Statistical Analysis of Partially Calibrated Bio-pathway Models

Bing Liu

NUS Graduate School for Integrative Sciences and Engineering,
National University of Singapore
lb@nus.edu.sg

Abstract. Understanding the functioning of complex biological systems is a major challenge. A standard formalism used to model bio-pathways is a system of ordinary differential equations (ODEs). Bio-pathways usually involve a large number of molecular species and bio-chemical reactions. The values of parameters (rate constants) are often unknown because they are very difficult to be measured. Thus, we have to resort to computational methods and estimate them from experimental data by collective fitting. As the noisy data are often insufficient to constrain all parameters to have a unique solution, we propose to perform parameter estimation multiple times with different starting points. We then can gain biological insights by statistically analyzing the estimated parameters. Although this process is very time-consuming, it can be easily parallelized. We tested our approach on Complement system and achieved 82.1 speedup by enabling our application on the Grid.

1 Introduction

Understanding the functioning of complex biological systems is one of the great challenges facing science. The rewards of success will range from better therapies to new engineering materials. In the past decades, many experimental techniques have been developed, such as sequencing, western blot, etc. These tools have helped biologists to identify what components are present in a cell and how they interact with each other. Though we already have a general picture of the basic constituents of a cell, it is still difficult to understand the functioning of biological systems because bio-molecules do not function alone but exist in complex assemblies and pathways. Thus, it is important to develop system-level approaches. This has motivated the emerging domain of *Systems Biology* that seeks to understand how the individual components of a biological system interact in time and space to realize the various functions of the system. Systems Biology integrates many disciplines such as biology, mathematics, physics, chemistry, computer science, and electrical engineering. A long-term vision of this field is to put all the relevant biological processes together and build a model that can simulate an entire organism. Such model will have substantial impact on our health care, since it will not only lead to a better understanding of the physiological mechanisms and human diseases, but also bring about a more effective drug development process by means of *in silica* experiments.

To achieve this long-term vision, one first requires quantitative models to describe the dynamics of fundamental intra and intercellular processes. Usually, those processes are described as biological pathways including metabolic pathways, signaling pathways, and gene regulatory networks. Specifically, signaling pathways describe how cells sense changes in their environment and make responses by modifying their metabolisms, transcriptional activities or cell fates. Upon receiving extracellular stimuli or internal signals, cells pass messages to their gene regulatory systems via cascades of complex bimolecular reactions. The dynamics of this process is crucial for achieving the final responses. However, obtaining an intuitive understanding is difficult because signaling networks often consist of many players and multiple feedback loops. In order to accurately describe bio-pathway dynamics, we need to resort to mathematical modeling.

Currently, the most successful means of modeling bio-pathways is through simultaneous ordinary differential equations (ODE) systems and their variants.

Bio-pathways usually involve a large number of molecular species and biochemical reactions. Hence the corresponding ODE model will involve many variables and parameters. The values of parameters (rate constants) are often unknown because they are very difficult to be measured through experiments. Thus, we have to resort to computational methods and estimate them from experimental data by collective fitting. However, parameter estimation in high dimensional space is very difficult because: (1) it require a huge amount of time; (2) the noisy and limited experimental data are often insufficient to constrain all parameters to have a unique solution. Therefore, instead of searching for the 'true' values of parameters, it is more reasonable to perform parameter estimation many times with different random starting points to get multiple sub-optimal solutions. Then statistical analysis of those estimated parameters will not only help us to understand the properties of parameter space but also will reflect the sensitivities of corresponding biochemical reactions and provide important biological insights.

Although generating many sub-optimal solutions is extremely time-consuming, this process can be easily parallelized and dramatically sped up by grid computing. In this project, We tested our approach on the Complement system. We enabled our application on TCG@NUS and achieve up to 82.1 speedup.

The rest of this report is organized as follows. We briefly introduce bio-pathway modeling and our analysis method in Section 2. The results of our case study on Complement system and the performance speedup by grid computing are presented in Section 3. In the final section, we summarize the report and discuss future work.

2 Methods

2.1 Bio-pathway modeling

To study the complex behavior of bio-pathways, many computational models have been proposed in the recent years, ranging from abstract models that emphasize on understanding some generic features of signaling pathways [1] to detailed models that quantitatively describe the dynamics of specific pathways in

specific organisms [2]. Currently, the most commonly used approach is via a system of ordinary differential equations (ODEs), which is the cores of many quantitative models such as XS-System [3], Hybrid Automata [4], Piecewise-affine hybrid model [4], etc. ODE kinetic modeling seeks to describe bio-molecular transformations such as covalent modification, intermolecular association and intracellular localization in terms of equations derived from established physical and chemical theories [5]. The general mathematical representation is as follows:

$$\dot{x_i} = \sum_{j=1}^{r} n_{ij} v_j \tag{1}$$

where $\dot{x_i}$ denotes the rate of production and consumption of individual bio-molecular species x_i, r is the number of reactions associated with x_i, the quantities n_{ij} denote the stoichiometric coefficients, and v_j describes the reaction rate of corresponding biological processes such as catalysis and assembly in terms of sophisticated kinetic laws such as Mass action law, Michaelis-Menten law, Power law, etc. For example, the reaction rate of a reversible binding process of two proteins $x_1 + x_2 \rightleftharpoons x_3$ can be modeled by mass action kinetics, which states that rates of an elementary reaction are proportional to the concentrations of the reacting species. Thus given rate constants k_1 and k_2, the forward and reverse rate can be expressed as $v_1 = k_1 x_1 x_2$ and $v_2 = k_2 x_3$, which results in an ODE system as follows:

$$\begin{aligned}
\dot{x_1} &= -v_1 + v_2 = -k_1 x_1 x_2 + k_2 x_3 \\
\dot{x_2} &= -v_1 + v_2 = -k_1 x_1 x_2 + k_2 x_3 \\
\dot{x_3} &= v_1 - v_2 = k_1 x_1 x_2 - k_2 x_3
\end{aligned}$$

Since each bio-chemical reaction involved in the signaling network can be described by certain rate equation, we can systematically convert the bio-pathway into a system of ODEs. Then the dynamics we are trying to study become the solution of such ODE system.

Parameter estimation However, it is very challenging to analyze large signaling networks because even the simplified models often possess tens or hundreds of unknown kinetic parameters whose values can significantly affect model behavior. As it is very difficult to determine these parameters through wet-lab experiments, a typical modeling effort might draw some values from the literature (possibly from different cell lines), which may introduce substantial inaccuracies. An alternative way is to resort to computational techniques to estimate parameters by optimizing the agreement between the model prediction and available experimental data (e.g. time courses of bio-molecular concentration). In this manner, parameter estimation is formulated as an optimization problem with differential algebraic constraints. Typically, the goodness-of-fit of a parameter combination (i.e. difference between prediction and experimental data) is measured by the objective function that is a weighted sum of square error:

$$f_{obj}(\mathbf{p}) = \sum_{i,j} \omega_i (x_{i,j} - y_{i,j}(\mathbf{p}))^2 \tag{2}$$

where $\mathbf{p}$ is the currently tested parameter set, $x_{i,j}$ is a point in the data set for species x_i at time point j, $y_{i,j}(\mathbf{p})$ is the corresponding simulated value, and ω_i is the normalization factor for x_i which is usually the inverse of the maximum value of x_i.

In order to search for the optimal solution $\mathbf{p}$ in parameter space which can give the lowest objective value f_{obj}, many algorithms such as Levenberg-Marquardt [6], Genetic Algorithm [7], Particle Swarm Optimization [8], etc, have been proposed in the past decades. The benchmarking tests done by Moles et al. and Fomekong-Nanfack et al. [9] [10] separately showed that a variation of evolutionary strategy algorithm namely Stochastic Ranking Evolutionary Strategy (SRES) [11] can give the best performance among popular algorithms in the settings of biological pathway models. Thus, we adopted SRES to estimate the unknown rate constants in our pathway models.

Model Analysis Once we have a system of ODEs with estimated parameter values, simulations via numerical integration methods can be performed. Model analyses can be conducted in many ways. For instance, we can monitor the dose-dependent response of a pathway to certain drug by varying the initial condition of simulation. Similarly, the effects of a kinase inhibitors or mutation can be simulated by modifying the rate constant [12]. By adding or removing links of the model structure, we can also test or generate hypotheses by comparing the obtained time courses with experimental data [13]. Further, one can also analyze the role of particular species or parameters. For example, Gallego et al. found an opposite role for *tau* in circadian rhythms [14], Lee et al. explained the importance of *axin* degradation in amplifying and sharpening the *Wnt* signal [15], and Birtwistle et al. suggested that phosphoinositol-3 kinase is a major regulator of post-peak but not pre-peak *EGF*-induced *ERK* activity [16]. The analysis of the role of parameters is usually via sensitivity analysis that evaluates the contribution of a particular parameter to the overall behaviors [17]. It is valuable in ascertaining which parameters should be the focus of direct measurement or experimental perturbation. Finally, studies can also be performed to analyze the network properties such as bistability [18] and oscillation [19] caused by multiple positive/negative feedback loops.

2.2 Statistical analysis with grid computing

Although SRES outperform other parameter estimation algorithms in general cases, they still fail to produce good results within acceptable time when we are dealing with large signaling networks. First, to evaluate the goodness-of-fit of a parameter solution requires simulating the whole ODE system for all time points. Since nonlinear differential equations have no analytical solutions in general, we

need to resort to *finite difference methods* that numerically approximate the solutions of differential equations using finite difference equations to approximate derivatives. Such methods are often very slow especially for stiff ODE systems. Second, although SRES may require less evaluations than others, searching for the optimal solution in a high dimensional space still require millions of evaluations that is very computational expensive. Third, due the limitations of the the quantity, quality and type of biological data, non-identifiability of the models may arise so that relationships among parameters and experimental observables are inherently non-unique. It has been shown that even a complete set of time-resolved measurements of all species in a complex reaction model is insufficient to fully constrain all parameters due to the sloppy' nature of the fitting procedure [20]. In other words, non-identifiability may induce a large regions of parameter space that are consistent with experimentally observed dynamics. Therefore, instead of researching for one optimal solution using limited and noisy data, we need to generate and analyze multiple sub-optimal solutions. By computing the width of the distributions in parameter values across those solutions, it would be valuable to know which parameters are well constrained, which co-vary and which are truly unconstrained because such information will help us to identify sensitive or robust reactions in the network and provide more biological insights.

To achieve this, we have to execute SRES many times with different initial starts. This is infeasible on a stand-alone personal computer since the computational demand will be huge. However, notice that the process of getting multiple sub-optimal solutions can be easily parallelized to many sub-tasks and executed on computer cluster. Furthermore, these sub-tasks only differ in initial conditions and do not require communications during execution. This motivated us to enable our program on Tera-Scale Campus Grid at NUS (TCG@NUS) using the Parametric Study program. Currently, there are 1554 computers on the TCG@NUS. To run an application in parallel mode on the TCG@NUS, one needs to have a wrapper program to split the application into multiple workunits. When a job is submitted, those workunits will be dispatched by the server to client agents if their load is below certain level. Then the result will be sent back to the server upon completion. We utilized Parametric Study Program as the wrapper since it allows user to run a program many times with different inputs. We then modified our parameter estimation program in order to fit its requirements about input and output.

3 Results

We now present our case study of a relatively large signaling and show the relevant performance results along with our grid-enabled implementation.

3.1 Modeling complement system

The complement system is a bio-chemical cascade which helps clear pathogens from an organism. It is one part of the larger immune system. The complement

system consists of a number of small proteins found in the blood, normally circulating as inactive zymogens, when simulated by one of several triggers, proteases in the system cleave specific proteins to release cytokines and initiate an amplifying cascade of further cleavages. The end result of this activation cascade is massive amplification of the response and activation of the cell-killing membrane attack complex. Over 20 proteins and protein fragments make up the complement system, including serum proteins, serosal proteins, and cell membrane receptors. These proteins are synthesized mainly in the liver, and they account for about 5% of the globulin fraction of blood serum. The bio-chemical pathways activate the complement system: the classical complement pathway, the alternative complement pathway and the mannose-binding lectin pathway. The differences between the three pathways are steps before the forming of C5convertase, i.e. three pathways converge at C5. Our research focuses on the crosstalk between classical and lectin pathways induced by local inflammation. The schematic graph is shown as Figure 1.

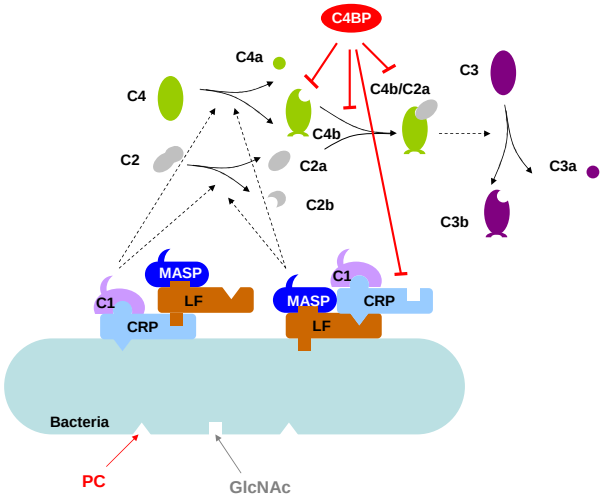


Fig. 1. The Schematic representation of the complement pathway

We constructed the ODE model using COPASI software [21] based on mass law and Michales-Menten kinetics. The resulting model has 42 differential equations and 52 unknown parameters.

3.2 Generating multiple parameter sets

We then implemented a program to translate the SBML file generated by COPASI to C code so that the ODE system can be solved by CVODA method and the objective function can be recognized by libSRES, a C library for stochastic ranking evolution strategy for parameter estimation. As TCG@NUS consists of resources available in multiple OS platforms such as Windows and Linux, we

compiled our code for Linux 64 platform. After uploading the generated executable to the grid server, the following jobs can be automatically distributed to the selected nodes with Linux platform.

Job ID	μ	λ	Gen	No. of Samples
9944	30	200	6000	99
10055	40	300	10000	99
10282	30	200	500	2000

Table 1. Experiments

Three experiments summarized in Table 1 were performed. Here μ, λ and Gen, standing for number of parents, number of mutants and number of generations respectively, are the parameters of SRES algorithm. The number of samples correspond the number of sub-optimal solutions generated. Since Job10055 has the largest $\{\mu, \lambda, Gen\}$, it is expected be able to find better solutions than the other two jobs. As Job10282 has the largest sample number, it is expected be able to generate the most accurate distributions. The probability distributions of objective values induced by parameter sets generated by the three experiments are shown in Figure 2. The distributions of Job9944 and Job10055 are very close, indicating that $\{\mu = 30, \lambda = 200, Gen = 6000\}$ is enough to estimated relatively good sets of parameters. Although global method like SRES can find the optimal theoretically, it is interesting to see that with arbitrary staring points, the qualities of estimated parameters can be quite different even we spend enough time for searching. On other hand, the mean of the distribution of Job10282 is much larger than the other two, suggesting that 500 generation is far from enough. However, those solutions are still quite useful, as they can help us to understand the properties of parameter space.

Moreover, we further investigate the distributions of each parameter values across all sub-optimal solutions. The results are shown in Figure 3. Among equally good fits (those that varied by less than experimental error), only a subset of parameter values were constant such as K49, whereas others such as K15 varied over the entire allowable range. One interpretation of this finding is that biological networks are quite robust to changes in rate constants and protein concentration, while the dynamics are shaped by the values of a small number of critical parameters.

3.3 Performance

We tested the performance of our program on both stand-alone PC and TCG@NUS. The running time of the three tasks for one sample (workunit) on stand-alone PC is reported in Table 2. For example, a single run of Job10055 took about 20 hours. Thus, hundreds of runs of Job10055 will require several month of running time which is infeasible. However, as the results shown in Table 3 indicated, with the help of grid computing. We can dramatically reduce the running time and achieve up to 82.1 speedup.

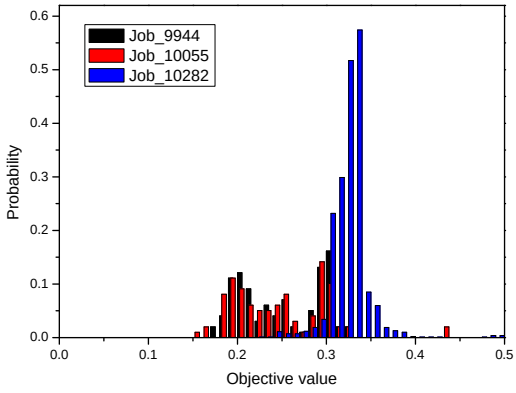


Fig. 2. Probability distributions of objective values of sub-optimal solutions generated by the three experiments.

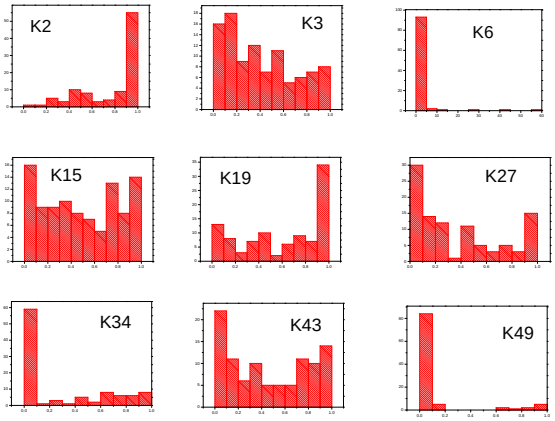


Fig. 3. Probability distributions of selected parameter values across sub-optimal solutions.

Job ID	No. of Workunits	Running Time
9944	1	481 min
10055	1	1208 min
10282	1	42 min

Table 2. Performance on single PC

Job ID	No. of Workunits	No. of Nodes	Total CPU Time	Running Time	Speedup
9944	99	96	44928 min	729 min	61.6
10055	99	100	108780 min	3245 min	33.5
10282	2000	91	67815 min	826 min	82.1

Table 3. Performance on the Grid

By comparing Table 2 and Table 3, we may observe that for light workload tasks such as Job10282, it is easy to achieve high speedup, while for heavy workload tasks such as Job10055, the speedup we can get is relatively low. The reason for this is that a long run on nodes of TCG@NUS will induce high risk of failure (e.g. the failure rate of Job10055 %35) since when the nodes are used by their owners for CPU intensive jobs, they are not available anymore. Thus, when we are using TCG@NUS, it is better to further parallel our applications so that the running time of workunits will not be too long.

4 Conclusion

We have proposed a statistical analysis method for bio-pathway modeling. We estimate multiply parameter sets instead of searching for one global optimal in order to cope the non-identifiability induced by limited and noisy data. As the distributions of each estimated parameters can reflect the sensitivities and robustness of corresponding reactions, biological insights about the pathway dynamics can be gained by such statistical analysis. We tested our approach on the Complement system. By enabling our application on TCG@NUS, we were able to achieve up to 82.1 speedup and make our application piratical. Currently, we are investigating more properties of Complement system using the parameters estimated by TCG@NUS, and we plan to apply this method to other bio-pathways in collaborating with biologists in NUS.

References

1. Papin, J., Palsson, B.: Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. Journal of Theoretical Biology **227**(2) (2004) 283–297
2. Vaseghi, S., Macherhammer, F., Zibek, S., Reuss, M.: Signal transduction dynamics of the protein kinase-A/phosphofructokinase-2 system in Saccharomyces cerevisiae. Metabolic Engineering **3**(2) (2001) 163–172
3. Antoniotti, M., Policriti, A., Ugel, N., Mishra, B.: XS-systems: extended s-systems and algebraic differential automata for modeling cellular behavior. In Sahni, S., Prasanna, V.K., Shukla, U., eds.: HiPC. Volume 2552 of Lecture Notes in Computer Science., Springer (2002) 431–442

4. Ghosh, R., Tomlin, C.: Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modelling: Delta-notch protein signalling. Systems Biology **1**(1) (2004) 170–183
5. Aldridge, B.B., Burke, J.M., Lauffenburger, D.A., Sorger, P.K.: Physicochemical modelling of cell signalling pathways. Nature Cell Biology **8** (2006) 1195–1203
6. Levenberg, K.: A method for the solution of certain nonlinear problems in least squares. Quart. Appl. Math. **1994** (2) 164–168
7. : Handbook of evolutionary computation. Oxford University Press (1997)
8. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: Proceedings of the Fourth IEEE International Conference on Neural Networks, Perth, Australia (1995) 1942 – 1948
9. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Research **13** (2003) 2467–2474
10. Fomekong-Nanfack, Y., Kaandorp, J.A., Blom, J.: Efficient parameter estimation for spatio-temporal models of pattern formation: case study of Drosophila melanogaster. Bioinformatics **23**(24) (2007) 3356–3363
11. Runarsson, T., Yao, X.: Stochastic ranking for constrained evolutionary optimization. IEEE Transactions on Evolutionary Computation **4** (2000) 284–294
12. Geva-Zatorsky, N., Rosenfeld, N., Itzkovitz, S., Milo, R., Sigal, A., Dekel, E., Yarnitzky, T., Liron, Y., Polak, P., Lahav, G., Alon, U.: Oscillations and variability in the p53 system. Molecular Systems Biology **2**(32) (2006) 1–16
13. Koh, G., Teong, H.F.C., Clement, M.V., Hsu, D., Thiagarajan, P.: A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk. Bioinformatics **22**(14) (2006) e271–e280
14. Gallego, M., Eide, E.J., Woolf, M.F., Virshup, D.M., Forger, D.B.: An opposite role for tau in circadian rhythms revealed by mathematical modeling. PNAS **103**(28) (2006) 10618–10623
15. Lee, E., Salic, A., Kruger, R., Heinrich, R., Kirschner, M.W.: The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. PLoS Biology **1**(1) (2003) 116–132
16. Birtwistle, M.R., Hatakeyama, M., Yumoto, N., Ogunnaike, B.A., Hoek, J.B., Kholodenko, B.N.: Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Molecular Systems Biology **144** (3) 1–16
17. Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D.S., Kell, D.B.: Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway. Molecular BioSystems **2** (2006) 640–649
18. Legewie, S., Bluthgen, N., Herzel, H.: Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. PLoS Computational Biology **2**(9) (2006) 120–133
19. Forger, D.B., Peskin, C.S.: A detailed predictive model of the mammalian circadian clock. PNAS **100**(25) (2003) 14806–14811
20. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology. PLoS Computational Biology **3**(10) (2007) 189
21. Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U.: COPASI - a COmplex PAthway SImulator. Bioinformatics **22**(24) (2006) 3067–3074