Data-driven, Mechanistic Modeling of Biochemical Reaction Networks

Jason M. Haugh\textsuperscript{1,*}, Timothy C. Elston\textsuperscript{2}, Murat Cirir\textsuperscript{1}, Chun-Chao Wang\textsuperscript{1}, Nan Hao\textsuperscript{2}, and Necmettin Yildirim\textsuperscript{2}

\textsuperscript{1} Department of Chemical & Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695
\textsuperscript{2} Department of Pharmacology, University of North Carolina, Chapel Hill, NC 27599

* To whom correspondence should be addressed: jason_haugh@ncsu.edu
At a certain level of abstraction, living cells are picoliter-sized reaction vessels in which thousands of biochemical reactions and intermolecular binding processes take place in a dynamic, coordinated, and highly regulated fashion. This is an energy intensive process. Intracellular enzyme activities are modulated by covalent modifications that are rapidly added and removed in a seemingly futile cycle in order to respond to changes in the cell’s external environment. These reactions are responsible for governing cell function, and their dysregulation and modulation by infectious agents constitute the molecular basis for human disease. From the perspective of chemical kinetics, the inner workings of the cell are fascinating, but we are still a long way from a mechanistic understanding of these reactions and quantitative characterization of their rates. In this chapter, we discuss how mathematical modeling is applied in tandem with biochemical measurements to achieve this goal.

Whether before, during, or after the collection of experimental data, quantitative modeling is a valuable approach for critically assessing and organizing hypotheses that integrate the many processes that might be at play [1]. And, to the extent that a model is trained on a sufficient amount of quantitative data and its mechanistic assumptions are sound, it may be used to predict the outcomes of novel experiments and thus generate new, hypothesis-driven research. Some experiments will inevitably contradict the model predictions, but as with conceptual, ‘arrow diagram’ models, one iteratively refines the model based on new data.

The examples presented here are focused on mechanisms of signal transduction in eukaryotic cells, which are responsible for controlling cell cycle progression, cell motility, responses to stress, programmed cell death, and differentiation of cell function [2, 3]. These reaction pathways transmit information about the cell’s external microenvironment, making them fundamentally distinct from metabolic pathways, which deal in currencies of energy and
reducing power. We further narrow our focus on modeling of cell signaling that is both data-driven and rooted in biochemical mechanisms. We distinguish data-driven models from purely theoretical models, where experimental data are either not available or not accessible with current technology, and mechanistic models are distinguished from purely phenomenological and purely statistical/correlative models. To supplement the topics presented here, the reader is referred to a number of reviews on the subject of modeling signal transduction processes [4-8].

Rather than presenting detailed recipes of experimental or modeling techniques, this chapter aims to shed light on the inherent relationship between the two in data-driven modeling. In Section I below, we discuss the advantages and shortcomings of different experimental methodologies from the standpoint of modeling and the formulation of models of the appropriate type and level of complexity. Emphasis is placed on the pressing need for model simplification and more systematic approaches for model parameter specification. Then, in Section II, we present two examples of how we have applied those modeling principles to understand specific cell signaling systems.

I. PRINCIPLES OF DATA-DRIVEN MODELING

Types of experimental data

Depending on one’s point of view, cell biology is currently either in a data-rich or data-deprived state. There is a wealth of genomic and proteomic data that have yielded mostly qualitative information about the connectivity of pathways, yet there is relatively little in the way of measurements characterizing their dynamics. Here, we briefly discuss the various quantitative experimental methodologies that define the current state of the art and weigh their advantages,
caveats, and limitations. We choose to classify measurement techniques in three categories: population endpoint, single-cell endpoint, and single-cell kinetic (Table 1). An endpoint measurement is one in which the experiment is stopped at a certain time, and the sample is prepared for analysis, whereas a kinetic measurement is one in which the biochemical readout is monitored in real time. Important considerations include dynamic range (the range of measured values from the lowest limit of detection to the upper limit of assay linearity), throughput (the number of conditions that can be compared in each independent experiment), the ability to multiplex (measure multiple readouts at once), and the ability to assess subcellular localization.

In population endpoint measurements, a large number of cells ($10^3$-$10^8$) are subjected to identical experimental conditions, and a lysate of the cell collective is prepared for analysis. Hence, information about individual cells is lost, and information about subcellular localization is at best indirect; depending on the method of lysis, the preparation can be subdivided based on density and/or detergent solubility into fractions representing different subcellular compartments (cytosol, plasma membrane, endosomes and Golgi, nuclei, etc.). Despite these shortcomings, this approach has several advantages, including potentially high sensitivity and throughput and broad versatility for measuring a variety of molecular readouts; all of these depend critically on the quality of the reagents used. The most common population endpoint measurements, such as immunoblotting, enzyme-linked immunosorbent assays (ELISAs), and *in vitro* enzymatic assays, involve protein immobilization and the use of antibodies for specific capture or detection. All things being equal, assays that involve an initial separation step (e.g., gel electrophoresis or the use of a capture antibody) tend to be more specific and therefore have a higher dynamic range. There is also a general trade-off between the ability to multiplex and throughput, as exemplified by antibody arrays [9] and especially current mass spectrometry technology [10, 11]. These
methods can also be used in conjunction with co-immunoprecipitation to assess protein-protein interactions; however, because of the work-up time involved, this approach is strongly biased to detect only very stable interactions. From the standpoint of experimental data, the inability to measure intracellular protein-protein interactions quantitatively is arguably the most significant limitation for data-driven modeling.

Single-cell endpoint measurements, which provide information about individual cells, include flow cytometry and immunofluorescence microscopy. Both involve incubation with antibodies, detection of fluorescence, and in the case of intracellular proteins, cell fixation and permeabilization. Flow cytometry offers high throughput in terms of assembling population statistics for each sample and moderate throughput in terms of comparing multiple samples. Immunofluorescence offers information about subcellular localization, but the analysis is tedious and therefore low in throughput.

Single-cell kinetic measurements generally involve microscopic imaging of live cells, in which case information about subcellular localization is obtained. Although this approach suffers from many of the same throughput issues as immunofluorescence, the ability to observe signaling kinetics in real time and in conjunction with cell behavior makes it unique [12, 13]. The basis for the measurement is the introduction of a biosensor, either genetically encoded or microinjected into the cell; genetically encoded biosensors are fusion proteins comprised of a protein or protein domain of interest, to which a fluorescent protein such as enhanced green fluorescent protein is attached. A limited degree of multiplexing is offered through the use of multiple biosensors labeled with different fluorophores. The dynamic range of the measurement is affected by which particular biosensor and microscopy modality (e.g., wide-field fluorescence, confocal fluorescence, or total internal reflection fluorescence (TIRF)) are used, and the basis for
the measurement (e.g., a shift in spectral properties of the fluorophore, as in calcium imaging, translocation to a particular membrane or intracellular compartment, or changes in Förster resonance energy transfer (FRET)). The most significant limitation of this approach is that there are currently only a small number of biosensors that work well for quantitative studies; another caveat of using biosensors is that they might significantly interfere with or otherwise modulate the signaling processes they were meant to detect.

Data processing and normalization

All data require some form(s) of processing prior to any sort of quantitative analysis. Some of these are obvious and routine, for example the subtraction of assay/image background and the linear rescaling of images for presentation. Typically, quantitative data are also normalized. The purpose of normalization is to adjust for sources of variability, so that the reproducibility of experimentally deduced trends may be compared in a statistically meaningful way. The manner in which this is done varies and is context-dependent (and, in some cases, arbitrary), and hence this topic is worthy of some discussion.

Variability arises because of both the biological system and the assay itself. Biological variability is significant in any measurement involving cells; this is because, no matter how carefully the parameters of the cell culture are controlled, the culture will vary from experiment to experiment. Assay variability arises from heterogeneity within a sample (e.g., from cell to cell in single-cell measurements) and in the preparation of samples, which affects the comparison of conditions within the same experiment, and also from temporal and lot-to-lot changes in the reagents used, which along with biological variability affect the comparison of independent experiments. Sample heterogeneity at the single-cell or population level is generally normalized
by dividing the signal by a second measurement that should not be affected by the perturbations being tested. For example, population endpoint measurements are typically normalized by the total amount of cellular protein in the sample or by the amount of an abundant species that should be invariant from sample to sample (e.g., actin or tubulin). This is especially important when comparing samples derived from the same cell line/strain but which have been differentially modified over some period of time, for example comparing control cells to cells in which over-expression of a wild-type gene or expression of a mutant gene has been introduced.

Day-to-day variability of the assay reagents and other assay conditions can be normalized by the measurement of a common standard sample; however, this approach is of little use in the typical case where biological variability is also prominent.

To normalize for biological variability, it is often appropriate to use a negative or positive control sample, acquired in each independent experiment. A pitfall of using a negative control for normalization (e.g., fold-induction) is that it often has the lowest and least reliable signal. For more complex data sets of the sort that is desirable for quantitative modeling, with measurements at multiple time points for a variety of experimental conditions, choosing how to normalize the data by a positive control condition (e.g., maximum stimulation of otherwise unperturbed cells) is subject to some ambiguity. Normalizing by the value at a particular time point is a common practice, but the choice of the time point might be considered arbitrary; normalizing by the maximum (peak) value in each experiment is less arbitrary but nonetheless tends to obscure comparisons between control and non-control conditions at time points other than in the vicinity of the peak. For such data sets, we contend that normalizing in a manner that incorporates all of the time-dependent data for the control condition is more appropriate.

Examples include normalizing by the mean value of the control time course, its “area under the
curve” (e.g., [14]), or by normalization factors that minimize its variance across all experiments, e.g., as assessed by the mean coefficient of variation. The latter approach, which we currently favor, is briefly described here and demonstrated in the two Examples presented in Section II.

Suppose there are \( n \) experiments for which data are collected at \( m \) time points. During each of the \( n \) experiments the same control is run. Let \( X_{ij} \) denote the experimental readout for the control in the \( i \)th experiment at the \( j \)th time point. Often the quantity of interest \( Y_{ij} \) (e.g. the concentration of chemical species) is related to \( X_{ij} \) by an unknown scale factor. That is, \( Y_{ij} = \alpha_i X_{ij} \). Under ideal conditions, the control would not vary from experiment to experiment. Therefore, we seek the set of \( \alpha_i \)'s that minimize a suitable quantity \( F \), for example

\[
F = \sum_{j=1}^{m} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2 \quad \text{or} \quad F = \sum_{j=1}^{m} \frac{1}{\bar{Y}_j} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2,
\]

where \( \bar{Y}_j \) is the mean value that results for time point \( j \). The minimization is subject to a constraint that eliminates the trivial solution, \( \alpha_i = 0 \) for all \( i \). Once the \( \alpha_i \) have been found, they are used to scale the experimental time series allowing the mathematical model to be fit to all the data simultaneously.

**Suitability of models used in conjunction with quantitative data**

In formulating a suitable mathematical description of a system, it is important to cast the model at an appropriate level of abstraction, which should be weighed carefully along with considerations of computational feasibility. While all models of biochemical processes are expected to include fundamentals of chemical reaction kinetics, they are expected to vary along two axes of increasing complexity: from deterministic to stochastic, and from well-stirred to spatially extended (Figure 1). In deterministic models, continuum variables such as species...
concentrations evolve according to ordinary or partial differential equations (ODEs and PDEs, respectively) and associated initial and boundary constraints, whereas in stochastic models, molecules and molecular complexes are modeled as discrete entities whose states are updated probabilistically [15, 16]. So-called hybrid models incorporate both continuum and discrete variables [17]. On the other axis, well-stirred models assume spatial homogeneity within the domain of interest, and any transport processes in the model (for example, trafficking between intracellular compartments [18]) are incorporated as reaction terms, whereas spatially extended molecules account for spatial gradients and therefore model the underlying transport processes explicitly, according to physicochemical principles [8].

For data-driven modeling of biochemical systems, the chosen complexity of the model should depend not only on what qualitative information is available in the literature, however reliable, but also in large part on the amount and type of quantitative, experimental data available. For instance, population endpoint measurements tend to be the most versatile and quantitative, yet they do not provide the kind of information that would justify a stochastic or spatially extended description of the model. Therefore, even though more complex models might be formulated, it is most appropriate to cast the model as a set of deterministic ODEs (see Section II, Examples 1 and 2). Data-driven stochastic models generally benefit from single-cell information, which is obtained most quantitatively (albeit without spatial information) from flow cytometry data [19-21], and spatially extended models must be driven almost exclusively by single-cell kinetic (live-cell microscopy) data [22-26].
Issues related to parameter specification and estimation

Another aspect of model complexity that must be carefully considered when making comparisons to data is the amount of molecular detail to include. A comprehensive model, explicitly including all of the “known” biochemistry, comes at the expense of having to identify a large set of parameter values (rate constants and initial concentrations) [27]. Prominent examples of signaling pathway/network models with ~ 100 or more adjustable parameters have been offered [28-31], and in such cases the parameter values are typically culled from published in vitro measurements using purified components (or assumed to be similar in magnitude to parameters for related interactions where such data are available) or adjusted by hand to reconcile the sparse biochemical data assembled in various cell types and laboratories. Although models using this approach have proven valuable, it must be recognized that there is a great deal of uncertainty associated with such a parameter specification exercise. Formulation of very detailed models also dictates a qualitative assessment, wherein the model is judged by its ability to correctly produce the gross kinetic features seen in a relatively small collection of measurements [1].

The other approach is to simplify the model so as to reduce the number of adjustable parameters, to the point where a more direct, quantitative comparison or fit to the data becomes feasible and adequately constrained. Thus, the degree of model simplification is largely determined by the variety of experimental conditions and biochemical readouts in the data set; this, we contend, is the art of data-driven modeling. Simplification of kinetic models is achieved in a number of ways, including the use of scaled, dimensionless variables and through knowledge or assumptions about fast versus slow rate processes. Another mode of simplification is the lumping of multiple processes into a single step, which is warranted when quantitative data
related to that particular step are absent or unattainable, or when its details are poorly characterized.

Supposing that a model with an appropriate level of granularity has been tailored for a particular set of measurements, how does one fit the model output to the data? This can be somewhat tricky, because even with appropriate simplification, a pathway/network model is going to have more than a handful of adjustable parameters. Indeed, it is becoming increasingly clear that the values of parameters in models with even modest complexity are not uniquely identifiable, even with near perfect kinetic data [32]. With that said, there are efficient methods for identifying a (non-unique) set of parameters that fit the data optimally well. One approach, which has been used to great effect in the modeling of the cell cycle, is the use of global optimization algorithms such as ODRPACK, which implements the Levenberg-Marquardt method with variable step size [33, 34]. Another strategy, which is gaining in popularity, involves Monte Carlo-based or “genetic” algorithms, wherein all of the parameter values are adjusted randomly, according to distributions centered on the current values, and the resulting parameter set is either accepted or rejected with certain probability or based on specified criteria related to the goodness of fit. The classic example of such an approach is the Metropolis algorithm [35] (Figure 2). In this method, parameter sets that improve the goodness of fit are always accepted, whereas sets that yield a poorer fit are accepted with a probability determined by a Boltzmann-like function; the overall error ($\chi^2$) is analogous to the energy, which is compared with a user-specified parameter that is analogous to the thermal energy scale or temperature (the lower the “temperature”, the lower the probability of acceptance). A commonly used variation is simulated annealing, in which the “temperature” is steadily reduced with time, making it more efficient for finding a global optimum [36, 37]. Regardless of the method used,
it is important to note that the units of the model and those of the measurement are rarely the same, and so a conversion/alignment factor for each data type must usually be assigned or used as a fit parameter.

Faced with the inherent problem of identifying unique parameter values, it might not be fruitful to seek one single, “best” solution to the parameter estimation problem; another approach, which we have demonstrated in Examples 1 and 2 below, is the ensemble or collective fitting approach [32, 38]. In this method, one accumulates a large number (potentially >> 1,000) of parameter sets (the ensemble) that fit the data almost equally well. Starting with a single, near-optimal parameter set, the Metropolis algorithm is suitable for collecting the ensemble. At least for ODE models, which are solved with very little computational effort, it is no large task to recompute the model output for each of these parameter sets; the output of the “model”, then, may be taken as the ensemble mean, with its standard deviation yielding a measure of the variability in the model fit or prediction. An advantage of this approach is that one can readily infer whether or not a particular parameter is well constrained by the fit by inspection of the distribution of its values across the ensemble. Arguably, this evaluation is more insightful than the typical sensitivity analysis, which only assesses how the model responds to small changes in the parameter values, made one parameter at a time.

II. EXAMPLES OF DATA-DRIVEN MODELING

Example 1: Systematic analysis of crosstalk in the PDGF receptor signaling network

Historically, intracellular signal transduction has been characterized in terms of pathways of sequential sequential activation processes, such as the canonical mitogen-activated protein
kinase (MAPK) cascades; a prominent example is the Ras $\rightarrow$ Raf $\rightarrow$ MEK $\rightarrow$ extracellular signal-regulated kinase (Erk) pathway in mammals, which is both a master integrator of upstream inputs and a master controllers of transcription factors and other effectors [39]. Although our current understanding of signal transduction networks includes more complex interactions, including those between the classically defined pathways (crosstalk) and those responsible for feedback regulation/reinforcement, such interactions have not yet been adequately characterized.

In an effort to quantify the relative contributions of classical and crosstalk interactions in a signaling network, population endpoint measurements and computational modeling were systematically combined to study signaling mediated by platelet-derived growth factor (PDGF) receptors in fibroblasts [40] (Figures 3 & 4). The PDGF receptor signaling network is important in dermal wound healing and embryonic development [41], stimulating directed cell migration, survival, and proliferation through the aforementioned Ras/Erk pathway and exceptionally robust activation of phosphoinositide 3-kinases (PI3Ks), which produce specific lipid second messengers at the plasma membrane [42-44].

Erk phosphorylation and PI3K-dependent Akt phosphorylation in PDGF-stimulated NIH 3T3 fibroblasts were measured by quantitative immunoblotting for an array of 126 experimental conditions, sampling different combinations of PDGF dose, stimulation time, and molecular manipulations; considering biological replicates and parallel determination of total Erk and Akt levels, this set of data comprises 2,772 total measurements. A selected portion of the Erk data shows that blocking the activity of either Ras or PI3K only partially reduces PDGF-stimulated Erk phosphorylation (Fig. 3A), whereas simultaneous inhibition of Ras and PI3K almost completely abolished PDGF-stimulated Erk phosphorylation (Fig. 3B), indicating that Ras and
PI3K are responsible for all of the major pathways from PDGF receptors to Erk, and at least one mode of PI3K-dependent crosstalk to Erk is independent of Ras. By comparison, the Akt phosphorylation results showed that the PI3K pathway is not significantly affected by perturbations affecting Ras and Erk; crosstalk is apparently unidirectional, from PI3K to Ras/Erk, in this network [40]. This conceptual model was further refined by additional experiments, which characterized two known negative feedback mechanisms and established that PI3K-dependent crosstalk affects the Erk pathway both downstream and upstream of Ras (Fig. 3C).

Motivated by the dynamics revealed in this unique data set, a kinetic model of the network was formulated and used to quantify the relative magnitudes of the PI3K-dependent and Ras-independent inputs collaborating to activate Erk. A total of 34 unspecified parameter values were estimated using the ensemble approach described in the previous section; taken together, the data force the model to reconcile time- and PDGF dose-dependent features of the network observed under the various experimental conditions tested (Fig. 3D).

Analysis of the parameter sets chosen by the algorithm revealed a consistent ratio of PI3K- and Ras-dependent contributions to the dual phosphorylation of MEK, the kinase activity directly upstream of Erk (Fig. 4A). We formulated a single number, the MEK activation comparator (MAC), which compares the capacities of the two pathways to generate dually phosphorylated MEK. Importantly, the MAC quantifies these inputs in a way that uncouples them from negative feedback effects. This analysis revealed that, whereas the PI3K-dependent MEK activation pathway is predicted to be intrinsically much less potent than the Ras-dependent pathway under maximal PDGF stimulation conditions, feedback regulation of Ras renders the
PI3K-dependent pathway somewhat more important (Fig. 4B). A similar analysis was performed relating the PI3K-dependent and -independent signaling modes upstream of Ras [40].

The computational approach was also used to generate hypothetical predictions with an eye towards future experiments. Whereas inhibition of PI3K affects crosstalk interactions both upstream and downstream of Ras, the model ensemble predicts unique kinetic signatures that might be expected if either mechanism were silenced selectively [40], which could help validate the point of action of a particular PI3K-dependent pathway on Erk.

Example 2: Computational analysis of signal specificity in yeast

Yeast is well recognized as an excellent model organism for systems level analysis [45]. Their ability to undergo efficient homologous recombination is particularly useful for studying the functional role of proteins in vivo, through gene disruption or gene replacement. Because of this property, the yeast pheromone response system is arguably the best-characterized signaling pathway of any eukaryote. This pathway bears strong similarities to signaling networks in mammals. In particular, the MAPK components share extensive sequence similarity with their human counterparts [46]. Another feature common to the yeast pheromone response pathway and response pathways of higher organisms is the sharing of signaling proteins among multiple systems. This property makes the pheromone response pathway an excellent system for studying signal specificity.

Depending on specific external cues, yeast cells initiate either a mating response or an invasive growth program. Mating is initiated when haploid cell types a and α secrete and respond to type-specific pheromones, which act through G protein-coupled receptors on cells of the opposite mating type [47]. Alternatively, invasive growth occurs in nutrient-poor conditions.
Combined genetic and biochemical studies revealed that both mating and invasive growth require a protein kinase cascade comprised of Ste20 (MAP4K), Ste11 (MAP3K), and Ste7 (MAP2K) (Fig. 5A). The pathways diverge at the level of the MAP kinase. Whereas deletion of one MAP kinase gene (KSS1) blocks invasive growth, deletion of a second MAP kinase gene (FUS3) impairs pheromone-induced cell-cycle arrest. Deletion of FUS3 leads to enhanced activity of Kss1 [49]. However, the mechanism by which this cross inhibition occurs was unknown.

We recently combined mathematical modeling with experimental analysis to investigate how Fus3 limits the activity of Kss1 [50]. Six mathematical models were developed to describe different hypothetical mechanisms of cross inhibition. All six models were fit to the time courses for Fus3 and Kss1 activation obtained from wild-type cells as well as from strains containing various genetic alterations. The experiments yielded a data set of over 300 measurements. To compare the performance of each of the 6 potential models, the Monte Carlo approach described above was used. Fig. 5B shows a plot of the sum of the squared differences (SSD) versus the number of accepted realizations in the Monte Carlo optimization process for each of the six models. After 800 accepted realizations the SSD converged for each model. Model I performed the best (minimum SSD) and the next best models were II, III, and V, which roughly performed equally well.

Each of the six models fall into one of two distinct cases (Fig. 5C): 1) active Fus3 inhibits Kss1 phosphorylation, and 2) active Fus3 increases Kss1 dephosphorylation. Models I and II are mathematically the simplest and demonstrate the key difference between the two hypothetical mechanisms of cross inhibition. To compare the two models, 100 parameter sets were randomly selected from those accepted by the Monte Carlo optimization routine. The
model equations were run using these parameter sets to generate a distribution of solutions. Fig. 6A shows comparisons between the models’ output and experimental data for Kss1 activity in WT cells (black circles) and cells in which the MAPK Fus3 has been deleted (red circles). Note that only Model I is able to capture the rapid increase in Kss1 activity seen in the fus3Δ strain. The confidence intervals presented in these plots indicate that this behavior is not a consequence of the specific choice of parameter values, but a general property of the models.

Motivated by these results, a simplified model of cross-inhibition was developed that captures the two general mechanisms by which Fus3 might regulate Kss1. Analysis of the simple model revealed that the two mechanisms of cross-inhibition have opposite effects on the rate at which the system relaxes to steady state. If Fus3 inhibits Kss1 phosphorylation, the relaxation rate is reduced; if Fus3 increases deactivation, the relaxation rate is increased. Consequently, a mechanism that increases the dephosphorylation rate of Kss1 is incompatible with the experimental data because it cannot simultaneously account for i) the large increase in maximum Kss1 activity seen in the fus3Δ strain and ii) the slow decline in Kss1 activity observed in wild-type cells.

Because the MAP2K Ste7 is feedback phosphorylated by Fus3 [51-53] and directly catalyzes Kss1 activation, this protein was considered to be the most likely target for Fus3-mediated cross inhibition of Kss1. The sites at which Fus3 phosphorylates Ste7 have been mapped, and a mutant lacking each of these phosphorylated residues has been described (Ste7A7) [54]. Consistent with the results of the computational investigations, Ste7A7 exhibits a significant elevation in the extent of Kss1 phosphorylation compared with wild-type Ste7 (Ste7Wt); furthermore, the mathematical model describing this scenario accurately predicts the extent and duration of the increase in Kss1 activation promoted by Ste7A7 (Fig. 6B) [50].
ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R01-GM067739 and R21-GM074711 to J.M.H. and R01-GM079271 and R01-GM073180 to T.C.E.
REFERENCES


Table 1. Capabilities and limitations of common experimental methodologies. Various measurement techniques are rated by typical performance in four categories: dynamic range/sensitivity, throughput (the number of conditions that can be compared in each independent experiment), the ability to multiplex (measure multiple readouts at once), and the ability to assess subcellular localization. * Immunoblotting using enhanced chemiluminescence and a high-sensitivity, cooled charge coupled device camera for imaging; the traditional method using photographic film for imaging gives a sigmoidal response over a much narrower dynamic range (contributing to the false notion that immunoblotting is generally not quantitative).
Figure 1. Two axes of mechanistic model complexity. Models can be characterized according to whether they are deterministic or stochastic and whether or not they explicitly account for spatial gradients. Roughly speaking, the degree of computational difficulty increases as one moves from the lower left to the upper right quadrant. In each corner, techniques used to implement such models are listed along with, in parentheses, the type of experimental data that might be described. Abbreviations: ODE, ordinary differential equation; PDE, partial differential equation; SDE, stochastic differential equation; BD, Brownian dynamics.
Figure 2. Parameter estimation using the Metropolis algorithm. A. Schematic of the algorithm. The values of all model parameters are adjusted at random, according to distributions centered on the previous values, and the resulting quality of fit determines the probability of accepting each successive parameter set. Alignment of the model output to the data is achieved through the assignment of conversion factors, which may be estimated in a separate subroutine. The performance of the algorithm is tuned by adjusting the values of $\alpha$, which characterizes how much the parameters change in each step, and $\beta$, the stringency of the acceptance criterion. B. Illustration of the algorithm run in a highly stringent mode, wherein each accepted move almost always results in a better fit (lower SSD), starting from random guesses of the parameter values. C. After achieving a near-minimum SSD value, the algorithm may be reinitiated with a relaxed stringency, allowing a large number of parameter sets to be collected in an ensemble. The average output of the parameter set ensemble constitutes the output of the model. Quantitative predictions are made through uniform changes (e.g., setting a particular parameter equal to zero) across the ensemble.
Figure 3. Data-driven model to characterize crosstalk in the PDGF receptor signaling network. A. A portion of the quantitative data set shows that inhibition of Ras (by expression of dominant-negative S17N Ras) or PI3K (using the LY compound) affects the dynamics of PDGF-stimulated Erk phosphorylation. B. Whereas inhibition of Ras or PI3K only partially blocks Erk phosphorylation, the double-inhibition experiment shows that Ras and PI3K account for all of the major pathways from PDGF receptors to Erk. C. Conceptual model of the PDGF receptor signaling network based on the entire data set. D. A coarse-grained kinetic model of the network is aligned directly to the data using a variation of the Metropolis algorithm and the parametric ensemble approach. All panels are adapted from [40].
Figure 4. Quantification of Ras- and PI3K-dependent MEK phosphorylation pathways in the PDGF receptor signaling network. A. For each parameter set in the model ensemble, the quantity $C_{xij}$ is defined as the maximum catalytic efficiency of pathway $i$ ($i = 1$, Ras-dependent; $i = 2$, PI3K-dependent) towards site $j$ on MEK divided by the catalytic efficiency of the corresponding phosphatase reaction. On the dashed line, the two pathways are equally potent by this measure. B. When MEK kinases and phosphatases are far from saturation, the steady-state fractions of MEK in the unphosphorylated, singly phosphorylated, and doubly phosphorylated states are readily calculated. The MEK Activation Comparator (MAC) is a ratio devised to compare the MEK phosphorylation capacity of PI3K-dependent signaling crosstalk to that of the classical Ras-dependent pathway.
Figure 5.
Figure 6.