4bl Optimization as a Tool for Predictive Modeling of Biological Systems and Mining of Large-Scale Data Sets

Mano R. Maurya and Shankar Subramaniam

Abstract:

Due to recent advances in instrumentation, data collection and data storage capabilities, large-scale data are available. Typical examples include microarray data, mass-spectrometric data, and other cellular phenotypic data. Two issues requiring immediate attention in the context of large-scale data are: (1) utilization of the data to develop quantitative models with predictive accuracy, and (2) identification of critical features in the data via clustering (Xiao et al., 2003) or classification (Androulakis, 2005; Dettling and Buhlmann, 2002). Optimization has contributed substantially towards addressing both these issues. For data modeling, regardless of whether a biochemical model or a linear/nonlinear model with specific canonical structure (e.g. state-space formulation) is used, parameters can be estimated by minimizing fit-error between the data and model predictions. Optimization has been used for clustering and classification (Androulakis, 2005). Indirect forms of optimization, such as neural networks, support-vector machine, self-organized maps, etc., and least-squares approach also have been extensively used. However, for most problems resulting in a nonlinear program, there has been growing emphasis on finding the global or pseudo-global solutions.

The optimization technique used depends on several factors including the problem domain, criticality of global vs. pseudo-global solution, availability of suitable optimization software(s) that can efficiently formulate the optimization problem, and computational-complexity. Below, three projects are described in which we have used optimization extensively for understanding and quantifying the biological phenomena.

Modeling the Calcium response in RAW 264.7 macrophages: Calcium (Ca^{2+}) is an important second messenger. In macrophage, Calcium response mediates phagocytosis. Using experimental data on the concentration of free cytosolic Calcium upon stimulation of the cell with various ligands available from the Alliance for Cellular Signaling (AfCS), a quantitative model has been developed. The parameters are estimated using a stochastic-search-based nonlinear optimizer. The main challenges to be addressed are: (1) estimation of many parameters (reaction-rate constants, concentration of certain proteins that bind to Ca^{2+}, etc.) from experimental data on a single species (cytosolic free Ca^{2+}), and (2) variation in response in different cells (repeats) for the same dose of the stimulant (ligand). So, some of the initial states also are optimized and data from multiple cells is used to increase the confidence in the estimates of the common parameters.

Model-reduction using mixed-integer nonlinear programming (MINLP): Even with the recent surge in the availability of computing power, estimation of parameters for a detailed biochemical model is challenging due to the nonlinear and complex interactions. Hence, majority of the analyses with detailed models utilize parameter-values available from legacy knowledge on comparable systems. Thus, a reduced-order model that can sufficiently describe the system under all the scenarios of interest including dynamics while maintaining the known modularity of the subsystems is useful. This is particularly true if a module is found in many pathways. The GTPase cycle-module of heterotrimeric G-protein signaling is one such module. The total number of distinct types of GTPase cycle-modules in human cells could be more than 20. In this project, stochastic-search-based MINLP is used to estimate the parameters and to identify the topology of the reduced reaction-network simultaneously (Maurya et al., 2005).

Reduction in the number of false-positives via model-reduction: Due to cancellation-effects, most of the empirical models based upon high-dimensional input data suffer from the drawback of high number of false-positives. Though correlation analysis such as principal component regression (PCR) can be used to develop models that include all important predictors, the problem of false-positives still remains. Model-reduction can be used as a guiding procedure to develop minimal model(s) so that at least one of these minimal models captures most of the important effects. This approach has been used to identify the important signaling pathways contributing to the production of cytokines in RAW 264.7 macrophages. Since the number of predictors was not large, an exhaustive combinatorial-search was used for model-reduction.

Key words: optimization, parameter-estimation, model-reduction, false-positive, signaling pathways, Calcium.

References:

Androulakis, I. P., "Selecting maximally informative genes", *Computers & Chemical Engineering*, **29**(3), 535-546, 2005.
Dettling, M. and P. Buhlmann, "Supervised clustering of genes", *Genome Biology*, **3**(12), RESEARCH0069.1-0069.15, 2002.
Maurya, M. R., S. J. Bornheimer, V. Venkatasubramanian and S. Subramaniam, "Reduced-Order Modeling of Biochemical Networks by Simultaneous Determination of Network Topology and Parameters", *Foundations of Systems Biology in Engineering FOSBE 2005, University of California, Santa Barbara, CA, Aug. 7-10*, 2005 (accepted for presentation).
Xiao, X., E. Dow, R. Eberhart, Z. Ben Miled and R. J. Oppelt, "Gene Clustering using Self-Organizing Maps and Particle Swarm Optimization", *Second IEEE International Workshop on High Performance Computational Biology, Nice, France, April 22*, 2003 (http://www.hicomb.org/papers/HICOMB2003-06.pdf, May 24, 2005).

[1] Corresponding author: E-mail: mano@sdsc.edu, Phone: (858) 822 5403, Fax: (858) 534 8303.