biologist-friendly interface. When a user inputs a single or multiple genes or a small molecule and selects certain databases, a graph is presented with data entities as nodes, with links between those nodes shown as edges. Using this graph, a user can intuitively investigate the connectivity map of genes or small molecules of interest in six different biological databases in the EBI-RDF platform. We used Cytoscape's web API to create graphs⁴⁵).

When a user enters a list of genes or small molecules, cMapper creates separate output graphs for each element in the list. In the second step, cMapper randomly takes two output graphs and merges them into a single graph using shared data entities between the two graphs. This step is repeated until all output graphs are merged into a single graph.

3.4 Results

The cMapper web portal provides an easy way to investigate the connectivity map of genes or small molecules in six independent databases in the EBI-RDF platform. The present size of cMapper database is 50 GB and it provides the following functionality to its users.

3.4.1 Browsing, Searching and Filtering

Input to cMapper is a set of genes or small molecules. Users can enter genes or small molecules directly or select them from the autocomplete list. The output of cMapper in response to a user query is a connectivity graph comprised of data entities connected

Table 3.2: Database filters

Database filter	Applicable databases	Function
Database filter	N.A	Allows users to select databases
Organism filter	All databases	Allows users to select species
Organ filter	Expression Atlas and BioSamples	Allows users to select organs of their interest (e.g., liver or lung)
Pathway filter	REACTOME	Allows users to select metabolic and signaling pathways

with inputted genes or small molecules. Data entities from different databases are differentiated by color.

Users can apply filters on the output graph. Table 3.2 summarizes the database filters and their functions in cMapper. Database filters help users to focus on the core connectivity map of their interest. For example, using the combination of database, organism, and organ filters a user may focus on changes in FGF19 gene expression in the liver of Homo sapiens. The Use case in the supplementary information provides a compact working example of a filter combination.

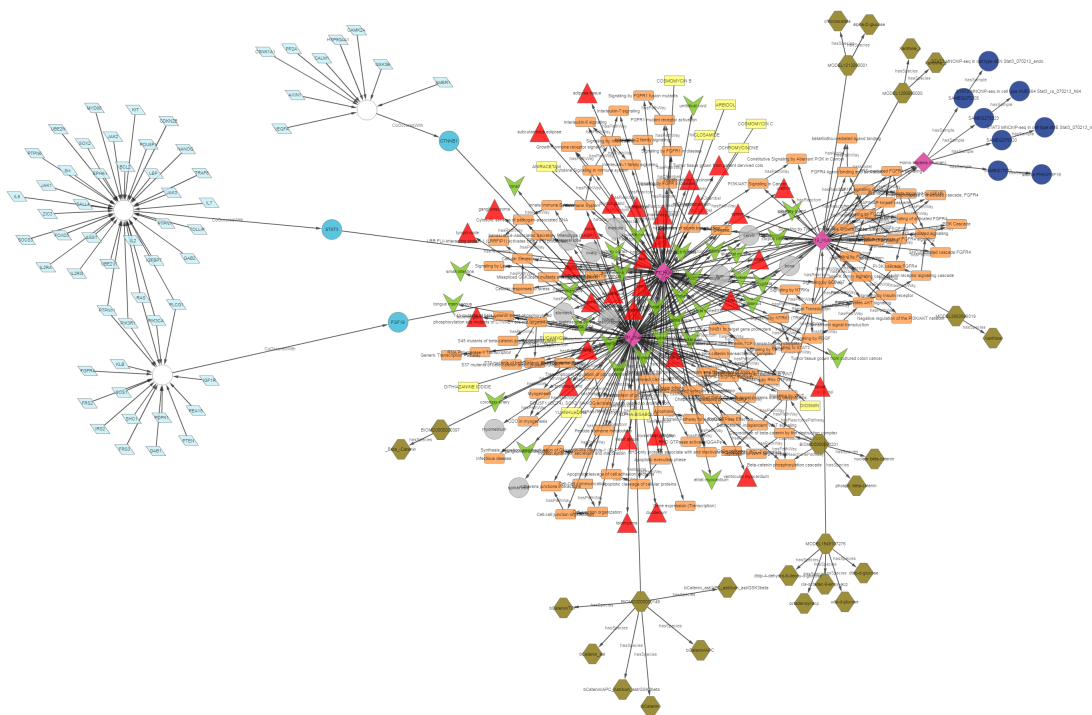


Figure 3.2: cMapper output for the user query FGF19, CTNNB1, and STAT3 with the shared connection filter disabled. The figure shows all data entities connected to FGF19, CTNNB1, and STAT3 in the EBI-RDF platform. Green, red, and silver nodes represent organs in which any of these genes are differentially expressed. Olive green nodes represent data entities from BioModels; orange nodes, those from REACTOME; yellow nodes, those from ChEMBL; and dark blue nodes, those from BioSamples.

Table 3.3: User Input query for Figure 3.3

Input Type	Input
Genes	CTNNB1; STAT3; FGF19
Databases Included	Associated Genes, UniProt, Expression Atlas, REACTOME, ChEMBL, BioModels, BioSamples
Databases Excluded	None
Organism Filter	Homo Sapiens
Organ Filter	All Organs
Pathway Filter	All Pathways
Graph Type	All Connections

3.4.2 Finding shared connections

Finding shared connections is one of the most useful functions of cMapper for biologists. In general, biologists search databases for hidden connections, such as a hidden direct or an indirect connection between two genes and between a gene expression change and a disease state.

Shared connection filter in cMapper enables users to investigate unknown or indirect connections between data entities of six independent databases in the EBI-RDF platform without possessing any programming skills. For example, when users input two genes, cMapper generates the connectivity map of these two genes in the EBI-RDF platform. Using the shared connection option, users can identify shared connections or pathways between these two genes. Using this functionality, researchers can identify unknown relationships between data entities in the EBI-RDF platform, thereby

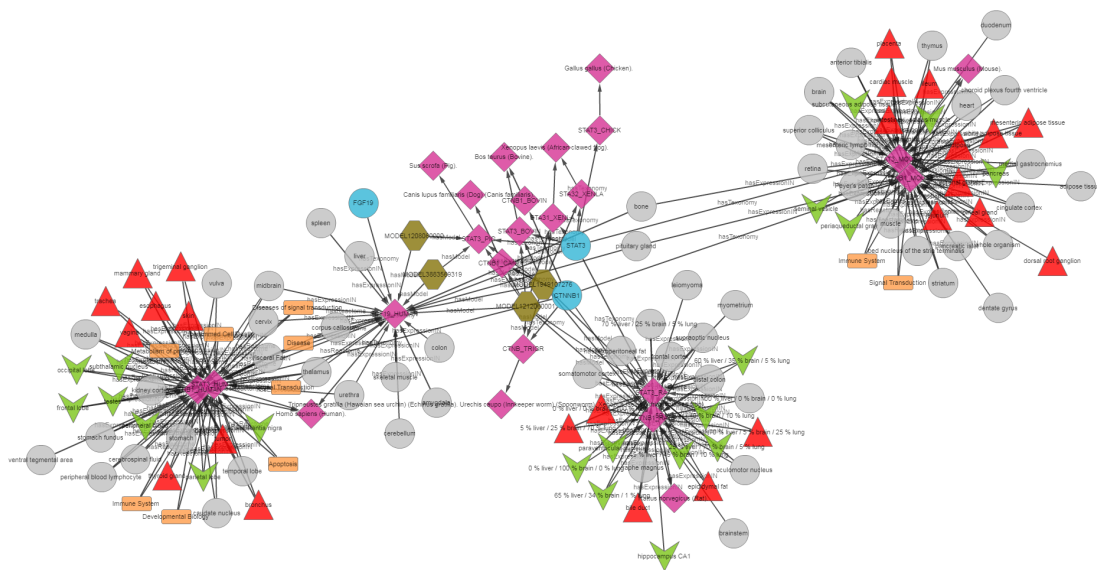


Figure 3.3: cMapper output for the user query FGF19, CTNNB1, and STAT3 with the filter for shared connections enabled. This connectivity map highlights shared data entities between inputted genes. Green and red nodes represent organs in which any two genes are co-expressed. Silver nodes represent organs in which any two genes are differentially expressed. Olive green nodes represent data entities from BioModels; orange nodes, those from REACTOME; yellow nodes, those from ChEMBL; and dark blue nodes, those from BioSamples.

proposing new hypotheses.

Figure 3.2 presents the output graph generated by cMapper as result of the user's query FGF19, CTNNB1, and STAT3 with an organism filter. Tables 3.3, 3.4, and 3.5 provide compact filter details applied for figures creation. The connectivity map consists of all data entities connected with any of the input genes in Homo sapiens (humans). Figure 3.3 presents the graph showing connectivity map for the same genes but with the organism filter disabled and the shared connections filter enabled. The connectivity map in Figure 3.3 consists of data entities that are shared between more

Table 3.4: User Input query for Figure 3.3

Input Type	Input
Genes	CTNNB1; STAT3; FGF19
Databases Included	UniProt, Expression Atlas, REACTOME, ChEMBL, BioModels, BioSamples
Databases Excluded	Associated Genes
Organism Filter	All Organisms
Organ Filter	All Organs
Pathway Filter	All Pathways
Graph Type	Shared Connections

than two genes in all organisms. Using the shared connection filter, users can highlight potential connections between genes of interest and can potentially put forward new hypotheses.

Figure 3.4 displays the output of the previous query including an additional pathway filter. The connectivity map in Figure 3.4 consists of data entities (genes) connected with the metabolic pathways of multiple inputted genes.

Using the filter for shared data entities, users can select a minimum number of genes or small molecules as a threshold for shared connections. For example, a user can filter data entities shared by two or three genes using the threshold. The default threshold value is two.

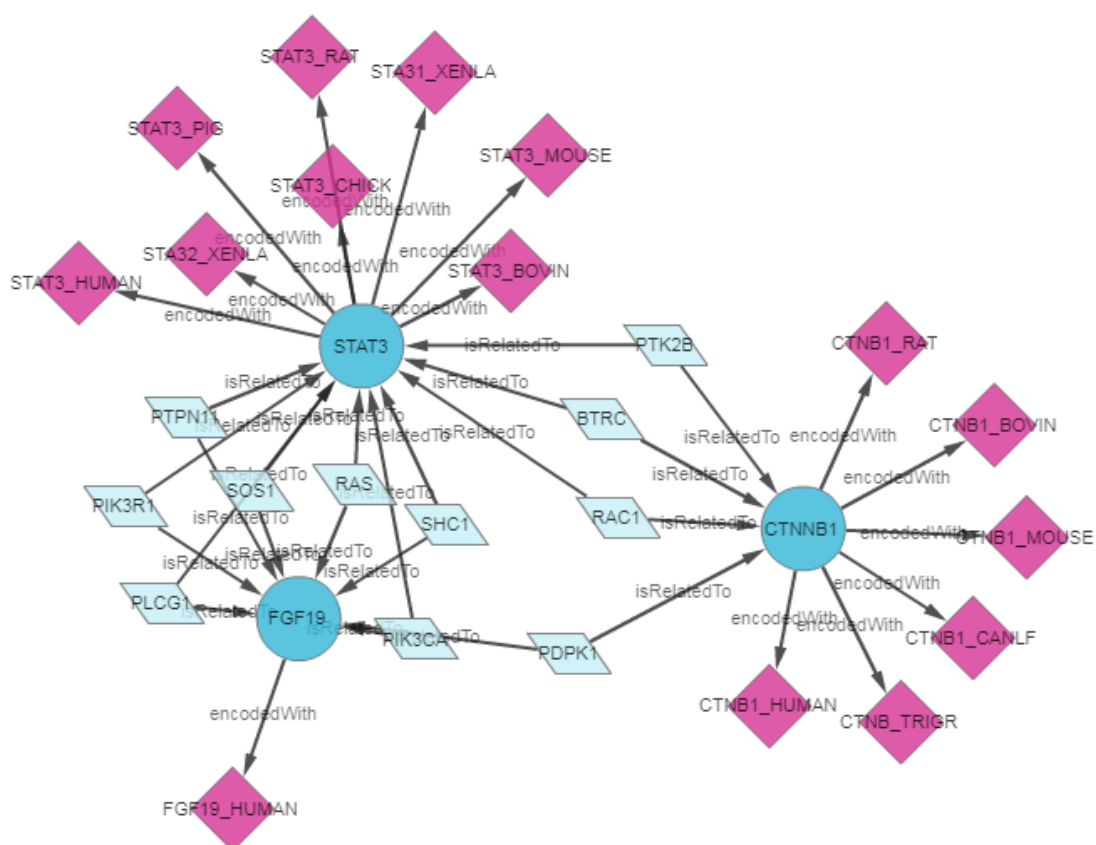


Figure 3.4: cMapper output for the user query FGF19, CTNNB1, and STAT3 with the filter for shared connections and the associated genes database filter enabled. In this map, connected data entities represent genes connected with multiple inputted genes through metabolic pathways

3.4.3 Downloading GraphML

Using the web interface of cMapper, users can navigate the connectivity map of selected genes or small molecules. Users can also save the connectivity map as an image or PDF document. The connectivity map can be downloaded in GraphML, JSON or CSV formats for further analysis or for generating high quality graphics. GraphML, JSON, and CSV files from cMapper contain information about nodes (representing data entities) and edges (representing connections between data entities). Each node

Table 3.5: User Input query for Figure 3.4

Input Type	Input
Genes	CTNNB1; STAT3; FGF19
Databases Included	Associated Genes
Databases Excluded	UniProt, Expression Atlas, REACTOME, ChEMBL, BioModels, BioSamples
Organism Filter	All Organisms
Organ Filter	All Organs
Pathway Filter	Metabolic Pathways
Graph Type	Shared Connections

contains information regarding data entities in the form of its label, database identifier, and URL. Labels are used to store display information and database identifiers for color-coding of nodes. Users can download GraphML or JSON files and then use them with any graph making software, e.g. Cytoscape.

3.5 cMapper Updater

cMapper updater is a Java Program with a set of SQL queries that is used to keep cMapper database up-to-date. Because cMapper database has been derived from databases in EBI-RDF platform, It was required to update cMapper database to keep it synchronize with its parent databases. As explained in section 2.1 and 2.2 of the manuscript, we have developed six different modules to create connectivity and data tables from six independent EBI-RDF databases because of schema diversity. This exercise of developing separate module also helped us in development of cMapper updater. Currently

for each EBI database, we manually check whether it has been updated. Once we know that a database has been updated, we download the new dump of respective database from the EBI-RDF website and run the corresponding module of cMapper updater. Each module of cMapper updater completes its process in following three steps. First of all it creates a new temporary database using the pre-defined schema. Temporary cMapper database only contains tables associated with the updated databases in EBI-RDF platform. For example when REACTOME database in EBI-RDF platform is updated, tables that store REACTOME data entities and their relationship with data entities of other databases are only created in the temporary database. cMapper updater creates new connectivity and data tables from updates dumps instead updating existing tables because finding updated entities in data dumps is more time intensive task then loading all entities into new tables. Second, it copies the tables from current cMapper database to backup database. We only preserve two most recent copies of each data and connectivity table because of the space limitation. Backup copies are used to restore cMapper database to its previous version when updater fails to complete update process because of any error. Finally, it deletes the respective tables from current database and move tables from temporary database to current cMapper database. In short cMapper updater (1) creates new connectivity tables and data tables, (2) creates backup of existing tables, and (3) replaces respective current cMapper database tables with new tables. Figures explaining cMapper updater workflow are available in the supplementary.

3.6 Case Study

In this section we explain how cMapper can be used to generate new hypothesis. We have been working on liver cancer genomics. We found a new potential role of FGF19 amplification in hepatocarcinogenesis (Ahn et al., Hepatology 2014). These days, PDL1 (CD274) is a hot issue because PDL1 is the target of immune checkpoint inhibitors, a new promising category of anti-cancer drug because tumor cells can express PDL1 to evade immunosurveillance.

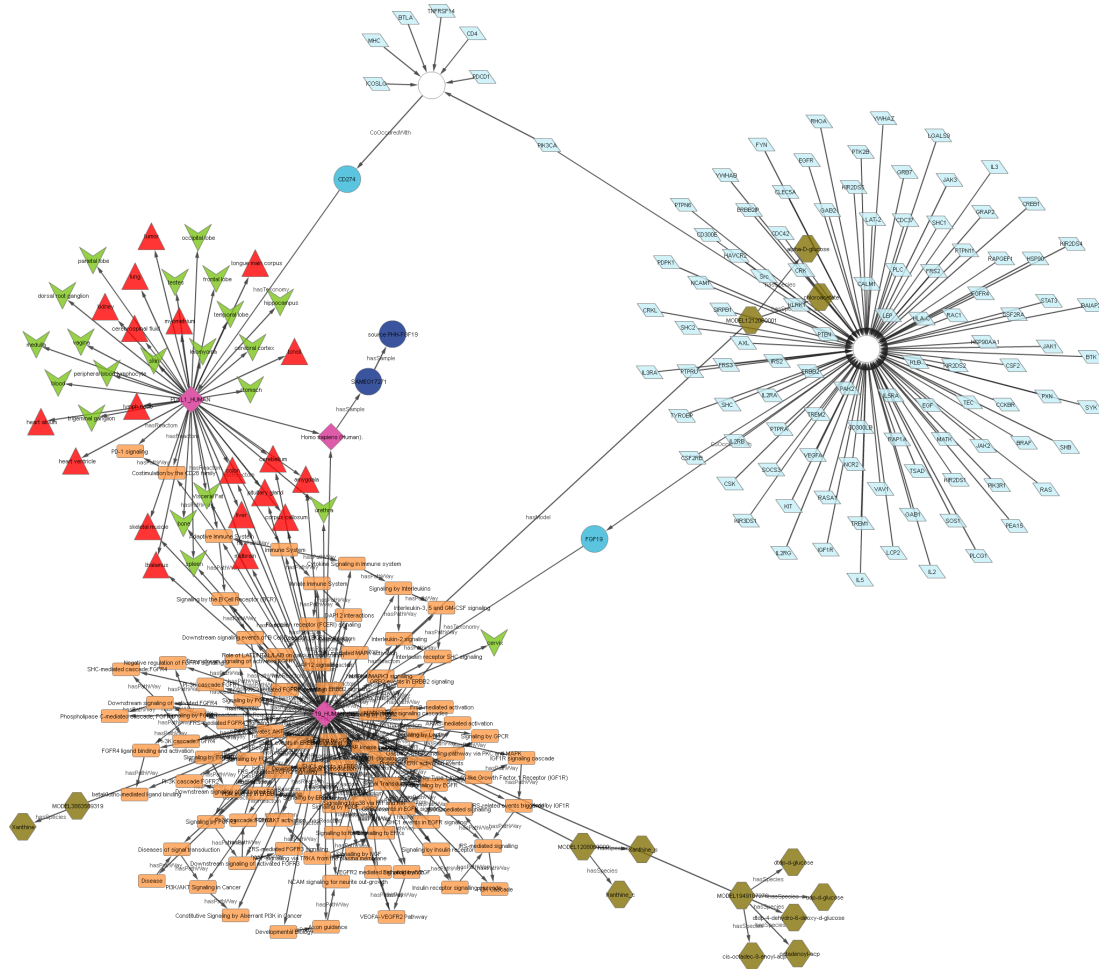


Figure 3.5: cMapper showing data entities connected with FGF19, and CD274, without applying any filter

We used cMapper to ask a simple question that FGF19 is involved in hepatocarcinogenesis, but can it have anything to do with PDL1? To answer this question we created graph of all data entities connected with FGF19 and PDL1(CD274) Figure 3.5 shows the network created by cMapper and shows general landscape of all data points connected with CD274 or FGF19.

The network showed in figure 3.5 is complex therefore we decided to focus on pathways. From the database menu, we select Pathway filters. The result of the query has been shown in figure 3.6

However this network shown in figure 3.6 is also complex and needs to be prune further therefore, we selected pathway database with shared connection filter enabled with the objective that cMapper will return the pathways that are common between

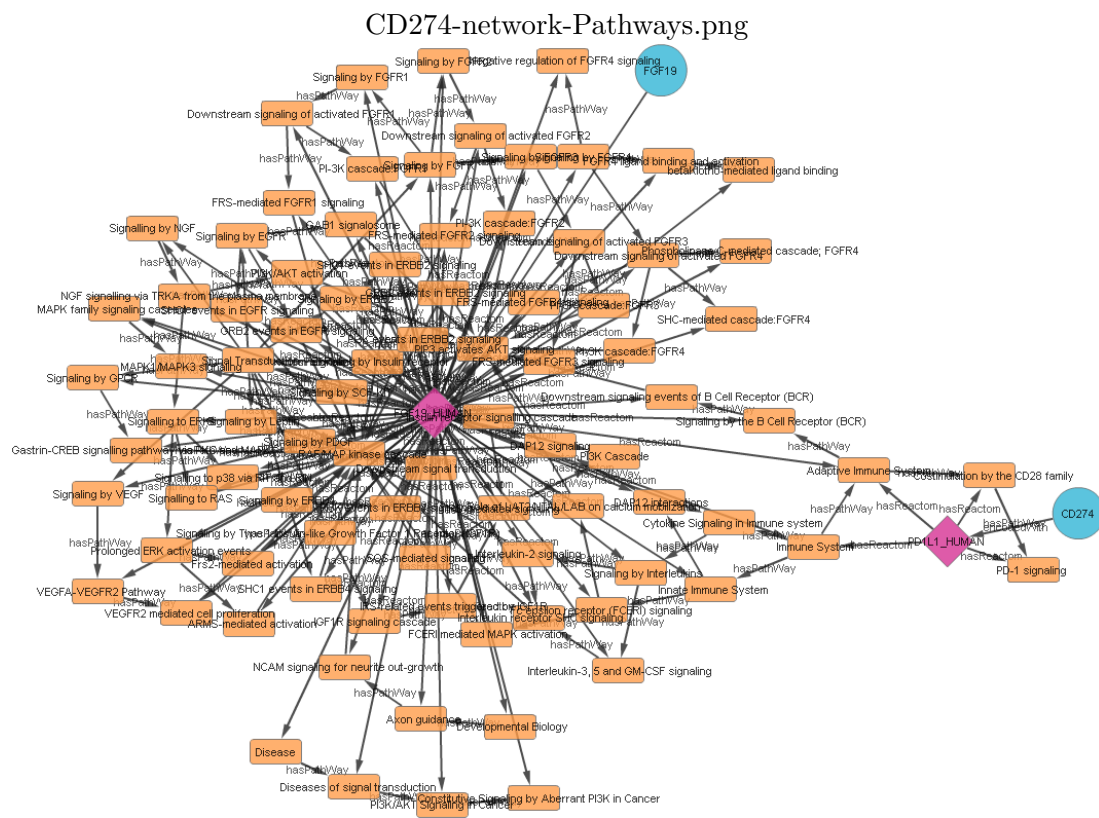


Figure 3.6: cMapper showing all FGF19, and CD274 pathways

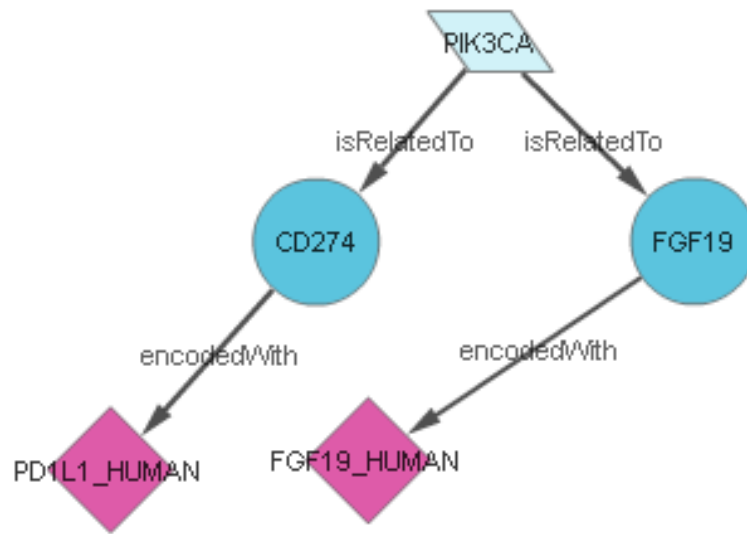


Figure 3.7: cMapper output showing common associated genes between PDL1 and FGF19 using shared connection filter.

PDL1(CD274) and FGF19.

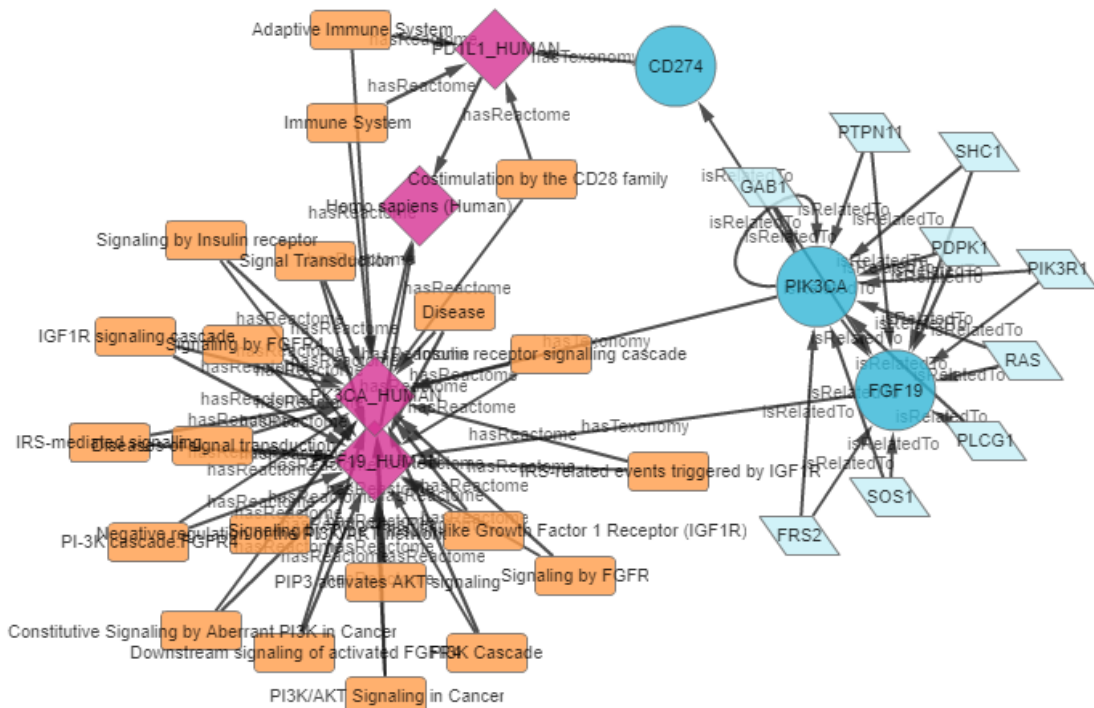


Figure 3.8: cMapper showing common pathways and associated genes between PDL1 and FGF19 and PIK3CA.

We were interested in knowing if FGF19 and PDL1 (CD274) has any associated pathways in common. Figure 3.6 gives the positive answer by finding two pathways that are associated with both PDL1 and FGF19. From figure we knew that both genes have some connection at pathway level and both genes are associated with Immune Systems and Adaptive Immune Systems. 4. Once we knew that both genes have common pathways, we wanted to know whether the two genes have any associated genes in common. So, we selected associated genes database and selected "shared connections".

We found that PIK3CA is related to both FGF19 and PDL1(CD274). Through some more literature search, we have come to a hypothesis that in our liver cancer model system, activation of FGF19-FGFR4 pathway may lead to PDL1 overexpression through PIK3CA pathway. In the next step we performed another search to find common pathways and associated genes between FGF19, CD274, and PIK3CA. From the result that is shown in figure 3.8 we noticed that FGF19 and PIK3CA has a lot of common pathways and associated genes however CD274 is connected with PIK3CA in only one pathway "*Costimulation by the CD28 family*" which leads towards a prospective hypothesis of investigating role of PIK3CA gene in CD28 Costimulation pathway.

3.7 Discussion

Semantic web technologies, such as RDF and SPARQL, allow for interoperability among heterogeneous databases. The EBI-RDF platform allows users to query six independent databases in an integrated fashion. In other words, users can query six independent databases to identify connections between data entities given an existing understanding

of RDF and SPARQL. Using the EBI-RDF platform, researchers can construct a query similar to that shown in Figure 4, which aims to find gene expression of CTNNB1 in all organs by connecting searches from the UniProt and Expression Atlas databases. Though this query is relatively simple, it still requires extensive knowledge of each database schema and the SPARQL query language. Schweiger and colleagues⁶⁷⁾ developed SPARGRAPH, a graphical SPARQL query builder for the EBI-RDF platform. It helps users to build a SPARQL query; however, users still need command over RDF and SPARQL.

cMapper enables data integration and interoperability at the EBI-RDF platform, allowing users to investigate the connectivity map of genes or small molecules of interest without in-depth knowledge of RDF and SPARQL.

In order to provide users with updated connections, we have developed a computational pipeline to update cMapper database. Currently, our plan is to update cMapper every three months, which seems sufficient to keep up with the update schedule of the EBI-RDF platform (refer to Table S4 for the latest update schedule of the EBI-RDF platform).

Chapter 4

Integrated Pharmacogenomic Platform of Human Cancer Cell Lines and Tissues

4.1 Abstract

Motivation: The exponential increase in multilayered data, including omics, pathways, chemicals, and experimental models, requires innovative strategies to identify new linkages between drug response information and omics features. Despite the availability of databases such as the Cancer Cell Line Encyclopedia (CCLE), the Cancer Therapeutics Response Portal (CTRP), and The Cancer Genome Atlas (TCGA), it is still challenging for biologists to explore the relationship between drug response and underlying genomic features due to the heterogeneity of the data. In light of this, the Integrated Pharmacogenomic Database of Cancer Cell Lines and Tissues (IPCT) has

been developed as a user-friendly way to identify new linkages between drug responses and genomic features, as these findings can lead not only to new biological discoveries but also to new clinical trials.

Results: The IPCT allows biologists to compare the genomic features of sensitive cell lines or small molecules with the genomic features of tumor tissues by integrating the CTRP and CCLE databases with the REACTOME, cBioPortal, and Expression Atlas databases. The input consists of a list of small molecules, cell lines, or genes, and the output is a graph containing data entities connected with the queried input. Users can apply filters to the databases, pathways, and genes as well as select computed sensitivity values and mutation frequency scores to generate a relevant graph. Different objects are differentiated based on the background color of the nodes. Moreover, when multiple small molecules, cell lines, or genes are input, users can see their shared connections to explore the data entities common between them. Finally, users can view the resulting graphs in the online interface or download them in multiple image or graph formats.

Availability and Implementation: The IPCT is available as a web application with an integrated MySQL database. The web application was developed using Java and deployed on the Tomcat server. The user interface was developed using HTML5, JQuery 3.1.0 , and the Cytoscape Graph API 1.0.4. The IPCT can be accessed at <http://ipct.ewostech.net>. The source code is available at <https://github.com/muhammadshoaib/ipct>.

4.2 Introduction

Advancements in pharmacogenomics through comprehensive next-generation sequencing studies have paved the way for developing effective therapeutics against cancer. The omics data of cancer cell lines and cancer tissues are now readily used for categorizing genomic diversity and identifying anti-cancer drug responses²²⁾. However, in the era of big data, biologists face new challenges in dealing with the large amount of segregated data available in different cancer genomic repositories^{71,11)}

In the past decade, data scientists have made efforts to facilitate biologists by developing numerous biological databases, which to some extent have helped biologists to analyze the underlying genetic mechanisms in cancer. NCI-60, the first cancer cell line database, remained a unique resource of in vitro drug discovery for many years⁸¹⁾. Recently, large pharmacogenomic databases including the Cancer Cell Lines Encyclopedia (CCLE), Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Therapeutics Response Portal (CTRP) have also emerged. The CCLE database provides genomic and transcriptomic information for 947 human cancer cell lines with drug response data of 24 compounds²⁾. GDSC and CTRP provide drug response information for more than 1000 cancer cell lines against 260 and 460 compounds, respectively³⁾. In addition to cell line data, omics data of thousands of cancer patients were also generated by The Cancer Genome Atlas (TCGA) and European Molecular Biology Laboratory (EMBL)⁷⁸⁾.

Unfortunately, the volume and heterogeneity of the data has prevented biologists from making effective use of these databases¹⁹⁾. Therefore, an efficient and biologist-friendly integration of these omics and pharmacogenomics databases is needed. This integration would help biologists generate accurate and practical hypotheses for iden-

tifying anti-cancer drug responses. The prime objective of this study was to provide a uniquely user-friendly platform for cancer biologists that they can use to investigate interlinked pharmacogenomics and cancer genomics data.

In this study, we have developed the Integrated Pharmacogenomics Platform of Cancer Cell Lines and Tissues (IPCT), which integrates major drug response information from the CTRP with omics data from the CCLE, cBioPortal¹⁹⁾, REACTOME¹⁴⁾, and Expression Atlas⁶¹⁾ databases. The IPCT is a biologist-friendly platform with numerous novel features, highlighting:

- the genomic features sensitive to specific drugs;
- the percentage of affected cancer patients sensitive to a drug;
- the pathways associated with the drug response;
- cancer cell lines that are true representatives of cancer tissues;
- user-friendly single-click access to multiple datasets, which facilitates the generation of new and practical hypotheses.

4.3 Materials and Methods

The CTRP portal contains quantitatively measured sensitivity for 461 small molecules in 860 deeply characterized cancer cell lines. Our tool (1) integrates the CTRP database with external biological databases and (2) allows biologists to query CTRP data in a graphical and integrated fashion. Biologists can start querying by entering a list of cell lines or small molecules and utilize the context of results of their search to generate new hypotheses.

The IPCT has been developed in three different steps: (1) construction of the database, (2) development of the database update pipeline and (3) web application. The database is an essential component of the IPCT, which stores all data points and connections among those data points in the CTRP, CCLE, cBioPortal, REACTOME Pathways and Expression Atlas. The update pipeline is a script written in Python that is used to update the database in real time. The web application is a GUI-based application that will be used by the end users to explore data points and connections.

The IPCT is a biological database that integrates data about cancer cell lines, small molecules, human pathways, experimental results, and cancer somatic mutations. Figure 4.1 and 4.2 show architecture of IPCT database and demonstrates how multiple databases have been integrated in IPCT database. We collected the cell line data from the CCLE dataset, the small molecule features from the CTRP dataset, the pathway data from REACTOME, the expression data from the Expression Atlas, the list of cancer genes from cancer genes census¹⁷⁾ and OncoKB⁹⁾, and the genomic features of cancer studies from cBioPortal. Our objective was to create an integrated database by connecting the data points in the above databases.

In the second step, we added genes to our network. To do this, a list of genes and their relationships to cell lines was required. We extracted genetic metadata from the NCBI website and connected the genes and cell lines based on genomic changes, which were present in the CCLE dataset in the form of mutations, copy number alterations, and gene expression. We extracted these for each cell line from the CCLE dataset and used this information to construct a small molecule–cell line–gene graph. A gene was included in the small molecule–cell line–gene network if it had mutations, copy number amplification, copy number deletion, high expression, or low expression in at least 10%

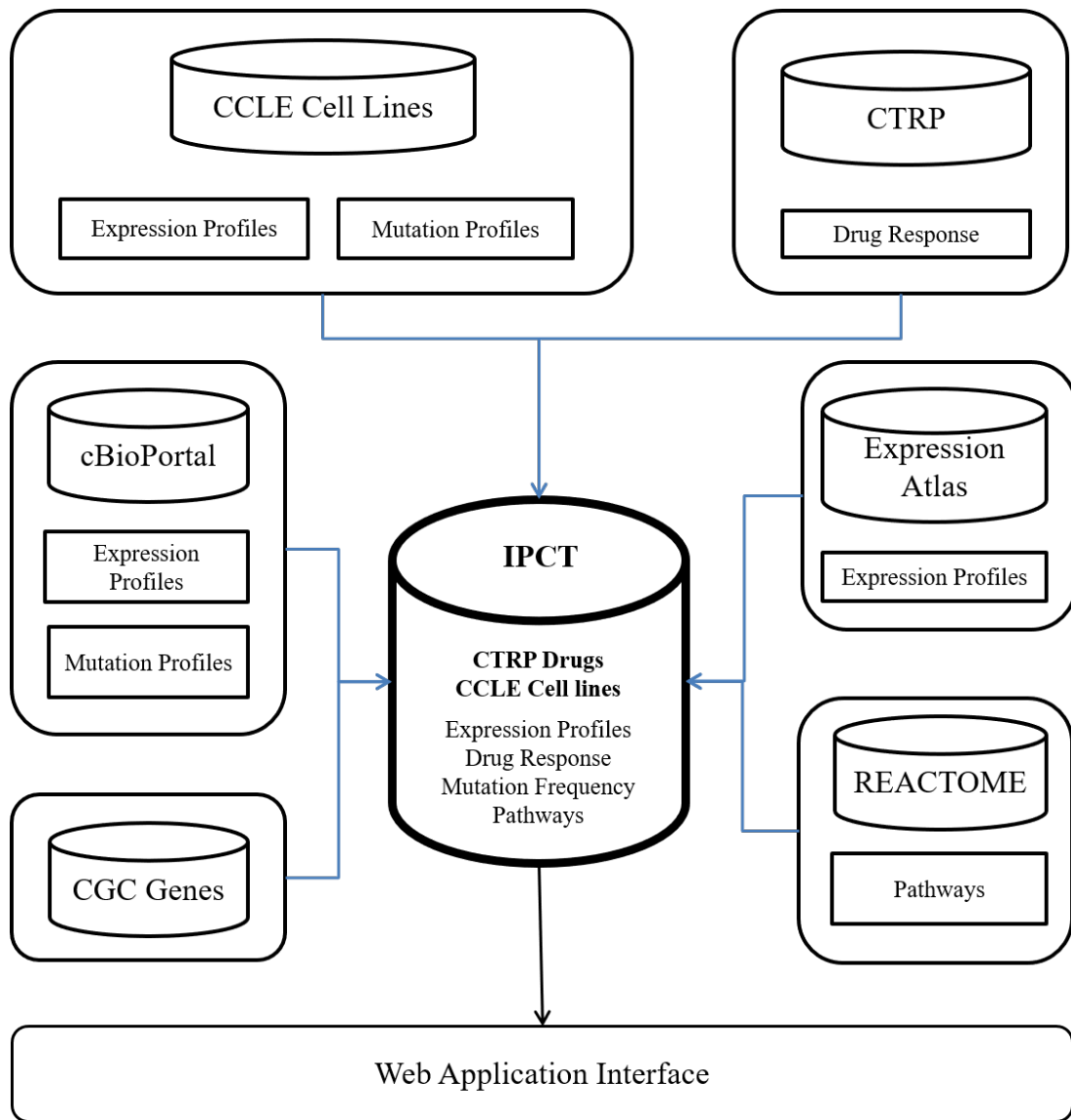


Figure 4.1: Overall Architecture of IPCT Platform. Figure demonstrates the system level flow of information. It shows which data point is connected from which database.

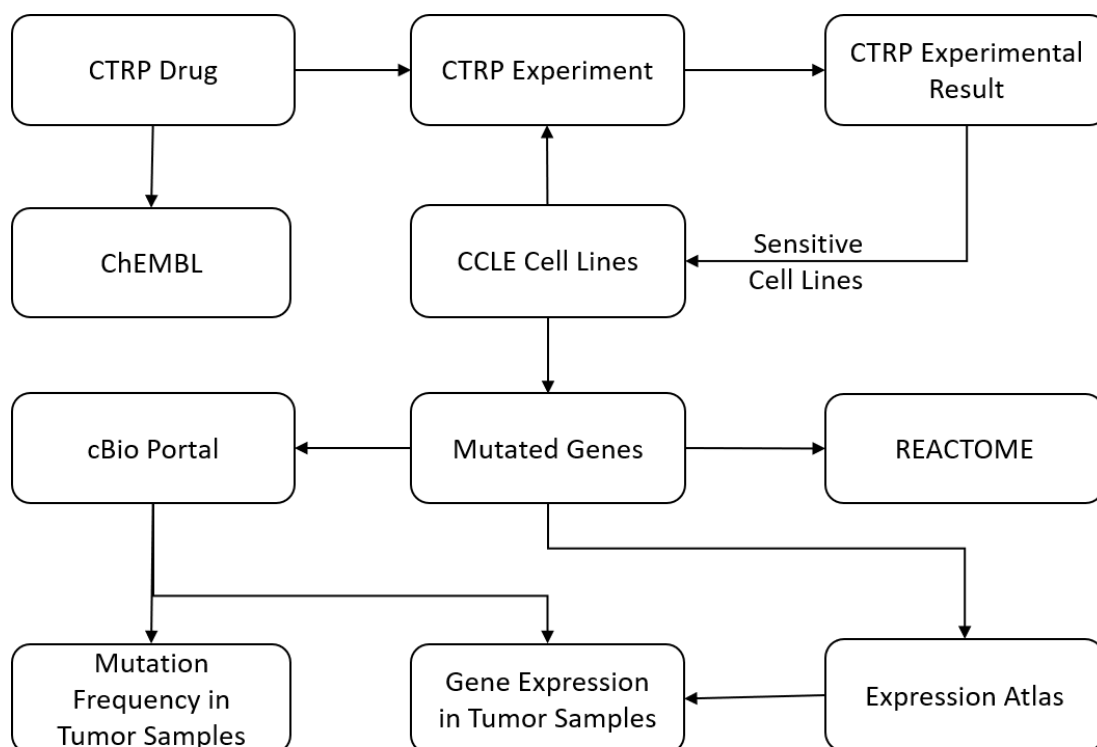


Figure 4.2: Entities connected in the IPCT Database. Figure explains how different objects can be linked with each other.

of the cell lines sensitive to an input small molecule. The IPCT, by default, connects only genes with genomic aberration in 20% of the cell lines sensitive to an input small molecule. However, users can relax or tighten these criteria as needed. Having already constructed a small molecule–cell line network, in this step we only had to connect the cell lines with the genes. To do this, we connected cell lines with the genes that had mutations or copy number alterations in the given cell lines. This process was repeated for all cell lines in the CCLE, which resulted in a cell line–gene network with genes and cell lines as nodes and mutations or copy number alterations as edges. The cell line–gene graph was then merged with the small molecule–cell line graph, which resulted in a small molecule–cell line–gene network. After this step, we could identify genes with mutations or copy number alterations in the cell lines that are sensitive to

a given set of small molecules.

Once we identified the mutated genes, in the third step, we added pathways of the mutated genes to our small-molecule-cell line-gene network. We collected pathway data from the REACTOME database. REACTOME pathways were connected using Entrez GeneIDs that were present in both databases.

The next step was to identify if the mutated genes had been reported as up-regulated or down-regulated in previous experiments. The Expression Atlas contains differential expression data from approximately 2500 experiments performed in different experimental conditions. However, the Expression Atlas uses Ensembl IDs instead of gene names or Entrez GeneIDs in its analyzed files. In the first step, we filtered only those experiments that were related to cancer, loaded them into the database, and removed insignificant records with p-value ≥ 0.05 and log fold change ≤ -1 and ≥ 1 . Records with log fold change ≥ 1 or ≤ -1 and p-value ≤ 0.05 were used for further processing. Next, we connected all Ensembl IDs with their Entrez GeneIDs using the R package "org.Hs.eg.db". We used this database to construct a gene-experiment network with genes and experiments as nodes and up-regulation or down-regulation as edges. This graph was then merged with the small molecule-cell line-gene network constructed in the previous step.

After construction of the small molecule-cell line-gene pathway graph, our next task was to identify if the mutated genes had any potential relationship with any cancer type in published cancer studies. To do this, we extracted data from cBioPortal. For each gene, we computed what percentage of samples were mutated, altered, up-regulated, and down-regulated in each study. In this way, we identified mutation and alteration frequencies for 30,000 genes in 151 cancer studies and 33 cancer types. Data from

cBioPortal were not used in network construction but are available as a separate entity for further investigation.

The IPCT can be accessed via the web application, which allows users to explore the connections between the data points of five biological databases in an integrated graphical fashion. When a user enters a small molecule, cell line, or gene, a graph is displayed with the data points as nodes and the relationships between the data points as edges. Using this graph, the user can intuitively investigate the connectivity of the given small molecules, cell lines, and genes. When a user enters multiple cell lines, small molecules, or genes, the IPCT first independently constructs a graph for each element in the list. Next, it takes two random graphs from among those and merges them using the common data points. This step is repeated until all the graphs are merged into one graph, which is ultimately displayed to the user.

4.4 Results

The IPCT comprises two major components: (1) the IPCT database and (2) the IPCT web portal. The IPCT web portal provides an easy way to investigate the connections between the data points available in the CTRP, CCLE, Expression Atlas, REACTOME, and cBioPortal databases in an integrated fashion. The IPCT database currently contains 860 cell lines, 481 small molecules, 2,500 differential expression studies, 2000 human pathways, and 151 cancer studies. Moreover, the IPCT contains 8,214,573 unique connections between the different data points (Table 4.2). The overall database size is 20 GB. The distinctive functionality and features of the IPCT are as follows:

1. Users can input up to ten cell lines, small molecules, or genes to find potential

connectivity with other data points.

2. Users can filter small molecules and cell lines sensitive to each other according to a minimum sensitivity score.
3. Users can apply a filter on genes if they want to view only cancer genes, exclude commonly mutated genes, or view all genes.
4. Users can apply filters if they want to see only mutated, copy number altered, or high- or low-expressed genes.
5. Users can check the mutation frequencies and differential expression frequencies in different cancer studies.
6. Users can highlight genes of their interest by applying a gene filter to the network.
7. Users can select if they want to show all connections or only shared connections when multiple cell lines, small molecules, or genes are entered.
8. Users can view the output in the web browser as a graph or table. Alternatively, users can download the graph and view it with Cytoscape version 1.0.4 or graph viewing tools that show JSON and CSV files.
9. Users can save the graphs in JSON, PNG, or PDF formats and table in CSV format.

4.4.1 Data Exploration

Users can start exploring the IPCT by entering small molecules, cell lines, or genes. If users enter a list of cell lines, the IPCT outputs graphs with small molecules that are sensitive to the queried cell lines and genes that are mutated or altered in the given

cell lines. If users enter a small molecule, the IPCT outputs a graph containing the cell lines sensitive to the given small molecule and genes mutated in the sensitive cell lines. If users enter genes, the IPCT outputs a graph of cell lines with mutations or copy number alterations in the given genes and the small molecules sensitive to those cell lines. Users can then expand their search by expanding the graph to include data points from the Expression Atlas or REACTOME. User can apply filters as explained in Table 4.1 and reduce number of entities in graph. Figure 4.3 illustrates the output generated by the IPCT for lapatinib with the shared pathway filter. By default, the IPCT shows the pathways associated with more than 20% of genes connected with the input drug, but users can modify this option to show all pathways if they want to see the pathways of connected genes. Supplementary Figure 4.4 shows the result of same query with all pathways. The IPCT also allows users to apply different filters to define the context of their search.

4.4.2 Comparison Between Cell Lines and Real Tissues

Since all cancer cell lines do not have equal values to the tumor models, comparison between genetic profiles of cell lines and real tumors is of importance. For example, when a mutation is found in a cell line, a first question can be if the specific mutation has also been reported in any of the cancer studies or not, and second, if the given gene has any reported differential expression or not. The IPCT, by integrating data from cBioPortal and Expression Atlas, provides answers to both question. When a user clicks on a mutated gene's node, he can explore cancer studies in which the given node is up-regulated or down-regulated and observe the mutation or alteration percentage in all cBioProtal cancer studies. With this exploration, a user can identify the cancer

Table 4.1: Database filters that can be applied to searches in the IPCT

Database filter	Applicable databases	Function
Compound Sensitivity	Small molecules	Allow users to set thresholds for small molecule sensitivity
Mutation Frequency	Genes	Allow users to set mutation frequency
Gene Filter	Genes	Allow user to select if he or she wants to see only cancer genes, exclude commonly mutated genes or see all genes
Pathway filter	REACTOME	Allows users to select metabolic and signaling pathways
Genomic aberration	Genes	Allow users to filter gene relationships based on mutations, copy number alterations, and gene expression

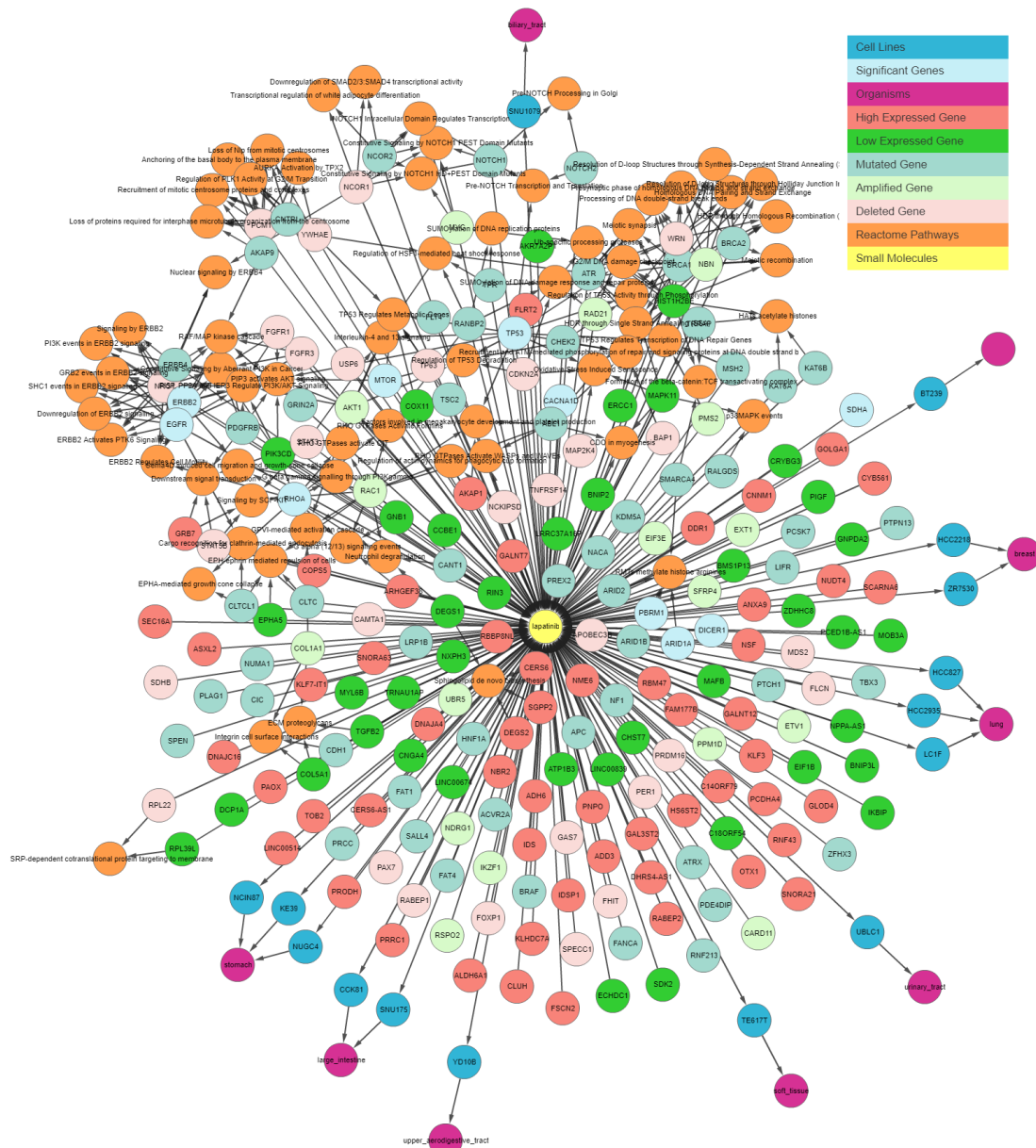


Figure 4.3: IPCT output for small molecule user query lapatinib with shared pathways. The graph shows all data points connected with lapatinib. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib; sky-blue nodes represent significant genes (those with multiple genomic aberrations); green and red nodes represent genes that are up-regulated and down-regulated in the sensitive cell lines, respectively; light green and light red represent the amplified and deleted genes in the sensitive cell lines, respectively; white nodes represent mutated genes; and orange nodes represent the REACTOME pathways of mutated genes.

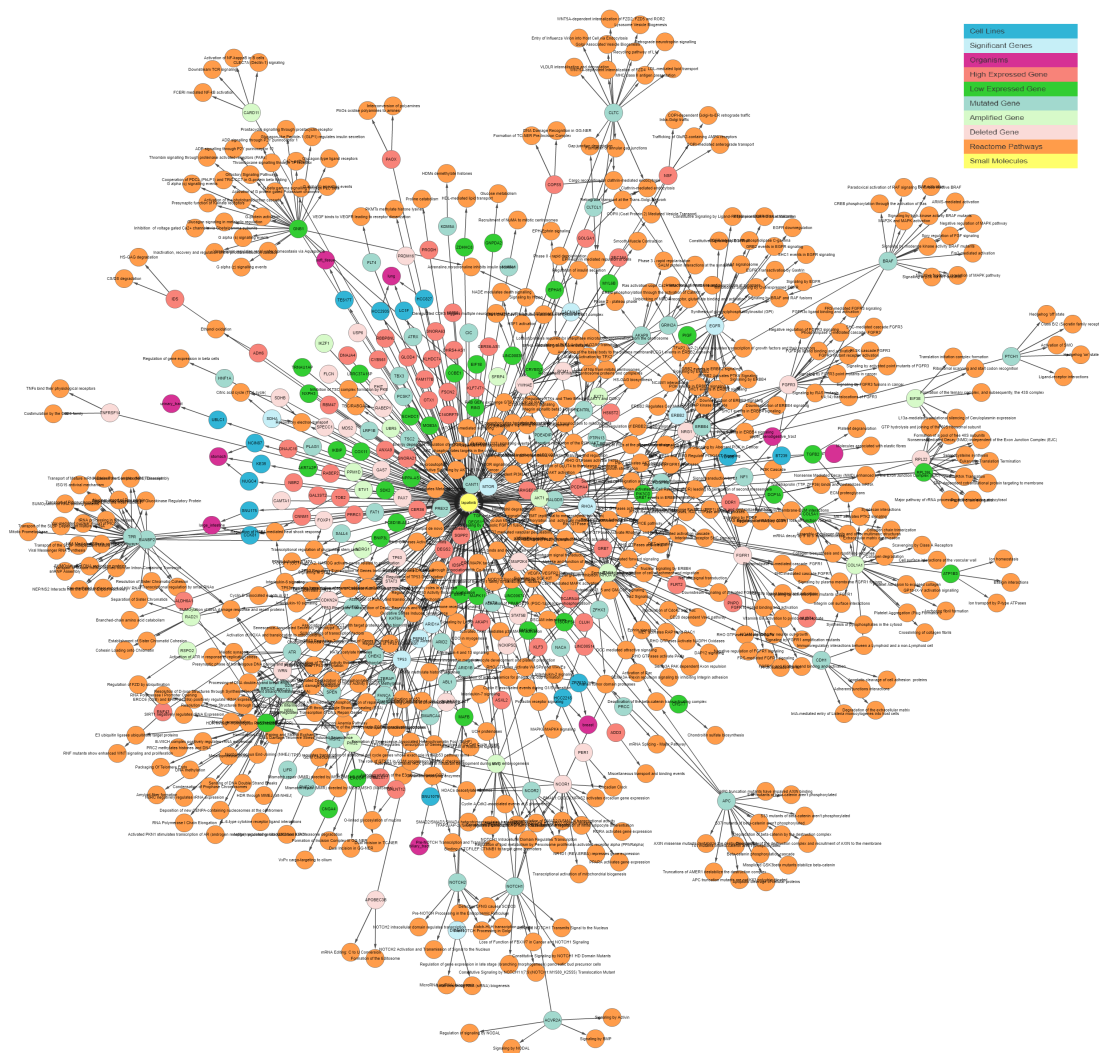


Figure 4.4: IPCT output for small molecule user query lapatinib with all pathways. The graph shows all data points connected with lapatinib. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib; sky-blue nodes represent significant genes (those with multiple genomic aberrations); green and red nodes represent genes that are up-regulated and down-regulated in the sensitive cell lines, respectively; light green and light red represent the amplified and deleted genes in the sensitive cell lines, respectively; white nodes represent mutated genes; and orange nodes represent the REACTOME pathways of mutated genes.

Table 4.2: User Input query for Figure 4.3

Input Type	Input
Small Molecule	LAPATINIB
Databases Included	Associated Genes, Cell Lines, Expression Atlas, REACTOME
Databases Excluded	None
Genes Filter	Cancer Genes Only
Drug Sensitivity	-1.50
Mutation Frequency	20%
Pathway Filter	Shared Pathways for figure 4.3 and All Pathways for figure 4.4
Graph Type	All Connections

type in which the selected gene has up-regulation and in which the selected gene has down-regulation. For example, in the previously illustrated (Figure 4.8) query, by investigating LAPATINIB, SORAFENIB, GEFITINIB and SUNITINIB together, we identified that FAT4 has mutations in 50% of cell lines sensitive to SORAFENIB, 24% of cell lines sensitive to GEFITINIB, 30% of cell lines sensitive to SUNITINIB and 27% of cell lines sensitive to LAPATINIB. A user can further investigate its frequency in real tumors. Figure 4.6 illustrates the results for mutations and the differential expression frequency of FAT4 in different cancer studies.

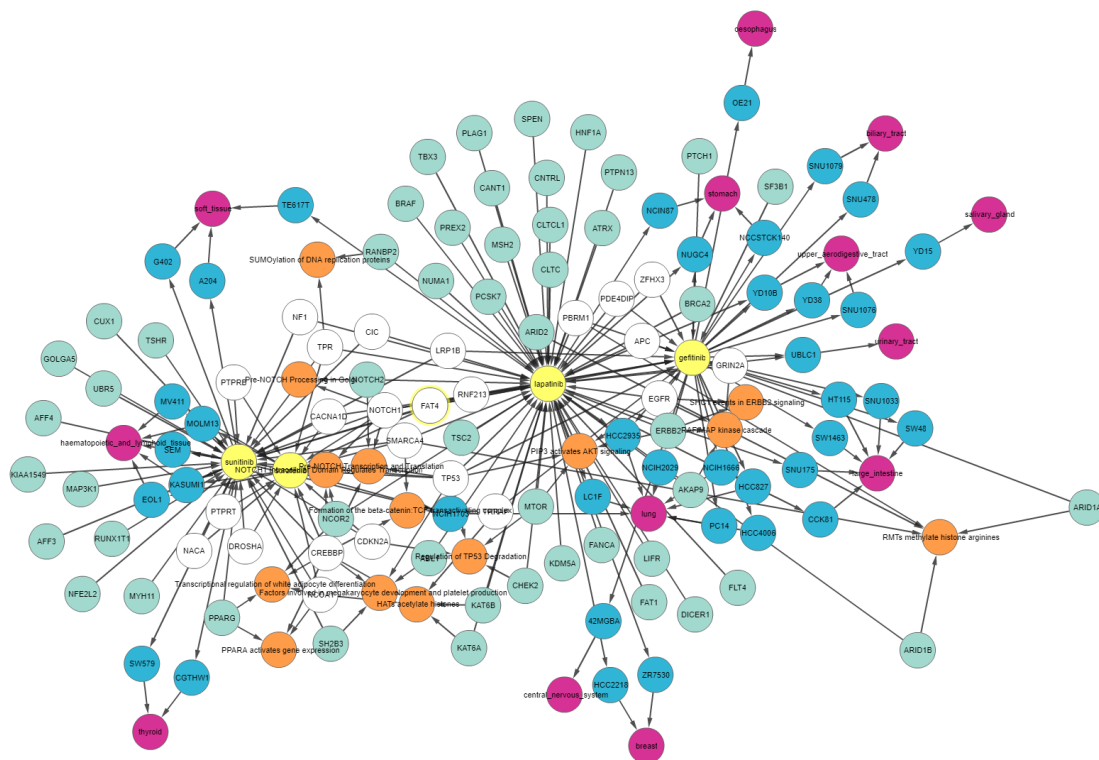


Figure 4.5: IPCT output for small molecule user query Lapatinib, Sorafenib, Gefitinib and Sunitinib to identify commonly mutated genes in cell lines sensitive to input small molecules. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to inputted small molecules; white nodes represent commonly mutated genes in multiple cell lines and sea-green nodes represent genes mutated in single cell line.

FAT4 MUTATIONS IN CBIOPORTAL

Gene Symbol	Cancer Type	Cancer Study	Mutations	Alterations
FAT4	Desmoplastic Melanoma	desm_broad_2015_sequenced	50%	50%
FAT4	Cutaneous Melanoma	skcm_broad_sequenced	34%	34%
FAT4	Primary CNS Lymphoma	pcnsj_mayo_2015_sequenced	30%	30%
FAT4	Cutaneous Melanoma	skcm_broad_dfarber_sequenced	28%	28%
FAT4	Mixed Cancer Types	cellline_nci60_cnaseq	25%	25%
FAT4	Colorectal Adenocarcinoma	coadread_genentech_sequenced	24%	24%
FAT4	Diffuse Large B-Cell Lymphoma	dlbc_broad_2012_sequenced	22%	22%
FAT4	Stomach Adenocarcinoma	stad_tcga_cnaseq	21%	22%
FAT4	Diffuse Large B-Cell Lymphoma	dlbc_tcga_cnaseq	21%	21%
FAT4	Small Cell Lung Cancer	scic_cicgp_sequenced	21%	21%

(a) FAT4 Mutations

FAT4 EXPRESSION IN CBIOPORTAL

Gene Symbol	Cancer Type	Cancer Study	Measurement Type	Up Regulated	Down Regulated
FAT4	Thymoma	thym_tcga	rna seq v2 mma median Zscores	8%	0%
FAT4	Head and Neck Squamous Cell Carcinoma	hnscc_tcga_pub	rna seq v2 mma median Zscores	10%	0%
FAT4	Colorectal Adenocarcinoma	coadread_tcga	rna seq v2 mma median Zscores	12%	0%
FAT4	Lung Squamous Cell Carcinoma	lusc_tcga_pub	rna seq mma median Zscores	8%	0%
FAT4	Prostate Adenocarcinoma	prad_siuc_2015	rna seq mma median Zscores	4%	0%
FAT4	Cervical Squamous Cell Carcinoma	cesc_tcga	rna seq v2 mma median Zscores	8%	0%
FAT4	Lung Squamous Cell Carcinoma	lusc_tcga	rna seq v2 mma median Zscores	7%	0%
FAT4	Endometrial Carcinoma	ucec_tcga	rna seq v2 mma median Zscores	9%	0%

(b) FAT4 Expression

Figure 4.6: FAT4's genetic profile in real tumors extracted from cBioPortal. (A) FAT4's mutation and alteration frequency in different cancer studies (B) FAT4's Differential expression in different cancer studies.

4.4.3 Filtering Genes

Cell lines have mutations in many genes; however, all mutated genes are not of interest for biologists. Biologists – most of the time – give high importance to mutations in oncogenes or tumor suppressor genes since their role in cancer is well-defined. To facilitate biologists, we have created a gene filter in the following three ways:

1. Cancer genes: Construct graphs in the context of only oncogenes or tumor suppressor genes.
2. Exclude commonly mutated genes: Construct graphs in the context of all genes but exclude genes that have mutations in more than 90% of cell lines.
3. All genes: Disable the filter and construct graphs in the context of all genes.

This filter facilitates the users when they only want to focus on cancer genes' mutations and are interested in further exploration of only cancer genes. This filter also helps users when they want to focus on rarely mutated genes by allowing them to exclude genes that are mutated in more than 90% of cell lines. Figure 4.7 illustrates the effect of applying a gene filter. Figure 4.7A shows only cancer genes that are mutated in sensitive cell lines, and Figure 4.7B shows the network with all genes excluding frequently mutated genes, i.e., genes that are mutated in fewer than 90% of overall cell lines.

4.4.4 Finding Shared Connections

Another important feature of the IPCT for biologists is to find shared connections between data entities. In general, biologists explore databases to find hidden connec-

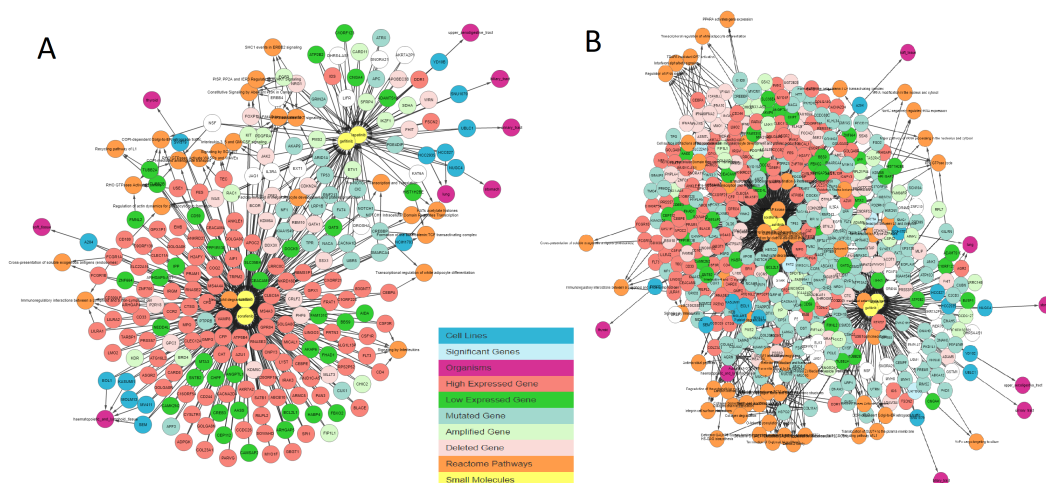


Figure 4.7: IPCT output for small molecule user query Lapatinib, Sorafenib, Gefitinib and Sunitinib after disabling REACTOME and Expression Atlas Databases and enabling Cell Lines and Mutated Genes only. (A) illustrates the results with gene filter = cancer genes, and (B) illustrates the value with gene filter = exclude common mutations.

tions between data entities such as hidden direct or indirect relationships between two cell lines and small molecule sensitivity or between small molecule sensitivity and gene mutations.

The shared connection filter in the IPCT allows users to investigate unknown or hidden relationships between data entities in five connected databases by simple clicks. For example, when a user inputs more than two small molecules, the IPCT constructs a graph with cell lines sensitive to the input small molecules, their mutated genes and data entities connected with mutated genes. By enabling the shared connection filter, the user can restrict the results to cell lines sensitive to both small molecules. Similarly, he can restrict the graph to genes mutated in more than one cell line and to pathways that are common between mutated genes. Using the shared connection filter,

Table 4.3: User Input query for Figure 4.7

Input Type	Input
Small Molecule	LAPATINIB, SORAFENIB, GEFITINIB and SUNITINIB
Databases Included	Associated Genes, Cell Lines
Databases Excluded	Expression Atlas, REACTOME
Genes Filter	(a) a) Cancer Genes Only, (b) All Genes
Drug Sensitivity	-1.50
Mutation Frequency	20%
Pathway Filter	All Pathways
Graph Type	All Connections

researchers can identify the unknown or hidden relationships between small molecules and genes.

Figures 4.5 and 4.8 and illustrates the output generated by the IPCT for LAPATINIB, SORAFENIB, GEFITINIB or SUNITINIB with the shared connection filter enabled. Since the shared connection filter was enabled, Figures 4.5 and 4.8 contains genes that are connected in more than two cell lines and pathways that are common between two or more than two genes.

4.5 Download Graph

The web interface of the IPCT allows users to navigate data entities connected with CTRP small molecules or CCLE cell lines. In addition to this, users can also download

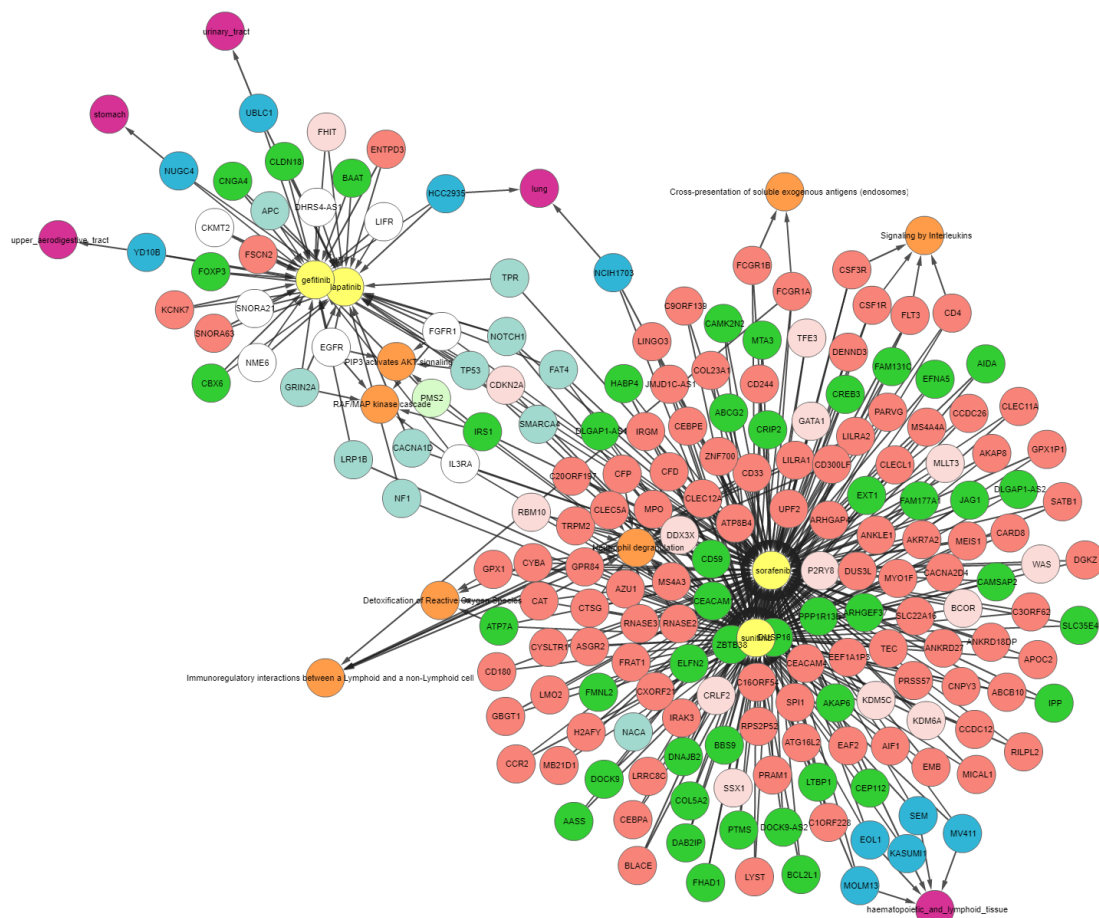


Figure 4.8: IPCT output for small molecule user query LAPATINIB, SORAFENIB, GEFITINIB and SUNITINIB with shared connection filter enabled. The graph shows only shared data points connected with LAPATINIB, SORAFENIB, GEFITINIB and SUNITINIB. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib; sky-blue nodes represent significant genes (those with multiple genomic aberrations); green and red nodes represent genes that are up-regulated and down-regulated in the sensitive cell lines, respectively; light green and light red represent the amplified and deleted genes in the sensitive cell lines, respectively; white nodes represent mutated genes; and orange nodes represent the REACTOME pathways of mutated genes

Table 4.4: User Input query for Figure 4.8

Input Type	Input
Small Molecule	LAPATINIB, SORAFENIB, GEFITINIB and SUNI-TINIB
Databases Included	Associated Genes, Cell Lines, Expression Atlas, REACTOME
Databases Excluded	None
Genes Filter	Cancer Genes Only
Drug Sensitivity	-1.50
Mutation Frequency	20%
Pathway Filter	All Pathways
Graph Type	Common Connections

output connectivity maps for future reference. The IPCT allows users to download an output connectivity maps in GraphML, JSON or CSV formats for further analysis and generation of high-quality graphs using external graph making tools such as Cytoscape. GraphML, CSV or JSON downloaded from the IPCT contain information about nodes representing data entities, and edges representing connection between nodes (data entities). Each edge is identified using its unique identifier and contains information in the form of its label, database identifier and URL. Labels store display information, and database identifiers are used to store color coding. Nodes representing small molecules contain additional information about their sensitivity to the connected cell lines. Similarly, nodes representing genes contain additional information about mutation frequency in sensitive cell lines.

4.6 Case Study

Lapatinib and Afatinib are two tyrosine kinase inhibitors that are effective in breast cancer. These drugs are usually effective in HER2 (ERBB2) mutation-positive patients⁶⁴, (Rimawi et al. 2015; Li et al. 2008). In this section, we demonstrate how the IPCT can be used to identify the mechanism of action of these two kinase inhibitors. For this purpose, in the first stage, we query lapatinib and afatinib in the IPCT. Figure 4.9 shows the graph containing the sensitive cell lines, associated genes, and their pathways generated by the IPCT as result of the query, without applying any filter. Genes associated with these drugs are colored and shaped based on their relationship; each color and shape represent a unique relationship between the genes and the cell lines sensitive to the input drug. As such, genes with certain colors and shapes can be classified as more important than other genes.

Next, we apply a filter to shortlist our gene set. We first apply the shared connection filter to see if any genes are associated with both drugs. Genes can have different associations with each drug, and the more important genes will be those that have the same association with both drugs. Figure 4.10 shows the resulting graph. The sky-blue genes are the most important ones, whereas those with a white background are the least important. The circled genes can be classified as the most relevant gene set due to the pathway clusters. Figure 4.10 shows that EGFR is amplified, NRG1 and FGFR1 are deleted, and AKAP9 and TP53 have mutations in cell lines sensitive to both drugs. These genes have been found to be relevant to lapatinib and afatinib in the literature (Forster et al. 2011; Leech et al. 2018; Li et al. 2008). ERBB2 is amplified and highly expressed in 95% of afatinib-sensitive cell lines and 100% of lapatinib-sensitive cell lines. Finally, we apply relationship filters to identify the most relevant results.

These filters are designed to filter genes that have multiple genomic aberrations with the queried drugs. Finally, Figure 4.11 demonstrates the relationship of ERBB2 with lapatinib and afatinib.

An other example of IPCT usage is identifying Dabrafenib and trametinib target gene. Dabrafenib and trametinib are BRAF inhibitors that are effective in melanoma skin cancer. These drugs are usually effective in BRAF mutation-positive patients. In Supplementary figure 4.12, we demonstrate how the IPCT can be used to identify the important gene i.e. BRAF that is associated with dabrafenib and trametinib. We have applied shared connection filter that gave us graph containing all genes that has some genomic association with cell lines sensitive to dabrafenib and trametinib. and highlight gene using gene find function.

4.7 Discussion and Conclusions

The recent advancements in pharmacogenomics through high-throughput sequencing have suggested that big data scientists develop innovative strategies to address the rapidly expanding biological data. However, the two major limitations to make effective use of this huge amount of information are extensive data heterogeneity and a lack of integration. To overcome these limitations, data scientists have been working on the development of integrated and biologist-friendly databases. For instance, EBI-RDF is a state of the art example that has enabled the integration of six different biological databases including UniProt, Expression Atlas, REACTOME, ChEMBL, BioModels and BioSamples. However, to the best of our knowledge, no large-scale efforts have been made to integrate pharmacogenomic features of cancer cell lines with cancer genomic

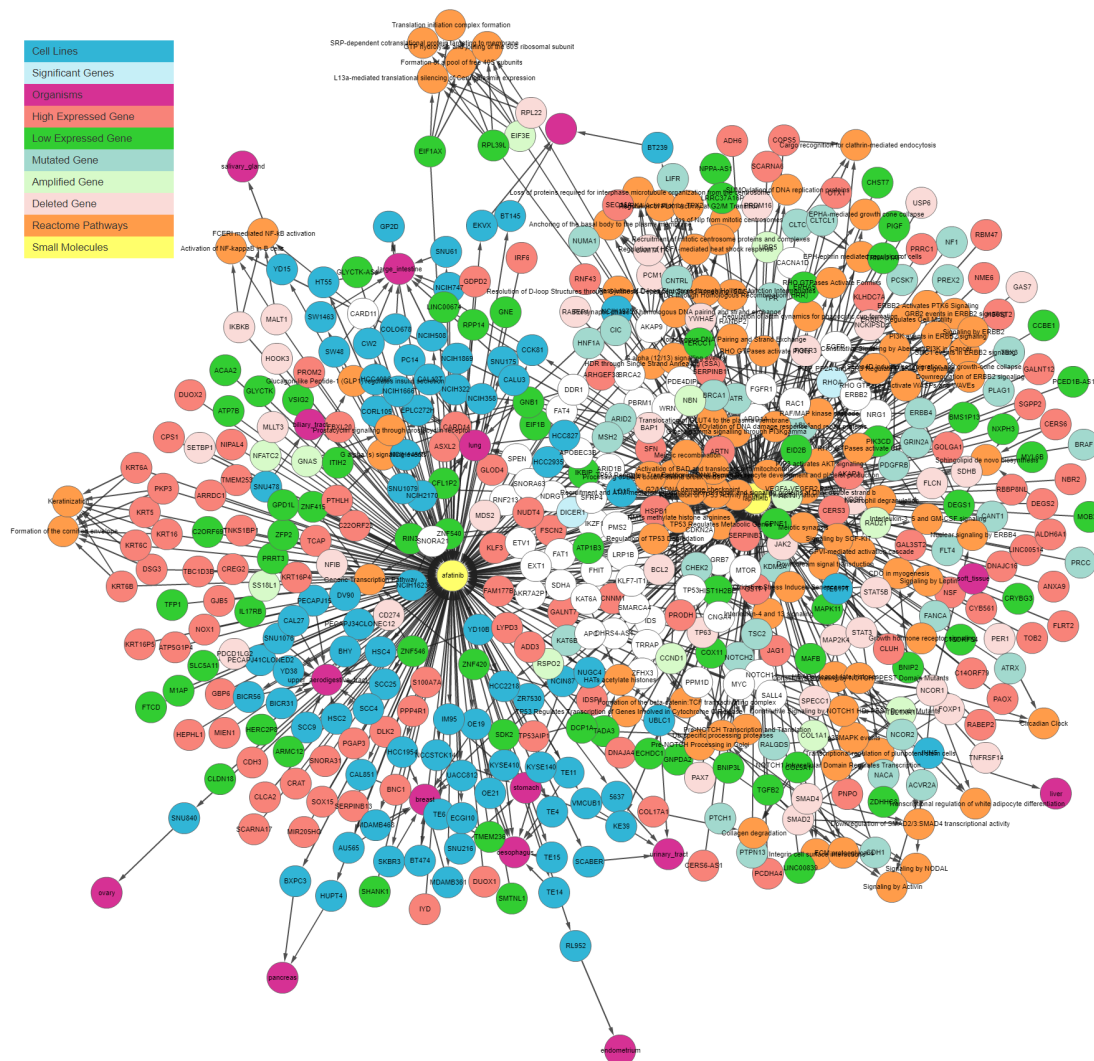


Figure 4.9: IPCT output for small molecule user query lapatinib and afatinib. The graph shows all data points connected with lapatinib and afatinib. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib and afatinib; sky-blue nodes represent significant genes (those with multiple genomic aberrations); green and red nodes represent genes that are up-regulated and down-regulated in the sensitive cell lines, respectively; light green and light red represent the amplified and deleted genes in the sensitive cell lines, respectively; white nodes represent mutated genes; and orange nodes represent the REACTOME pathways of mutated genes.

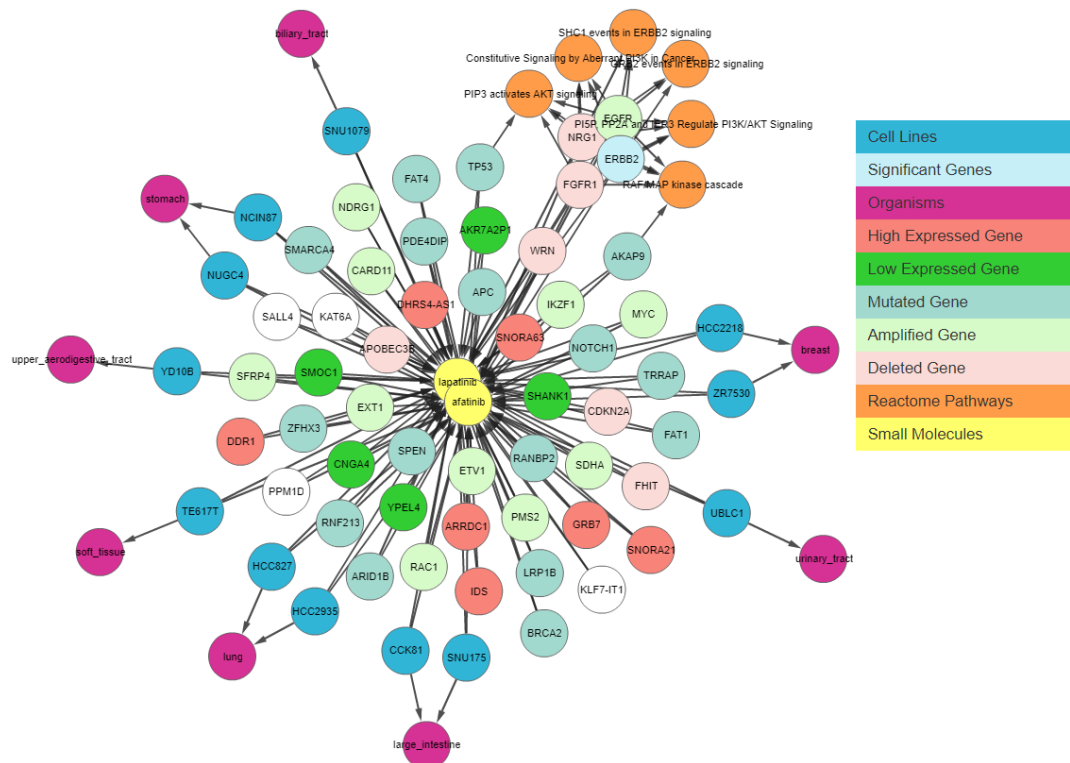


Figure 4.10: IPCT output for small molecule user query lapatinib and afatinib with the shared connection filter enabled. The graph shows all data points connected with lapatinib and afatinib. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib and afatinib; sky-blue nodes represent significant genes (those with multiple genomic aberrations); green and red nodes represent genes that are up-regulated and down-regulated in the sensitive cell lines, respectively; light green and light red represent the amplified and deleted genes in the sensitive cell lines, respectively; white nodes represent mutated genes; and orange nodes represent the REACTOME pathways of mutated genes.

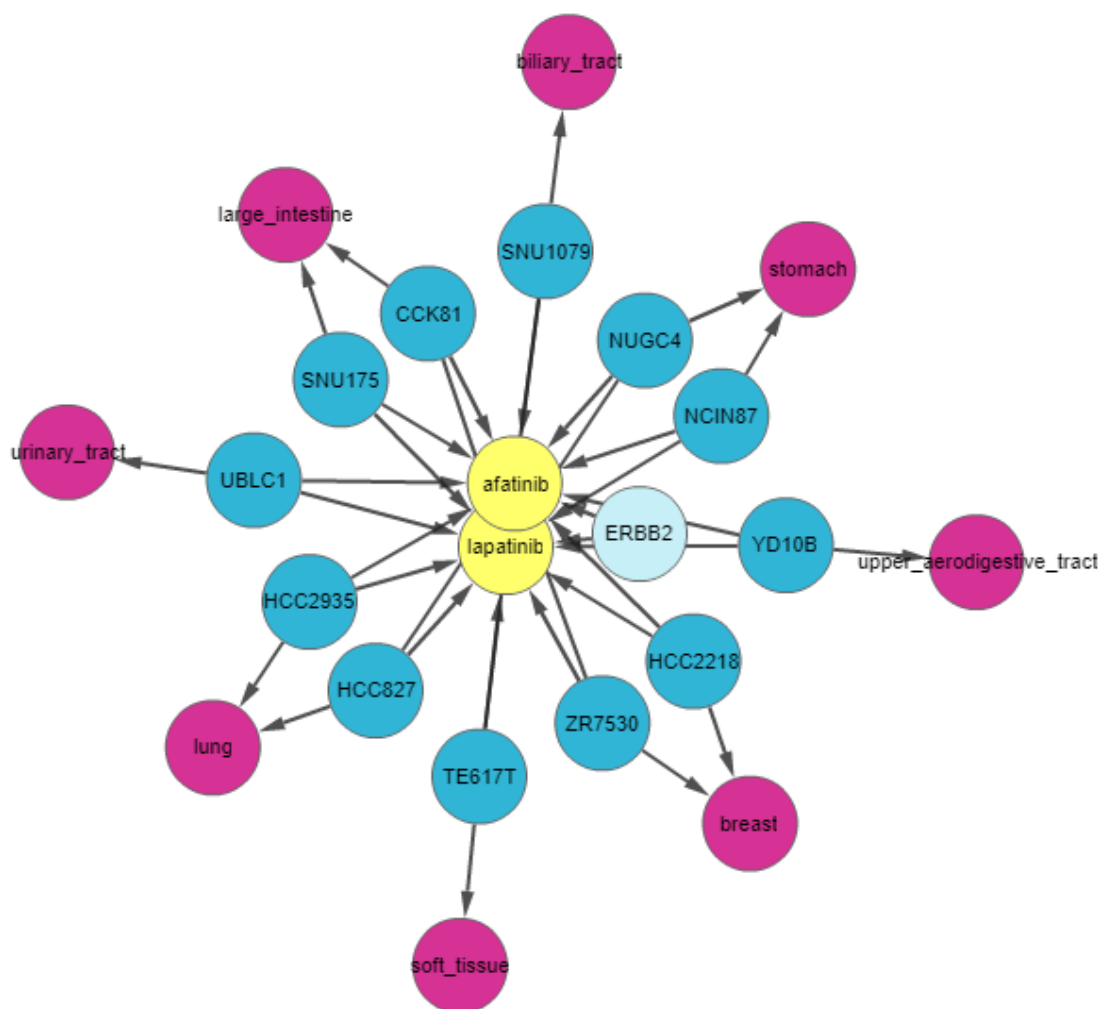


Figure 4.11: IPCT output for small molecule user query lapatinib and afatinib with the shared connection filter and the relationship filter enabled. The graph shows all data points connected with lapatinib and afatinib. Yellow nodes represent small molecules; blue nodes show cell lines sensitive to lapatinib and afatinib; and sky-blue nodes represent significant genes (those with multiple genomic aberrations).

Table 4.5: Genes associated with Lapatinib. These genes have genomic aberrations in more than 20% cell lines sensitive to Lapatinib

	Gene ID	Gene Symbol	Mutation %	Amplification %	Deletion %	High Expression %	Low Expression %	Score
1	2064	ERBB2	27	100	0	100	0	128
2	7157	TP53	53	0	100	0	0	104
3	387	RHOA	0	0	100	17	83	84
4	55193	PBRM1	27	0	100	0	0	78
5	1956	EGFR	27	100	0	0	0	78
6	776	CACNA1D	20	0	100	0	0	71
7	8289	ARID1A	20	0	100	0	0	71
8	2475	MTOR	20	0	100	0	0	71
9	6389	SDHA	20	75	0	0	0	59
10	23405	DICER1	20	75	0	0	0	59
11	8416	ANXA9	0	0	0	100	0	51
12	3084	NRG1	0	0	100	0	0	51
13	10289	EIF1B	0	0	0	0	100	51
14	100233209	PCED1B-AS1	0	0	0	0	100	51
15	51274	KLF3	0	0	0	100	0	51

Table 4.6: Genes associated with Afatinib. These genes have genomic aberrations in more than 20% cell lines sensitive to Afatinib

	Gene ID	Gene Symbol	Mutation %	Amplification %	Deletion %	High Expression %	Low Expression %	Score
1	2064	ERBB2	0	95	0	93	7	92
2	53353	LRP1B	35	0	100	0	0	86
3	7157	TP53	76	0	0	0	0	77
4	338324	S100A7A	0	0	0	100	0	51
5	163351	GBP6	0	0	0	100	0	51
6	89778	SERPINB11	0	0	0	100	0	51
7	5275	SERPINB13	0	0	0	100	0	51
8	286887	KRT6C	0	0	0	100	0	51
9	3664	IRF6	0	0	0	100	0	51
10	4087	SMAD2	0	0	100	0	0	51
11	7994	KAT6A	0	0	100	0	0	51
12	6665	SOX15	0	0	0	100	0	51
13	63970	TP53AIP1	0	0	0	100	0	51
14	3551	IKBKB	0	0	100	0	0	51
15	4089	SMAD4	0	0	100	0	0	51

Table 4.7: Genes association score with Lapatinib and Afatinib

	Gene ID	Gene Symbol	Lapatinib Score	Afatinib Score
1	2064	ERBB2	128	92
2	7157	TP53	104	77
3	1956	EGFR	78	46
4	6389	SDHA	59	49
5	3084	NRG1	51	50
6	5395	PMS2	51	48
7	2886	GRB7	51	47
8	84433	CARD11	51	48
9	2272	FHIT	51	50
10	10320	IKZF1	51	45
11	780	DDR1	51	39
12	6424	SFRP4	51	44
13	2115	ETV1	51	44
14	1029	CDKN2A	51	51
15	7486	WRN	51	51
16	5879	RAC1	51	47
17	8493	PPM1D	51	24
18	4609	MYC	41	50
19	2260	FGFR1	41	37
20	10397	NDRG1	39	49
21	2131	EXT1	39	49

Table 4.8: Pathways associated with genes having genomic changes in cell lines sensitive to Lapatinib and Afatinib

Reactome ID	Pathway Name	No of Genes
1	R-HSA-1257604 PIP3 activates AKT signaling	11
2	R-HSA-5673001 RAF/MAP kinase cascade	11
3	R-HSA-6811558 PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	9
4	R-HSA-2219530 Constitutive Signaling by Aberrant PI3K in Cancer	8
5	R-HSA-5628897 TP53 Regulates Metabolic Genes	8
6	R-HSA-3214858 RMTs methylate histone arginines	7
7	R-HSA-5689880 Ub-specific processing proteases	7
8	R-HSA-5693565 Recruitment and ATM-mediated phosphorylation o...	7
9	R-HSA-6785807 Interleukin-4 and 13 signaling	7
10	R-HSA-6798695 Neutrophil degranulation	7
11	R-HSA-6804756 Regulation of TP53 Activity through Phosphoryl...	7
12	R-HSA-6805567 Keratinization	7
13	R-HSA-6809371 Formation of the cornified envelope	7
14	R-HSA-69473 G2/M DNA damage checkpoint	7

Table 4.9: Genes associated with Dabrafenib. These genes have genomic aberrations in more than 20% cell lines sensitive to Dabrafenib

	Gene ID	Gene Symbol	Mutation %	Amplification %	Deletion %	High Expression %	Low Expression %	Score
1	673	BRAF	88	100	0	0	0	139
2	4286	MITF	0	100	0	85	15	86
3	1029	CDKN2A	24	0	100	0	0	75
4	2313	FLI1	0	0	100	0	0	51
5	1464	CSPG4	0	0	0	100	0	51
6	23365	ARHGEF12	0	0	100	0	0	51
7	2315	MLANA	0	0	0	100	0	51
8	2272	FHIT	0	0	100	0	0	51
9	3762	KCNJ5	0	0	100	0	0	51
10	5270	SERPINE2	0	0	0	100	0	51
11	126321	MFSD12	0	0	0	100	0	51
12	81138	OR7E101P	0	0	0	100	0	51
13	84767	TRIM51	0	0	0	100	0	51
14	419	ART3	0	0	0	100	0	51

Table 4.10: Genes associated with Tramentinib. These genes have genomic aberrations in more than 20% cell lines sensitive to Tramentinib

	Gene ID	Gene Symbol	Mutation %	Amplification %	Deletion %	High Expression %	Low Expression %	Score
1	7157	TP53	59	0	0	0	0	60
2	2531	KDSR	0	0	100	0	0	51
3	596	BCL2	0	0	100	0	0	51
4	2272	FHIT	0	0	100	0	0	51
5	4089	SMAD4	0	0	100	0	0	51
6	10892	MALT1	0	0	100	0	0	51
7	1029	CDKN2A	0	0	98	0	0	50
8	2115	ETV1	0	93	0	0	0	48
9	29126	CD274	0	0	91	0	0	47
10	80380	PDCD1LG2	0	0	91	0	0	47
11	4300	MLLT3	0	0	92	0	0	47
12	11168	PSIP1	0	0	91	0	0	47
13	3717	JAK2	0	0	92	0	0	47
14	4781	NFIB	0	0	88	0	0	45
15	475	ATOX1	0	0	0	87	13	38

Table 4.11: Genes association score with Dabrafenib and Trametinib

	Gene ID	Gene Symbol	Dabrafenib Score	Trametinib Score
1	673	BRAF	139	26
2	1029	CDKN2A	75	50
3	2272	FHIT	51	51
4	25825	BACE2	51	34
5	692099	FAM86DP	51	22
6	283652	SLC24A5	51	20
7	23500	DAAM2	51	31
8	4359	MPZ	51	23
9	4644	MYO5A	51	19
10	3717	JAK2	51	47
11	2115	ETV1	51	48
12	4781	NFIB	51	45
13	5354	PLP1	51	27
14	4300	MLLT3	51	47
15	11168	PSIP1	51	47
16	29126	CD274	51	47
17	80380	PDCD1LG2	51	47
18	399694	SHC4	51	16
19	22876	INPP5F	44	27
20	100190799	LDHAP2	43	34
21	475	ATOX1	43	38

Table 4.12: Pathways associated with genes having genomic changes in cell lines sensitive to Dabrafenib and Trametinib

	Reactome ID	Pathway Name	No of Genes
1	R-HSA-6802952	Signaling by BRAF and RAF fusions	7
2	R-HSA-5689880	Ub-specific processing proteases	5
3	R-HSA-1442490	Collagen degradation	4
4	R-HSA-1650814	Collagen biosynthesis and modifying enzymes	4
5	R-HSA-6785807	Interleukin-4 and 13 signaling	4
6	R-HSA-6798695	Neutrophil degranulation	4
7	R-HSA-6802946	Signaling by moderate kinase activity BRAF mut...	4
8	R-HSA-6802949	Signaling by RAS mutants	4
9	R-HSA-6802955	Paradoxical activation of RAF signaling by kin...	4
10	R-HSA-983231	Factors involved in megakaryocyte development ...	4
11	R-HSA-216083	Integrin cell surface interactions	3
12	R-HSA-3000178	ECM proteoglycans	3
13	R-HSA-2173795	Downregulation of SMAD2/3:SMAD4 transcriptiona...	2
14	R-HSA-3000170	Syndecan interactions	2

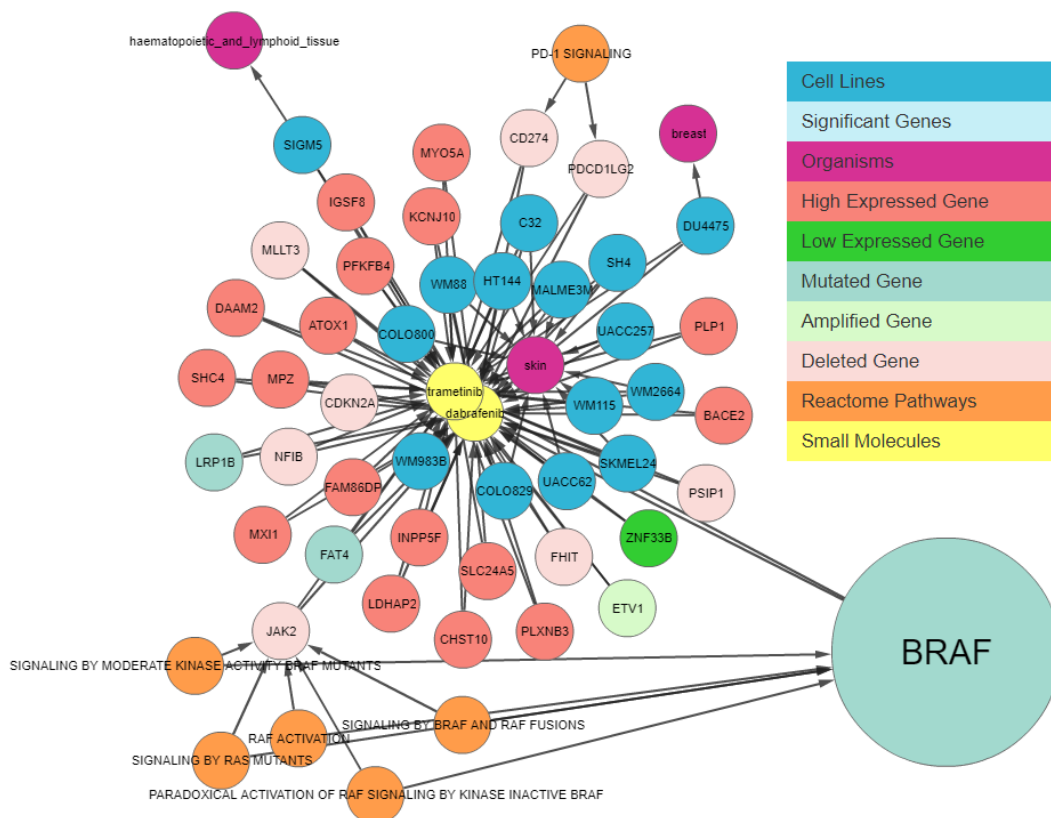


Figure 4.12: IPCT output for small molecule user query dabrafenib and trametinib with the shared connection filter and the relationship filter enabled. The graph shows all data points connected with dabrafenib and trametinib.

features of real cancer patients.

One of the key questions for any biologist is whether genomic features of cancer cell lines that are sensitive to drugs are also present in real cancer tissues. Earlier biologists had to search through multiple heterogeneous databases, which is a challenging job for researchers with limited computational skills. Recently, Elena Piñero-Yáñez et al. (Piñero-Yáñez, et al., 2018) developed PanDrugs to prioritize anticancer drug treatments depending on patients' genomic profiles. PanDrugs majorly focuses on clinical aspects of cancer genomics, in contrast to IPCT which is designed to help researchers in generation and in-silico testing of hypothesis on pharmacogenomics data of human

cell lines and genomic data of human tumor samples. IPCT enables data integration and interoperability at CTRP, CCLE, Expression Atlas, REACTOME and cBioPortal, allowing users to investigate the connectivity map of cell lines, small molecules and genes of interest in a user-friendly fashion.

In conclusion, IPCT enables biologists to investigate the connectivity map of small molecules and genomics features in relationship with cancer cell lines and real cancer tissues. It also highlights the genomic features sensitive to a specific drug and the percentage of cancer patients affected by that drug. Of note, IPCT can also identify cancer cell lines that are truly representative of real cancer tissues. In conclusion, the integration of these five major databases in a biologist-friendly manner will facilitate researchers in the generation of new and tangible hypotheses, leading to new clinical trials.

Chapter 5

Discussion

The origin of science starts with experimental science where scientists used to perform experiments to study natural phenomena. Experimental science also existed centuries ago which was then superseded by theoretical science in the 18th century. The era of theoretical science starts with Kepler's Laws and Newton's Laws of Motion. The basis of theoretical science was developing theoretic models and proving them analytically. However with the passage of time, because of complexity in nature, these models grew too complex that it became really difficult to solve and prove this model analytically. This leads the science to its third paradigm known as computational science in which computational simulations were designed for those problems which are too complex and it was not possible to solve them analytically.

In recent years, experimental side also progressed and now with the help of latest technology it has become possible to generate a massive amount of experimental data. In molecular biology, for example, mass spectrometry, microarray, and next-generation sequencing technologies have been producing a large amount of Genomics and Pro-

teomics data on a daily basis. The latest development in information technology has also made literature sharing faster than it was in the past. This has led to the problem of extracting knowledge from the scientific literature. These developments lead towards the creation of biological data repositories for experimental data, theoretical data, and scientific data to store these data in a more structured and organized fashion. Since these repositories contain overlapping information, therefore, more scientific insights can be explored by integrating these data repositories using common data entities among them.

Integrating biological databases has been a challenge because of heterogeneity and diversity in the structure of biological databases. At the same time integration of biological databases is essential because biologists have to ask questions in different contexts. Techniques that have been used previously – creating one large integrated database – are not realistic in the current era of biological big data because of growing data volume. On the other side researchers in Web Science solved the same problem by using Linked Open Data in which each resource is assigned an Internationalized Resource Identifier (IRI). They developed the Resource Description Framework (RDF) which allows the representation of data entities using a pre-defined schema and linked them with each other using URIs. In RDF each data entity is represented as a resource and its attributes are known as properties of those resources. RDF has been accepted as the standard for Linked Open Data by World Wide Consortium (W3C). It has been in practice from a decade to facilitate the process of data integration and ontology development. Like the other domains, in the domain of bioinformatics and computational biology, ontologies have been used to express biological knowledge. Since biological data is networked, therefore, Ontologies can represent the biological knowledge in a

more natural way with minimum loss of information.

In addition to this ontology also help in integrating data which more challenging for bioinformatics because of its growth and diversity. EBI-RDF is an example of Semantic Web (Ontology) based biological data integration, which integrates six biological databases namely UniProt, Expression Atlas, REACTOME, ChEMBL, BioModels, and BioSamples. However, one practical problem is that in order to extract real value from these databases one must have good knowledge of SPARQL query language and underlying ontological schema. The main goals of this dissertation are (1) to overcome this practical limitation by providing a gene-centric user-friendly interface which will allow users to explore EBI-RDF platform in a gene-centric fashion and (2) to extend EBI-RDF platform by including CCLE Cell lines and CTRP Drugs to allow biologists exploration of CTRP experiments in the context of EBI-RDF databases. Our research will provide biologist with a user-friendly way of exploring integrated biological data in a networked fashion.

To overcome this problem of understanding baseline technologies and schemas in this research we have designed and developed a framework for integration and visualization of biological databases and had shown how EBI-RDF platform can be extended to fulfill customized needs of biologists. The contributions of this dissertation are (that (1) it has enabled biologists friendly gene-centric access to EBI-RDF platform which allow biologist to search among six EBI-RDF databases in a gene-centric fashion using user friendly web interface and (2) it has shown how EBI-RDF platform can be extended and customized by connecting its four databases with The Cancer Therapeutics Response Portal (CTRP), Cancer Cell Line Encyclopedia (CCLE) and cBioPortal. The resultant framework will allow biologists to search EBI-RDF platform, CTRP drugs,

and CCLE cell lines and data entities connected with them in EBI-RDF platform, cBioPortal, CCLE, and CTRP. In addition to this, it will also allow users to apply filters on data entities while searching to facilitate biologists for asking questions in their specific context.

This research has been carried out in two phases, in the first phase we designed and develop the framework which allows biologist query EBI-RDF database in a gene-centric way. Users can start exploring EBI-RDF platform by entering a set of genes or small molecules and then expand their context by adding filters. In the second phase, we extended our framework by adding three more databases to it (1) the Cancer Therapeutics Response Portal (CTRP), (2) Cancer Cell Line Encyclopedia (CCLE) and (3) cBioPortal. This allows biologist to explore CTRP and CCLE database in an integrated way with EBI databases. The framework connects CTRP, CCLE, and cBioPortal with UniProt, REACTOME pathways, Expression Atlas and ChEMBL (database of small molecules and drugs).

cMapper and IPCT frameworks have two components (1) database and (2) web application. cMapper and IPCT databases were developed using MySQL database. We used RDMS because RDF triple stores are not suitable for efficient querying of large databases. This means users cannot get answers of on-the-fly queries in real time. Databases contain two types of tables (1) connectivity tables and (2) data tables. Connectivity tables store connections between all databases in database and data tables store basic information about each data entities used in connectivity tables. We created connectivity tables by identifying identical data entities across databases using string matching, owl-same as RDF property and named entity recognition (NER). In cMapper database, UniProtKB identifiers were used as a central identifier to identify entities from

other databases, and connect them with each other using UniProtKB identifiers. In IPCT database CTRP Drug ID, CCLE cell lines IDs and Entrez Gene IDs was used as bridged properties to connect data entities across databases.

cMapper and IPCT have been developed as a web application with an integrated MySQL database. The web application was developed using Java and deployed on Wildfly Web Application Server. We developed the user interface using HTML5, JQuery, and the Cytoscape Graph API. By using cMapper users can construct a graph of data entities connected with Genes or Small molecules. Similarly by using IPCT users can construct a graph of data entities connected with small molecules, or cell lines. Both frameworks allow users to (1) create connected graph of data entities, (2) view graph on browsers, (3) download graph as GraphML, image or PDF for further analysis, (4) apply filter on graph to perform a context-specific search (5) identify common connections between data entities in the graph by enabling shared connection filter.

Bibliography

- [1] Alberto Anguita, Miguel García-Remesal, Diana de la Iglesia, and Victor Maojo. Ncbi2rdf: Enabling full rdf-based access to ncbi databases. *BioMed research international*, 2013, 2013.
- [2] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.
- [3] Amrita Basu, Nicole E Bodycombe, Jaime H Cheah, Edmund V Price, Ke Liu, Giannina I Schaefer, Richard Y Ebright, Michelle L Stewart, Daisuke Ito, Stephanie Wang, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, 2013.
- [4] Corine M Beaufort, Jean CA Helmijs, Anna M Piskorz, Marlous Hoogstraat, Kirsten Ruigrok-Ritstier, Nicolle Besselink, Muhammed Murtaza, Wilfred FJ van IJcken, Anouk AJ Heine, Marcel Smid, et al. Ovarian cancer cell line panel (occp): clinical importance of in vitro morphological subtypes. *PloS one*, 9(9):e103988, 2014.
- [5] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and

- Jean Morissette. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716, 2008.
- [6] Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer Fostel, Yongqun He, Phillip W Lord, James Malone, Helen E Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al. Modeling biomedical experimental processes with obi. *J. Biomedical Semantics*, 1(S-1):S7, 2010.
- [7] Alison Callahan, José Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data. pages 200–212, 2013.
- [8] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Data complexity of query answering in description logics. *Artificial Intelligence*, 195:335–360, 2013.
- [9] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO precision oncology*, 1:1–16, 2017.
- [10] Dhananjay Chitale, Yixuan Gong, Barry S Taylor, Stephen Broderick, Cameron Brennan, Romel Somwar, Benjamin Golas, Lu Wang, Noriko Motoi, Janos Szoke, et al. An integrated genomic analysis of lung cancer reveals loss of dusp4 in egfr-mutant tumors. *Oncogene*, 28(31):2773, 2009.
- [11] Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium, et al. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84, 2015.

- [12] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.
- [13] Kelsy C Cotto, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, and Malachi Griffith. Dgidb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic acids research*, 46(D1):D1068–D1073, 2017.
- [14] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.
- [15] Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2015. *Nucleic acids research*, 43(D1):D662–D669, 2014.
- [16] André Freitas, Edward Curry, Joao Gabriel Oliveira, and Sean O’Riain. Querying heterogeneous datasets on the linked data web: challenges, approaches, and trends. *IEEE Internet Computing*, 16(1):24–33, 2012.
- [17] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature reviews cancer*, 4(3):177, 2004.
- [18] Matthias E Futschik, Gautam Chaurasia, and Hanspeter Herzel. Comparison of human protein–protein interaction maps. *Bioinformatics*, 23(5):605–611, 2007.

- [19] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbiportal. *Sci. Signal.*, 6(269):p11–p11, 2013.
- [20] Arti Gaur, David A Jewell, Yu Liang, Dana Ridzon, Jason H Moore, Caifu Chen, Victor R Ambros, and Mark A Israel. Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer research*, 67(6):2456–2468, 2007.
- [21] Michael Gillam, Craig Feied, Jonathan Handler, Eliza Moody, Ben Shneiderman, Catherine Plaisant, Mark Smith, and John Dickason. The healthcare singularity and the age of semantic medicine. 2009.
- [22] Jean-Pierre Gillet, Sudhir Varma, and Michael M Gottesman. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute*, 105(7):452–458, 2013.
- [23] Vladimir Gligorijević, Vuk Janjić, and Nataša Pržulj. Integration of molecular network data reconstructs gene ontology. *Bioinformatics*, 30(17):i594–i600, 2014.
- [24] Carole Goble and Robert Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–693, 2008.
- [25] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merckenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(Suppl 2):I1, 2014.

- [26] Casey S Greene, Jie Tan, Matthew Ung, Jason H Moore, and Chao Cheng. Big data bioinformatics. *Journal of cellular physiology*, 229(12):1896–1900, 2014.
- [27] Michael Hamacher, Friedrich Herberg, Marius Ueffing, and Helmut E Meyer. Seven successful years of omics research: the human brain proteome project within the national german research network (ngfn). *Proteomics*, 8(6):1116–1117, 2008.
- [28] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.
- [29] Ian Harrow, Wendy Filsell, Peter Woollard, Ian Dix, Michael Braxenthaler, Richard Gedye, David Hoole, Richard Kidd, Jabe Wilson, and Dietrich Rebholz-Schuhmann. Towards virtual knowledge broker services for semantic integration of life science literature and data sources. *Drug discovery today*, 18(9):428–434, 2013.
- [30] Leland H Hartwell, Philippe Szankasi, Christopher J Roberts, Andrew W Murray, and Stephen H Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278(5340):1064–1068, 1997.
- [31] Matthew Edwin Holford. *Using Semantic Web Tools to Create An Integrated Framework For Biomedical Research*. PhD thesis, YALE UNIVERSITY, 2014.
- [32] Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.

- [33] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.
- [34] Guangxu Jin, Changhe Fu, Hong Zhao, Kemi Cui, Jenny Chang, and Stephen TC Wong. A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer research*, 72(1):33–44, 2012.
- [35] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, et al. The ebi rdf platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014.
- [36] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, et al. The ebi rdf platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014.
- [37] A Kaplun, JD Hogan, F Schacherer, AP Peter, S Krishna, BR Braun, R Nambudiry, MG Nitu, R Mallelwar, and A Albayrak. Pgmdb: a comprehensive manually curated pharmacogenomic database. *The pharmacogenomics journal*, 16(2):124, 2016.
- [38] Shin Kawano, Tsutomu Watanabe, Sohei Mizuguchi, Norie Araki, Toshiaki Katayama, and Atsuko Yamaguchi. Togotable: cross-database annotation system using the resource description framework (rdf) data model. *Nucleic acids research*, 42(W1):W442–W448, 2014.

- [39] Gerald T Keusch. What do-omics mean for the science and policy of the nutritional sciences?-. *The American journal of clinical nutrition*, 83(2):520S–522S, 2006.
- [40] Shivaani Kummar, Martin Gutierrez, James H Doroshov, and Anthony J Murgo. Drug development in oncology: classical cytotoxics and molecularly targeted agents. *British journal of clinical pharmacology*, 62(1):15–26, 2006.
- [41] Etienne Y Lasfargues and Luciano Ozzello. Cultivation of human breast carcinomas. *Journal of the National Cancer Institute*, 21(6):1131–1147, 1958.
- [42] Jiao Li, Si Zheng, Bin Chen, Atul J Butte, S Joshua Swamidass, and Zhiyong Lu. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, 17(1):2–12, 2015.
- [43] Hongfang Liu, Petula D’Andrade, Stephanie Fulmer-Smentek, Philip Lorenzi, Kurt W Kohn, John N Weinstein, Yves Pommier, and William C Reinhold. mrna and microrna expression profiles of the nci-60 integrated with drug activities. *Molecular cancer therapeutics*, 9(5):1080–1091, 2010.
- [44] Ying Liu and Walter F Bodmer. Analysis of p53 mutations and their expression in 56 colorectal cancer cell lines. *Proceedings of the National Academy of Sciences*, 103(4):976–981, 2006.
- [45] Christian T Lopes, Max Franz, Farzana Kazi, Sylva L Donaldson, Quaid Morris, and Gary D Bader. Cytoscape web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2348, 2010.
- [46] Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, and Peter Tarczy-Hornoch. Data integration and genomic medicine. *Journal of biomedical informatics*, 40(1):5–16, 2007.

- [47] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.
- [48] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302, 2016.
- [49] M Scott Marshall, Richard Boyce, Helena F Deus, Jun Zhao, Egon L Willighagen, Matthias Samwald, Elgar Pichler, Janos Hajagos, Eric Prud’hommeaux, and Susie Stephens. Emerging practices for mapping and linking life sciences data using rdf—a case series. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:2–13, 2012.
- [50] Sebastian Mate, Felix Köpcke, Dennis Toddenroth, Marcus Martin, Hans-Ulrich Prokosch, Thomas Bürkle, and Thomas Ganslandt. Ontology-based data integration between clinical and research systems. *PloS one*, 10(1), 2015.
- [51] Jomol P Mathew, Barry S Taylor, Gary D Bader, Saiju Pyarajan, Marco Antonioti, Arul M Chinnaiyan, Chris Sander, Steven J Burakoff, and Bud Mishra. From bytes to bedside: Data integration and computational biology for translational cancer research. *PLoS computational biology*, 3(2):e12, 2007.
- [52] Katina Michael and Keith W Miller. Big data: New opportunities and new challenges [guest editors’ introduction]. *Computer*, 46(6):22–24, 2013.
- [53] Magali Michaut, Suet-Feung Chin, Ian Majewski, Tesa M Severson, Tycho Bis-

- meijer, Leanne De Koning, Justine K Peeters, Philip C Schouten, Oscar M Rueda, Astrid J Bosma, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific reports*, 6:18517, 2016.
- [54] Lorenzo Moja, Ludovica Tagliabue, Sara Balduzzi, Elena Parmelli, Vanna Pistotti, Valentina Guarneri, and Roberto D’Amico. Trastuzumab containing regimens for early breast cancer. *Cochrane database of systematic reviews*, (4), 2012.
- [55] Peter Mork, Alon Halevy, and Peter Tarczy-Hornoch. A model for data integration systems of biomedical data applied to online genetic databases. In *Proceedings of the AMIA Symposium*, page 473. American Medical Informatics Association, 2001.
- [56] National Institutes of health et al. Genetics home reference, 2012.
- [57] Tim O’Reilly, Mike Loukides, Julie Steele, and Colin Hill. *How data science is transforming health care.* ” O’Reilly Media, Inc.”, 2012.
- [58] Jeff Z Pan. Resource description framework. In *Handbook on ontologies*, pages 71–90. Springer, 2009.
- [59] Claude Pasquier. Biological data integration using semantic web technologies. *Biochimie*, 90(4):584–594, 2008.
- [60] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry’s grand challenge. *Nature reviews Drug discovery*, 9(3):203, 2010.
- [61] Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Bar-

- rera, Tony Burdett, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic acids research*, 44(D1):D746–D752, 2015.
- [62] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, and Lars Juhl Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 2014.
- [63] Raajit Rampal, Fatima Al-Shahrour, Omar Abdel-Wahab, Jay P Patel, Jean-Philippe Brunel, Craig H Mermel, Adam J Bass, Jennifer Pretz, Jihae Ahn, Todd Hricik, et al. Integrated genomic analysis illustrates the central role of jak-stat pathway activation in myeloproliferative neoplasm pathogenesis. *Blood*, 123(22):e123–e133, 2014.
- [64] Mothaffar F Rimawi, Sabina B Aleixo, Ashley Alarcon Rozas, João Nunes de Matos Neto, Maira Caleffi, Alicardo Cesar Figueira, Sulene Cunha Souza, Andre B Reiriz, Carolina Gutierrez, Heloisa Arantes, et al. A neoadjuvant, randomized, open-label phase ii trial of afatinib versus trastuzumab versus lapatinib in patients with locally advanced her2-positive breast cancer. *Clinical breast cancer*, 15(2):101–109, 2015.
- [65] Daniel L Rubin, Nigam H Shah, and Natalya F Noy. Biomedical ontologies: a functional perspective. *Briefings in bioinformatics*, 9(1):75–90, 2007.
- [66] Andreas Schultz, Andrea Matteini, Robert Isele, Pablo N Mendes, Christian Bizer, and Christian Becker. Ldif-a framework for large-scale linked data integration. In

- 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France, 2012.*
- [67] Dominik Schweiger, Zlatko Trajanoski, and Stephan Pabinger. Sparqlgraph: a web-based platform for graphically querying biological semantic web databases. *BMC bioinformatics*, 15(1):279, 2014.
- [68] David B Searls. Data integration: challenges for drug discovery. *Nature reviews Drug discovery*, 4(1):45, 2005.
- [69] Sreenath V Sharma, Daniel A Haber, and Jeff Settleman. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature reviews cancer*, 10(4):241, 2010.
- [70] Ie-Ming Shih and Tian-Li Wang. Apply innovative technologies to explore cancer genome. *Current opinion in oncology*, 17(1):33–38, 2005.
- [71] Robert H Shoemaker. The nci60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813, 2006.
- [72] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.
- [73] Larisa N Soldatova and Ross D King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006.
- [74] Cristian Taccioni, Giovanni Sorrentino, Alessandro Zannini, Jimmy Caroli, Domenico Beneventano, Laura Anderlucci, Marco Lolli, Silvio Biccato, and Gian-

- nino Del Sal. Mdp, a database linking drug response data to genomic information, identifies dasatinib and statins as a combinatorial strategy to inhibit yap/taz in cancer cells. *Oncotarget*, 6(36):38854, 2015.
- [75] Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto, and Yoshihiro Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, 2012.
- [76] Melanie B Thomas, James P O’Beirne, Junji Furuse, Anthony TC Chan, Ghassan Abou-Alfa, and Philip Johnson. Systemic therapy for hepatocellular carcinoma: cytotoxic chemotherapy, targeted therapy and immunotherapy. *Annals of surgical oncology*, 15(4):1008–1014, 2008.
- [77] Caroline F Thorn, Teri E Klein, and Russ B Altman. Pharmgkb: the pharmacogenomics knowledge base. In *Pharmacogenomics*, pages 311–320. Springer, 2013.
- [78] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.
- [79] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.
- [80] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblas-

- toma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.
- [81] John N Weinstein. Spotlight on molecular profiling: “integromic” analysis of the nci-60 cancer cell lines. *Molecular cancer therapeutics*, 5(11):2601–2605, 2006.
- [82] Gio Wiederhold. Mediators in the architecture of future information systems. *Computer*, 25(3):38–49, 1992.
- [83] Jennifer L Wilding and Walter F Bodmer. Cancer cell lines for drug discovery and development. *Cancer research*, 74(9):2377–2384, 2014.
- [84] Satya P Yadav. The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT*, 18(5):277, 2007.
- [85] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

국내 요약

생물학의 기술적 진보는 엄청난 양의 다중 오믹스 데이터 집합들을 양산하였다. 생물학적 데이터베이스와 다중 오믹스 데이터 집합 간의 통합은 생물학적 개체 간의 숨겨진 연결을 밝힐 수 있기 때문에 생물 학자에게는 필수적이다. 그러나 데이터 집합들 간의 통합 프로세스는 다양하고 이질적인 특성으로 인해 매우 복잡하고 각 생물학적 데이터베이스와 오믹스 데이터 세트는 특정 생물학적 영역 및 오믹스 영역을 다루기 위해 독립적으로 개발되고 생성되므로 그 구조 (데이터 구성 방법)는 서로 다른 형태를 취하고 있다. 이러한 이질적 특성으로 인해 오믹스 데이터베이스의 통합은 과학자들에게 어려운 과제 중 하나였다.

RDF (Resource Description Framework)는 트리플 형태로 데이터를 게시할 수 있는 통일된 메커니즘을 제공하여 서로 이질적인 리소스들을 연결할 수 있게 해준다. 트리플 데이터가 포함된 데이터베이스를 트리플 스토어라고 하는데 EBI-RDF 플랫폼은 RDF 기술을 사용하여 트리플 스토어를 게시함으로써 6개의 독립된 생물학적 데이터베이스를 해석하고 통합된 액세스를 가능하게 했다. 그러나 이러한 트리플 스토어를 쿼리하려면 스키마와 SPARQL 쿼리 언어에 대한 심층적인 지식이 필요하다. 이러한 한계를 극복하기 위해 이 논문의 첫번째 부분에서는 생물 학자들에게 친화적인 방식으로 통합된 생물학적 데이터베이스를 시각화하는 유전자 중심 플랫폼인 cMapper를 제시하고 있다. cMapper는 생물 학자가 RDF 및 SPARQL 쿼리 언어에 대한 기술적 지식없이 통합 방식으로 (1) UniProt, (2) Expression Atlas, (3) REACTOME, (4) ChEMBL, (5) BioModels 및 (6)

Biosamples 등 6 개의 생물학적 데이터베이스를 쿼리 할 수 있게 해준다.

논문의 두 번째 부분은 pharmacogenomics 데이터를 다른 생물학적 데이터베이스와 통합하는 프레임 워크인 cMapper의 확장 버전인 IPCT에 대한 것이다. IPCT는 CCLE와 cBioPortal 에서의 암 세포주 및 암 조직 유전적 이상, CTRP에서의 약물 반응 데이터, Expression Atlas에서의 차별발현유전자의 실험 조건 그리고 REACTOME의 생물학적 경로들을 통합한다. IPCT는 생물학자들이 관심있는 약물에 민감한 암 세포주의 유전적 이상을 탐색할 수 있도록 해줄 뿐만 아니라 관심있는 세포주에 민감한 약물을 검색하는 기능도 제공한다. 또한 IPCT는 데이터 통합에 의해 사용자들에게 암세포주와 조직에서의 유전적 이상을 비교하는 기능도 제공한다.

cMapper와 IPCT는 사용자들에게 관심있는 항목에 필터를 적용할 수 있도록 해준다. 사용자가 하나 이상의 유전자 및 작은 분자 또는 세포주들을 입력하면 이와 연결된 일반적인 생물학적 항목을 찾는 옵션을 선택할 수 있다. 또한 두 플랫폼 모두 사용자에게 화면에서 그래프를 시각화 하거나 PNG 또는 GraphML 형식으로 다운로드 할 수 있는 기능을 제공하고 IPCT를 통해 사용자는 CSV 및 JSON 형식의 데이터를 다운로드하여 추가 분석을 수행할 수 있다. 결론적으로 이 논문에서 수행된 연구는 생물학에서의 데이터 통합 문제를 다루며, 생물 학자들에게 친숙한 방식으로 통합된 생물학적 데이터 제시를 위해 사용할 수 있는 현대 데이터 계산 방법을 보여줌으로써 생물 학자들로 하여금 각 개체 간의 잠재성 있는 숨겨진 관계 확인하고 이를 이용하여 자신의 가설을 세울 수 있도록 해준다.

