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) [11], 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 [1]. 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 [5] or BioNetGen [6]. 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

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 [10]. 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 [3] 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 [7] an important source for drug discovery is the study of PPI. This is also evident from the fact that, as per [8] 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 application.

An important step that would help researchers to study protein interactions is to create a repository of the PPI rules. The design of the PPI repository should be such that it would be easily retrievable by the researchers based on the relevant search term. This motivates the creation of a Protein-Protein interaction (PPI) database. There have been several databases in the past that have tackled the problem of collecting the PPI rules. 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. The database created in this project, is a novel approach at assimilating the relationship between molecules of proteins based on Kappa rules.

PPI interactions were initially modeled as Ordinary Differential Equations (ODEs) [11], which have now progressed to a rule-based approach due to their ease and succinctness of expression. This project is aimed at creating a database for Protein-Protein interactions stored in the Kappa rule format [5]. 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 [2]. 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) [4]. KaSim is an implementation, of an algorithm called continuous time Monte-Carlo (CTMC), which is created for systems based on rules.

Rule-based methods have several applications some of which are assessing the druggability of proteins [12] and drug effect pathway analysis [9]. These will be dealt in greater detail, in the background section.

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. 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

Chapter 3

Method

A dissertation usually contains several chapters.

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] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.