SAPIENZA
UNIVERSITÀ DI ROMA

AI-driven analysis of molecular pathways

Facoltà di Ingegneria dell'informazione, informatica e statistica
Informatica

Ionuț Cicio
ID number 2048752

Advisor	Co-Advisor
Prof. Mancini	Prof. Tronci

Academic Year 2024/2025

AI-driven analysis of molecular pathways
Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author’s email: cicio.2048752@studenti.uniroma1.it

Abstract

Contents

Chapter 1

Introduction

Biological systems such as organisms, cells, or biomolecules are highly organized in their structure and function, [2]. [Such structures can help study the and compare biological functions of different organisms]. [In order to understand, formalize and abstract these structures some kind of modeling is needed, be it mathematical or not (what not?)].

In this report is presented a proof of concept which generates biological models by using the concept of reachability and databases of biochemical reactions.

[A fundamental component of bioinformatics is data integration, (a problem is) i.e. partial information in distributed databases is needed]. [One such data source is Reactome [1], which is a qualitative network database] REACTOME is an open-source, open access, manually curated and peer-reviewed pathway database. [The goal with this work is to use qualitative data in the Reactome database to generate quantitative models, and use BBO (cite something) techinques to "validate" (validate is not good, find something else) these models]

[Validating complex biological models is a computationally intensive task (cite something) thus HPC clusters are required in order.] Such clusters may not always be fully available (i.e. there are multiple users and multiple experiments running on the cluster + there are some limits on job running times), so in order to better distribute the load of the validation task some infrastructure work is needed.

1.1 Preliminaries

The set of non-zero natural numbers is denoted with $\mathbb{N}_+ = \mathbb{N} - \{0\}$.

1.2 Biochemical networks

Biochemical networks are one of the main concepts often used to model biological systems [3]. Such networks are made up of relations between physical entities (small molecules, proteins, nucleic acids, chemical compounds, complexes, larger macromolecular assemblies, atoms, electrons, and photons [6]) and reactions (such as standard chemical reactions, translocation of molecules from one compartment to another, association of molecules to form complexes and polymers, etc... [6]). More often than not biochemical networks are organized in pathways, sets of causally connected reactions [6]; such pathways, in pathway databases like Reactome (Figure .1) are organized hierarchically: Reactome has top-level pathways, like *Signal transduction* (reactions where extracellular signals elicit changes in cell state and activity), *Muscle contraction* (Figure .2) and the *Neuronal system*, each composed of other smaller pathways or reactions.

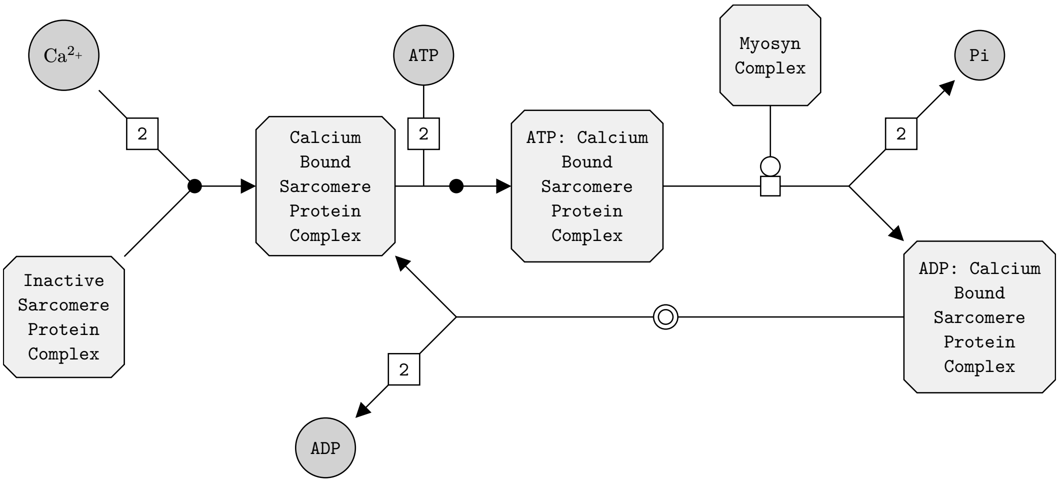

Ca2+
ATP
Myosyn Complex
Pi
2
2
2
Calcium Bound Sarcomere Protein Complex
ATP: Calcium Bound Sarcomere Protein Complex
Inactive Sarcomere Protein Complex
ADP: Calcium Bound Sarcomere Protein Complex
2
ADP

Figure 1.1. *Striated muscle contraction* pathway within the *Muscle contraction* pathway in the Reactome PathwayBrowser [5]

The biochemical network in Figure 1.1 represents the *Striated muscle contraction* pathway in the *Cytosol* compartment, a process where force is generated within striated muscle tissue, resulting in muscle movement [5]. Here the pathway has the following reactions:

1. Ca^{2+} (stoichiometry 2) binds with an *Inactive Sarcomere Protein Complex*
2. ATP (stoichiometry 2) associates with the bound "Ca^{2+} - *Calcium Sarcomere Protein Complex*"
3. The *Myosyn Complex* acts as a catalyst in order to activate the reaction that produces *Inorganic Phosphate* and the "ADP - *Calcium Sarcomere Protein Complex*" bound

4. The ADP dissociates from the "Ca^{2+} - *Sarcomere Protein Complex*" bound, producing two units of ADP (stoichiometry 2)

In particular, Ca^{2+}, ATP, *Inactive Sarcomere Protein Complex* and *Myosyn Complex* are inputs of the biochemical network (no reaction within the network produces these species), while Pi and ADP are outputs (no reaction within the network consumes these species).

It's important to note that since the *Myosyn Complex* is a catalyst (modifier) of the reaction, it is not consumed in the reaction.

Biochemical networks as presented in pathway databases are qualitative models, they describe only the roles of species in reactions and the structure of the network. Other than the stoichiometries there are no informations about typical quantities of species, reactions speeds, compartments sizes etc...

Definition 1 (Biochemical network). A biochemical network G is a tuple (S, R, E, s) where

- $S = U \cup X \cup Y$ is the set of species of the biochemical network
  - U is the set of input species of the network
  - X is the set of species which are neither inputs or outputs
  - Y is the set of output species of the network
- R is the set of reactions in the biochemical network
- $E = E_{\text{reactants}} \cup E_{\text{products}} \cup E^{+}_{\text{modifiers}} \cup E^{-}_{\text{modifiers}} \subseteq S \times R$ is the set of relations between species and reactions
- $s : E_{\text{reactants}} \cup E_{\text{products}} \to \mathbb{N}_{+}$ is the stoichiometry of the species in the reaction

The definition abstracts different reaction types. It also allows for a species to be have multiple roles in a reaction, since in the Reactome pathways database there are some reactions in which, for example, modifiers are also inputs or outputs of the reaction. (TODO: reference cypher query or species)

1.3 Dynamics of biochemical networks

Systems of ordinary differential equations (ODE) are one of the approaches used in order to describe the dynamic behaviour of biochemical networks [2].

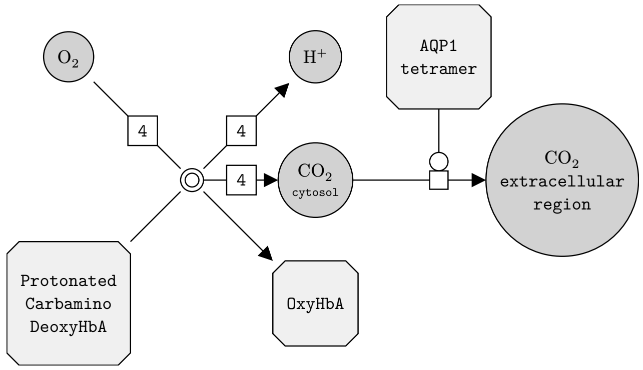


Figure 1.2. *Hemoglobin A* binds O_2, releasing H^+ and CO_2, then AQP1 acts as a catalyst and transports CO_2 from the cytosol to the extracellular region

Let $G = (S, R, E, s)$ be the biochemical network in Figure 1.2

- $S = U \cup X \cup Y$ where
  - $U = \{O_2, Hb_1, A\}$ with Hb_1 and A being respectively *Hemoglobin A* before the dissociation and AQP1
  - $X = \{CO_2^c\}$, with CO_2^c being CO_2 in the cytosol
  - $Y = \{Hb_2, H^+, CO_2^e\}$ with Hb_2 and CO_2^e being respectively *Hemoglobin A* after the dissociation and CO_2 in the extracellular region
- $R = \{R_1, R_2\}$ where R_1 and R_2 are respectively the dissociation reaction and the transport reaction

$$\begin{cases} a \\ b \end{cases}$$

Definition 2 (Dynamic biochemical network). Given a biochemical network $G = (S, R, E, s)$ let $D = (G, theta)$

Chapter 2

Biological models

We denote by $\mathbb{R}, \mathbb{R}_{0+}, \mathbb{Z}, \mathbb{N}$ the sets of, respectively, real, non- negative real, integer, and non-negative integer numbers.

The terms "physical entity" and "species" are used interchangibly in this document, since Reactome (TODO: cite reactome document model / glossary) uses "PhysicalEntity" to reference (TODO: biological species?) and SBML (TODO: cite SBML documentation) uses "Species".

Reactome uses "ReactionLikeEvent" to refer to generic reactions (TODO: cite glossary and use glossary to tell different types of reactions)

Definition 3 (Biological network). A biological network G is a tuple (S, R, E, σ) where

- $S = U \cup X \cup Y$ is the set of species of the biological network, where
  - U is the set of input species
  - Y is the set of output species
  - X is the set of other species in the network
- R is the set of reactions in the biological network
- $E \subseteq S \times R$

2.1 Reachability

Given a set of target species, a set of constraints on the target species (constraints which model a scenario that could present, for example, in a disease) and by taking into account all the reactions within a set target pathways that lead to the production, both directly and indirectly, of the target species, the goal is to find a subset of virtual patients for the described scenario.

2.2 Scenario definition

(TODO: define what is an expansions, why do we need a scenario, etc...it might be important to study what a subsection which contains two specific species behaves)

(TODO: define what is a Pathway, and what a Pathway is in terms of Reactome)

(TODO: maybe do a chapter about Reactome, or something simpler before about the generation, maybe about "Reachability", for reachability you need a definition of a network)

Definition 4 (Biological scenario). A
A scenario is defined by

- a set of physical entities from which to start the expansions
- a set ot pathways to which to limit (constraint?) reactions to
- a max depth for recursion (/reachability) (TODO: a max depth in terms of nodes in the path, not in the number of reactions in the path, for that apoc is needed)
- a set of physical entities to exclude from reachability
- a partial order of the species

(as per figure ... of UML etc...)

2.3 lll

2.4 Satisfiability problem definition

TODO: biochemical network → biochemical network with dynamics → satisfiability problem → optimization problem

Definition 5 (Biological model satisfiability problem). Given a dynamic biochemical network $D = (S, R, E, s)$ let

- $\mathcal{S} = \{C_s | s \in S\}$
- $\mathcal{S}_{\text{avg}} = \{C_{\text{avg}} | s \in S\}$
- ll

Chapter 3

Blackbox optimization architecture

3.1 Blackbox optimization

Given a function $f : X \to Y$, which is expensive to compute, and an optimization problem of the type $\operatorname{argmin}_{x \in X} f(x)$ is a blackbox optimization problem if no information about the derivative of f

3.2 OpenBox

OpenBox is an efficient open-source system designed for solving gener alized black-box optimization (BBO) problems. It can be used either as a Standalone python package or Online BBO service [4].

3.3 Orchestrator-worker infrastructure

OpenBox is an efficient open-source system designed for solving generalized black-box optimization (BBO) problems. It can be used either as a Standalone python package or Online BBO service @open-box.

OpenBox has a great support for bayesian optimization, so that will be the main subject of the analysis @open-box-automatic-algorithm-selection.

3.4 Scalability analysis

Chapter 4

Experiments

Chapter 5

Appendix

.1 Neo4j Graph databse

Neo4j is a native graph database, which means that it implements a true graph model all the way down to the storage level. Instead of using a "graph abstraction" on top of another technology, the data is stored in Neo4j in the same way you may whiteboard your ideas.

https://neo4j.com/docs/getting-started/whats-neo4j/

.2 Reactome

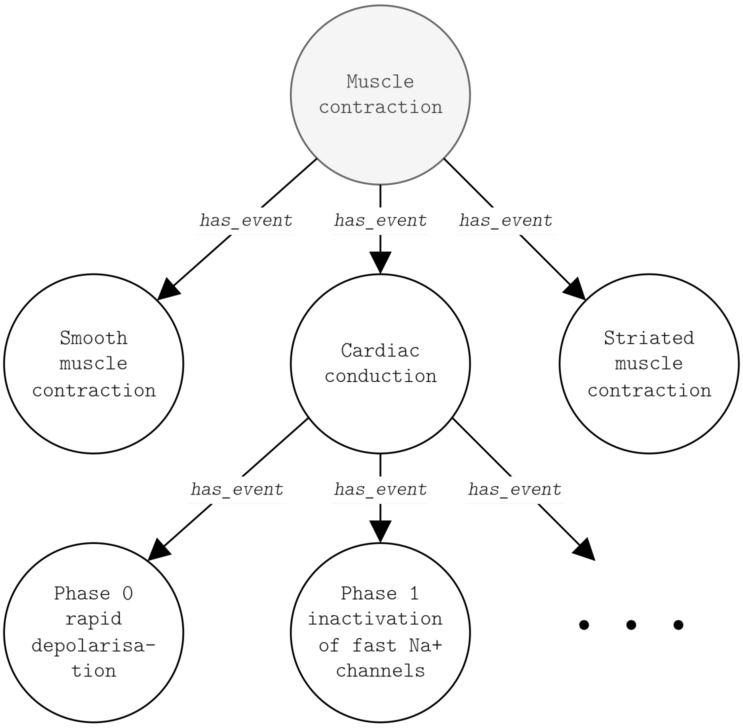

Muscle contraction
has_event
has_event
has_event
Smooth muscle contraction
Cardiac conduction
Striated muscle contraction
has_event
has_event
has_event
Phase 0 rapid depolarisa-tion
Phase 1 inactivation of fast Na+ channels
. . .

Figure .2. Muscle contraction pathway hierarchy [5]

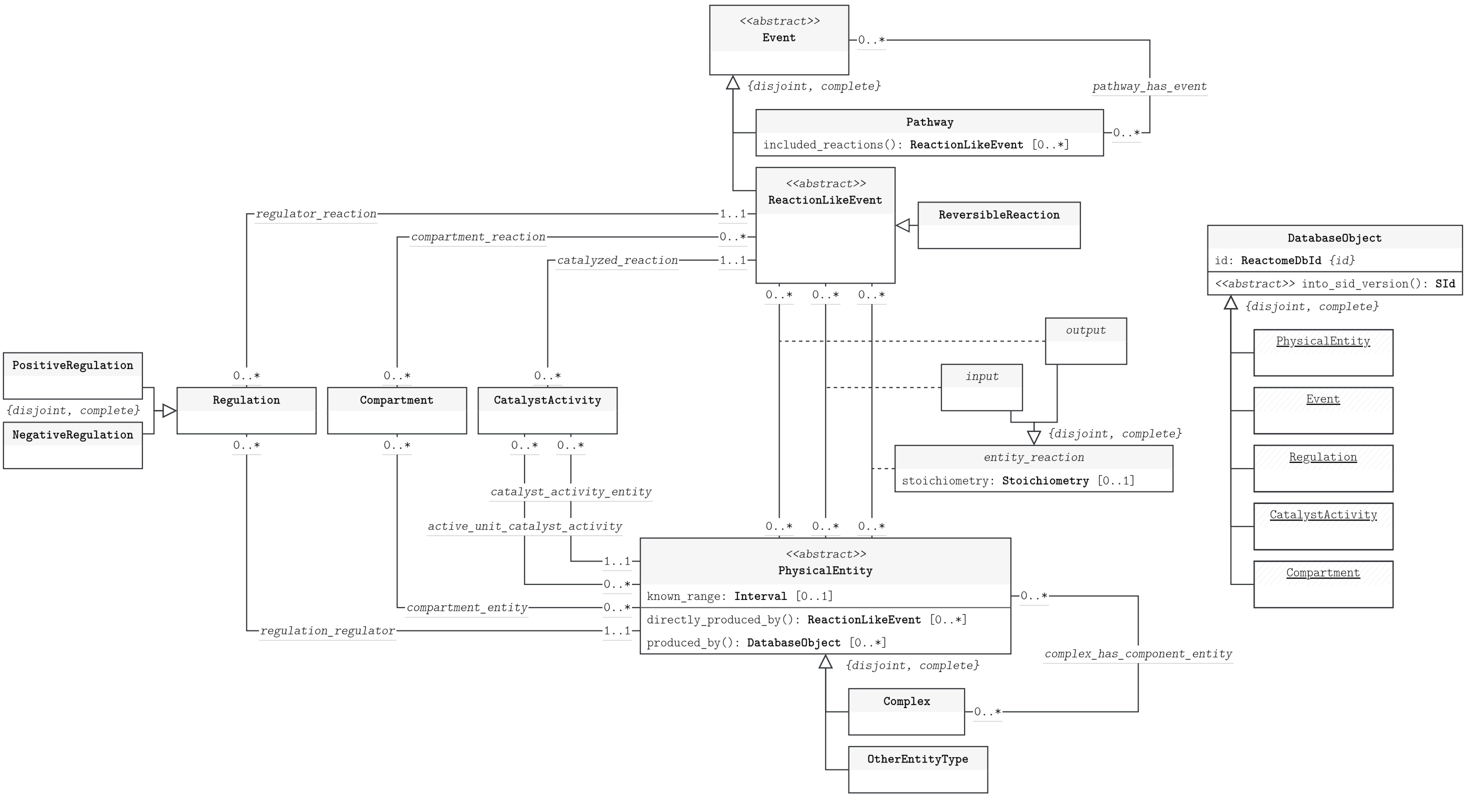
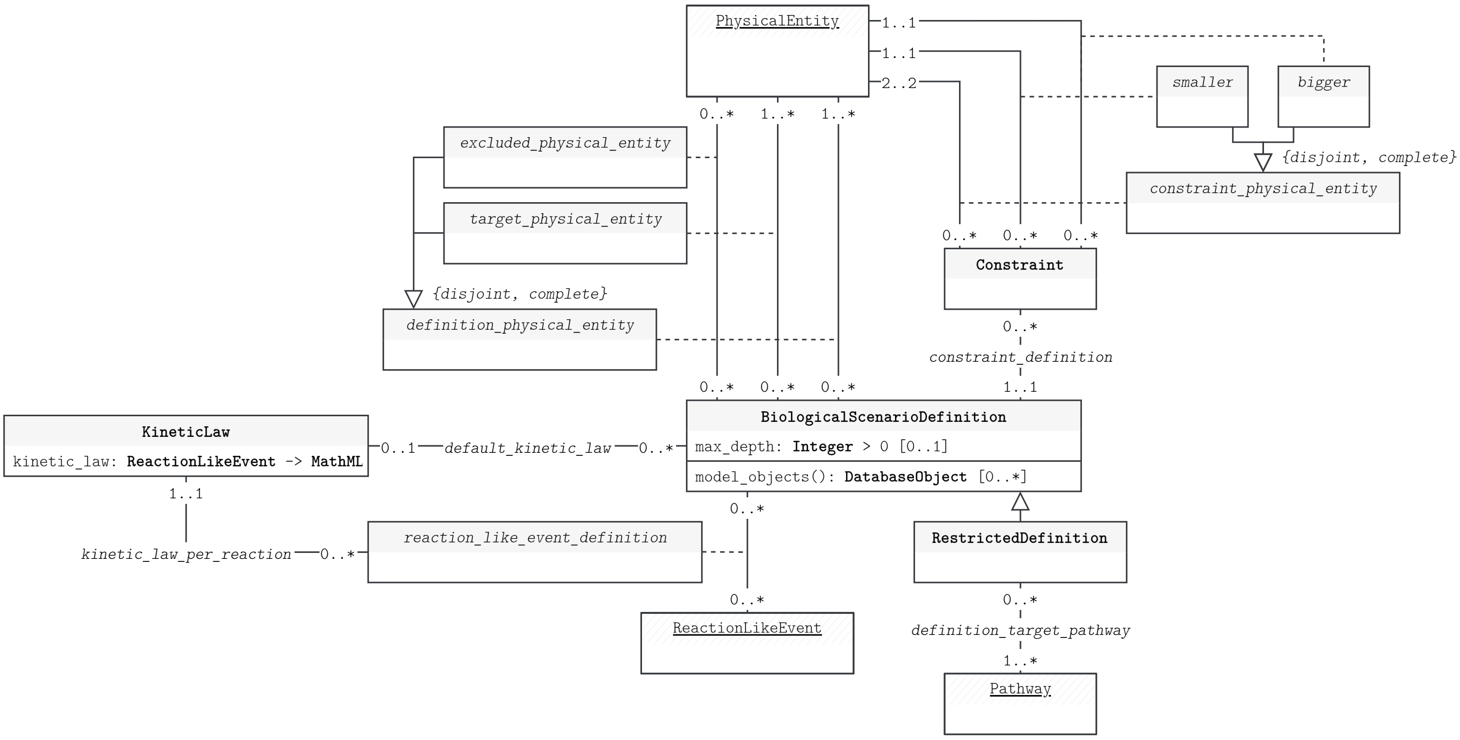


Figure .1. UML model of a portion of the Reactome database



```
MATCH
path =
    (reactionLikeEvent:ReactionLikeEvent)-[:catalystActivity]->
    (:CatalystActivity)-[:physicalEntity]->
    (physicalEntity:PhysicalEntity)
WHERE
EXISTS {
    MATCH (reactionLikeEvent)-[:input]->(physicalEntity)
}
RETURN COUNT(DISTINCT path)
```

.3 SBML simulation with Roadrunner

.4 HPC Cluster emulation with Docker Compose

Bibliography

[1] Antonio Fabregat et al. “Reactome graph database: Efficient access to complex pathway data”. In: *PLoS Computational Biology* 14.1 (2018), e1005968. DOI: 10.1371/journal.pcbi.1005968.

[2] Edda Klipp et al. *Systems Biology: A Textbook*. 2nd. Wiley-Blackwell, 2016, p. 504. ISBN: 9783527336364.

[3] M Koutrouli et al. “Erratum: A Guide to Conquer the Biological Network Era Using Graph Theory”. In: *Frontiers in Bioengineering and Biotechnology* 11 (Mar. 2023), p. 1182500. DOI: 10.3389/fbioe.2023.1182500. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7004966/.

[4] Yang Li et al. “Openbox: A generalized black-box optimization service”. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 3209–3219.

[5] Rush MG Marc E. Gillespie. *Striated Muscle Contraction*. Accessed: 2025-11-25. 2025. URL: https://reactome.org/PathwayBrowser/#/R-HSA-390522&PATH=R-HSA-397014.

[6] Reactome. *Data Model Glossary*. 2025. URL: https://download.reactome.org/documentation/DataModelGlossary_V90.pdf.