Published in final edited form as:

Semin Cell Dev Biol. 2014 November ; 35: 109–123. doi:10.1016/j.semcdb.2014.06.017.

Making models match measurements: Model optimization for morphogen patterning networks

JB Hengenus^a, MR Gribskov^a, AE Rundell^b, and DM Umulis^{a,b,1}

^aWeldon School of Biomedical Engineering, Purdue University. 206 S. Martin Jischke Drive West Lafayette, IN 47907

^bDepartment of Agricultural and Biological Engineering, Purdue University. 225 S. University Street West Lafayette, IN 47907

^cDepartment of Biological Sciences, Purdue University. 247 S. Martin Jischke Drive West Lafayette, IN 47907

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.

© 2014 Elsevier Ltd. All rights reserved.

¹Correspondence: dumulis@purdue.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Developmental biology; mathematical modeling; dynamic modeling; objective functions; parameter estimation; morphogens

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 the parameter space while ignoring biologically impossible parameter value combinations. Third, resultant parameter sets are ranked according to the chosen objective function. Finally, investigators decide whether the fit is “good enough.” If it is, they stop; if not, they repeat the above search.

The magnitude of these difficulties grows with the size of the model and the increasing number of unknown parameters. As the number of parameters increase linearly, the feasible region grows exponentially. This is known as the curse of dimensionality. When an objective function is mapped to feasible parameter combinations, the resulting map is called a cost landscape (see Fig. 1A for a one-parameter landscape). Searches seek the lowest point – the global minimum – to minimize the cost [6,7]. Local search methods begin at an expert's “best guess” and proceed “downhill” to the nearest minimum (black arrows, Fig. 1A). Because parameter values are uncertain, these methods usually lead to a local rather than the global minimum. To avoid such traps, randomized and incomplete global searches are used to sample the whole feasible region[12]. However, the curse of dimensionality means that smaller proportion of the space is covered. Parameter estimation scales poorly with model complexity. As a result, parameters reported in the literature almost always represent local rather than global minima. Multiple global searches are sometimes run to sample multiple local minima.

When searches return optimized parameter sets, these sets are ranked by their objective values. Those with the lowest objective values are retained, though investigators often inject an element of subjectivity by ranking the lowest cost parameters “by eye.” This expert evaluation, coupled with the probabilistic nature of the search makes replicability of model analyses more difficult.

2.1. Bicoid gradient formation models

In *Drosophila melanogaster*, initial expression of gap genes takes positional cues from heterogeneously deposited maternal mRNAs and their resulting transcription factors [13]. The anterior determinant, Bicoid (Bcd), is expressed from anteriorly distributed mRNA and forms a decreasing concentration gradient from the anterior to posterior of the embryo [14]. Though Bcd has been studied since the 1980s[14–21], mathematical modeling of its gradient formation is a more recent enterprise [15,17,22–24]. These models represent the anterior-posterior (AP) axis of the embryo as a 1D spatial domain. In these early models, the anterior end of the line was a point source for Bcd production and expressed protein would diffuse toward the posterior. To avoid saturation of the AP axis by accumulating Bcd, Bcd degradation was modeled by a first order decay term [15,17]. Collectively, these source-diffusion-decay (SDD) models were numerically easy to calculate and contained parameters with explicit biophysical meanings (e.g., diffusion constants). The models replicated the stationary Bcd gradient observed in data, though the time required to reach this state did not agree with comparable experimental time course observations. These discrepancies prompted further model development, demonstrating the utility of imperfect model results. Ibanez and Belmonte provide an excellent review of these SDD modeling approaches [25].

Though Bcd is relatively simple to model, it is vital for the establishment of the gap gene patterning. The gap genes represent considerably more complex modeling approaches and are discussed in Section 2.2.

2.2. Gap gene patterning

Modeling has played a large role in the evaluation of the *Drosophila melanogaster* gap gene system. Driven in part by maternal mRNAs such as Bcd, this system of genetically interacting transcription factors forms increasingly specific expression bands along the fly embryo's anterior-posterior axis [13]. First experimentally characterized in the 1980s, decades of experimental work have gathered a wealth of data with which to fuel model-driven discovery [26–32]. Varied modeling approaches have been applied to this system over the last two decades, each presenting different challenges to model fitting. Early models (*e.g.*, Sanchez and Thierry [33]) incorporated pre-defined genetic regulatory networks (GRNs) inferred from expert interpretation of mutant data rather than using parameter estimation to infer GRNs. To compensate for contemporary computational limitations, these simulations frequently incorporated simplifying assumptions such as discrete protein levels, discrete spatial domains, and discrete time updates. The qualitative nature of the model output limited error calculations; protein data might be classified as high, low, or absent and compared to similar model output.

The work of Jaeger *et al.* [34,35] initiated quantitative modeling for the purpose of GRN inference via parameter estimation. This approach led to numerous analyses [36–43], but also brought new challenges to model fitting and parameter identification. Building on the partial differential equation simulation methods of Mjolsness and Reinitz [24], these models moved gap genes from discrete time and concentration simplifications to simulation of continuous concentrations and time on a 1D domain. Rather than building a model with *ad hoc* assumptions about the nature of GRN interactions, Jaeger *et al.* built a general model framework in which every gene had the potential to enhance or inhibit the expression of every other gap gene; they then fit the model to immunofluorescence expression data and observed which optimized GRNs minimized the unscaled sum of squared error between model and data. This approach generated a set of similar GRNs that recapitulated expression patterns and agreed with genetic interactions previously proposed from mutant data. While the computationally inferred GRNs are consistent with expectations, a major shortcoming remains: When gap genes were knocked out in simulations, the resulting protein distributions did not match available mutant data. This suggests that additional regulatory genes may be missing from the model, that the large number of parameters is causing overfitting, and provides further avenues for study.

All of the preceding models (except the early discrete variants) attempt to fit 25+ parameters; this high-dimensional parameter space *may* lead to two related problems: overfitting and non-unique solutions. In statistical models, overfitting refers to overly parameterized models which predict noise rather than underlying trends [10]. Similarly, overfitting of dynamic models involves the distribution of error among many parameters during fitting; this may lead to spurious inferences from parameter estimation. While parameter estimation seeks the global optimal point – the best possible fit in the parameter space – high dimensional parameter spaces may contain many locally optimal parameter sets that produce equivalent fits (Fig. 1A). Indeed, when two parameters affect the same model output (*e.g.*, when two genes A and B enhance expression of a target gene C), the parameters regulating $A \rightarrow C$ and $B \rightarrow C$ are said to be correlated and unique parameter values

will not be identifiable [12,44]. Because the approach of Jaeger *et al.* allows all gap proteins to interact (thus estimating 36+ interaction parameters), there is no guarantee of the uniqueness of estimated parameters and overfitting is likely. Fomekong-Nanfack *et al.* found that it was impossible to distinguish between enhancer and repressor activity in fitting results because of parameter correlation [36,40]. The sensitivity analysis of Bieler *et al.* also indicates correlation among model parameters, limiting the scope of model inference [45].

Finally, it must be noted that all of the preceding studies have utilized either qualitative fits (in the case of discrete models) or variants of the sum of squared error measure (*e.g.*, MSE, SSE, RMSE, and weighted SSE). The limitations of these measures in fitting relative spatial positioning data (such as adjacent gap genes) will be discussed in Section 3.

3. Objective functions

Objective functions compare model and data to produce lower residual scores for data-model agreement. In concrete terms, different objective functions compare different features of the model and experimental data. Two different metrics may produce conflicting rankings for the same sets of parameter values. As a result, the “terrain” of a cost landscape is determined by the choice of metric leading to different possible outcomes depending on the initial design choice. To demonstrate how each objective function responds to different types of model-data mismatches, data of the *Drosophila* gap gene *Knirps* (Figs. 2–4, brown diamonds) is compared against a set of artificially prepared erroneous distributions (Figs. 2A–4A, blue and red lines). To allow for ready production of the different types of mismatches, these distributions were generated from Gaussian distributions as a proxy for models described in Section 2.2. This enabled manipulation of spatial positioning (Figs. 2A–4A), width (Figs. 2B–4B), and aspects of magnitude (Figs. 2CD–4CD).

3.1. Pairwise measures

The most commonly used cost functions are drawn from a class of pairwise measures. These objectives are computed by comparing corresponding pairs of experimental and simulated data. The final objective is calculated using an aggregate of the individual pairwise residuals (or a subset of those residuals). Example equations for pairwise objectives are provided in Box 1.

The sum of squared error (SSE), its square root the Euclidean distance, and its sample-normalized variant the root mean squared error (RMSE) are the most frequently used error measures in biological modeling studies. As its name implies, the SSE (also called ordinary least squares, OLS) is the summation of each squared pairwise residual. Squaring the pairwise residuals serves two purposes. It prevents positive and negative residuals from partially canceling in the summation (thus underestimating the model-data mismatch). It also emphasizes larger residuals due to the super-linear growth of the square function. This intrinsic weighting translates to an error measure that is more tolerant of small residuals (*e.g.*, experimental noise at the optimal fit, Fig. 2B,1-4, data vs. optimal fit), while penalizing larger model-data disagreement with a weight proportional to the square of the mismatch (extrema of Fig. 2B,1-4). For global parameter searches that use gradient information from the cost landscape, this approach “smoothes” the landscape (cf. Figs.

1B,SSE and 1D,K-S). In ideal situations, this smoothing may improve convergence toward minima, whether local or global. The data-averaged variant is the mean square error (MSE), calculated by dividing the SSE by the number of data points.

The square root of the SSE is the intuitive Euclidean distance metric (also known as the ℓ_2 norm and straight line distance). In everyday contexts, this is the familiar quadrature sum used to calculate the distance between points in space. Unlike the SSE, the Euclidean distance does not incorporate intrinsic residual-dependent weighting. This is exemplified in the one-parameter landscapes displayed in Fig. 1. SSE has lower relative costs corresponding to negative parameter values; the relative cost increases in Euclidean distance (cf. Fig. 1B,SSE and 1B,Euclidean). Because the SSE is always nonnegative in practice, the cost landscape of the Euclidean distance will always share a global minimum with the SSE. However, the contours of the cost landscape will differ in that small errors are not smoothed from the landscape. This is shown in Fig. 1A,1,4 in which SSE ignores the low shoulders of the widened peak whereas Euclidean error takes them into account. Once global search algorithms have converged on the neighborhood of a minimum, the Euclidean metric provides a more stringent criterion by which local optimization homes in on the exact minimum value. A common normalization of the Euclidean distance is the root mean squared error (RMSE). The RMSE is computed by dividing the SSE by the number of data points before taking the square root. As demonstrated by Fig. 2A,C, this scalar normalization leads to identical landscapes for RMSE and Euclidean distance.

Though calculation time of the objective function is often negligible compared to the evaluation time of complex models, efficient objective calculations are sometimes called for. One such measure is the Manhattan distance (also known as the ℓ_1 norm, absolute distance, city block distance, and taxi cab distance), so named because it represents the distance traveled between two points on a regular grid such as New York's street system (see Box 1 for an illustrated example). As such, the Manhattan distance is always greater than or equal to the corresponding Euclidean distance. In practice, this objective is computed by summing the absolute values of each pairwise error. This serves the same purpose as squaring in the SSE; it prevents partial cancelation of positive and negative errors in the final distance value. The mean absolute error (MAE), the related per-datum error, is computed by dividing the Manhattan distance by the number of data points. The Manhattan distance and MAE are not without drawbacks. They are not differentiable, making it impossible to analytically determine which direction on the cost landscape leads “downhill” to minima. These derivatives are used in some optimization algorithms. In practice, the derivatives of these error measures must be numerically estimated by evaluating the model multiple times. When using these algorithms, this numerical estimation may negate the increased efficiency derived by use of the Manhattan distance. The Manhattan distance has landscape behavior qualitatively similar to that of Euclidean and RMSE; this is demonstrated in Figs. 1B and 2F,1-4.

In practice, any ℓ_j norm may be computed for real numbers by summing the absolute residuals raised to the j th power and then taking the j th root of the result (see Box 1). As indicated in discussion of the SSE and Euclidean distances, larger values of j place larger intrinsic weights on larger residuals. The logical conclusion of this trend is the ℓ_∞ norm, or

Chebyshev distance. This objective places all the weight on the largest residual and no weight on lesser residuals. The Chebyshev distance is computed by calculating the absolute residuals and then returning the largest residual. The objective shares the Manhattan distance's drawback of non-differentiability. Determining the maximum residual is also computationally expensive for large datasets. The resulting cost landscape is often extremely flat, making optimization difficult. However, the Chebyshev distance is useful for mini-max optimization strategies. This entails minimization of the maximum residual between simulation and data, effectively finding the “best” worst-case parameter set and is reflected in a landscape that differs qualitatively from other pairwise measures (Figs. 1B, extrema of 2E,1, overall shape of 2E,2–4).

In each of the preceding error measures, weighted and/or normalized variants are also in use. These are described in Section 3.5.

3.2. Whole-dataset measures

Pairwise error measures incorporate transformations and summations of residuals to report a scalar error value. In contrast, whole-dataset measures are computed using properties of the entire experimental and simulated data sets such as arithmetic mean and variance. These measures are most commonly encountered in statistical contexts (*e.g.*, the Pearson correlation coefficient) and machine learning and bioinformatics applications (the cosine similarity), but are rarely seen in dynamic model fitting. Implemented in isolation, these methods have limitations that restrict their use. These measures sometimes produce qualitatively similar landscapes (Fig. 1C), indicating that they share some of the same limitations.

The Pearson correlation coefficient (PCC) is often associated with linear regression. It is computed as the covariance of model and experimental data divided by the product of their standard deviations (see Box 2). This objective quantifies the variability of paired (experimental, model) data points relative to the least squares linear regression of the data; a value of one represents perfect correlation while negative one indicates perfect inverse correlation. As such, a large positive value of this metric *might* indicate a linear relationship between the experiment and model (Fig. 3C,4) while the slope of the regression line provides a measure of linear scaling between the two sets. The chief limitation of this objective is its dependence on a linear relationship between simulated and experimental data. The variability measured by this objective may represent normally distributed residuals about the regression line. However, nonlinear errors may still produce relatively large correlation coefficients (Fig. 3C,2). Anscombe presented four examples that illustrate the measure's insensitivity to nonlinearities[46]; his results are reproduced in Box 2. Due to this limitation, correlation measures provide additional information about variability when used in conjunction with other measures. Used in isolation, PCC may lead to unrealistic parameter estimation.

The cosine (or angular) distance and its relative the cosine similarity provide another means of comparing two datasets. It is often used in machine learning and clustering algorithms as a means of measuring distances between vectors of features used for classification; it is not often used to fit biological dynamic models. Rather than pairwise comparisons in two

dimensions (model and experimental data), this objective considers each of the n paired data to represent a dimension in an n -dimensional data space. The model and experimental data are represented by two respective vectors in this space and the angle between these vectors is computed. The cosine of this angle will be one if the vectors point in the same direction and negative one if they point in the opposite direction, though this cosine similarity disregards scaling. It thus compares the overall shape of spatial morphogen distributions while lessening penalties for differences in total concentrations (Fig. 3B,2–4). Because cost functions represent poor fits with larger objective values, cosine similarity is often normalized from $(-1,1)$ to $(0,1)$ and then subtracted from one (see Box 2). This produces a measure that is one when the data vectors point in opposite directions of the data space and zero when the vector point in the same direction. The main shortcoming of this measure is its blindness to the relative magnitudes of model and experimental data. If a parameter set leads to a simulated dataset in which all morphogen concentrations are near zero, computational rounding errors and limited machine precision create varied objective results, which may be erroneously low – representing a good fit where none exists.

3.3. Information-theoretic measures

While the whole-dataset objectives are used to compare data vectors (albeit not always in the context of biology), this next class of information-theoretic objectives is used for comparison of probability density functions (PDFs). Here, we treat morphogen distributions as analogous to PDFs over an organism's spatial domain. In typical use, experimental or simulated sampling provides data to estimate the PDF(s) of interest. When the information measures are used as objective functions for model and experimental data, the intermediate sampling step and PDF estimation are not necessary. In place of sampling, the model and experimental data are treated as morphogen concentration density functions over the spatial domain of the. With these PDF analogues in hand, the information measures may be brought to bear, though the PDF analogues do not always meet conditions for a robust information-theoretic interpretation.

A common measure for PDF comparison is the Kolmogorov-Smirnov D-statistic (K-S statistic). The procedure for calculating the K-S statistic is similar to that of the Chebyshev distance (Section 3.1). Though the Chebyshev distance corresponds to the maximum absolute residual of the model and experimental data, the K-S first transforms the data- and model-sets y and $\hat{y}$ to the cumulative density functions (CDFs) F_y and $F_{\hat{y}}$ (Box 3). The K-S statistic is the maximum absolute residual between these CDFs. Unlike the Chebyshev distance, K-S incorporates an implicit linear scaling. Regardless of relative magnitudes of the model and experimental data, transformation to CDFs bounds the functions from zero at one end of the spatial domain to one at the other. Like cosine distance (Section 3.2), the scaling emphasizes the overall shape of the data while relaxing the requirement that concentration values match exactly. This statistic still has limitations. As indicated in Figs. 2E, 3B and 4B, Chebyshev and cosine distances penalize common errors differently. The intrinsic scaling in the K-S objective *should* penalizes mismatches in absolute concentration less harshly, though the estimation of CDFs F_y and $F_{\hat{y}}$ is very sensitive to experimental noise. Figure 4B,1–2 demonstrates the effect of noisy data: the minimal K-S values do not even align with the best-fit distribution. Objectives like cosine emphasize the

overall shape of the model and experimental distributions. Meanwhile, mini-max strategies like Chebyshev reduce the largest error – sometimes at the expense of an overall consistent fit.

Every distance measure discussed so far has been symmetric—the distance from y to $\hat{y}$ is identical to the distance from $\hat{y}$ to y . However, the relative entropy (also known as the Kullback–Leibler divergence) is asymmetric; these two distances are unequal. In the context of information theory, entropy is a measure of information. The relative entropy may be considered the amount of information lost if the model data $\hat{y}$ is used to approximate the true experimental distribution y [47]. To compute the relative entropy, the two data sets are each normalized so that the data in each distribution sum to one (satisfying the criteria for a discrete probability distribution). Rather than using residuals, each comparison between experimental and model datum i ($i = 1, \dots, n$) is performed by multiplying y_i by the logarithmic ratio of y_i over $\hat{y}_i$. The summation of these values yields the relative entropy. A visualization of this measure is found in Box 3.

This method is not without limitations; it is only defined if both $\hat{y}$ and y each sum to one (hence the normalization) and if y_i equals zero for any i where $\hat{y}_i$ equals zero. This limits parameter estimation: If the model dataset predicts no morphogen at a point in space that has nonzero experimental data, the relative entropy will be undefined due to division-by-zero in the log ratio. This means that some parameter sets (which cannot be identified in advance) will not have an associated error value. If y_i and $\hat{y}_i$ equal zero, the comparison becomes $0 \ln(0/0)$. If y_i is zero and $\hat{y}_i$ is not, $0 \ln(0)$ appear in the calculation. These two values are normally considered to be undefined. For the purpose of relative entropy calculation, both are considered to be zero; in each case, the limit of $x \ln(x)$ and $x \ln(x/x)$ is zero as x goes to zero.

3.4. Composite cost functions

Pairwise, whole-dataset, and information-theoretic error measures all have strengths and weaknesses. It is common for computational biologists to assemble *ad hoc* combinations of objectives into a single composite cost function. These functions are tailored to emphasize the data features pertinent to the current research or to combine models and/or experiments. For example, we may wish to combine the qualitative shape matching of cosine distance with the residual-dependent weighting of SSE. The resulting composite function may appear as follows.

$$\text{Composite Cost} = w_1 \text{Cosine Error} + w_2 \text{SSE}$$

For a given parameter set, *Cosine Error* and *SSE* represent the two objective functions. Each is weighted by positive constant values w_1 and w_2 . These weights serve multiple purposes. Consider that the cosine distance ranges from zero to one while SSE values are dependent on the magnitude of the data. If weights are not applied within the composite function, the either objective may dominate the composite value. Depending on the magnitude of the data,

one must choose a ratio of w_1 to w_2 to give each component objective comparable contributions to the final error.

Conversely, weights may be chosen to emphasize one component objective over another. Perhaps the foremost interest is in finding a qualitative shape match between experimental and simulated morphogen distributions. At the same time, we wish to avoid the inaccurately low error due to the cosine distance's blindness to magnitude differences. We may choose weights w_1 and w_2 such that the cosine distance contributes more to the composite cost than SSE. The exact ratio is difficult to identify in advance of parameter estimation. Often, we generate a set of synthetic experimental data (model data generated by a known parameter set), and weight ratios are tested to determine which ratio recovers the known parameter set most efficiently.

3.5. The importance of transformation and normalization

A common transformation of data is unit conversion. For example, immunofluorescent intensities are assumed to correlate with the concentration of the antibody targets. When a model simulates concentration and experimental data is recorded in intensity, we might perform a linear transformation ($c = ax - b$) on data. Here, c is the estimated concentration, a is a linear scaling factor, x is the intensity data, and b accounts for background fluorescence in the micrograph. This simple example assumes uniform background intensity. This assumption is not always satisfied in practice due to varied experimental limitations. Though we limit our discussion to linear transformations, the transformation is assay-dependent and may involve more complex functions.

Experimental data does not always require linear or nonlinear transformation. Some computational biologists design their models and parameter ranges such that the simulated morphogen values are directly comparable to data. This approach precludes the need for mapping experimental units to simulated units; instead, subtraction of a uniform or nonuniform background value is the only modification to experimental data.

Error values may also be transformed. Some cost landscapes are very “rough,” leading to narrow basins of attraction separated by high “peaks” Figs. 1B–D. Some optimization techniques are more tolerant of “smoother” cost landscapes [48], and transformations may be used to reduce the magnitudes of the peaks in the landscape. Often, a logarithmic transform is applied to the objective function values; it acts to compress error values, which may be orders of magnitude apart before transformation, to the same magnitude.

The weighted SSE (wSSE) scheme calls for division of each residual by the replicate standard deviation of its experimental data point. Formulaically, the i th squared residual has a weight $w_i = 1/\sigma_i$. Here, σ_i is the standard deviation of experimental datum y_i . When the replicate values for a given experimental datum are highly variable, the residual is divided by a large standard deviation and the contribution of the uncertain data to the error is minimized. Other *ad hoc* weighting schemes may be used depending on our needs. Normalization methods are also employed in problem-specific contexts. For example, fractional or normalized RMSE (nRMSE) divides the RMSE by the range of predicted

values; $w = \hat{y}_{max} - \hat{y}_{min}$. The resulting weighted residuals may be interpreted as a fractional or percentage errors in each simulated datum.

A special case of weighting addresses the comparison of model and experimental data spanning several types of chemical species. Intercellular morphogens and intracellular signaling molecules may have respective molarities differing by several orders of magnitude. In cases of pairwise measures without implicit linear scaling, this may lead to unequal contributions from each chemical species to the final cost value. For example, errors in a morphogen profile spanning a 100 nM range could completely mask fluctuating error contributions from a protein distribution with a 10 nM range. Similarly, comparison of untransformed immunofluorescence or histochemical data with compatible models may be hindered by differences in antibody affinity among measured species. Different chemical species' errors may contribute unequally to the error value. One solution to this problem has been noted above. The nRMSE uses fractional residuals. Each residual is weighted by the range of the simulated chemical species. For morphogen $M1$, each residual is multiplied by a weight $w_{M1} = \hat{y}_{M1,max} - \hat{y}_{M1,min}$. Likewise, all residuals for morphogen $M2$ are weighted $w_{M2} = \hat{y}_{M2,max} - \hat{y}_{M2,min}$, and so on for each additional chemical species. Substituting fractional residuals in place of residuals, thus ensuring that they contribute equally to the overall error.

When sampling a large parameter space, it can be difficult even to find qualitatively useful features like peaks, let alone peaks that align with experimental data. Though some objectives tolerate spatial changes better than others, all surveyed objectives save the Kolmogorov-Smirnov statistic have similar responses to spatially shifted peaks (Figs. 2C–F, 3B–C, and 4B–C). Dynamic space warping (DSW) provides a means of protecting qualitative agreement from positional mismatch. DSW is a nonlinear distortion of model values rather than a linear transform. It is an alignment algorithm that warps the spatial coordinates of two signals (such as immunofluorescence intensities) to minimize mismatch between the signals; this allows slightly offset peaks to be aligned before the objective is calculated and reduces penalties when evaluating distribution shapes. To demonstrate how the procedure, consider data and slightly offset model output (Fig. 5A). DSW traverses the length of the domain and calculates the residual of each model-datum pair in spatial order (an unwrapped path would be the diagonal of Fig. 5B). It iteratively calculates the residuals of the residuals at spatially adjacent points and then moves to the smallest residual value (Fig. 5C). This results in an alignment (white line in Fig. 5B) that minimizes errors of position (but not errors of magnitude) between model and data (Fig. 5D). To prevent drastic warping, the method is often constrained to a fraction of the total domain length. A concise introduction is Müller's "Information Retrieval for Sound and Music"[49], which describes the mathematically analogous technique of dynamic space warping.

Finally, a more flexible solution to the above problem is to employ a composite cost function. Rather than incorporating different measures of the same data, we may partition the data and compute the same objective function for each species. The weight ratios noted in Eqn.1 may simply equalize the contributions (equivalent to the nRMSE approach) or may be further modified to preferentially penalize some species if the research question calls for it.

4. Strategies for improved parameter estimation

Today, many modeling studies rely on “off-the-shelf” objective functions such as SSE. These objectives are not always appropriate for the biological question being addressed. Here, we provide a brief schema to screen objectives before committing computational resources and time to a large-scale search of the parameter space.

4.1. Identify appropriate error measures

The selection of appropriate error measures depends upon the type of data available (here assumed to be semi-quantitative), whether unit rescaling or normalization matters for the research problem, and the type of fit required (exact fit, shape matching, relative positioning, etc.).

The assumption of semi-quantitative data provides two important considerations. The data do not contain absolute concentrations of the values of interest, but rather recordings of some observable signal (immunofluorescent image intensity). Further, this implies that model outputs, if designed to accurately represent the underlying units of interest, may differ in units and by orders of magnitude relative to experimental observables. In some cases, this will necessitate transformation/scaling to bring the two datasets' units into agreement.

In the *Drosophila* gap gene case study above, the model parameters were scaled such that model output was bounded between 0 and 255 arbitrary concentration units. This was decided so that model output could be compared directly to experimental image intensity data; in 8-bit digital imaging, each pixel's unitless intensity ranges from 0 to 255. This type of model does not require transformation of data because the model was designed to avoid issues of unit and magnitude mismatch. This is not always straightforward in model design. The construction of large models often involves surveying the literature of chemical rate constants – typically incorporating units of molarity – resulting in molar outputs. As noted in Section 3.5, a linear transformation is used to convert intensity to molarity. In the absence of calibration standards, the choice of linear transformation parameters must be optimized before each comparison of model and experimental data. An alternative method renders both datasets unitless by dividing each set by its maximum value; this also scales the data range from zero to one. Finally, this preprocessing step may be avoided by use of implicitly-scaled error measures such as cosine distance or the information measures. In addition, if the relative spatial positions of biochemical species are an important criterion for optimization, data pre-preprocessing via dynamic time warping may serve to align spatially offset data before objectives are applied. A schema for objective function selection is shown in Table 1. Note the many more objective functions exist than are listed here; the Encyclopedia of Distance Measures provides brief descriptions of a greater number than can be discussed here [50].

4.2. Perform a basic global sensitivity analysis

To choose a well-behaved measure for the objective function, we first carry out a sensitivity analysis. Sensitivity analyses (SA) systematically perturb parameters to characterize a model's response to changes in parameter values. If the modeled behavior of a morphogen

changes drastically in response to perturbation of its diffusion constant D , the morphogen is said to be sensitive to changes in D [51]. Rather than determining the sensitivities of biochemical species to parameter perturbations, we suggest a simple SA to determine the sensitivity of cost functions to parameters.

Global sensitivity analyses reveal trends in objective values over the entire parameter search space (e.g., parameter regions where objective values are lower), but more informative analyses require additional model simulations. To save time, we recommend using an extremely simple global sensitivity approach: Randomly sample the parameter space and visually compare changing objective values versus each parameter. This will reveal direct effects of parameters on objectives, but not parameter interactions.

We recommend the Latin Hypercube Sampling (LHS) algorithm for fast sample selection. It is quasi-random and guarantees more uniform sampling of the parameter space than uniform random sampling [51]. The number of points in the sample will vary depending on the computational cost of the model and the size of the space. This method serves a preliminary screen before time-intensive parameter estimation, so we err on the side of efficiency rather than completeness: We use a heuristic of at least $1000d$ sampled parameter sets, where d is the number of parameters. (For the 1D *Drosophila* gap gene model, 42,000 evaluations take approximately six hours.) When the model in question is particularly fast, we may choose to increase the number. For models that require hundreds of seconds to evaluate, we may omit this step.

After sample parameter sets are selected and the model is evaluated, we compute all the objective functions chosen in Section 4.1 at each sample point. This creates a list of objective values that corresponds to each parameter set. With this information, we plot the each objective value against each model parameter. Because objective values are only plotted against one parameter at a time, the variability of other parameters adds noise to the plot; visible trends in these plots indicate global trends in objective sensitivity.

Regions of distinct high and low objective values are indicative of promising search areas (low cost regions) within the parameter space. These objective functions should be retained. Conversely, some objectives may produce only uncorrelated scatterplots; while minima exist in the corresponding cost landscapes, the surface is likely “flat” and “rough.” This may present difficulties during optimization[52]

4.3. Choose an objective or objectives

Using the above information, we may make an informed choice of objective functions rather than defaulting to SSE or RMSE. If visual analysis yields cost functions with global trends these are recommended for use in parameter optimization.

Scatterplots without any global trends do not indicate that the selected cost functions are inappropriate for parameter estimation. The sampling performed by random or LHS sampling is incredibly sparse when compared to the full-factorial sample. There may not be enough coverage to identify global trends. While it is not necessary to exclude objective functions without global trends, we consider removing them from subsequent steps unless

(a) all objectives are uncorrelated or (b) objectives showing global trends fail in subsequent steps. Here we may choose to retain at least one stringent cost (*i.e.*, a magnitude matching function) and an error tolerant (shape matching) function.

4.4. Parameter optimization by sequential optimization

With objectives in hand, it is now possible to search for optimal parameter estimates. This step costs the most in terms of computation terms and we have undertaken the previous steps to identify maximally effective objective functions used in this step. We recommend a three-step process: It begins with another stochastic sampling to screen for promising parameter sets. Then these initial sampled points are used with a stochastic global search method (*e.g.*, genetic algorithms (GAs) or simulated annealing (SA))[53]. Finally, the best point(s) are locally optimized.

For the first task, choose the least stringent objective function you have retained from previous steps. The purpose of this global screen is to generate candidate parameter sets for further refinement. To do so, repeat the stochastic sampling discussed in Section 4.2. The number of points required is dependent on the global search algorithm you choose. The choice of a lenient objective ensures that promising minima and basins of attraction are not discarded prematurely. Dynamic space warping (DSW), normally used in face- and handwriting-recognition may also be used if appropriate to the research goals[54]. From the results of the screen, identify the lowest-error point(s) as required for your chosen global search algorithm (such as the initial population of a GA or the starting point of SA).

The second task is the time-intensive parameter estimation step that would typically begin with a pre-chosen objective such as SSE or RMSE. Here, the promising parameter dataset(s) from the previous screen are used to initialize the stochastic search of the parameter space. Depending on available resources, this may be run once or multiple times to generate one or more optimal parameter sets for analysis.

For the third task, screen the resulting point(s) with a follow-up local search. The stopping criteria of global optimization methods generally stop the search short of the minimum of a basin. The purpose of the subsequent local search is to obtain the nearest minimum in the region of the solution found by the global search. We recommend switching to a more stringent objective for this task. The retained parameter sets should already be in the basins of the minima we wish to reach, so the risk of entrapment in another local minimum is lessened. The more stringent objective will enforce a closer fit between model and data.

5. Conclusions

Models that have been rigorously constrained to the available data can be used in a number of ways that complement the existing tools for experimental discovery.

Two primary motivations for using modeling as part of the repertoire of experimental methods are to test new hypotheses with the model or to use the data obtained to reverse engineer the networks and/or parameters of signal regulation. For instance, while many mechanisms are testable *in principle* and may serve as inspiration for further

experimentation, technical limitations such as the lack of appropriate antibodies, lack of equipment, and reagent costs sometimes make testing impractical. Using parameter estimation, mechanistic models may act as surrogate systems for interrogation when direct experiments are impractical. These surrogates might indicate an informative subset of possible experiments on which to expend limited resources.

Whereas hypothesis testing typically proceeds from identification of falsification criteria to experiment, dynamic models permit alternative approaches. When multiple explanations for observed behavior are under consideration and it is impractical to measure key concentration or kinetic values, each mechanism may be encoded in a separate model structure. So long as models simulate overlapping observable concentrations and kinetics, they may be fit to previously collected data. The hypotheses (models) are then ranked according to agreement with data [55]. Because these models each describe a single hypothesis, this surrogate approach is most useful for testing the sufficiency of proposed model mechanism(s) to explain observations. However, its scope is limited compared to direct experimentation. Simulated results may be used to rule out insufficient mechanisms but are only valid assuming the model is true. Experiments provide new data about the system even if they lead to falsification of prevailing hypotheses.

After models are ranked according to their ability to recapitulate existing data, analysis can proceed in two directions. If none of the hypotheses produce agreement with data, further hypothesis generation is called for. If one or more models are consistent with available data, the results can suggest additional experimental falsification criteria to discriminate between consistent models. For example, the competing viable models may predict distinguishable dynamics for observable but unmeasured molecular species. In this case, models drive experimental design by identifying the experiments that will have the greatest discriminatory power to resolve differences in the data-consistent networks, a process known in engineering as Model-Based Optimal Design of Experiments (MB-ODE) [1,56].

When parameter estimation is used to fit multiple models to data, it allows us to judge the sufficiency of competing mechanisms to explain observations. When investigators are confident in a particular mechanism, parameter estimation provides a framework to make holistic inferences about the underlying biology. The last decade has evidenced increasing enthusiasm for systems biology with over 15000 instances present in the NCBI PubMed database (8500 published since 2010)[57]. Despite the promise of whole-system analysis, the cost of comprehensive data collection and bottlenecks in data analysis present challenges to its implementation. Here again, models may serve as surrogates for costly batteries of high-throughput perturbation experiments.

Once estimated parameter values are selected that approximate observed data, sensitivity analyses are used to test the fragility and robustness of the system to parameter changes and state perturbations. Clinically, fragile parameters may suggest targets for therapeutic intervention in signaling models of disease [58]. Conversely, robustness may be a design criterion in synthetic biology [59]. In developmental biology, robustness is related to the concept of canalization [60]. More generally, these predictions may inform experimental design for hypothesis falsification: Models predict which experimentally observable species

are sensitive to small experimental perturbations to suggest experimental targets for model (hypothesis) discrimination.

Beyond model discrimination and inference about system behavior, parameter estimation may be used to infer the structure of biological pathways and genetic regulatory networks [34,35,37,39,61,62]. Reverse engineering of the gap gene network in *Drosophila* demonstrates both the utility and the challenges of inferring gene regulatory networks from image databases and a mathematical model. The richness of the existing data and tools available to acquire new data to inform models, and the evolution of algorithmic approaches to include diverse data in model optimization continue to drive the systems biology of development forward. Looking forward, our hope is that mechanistic models integrate more closely with experimental inquiry, and that they are more widely used alongside other tools in the biologist's toolbox for experimental discovery.

Glossary

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

References

1. Chakrabarty A, Buzzard GT, Rundell AE. Model-based design of experiments for cellular processes. *Wiley Interdisciplinary Reviews-Systems Biology and Medicine*. 2013; 5:181–203. [PubMed: 23293047]
2. Strogatz, SH. *Nonlinear dynamics and chaos : with applications to physics, biology, chemistry, and engineering*. Westview Press; 2001. p. 512
3. Tomlin CJ, Axelrod JD. *Biology by numbers: mathematical modelling in developmental biology*. *Nature Reviews Genetics*. 2007; 8:331–340.
4. Umulis D, O'Connor M, Blair S. The extracellular regulation of bone morphogenetic protein signaling. *Development*. 2009; 136:3715–3728. [PubMed: 19855014]
5. Lauffenburger, DA.; Linderman, J. *Receptors: Models for Binding, Trafficking, and Signaling*. Oxford University Press; 1996. p. 376
6. Voit, E. *A First Course in Systems Biology*. Garland Science; 2012. p. 496
7. Voit, E. *Computational Analysis of Biochemical Systems*. Cambridge University Press; New York, NY: 2000. p. 532
8. Ellner, SP.; Guckenheimer, J. *Dynamic Models in Biology*. Princeton University Press; Princeton, NJ: 2006. p. 329
9. Perrot N, Trelea IC, Baudrit C, Trystram G, Bourguine P. Modelling and analysis of complex food systems: State of the art and new trends. *Trends in Food Science & Technology*. 2011; 22:304–314.
10. Kutner, M.; Nachtsheim, C.; Neter, J.; Li, W. *Applied Linear Statistical Models*. Fifth ed.. McGraw Hill; New York, NY: 2005. p. 1396
11. Pargett M, Umulis DM. Quantitative model analysis with diverse biological data: Applications in developmental pattern formation. *Methods*. 2013; 62:56–67. [PubMed: 23557990]
12. Schaber J, Klipp E. Model-based inference of biochemical parameters and dynamic properties of microbial signal transduction networks. *Current Opinion in Biotechnology*. 2011; 22:109–116. [PubMed: 20970318]
13. Jaeger J. The gap gene network. *Cellular and Molecular Life Sciences*. 2011;243–274. [PubMed: 20927566]
14. Driever W, Nusslein-Volhard C. A gradient of Bicoid protein in *Drosophila* embryos. *Cell*. 1988; 54:83–93. [PubMed: 3383244]
15. Gregor T, Wieschaus E, McGregor A, Bialek W, Tank D. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*. 2007;141–152. [PubMed: 17632061]
16. Jaeger J, Sharp D, Reinitz J. Known maternal gradients are not sufficient for the establishment of gap domains in *Drosophila melanogaster*. *Mechanisms of Development*. 2007;108–128. [PubMed: 17196796]
17. Gregor, T.; Bialek, W.; van Steveninck, R.; Tank, D.; Wieschaus, E. Diffusion and scaling during early embryonic pattern formation. *Proceedings of the National Academy of Sciences of the United States of America*; 2005. p. 18403-18407.
18. Driever W, Siegel V, Nusslein-Volhard C. Autonomous determination of anterior structures in the early *Drosophila* embryo by the Bicoid morphogen. *Development*. 1990;811–820. [PubMed: 2226200]
19. Driever W, Thoma G, Nusslein-Volhard C. Determination of spatial domains of zygotic gene-expression in the *Drosophila* embryo by the affinity of binding-sites for the Bicoid morphogen. *Nature*. 1989;363–367. [PubMed: 2502714]

20. St. Johnston D, Driever W, Berleth T, Richstein S, Nusslein-Volhard C. Multiple steps in the localization of Bicoid RNA to the anterior pole of the *Drosophila* oocyte. *Development*. 1989;13–19. [PubMed: 2483989]
21. Driever W, Nusslein-Volhard C. The Bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell*. 1988;95–104. [PubMed: 3383245]
22. Wu Y, Myasnikova E, Reinitz J. Master equation simulation analysis of immunostained Bicoid morphogen gradient. *BMC Systems Biology*. 2007
23. Alves F, Dilao R. Modeling segmental patterning in *Drosophila*: Maternal and gap genes. *Journal of Theoretical Biology*. 2006; 241:342–359. [PubMed: 16427090]
24. Reinitz J, Mjolsness E, Sharp DH. Model for cooperative control of positional information in *Drosophila* by Bicoid and maternal Hunchback. *Journal of Experimental Zoology*. 1995; 271:47–56. [PubMed: 7852948]
25. Ibanes M, Belmonte JCI. Theoretical and experimental approaches to understand morphogen gradients. *Molecular Systems Biology*. 2008; 4
26. Jackle H, Tautz D, Schuh R, Seifert E, Lehmann R. Cross-regulatory interactions among the gap genes of *Drosophila*. *Nature*. 1986; 324:668–670.
27. Hulskamp M, Pfeifle C, Tautz D. A morphogenetic gradient of Hunchback protein organizes the expression of the gap genes Kruppel and Knirps in the early *Drosophila* embryo. *Nature*. 1990; 346:577–580. [PubMed: 2377231]
28. Bronner G, Jackle H. Control and function of terminal gap gene activity in the posterior pole region of the *Drosophila* embryo. *Mechanisms of Development*. 1991:205–211. [PubMed: 1768621]
29. Eldon E, Pirrotta V. Interactions of the *Drosophila* gap gene Giant with maternal and zygotic pattern-forming genes. *Development*. 1991; 111:367–378. [PubMed: 1716553]
30. Kraut R, Levine M. Mutually repressive interactions between the gap genes Giant and Kruppel define the middle body regions of the *Drosophila* embryo. *Development*. 1991; 111:611–621. [PubMed: 1893878]
31. Capovilla M, Eldon E, Pirrotta V. The giant gene of *Drosophila* encodes a B-zip DNA-binding protein that regulates the expression of other segmentation gap genes. *Development*. 1992; 114:99–112. [PubMed: 1576969]
32. Struhl G, Johnston P, Lawrence P. Control of *Drosophila* body pattern by the Hunchback morphogen gradient. *Cell*. 1992; 69:237–249. [PubMed: 1568245]
33. Sanchez L, Thieffry D. A logical analysis of the *Drosophila* gap-gene system. *Journal of Theoretical Biology*. 2001:115–141. [PubMed: 11419955]
34. Jaeger J, Blagov M, Kosman D, Kozlov K, Manu, et al. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. *Genetics*. 2004:1721–1737. [PubMed: 15342511]
35. Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, et al. Dynamic control of positional information in the early *Drosophila* embryo. *Nature*. 2004:368–371. [PubMed: 15254541]
36. Fomekong-Nanfack Y, Kaandorp J, Blom J. Efficient parameter estimation for spatio-temporal models of pattern formation: case study of *Drosophila melanogaster*. *Bioinformatics*. 2007; 23:3356–3363. [PubMed: 17893088]
37. Ashyraliyev M, Jaeger J, Blom J. Parameter estimation and determinability analysis applied to *Drosophila* gap gene circuits. *BMC Systems Biology*. 2008; 2 doi:10.1186/1752-0509-1182-1183.
38. Ishihara S, Shibata T. Mutual interaction in network motifs robustly sharpens gene expression in developmental processes. *Journal of Theoretical Biology*. 2008:131–144. [PubMed: 18342890]
39. Ashyraliyev M, Siggens K, Janssens H, Blom J, Akam M, et al. Gene Circuit Analysis of the Terminal Gap Gene huckebein. *PLoS Computational Biology*. 2009; 5 doi:10.1371/journal.pcbi.1000548.
40. Fomekong-Nanfack Y, Postma M, Kaandorp J. Inferring *Drosophila* gap gene regulatory network: a parameter sensitivity and perturbation analysis. *BMC Systems Biology*. 2009
41. Manu, Surkova S, Spirov A, Gursky V, Janssens H, et al. Canalization of Gene Expression and Domain Shifts in the *Drosophila* Blastoderm by Dynamical Attractors. *PLoS Computational Biology*. 2009; 5 DOI: 10.1371/journal.pcbi.1000303.

42. Hengenius J, Gribskov M, Rundell A, Fowlkes C, Umulis D. Analysis of Gap Gene Regulation in a 3D Organism-Scale Model of the *Drosophila melanogaster* Embryo. *Plos One*. 2011; 6
43. Kozlov K, Surkova S, Myasnikova E, Reinitz J, Samsonova M. Modeling of Gap Gene Expression in *Drosophila* Kruppel Mutants. *PLoS Computational Biology*. 2012; 8 DOI: 10.1371/journal.pcbi.1002635.
44. Jimenez-Hornero JE, Santos-Duenas IM, Garcia-Garcia I. Optimization of biotechnological processes. The acetic acid fermentation. Part II: Practical identifiability analysis and parameter estimation. *Biochemical Engineering Journal*. 2009; 45:7–21.
45. Bieler J, Pozzorini C, Naef F. Whole-embryo modeling of early segmentation in *Drosophila* identifies robust and fragile expression domains. *Biophysical Journal*. 2011; 101:287–296. [PubMed: 21767480]
46. Anscombe F. Graphs in Statistical Analysis. *The American Statistician*. 1973; 27:17–21.
47. Cover, T.; Thomas, J. *Elements of Information Theory*. Wiley Interscience; 1991. p. 576
48. Guinee, RA.; Lyden, C.; Aacc, Aacc, A. Motor parameter identification using response surface simulation and analysis. *Proceedings of the American Control Conference*; Arlington, Va.. 2001. p. 4597-4602.
49. Müller, M. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc.; 2007.
50. Deza, M.; Deza, E. *Encyclopedia of Distances*. 2 ed.. Springer-Verlag; Berlin, Heidelberg: 2009.
51. Saltelli, A.; Ratto, M.; Andres, T.; Compolongo, F.; Cariboni, J., et al. Chichester. Wiley; West Sussex, UK: 2008. *Global Sensitivity Analysis: The Primer*; p. 282
52. Kauffman SA. The origins of order: Self-organization and selection in evolution. *The origins of order: Self-organization and selection in evolution*. 1993:xviii+709p–xviii+709p.
53. Spall, JC. *Introduction to Stochastic Search and Optimization*. Wiley; Hoboken, N.J.: 2003.
54. Alajlan N. HopDSW: An approximate dynamic space warping algorithm for fast shape matching and retrieval. *Journal of King Saud University - Computer and Information Sciences*. 2011; 23:7–14.
55. Umulis D, Shimmi O, O'Connor M, Othmer H. Organism-Scale Modeling of Early *Drosophila* Patterning via Bone Morphogenetic Proteins. *Developmental Cell*. 2010; 18:260–274. [PubMed: 20159596]
56. Bazil JN, Buzzard GT, Rundell AE. A Global Parallel Model Based Design of Experiments Method to Minimize Model Output Uncertainty. *Bulletin of Mathematical Biology*. 2012; 74:688–716. [PubMed: 21989566]
57. [PubMed.gov](http://pubmed.gov). National Center for Biotechnology Information, U.S. National Library of Medicine; Bethesda, MD: [PubMed.gov](http://pubmed.gov)
58. Noble SL, Sherer E, Hannemann RE, Ramkrishna D, Vik T, et al. Using adaptive model predictive control to customize maintenance therapy chemotherapeutic dosing for childhood acute lymphoblastic leukemia. *Journal of Theoretical Biology*. 2010; 264:990–1002. [PubMed: 20138060]
59. Chen BS, Lin YP. A Unifying Mathematical Framework for Genetic Robustness, Environmental Robustness, Network Robustness and their Trade-offs on Phenotype Robustness in Biological Networks. Part III: Synthetic Gene Networks in Synthetic Biology. *Evolutionary Bioinformatics*. 2013; 9:87–109.
60. Siegal, ML.; Bergman, A. Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Sciences of the United States of America*; 2002. p. 10528-10532.
61. Perkins T, Jaeger J, Reinitz J, Glass L. Reverse engineering the gap gene network of *Drosophila melanogaster*. *PLoS Computational Biology*. 2006; 2 DOI: 10.1371/journal.pcbi.0020051.
62. MacNamara A, Terfve C, Henriques D, Bernabe BP, Saez-Rodriguez J. State-time spectrum of signal transduction logic models. *Physical Biology*. 2012; 9:045003. doi: 045010.041088/041478-043975/045009/045004/045003. [PubMed: 22871648]

Highlights

- Fitting dynamic models to data is a crucial first step to obtaining model-based insights into biological systems. This requires some measure of the “distance” between experimental data and model simulations. To date, models in developmental signaling have used variants of the sum of square error almost exclusively.
- We investigate three broad classes of distance measures. The pairwise distances are most prevalent in the field. Whole-dataset measures are typically used in statistical modeling and informatics; we consider their application to spatiotemporal data in model-fitting. Finally, information-theoretic methods provide measures of dissimilarity between probability distributions. We adapt them for use with morphogen concentration data.
- The *Drosophila* gap gene network has been modeled by multiple investigators. We review the history of parameter optimization methods and metrics in this experimental system. We use the gap genes as a case study to characterize eight representative objective functions –many of which have never been utilized in this context.
- To aid newcomers to dynamic modeling and optimization, we suggest best practices for selection of objective functions. Emphasis is placed on those functions that complement available data formats and meet research-specific requirements.

Box 1: Pairwise Metrics

Sum of squared error (SSE)

The SSE is the summation of squared residuals. The RMSE is the square root of the mean SSE. between experimental and simulated data. For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), SSE is computed as

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Euclidean Distance (l_2 -Norm)

The Euclidean distance is the square root of the SSE. For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), the Euclidean distance is computed as

$$\sqrt{SSE} = \sqrt{\sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Chebyshev Distance (l_∞ -Norm)

The Chebyshev distance is the maximum absolute residual. For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), the Chebyshev distance is computed as

$$\max_i (|\hat{y}_i - y_i|)$$

Weighting Approaches

The squared and absolute residuals in the preceding metrics may be weighted or normalized to reflect uncertainty in data and differences in data magnitude (e.g., where two proteins' concentrations orders of magnitude apart are being summed in the same

function). $w_i |\hat{y}_i - y_i|$ or $w_i (\hat{y}_i - y_i)^2$ where w_i may be $w_i = \frac{1}{y_{max} - y_{min}}$, $w_i = \frac{1}{\sigma_{y,i}}$, or other functions.

Root means squared error (RMSE)

The RMSE is the square root of the mean SSE. For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), RMSE is computed as

$$\sqrt{\frac{SSE}{n}} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

Manhattan Distance (l_1 -Norm)

The Manhattan distance is the summation of absolute error between residuals. For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), the Euclidean distance is computed as

$$\sqrt{\sum_{i=1}^n |\hat{y}_i - y_i|}$$

l_j -Norms

For an equal number ($i = 1, 2, \dots, n$) of model and experimental data (y_i and $\hat{y}_i$, respectively), the l_j -norm is computed as

$$\sqrt[j]{\sum_{i=1}^n |\hat{y}_i - y_i|^j}$$

Visual Comparison of Pairwise Measures Simulated ($\hat{y}_i$) vs. Experimental Data (y_i)

Dotted and solid lines show the relative lengths of Manhattan and Euclidean distances between two points. The bold dash-dotted line indicates the Chebyshev distance

Box 2: Whole Dataset Metrics

Pearson Correlation Coefficient (PCC)

The Pearson correlation uses the mean and variance of the model ($\hat{y}$) and experimental (y) data (with n data) to calculate the variability of model-experimental points predictions relative to a regression line fit to the data.

$$\frac{\text{cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} = \frac{\sum_{i=1}^n (y_i - E[y]) (\hat{y}_i - E[\hat{y}])}{\sum_{i=1}^n (y_i - E[y])^2 (\hat{y}_i - E[\hat{y}])^2}$$

The PCC increases with improved fit, but error metrics must decrease. To rescale the and invert the PCC, it is normalized as follows.

$$1 - \left(\frac{PCC+1}{2} \right)$$

This distance measure is zero for perfect correlation and one for anti-correlation.

The Anscombe Quartet

The limitation of assuming linearity was demonstrated by Anscombe in 1973. He reported four distributions that returned respectable PCCs of 0.816 despite nonlinearity. This occurrence translates into false positives when optimizing parameters.

Cosine (Angular) Distance Metric

The cosine (or angular) similarity is computed by considering a data space n dimensions corresponding to n data. Model and experimental data are each represented by an vector in this space; the magnitude of the n th datum is the component of the vector in the n th data dimension. Cosine similarity computes the cosine of the angle θ between the two vectors as the crossproduct divided by the vectors' norms:

$$\cos(\theta) = \frac{y \cdot \hat{y}}{\|y\| \|\hat{y}\|} = \frac{\sum_{i=1}^n y_i \times \hat{y}_i}{\sqrt{\sum_{i=1}^n y_i^2 \times \sum_{i=1}^n \hat{y}_i^2}}$$

As with correlation, cosine similarities of one represent perfect agreement while values of negative one indicate vectors pointed 180° apart. Like the PCC, it is normalized to a range of zero to one and subtracted from one to provide a cost metric:

$$1 - \left(\frac{2 \cos^{-1} [\cos(\theta)]}{\pi} \right)$$

This measure is zero for vectors pointing in the same direction and one for vectors pointed in opposite directions.

Box 3: Information Metrics

Kolmogorov-Smirnov Metric

The Kolmogorov-Smirnov (K-S) metric applies a mini-max optimization strategy similar to the Chebyshev distance. Rather than determining the maximum residual error of the data distributions ($y, \hat{y}$), the cumulative distribution functions ($F_y, F_{\hat{y}}$) are computed first. Then the Chebyshev distance is computed for these CDFs. For $i = 1, 2, \dots, n$ data the absolute residuals of F_{y_i} and $F_{\hat{y}_i}$ are computed. The maximum residual is returned:

$$\max_i (|F_{\hat{y}_i} - F_{y_i}|)$$

A K-S value of zero indicates perfect agreement; the calculation ignores scaling like the cosine distance. A value of one indicates a large difference in distributions. Despite the similarity of the procedures, Chebyshev distance and K-S penalize functions differently. This is illustrated below.

Relative Entropy (RE; Kullback-Leibler divergence)

The relative entropy is used to compare the information content of two empirical distributions. Here, we take the model and experimental data to be probability distributions of protein concentration on the spatial domain. Under this assumption, relative entropy measures the information loss experienced if an incorrect modeled distribution is substituted for the true experimental distribution. It is calculated as

$$D_{RE}(y||\hat{y}) = \sum_{i=1}^n \ln \left(\frac{y_i}{\hat{y}_i} \right) y_i,$$

where n is the number of data points, y_i is the i th experimental point, and $\hat{y}_i$ is the i th model point. Before computing the error, both distributions are normalized such that the area under the curve is one (i.e., probability density functions).

An important characteristic of the RE is the asymmetric nature of the measure. This may be explained by expanding the formula:

$$\sum_{i=1}^n \ln(y_i) y_i - \ln(\hat{y}_i) y_i$$

The information lost substituting model $\hat{y}_i$ for data y_i is not identical to the opposite substitution:

$$\sum_{i=1}^n \ln(\hat{y}_i) \hat{y}_i - \ln(y_i) \hat{y}_i$$

This is illustrated in the figures at left and below. The integral (shaded area) is the relative entropy.

Figure 1. Objective choice affects the cost landscape and parameter estimation efficiency

When objective functions are mapped to parameter values, a high-dimensional cost surface is revealed. (A) Plots a simple 1D surface. Maxima (red) and minima (blue) dot its surface; parameter estimation seeks the global minimum, or best fit, though this is not always possible. Each minimum has an associated basin of attraction; local search must start with the global minimum's basin to converge at the desired point. Because basins may be narrow and difficult to reach, global searches are used to survey the cost surface for low-cost points from which to launch local searches. In the full version of the *Drosophila* gap gene model, different objectives produce different landscapes (B–D), though all shared a similar minimal value near a predetermined local minimum (-0.034) of the $Kr \rightarrow Kni$ interaction parameter. Excepting Chebyshev, all of the pairwise measures (B) share qualitative trends. Whole dataset measures (C) also share a characteristic landscape. Information-theoretic measures (D) produced two qualitatively different landscapes when applied to this model.

Figure 2. Response of pairwise objective functions to common spatial errors

Each column represents a type of model-data mismatch; row A represents spatial distributions of the gap gene *Knirps* (*Kni*); rows B–F represent the responses of pairwise objective functions to each type of error. The *Knirps* (*Kni*) expression data in row A (brown diamonds) is from the lateral AP axis of embryonic *Drosophila melanogaster*. Relative to the data, the protein expression peak may be shifted anteriorly (A1, blue line) or posteriorly (A1, red line), it may narrow (A2, blue line) or widen (A2, red line), it may reach a uniform spatial distribution at varying concentrations (A3, blue line-red line), or it may decrease (A4, blue line) or increase (A4, red line) in concentration. Black lines in row A represent optimal model fits. Blue-to-red graded arrows in row A each represent a sweep across the range of erroneous models, with blue lines (E_0) at the beginning and red lines (E_1) of each sweep. Rows B–F plot the response of each objective function as the sweep progresses.

Figure 3. Response of whole-dataset objective functions to common spatial errors

Each column represents a type of model-data mismatch; row A represents spatial distributions of the gap gene *Knirps* (*Kni*); rows B–C represent the responses of whole-dataset objective functions to each type of error. The *Knirps* (*Kni*) expression data in row A (brown diamonds) is from the lateral AP axis of embryos. Relative to the data, the protein expression peak may be shifted anteriorly (A1, blue line) or posteriorly (A1, red line), it may narrow (A2, blue line) or widen (A2, red line), it may reach a uniform spatial distribution at varying concentrations (A3, blue line-red line), or it may decrease (A4, blue line) or increase (A4, red line) in concentration. Black lines in row A represent optimal model fits. Blue-to-red graded arrows in row A each represent a sweep across the range of erroneous models, with blue lines (E_0) at the beginning and red lines at the end (E_1) of each sweep. Rows B–C plot the response of each objective function as the sweep progresses.

Figure 4. Response of information-theoretic objective functions to common spatial errors
Each column represents a type of model-data mismatch; row A represents spatial distributions of the gap gene *Knirps* (*Kni*); rows B–C represent the responses of information objective functions to each type of error. The *Knirps* (*Kni*) expression data in row A (brown diamonds) is from the lateral AP axis of embryos. Relative to the data, the protein expression peak may be shifted anteriorly (A1, blue line) or posteriorly (A1, red line), it may narrow (A2, blue line) or widen (A2, red line), it may reach a uniform spatial distribution at varying concentrations (A3, blue line-red line), or it may decrease (A4, blue line) or increase (A4, red line) in concentration. Black lines in row A represent optimal model fits. Blue-to-red graded arrows in row A each represent a sweep across the range of erroneous models, with blue lines (E_0) at the beginning and red lines at the end (E_1) of each sweep. Rows B–C plot the response of each objective function as the sweep progresses.

Figure 5. Schematic explanation of dynamic space warping

Dynamic space warping is a spatial alignment of two signals to minimize the disagreement between them. Consider a dataset (A, blue line) and model simulation that is partially misaligned (A, red line). A pointwise array of the two signals' data (B) is formed and traversed from beginning to end of each sequence (B, white arrows and line). At the i th, j th position on the grid, the residual values are computed for the next model-datum pair at $i+1, j+1$, but the model is also compared to spatially adjacent data points $i+1, j$ and $i, j+1$. The traversal iterates by choosing the lowest of the three residual errors at each step (C). The mismatch between each model-datum pair is shown in B, with darker regions representing lower errors. The result is a warping that minimizes spatial translation and peak width errors.

Table 1

Cost function criteria		
Error selection criteria	Objective function(s)	Scaling or normalization
Exact fit of concentrations	SSE, MSE, Euclidean, RMSE, Manhattan, MAE, ℓ_2 norm	Transformation to bring units and magnitudes into agreement
Emphasis on spatial coordinates and “shape” of distributions	Cosine distance, relative entropy	Rescaling is implicit in these functions; Dynamic space warping may be required
Relative spatial positioning with exact concentrations	SSE, MSE, Euclidean, RMSE, Manhattan, MAE, ℓ_2 norm with DSW	Transformation to bring units and magnitudes into agreement; DSW to align spatial offsets
Relative spatial positioning with emphasis on “shape”	Relative entropy, pairwise or whole-dataset method with DSW	Rescaling is implicit in these functions; DSW to align spatial offsets
Mini-max strategy for concentration fitting	Chebyshev	Transformation to bring units and magnitudes into agreement