Maximum entropy framework for inference of cell population heterogeneity in signaling network dynamics

Purushottam D. Dixit*[1], Eugenia Lyashenko*[1], Mario Niepel[2], and Dennis Vitkup[1,3,4]

Correspondence should be addressed to P. D. (dixitpd@gmail.com) or D.V. (dv2121@columbia.edu).
*These authors contributed equally

[1]Department of Systems Biology, Columbia University
[2]Department of Systems Biology, Harvard Medical School
[3]Department of Biomedical Informatics, Columbia University
[4]Center for Computational Biology and Bioinformatics, Columbia University

Predictive models of signaling networks are essential tools for understanding cell population heterogeneity and designing rational interventions in disease. However, using network models to predict signaling dynamics heterogeneity is often challenging due to parameter non-identifiability and the variability of network parameters across cell populations. To address these challenges, we present here an inference framework based on the principle of maximum entropy (ME). The developed approach allows us to estimate the joint probability distribution over network parameters that is consistent with experimentally observed cell-to-cell variability in abundances of network species. We apply the inference framework to investigate the heterogeneity in the signaling network activated by the epidermal growth factor (EGF) and leading to phosphorylation of protein kinase B (Akt). Using the inferred parameter distribution, we predict heterogeneity of phosphorylated Akt levels and the distribution of EGF receptor abundance hours after EGF stimulation. We also discuss how the developed framework can be generalized and applied to problems beyond modeling of heterogeneous signaling dynamics.

Significance statement

Cells in a population respond heterogeneously to extracellular stimuli. Predictive models of cell population heterogeneity are key to designing rational interventions in disease. While multiple experimental approaches make it possible to measure species abundances at the single cell level, estimating heterogeneity in mechanistic signaling network parameters from these data has remained challenging. We present here a maximum entropy framework to infer the joint distribution over network parameters from single cell data. The framework accurately predicts experimentally measured cell-to-cell variability in proteins in the protein kinase B (Akt) pathway activated by the epidermal growth factor (EGF). We discuss how the framework can be generalized to analyze cell population heterogeneity using multiple types of single cell data as well as problems beyond signaling networks.

Introduction

Signaling cascades in genetically identical cells often respond in a heterogeneous manner to extracellular stimuli (1). This heterogeneity is due to cell-to-cell variability in network parameters such as reaction rates and species abundances. Notably, variable network responses have important functional consequences. For example, they mediate stochastic transitions in development (2) and are responsible for fractional killing of cancer cells treated with chemotherapeutic drugs (3).

Several experimental techniques such as flow cytometry (4), immunofluorescence (4), and live cell assays (5) have been developed to investigate the variability of network species abundances. However, it is often difficult to estimate the distributions over network parameters that are consistent with these experimental measurements. The reasons for this challenge are primarily twofold. First, parameters such as protein abundances and biochemical rates vary substantially across cells in a population (1). For example, previous studies have reported the coefficients of variation of protein abundances in the range 0.1-0.6 (6). Similarly, reaction rate constants also vary between cells by several orders of magnitude (7). Second, many network parameters are usually not completely identifiable given limited experimental measurements of a few species at several time points (8-10). The non-identifiability is widespread as network dynamics is often insensitive to correlated variations in parameters with opposite effects, such as species association rates and the corresponding dissociation rates.

Over the last decade, several computational methods have been developed to estimate the joint distribution of network parameters consistent with experimentally measured cell-to-cell variability in network species (11-14). However, these methods usually rely on several simplifying assumptions. For example, the parameter distributions are restricted to a pre-defined functional family, such as the multivariate log-normal distribution (12). Alternatively, simplifying restrictions are imposed on the underlying dynamical model with reaction propensities assumed to be linear in network parameters (14).

Building on our previous work (9, 15), we here present a framework based on the maximum entropy (ME) principle to make it possible to infer the joint distribution over

network parameters consistent with single cell experimental measurements. The developed approach addresses the many aforementioned challenges. ME principle was first introduced more than a century ago in statistical physics (16). Among all candidate distributions that agree with imposed constraints, ME selects the one with the least amount of bias. ME-based approaches have been successfully applied previously to a variety of biological problems, including protein structure prediction (17), protein sequence evolution (18), neuron firing dynamics (19), molecular simulations (20-22), and dynamics of biochemical reaction networks (23).

After describing the ME approach, we apply the developed framework to study the heterogeneity in the signaling network leading to phosphorylation of protein kinase B (Akt). Epidermal growth factor (EGF)-induced Akt phosphorylation governs key intracellular processes (24) including, metabolism, apoptosis, and cell cycle entry. Due to its central role in mammalian signaling, aberrations in the Akt pathway are implicated in multiple diseases (24, 25). We use the ME approach to infer the distribution over network parameters using experimentally measured phosphorylated Akt (pAkt) and cell surface EGFR (sEGFR) levels in MCF 10A cells following EGF stimulation (26). We then demonstrate that the obtained parameter distribution allows us to accurately predict the heterogeneity in single cell pAkt levels at late time points, as well as heterogeneity in cell surface EGFRs in response to EGF stimulation. Finally, we also discuss generalizations of the framework to study problems beyond modeling heterogeneity in signaling networks.

Results

Outline of the ME framework

We consider a signaling network comprising N chemical or biological species whose intracellular abundances we denote by $\bar{X} = \{X_1, X_2, \ldots, X_N\}$. We assume that the molecular interactions among the species are described by a system of ordinary differential equations

$$\frac{d}{dt}\bar{X}(t,\bar{\theta}) = f(\bar{X},\bar{\theta}) \tag{1}$$

where $f\left(\bar{X},\bar{\theta}\right)$ is a function of species abundances, $\bar{X}$. Here $\bar{\theta}=\left\{\theta_1,\theta_2,\ldots\right\}$ is a vector of parameters that describe the dynamics of the signaling network. We denote by $x_a\left(t,\bar{\theta}\right)$ the solution of equations 1 for species *a* with parameters $\bar{\theta}$.

Our computational approach is illustrated in Figure 1. We use experimentally measured the cell-to-cell variability of protein species x_a at multiple time points to constrain the parameter distribution $P\left(\bar{\theta}\right)$ (illustrated by histograms in Figure 1). Specifically, we quantify the measured variability by estimating fraction ϕ_{ik} of cells that belong to a particular 'bin' in the abundance distribution histogram of each time point; in our notation, *i* indicates time and *k* indicates bin number. Every distinct dynamical trajectory $x_a\left(t,\bar{\theta}\right)$ (illustrated by red and blue curves in Figure 1) generated by specific parameter values $\bar{\theta}$ passes through a unique set of bins (red curve through red bins and blue curve through blue bins in Figure 1) at multiple time points. Using the ME framework, we find a corresponding probability distribution for network parameters $P\left(\bar{\theta}\right)$ (see inset, Figure 1) such that the distribution over trajectories $P\left[x_a\left(t,\bar{\theta}\right)\right]$ is consistent with all experimentally measured bin fractions. Notably, because we simultaneously identify all temporal bins crossed by a specific trajectory $x_a\left(t,\bar{\theta}\right)$, our approach naturally accounts for statistical correlations between data at different time points and thus avoids over-constraining the probability distribution. Below we present our development to derive the functional form of $P\left(\bar{\theta}\right)$ consistent with experimental data.

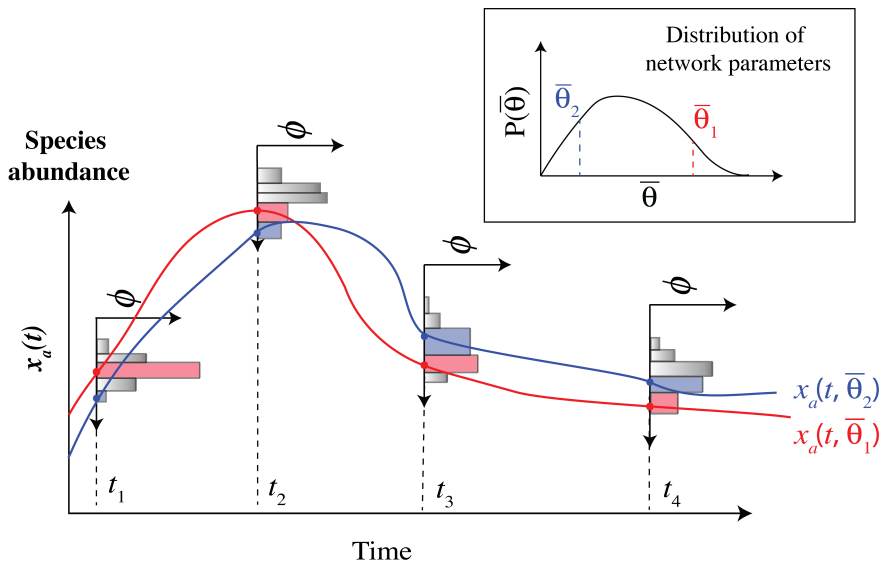

Figure 1. **Illustration of our inference approach**. Cell-to-cell variability in a protein x_a is measured at 4 time points t_1, t_2, t_3, and t_4. From the data, we determine the fraction ϕ_{ik} of cells that populate the kth abundance bin at the ith time point by binning the cell-to-cell variability data in B_i bins. The histograms show the bin fractions ϕ_{ik} at multiple time points. We find $P(\bar{\theta})$ with the maximum entropy while requiring that the distribution $P\left[x_a(t,\bar{\theta})\right]$ of simulated trajectories of $x_a(t,\bar{\theta})$ simultaneously reproduce all bin fractions ϕ_{ik}s.

Derivation of $P(\bar{\theta})$ consistent with experimental data

For simplicity, we first consider the case when the distribution of cell-to-cell variability in one species x is available only at time point t (for example, $t = t_1$ in Figure 1). We denote by $\bar{\phi} = \{\phi_1, \phi_2, \ldots, \phi_B\}$ the fraction of cells whose experimental measurement of x lies in the kth bin. Given a particular parameter distribution $P(\bar{\theta})$, the *predicted* fractions $\bar{\psi} = \{\psi_1, \psi_2, \ldots, \psi_B\}$ can be obtained as follows. Using Markov chain Monte Carlo (MCMC), we generate multiple parameter sets $\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3, \cdots$ from $P(\bar{\theta})$. For each generated parameter set $\bar{\theta}$, we numerically solve equations 1 and find $x_a(t_1,\bar{\theta})$, i.e. the predicted value of the abundance at time t_1. Using the samples from the ensemble of trajectories, we estimate ψ_k as the fraction of sampled trajectories where $x_a(t_1,\bar{\theta})$ belonged to the kth bin. Mathematically,

$$\psi_k = \int I_k\left(x_a\left(t_1,\bar{\theta}\right)\right) \bullet P\left(\bar{\theta}\right) d\bar{\theta} \tag{19}$$

where $I_k(x)$ is an indicator function; i.e. $I_k(x)$ is equal to one if x lies in the kth bin and zero otherwise.

The central idea behind our approach is to find the maximum entropy distribution $P\left(\bar{\theta}\right)$ over parameters such that all predicted fractions ψ_k agree with those estimated from experiments ϕ_k. Formally, we seek $P\left(\bar{\theta}\right)$ with the maximum entropy,

$$S = -\int P\left(\bar{\theta}\right) \bullet \log \frac{P\left(\bar{\theta}\right)}{q\left(\bar{\theta}\right)} d\bar{\theta} \tag{3}$$

subject to constraints $\psi_k = \phi_k$ and normalization, $\int P\left(\bar{\theta}\right) d\bar{\theta} = 1$. Here, $q\left(\bar{\theta}\right)$ plays a similar role to the prior distribution in Bayesian approaches (27). In this work, we choose $q\left(\bar{\theta}\right)$ to be a uniform distribution within literature-derived ranges of parameters, but other choices can be implemented as well.

To impose aforementioned constraints and perform the entropy maximization, we use Lagrange multipliers. To that end, we use an unconstrained optimization function

$$L = S + \beta\left(\int P\left(\bar{\theta}\right) d\bar{\theta} - 1\right) - \sum_{k=1}^{B} \lambda_k \left(\int I_b\left(x_k\left(t_1,\bar{\theta}\right)\right) \bullet P\left(\bar{\theta}\right) d\bar{\theta} - \phi_k\right) \tag{4}$$

where β is the Lagrange multiplier associated with normalization and λ_k are the Lagrange multipliers associated with fixing the predicted fractions ψ_k to their experimentally measured values ϕ_k in all bins $k \in \{1,2,\ldots,B\}$. Differentiating equation 4 with respect to $P\left(\bar{\theta}\right)$ and setting the derivative to zero, we obtain

$$P\left(\bar{\theta}\right) = \frac{1}{\Omega} q\left(\bar{\theta}\right) \exp\left(-\sum_{k=1}^{B} \lambda_k I_k\left(x_a\left(t_1,\bar{\theta}\right)\right)\right) \tag{5}$$

where $\Omega = \int q\left(\bar{\theta}\right) \exp\left(-\sum_{k=1}^{B} \lambda_k I_k\left(x_a\left(t_1,\bar{\theta}\right)\right)\right) d\bar{\theta}$ is the partition function that normalizes the probability distribution.

The generalization of equation 5 when abundances of multiple biological or chemical species are measured at several time points is straightforward. Specifically, if we constrain distributions of cell-to-cell abundance variability of n species (x_1, x_2, x_3, …

x_n) measured at times t_{ij} (where i denotes species and j denotes time points) the maximum entropy probability distribution over parameters $P(\bar{\theta})$ can be written as

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp\left(-\sum_{i=1}^{n} \sum_{j=1}^{T_i} \sum_{k_{ij}=1}^{B_{ij}} \lambda_{k_{ij}} I_{k_{ij}} \left(x_i \left(t_{ij}, \bar{\theta} \right) \right) \right) \tag{6}$$

where $x_i\left(t_{ij},\bar{\theta}\right)$ is the solution of equation 1 for the ith species measured at the jth time point. In equation 6, the experimentally measured distributions of abundances for species i at time j are represented by B_{ij} bins and the Lagrange multipliers corresponding to the k_{ij}^{th} bin are λ_{ij} respectively.

The Lagrange multipliers λs in equations 6 need to be numerically optimized such that the predicted bin fractions are consistent with the experimentally estimated ones. Notably, the search for the Lagrange multipliers is a convex optimization problem and we solve it using an iterative algorithm proposed in (28) (see Figure 2). We start from a randomly chosen point in the space of Lagrange multipliers. In the nth iteration of the optimization algorithm, using the current vector of the Lagrange multipliers $\bar{\lambda}_n$, we estimate the predicted bin fractions $\bar{\psi}_n$ using Markov chain Monte Carlo (MCMC). Next, we estimated the error vector $\bar{\Delta}_{n,i} = \left(\bar{\psi}_{n,i} - \bar{\phi}_{n,i}\right) / \bar{\phi}_{n,i}$ for the nth iteration. We then update the multipliers for the n+1st iteration as $\bar{\lambda}_{n+1} = \bar{\lambda}_n + \alpha_n \bar{\Delta}_n$ (see Figure 2) where the positive "learning rate" α_n is chosen to minimize the error $\left|\bar{\Delta}_{n+1}\right|$ (see SI section III for details).

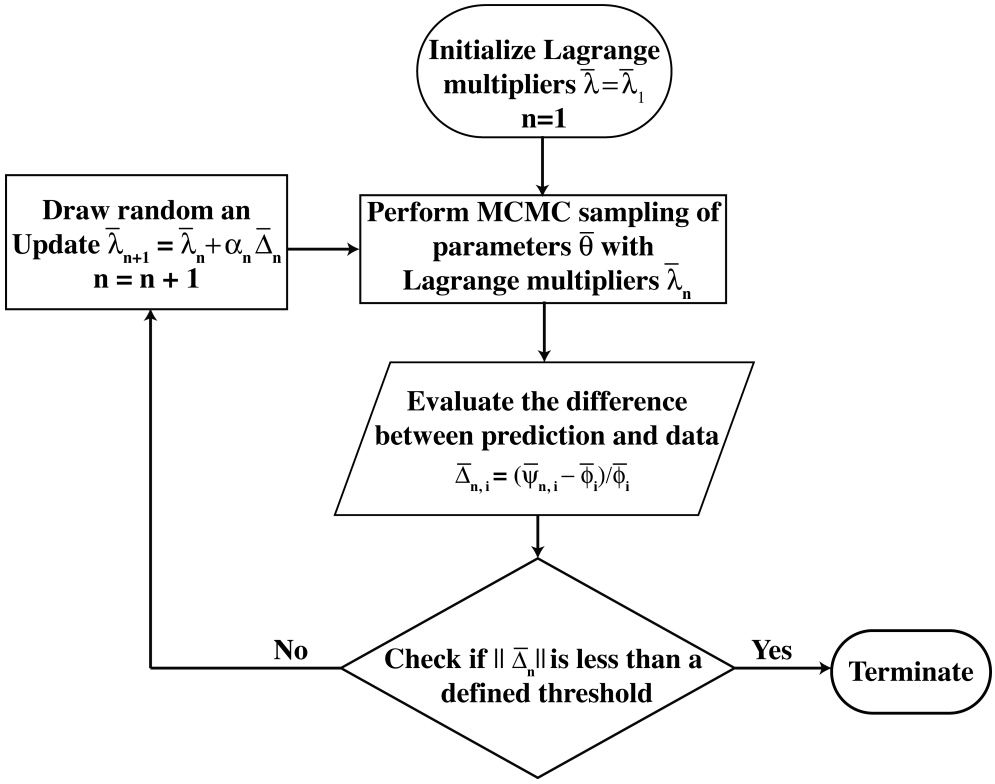

Figure 2. The workflow to numerically determine the values of Lagrange multipliers. In each iteration we evaluated the error vector $\bar{\Delta}_n$ between predicted bin fractions $\bar{\psi}_n$ and the experimentally measured bin fractions $\bar{\phi}$ using MCMC. We proposed a new set of Lagrange multipliers based on the error vector. We repeated until the error reached below a predefined accuracy cutoff.

Computational model of the EGFR/Akt signaling network

Signal transduction in the EGFR/Akt network is illustrated in Figure 3. Following stimulation of cells with EGF, the ligand binds to cell surface EGFRs. Ligand-bound receptors dimerize with other ligand-bound as well as ligand-free receptors. EGFR dimers then phosphorylate each other and phosphorylated receptors (active receptors, pEGFRs) on the cell surface lead to downstream phosphorylation of Akt (pAkt). Both active and unphosphorylated (inactive) receptors are internalized with different rates from the cell surface through receptor endocytosis. After addition of EGF in the extracellular medium, pAkt levels increase transiently within minutes and then, as a results of receptor endocytosis and action of phosphatases, both pAkt and surface EGFR (sEGFR) levels decrease within hours after EGF stimulation (29).

To explore the cell-to-cell variability in this pathway, we used a dynamical model of EGF/EGFR dependent Akt phosphorylation based on Chen *et al*. (29). The model includes reactions describing EGF binding to EGFR and subsequent dimerization, phosphorylation, dephosphorylation, internalization, and degradation of the receptors. To keep the model relatively small, we simplified pEGFR-dependent phosphorylation of Akt by assuming a single step activation of pAkt by pEGFR with an effective rate constant (see SI section I for the model details).

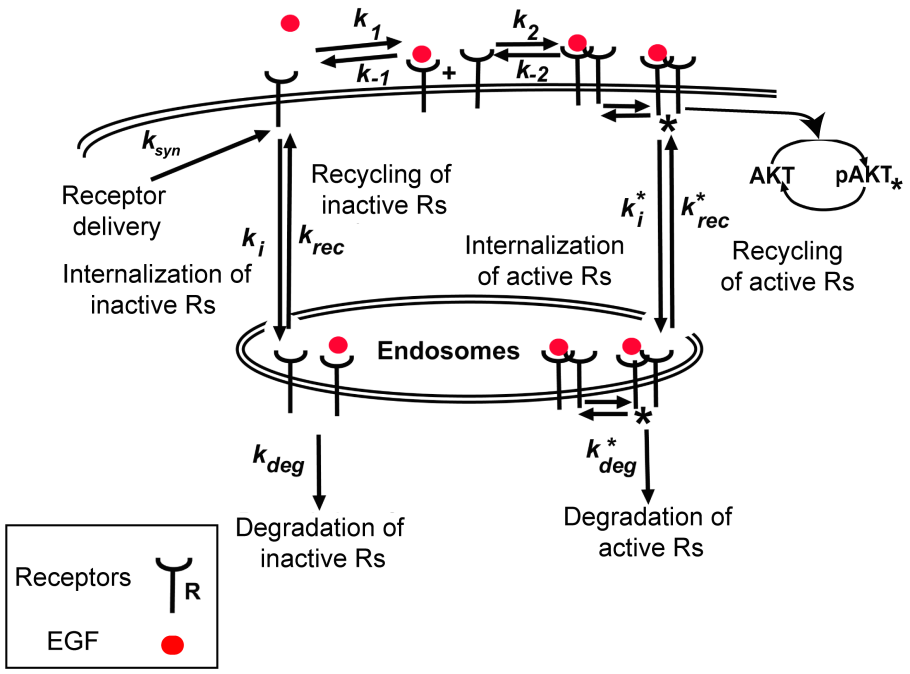

Figure 3. A schematic of the EGF/EGFR pathway leading to phosphorylation of Akt. Extracellular EGF binds to cell surface EGFRs leading to their dimerization. Dimerized EGFRs are autophosphorylated and in turn lead to phosphorylation of Akt. Receptors are also removed from cell surface through internalization into endosomes. See SI section I for details of the model.

Numerical inference of the parameter distribution

To derive the network parameter distribution consistent with experimental data using equation 6, we used experimentally measured cell-to-cell variability in pAkt and sEGFR levels at early times after EGF stimulation. Specifically, we used measured pAkt levels after stimulation with five different EGF doses (0.1, 0.316, 3.16, 10, and 100 ng/ml) at 4 early time points (5, 15, 30, and 45 minutes) (30). Additionally, we used sEGFR levels without EGF stimulation and after 3 hours of EGF stimulation at 0 ng/ml

and 1 ng/ml (see SI section II and III for details). We used 11 bins to represent each experimentally measured distribution; the bin sizes and locations were chosen to cover the entire range of observed variability. We numerically determined the Lagrange multipliers corresponding to the bind fractions using the procedure described above (see equation 6 and Figure 2).

Notably, the best-fit Lagrange multipliers accurately reproduced the experimentally measured bin fractions (Pearson's $r^2 = 0.84$, $p < 10^{-10}$, median relative error ~ 12%). Furthermore, fitted bin fractions obtained in two independent calculations showed excellent agreement with each other as expected for a convex optimization problem (Pearson's $r^2 = 0.98$, $p < 10^{-10}$, see SI Figure 1). In Figure 4, we show the temporal profile of measured cell-to-cell variability in pAkt levels (colored circles) at EGF stimulation of 10 ng/ml and the corresponding fits (dashed black lines) based on the inferred parameter distribution.

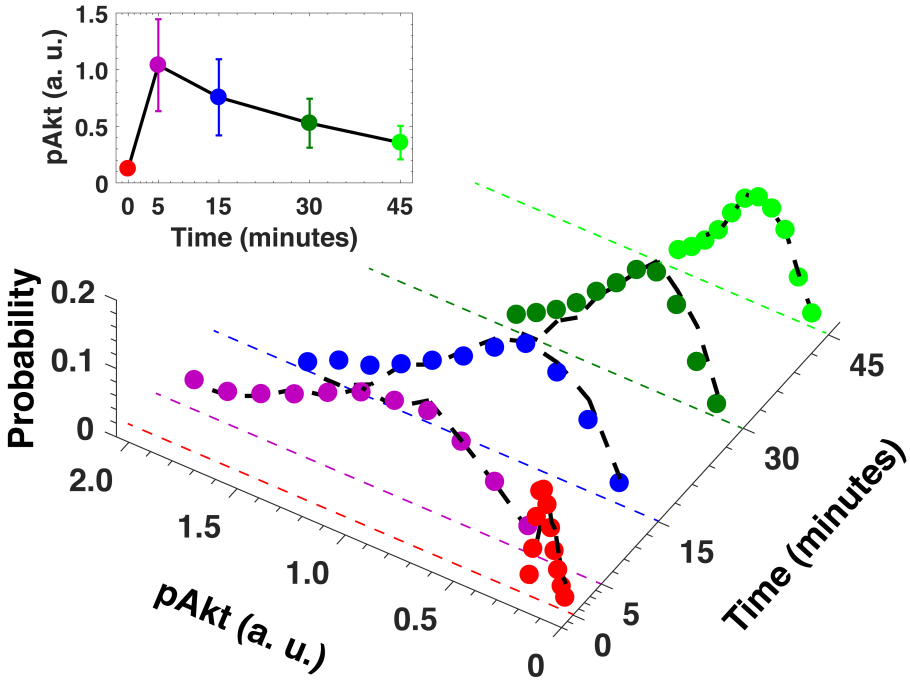

Figure 4. Experimental estimated cell-to-cell variability in pAkt levels used to infer the parameter distribution. We show distribution of pAkt levels at 0, 5, 15, 30, and 45 minutes after exposure to 10 ng/ml EGF. The colored circles represent the experimentally measured pAkt distributions used in the inference of the parameter distribution. The black dashed lines represent

fitted distributions. The inset shows the experimentally measured population average pAkt levels at multiple time points. Error bars in the inset represent population standard deviations.

Prediction of single cell dynamics

Akt is a key hub of mammalian cell signaling (24). Naturally, sustained activity of phosphorylated Akt (pAkt) is implicated in diverse human diseases, such as psychiatric disorders (31) and cancer (32, 33). Using the developed approach, we investigated whether we could predict pAkt levels hours after EGF stimulation using the parameter distribution inferred from pAkt variability at early times after EGF stimulation. To that end, we numerically sampled multiple parameter sets using the inferred parameter distribution and predicted pAkt levels at late time across a range of EGF stimulation levels corresponding to each parameter set. We compared predicted and experimentally observed distribution of pAkt levels across cells at late times (180 minutes) after sustained EGF stimulation (Figure 5a, b). Our simulations correctly predicted that a significant fraction of cells have high pAkt levels hours after stimulation; the predicted and observed coefficient of variation (CV) of the pAkt distributions in cells stimulated with 10 ng/ml EGF for 180 minutes were in good agreement, 0.41 and 0.37 respectively. Notably, the inferred parameter distribution accurately captures the population mean and variability (Figure 5c) in pAkt levels at late times across four orders of magnitude of EGF concentrations used to stimulate cells.

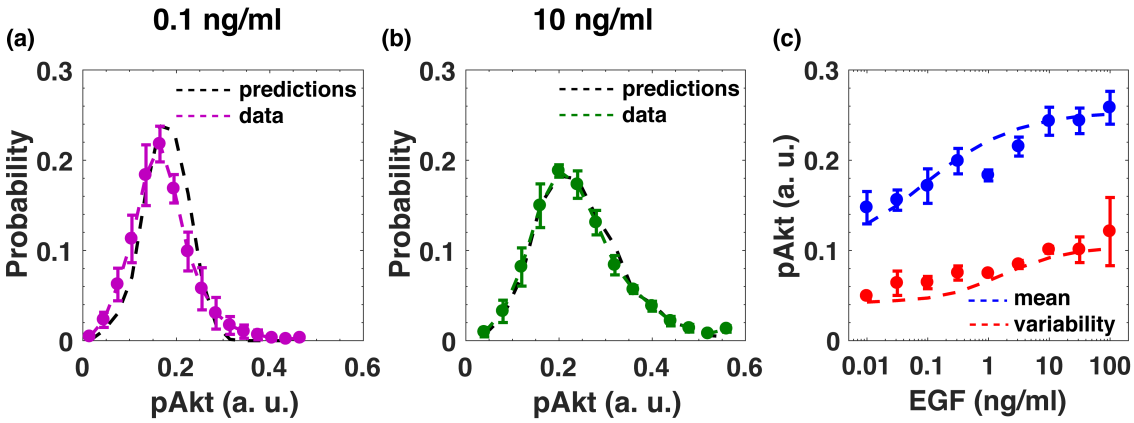

Figure 5. Prediction pAkt levels at late times. (a and b) Measured distributions (colored circles and lines) and the corresponding predictions (dashed black lines) of cell-to-cell variability in

pAkt levels at 180 minutes after stimulation with 0.1 ng/ml and 10 ng/ml EGF respectively. (c) Measured mean pAkt levels (blue circles) and measured standard deviation in pAkt levels (red circles) at 180 minutes after sustained stimulation with EGF (x-axis) and the corresponding predictions (dashed blue line and dashed red line respectively). The error bars represent standard deviation estimated from four replicates.

EGF stimulation leads to internalization and degradation of EGFRs (34), and cell surface receptor levels reach an EGF-dependent steady state within hours of continuous stimulation (30). The steady state abundance of EGFRs determines cells' ability to phosphorylate downstream signaling targets in response to subsequent EGF stimulation (35). Therefore, we applied the developed computational approach to evaluate the heterogeneity of EGFR levels after prolonged stimulation (180 minutes) with EGF. We then compared the predicted and the experimentally measured the steady state EGFR levels across EGF stimulation doses. Similar to pAkt, the simulations accurately captured both the population mean and variability of the EGFR receptor levels across multiple doses of EGF stimulations (Figure 6c). The simulations and experiments demonstrate that there is a significant fraction of cells with relatively high levels of EGFR even hours after the growth factor stimulation (Figure 6a, b). Interestingly, across simulated trajectories, the steady state cell surface EGFR levels showed the highest correlation with sustained pAkt level among all receptor-related parameters (Pearson $r = 0.44$, $p < 10^{-10}$, EGF stimulation 10 ng/ml), suggesting that cells with high levels of steady state surface EGFR levels likely predominantly contribute to the cell sub-population with high steady state pAkt levels.

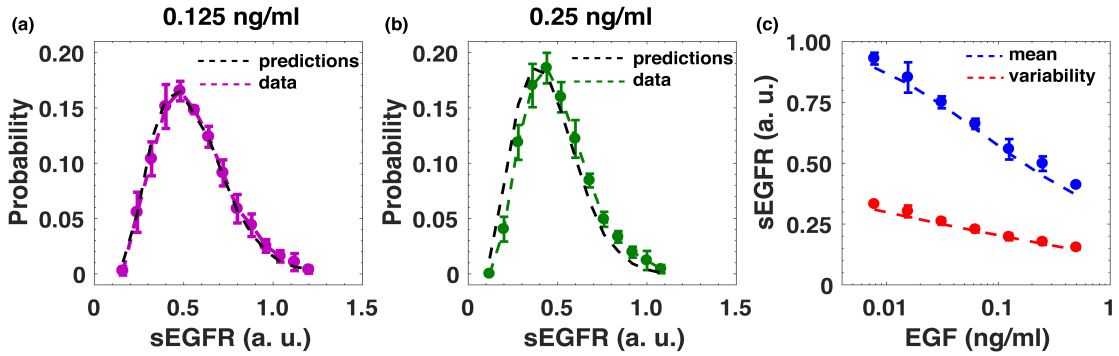

Figure 6. Prediction of sEGFR levels at late times. (a) and (b) Measured distributions (colored circles lines) and the corresponding predictions (dashed black lines) of cell-to-cell variability in sEGFR levels at 180 minutes after stimulation with 0.125 ng/ml and 0.25 ng/ml EGF respectively. (c) Measured mean sEGFR levels (blue circles) and measured standard deviation in

sEGFR levels (red circles) at 180 minutes after sustained stimulation with EGF (x-axis) and the corresponding predictions (dashed blue line and dashed red line respectively). The error bars represent standard deviation estimated from three replicates.

Possible extensions of the ME framework

Importantly, a straightforward extension makes it possible to use the developed framework for networks when the time evolution of species abundances $\bar{X}$ is intrinsically stochastic. For example, transcriptional networks and prokaryotic signaling networks with relatively small species abundances (1). To that end, we can modify the definition of the predicted fraction $\psi_b = \int P\left(x\left(t,\bar{\theta}\right)=x\middle|\bar{\theta}\right)\bullet I_b\left(x\right)dx$ of a chemical species *x,* where $P\left(x\left(t,\bar{\theta}\right)=x\middle|\bar{\theta}\right)$ is the distribution of *x* values at time *t* with parameters $\bar{\theta}$. The distributions can be obtained numerically using Gillespie's stochastic algorithm (36) or approximated using moment closure techniques (37).

The presented framework can also be used to infer parameter distributions when, instead of the entire abundance distributions only a few moments of the distribution are available, such as average protein abundances measured using quantitative western blots or mass spectrometry (38). For example, we consider the case where the population mean *m* and the variance *v* of one species *x* are measured at a fixed time point *t*. Instead of constraining fractions ψ_b that represent cell-to-cell variability in different bins of the relevant abundance distribution, we can constrain the population mean $\mu_1 = \int x\left(t,\bar{\theta}\right)P\left(\bar{\theta}\right)d\bar{\theta}$ and the second moment $\mu_2 = \int x\left(t,\bar{\theta}\right)^2 P\left(\bar{\theta}\right)d\bar{\theta}$ to their experimentally measured values *m* and $v+m^2$ respectively. Entropy maximization can then be carried out with these constraints. In this case, we have

$$P\left(\bar{\theta}\right)=\frac{1}{\Omega}q\left(\bar{\theta}\right)\exp\left(-\lambda_1 x\left(t,\bar{\theta}\right)-\lambda_2 x\left(t,\bar{\theta}\right)^2\right) \tag{7}$$

Lastly, we can use the ME framework to infer parameters from experiments where dynamics of species abundances within single cells are measured using live cell imaging (5). For example, consider that the time evolution of a species $x\left(t\right)$ is measured

in n_c cells from time $t=0$ to $t = T$. We can discretize the n_c continuous time observations into K discrete times $t = \{t_1, t_2, \ldots, t_K\}$. At each time point t_i, we can then divide the range of observed abundances in B_i bins. Then, each individual dynamical trajectory $x(t)$ can be characterized by a vector of discrete indices $x(t) \sim \{B_{1a_1}, B_{2a_2}, B_{3a_3}, \ldots, B_{Ka_K}\}$ where B_{ia_i} is the index of the abundance distribution bin through which the trajectory $x(t)$ passed at time point t_i. Given a sufficiently large number of trajectories, we then can constrain the fraction of trajectories that populate a given sequence of bins to infer the parameter distribution.

Discussion

In this work we presented a maximum entropy framework to infer the joint distribution $P(\bar{\theta})$ over signaling network parameters based on experimentally measured cell-to-cell variability in species abundances. We demonstrated that the obtained parameter distribution allow accurate prediction of the time evolution of cell-to-cell variability in species abundances. The inferred distribution takes into account both parameter non-identifiability and their cell-to-cell variability. The contribution due to parameter non-identifiability can be further minimized by explicitly incorporating known constraints in parameter values using prior (27). Notably, the maximum entropy framework naturally avoids over-constraining the parameter distribution as redundant constraints lead to redundant Lagrange multipliers (39).

Recent developments in cytometry (40) and single cell RNA sequencing (41) make it possible to simultaneously quantify multiple species abundances in single cells. Elegant statistical approaches have been developed to reconstruct trajectories of intracellular species dynamics consistent with time-stamped single cell abundance data (27, 42, 43). Complementary to these statistical methods, our approach (1) allows us to infer the distribution over parameters that describe mechanistic interactions in the signaling network and moreover (2) the inferred parameter distribution can be used to predict the ensemble of single cell trajectories for time intervals and experimental conditions beyond the measured abundance distributions.

In this work, we applied the developed framework to signaling network data. However, it can also be used in other diverse research contexts. For example, the framework can be applied to computationally reconstruct the distribution of longitudinal behaviors from cross-sectional time-snapshot data in fields such public health, economics, and ecology or to estimate parameter distributions from lower dimensional statistics (44).

Acknowledgment and Funding

We would like to thank Jan Hasenauer and Caroline Loos for suggestions about the manuscript. We acknowledge funding from NIH grants R01CA201276-03 and U54CA209997-02.

References

1. Raj A & van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135(2):216-226.
2. Chastanet A, *et al.* (2010) Broadly heterogeneous activation of the master regulator for sporulation in Bacillus subtilis. *Proc Natl Acad Sci U S A* 107(18):8486-8491.
3. Spencer SL, Gaudet S, Albeck JG, Burke JM, & Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 459(7245):428-432.
4. Wu M & Singh AK (2012) Single-cell protein analysis. *Curr Opin Biotech* 23(1):83-88.
5. Meyer R, *et al.* (2012) Heterogeneous kinetics of AKT signaling in individual cells are accounted for by variable protein concentration. *Front Physiol* 3:451.
6. Niepel M, Spencer SL, & Sorger PK (2009) Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr Opin Chem Bio* 13(5-6):556-561.
7. Snijder B, *et al.* (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* 461(7263):520-523.
8. Raue A, *et al.* (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 25(15):1923-1929.
9. Eydgahi H, *et al.* (2013) Properties of cell death models calibrated and compared using Bayesian approaches. *Mol Syst Biol* 9:644.
10. Transtrum MK, *et al.* (2015) Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys* 143(1):010901.
11. Waldherr SH, J. Allgower, F. (2009) Estimation of biochemical network parameter distributions in cell populations. *Proceedings of the 15th IFAC Symposium on System Identification*:1265-1270.

12. Hasenauer J, *et al.* (2011) Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics* 12:125.
13. Zechner C, *et al.* (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci U S A* 109(21):8340-8345.
14. Zechner C, Unger M, Pelet S, Peter M, & Koeppl H (2014) Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat Methods* 11(2):197-202.
15. Dixit PD (2013) Quantifying extrinsic noise in gene expression using the maximum entropy framework. *Biophys J* 104(12):2743-2750.
16. Dixit PD, *et al.* (2018) Perspective: Maximum caliber is a general variational principle for dynamical systems. *J Chem Phys* 148(1):010901.
17. Weigt M, White RA, Szurmant H, Hoch JA, & Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 106(1):67-72.
18. Mora T, Walczak AM, Bialek W, & Callan CG, Jr. (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci U S A* 107(12):5405-5410.
19. Schneidman E, Berry MJ, 2nd, Segev R, & Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007-1012.
20. Dixit PD, Jain A, Stock G, & Dill KA (2015) Inferring Transition Rates of Networks from Populations in Continuous-Time Markov Processes. *J Chem Th Comp* 11(11):5464-5472.
21. Dixit PD & Dill KA (2014) Inferring Microscopic Kinetic Rates from Stationary State Distributions. *J Chem Th Comp* 10(8):3002-3005.
22. Tiwary P & Berne BJ (2016) Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc Natl Acad Sci U S A* 113(11):2839-2844.
23. Dixit PD (2018) Communication: Introducing prescribed biases in out-of-equilibrium Markov models. *J Chem Phys* 148(9):091101.
24. Manning BD & Toker A (2017) AKT/PKB Signaling: Navigating the Network. *Cell* 169(3):381-405.
25. Herbst RS (2004) Review of epidermal growth factor receptor biology. *Int J Radiat Oncol Biol Phys* 59(2):21-26.
26. Soule HD, *et al.* (1990) Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer research* 50(18):6075-6086.
27. Mukherjee S, Stewart D., Stewart W., Lanier L. L, & Das J (2017) Connecting the dots across time: reconstruction of single-cell signalling trajectories using time-stamped data. *Royal Soc Op Sci* 4:170811.
28. Tkacik G, Schneidman E, Berry MJ, 2nd., & Bialek W (2006) Ising models for networks of real neurons. *arXiv* q-bio/0611072.
29. Chen WW, *et al.* (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol Syst Biol* 5:239.
30. Lyashenko E, *et al.* (2017) Receptor-based mechanism of relative sensing in mammalian signaling networks. *bioRxiv*:10.1101/158774.

31. Gilman SR, *et al.* (2012) Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat Neurosci* 15(12):1723-1728.
32. Vivanco I & Sawyers CL (2002) The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat Rev Cancer* 2(7):489-501.
33. Nicholson KM & Anderson NG (2002) The protein kinase B/Akt signalling pathway in human malignancy. *Cell Signal* 14(5):381-395.
34. Wiley HS, *et al.* (1991) The role of tyrosine kinase activity in endocytosis, compartmentation, and down-regulation of the epidermal growth factor receptor. *J Biol Chem* 266(17):11083-11094.
35. Normanno N, *et al.* (2006) Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* 366(1):2-16.
36. Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* 58:35-55.
37. Gillespie CS (2009) Moment-closure approximations for mass-action models. *IET Syst Biol* 3(1):52-58.
38. Shi T, *et al.* (2016) Conservation of protein abundance patterns reveals the regulatory architecture of the EGFR-MAPK pathway. *Sci Signal* 9(436):rs6.
39. Dixit PD & Dill KA (2018) Caliber Corrected Markov Modeling (C2M2): Correcting Equilibrium Markov Models. *J Chem Theory Comput* 14(2):1111-1119.
40. Chattopadhyay PK, Gierahn TM, Roederer M, & Love JC (2014) Single-cell technologies for monitoring immune systems. *Nat Immunol* 15(2):128-135.
41. Saliba AE, Westermann AJ, Gorski SA, & Vogel J (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 42(14):8845-8860.
42. Gut G, Tadmor MD, Pe'er D, Pelkmans L, & Liberali P (2015) Trajectories of cell-cycle progression from fixed cell populations. *Nat Methods* 12(10):951-954.
43. Mukherjee S, *et al.* (2017) In silico modeling identifies CD45 as a regulator of IL-2 synergy in the NKG2D-mediated activation of immature human NK cells. *Sci Signal* 10(485).
44. Das J, Mukherjee S, & Hodge SE (2015) Maximum Entropy Estimation of Probability Distribution of Variables in Higher Dimensions from Lower Dimensional Data. *Entropy (Basel)* 17(7):4986-4999.