Identification of cell signaling pathways based on biochemical reaction kinetics repositories

Gustavo Estrela de Matos

Dissertation presented
to
Institute of Mathematics and Statistics
of the
University of São Paulo
For
The Degree Of Master Of Science

Field of knowledge: Computer Science

Advisor: Dr. Marcelo da Silva Reis

Center of Toxins, Immune-Response and Cell Signaling (CeTICS)

Special Laboratory of Cell Cycle, Butantan Institute

During the development of this work the author received financial support of FAPESP.

São Paulo, March 25, 2019

Abstract

Cell signaling pathways are composed of a set of biochemical reactions that are associated with signal transmission within the cell and its surroundings. Traditionally, these pathways are identified through statistical analyses on results from biological assays, in which involved chemical species are quantified. However, once generally it is measured only a few time points for a fraction of the chemical species, to effectively tackle this problem it is required to design and simulate functional dynamic models. Recently, it was introduced a method to design functional models, which is based on systematic modifications of an initial model through the inclusion of biochemical reactions, which in turn were obtained from the interactome repository KEGG. Nevertheless, this method presents some shortcomings that impair the estimated model; among them are the incompleteness of the information extracted from KEGG, the absence of rate constants, the usage of sub-optimal search algorithms and an unsatisfactory overfitting penalization. In this project, we propose a new methodology for identification of cell signaling pathways, which will make use of a myriad of public interactome and biochemical reaction kinetics repositories to deal with the incompleteness of a priori information. Moreover, we will use optimal algorithms for model selection, as well as more effective cost functions for overfitting penalization. The new methodology will be tested on artificial instances and also on cell signaling pathways identification in our case study, the Y1 mouse adrenocortical tumor cell line. (AU)

Contents

Chapter 1

Introduction

Cell Signaling pathways are cascades of chemical interactions that allow the communication between the cell environment and the cell itself. These pathways are also able to regulate many cell functions, including DNA replication, cell division and cell death. We can observe the functioning of signaling pathways as a mechanism that can conform the cell behavior with signals that come from the environment conditions in which the cell is placed. The studies of cell signaling pathways can lead to determining how cells can respond to different stimuli; for instance, with the studies of signaling pathways activated by a chemical species, one could determine how an unhealthy cell would respond to a drug containing this species.

It's possible to construct mathematical models to represent a set of chemical reactions and consequently a signaling network. One approach on the modeling of those interactions is based on the law of mass action. This law proposes that the rate of a chemical reaction is proportional to the product of reactants concentrations, i.e we can calculate the concentration change rate of a species in an interaction by calculating the product of reactants concentrations, up to a multiplying constant. If we consider the set of interactions of a signaling pathway, we can then come up with a system of ordinary differential equations (ODEs) that can model the dynamics of the concentration of each chemical species from the pathway. Generally, these systems are complex and cumbersome, if not impossible, to be solved analytically, therefore we resort on computational models that apply numerical methods to approximate solutions of these systems.

In this work, we are interested in computational models that can reproduce the behavior of signaling networks, comparing experimental measures—generally based on Western blot data—to simulated results. The figure 1.1 shows a set of interactions as well as parameters of a model of a signaling network. To create these computational models, two main tasks need to be accomplished.

The first task one must complete to create a model is to determine a set of interactions to consider in the ODE system. Searching for pathway maps on the Kyoto Encyclopedia of Genes and Genomes (KEGG) [KG00] is a good start. The KEGG PATHWAY Database provides manually drawn diagrams that represent signaling networks created with experimental evidences. However, it's possible that there's no pathway on KEGG that is able to correctly represent the biological experiment of interest; for those situations, it's necessary to modify the pathway by adding or even removing interactions. One might reason that we should use as many interactions as we can to get a better simulation, however, this usually implies in poor or computationally infeasible models because of two reasons: first, complex models will require more time in the numerical solution computation, which may be infeasible due to limited computational resources; and second, when considering many interactions, we are also placing many parameters (multiplying constants of the differential equations) on the model, and finding

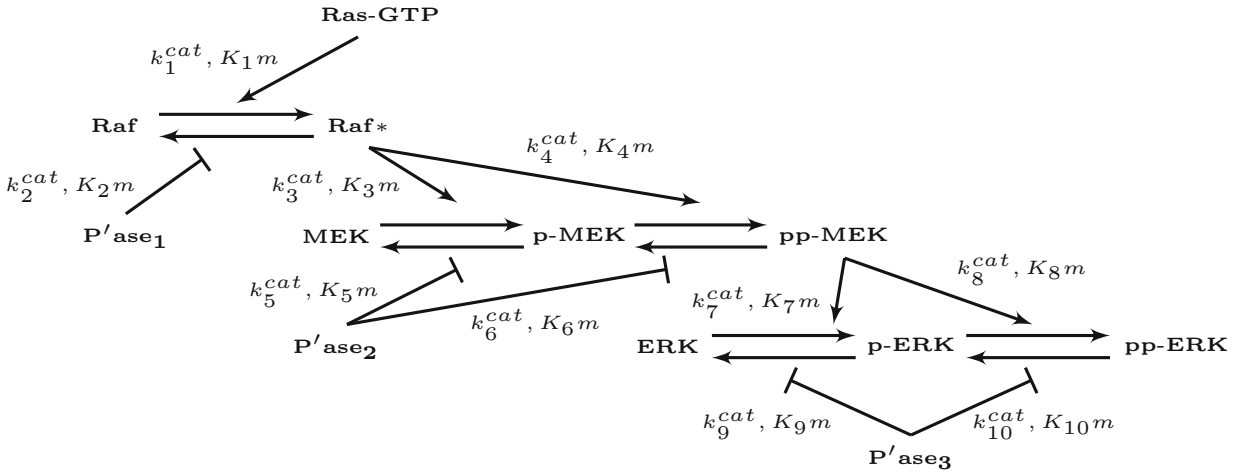

Figure 1.1: The above diagram show a hypothesis for a signaling pathway that flows through Raf-MEK-ERK cascade. Names in bold represent chemical species. Names in italic represent parameters of the ordinary differential equation of each interaction.Horizontal arrows represent phosphorylation when directed from left to right or dephosphorilation when directed in the opposite direction. Other arrows represent positive feedback if they are directed downwards or negative feedback otherwise. Image copied from Marcelo S. Reis et. al (2017) [Rei+17].

appropriate values for them becomes harder as we increase the number of parameters.

The second task to create a model is to find values for all the system parameters. There are two approaches for this task, you can either fetch values for these constants from the literature or you can find values that makes the model output approximate the experimental observations. For the first approach, repositories such as BioModels [LN+06] can be used; for the second approach, statistical and optimization methods can be used. For optimization, it is necessary to define a metric that can evaluate how close the parameters brings the model output to the experimental observation so that you can search for the optimal parameter in the parameter space. Statistical inference, in the other hand, usually tries to maximize some likelihood function, which is defined to represent the probability of a model, with a set of parameter values, to reproduce the experimental observation.

After completing both tasks, however, as we mentioned before, we might still not have found a model that fairly approximates the biological experiment of interest. That could indicate that the set of chemical interactions chosen for the model is incomplete or has interactions that are not relevant for the biological experiment. Therefore, it is desirable to construct a systematic method of modifying the set of chemical reactions in order to find the optimal set.

With the title "A method to modify molecular signaling networks through examination of interactome databases" [Wu15] Lulu Wu presented in her masters dissertation a methodology to systematically modify computational models of signaling networks to better simulate biological experiments. Starting with a model that does not approximate well the biological data, this methodology proposes to add a set of interactions to the model topology in order to better simulate the signaling pathway of interest. This set of interactions is a subset of interactions from a database created by Lulu Wu, joining information from many static maps available on KEGG. The choice of this subset can be modeled as a combinatorial optimization problem in which the search space is the set of all possible subsets of interactions to be added. The cost function of this problem, however, is not as simple to define as the search space. Note that the search space itself does not define models, but only the topology, i.e. the set of interactions, therefore, to consider the second task of creating computational models for signaling networks,

the cost function must take into account the set of values for the model parameters. As an example, we could define the cost as the minimum distance between model and experimental measures considering all possibilities of parameter values; however, unfortunately, finding the minimizing parameter values is a hard problem.

Since this is a hard problem, the method presented by Lulu Wu implements a heuristic version of this cost function. This heuristics is based on a Simulated Annealing procedure that searches for a set of parameter values trying to minimize (as much as possible) the distance between model and experimental measures; the best found distance is then considered as the cost of the model. Once the number of possible model topologies grows exponentially on the number of interactions from the database, the algorithm used to traverse the search space of subset of interactions is also a heuristic, and it is based on the greedy algorithm called Sequential Forward Selection (SFS) [Whi71]. The heuristic implemented by Lulu Wu selects a fixed number of interactions from the database and then creates candidate models by adding to the current solution the respective interaction; then, after evaluating the cost function for each model, the algorithm moves to the best candidate.

Chapter 2

Fundamental Concepts

Chapter 3

Conclusion

Bibliography

[KG00] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 28.1 (2000), pp. 27–30.

[Lic13] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: http://archive.ics.uci.edu/ml.

[LN+06] Nicolas Le Novere, Benjamin Bornstein, Alexander Broicher, Melanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, et al. "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems". In: *Nucleic Acids Research* 34.suppl_1 (2006), pp. D689–D691.

[Rei+17] Marcelo S. Reis, Vincent Noël, Matheus H. Dias, Layra L. Albuquerque, Amanda S. Guimarães, Lulu Wu, Junior Barrera, and Hugo A. Armelin. "An Interdisciplinary Approach for Designing Kinetic Models of the Ras/MAPK Signaling Pathway". In: *Methods in Molecular Biology*. Springer New York, 2017, pp. 455–474. DOI: 10.1007/978-1-4939-7154-1_28. URL: https://doi.org/10.1007/978-1-4939-7154-1_28.

[Whi71] A.W. Whitney. "A Direct Method of Nonparametric Measurement Selection". In: *IEEE Transactions on Computers* C-20.9 (1971), pp. 1100–1103. DOI: 10.1109/t-c.1971.223410. URL: https://doi.org/10.1109/t-c.1971.223410.

[Rei12] M. S. Reis. "Minimization of decomposable in U-shaped curves functions defined on poset chains – algorithms and applications". PhD thesis. University of Sao Paulo, 2012.

[Wu15] Lulu Wu. "Um método para modificar vias de sinalização molecular por meio de análise de banco de dados de interatomas". MA thesis. University of Sao Paulo, 2015.