Building a database to store Protein-Protein Interactions (PPI) in a rule based format

Ayush Das

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2019

Abstract

The study of Protein-Protein interactions (PPI) involves the analysis and identification of complexes that may form under a variety of reaction conditions. These reactions were initially modeled as Ordinary Differential Equations (ODEs) [25], which is now progressing to a rule-based modeling approach. This is because the interacting biomolecules have the potential to interact in a myriad different ways. The number of possible post-translational modifications and complexes grow exponentially when considering the binary interactions within the reaction network. Using traditional methods like ODEs to model PPI requires large amounts of reaction specific details, and the chemical kinetics of the interactions within network requires explicit mention of the network conditions [9]. A rule-based model on the other hand comprises of, a set of rules where the network specification is implicit. These rules can specified using model specification languages like Kappa [14] or BioNetGen [16]. Software tools enable researchers to model these interactions using different objectives like deterministic or stochastic modeling. Hence, researchers in bioinformatics have spent tremendous efforts in collecting the Protein-Protein interaction rules and the purpose of this project is to create and load a database with the PPI rules stored in a rule-based format. This will enable researchers to readily access PPI rules which when fed to a simulator will enable study of the protein interactions and draw conclusions based on their observations.

Acknowledgements

I would like to thank my supervisor Oksana Sorokina for her continued guidance and support during all stages of this project. The insightful feedback and guidance helped in improving the project to a large extent. I would also like to thank, Anatoly Sorokin and Douglas Armstrong for their insightful feedback and support during the course of this project.

Finally, I would like to thank my family for their support and guidance.

Table of Contents

1 Introduction **1**

2 Background **3**

2.1 Protein-Protein Interaction (PPI) Databases 4

2.2 Motivation for constructing the rule based database 5

2.3 Understanding rule based specification of protein interactions 5

2.3.1 General Understanding of rule based languages 5

2.3.2 Syntax and semantics of Kappa rules 7

2.3.3 Analysis of KaSim outputs 8

2.4 Application of Protein-Protein Rule Interaction Database 10

3 Methodology **12**

3.1 Methodology for creating the database 12

3.1.1 Entity Relationship Diagram: 13

3.1.2 Description of main tables 14

3.1.3 Description of relationship tables 15

3.1.4 Advantages and disadvantages of the database structure 16

3.2 Steps for adding new entries to the database 16

3.2.1 Agent 16

3.2.2 Domain 17

3.2.3 Domain-Agent Relationship 17

3.2.4 Rule 17

3.2.5 Rule-Domain-Agent 17

3.3 Description of Stored Procedures 18

3.3.1 GetRulesFromAgentName 18

3.3.2 GetRulesFromDomainNamePair 18

3.4 Description of the project folder and its files 19

3.4.1 Project folder source 19
3.4.2 Project folder structure and files 19
3.5 Web Application for accessing the stored procedures 22
3.6 Deployment steps 22
3.6.1 Database Scripts 22
3.6.2 Web Application 22

4 Results and Discussion **23**

5 Conclusions **24**
5.1 Final Reminder 24

Bibliography **25**

A Software Version **29**
A.1 Operating System 29
A.2 MySQL Version/ Client Version 29
A.3 Python Version 29
A.4 MySQL Connector Version in Python scripts 29
A.5 Django Version 29
A.6 Data Tables Version 29

Chapter 1

Introduction

Protein is an important component of the cells in the human body. It is an important component of bones, muscles, cartilage and so on. Decades of research in the field of biology have produced a vast repository of knowledge on individual protein molecules. Examples of such knowledge base include UniProt [21]. However, in order to further explore the relationships of complex molecular species it is imperative to understand the interactions that take place between them and their governing rules.

As per [13] Protein-Protein Interactions (PPI) are defined as the physical contacts with molecular docking between the protein molecules that occur in a living organism or cell. PPI interactions play a vital role as they dictate cellular activities which are responsible for good health or diseases. Achieving an in-depth understanding of protein interactions will help researchers improve the existing quality of medicine and health care in general. According to [17] an important source for drug discovery is the study of PPI. This is also evident from the fact that, as per [18] in recent times the study of PPI has gained momentum for research in the field of anti-cancer therapy. All these facts suggest the importance of PPI and it's applications.

Known PPIs are stored in PPI repositories designed to be easily retrievable by the researchers based on the relevant search term. There have been several databases in the past that have tackled the problem of collecting the PPIs. Such databases are of different varieties, based on their method of organizing and structuring the data. These kinds of databases are covered in greater detail, in Chapter 2.

PPI interactions are used to model the dynamics of protein complexes in the cell. Initially they were modeled as Ordinary Differential Equations (ODEs) [25], which have now progressed to a rule-based approach due to their ease and succinctness of expression. Rule-based methods have several applications some of which are assessing

the druggability of proteins [26] and drug effect pathway analysis [19]. These will be dealt in greater detail, in the background section.

This project is aimed at creating a database for Protein-Protein interactions stored in the Kappa rule format [14]. The database would allow the PPI interactions to be retrieved based on certain conditions that are elaborated in Chapter 3. Rule based simulation of protein interaction can either be performed based on Stochastic Simulation Algorithm (SSA) or using Ordinary differential equations [10]. In this project feeding the Kappa rules, to a Kappa simulator will help in visualizing the interactions of protein molecules in the Kappa simulator(KaSim) [15]. KaSim is an implementation, of an algorithm called continuous time Monte-Carlo (CTMC), which is created for systems based on rules.

This work is divided into chapters and we present a brief summary of each of these chapters. Chapter 2 covers, the kinds of PPI database that exist in literature, followed by a description of the kappa rules which encapsulates their syntax and semantics. The application of rule-based methods are also further elaborated in Chapter 2. Chapter 3 covers, the work that has been undertaken. This section elucidates the methodology used to create the database, the python scripts used to extract the relevant information from the assimilation of data collected by researchers. This section also elaborates on the SQL stored procedures used to extract the PPI rules and the user interface for accessing those rules. Furthermore, this chapter elaborates on the steps for deploying the database scripts and the web application. This chapter also walks through the process of adding a new rule to the database, to enable future PPI additions. In Chapter 4, the validation pipeline for the data within the database is defined concisely. In this chapter we retrieve some of the PPI rules and validate the result set with the provided data. Chapter 5 presents the conclusion with future improvements and proposal for work that can be extended from the project.

Chapter 2

Background

Protein-Protein interactions play a vital role in the regular functioning of life processes. Hence the study of these interactions, plays a crucial role in improving our understanding of diseases and the life processes.

Historically, PPI interactions were modeled using ordinary differential equations (ODE) [25]. However, as per [9], this traditional approach of modeling PPI through ODE had several limitations due to the following reasons.

- The protein molecules can potentially interact in an exponential number of ways.
- Due to the exponential number of possibilities only large reaction networks can capture them.
- This is a problem because traditional approaches like ODE require explicit network specification.

This problem is overcome by the use of local rules where the network specification is implicit. As a result the specification of the model is concise. These rules can be specified using languages for model specification like Kappa [14] and BioNetGen[16]. Specialized software tools then enable researchers to visualize the PPI interactions and run the simulation of the model in stochastic or deterministic way.

In the following sub-sections we will first explore the kinds of Protein interaction databases, followed by a section that develops on the understanding of rule based specification of protein interactions. The latter is followed by a section that deals with the motivation of constructing the rule based database system and application of protein interaction database in biology.

2.1 Protein-Protein Interaction (PPI) Databases

As per [32], the kinds of protein-protein interaction databases can be divided into three types.

- **Pathway:** In such databases researchers and domain experts collect pathway information that are generally agreed upon by the scientific community. This information is manually curated and cover association with diseases, stoichiometry of reactions and so on. Due to the requirement of manual intervention and the aim to achieve a high accuracy, construction of such databases is a laborious process.
  Examples of such databases include KEGG [22] and Reactome [11].

- **Experimentally Verified:** Such databases contain an assimilation of the protein interaction rules that have been experimentally verified. In other words such databases contain experimentally observed (verified) PPI rules. The method of the rule organization and the amount of information carried varies from one database to another.
  Examples of such databases include IntAct [23] and BioGrid [30].

- **Experimentally Verified or Computationally predicted:** Such databases contain PPI rules that are either experimentally observed or are computationally predicted. These PPIs however, assume minor manual curation. Thus, the computationally obtained PPIs may contain false positives and hence, to improve the accuracy a confidence score is often attached to them. In addition to using computational methods to obtain PPI rules, Natural Language Processing and text mining methods are also used in order to extract PPI rules from research literature.
  Examples of such databases include STRING [31] and GeneMANIA [33].

While the three types of databases mentioned above, serve as the primary categorization of PPI databases, there also exists categorization of PPI databases based on diseases, organisms of particular kind and so on.

According to [32], there are over hundreds of databases that aim to collect and store protein interactions. However, none of these databases capture the complexity of biological systems in it's entirety. These kinds of details include - temporal dependencies, spatial dependencies, protein isoforms and so on.

2.2 Motivation for constructing the rule based database

As per [7], historically the protein interactions were used to simulte the association of the protein into the complexes through the method of deterministic models , based on differential equations. These models aimed to capture a myriad of information like post-translational, structural information an so on. Such a vast domain of knowledge was susceptible to errors or missing values, which cast a doubt upon their accuracy, as per [7]. Such models also had maintainability issues because besides being difficult to build, they were also difficult to be updated with the ever evolving knowledge base that dictate these protein interactions. The human understanding of protein interactions matures and evolves with time and further research. Hence it is imperative to have a method of expressing PPI rules and building PPI models that make it easy to be read, stored and retrieved. As per [7], kappa rules, used to express the PPI interactions have the ability to encode and express information about the protein molecules (agents) and the sites that take part in the reaction. These rule summarize the pre and post conditions of the protein interaction without venturing into a detailed version of their structural analysis. Hence the rule based format of expressing PPI interaction is more concise, the stochastic models built using Kappa rules enable the description of large protein complexes.

With this motivation in mind we have ventured to create a database of Protein-Protein interactions stored in a rule based format. The database created also has stored procedure routines that enable extraction of PPI rules based on certain conditions like agent name and domain name. The stored procedures are accessible via a web application built in django, with a user interface (UI), that enables the extracted rules to be printed or exported as CSV or Excel files. These files can be further analyzed and processed, fed to the kappa simulator [15] and generate visualizations of the protein interactions.

2.3 Understanding rule based specification of protein interactions

2.3.1 General Understanding of rule based languages

In rule based languages like kappa [14] and BioNetGen [16], the agent is a conceptualization of the protein molecule. The protein molecules (agents) connect to form site

graphs via the protein sites. The PPI rule as per the rule based format consists of a left hand denoted by L_r side and a right hand side denoted by R_r. The L_r and R_r contain the site graphs which mention only the necessary sites for the protein interaction $L_r \rightarrow R_r$.

Furthermore, $\mathfrak{M}$ denotes the state of a system which is also called the reaction mixture. Each disconnected graph denotes one molecular species and the state of the system $\mathfrak{M}$, is a collection of disconnected graphs. The execution of a certain rule 'r' implies the replacement of the mixture matched to L_r, with R_r, as shown in the Figure 2.1. A model as per [7] a collection of rules. Reasoning at the level of rules helps to introduce a level of compactness essential for the succinct expression of the protein interaction.

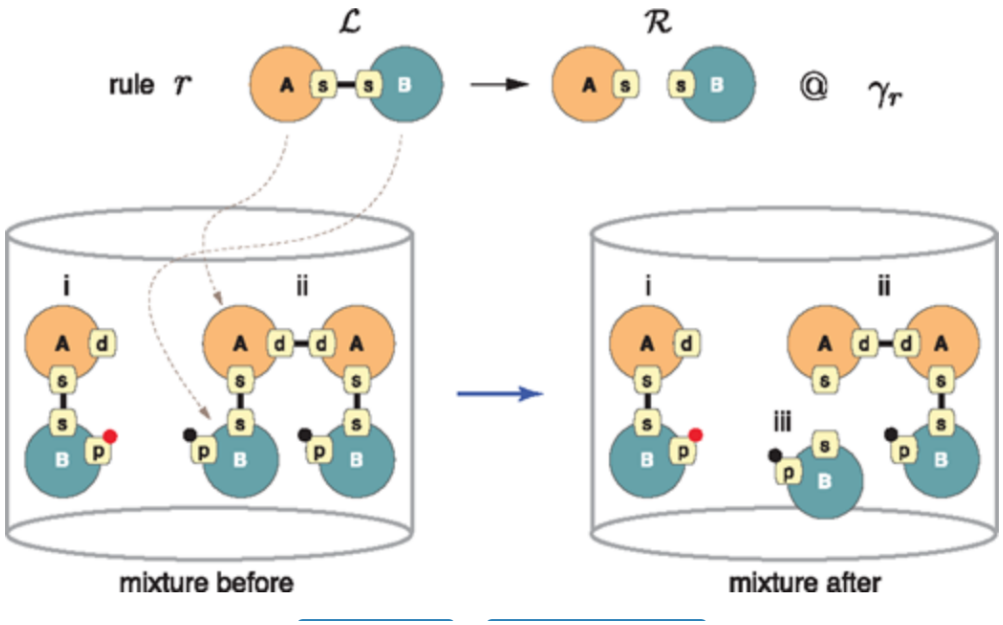

Figure 2.1: Application of rule 'r' to a reaction mixture $\mathfrak{M}$, comprising of two agents. L_r and R_r represent the state of the site graph before and after the application of rule 'r' respectively. Source: [7]

Modeling tools built using latest technology help to visualize the protein interactions. Some examples of this include Virtual Cell [28] and Kappa Simulator KaSim [15], which provide an environment for the modeling and simulation of cellular interactions. The Kappa rules can be fed into KaSim, which is a protein interaction simulator that implements continuous time MonteCarlo algorithm (CTMC). Similarly, the rules created in BioNetGet [16] format can be simulated with NFsim [29]. KaSim and NFsim employ the process of stochastic modeling that can be used to describe highly complex interactions. Besides them there also exist other stochastic modeling

tools like STOCHSIM [24], which can be used by researchers, to compare the results.

In addition to KaSim, the Kappa simulator, there are tools like KaSA (Kappa Static Analyzer) and KaSTOR (Kappa Story Extractor). As per [7], KaSA helps in analyzing the static properties of the model that can help in debugging, as it can efficiently detect discrepancies between actual and expected behavior. KaSA relies on a technique called 'abstract interpretation' described in [7] to achieve this task. KaSTOR helps by providing an insight as to how an event of interest EOI was obtained in a particular event trace. EOI is obtained by the simulation of a model. KaSTOR works on the concept of mechanistic causality in molecular systems. KaSim generates an output called 'trace' which is a sequence of events generated as a result of the simulation. Algorithms such as the one stated in [12] aid this process. This is dealt in greater detail, in the Analysis of KaSim outputs, subsection.

2.3.2 Syntax and semantics of Kappa rules

Kappa rules are formally documented well. Their specification and syntax can be found in [15] and [1]. In rule based specifications graphs are formally specified as objects which have been converted to a notation of textual form, for convenience. As per these rules an agent site denoted by 's', has a binding state 'n' can be specified as s[n]. Here 'n' can be any positive integer or may be replaced by a '.', which indicates the site is not bound. In case 'n' is a positive integer then it implies that the site is bound to another unique site having the same binding site 'n', within the same PPI expression. Hence, a sub expression like, (Agent1 [site1 [1]], Agent2 [site2 [1]]) indicates that Agent1 is bound to Agent2 through site1 of Agent1 and site2 of Agent2. The protein sites may also have their own internal states which is specified within curly brackets '{}'. Hence Agent1(site1{p}[.]), implies that Agent1 has an unbound site name site1 that is in an internal phosphorylated state (denoted by 'p').

Figure 2.2, depicts the process of obtaining a Protein interaction rule from the textual information of research literature. These rules when passed to KaSim, the Kappa simulator help to visualize the PPI interactions.

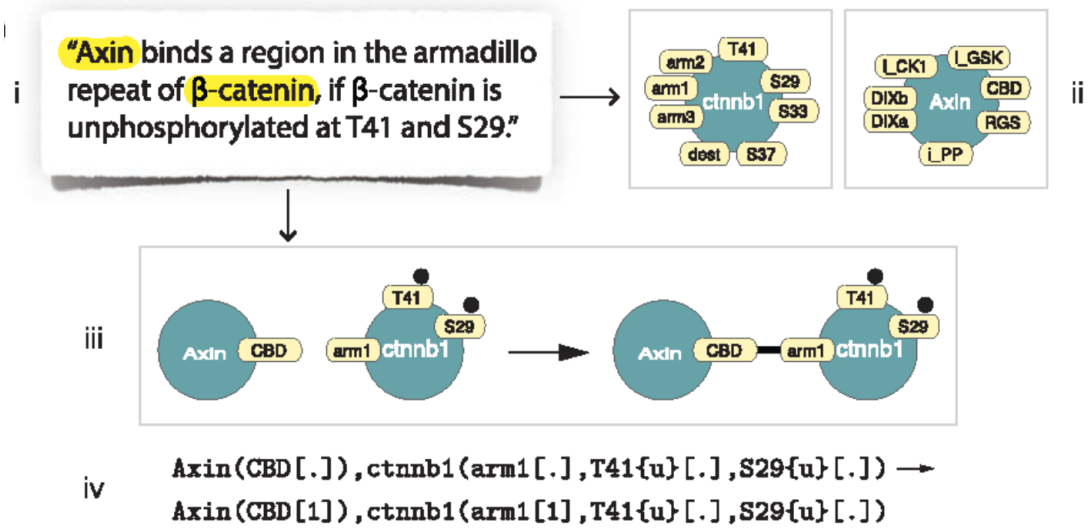

Figure 2.2: Depiction of obtaining a Protein-Protein interaction (PPI) rule from research literature. Source: [7]

2.3.3 Analysis of KaSim outputs

As per [7], KaSim generates as it's output the whole sequence of events called trace. The analysis of KaSim output includes discovery of the path by which a particular event of interest (EOI) has occurred. The problem can be formally stated as follows: Provided a trace of events denoted by τ, where $\tau = e_1, e_2, ..e_n$ and e_n denotes an individual event. The problem is to find a suitable explanation for an EOI, where the EOI is an event, $e_n \in \tau = e_1, e_2, ..e_n$. This process is called causal analysis and in Kappa platform KaSTOR, the software agent is utilized for this purpose. The working of this software and the process of causal analysis, is elucidated using the following figure, 2.3

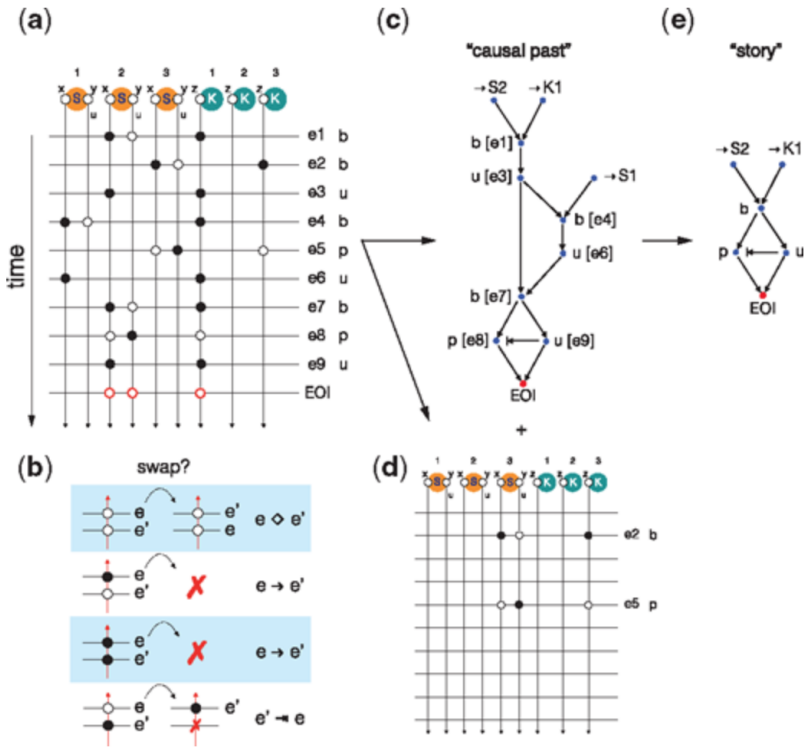

Figure 2.3: Explanation of causal analysis. Source: [7]

In figure 2.3 (a), the vertical lines from each site, of every agent within the mixture denotes a thread. This thread contains the record whether an event tested or modified the state of a site. It should be noted that in kappa terminology a modification of site also implies a test. The events are shown with the horizontal markings. The black discs denote that the thread was modified by its corresponding horizontal event and the white circle denotes that the thread was tested.

As per [7], causality comprises of a relation between events. The algorithm begins by reconstructing the causal past of '*', the identity rule. The figure 2.3 (b) depicts an event 'e' that is followed by event ' e′ '. and iteratively quries if the modification or test of the site in e′ could have occurred before 'e'. There are certain rules for the query which are detailed in [7], under causal linages and compression. This procedure results in a directed acyclic graph (DAG) as shown in 2.3 (c) which represents the precedence structure of the causal past. The procedure depicted in 2.3 (b) also eliminates from within the trace, any events irrelevant to the EOI. This is shown in figure 2.3 (d). The problem with obtaining the precedence structure is that it may violate a property called 'necessity'. A method to tackle this problem involves an algorithm called minimization causal compression elaborated in [12]. KasTOR translates the causal past into a boolean expression where each event is associated with a Boolean variable. The pro-

cessing within KasTOR results to a compressed form of the causal past as shown in the figure 2.3 (e).

2.4 Application of Protein-Protein Rule Interaction Database

The databse for PPIs stored in a rule based format will have several applications. Some of these are covered in this section.

- **Building the rule based model in a Kappa format:** The database will allow to build the executable rule based model based on the proteins of interest. For that, the list of proteins submitted to the database would result in a list of the relevant rule PPIs retrived as a csv file. Those rules could be built into Kappa model and simulated by KaSIM to obtain the dynamics of the molecular complex of the interest. This would allow the following more specific applications.

- **Assessment of protein druggability:** As per [27], ruled based PPI stored in a database are useful in the development of a rule-based model for the assessment of protein druggability. Selection of target is an important step for the discovery of drugs. The effectiveness of a drug target depends upon factors like its chemical influence and biological importance. Druggability is defined as the ability of a target to bind a drug-like molecule with an affinity that is at a therapeutically useful level. [27] develops a set of simple rules that govern the process of durggability which can be applied in the future to get an idea of the chemical influence of prospective targets. These rules are based on the property space of druggable pockets like volume, depth and so on.

- **Drug effect pathway analysis:** Pathway analysis involves the study of specific molecular pathways using formalism that are qualitative and quantitative. This domain provides a tremendous computational complexity and experimental approaches that incur a high cost. Research conducted in [20], included extraction of about 200 rules, related to type 2 diabetes obtained from varied sources like literature, pathway databases and conversion from different kinds of models. Using these rules [20] a multi-scale rule-based modeling platform was constructed. This modeling platform was then used for simulation of drug effect pathways of type 2 diabetes drugs and check for it's efficacy. The research concluded that their simulation helped in identifying effective drug combinations and provide a new way to effectively apply existing drugs for new targets.

- **Application of rule based models in biology:** Rule based models can be applied to describe the dynamics of population in a predator-prey ecosystem and simulate rhythm changes that are circadian ([8]).

All these research applications help in substantiating the claim that database for storing PPI in a rule format is an important tool for the study the protein complexes, drug discovery and pathway analysis aimed to improve our understanding of complex biological systems.

Chapter 3

Methodology

This chapter includes the methodology adopted for the creation of rule-based Protein-Protein Interaction (PPI) database and the steps to add a new rule to it. It also contains a description of the stored procedures used to extract protein interaction rules based on either agent name or domain name pair. The source code for this project is available on an online repository (Github). This chapter contains a description of the files and folder structure within it.

As a part of this project, a web application using Django [2] was created to access the PPI rules by calling the stored procedures from within it. Chapter 3 includes the design and implementation details for the web application used to access the PPI rules. The PPI rules can be accessed from the web application, either based on the agent name or a domain name pair.

This chapter also includes the steps for deploying the database scripts to a database server and the web application to a web server. During the process of deployment, it is often useful to have the software versions that were initially used to create the software. These are included in Appendix A, under the Software Version section. Having these details will enable developers to debug when deploying the application.

3.1 Methodology for creating the database

PPI rules were collected and provided by researchers in the form of an Excel Sheet (rule_baseV0.2.xlsx). To extract the agent , domain names and rules from the dataset, excel sheet parser scripts were created in Python. Python scripts were also written to create and validate entity relations within the database. Description of these scripts are elucidated in Section: Description of the Project Folder and Files.

To construct the PPI database it was decided to structure the data into separate tables consisting of the agent, domain and the PPI rules. In order to connect the data in these three main tables, 2 relationship tables consisting of domain_agent and rule_domain_agent were created.

It should be noted that main tables contain certain 'meta data columns'. These meta data column help in keeping track of the associated data for an entity within the table. For example, the domain table has agent_name column, which keeps a record of all the agent_names that it associates to. However, the correct procedure to obtain the agent_names that associate with a particular domain is to use the domain_agent relationship table. The Entity Relationship (ER) diagram of the database and the description of the database tables are presented in the following sections.

3.1.1 Entity Relationship Diagram:

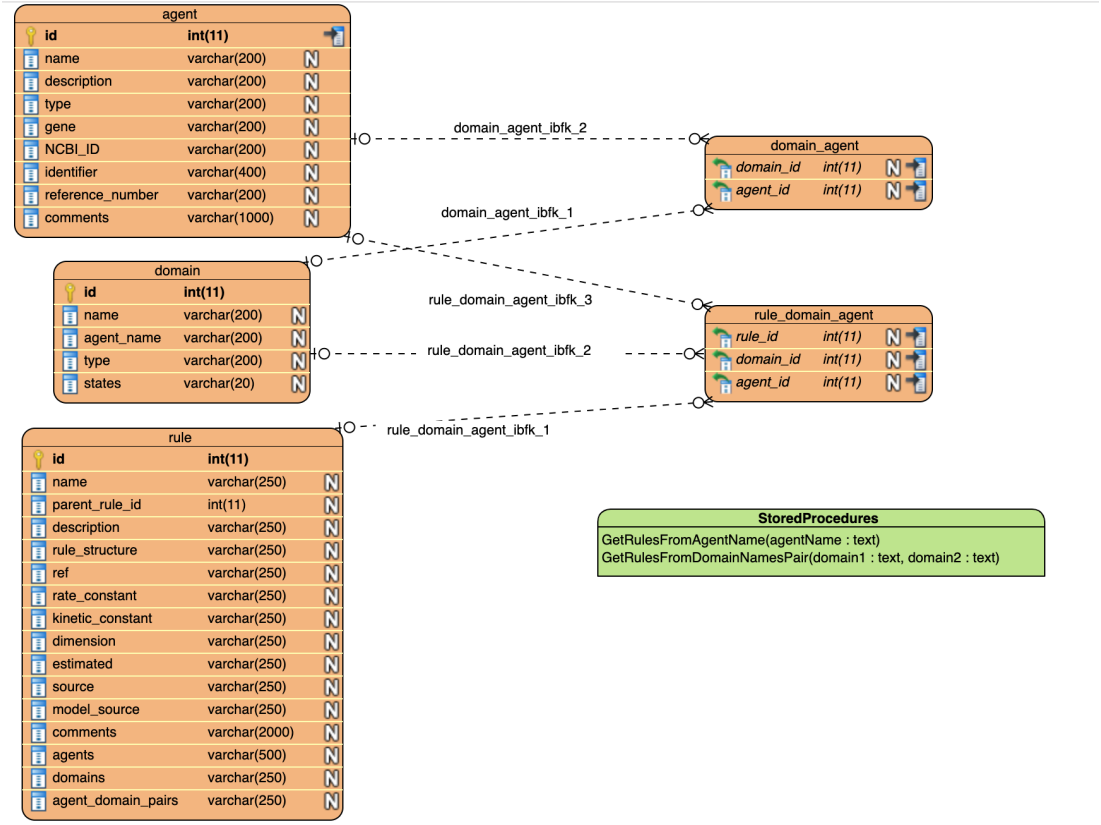

Figure 3.1: Database ER Diagram created using software Visual Paradigm [3]

In Fig 3.1, the dotted lines depict the primary-foreign key relations between the tables. The green box depicts the names of the stored procedures that are part of the database.

3.1.2 Description of main tables

1. **agent:-** Provides a description of the protein molecule (the agent). This table contains the following fields: id, name, description, type, gene, NCBI_ID, identifier, reference_number and comments.

   (a) id: The unique identifier for the agent molecule.

   (b) name: The name of the agent.

   (c) description: A description of the agent.

   (d) type: The type/family of the agent.

   (e) gene: The particular gene

   (f) NCBI_ID, identifier, reference_number: Provided within the dataset and contains references to the Uniprot Prosite Database [4] and literature.

   (g) comments: Contains additional comments about the agent.

2. **domain:-** Provides a description of the domains that associate with the agents. This table contains the following fields: id, name, agent_name, type, states.

   (a) id: The unique identifier for the domain name.

   (b) name: The name of the domain.

   (c) agent_name: A 'meta data column' that contains the names of agent molecules that associate with the particular domain.

   (d) type: The type of the domain.

   (e) states: Contains the possible states of the domain.

3. **rule:-** Provides a description of the PPI rules within the database. This table contains the following fields: id, name, parent_rule_id, description, rule_structure, ref, rate_constant, kinetic_constant, dimension, estimated, source, model_source, comments, agents, domains and agent_domain_pairs.

   (a) id: The Unique identifier for the rule

   (b) name: The name of the rule

   (c) parent_rule_id: In some cases the rules can be traced to a parent. This field stored the id of that parent rule. By default it is 0, which signifies no parent rule.

(d) description: A description of the rule.

(e) rule_structure: The structure of the rule.

(f) ref, source: A reference to the rule.

(g) rate_constant: The rate constant for the reaction.

(h) kinetic_constant: The kinetic constant may contain numerical value or a formula to which the rate_constant can be plugged to obtain the result.

(i) dimension: The dimension of the reaction.

(j) estimated: The estimated value of the reaction.

(k) model_source: A reference to the model from which the PPI is obtained.

(l) comments: Any additional comments about the rule.

(m) agents: A 'meta data column' containing the names of the agent molecules that take part in the rule.

(n) domains:A 'meta data column' containing the names of the domains that take part in the rule.

(o) agent_domain_pairs: A 'meta data column' containing the names agent and associated domains that form a part of the protein interaction.

3.1.3 Description of relationship tables

1. **domain_agent:-** Each row contains the id of a domain and the id of an agent, obtained by referencing the domain and agent table respectively. By obtaining the domain id from the domain table, the corresponding agent ids that associate with the domain can be obtained. This can then be used to obtain the agent names that associate with that particular domain.

   (a) domain_id: The Id of a domain (from domain table).

   (b) agent_id: The Id of an agent (from agent table).

2. **rule_domain_agent:-** Each row contains the id of a rule, id of a domain and the id of an agent, obtained by referencing the rule, domain and agent table respectively. Hence, this table connects the agents and domains to the PPI rules from the rule table in the database.

   (a) rule_id: The Id of a rule (from rule table).

(b) domain_id: The Id of a domain (from domain table).

(c) agent_id: The Id of an agent (from agent table).

3.1.4 Advantages and disadvantages of the database structure

3.1.4.1 Advantages

The database structure is optimized for fast read operations. This is because, by using the id structure to uniquely identify entities in the database, we can leverage the power of indexing [5], which can be used to obtain the results faster. As per [5], in section 8.3.1, most MySQL indexes like primary key, unique key are stored in B Trees. Due to this, it is possible to obtain rows matching the where clause quickly and even leverage the power of Multiple-Column Indexes, as per [5] section 8.3.6. It was also verified in MySQL workbench that the relationship tables (domain_agent and rule_domain_agent) had each of the columns indexed as a B Tree. Hence MySQL would internally be able to work more efficiently with this structure than we if we employed a method that included string regex based searching.

3.1.4.2 Disadvantages

The disadvantage of this database design is that it makes the write operations a multi-step process. For example, if we wanted to add a new rule to the database then besides performing the regular checks, several entries may have to be made in different tables within the database. This is however not a major disadvantage as the database is expected to have many more read operations than write, also the next section states the steps to add a new PPI rule to the database. While trying to make a new entry within the database, following these rules will help in reducing the errors.

3.2 Steps for adding new entries to the database

3.2.1 Agent

Query the agent table and check if an agent exists with the given name. It should be noted that the agent name is case sensitive. In case it exists then there is no need to add a new entry to the database. However, if it does not exist then a new entry can be made in agent table filling in the relevant fields.

3.2.2 Domain

Query the domain table and check if a domain exists with the given name. It should be noted that the domain name is case sensitive. In case it exists then there is no need to add a new entry to the database. However, if it does not exist then a new entry can be made in domain table filling in the relevant fields.

3.2.3 Domain-Agent Relationship

When the Agent/Domain is added to the database to connect them we have to make entries within the domain_agent table. Based on the entries made in the Agent and Domain table we can obtain the corresponding ids and make entries containing the (domain id, agent id) pairs. If a new agent is added then all its corresponding domains need to be entered as separate rows where the agent id will remain the same. If a new domain is added then all its corresponding agents need to be added as separate rows where the domain id will remain the same. It should be kept in mind that if a new agent is added to the agent table and it has a new corresponding domain added to the domain table, then that pair must be entered only once within the domain_agent table.

3.2.4 Rule

To add a new rule to the database, first check if the rule already exists within the database. If not, then obtain the agents and domains that take part in the rule. Verify that the necessary agents, domains and domain_agent entries exist within the database. If not, then first make the necessary entries in these three tables as per guidelines in the previous sections. Then add the PPI rule to the rule table in the database.

3.2.5 Rule-Domain-Agent

After following the steps laid out in Rule subsection, the three pair tuple/s consisting of the rule_id, agent_id for each of the agents in the rule and the domain_id for each of the domains that associate with the corresponding agent within rule, need to be entered in the rule_domain_agent table as (rule id, domain id, agent id).

3.3 Description of Stored Procedures

In Fig 3.1, the green box displays stored procedures defined within the database. The two stored procedures created are GetRulesFromAgentName and GetRulesFromDomainNamePair. The expected input, output and algorithm for each of the stored procedure is mentioned in the following subsections.

3.3.1 GetRulesFromAgentName

Input: An agent name (data type: text).
Output: PPI rules in which the agent name, provided as an input occurs.
Algorithm

1. The agent id corresponding to the agent name is obtained from the agent table.
2. A list of the rules, in the form of rule ids occurring along with the agent id (obtained in step 1) is extracted from the rule_domain_agent table. This is a list of distinct rule ids, which implies that none of the rule ids are repeated in the output.
3. Using the list of distinct rule ids, the rules are extracted from the rule table and provided as output of the stored procedure.

3.3.2 GetRulesFromDomainNamePair

Input: Two domain names (data type of each being: text).
Output: PPI rules in which both the domain names, provided as an input occurs.
Algorithm

1. The domain id for each of the domains are extracted from the domain table and stored in two separate variables within the stored procedure, called domain1 and domain2.
2. From the table, rule_domain_agent those rule ids are extracted that co-occur with the domain2.
3. The rule ids are extracted which co-occur with domain1 and also occur in the list of rule ids extracted in the previous step. Hence implying that the rules ids occuring in this step are the rules that co-occur with domain1 and domain2. This is a distinct list of rule ids.

4. Using the list of distinct rule ids, the rules are extracted from the rule table and provided as output of the stored procedure.

3.4 Description of the project folder and its files

3.4.1 Project folder source

This project is hosted on Github, as a public repository in the link [6], and is open source.

3.4.2 Project folder structure and files

The main folder is called PPI_DB, which contains within it the following sub-folders:

1. django-app: Contains files pertaining to the web application for accessing the PPI rules. The relevant files and folder structure are elucidated in the next section.

2. Parser: Contains files pertaining to extracting the agents, domains and PPI rules from the dataset (rule_baseV0.2.xlsx). This folder has the following files and sub-folder.

   (a) PARSER_POPULATE_TABLE_AGENT.py: Accesses the Agents Sheet of the dataset and generates the table insertion script for the SQL table, agent.

   (b) PARSER_POPULATE_TABLE_DOMAIN.py: Accesses the Agents Sheet of the dataset and generates the table insertion script for the SQL table, domain.

   (c) PARSER_POPULATE_TABLE_RULE.py: Accesses the Rules Sheet of the dataset and generates the table insertion script for the SQL table, rule.

   (d) Extractors: This folder contains python scripts used to obtain the agents, domains, agent-domain pairs and kinetic constants for each of the PPI rules, from the Rules Sheet of the dataset. These extracted columns are then appended to the Rules Sheet of the dataset, as additional columns and included into the SQL rule table. The ‘meta data columns’ like agents, domain and agent-domain pairs provide additional information about the rules.

(e) DB_POPULATE_TABLE_DOMAIN_AGENT.py: This script connects to the database using the credentials mentioned within the script and generates the table insertion script for the SQL table, domain_agent.

To summarize, this script works by first obtaining a row from the domain table created in step (b). The agent_name field in the row gives names of all the agents to which the domain associates itself with. The id of each of these agents is obtained from the agents table and the id of the domain is included in the row extracted from the domain table. Then insert table commands of the form (domain_id, agent_id) are generated, to populate the domain_agent table.

(f) DB_POPULATE_TABLE_RULE_DOMAIN_AGENT.py: This script connects to the database using the credentials mentioned within the script and generates the table insertion script for the SQL table, rule_domain_agent.

To summarize, this script works by first obtaining a row from the rule table created in step (c). The agent_domain_pairs field in the row gives all the agent-domain pairs within that PPI rule. The id of each of these agents is obtained from the agents table and the id of the domain is obtained from the domain table (both obtained by querying from the database). Then insert table commands of the form (rule_id, domain_id, agent_id) are generated, to populate the rule_domain_agent table.

3. Stored Procedure: Contains two files that define stored procedures.

   (a) RULES_FROM_AGENT_NAME.sql: Defines the stored procedure that returns the PPI rules based on the agent name as input.

   (b) RULES_FROM_DOMAIN_NAME_PAIR.sql: Defines the stored procedure that returns the PPI rules based on the pair of domain names as input.

4. Table Populate Scripts: Contains database scripts that are used to populate the PPI database. These table insertion scripts are written in MySQL format and tested using the MySQL client called MySQLWorkbench. It contains the following files:

   (a) AGENT.sql: Script to populate the agent table.

   (b) DOMAIN.sql: An initial Script to populate the domain table. (included in the projects file only for reference)

(c) DOMAIN_V2.sql: Script to populate the domain table.

(d) RULE.sql: Script to populate the rule table.

(e) DOMAIN_AGENT.sql: Script to populate the domain_agent table.

(f) RULE_DOMAIN_AGENT.sql: Script to populate the rule_domain_agent table.

5. Tables/Database Creation Scripts: Contains the database scripts used to define and create the database and tables within it. These scripts are written in MySQL format and tested using the MySQL client called MySQLWorkbench. It contains the following files:

   (a) CREATE_DATABASE_PPI.sql: Script to define and create the PPI Database. This serves like a container to store the PPI database tables.

   (b) CREATE_TABLE_AGENT.sql: Script to define and create the agent table.

   (c) CREATE_TABLE_DOMAIN.sql: Script to define and create the domain table.

   (d) CREATE_TABLE_RULE.sql: Script to define and create the rule table.

   (e) CREATE_TABLE_DOMAIN_AGENT.sql: Script to define and create the domain_agent table.

   (f) CREATE_TABLE_RULE_DOMAIN_AGENT.sql: Script to define and create the rule_domain_agent table.

6. Rectified Rules: Verification steps were adopted after the initial population of the database. A report was generated based on any discrepancies within the data in the database. For example, it is possible that while inserting a rule it is observed that the domain that the agent occurs with was not entered in the database before as it was not observed in the Agents sheet, in the dataset. Based on the report and subsequent discussions the table population scripts were updated. Moe information regarding the verification pipeline is presented in the 'Results and Discussion Section'. This folder contains the following two files:

   (a) problematic_agents_domains.txt: The initial report on the the observed discrepancies.

   (b) rule_baseV0.2_fixed.xlsx: The dataset with the updated values for agents, domains and rules based on the discrepancies reported in

problematic_agents_domains.txt. It is an updated version of the dataset contained in the Parser folder.

3.5 Web Application for accessing the stored procedures

3.6 Deployment steps

3.6.1 Database Scripts

3.6.2 Web Application

Chapter 4

Results and Discussion

Chapter 5

Conclusions

5.1 Final Reminder

The body of your dissertation, before the references and any appendices, *must* finish by page 40. The introduction, after preliminary material, should have started on page 1.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default 1.5 spacing). Over length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

Bibliography

[1] https://kappalanguage.org/documentation.

[2] https://docs.djangoproject.com/en/2.2/.

[3] https://www.visual-paradigm.com/.

[4] https://www.uniprot.org/database/DB-0084.

[5] https://dev.mysql.com/doc/refman/8.0/en/optimization-indexes.html.

[6] https://github.com/ayushdas/PPI_DB.

[7] Pierre Boutillier, Mutaamba Maasha, Xing Li, Hctor F Medina-Abarca, Jean Krivine, Jrme Feret, Ioana Cristescu, Angus G Forbes, and Walter Fontana. The Kappa platform for rule-based modeling. *Bioinformatics*, 34(13):i583–i592, 06 2018.

[8] Álvaro Bustos, Ignacio Fuenzalida, Rodrigo Santibáñez, Tomás Pérez-Acle, and Alberto JM Martin. Rule-based models and applications in biology. In *Computational Cell Biology*, pages 3–32. Springer, 2018.

[9] Lily A. Chylek, Leonard A. Harris, Chang-Shung Tung, James R. Faeder, Carlos F. Lopez, and William S. Hlavacek. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdiscip Rev Syst Biol Med*, 6(1): 1336, 2014.

[10] Lily A Chylek, Leonard A Harris, Chang-Shung Tung, James R Faeder, Carlos F Lopez, and William S Hlavacek. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 6(1):13–36, 2014.

[11] David Croft, Gavin OKelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697, 2010.

[12] Vincent Danos, Jérôme Feret, Walter Fontana, Russell Harmer, Jonathan Hayman, Jean Krivine, Chris Thompson-Walsh, and Glynn Winskel. Graphs, rewriting and pathway reconstruction for rule-based models. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[13] Javier De Las Rivas and Celia Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, 2010.

[14] Danos V. et al. (2007a) rule-based modelling of cellular signalling. *In: Proceedings of the Eighteenth International Conference on Concurrency Theory, CONCUR 2007*, Vol. 4703 of Lecture Notes in Computer Science. Lisbon, Portugal, Springer-Verlag Berlin Heidelberg, pp. 1741., 2007.

[15] Danos V. et al. (2007b) scalable simulation of cellular signaling networks. In *Asian Symposium on Programming Languages and Systems*, pages 139–157. Springer, 2007.

[16] Faeder J.R. et al. Rule-based modeling of biochemical systems with bionetgen. *In: Maly I.V. (ed.) Methods in Molecular Biology*, Systems Biology , Vol. 500, Springer-Verlag Berlin Heidelberg, Springer, pp. 113167, 2009.

[17] David C Fry. Targeting protein-protein interactions for drug discovery. In *Protein-Protein Interactions*, pages 93–106. Springer, 2015.

[18] Alexander Goncearenco, Minghui Li, Franco L Simonetti, Benjamin A Shoemaker, and Anna R Panchenko. Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. In *Proteomics for Drug Discovery*, pages 221–236. Springer, 2017.

[19] Woochang Hwang, Yongdeuk Hwang, Sunjae Lee, and Doheon Lee. Rule-based multi-scale simulation for drug effect pathway analysis. In *BMC medical informatics and decision making*, volume 13, page S4. BioMed Central, 2013.

[20] Woochang Hwang, Yongdeuk Hwang, Sunjae Lee, and Doheon Lee. Rule-based multi-scale simulation for drug effect pathway analysis. In *BMC medical informatics and decision making*, volume 13, page S4. BioMed Central, 2013.

[21] Eric Jain, Amos Bairoch, Severine Duvaud, Isabelle Phan, Nicole Redaschi, Baris E Suzek, Maria J Martin, Peter McGarvey, and Elisabeth Gasteiger. Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC bioinformatics*, 10(1):136, 2009.

[22] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl_1):D355–D360, 2009.

[23] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. The intact molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846, 2011.

[24] Nicolas Le Novere and Thomas Simon Shimizu. Stochsim: modelling of stochastic biomolecular processes. *Bioinformatics*, 17(6):575–576, 2001.

[25] Le Novere N and Endler L. Using chemical kinetics to model biochemical pathways. *Methods Mol Biol.*, 1021:147-67., 2013.

[26] Emanuele Perola, Lee Herman, and Jonathan Weiss. Development of a rule-based method for the assessment of protein druggability. *Journal of chemical information and modeling*, 52(4):1027–1038, 2012.

[27] Emanuele Perola, Lee Herman, and Jonathan Weiss. Development of a rule-based method for the assessment of protein druggability. *Journal of chemical information and modeling*, 52(4):1027–1038, 2012.

[28] Boris M Slepchenko and Leslie M Loew. Use of virtual cell in studies of cellular dynamics. In *International review of cell and molecular biology*, volume 283, pages 1–56. Elsevier, 2010.

[29] Michael W Sneddon, James R Faeder, and Thierry Emonet. Efficient modeling, simulation and coarse-graining of biological complexity with nfsim. *Nature methods*, 8(2):177, 2011.

[30] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, et al. The biogrid interaction database: 2011 update. *Nucleic acids research*, 39(suppl_1):D698–D704, 2010.

[31] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1):D561–D568, 2010.

[32] Damian Szklarczyk and Lars Juhl Jensen. Protein-protein interaction databases. In *Protein-Protein Interactions*, pages 39–56. Springer, 2015.

[33] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220, 2010.

Appendix A

Software Version

This section includes software versions that were used to create and deploy the application onto a local machine. While deploying to a database/web server it may be helpful to have these details.

A.1 Operating System

A.2 MySQL Version/ Client Version

A.3 Python Version

A.4 MySQL Connector Version in Python scripts

A.5 Django Version

A.6 Data Tables Version