School Of Electronics and Computer Science
Faculty of Physical and Applied Sciences
University of Southampton

Argyris Zardilis

December 8, 2012

Tool for parameter inference in dynamic biological systems

Project Supervisor: Dr. Srinandan Dasmahapatra
Second Examiner: Dr. Markus Brede

A progress report submitted for the award of
BSc Computer Science

Abstract

The increase in the use of theoretical mathematical models to describe biological systems has led to an increasing need for computational tools to assist in the process of constructing those models and estimating their parameters from available experimental data. Although there is a rich literature on parameter estimation using a number of different techniques, very few attempts have been made to produce computational tools that systematically attack the parameter estimation problem. From those, almost all of them attempt to reproduce experimental data, disregarding qualitative features of the systems and other requirements we might have from the model arising from the dynamic behaviour of such systems as they evolve or respond to external or internal stimuli.

The aim of this project it to produce a computational tool for automatic parameter estimation in generic dynamic biological systems taking into account the dynamic behaviour of such systems.

Contents

1 Problem and Project goals

Although mathematical modeling for biological systems is not a new topic, with the advent of microarray experiments that enabled us to take measurements of the expression levels of a big number of genes at different time points, there has been a big grow in the attempts to mathematically capture the behaviour of such systems. These measurements led to the discovery of common patterns and networks in cell behaviour. Increasingly complex models have been employed to describe the behaviour of such networks. The use of models in this case is beneficial as they provide a great intuition and increased understanding into the dynamic nature of such systems and can even uncover and predict new behaviour which can lead to experimental confirmation. These models are often a system of Ordinary Differential Equations (ODEs) which are the natural way to capture dynamic and changing behaviour. As the complexity of these models grew the problem of estimating the parameters of these models, in this case primarily rate constants and kinetic parameters, became a major task for modelers as those parameters are often infeasible to experimentally measure. This fact along with the noisy and often scarce data made people resort to manual search of parameter space and/or mathematical techniques to assist in the task.

The last few years and with the rapid technological advancements, the computational cost associated with such techniques has fallen and there have been some attempts to automate this process with the production of computational tools that tackle the parameter estimation problem using mathematical methods borrowed from other fields. Unlike other fields however the requirements of models of biological systems are greater than just a simple reproduction of experimental data. These systems have qualitative features that are often very important and their behaviour changes as they evolve or respond to stimuli. If the proposed model does not take into account such behaviour then the model does not capture the precise behaviour and although it can sometimes be useful, it is incomplete. The main goal of this project is the production of a computational tool that systematically attacks the problem of parameter estimation by automating the estimation process using mathematical methods. This tool will take into account the dynamic behaviour of biological systems and use these criteria in the parameter selection process. This new aspect will make the tool more powerful and the models produced more complete and closer to the true behaviour as exhibited in nature. I believe that mathematical models that capture all aspects of these systems will increase the confidence in modeling and the use of mathematical tools in biological research just like they are being used successfully in other scientific disciplines such as Physics.

2 Background

Recent advances in recombinant DNA and microarray technology have led to greater understanding of the networks that govern different cell processes. In these networks a number of different genes and proteins interact to carry out a biochemical process. Examples of well studied networks are metabolic pathways and circadian oscillators[1]. The traditional way to describe these networks is by means of diagrams that show the signals and interaction between entities (genes, proteins). As these systems grow in complexity the diagrammatic way of describing the system is increasingly proved to be a poor way to do it as it does not give any intuition into the inner workings of the system nor does it capture the dynamic nature that such systems exhibit. Moreover as the complexity of the network grows it become increasingly difficult to predict the future behaviour of the system especially quantitatively. The advantages and the need for using quantitative mathematical models to describe such systems has been recognised and the perception of their importance is growing[6]. Also the number of attempts to model systems that are thought to be well understood like circadian oscillators is ever increasing[2, 9]. Chemical kinetic laws such as the Law of mass-action or the Michaelis-Menten kinetics can be easily translated to simple systems of ODEs to construct simple components or commonly found patterns in biochemical networks such as switches, buzzers, blinkers[14] the inspiration for the decomposition into electronic-like components coming from the resemblance that these networks show at the systems level with electronic circuits. By combining these simple motifs more complex networks can be built. Although the underlying behaviour at the molecular level is highly stochastic the models are deterministic systems of ODEs because the underlying stochasticity gives rise to deterministic behaviour at the systems level.

A typical systems can be described with a set of ODEs like so: $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t, \theta)$. Vector $\mathbf{x}$ is usually functions describing the change of the concentration of a substance (like mRNA, protein etc.) over time with θ being the parameter vector. And altough the construction of such models can be done by using simple laws (law of mass-action, Michaelis-Menten kinetics) or from just knowledge of the behaviour of the system in question, finding the parameters θ that will make these systems behave according to observations is an equally if not more difficult task since experimentally measuring these parameters like rate constants can be next to impossible. The task therefore becomes from a set of observations over time $\mathbf{Y} = \{\mathbf{y}_{t_1}, \mathbf{y}_{t_1}, \ldots, \mathbf{y}_{t_n}\}$ to find the parameter vector θ^* such that the distance between the simulated dataset from $f(\mathbf{x}, t, \theta^*)$ and the actual dataset $\mathbf{Y}$ is minimum. This is a common formulation of the problem of the inverse problem which reduces the problem to an optimisation problem which is a well studied area[4]. Common local and global optimisation techniques have been used in this area as well[10].

Another way to look at the problem is from a statistical/probabilistic point of view. The goal in this case is to increase the likelihood $L(\theta) = f(Y|\theta)$ where Y are observations and θ is the parameter values[3]. Then the problem becomes to find the values of parameters θ that maximise the likelihood function $L(\theta)$. Maximum likelihood estimation is again a well studied area and techniques used in other areas have been employed in this particular case as well.Other statistical methods have also been employed to tackle this problem, like Bayesian inference techniques. The traditional Bayesian inference in this case becomes $\pi(\theta|Y) = f(Y|\theta)\pi(\theta)$. [13]. The likelihood given here is not computable so that gave rise to new set of techniques that simulated the Bayesian inference by repeatedly sampling from some distribution accepting only samples θ^* such that the simulated dataset is close to the original dataset Y. That way the output is sample from the posterior $\pi(\theta|Y)$. Generating a sample to simulate an unknown or difficult distribution is not a new technique as it is the main theme in traditional Monte Carlo techniques such as Metropolis-Hastings[15]. Because computer simulation is more feasible nowadays these techniques commonly referred as Approximate Bayesian Computation techniques (approximate because they do not really do the Bayesian inference computation) have gained popularity. A range of different algorithms have been employed, from simple rejection samplers[12] to Markov Chain Monte Carlo methods in the spirit of Metropolis-Hastings[8] and Sequential Monte Carlo methods which use a series of distributions that at every stage are closer to the true posterior[13].

Tools that assist in the modeling process have been created in recent years(XPP-Auto, GRIND etc.). These tools are widely used because they allow modelers to get an intuition and greater understanding of the model in question usually by graphical means such as plotting bifurcation diagrams and allowing them to see the changes in the dynamics of the system as parameters change. And although they help in parameter estimation they do not address the problem directly. One other attempt that has been made that addresses the problem directly is the ABC-Sysbio[7]. This tool does automatic parameter estimation using Sequential Monte Carlo method.

All these methods and these tools have as their main aim to find the set of parameters that reconstructs the experimental data. They do not take into account other features of the system in question that the model must capture to correctly capture its dynamics. However systems have a dynamic nature that has to be taken into account. They respond to signal in a certain way, they evolve. For example circadian oscillators that exist in many tissues get synchronised with the main clock in the SCM of the brain by getting an entrainment signal from it. When perturbed in that way their signal changes in a specific way as described by their Phase Response curves[11]. Other studies have also been made in using Fourier Analysis which is a widely used technique in engineering applications to model selection by comparing the Fourier Transform as produced by different models with the Fourier Transform of the experimental data instead of actually comparing the data points of simulated versus real dataset[5]. That way the model selection process captures other qualitative features of the model that are not directly observable from the data.

3 Work

3.1 Results

4 Future Work

References

[1] J. Bass and J.S. Takahashi. Circadian integration of metabolism and energetics. *Science Signalling*, 330(6009):1349, 2010.

[2] S. Becker-Weimann, J. Wolf, H. Herzel, and A. Kramer. Modeling feedback loops of the mammalian circadian oscillator. *Biophysical journal*, 87(5):3023–3034, 2004.

[3] S. Filippi, C. Barnes, J. Cornebise, and M. P. H Stumpf. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *ArXiv e-prints*, June 2011.

[4] D. Gonze. Modeling circadian clocks: From equations to oscillations. *Central European Journal of Biology*, 6(5):699–711, 2011.

[5] T. Konopka and M. Rooman. Gene expression model (in) validation by fourier analysis. *BMC systems biology*, 4(1):123, 2010.

[6] Y. Lazebnik. Can a biologist fix a radio?–or, what i learned while studying apoptosis. *Cancer cell*, 2(3):179, 2002.

[7] Juliane Liepe, Chris Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, and Michael P.H. Stumpf. Abc-sysbio approximate bayesian computation in python with gpu support. *Bioinformatics*, 26(14):1797–1799, 2010.

[8] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[9] H.P. Mirsky, A.C. Liu, D.K. Welsh, S.A. Kay, and F.J. Doyle. A model of the cell-autonomous mammalian circadian clock. *Proceedings of the National Academy of Sciences*, 106(27):11107–11112, 2009.

[10] C.G. Moles, P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, 13(11):2467–2474, 2003.

[11] B. Pfeuty, Q. Thommen, and M. Lefranc. Robust Entrainment of Circadian Oscillators Requires Specific Phase Response Curves. *Biophysical Journal*, 100:2557–2565, June 2011.

[12] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

[13] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009.

[14] J.J. Tyson, K.C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

[15] B. Walsh. Markov chain monte carlo and gibbs sampling, 2004.