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 28, 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 is 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 is possible that there is no pathway on KEGG that is able to correctly represent the biological experiment of interest; for those situations, it is 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

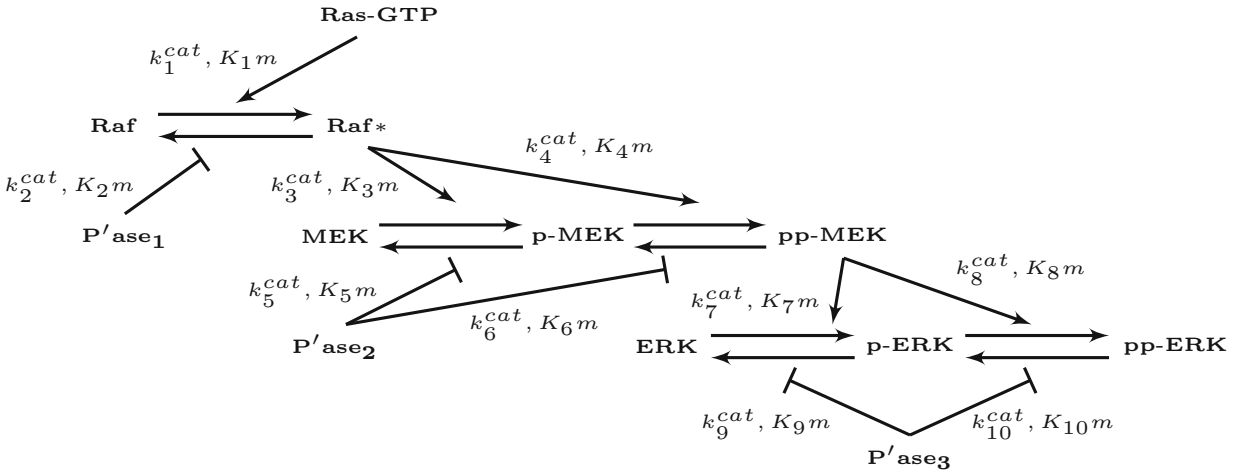

Figure 1.1: The above diagram show a hypothesis for a signaling pathway that flows through Raf-MEK-ERK cascade. Names in bold represent chemical species. Names in italic represent parameters of the ordinary differential equation of each interaction.Horizontal arrows represent phosphorylation when directed 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+17a].

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 of the model in order to find a good set to represent the signaling network.

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 of signaling networks available on KEGG. The choice of this subset can be modeled as a combinatorial optimization problem, the feature selection problem, in which the search space is the set of all possible subsets of interactions (features) 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, moreover, the algorithm used to traverse the search space is also a heuristics. The cost function 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. The size of the search space is exactly the number of all possible subsets of interactions to be added, and this number grows exponentially on the number of interactions from the database. That explains why Lulu Wu used a heuristic to traverse the search space. This heuristic 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.

The results presented on Lulu Wu's dissertation show that the method is useful when there are only a few differences between the starting model and a model that closely approximates the biological experiment. This limitation could be explained by the intrinsic difficulty of the problem, which demands fitting complex models with few experimental data; however we would like to highlight three aspects of the work that contributes to its limitations. The first aspect is that the constructed database could be more nearly complete, adding information from other interactome databases, such as STRING [Szk+10], and also by adding information about model parameters, i.e. chemical reaction constants, that are available in other databases, e.g. the SABIO-RK [Wit+11] database. Second, the search algorithm used to modify the models can only add interactions, therefore, if the algorithm starts with (or add along the search) a spurious interaction on the topology, then the algorithm will not be able to "regret" that interaction even though there might be similar solutions without it with better fit. Third, the cost function does not include a proper penalization of complex models; the used penalization is based on a execution time limit on the simulated annealing procedure, implying on a random penalization for more complex models, which typically demand more execution time. Without a proper penalization, the algorithm is doomed to select overly complex models that, even with a good fit to the experimental data, are not likely to reproduce the same experiment conduced with any kind of perturbation to the biological environment or to the data collected.

We propose on this project to create a new method for modifying models of signaling networks, based on the work of Lulu Wu, and including possible solutions to the three aspects mentioned on the last paragraph. To the first aspect, we propose to create a database that includes interactions informations gathered from KEGG and STRING, and also that includes reaction constants values, which can be fetched from SABIO-RK, Brenda [Sch04] or BioNumbers [Mil+09]. To the second aspect, we propose to create new search algorithms that are more general than SFS, and to test and compare these new algorithms we intend to use the featsel framework [Rei+17b]. For the last aspect, about the cost function, we intend to use Bayesian approaches to rank models [VG07] based on the likelihood of them to reproduce the observed data; if we say M is a model with parameter space Θ and D is a set of observations, then we

would like to estimate

$$p(D|M) = \int_{\theta \in \Theta} p(D|\theta, M) p(\theta|M) d\theta,$$

where $p(D|\theta, M)$ is the likelihood of model M, with parameters θ, to reproduce the experimental data D; $p(\theta|M)$ is the prior probability of θ; and finally, $p(D|M)$ is the probability of the data being generated by model M. This cost function has as an advantage the fact that models are not ranked using a single value for parameters, instead, the cost considers all possibilities of parameters, integrating over the parameter space. Another advantage of this cost function is that, since it is based on likelihoods of the model to reproduce data, overly complex models are automatically penalized.

The method developed on this work for model selection will then be applied in biological models, mainly related to tumor cells, that are relevant for Center of Toxins, Immune-response and Cell Signaling (CeTICS).

1.1 Objectives

In this work we propose to create a software that applies a method, based on the work o Lulu Wu, that is able to select a model to reproduce some biological experiment. This method should be able to consult a vast database of interactions and modify a starting model by removing or adding those interactions in order to construct a model that approximates more closely the biological experiment. The construction of this database and how the method consults it is also part of this work. We also propose to validate the methodology with real cell data. To achieve these goals, we should complete the following tasks:

1. **Define a cost function for models**. We propose to use a Bayesian approach to implement an algorithm that allows us to estimate the value of $p(D|M)$, which should be used as the "score" of the model. To complete this activity we will use as reference the Bayesian inference-based modeling method (BIBm) [Xu+10] and also the software ABC-SysBio [Lie+14].

2. **Build a database of interactions**. This database should include interactions gathered from KEGG and STRING, and reaction rate constants of interactions, which are available on other databases such as SABIO-RK, BioModels and BioNumbers.

3. **Formulate the systematic modifications on the models as a feature selection search space**. Given the initial model, we should be able to identify this model as a node of the search space, and we also should be able perform valid jumps from one node to another, in other words, we should be able to perform valid modifications to the model.

4. **Define search algorithms on the feature selection problem**. Given that we successfully structured the modifications of the model as a Boolean lattice, we should define algorithms to determine how to traverse this space in order to find a model with the least possible cost in a reasonable amount of execution time.

5. **Test feature selection algorithms**. Using artificial and then real data, we should test if the methodology can select a model that is able to reproduce the behaviour measured on the signaling network.

6. **Apply the methodology on a real case**. Finally, with a tested implementation of the methodology, we should help researchers from CeTICS to identify cell signaling networks of cells that are relevant to their research.

1.2 Organization

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.

[Lie+14] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael P H Stumpf. "A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation". In: *Nature Protocols* 9.2 (Jan. 2014), pp. 439–456. DOI: 10.1038/nprot.2014.025. URL: https://doi.org/10.1038/nprot.2014.025.

[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.

[Mil+09] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. "BioNumbers—the database of key numbers in molecular and cell biology". In: *Nucleic Acids Research* 38.suppl_1 (Oct. 2009), pp. D750–D753. DOI: 10.1093/nar/gkp889. URL: https://doi.org/10.1093/nar/gkp889.

[Rei+17a] 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.

[Rei+17b] Marcelo S. Reis, Gustavo Estrela, Carlos Eduardo Ferreira, and Junior Barrera. "featsel: A framework for benchmarking of feature selection algorithms and cost functions". In: *SoftwareX* 6 (2017), pp. 193 –197. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2017.07.005. URL: http://www.sciencedirect.com/science/article/pii/S2352711017300286.

[Sch04] I. Schomburg. "BRENDA, the enzyme database: updates and major new developments". In: *Nucleic Acids Research* 32.90001 (Jan. 2004), pp. 431D–433. DOI: 10.1093/nar/gkh081. URL: https://doi.org/10.1093/nar/gkh081.

[Szk+10] D. Szklarczyk et al. "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored". In: *Nucleic Acids Research* 39.Database (Nov. 2010), pp. D561–D568. DOI: 10.1093/nar/gkq973. URL: https://doi.org/10.1093/nar/gkq973.

[VG07] Vladislav Vyshemirsky and Mark A. Girolami. "Bayesian ranking of biochemical system models". In: *Bioinformatics* 24.6 (Dec. 2007), pp. 833–839. DOI: 10.1093/bioinformatics/btm607. URL: https://doi.org/10.1093/bioinformatics/btm607.

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

[Wit+11] U. Wittig et al. "SABIO-RK–database for biochemical reaction kinetics". In: *Nucleic Acids Research* 40.D1 (Nov. 2011), pp. D790–D796. DOI: 10.1093/nar/gkr1046. URL: https://doi.org/10.1093/nar/gkr1046.

[Xu+10] Tian-Rui Xu et al. "Inferring Signaling Pathway Topologies from Multiple Perturbation Measurements of Specific Biochemical Species". In: *Science Signaling* 3.113 (2010), ra20–ra20. ISSN: 1945-0877. DOI: 10.1126/scisignal.2000517. eprint: http://stke.sciencemag.org/content/3/113/ra20.full.pdf. URL: http://stke.sciencemag.org/content/3/113/ra20.

[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.