Seminars in Cell & Developmental Biology 35 (2014) 109–123

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Making models match measurements: Model optimization for morphogen patterning networks

J.B. Hengenius[c], M. Gribskov[c], A.E. Rundell[a], D.M. Umulis[a,b,*]

[a] Weldon School of Biomedical Engineering, Purdue University, 206 S. Martin Jischke Drive, West Lafayette, IN 47907, United States
[b] Department of Agricultural and Biological Engineering, Purdue University, 225 S. University Street, West Lafayette, IN 47907, United States
[c] Department of Biological Sciences, Purdue University, 247 S. Martin Jischke Drive, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Available online 9 July 2014

Keywords:
Developmental biology
Mathematical modeling
Dynamic modeling
Objective functions
Parameter estimation
Morphogens

ABSTRACT

Mathematical modeling of developmental signaling networks has played an increasingly important role in the identification of regulatory mechanisms by providing a sandbox for hypothesis testing and experiment design. Whether these models consist of an equation with a few parameters or dozens of equations with hundreds of parameters, a prerequisite to model-based discovery is to bring simulated behavior into agreement with observed data *via* parameter estimation. These parameters provide insight into the system (*e.g.*, enzymatic rate constants describe enzyme properties). Depending on the nature of the model fit desired – from qualitative (relative spatial positions of phosphorylation) to quantitative (exact agreement of spatial position and concentration of gene products) – different measures of data-model mismatch are used to estimate different parameter values, which contain different levels of usable information and/or uncertainty. To facilitate the adoption of modeling as a tool for discovery alongside other tools such as genetics, immunostaining, and biochemistry, careful consideration needs to be given to how well a model fits the available data, what the optimized parameter values mean in a biological context, and how the uncertainty in model parameters and predictions plays into experiment design. The core discussion herein pertains to the quantification of model-to-data agreement, which constitutes the first measure of a model's performance and future utility to the problem at hand. Integration of this experimental data and the appropriate choice of objective measures of data-model agreement will continue to drive modeling forward as a tool that contributes to experimental discovery. The *Drosophila melanogaster* gap gene system, in which model parameters are optimized against *in situ* immunofluorescence intensities, demonstrates the importance of error quantification, which is applicable to a wide array of developmental modeling studies.

Contents

Abbreviations: Bcd, Bicoid; Gt, Giant; Kni, Knirps; AP, anterior–posterior; SDD, source diffusion decay; GRN, genetic regulatory network; DV, dorsal ventral; RMSE, root mean square error; SSE, sum of square error; OLS, least squares error; MSE, mean square error; MAE, mean absolute error; PCC, Pearson correlation coefficient; PDF, probability density function; *K–S* statistic, Kolmogorov–Smirnov statistic; CDF, cumulative density function; RE, relative entropy; *K–L* divergence, Kullbeck–Leibler divergence; wSSE, weighted sum of square error; nRMSE, normalized root mean square error; DSW (DTW), dynamic space (time) warping; SA, sensitivity analysis; LHS, Latin hypercube sampling; MB-ODE, model based optimal design of experiments.

* Corresponding author at: Department of Agricultural and Biological Engineering, Purdue University, 225 S. University Street, West Lafayette, IN 47907, United States. Tel.: +1 7654941223.
E-mail address: dumulis@purdue.edu (D.M. Umulis).

http://dx.doi.org/10.1016/j.semcdb.2014.06.017
1084-9521/

1. Introduction

Mathematical models of complex networks in development exist in an "uncanny valley"; many models look and behave *almost* like the natural systems they are designed to simulate, but they display imperfections that make their predictions suspect. The disparity between a model result and the actual system may be a small yet systematic mismatch, the complete absence of frequently observed experimental features, or the prediction of unviable conditions (*e.g.*, fatal pH) despite good agreement with experimental data. This "uncanny valley" for models might suggest that modeling is a distraction that interferes with experimental discovery because the model attempts to show how the system works in quantitative detail, yet models are always deficient. Among model-builders it is understood that simulations will always be simplifications incapable of reproducing all experimental behaviors; however, imperfect models still promote greater understanding and have, more recently, been informing experimental design and testing assumptions when experiments are infeasible [1].

Central to modeling are the needs to quantify how well a model agrees with experimental data and to identify where it might disagree. Quantification of model-data agreement is determined by an objective function that measures the "error" of the model; however there are many ways to measure the error and the choice of objective to measure model-data differences depends on the type of data, the type of model, and the question being asked. Herein we review diverse objective functions for the calculation of model-data error and identify each function's strengths and weaknesses in the context of developmental pattern formation by morphogens.

Mathematical models of varying complexity are used to represent diverse dynamic phenomena in the biological sciences. The specific type of model determines both the type of data needed to inform the model and the optimal objective functions to relate the model to the data. A dynamic model describes change in the system state over a time course of interest; it contains explicit mechanistic descriptions of the system and rules for updating the state of the system in time [2]. Independent of the mechanistic description, the behavior of the model depends on the initial conditions of the system (*e.g.*, simulated molecular concentrations at time zero). Developmental models often simulate spatially heterogeneous systems; in these cases the shape of the spatial domain also affects outcome. Mechanistic dynamic models are parametric [3]. In addition to the state of the system and its domain, parameters are constant values that define the behavior of the system and often have biophysical interpretations. For example, binding rate constants are parameters of receptor binding models [4,5]. To determine the validity of a model, parameter estimation must be used to bring the model into agreement with data [6]. This often involves iteratively simulating the model with different parameter values and comparing the resultant simulation to data. Parameters that yield simulated values minimally different (or maximally similar) to data are retained [6–8]. The difficulty of this parameter search depends on the range each parameter is allowed to assume, the number of parameters to be estimated, and the covariance of parameters with model output [9].

Mechanistic models should not be confused with statistical models (sometimes known as phenomenological models). Statistical models (*e.g.*, linear or logistic regression) quantify correlation among observable data. This knowledge often proves useful in hypothesis generation, but the predictive power of statistical models is limited to interpolation within the range of existing data [10]. Conversely, mechanistic models encode suppositions about the nature of the underlying system. As such, they may be used to extrapolate beyond the range of current data and provide predictions *given that the modeled mechanism is accurate*. Mechanistic models are the primary context for the comparison of fitness metrics herein.

The quality of the model and the uncertainty of its predictions depend on the type and quality of the data used for the training and optimization of the model. Experimental data common in the analysis of morphogen signaling systems may take several forms depending on the nature of the assays used. Specifically, qualitative data encodes nonnumeric descriptors of the morphogen and targets of interest; semi-quantitative data is predominantly ratiometric such as the relative intensity of a stained molecule or intensity of a western blot; and quantitative data provides information of specific, measured quantities with associated uncertainty. As the quantitative content of the data increases, the associated uncertainties typically decrease. This provides more stringent constraints that improve the resulting model (see Pargett et al., 2013 for further details [11]).

Once mechanistic models are trained or optimized to the supporting data, they can be used to address a number of important questions. Specifically, a parameterized model can be used to infer the behavior of hard-to-observe molecules, perform quantitative simulations of qualitative hypotheses, or generate new hypotheses based on model behavior. In Section 2 we focus on the challenges that exist in most model optimization problems and then utilize a specific example in the *Drosophila* gap gene network as an illustrative case study.

2. Model complexity and parameter estimation

Dynamic modeling allows insight into systems' behaviors, but this insight requires optimized physiochemical parameter values. Several challenges stand between a newly defined mechanistic model and the parameter values that make it biologically relevant. This parameter estimation problem grows exponentially as the number of modeled (and parameterized) biochemical interactions grows.

Model and objective function in hand, optimization proceeds in several steps. First, the unknown parameter values are enumerated and constrained to biologically feasible ranges (*e.g.*, a kinetic constant or diffusion constant cannot be negative). Second, a stochastic and incomplete search is performed within this feasible region of