Chu et al. BMC Systems Biology (2017) 11:14
DOI 10.1186/s12918-017-0394-4

BMC Systems Biology

RESEARCH ARTICLE Open Access

Markov State Models of gene regulatory networks

Brian K. Chu[1], Margaret J. Tse[1], Royce R. Sato[1] and Elizabeth L. Read[1,2*]

Abstract

Background: Gene regulatory networks with dynamics characterized by multiple stable states underlie cell fate-decisions. Quantitative models that can link molecular-level knowledge of gene regulation to a global understanding of network dynamics have the potential to guide cell-reprogramming strategies. Networks are often modeled by the stochastic Chemical Master Equation, but methods for systematic identification of key properties of the global dynamics are currently lacking.

Results: The method identifies the number, phenotypes, and lifetimes of long-lived states for a set of common gene regulatory network models. Application of transition path theory to the constructed Markov State Model decomposes global dynamics into a set of dominant transition paths and associated relative probabilities for stochastic state-switching.

Conclusions: In this proof-of-concept study, we found that the Markov State Model provides a general framework for analyzing and visualizing stochastic multistability and state-transitions in gene networks. Our results suggest that this framework—adopted from the field of atomistic Molecular Dynamics—can be a useful tool for quantitative Systems Biology at the network scale.

Keywords: Multistable systems, Stochastic processes, Gene regulatory networks, Markov State Models, Cluster analysis

Background

Gene regulatory networks (GRNs) often have dynamics characterized by multiple attractor states. This multistability is thought to underlie cell fate-decisions. According to this view, each attractor state accessible to a gene network corresponds to a particular pattern of gene expression, i.e., a cell phenotype. Bistable network motifs with two possible outcomes have been linked to binary cell fate-decisions, including the lysis/lysogeny decision of bacteriophage lambda [1], the maturation of frog oocytes [2] and a cascade of branch-point decisions in mammalian cell development (reviewed in [3]). Multistable networks with three or more attractors have been proposed to govern diverse cell fate-decisions in tumorigenesis [4], stem cell differentiation and reprogramming [5–7], and helper T cell differentiation [8]. More generally, the concept of a rugged, high-dimensional epigenetic landscape connecting every possible cell type has emerged [9–11]. Quantitative models that can link molecular-level knowledge of gene regulation to a global understanding of network behavior have the potential to guide rational cell-reprogramming strategies. As such, there has been growing interest in the development of theory and computational methods to analyze global dynamics of multistable gene regulatory networks.

Gene expression is inherently stochastic [1, 12–14], and fluctuations in expression levels can measurably impact cell phenotypes and behavior. Numerous examples of stochastic phenotype transitions have been discovered, which diversify otherwise identical cell-populations. This spontaneous state-switching has been found to promote survival of microorganisms or cancer cells in fluctuating environments [15–17], prime cells to follow alternate developmental fates in higher eukaryotes [18, 19], and generate sustained heterogeneity (mosaicism) in a homeostatic mammalian cell population [20]. These findings have motivated theoretical studies of stochastic state-switching in

* Correspondence: elread@uci.edu
[1]Department of Chemical Engineering and Materials Science, University of California Irvine, Irvine, CA, USA
[2]Department of Molecular Biology and Biochemistry, University of California Irvine, Irvine, CA, USA

gene networks, which have shed light on network parameters and topologies that promote the stability (or instability) of a given network state [20]. Characterizing the global stability of states accessible to a network is akin to quantification of the "potential energy" landscape of a network. Particularly, with the advent of stem-cell reprogramming techniques, there has been renewed interest in a quantitative reinterpretation of Waddington's classic epigenetic landscape [21], in terms of underlying regulatory mechanisms [10, 22].

A number of mathematical frameworks exist for modeling and analysis of stochastic gene regulatory network (GRN) dynamics (reviewed in [23, 24]), including probabilistic Boolean Networks, Stochastic Differential Equations, and stochastic biochemical reaction networks (i.e., Chemical Master Equations). Of these, the Chemical Master Equation (CME) approach is the most complete, in that it treats all biomolecules in the system as discrete entities, fully accounts for stochasticity due to molecular-level fluctuations, and propagates dynamics according to chemical rate laws. The CME is analytically intractable for GRNs except in some simplified model systems [25–29], but trajectories can be simulated by Monte Carlo methods such as the Stochastic Simulation Algorithm (SSA) [30]. Alternatively, methods for reducing the dimensionality of the CME, enabling numerical approximation of network behavior by matrix methods, have been developed [31–35].

Analysis of multistability and global dynamics of discrete, stochastic GRN models remains challenging. In this study, we define multistability in stochastic systems as the existence of multiple peaks in the stationary probability distribution. In such systems, the GRN dynamics can be considered somewhat analogous to that of a particle in a multi-well potential [3]. (Peaks in the probability distribution—or alternatively, basins in the potential—may or may not correspond to stable fixed points of a corresponding ODE model, as discussed in more detail further on.) Stochastic multistability is often assessed by plotting multi-peaked steady-state probability distributions (obtained either from long stochastic simulations [5, 36, 37] or from approximate CME solutions [35, 38, 39]), projected onto one or two user-specified system coordinates. However, even small networks generally have more than two dimensions along which dynamics may be projected, meaning that inspection of steady-state distributions for a given projection may underestimate multistability in a network. For example, the state-space of a GRN may comprise different activity-states of promoters and regulatory sites on DNA, the copy-number of mRNA transcripts and encoded proteins, and the activity- or multimer-states of multiple regulatory molecules or proteins. Furthermore, while steady-state distributions give a global view of system behavior, they do not directly yield dynamic information of interest, such as the lifetimes of attractor states.

In this paper, we present an approach for analyzing multistable dynamics in stochastic GRNs based on a spectral clustering method widely applied in Molecular Dynamics [40, 41]. The output of the approach is a Markov State Model (MSM)—a coarse-grained model of system dynamics, in which a large number of system states (i.e., "microstates") is clustered into a small number of metastable (that is, relatively long-lived) "macrostates", together with the conditional probabilities for transitioning from one macrostate to another on a given timescale. The MSM approach identifies clusters based on separation of timescales, i.e., systems with multistability exhibit relatively fast transitions among microstates within basins and relatively slow inter-basin transitions. By neglecting fast transitions, the size of the system is vastly reduced. Based on its utility for visualization and analysis of Molecular Dynamics, the potential application of the MSM framework to diverse dynamical systems, including biochemical networks, has been discussed [42].

Biochemical reaction networks present an unexplored opportunity for the MSM approach. Herein, we applied the method to small GRN motifs and analyzed their global dynamics using two frameworks: the quasipotential landscape (based on the log-transformed stationary probability distribution), and the MSM. The MSM approach distilled network dynamics down to the essential stationary and dynamic properties, including the number and identities of stable phenotypes encoded by the network, the global probability of the network to adopt a given phenotype, and the likelihoods of all possible stochastic phenotype transitions. The method revealed the existence of network states and processes not readily apparent from inspection of quasipotential landscapes. Our results demonstrate how MSMs can yield insight into regulation of cell phenotype stability and reprogramming. Furthermore, our results suggest that, by delivering systematic coarse-graining of high-dimensional (i.e., many-species) dynamics, MSMs could find more general applications in Systems Biology, such as in signal-transduction, evolution, and population dynamics. In our implementation, the MSM framework is applied to the CME, thus mapping all enumerated molecular states onto long-lived system macrostates. We anticipate that the method could in future studies be used to analyze more complex systems where enumeration of the CME is intractable, if implemented in combination with stochastic simulation or other model reduction approaches.

Methods

Gene regulatory network motifs

We studied two common GRN motifs that are thought to control cell fate-decisions. The full lists of reactions and associated rate parameters for each network are given in the Additional file 1. Both motifs consist of two mutually-inhibiting genes, denoted by A and B. In the Exclusive Toggle Switch (ETS) motif, each gene encodes a transcription factor protein; the protein forms homodimers, which are capable of binding to the promoter of the competing gene, thereby repressing its expression. One DNA-promoter region controls the expression of both genes; when a repressor is bound, it excludes the possibility of binding by the repressor encoded by the competing gene. Therefore, the promoter can exist in three possible binding configurations, P_{00}, P_{10}, and P_{01}, denoting the unbound, a_2-bound, or b_2-bound states, respectively. Production of new protein molecules (including all processes involved in transcription, translation, and protein synthesis) occurs at a constant rate, which depends on the state of the promoter. When the gene is repressed, the encoded protein is produced at a low rate, denoted g_0. When the gene is not repressed, protein is produced at a high rate, g_1. For example, when the promoter state is P_{10} the a protein is produced at rate g_1, and the b protein is produced at g_0. When the promoter is unbound, neither gene is repressed, causing both proteins to be produced at rate g_1.

In the Mutual Inhibition/Self-Activation (MISA) motif, each homodimeric transcription factor also activates its own expression, in addition to repressing the other gene. The A and B genes are controlled by separate promoters, and each promoter can be bound by repressor and activator simultaneously. Therefore, the A-promoter can exist in four possible states, A_{00}, A_{10}, A_{01} and A_{11}, denoting unbound, a_2-activator bound, b_2-repressor bound, and both transcription factors bound, respectively (and similarly for the B-promoter). Proteins are produced at rate g_1 only when the activator is bound and the repressor is unbound. For example, the A_{10} promoter state allows a protein to be produced at g_1. The other three A promoter states result in a protein being produced at rate g_0. Similarly, the rate of b protein production depends only on the binding configuration of the B-promoter. In both the ETS and MISA networks, protein dimerization is assumed to occur simultaneously with binding to DNA. All rate parameters are given in Additional file 1: Tables S1 and S2.

Chemical master equation

The stochastic dynamics are modeled by the discrete, Markovian Chemical Master Equation, which gives the time-evolution of the probability to observe the system in a given state over time. In vector–matrix form, the CME can be written

$$\frac{d\mathbf{p}(\mathbf{x},t)}{dt} = \mathbf{K}\mathbf{p}(\mathbf{x},t)$$

where $\mathbf{p}(\mathbf{x},t)$ is the probability over the system state-space at time t, and $\mathbf{K}$ is the reaction rate-matrix. The off-diagonal elements K_{ij} give the time-independent rate of transitioning from state $\mathbf{x}_i$ to $\mathbf{x}_j$, and the diagonal elements are given by $K_{ii} = -\sum_{j \neq i} K_{ji}$. We assume a well-mixed system of reacting species, and the state of the system is fully specified by $\mathbf{x} \in \mathbb{N}^S$, a state-vector containing the positive-integer values of all S molecular species/configurations. We hereon denote these state-vectors as "microstates" of the system. In the ETS network, $\mathbf{x} = [n_A, n_B, P_{ab}]$, where n_A is the copy-number of a molecules (protein monomers expressed by gene A, and similar for B), and P_{ab} indexes the promoter binding-configuration. In the MISA network, $\mathbf{x} = [n_A, n_B, A_{ab}, B_{ba}]$, which lists the protein copy numbers and promoter configuration-states associated with both genes.

The reaction rate matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ is built from the stochastic reaction propensities (Additional file 1: Eq. 1), for some choice of enumeration over the state-space with N reachable microstates. In general, if a system of S molecular species has a maximum copy number per species of $n_{\max}$, then $N \sim n_{\max}^{S}$. To enumerate the system state-space, we neglect microstates with protein copy-numbers larger than a threshold value, which exceeds the maximum steady-state gene expression rate, g_1/k, (where g_1 is the maximum production rate of protein and k is the degradation rate), as these states are rarely reached. This truncation of the state-space introduces a small approximation error, which we calculate using the Finite State Projection method [31] (Additional file 1: Figure S1).

Stochastic simulations

Stochastic simulations were performed according to the SSA method, implemented by the software package StochKit2 [43].

Quasipotential landscape

The steady-state probability $\boldsymbol{\pi}(\mathbf{x})$ over N microstates is obtained from $\mathbf{K}$ as the normalized eigenvector corresponding to the zero-eigenvalue, satisfying $\mathbf{K}\boldsymbol{\pi}(\mathbf{x}) = 0$ [44]. Quasipotential landscapes were obtained from $\boldsymbol{\pi}(\mathbf{x})$ using a Boltzmann definition, $U(\mathbf{x}) = -\ln(\boldsymbol{\pi}(\mathbf{x}))$ [22]. All matrix calculations were performed with MATLAB [45].

Markov State Models: mathematical background

The last 15 years have seen continual progress in development of theory, algorithms, and software implementing

the MSM framework. We briefly summarize the theoretical background here; the reader is referred to other works (e.g., [41, 46–49]) for more details.

The MSM is a highly coarse-grained projection of system dynamics over N microstates onto a reduced space of selected size C (generally, $C \ll N$). The C states in the projected dynamics are constructed by clustering together microstates that experience relatively fast transitions among them. The C clusters, also called "almost invariant aggregates" [48], are hereon denoted "macrostates".

The MSM approach makes use of Robust Perron Cluster Analysis (PCCA+), a spectral clustering algorithm that takes as input a row-stochastic transition matrix, $\mathbf{T}(\tau)$ which gives the conditional probability for the system to transition between each pair of microstates within a given lagtime τ. The lagtime determines the time-resolution of the model, as expressed by the transition matrix. Off-diagonal elements T_{ij} give the probability of finding the system in microstate j at time $t + \tau$, given that it was in microstate i at time t. Diagonal elements T_{ii} give the conditional probability of again finding the system in microstate i at time $t + \tau$, and thus rows sum to 1. $\mathbf{T}(\tau)$ is directly obtained from the reaction rate matrix by [50]:

$$\mathbf{T}(\tau) = \exp\left(\tau \mathbf{K}^{\mathrm{T}}\right),$$

(where exp denotes the matrix exponential). The evolution of the probability over discrete intervals of τ is given by the Chapman-Kolmogorov equation,

$$\boldsymbol{p}^{T}(\mathbf{x}, t + k\tau) = \boldsymbol{p}^{T}(\mathbf{x}, t)\mathbf{T}^{k}(\tau).$$

For an ergodic system (i.e., any state in the system can be reached from any other state in finite time), $\mathbf{T}(\tau)$ will have one largest eigenvalue, the Perron root, $\lambda_1 = 1$. The stationary probability is then given by the normalized left-eigenvector corresponding to the Perron eigenvalue,

$$\boldsymbol{\pi}^{\mathrm{T}}(\mathbf{x})\mathbf{T}(\tau) = \boldsymbol{\pi}^{\mathrm{T}}(\mathbf{x}).$$

If the system exhibits multistability, then the dynamics can be approximately separated into fast and slow processes, with fast transitions occurring between microstates belonging to the same metastable macrostate, and slow transitions carrying the system from one macrostate to another. Then $\mathbf{T}(\tau)$ is nearly decomposable, and will exhibit an almost block-diagonal structure (for an appropriate ordering of microstates) with C nearly uncoupled blocks. In this case, the eigenvalue spectrum of $\mathbf{T}(\tau)$ shows a cluster of C eigenvalues near $\lambda_1 = 1$, denoting C slow processes (including the stationary process), and for $i > C$, $\lambda_i \ll \lambda_C$, corresponding to rapidly decaying processes. The system timescales can be computed from the eigenvalue spectrum according to $t_i = -\tau/\ln|\lambda_i(\tau)|$.

The PCCA+ algorithm obtains fuzzy membership vectors $\boldsymbol{\chi} = [\chi_1, \chi_2, \ldots, \chi_C] \in \mathbb{R}^{N \times C}$, which assigns microstates $i \in \{1, \ldots, N\}$ to macrostates $j \in \{1, \ldots, C\}$ according to grades (i.e., probabilities) of membership, $\chi_j(i) \in [0, 1]$. The membership vectors satisfy the linear transformation:

$$\boldsymbol{\chi} = \boldsymbol{\psi}\mathbf{B}$$

Where $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_C]$ is the $N \times C$ matrix constructed from the C dominant right-eigenvectors of $\mathbf{T}(\tau)$, and $\mathbf{B}$ is a non-singular matrix that transforms the dominant eigenvectors into membership vectors. The coarse-grained $C \times C$ transition matrix $\tilde{\mathbf{T}}(\tau) \in \mathbb{R}^{C \times C}$ (i.e., the Markov State Model) is then obtained as the projection of $\mathbf{T}(\tau)$ onto the C sets by:

$$\tilde{\mathbf{T}}(\tau) = \tilde{\boldsymbol{D}}^{-1}\boldsymbol{\chi}^{T}\boldsymbol{D}\mathbf{T}(\tau)\boldsymbol{\chi}$$

where $\boldsymbol{D}$ is the diagonal matrix obtained from the stationary probability vector, $\boldsymbol{D} = \mathrm{diag}(\pi_1, \ldots, \pi_N)$. The coarse-grained probability $\tilde{\boldsymbol{\pi}}(x)$ is obtained by $\tilde{\boldsymbol{\pi}}(\mathbf{x}) = \boldsymbol{\chi}^{\mathrm{T}}\boldsymbol{\pi}(\mathbf{x})$, and $\tilde{\boldsymbol{D}} = \mathrm{diag}(\tilde{\pi}_1, \ldots, , \tilde{\pi}_C)$. The elements of the linear transformation matrix $\mathbf{B}$ are obtained by an optimization procedure, with "metastability" of the resultant coarse-grained projection as the objective function to be maximized. The trace of the coarse-grained transition matrix, $trace[T]$ has been taken to be the measure of metastability, because it expresses the probabilities for the system to remain in metastable states over the lagtime (i.e., maximizing the sum over the diagonal elements). The original PCCA method [48] used the sign structure of the eigenvectors to identify almost invariant aggregates (instead of this optimization procedure), and more recent work has identified an alternative objective function [49]. The results of this paper were generated using the PCCA+ implementation of MSMBuilder2 [51].

Construction of Markov State Models and pathway decomposition

The PCCA+ algorithm generates a fuzzy discretization. We convert fuzzy values into a so-called "crisp" partitioning of N states into C clusters, which entirely partitions the space with no overlap, by assigning $\chi_j^{crisp}(i) \in \{0, 1\}$. That is, $\chi_j^{crisp}(i) = 1$ if the jth element of the row vector $\chi(i)$ is maximal, and 0 otherwise. Transition probabilities are estimated over the C coarse-grained sets by summing over the fluxes, or equivalently:

$$\tilde{\mathbf{T}}(\tau) = \tilde{\boldsymbol{D}}^{-1}\boldsymbol{\chi}^{T}\boldsymbol{D}\mathbf{T}(\tau)\boldsymbol{\chi},$$

where $\tilde{\mathbf{T}}(\tau) \in \mathbb{R}^{C \times C}$ is the coarse-grained Markov State Model and $\boldsymbol{D}$ is the diagonal matrix obtained from the

stationary probability vector, $\boldsymbol{D} = \mathrm{diag}(\pi_1, \ldots, \pi_N)$. The coarse-grained probability $\tilde{\boldsymbol{\pi}}(\mathbf{x})$ is obtained by $\tilde{\boldsymbol{\pi}}(\mathbf{x}) = \boldsymbol{\chi}^T \boldsymbol{\pi}(\mathbf{x})$, and $\tilde{\boldsymbol{D}} = \mathrm{diag}(\tilde{\pi}_1, \ldots, , \tilde{\pi}_C)$.

The Markov State Model is visualized using the PyEmma 2 plotting module [46], where the magnitude of the transition probabilities and steady state probabilities are represented by the thickness of the arrows and size of the circles, respectively.

Upon construction of the Markov State Model, transition-path theory [52–54] was applied in order to compute an ensemble of transition paths connecting two states of interest, along with their relative probabilities. This was achieved by applying a pathway decomposition algorithm adapted from Noe, et al. in a study of protein folding pathways [54] (details in Additional file 1). A summary of the workflow used in generating the results of this paper is included in the Additional file 1: Supplement S5.

Results

Eigenvalues and Eigenvectors of the stochastic transition matrix reveal slow dynamics in gene networks

In order to explore the utility of the MSM approach for analyzing global dynamics of gene networks, we studied common motifs that control lineage decisions. The MISA network motif (Fig. 1a, Additional file 1: Supplement S1, and Methods) has been the subject of previous theoretical studies and is thought to appear in a wide variety of binary fate-decisions [5, 55, 56]. In the network model, the *A/B* gene pair represents known antagonistic pairs such as Oct4/Cdx2, PU.1/Gata1, and GATA3/T-bet, which control lineage decisions in embryonic stem cells, common myeloid progenitors, and naïve T-helper cells, respectively [9, 57, 58]. In general, a particular cell lineage will be associated with a phenotype in which one of the genes is expressed at a high level, and the other is expressed at a low (repressed) level. The MISA network as an ODE model has been reported to have

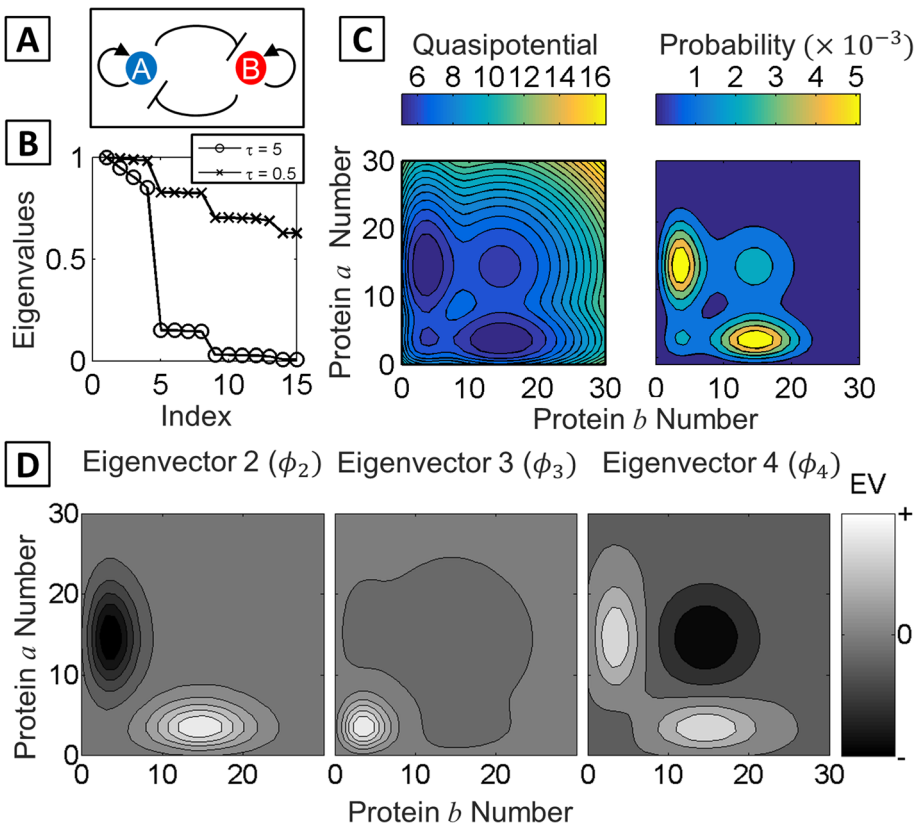

Fig. 1 Eigenvalue and eigenvector analysis of the Mutual Inhibition/Self Activation (MISA) network. **a** Schematic of the MISA network motif. **b** The fifteen largest eigenvalues of the stochastic transition matrix $\mathbf{T}(\tau)$, indexed in descending order, for $\tau = 5$ (*circles*) and $\tau = 0.5$ (*crosses*) (time units of inverse protein degradation rate, k^{-1}). Gaps indicate separation between processes occurring on different timescales. Network parameter values are listed in Additional file 1: Table S1. **c** The quasipotential landscape (*left*) and probability landscape (*right*) for the MISA motif, projected onto the A vs. B protein copy number subspace, showing four visible basins Landscapes were obtained from $\boldsymbol{\phi}_1$, the eigenvector associated with the largest eigenvalue of $\mathbf{T}(\tau)$. **d** *Left* to *right*: second, third, and fourth eigenvectors ($\boldsymbol{\phi}_2$, $\boldsymbol{\phi}_3$, $\boldsymbol{\phi}_4$) of $\mathbf{T}(\tau)$. The sign structure reveals the nature of the slowest dynamical processes (see text)

up to four stable fixed-points corresponding to the *A/B* gene pair expression combinations Lo/Lo, Lo/Hi, Hi/Lo, and Hi/Hi. We computed the probability and quasipotential landscape of the MISA network. For a symmetric system with sufficiently balanced rates of activator and repressor binding and unbinding from DNA, four peaks (or basins) can be distinguished in the steady state probability (quasipotential) landscape, plotted as a function of protein *a* copy number vs. protein *b* copy number (Fig. 1a, b). Quasipotentials computed from $\boldsymbol{\pi}(\mathbf{x})$, the Perron eigenvector of the transition matrix (see Methods) and from a long stochastic simulation showed agreement (Additional file 1: Figure S2).

The Markov State Model framework has been applied in studies of protein folding, where dynamics occurs over rugged energetic landscapes characterized by multiple long-lived states (reviewed in [40, 41]). Therefore, we reasoned that the approach could be useful for studying global dynamics of multistable GRNs. The method identifies the slowest system processes based on the dominant eigenvalues and eigenvectors of the stochastic transition matrix, $\mathbf{T}(\tau)$, which gives the probability of the system to transition from every possible initial state to every possible destination state within lagtime τ (with τ having units of k^{-1} and k being the rate of protein degradation). Inspection of the eigenvalue spectrum of $\mathbf{T}(\tau = 5)$ for the MISA network in Fig. 1b reveals four eigenvalues near 1 followed by a gap, indicating four system processes that are slow on this timescale. Decreasing τ to 0.5 reveals a step-structure in the eigenvalue spectrum, suggesting a hierarchy of system timescales. The timescales are related to the eigenvalues according to $t_i = -\tau/\ln|\lambda_i(\tau)|$. The Perron eigenvalue $\lambda_1 = 1$ is associated with the stationary (infinite time) process, and the lifetimes t_2 through t_5 are computed to be {95.6, 49.4, 30.8, 2.6} (in units of k^{-1}). Thus, the first gap in the eigenvalue spectrum arises from a more than ten-fold separation in timescales between t_4 and t_5. The original PCCA method [48] used the sign structure of the eigenvectors to assign cluster memberships. Plotting the left-eigenvectors corresponding to the four dominant eigenvalues in the MISA network is instructive: the stationary landscape is obtained from the first left-eigenvector ($\phi_1 = \boldsymbol{\pi}(\mathbf{x})$), which is positive over all microstates, while the opposite-sign regions in ϕ_2, ϕ_3, ϕ_4 reveal the nature of the slow processes (Fig. 1d). An eigenvector with regions of opposite sign corresponds to an exchange between those two regions (in both directions, since eigenvectors are sign-interchangeable). For example, the slowest process corresponds to exchange between the $a > b$ and $b > a$ regions of state-space, i.e., switching between *B*-gene dominant and *A*-gene-dominant expression states. Eigenvectors ϕ_3 and ϕ_4 show that somewhat faster timescales are associated with exchange in and out of the Lo/Lo and Hi/Hi basins.

The Markov State Model approach identifies multistability in GRNs

Reduced models of the MISA network

The MSM framework utilizes a clustering algorithm known as PCCA+ (see Methods and Additional file 1) to assign every microstate in the system to a macrostate (i.e., a cluster of microstates) based on the slow system processes identified by the eigenvectors and eigenvalues of $\mathbf{T}(\tau)$. Applying the PCCA+ algorithm to the MISA network for the parameter set of Fig. 1 resulted in a mapping from $N = 15,376$ $(31 \times 31 \times 4 \times 4)$ microstates onto $C = 4$ macrostates. The N microstates were first enumerated by accounting for all possible system configurations with $0 \le a \le 30$ and $0 \le b \le 30$. This enumeration assumes a negligible probability for the system to ever exceed 30 copies of either protein, which introduces a small approximation error of $1E-5$ (details in Additional file 1: Figure S1). Because the promoters of each gene can take four possible configurations—that is, two binding sites (for the repressor and activator) that can be either bound or unbound—a total of 16 gene configuration states are possible, giving $N = 15,376$ enumerated microstates. For this parameter set, the highest probability densities within the four macrostates obtained correspond closely to the visible peaks (basins) in the probability (quasipotential) landscape. This can be seen by the ellipsoids in Fig. 2a, which show the highest probability-density regions of each macrostate (according to the stationary probability), projected onto the protein subspace. The average expression levels of proteins in each macrostate indicate the four distinct cell phenotypes (Lo/Lo, Lo/Hi, Hi/Lo, Hi/Hi). The complete microstate-to-macrostate mapping is detailed in Additional file 1: Figure S3 and Table S3. In this parameter regime, since the protein binding and unbinding rates are slow relative to protein production and degradation, the promoter configurations determine the macrostate assignment exactly. That is, the algorithm partitions microstates according to the promoter configuration, rather than the protein copy number. Each of the four macrostates contains microstates from four distinct promoter configurations out of the possible sixteen, along with microstates with all possible protein copy number (a/b) combinations. A representative gene promoter configuration for each macrostate (i.e., the configuration contributing the most probability density to each macrostate) is shown schematically (Fig. 2b).

Parameter-dependence of landscapes and MSMs

To determine whether the MSM approach can robustly identify gene network macrostates, we applied it over a

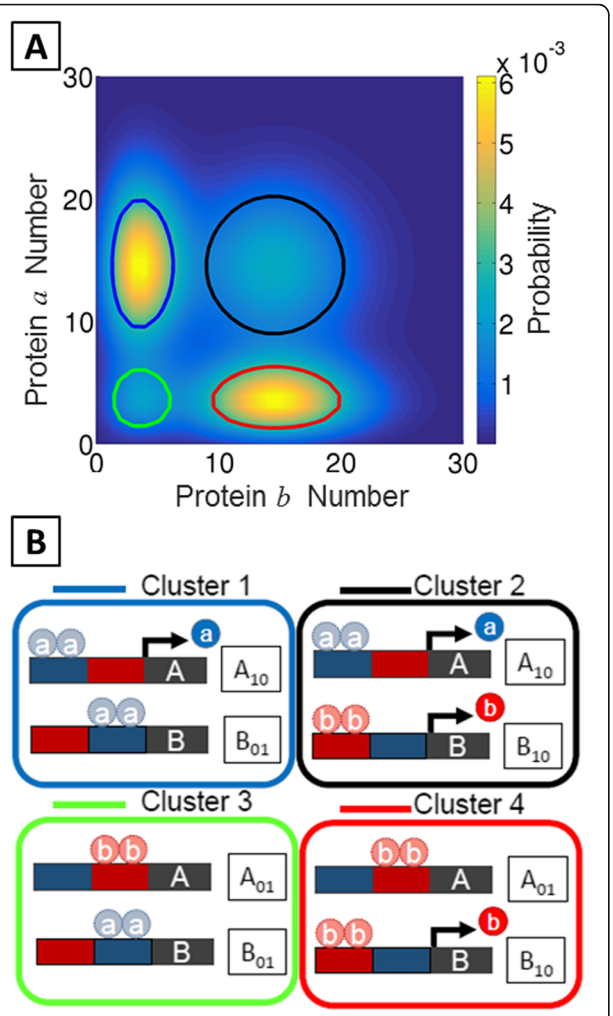

Fig. 2 Four metastable clusters, or network "macrostates", identified for the MISA network by the Markov State Model approach. (Rate parameters same as Fig. 1) **a** Macrostate centers located by their respective 50% probability contours (*ellipsoids*), corresponding to visible peaks in the probability landscape. The locations of the ellipsoids are determined by grouping the most-probable, rank-ordered microstates within each macrostate, until the total probability of grouped microstates is 50% of the total macrostate probability. **b** Schematics of the most probable gene promoter configuration for each metastable cluster

range of network parameters by varying the repressor unbinding rate f_r (all parameters defined in Additional file 1: Table S1). Increasing f_r relative to other network parameters modulates the quasipotential landscape by increasing the probability of the Hi/Hi phenotype, in which both genes express at a high level simultaneously (Fig. 3b). This occurs as a result of weakened repressive interactions, since the lifetimes of repressor occupancy on promoters are shortened when f_r is increased. The eigenvalue spectra show a corresponding shift: when $f_r = 1E-3$, four dominant eigenvalues are present. When f_r is increased to $f_r = 1$, the largest visible gap in the eigenvalue spectrum shifts to occur after the first eigenvalue ($\lambda = 1$), indicating loss of multistability on the timescale of τ (here, $\tau = 5$) (Fig. 3a). Correspondingly, for this parameter set, the landscape shows only a single visible Hi/Hi basin.

The PCCA+ algorithm seeks C long-lived macrostates, where C is user-specified. We constructed Markov State Models for the MISA network over varying f_r, specifying four macrostates. The MSMs are shown graphically in Fig. 3d. The sizes of the circles are proportional to the relative steady-state probability of the macrostate, and the thickness of the directed edges are proportional to the relative transition probability within τ. In agreement with the landscapes, the MSMs over this parameter regime show increasing probability of the Hi/Hi state, as a result of an increasing ratio of transition probability "into" versus "out of" the Hi/Hi state. The locations of the clusters in the state-space (according to 50% (of the total) stationary probability contours) do not change appreciably. The choice of lagtime τ sets the timescale on which metastability is defined in the system. However, in practice, the PCCA+ seeks an assignment of C clusters regardless of whether C metastable states exist in the system on the τ timescale, and the resulting aggregated macrostates are generally invariant to τ. Thus, for $f_r = 1$, the algorithm locates four macrostates, although the (low-probability) Hi/Lo, Lo/Lo, and Lo/Hi macrostates are likely to experience transitions away, into the Hi/Hi macrostate, within τ. These low-probability states appear in the landscape as shoulders on the outskirts of the Hi/Hi basin. Overall, Fig. 3 demonstrates that, for this parameter regime, the quasipotential landscape and the MSM yield similar information on the global system dynamics in terms of the number and locations of long-lived states, and their relative probabilities as a function of the unbinding rate parameter f_r. The MSM further provides quantitative information on the probabilities (and thus timescales) of transitioning between each pair of macrostates.

MSM identifies purely stochastic multistability

Multistability in gene networks is often analyzed within an ordinary differential equation (ODE) framework, by graphical analysis of isoclines and phase portraits, or by linear stability analysis [4, 8]. ODE models of gene networks treat molecular copy numbers (i.e., proteins, mRNAs) as continuous variables and apply a quasi-steady-state approximation to neglect explicit binding/unbinding of proteins to DNA. This approximation is valid in the so-called "adiabatic" limit, where binding and unbinding of regulatory proteins to DNA is fast, relative to protein production and degradation. Previous studies have shown that such ODE models can give rise to landscape structures that are qualitatively different

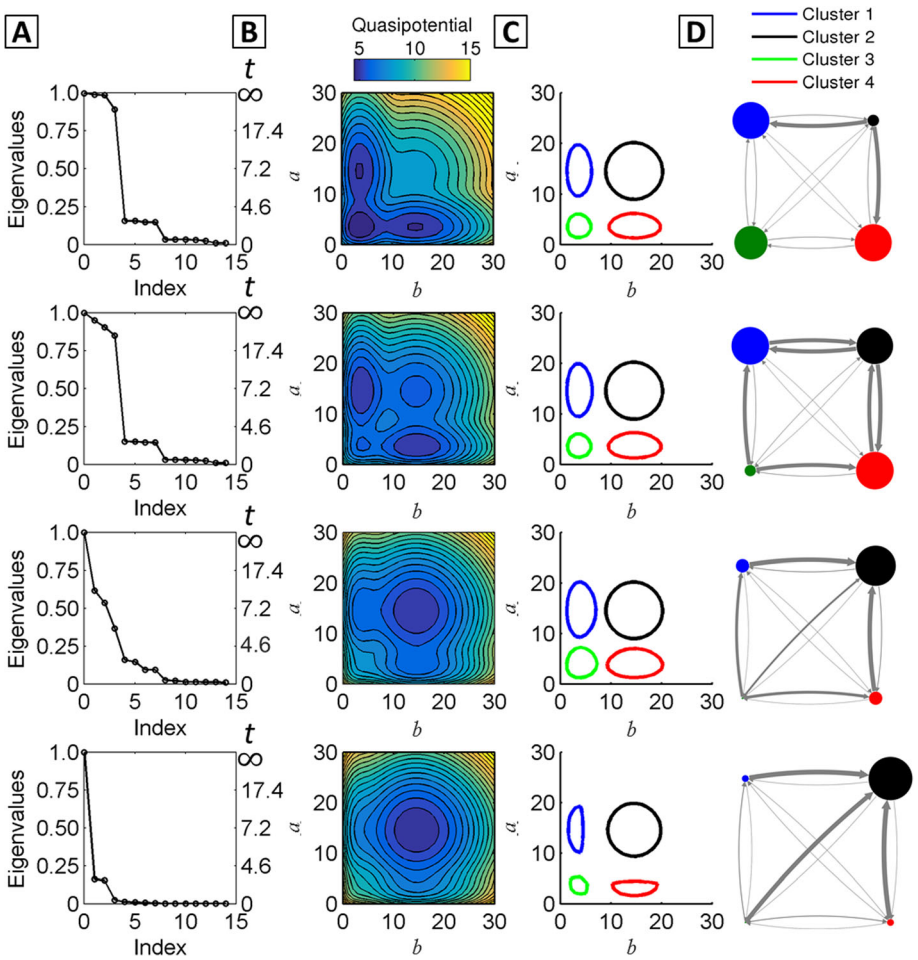

Fig. 3 Dependence of the MISA network eigenvalues, landscape, and MSM on the repressor unbinding parameter f_r . Top to Bottom: increasing $fr = \{1E-3, 1E-2, 1E-1, 1\}$ in units of protein degradation rate, k^{-1} (complete parameter list in Additional file 1: Table S1). **a** The eigenvalue spectrum of $\mathbf{T}(\tau)$ for $\tau = 5$, and associated timescales. **b** The quasipotential landscape. **c** The Markov State Model with four macrostates, visualized by the 50% probability contour for each metastable state. **d** The state transition graph. Nodes and edges denote macrostates and transition probabilities, respectively. The size of each node is proportional to the steady-state probability, and edge thickness is proportional to the probability of transition within $\tau = 5$

from those of their corresponding discrete, stochastic networks. For example, multistability in an ODE model of the genetic toggle switch requires cooperativity—i.e., multimers of proteins must act as regulators of gene expression [59]. However, it was found that monomer repressors are sufficient to give bistability in a stochastic biochemical model [55, 60]. We compared the dynamics of the monomer ETS network (shown schematically in Fig. 4a) as determined by analysis of the ODEs, along with the corresponding stochastic quasipotential landscape and the MSM. In a small-number regime, the ODEs predict monostability (Fig. 4c), while the stochastic landscape shows tristability—that is, three basins corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi expressing phenotypes (Fig. 4d) (The dominant eigenvectors are shown in Additional file 1: Figure S4). This type of discrepancy has been shown to occur in systems with small number effects, i.e., extinction at the boundaries [55] or slow transitions between expression states [29].

The MSM approach identifies three metastable macrostates for the monomer ETS in this parameter regime, as seen in the eigenvalue spectrum, which shows a gap after the third index. The reduced Markov State Model constructed for this network thus reduces the system from $N = 7,803$ $(51 \times 51 \times 3)$ microstates to $C = 3$ macrostates (Fig. 4b), corresponding to the same Hi/Lo, Hi/Hi, and Lo/Hi metastable phenotypes seen in the quasipotential landscape. Figure 4 demonstrates that the MSM approach can accurately identify purely stochastic multistability in systems where continuous models predict only a single stable fixed-point steady state. Similar results were found for a self-regulating, single-gene network (Additional file 1: Figure S5 and Table S4). This network, which has been solved analytically, gives rise to

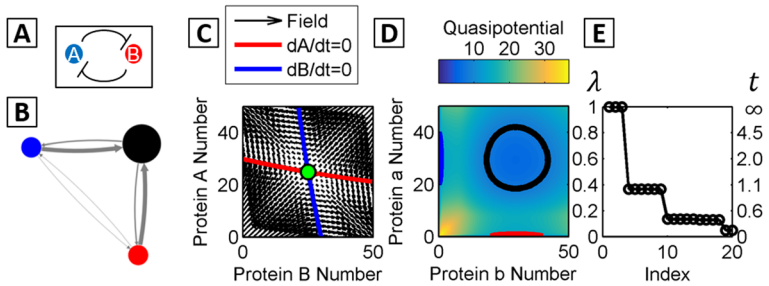

Fig. 4 Comparison of ODE and MSM analysis of the monomer Exclusive Toggle Switch (ETS) network. **a** Schematic of the ETS network motif. **b** The Markov State Model identifies three macrostates corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi phenotypes. Parameter values are listed in Additional file 1: Table S2. **c** The nulllclines and vector field of the deterministic ODEs show a single fixed point steady-state, with both genes expressing at the maximum rate (Hi/Hi phenotype). **b**, **d**, **e** The corresponding landscape and MSM show tristability: **d** The quasipotential landscape shows three visible basins corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi phenotypes. Macrostate centers located by their respective 50% probability contours (ellipsoids), as in Fig. 2. **e** The 20 dominant eigenvalues reveal timescale separation, including a gap after λ_3

a bimodal or monomodal stationary distribution depending on the protein binding/unbinding rates [28, 29, 61].

Analyzing global gene network dynamics with the Markov State Model

MSM provides good approximation to relaxation dynamics from a given initial configuration

Figures 1, 2, 3 and 4 demonstrate the utility of the MSM approach for analyzing stationary properties of networks—that is, for identifying the number and locations of multiple long-lived states. Additionally, the MSM can be used to make dynamic predictions about transitions among macrostates. Dynamics for either the "full" transition matrix (with all system states enumerated up to a maximum protein copy number) or reduced transition matrix (i.e., the MSM) is propagated according to the Chapman-Kolmogorov equation (see Methods and Additional file 1). We sought to determine the accuracy of the dynamic predictions obtained from the MSM. Applying the methods proposed by Prinz, et al. ([47]) (details in Additional file 1), we compared the dynamics propagated by the fully enumerated transition matrix $\mathbf{T}(\tau)$, which is then projected onto the coarse-grained macrostates, to the dynamics of the coarse-grained system propagated by $\tilde{\mathbf{T}}(\tau)$ (i.e., the MSM). We thus computed the error in dynamics of relaxation out of a given initial system configuration. The system relaxation from a given initial microstate can also be computed by running a large number of brute force SSA simulations. Relaxation dynamics for the full, brute-force, and reduced MSM methods, applied to the MISA with $f_r = 1E-2$, all show good agreement (Fig. 5a, b, and c). The error computed between the reduced MSM vs. full dynamics (i.e., $\tilde{\mathbf{T}}(\tau)$ vs $\mathbf{T}(\tau)$), is maximally $7.8E-3$, varies over short times, and decreases continuously after time $t = 140$. Alternatively, the error of the MSM can be quantified by comparing the autocorrelation functions of the MSM and brute force simulation [50, 62]. In Additional file 1: Figure S6, we show that the derived autocorrelation functions of the MSM and brute force, and the relaxation constants τ_r, which describes the amount of time to reach equilibrium, are close in value ($\tau_r = 1E3$, for the MSM, and $\tau_r = 1.1E3$ for the brute force). Overall, these results demonstrate that the most accurate predictions of the coarse-grained MSM can be obtained on long

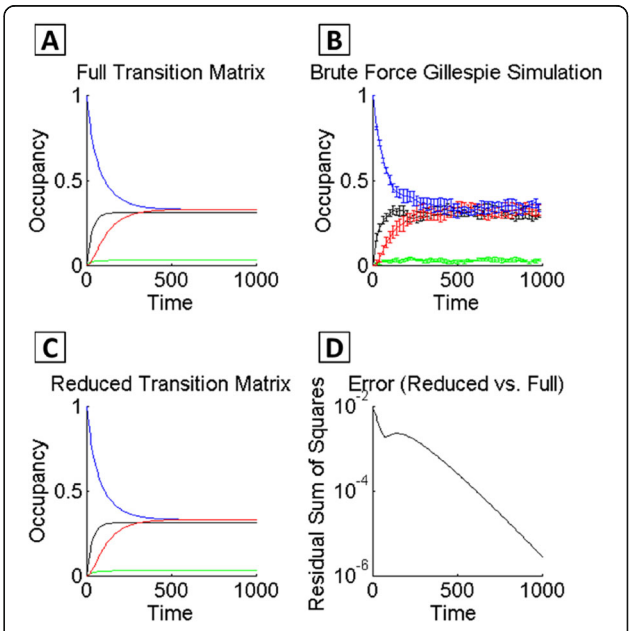

Fig. 5 MSM approximation error for the MISA motif. Relaxation of the system from a particular initial configuration (see text), as obtained from **a** the full transition matrix, **b** brute force SSA simulation, and **c** the reduced transition matrix obtained from the MSM. Color-coding is according to the macrostates, as in Figs. 1, 2 and 3: *blue*, *black*, *red*, *green* correspond to A/B expression phenotypes Hi/Lo, Hi/Hi, Lo/Hi, and Lo/Lo, respectively. **d** Calculated approximation error as a function of time, comparing the reduced MSM to the full CME dynamics. Network parameter values are same as Figs. 1 and 2

timescales, but dynamic approximations with reasonable accuracy can also be obtained for short timescales.

Parameter-dependence of MSM error

The accuracy of the MSM dynamic predictions depends on whether inter-macrostate transitions can be treated as memory-less hops. Previous theoretical studies of gene network dynamics found that the height of the barrier separating phenotypic states, and the state-switching time associated with overcoming the barrier, depends on the rate parameters governing DNA-binding by the protein regulators [5, 6, 55, 63]. We reasoned that a larger timescale separation between intra- and inter-basin transitions (corresponding to a larger barrier height separating basins) should result in higher accuracy of the MSM approximation. Thus, we hypothesized that the accuracy of the MSM dynamic predictions should depend on the DNA-binding and unbinding rate parameters. We demonstrated this using the dimeric ETS motif, by computing the error of the MSM approximation for a range of repressor unbinding rates f. We varied the binding kinetics without changing the overall relative strength of repression, by varying f together with the repressor binding rate h, to maintain a constant binding equilibrium $\left(X_{eq} = \frac{f}{h} = 100\right)$. By varying f and h in this way over eight orders of magnitude, we found that the barrier height and timescale of the slowest system process (t_2) had a non-monotonic dependence on the binding/unbinding parameters. Thus, the fastest inter-phenotype switching was observed in the regime with intermediate binding kinetics, in agreement with previous work [5]. The system also exhibits a shift from three visible basins in the quasipotential landscape in the small f regime to two basins in the large f regime. We performed clustering by selecting $C = 2$ (dashed lines, Fig. 6) and $C = 3$ clusters (solid lines, Fig. 6), and computed the total error over all choices of system initialization, as well as the error associated with relaxation from a particular system microstate. In general, we find that the 3-state MSM approximation is more accurate than the 2-state partitioning. The 3-state MSM dynamic predictions are highly accurate when the DNA-binding/unbinding kinetics is slow. As such, in this regime the Markovian assumption of memory-less transitions between the three phenotypic states is most accurate. As hypothesized, the accuracy of the MSM approximation is lowest (highest error) when the lifetime t_2 is shortest (intermediate regime, $f = 1$), and the error decreases modestly with further increase in f (i.e., increase in t_2).

Decomposition of state-transition pathways in gene networks using the MSM framework

Quantitative models of gene network dynamics can shed light on transition paths connecting phenotypic states. The MSM approach coupled with transition path theory [52, 53, 64] enables decomposition of all major pathways linking initial and final macrostates of interest. This type of pathway decomposition has previously shed light on mechanisms of protein folding [54]. We demonstrate this pathway decomposition on the MISA network, by computing the transition paths linking the polarized A-dominant (Hi/Lo) and B-dominant (Lo/Hi) phenotypes. Multiple alternative pathways linking these phenotypes are possible: for the 4-state coarse-graining, the system can alternatively transit through the Hi/Hi or Lo/Lo phenotypes when undergoing a stochastic state-transition from one polarized phenotype to the other. Not all possible paths are enumerated since only transitions with net positive fluxes are considered (see Additional file 1: Equation S18). The hierarchy of pathway probabilities for successful transitions depends on the kinetic rate parameters (Fig. 7a). It could be tempting to intuit pathway intermediates based on visible basins in the quasipotential landscape. However, we found that the steady-state probability of an intermediate macrostate (i.e., the Hi/Hi or Lo/Lo states) does not accurately predict if it serves as a pathway intermediate for successful transitions, because parameter regimes are possible in which successful transitions are likely to transition through intermediates with high potential/low probability (Fig. 7c). This occurs because the relative probability of transiting through one intermediate macrostate versus another is based on the balance of probabilities for entering and exiting the intermediate: intermediate states that can be easily reached—but not easily exited—as a result of stochastic fluctuations can act as "trap" states. Therefore, it is shown that the pathway probability cannot be inferred from the steady state probability of the intermediates alone.

MSMs can be constructed with different resolutions of coarse-graining

The eigenvalue spectrum of the MISA network shows a step-structure, with nearly constant eigenvalue clusters separated by gaps. These multiple spectral gaps suggest a hierarchy of dynamical processes on separate timescales. A convenient feature of the MSM framework is that it can build coarse-grained models with different levels of resolution by PCCA+, in order to explore such hierarchical processes. We applied the MSM framework to a MISA network with very slow rates of DNA-binding and unbinding ($f_r = 1E-4$, $h_r = 1E-6$), comparing the macrostates obtained from selecting $C = 4$ versus $C = 16$ clusters. For $\mathbf{T}(\tau = 1)$, a prominent gap occurs in the eigenvalue spectrum between λ_{16} and λ_{17}, corresponding to an almost 30-fold separation of timescales between $t_{16} = 27.8$ and $t_{17} = 0.99$ (Fig. 8a). Applying PCCA+ with $C = 16$ clusters uncovered a 16-macrostate network with four highly-interconnected subnetworks consisting of

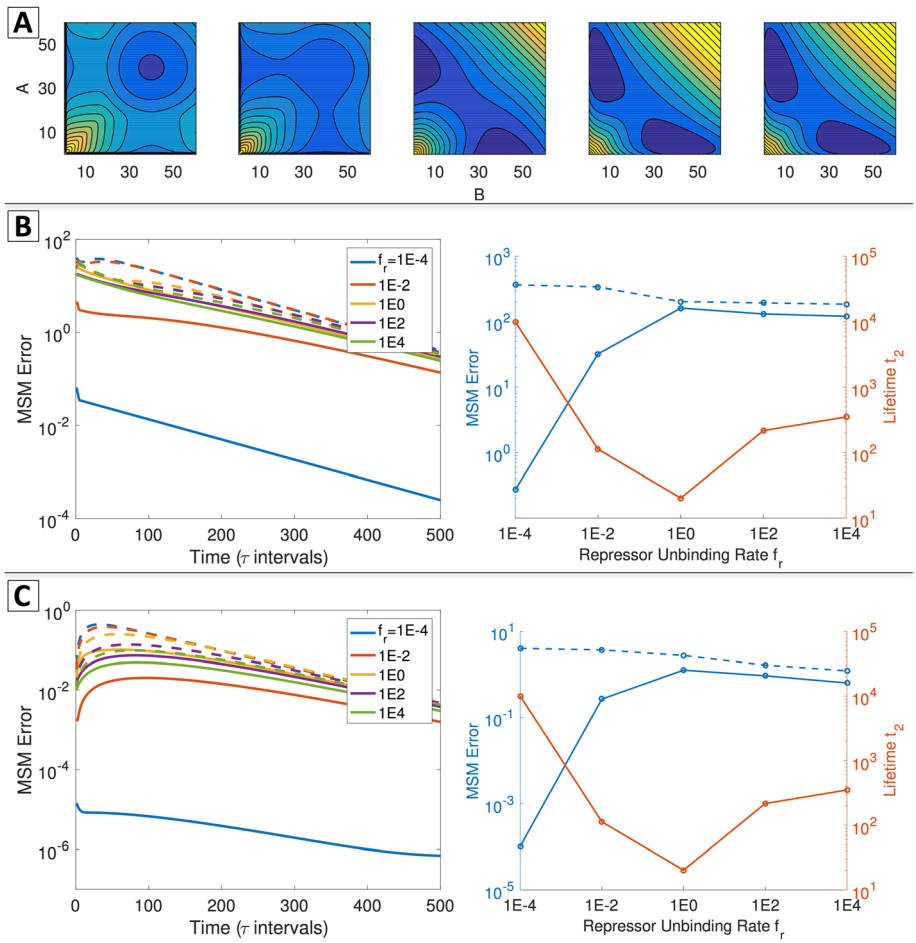

Fig. 6 The MSM approximation accuracy for the ETS motif depends on rate parameters and number of macrostates in the reduced model. **a** Quasipotential landscape for the exclusive dimeric repressor toggle switch, with increasing DNA-binding rates (left to right: $fr = \{1E-4, 1E-2, 1E0, 1E2, 1E4\}$, all parameter values listed in Additional file 1: Table S2), demonstrating the dependence of basin number and barrier height on network parameters. **b** Global error of the MSM approximation. *Left*: Global error as a function of time (in intervals of τ) for different f_r and numbers of macrostates. *Solid lines*: global error of the 3-state MSM. Dashed lines: global error of the 2-state MSM. *Right*: Total global error over kτ, $k = 0$ to 500, for a 3-state (*solid blue*) or 2-state (*dashed blue*) MSM. *Solid orange line*: the longest system lifetime t_2. **c** Error of the MSM approximation when the system is initialized in a particular microstate. *Left*: Error as a function of time (in intervals of τ) for different adiabaticities and different numbers of macrostates. *Solid lines*: error of the 3-state MSM. *Dashed lines*: error of the 2-state MSM. *Right*: Total error from a particular microstate over kτ where $k = 0$ to 500, for a 3-state (*solid blue*) or 2-state (*dashed blue*) MSM. *Orange line*: the longest system lifetime t_2

four states each (Fig. 8c). The identities of the sixteen macrostates showed an exact correspondence to the sixteen possible *A/B* promoter binding configurations. This correspondence reflects the fact that, in the slow binding/unbinding, so-called non-adiabatic regime [65], the slow network dynamics are completely determined by unbinding and binding events that take the system from one promoter configuration macrostate to another, while all fluctuations in protein copy number occur on much faster timescales.

Each subnetwork in the MSM constructed with $C = 16$ corresponds to a single macrostate in the MSM constructed with $C = 4$. Thus, in the $C = 4$ MSM, four different promoter configurations are lumped together in a single macrostate, and dynamics of transitions among them is neglected. Counterintuitively, the locations of the $C = 4$ macrostates do not correspond directly to the four basins visible in the quasipotential landscape (Fig. 8b, d). Instead, the clusters combine distinct phenotypes—e.g., the red macrostate combines the *A/B* Lo/Lo and Lo/Hi phenotypes, because it includes the promoter configurations $A_{01}\ B_{10}$ and $A_{11}\ B_{10}$ (corresponding to Lo/Hi expression) and $A_{01}\ B_{00}$ and $A_{11}\ B_{00}$ (corresponding to Lo/Lo expression) (Fig. 8b, Additional file 1: Table S5 and Figure S7). This result demonstrates that the barriers visible in the quasipotential landscape do not reflect the slowest timescales in the system. This occurs because of the loss of information inherent to visualizing global dynamics via the quasipotential landscape, which often projects

A

	Pathway Probability (%)				Stationary Probability (%)			
f_r	🔵→🔴	🔵→🟢→🔴	🔵→⚫→🔴	🔵→⚫→🟢→🔴	🔵	⚫	🟢	🔴
5E-4	1.0	61.8	36.9	2.3E-01	27.4	1.3	44.0	27.4
1E-3	1.5	44.7	53.5	3.3E-01	34.5	3.2	27.8	34.5
5E-3	2.3E-01	13.8	83.4	5.0E-01	37.9	17.9	6.3	37.9
1E-2	2.4E-01	7.3	89.7	5.2E-01	33.0	31.1	2.9	33.0
1.5E-2	2.5E-01	4.9	92.1	5.2E-01	28.8	40.7	1.7	28.8

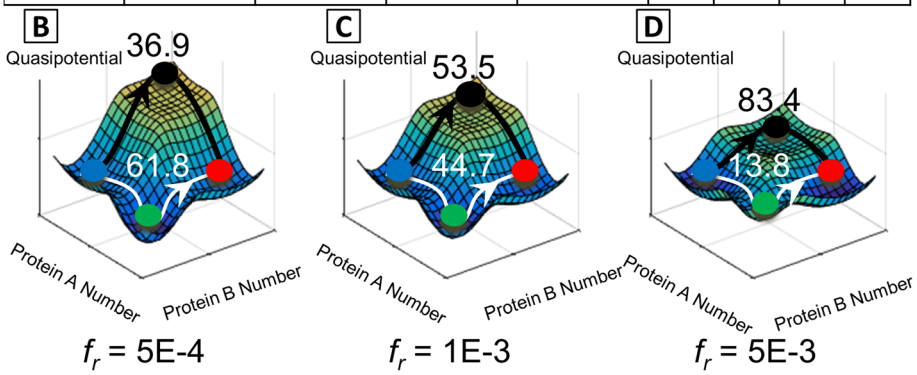

Fig. 7 Dependence of stochastic transition paths on the repressor unbinding rate parameter f_r in the MISA network (parameter values listed in Additional file 1: Table S1). **a** Table of all possible transition paths starting from the Hi/Lo (blue) and ending in the Lo/Hi (red) macrostate (color coding is same as Figs. 1, 2, 3 and 5). Relative probabilities of traversing a given path are shown, along with the stationary probabilities of the system to be found in a given macrostate. **b-d** Dominant transition paths superimposed on the 3D quasipotential surfaces for $fr = \{5E-4, 1E-3, 5E-3\}$, demonstrating how dominant paths can traverse high-potential areas of the landscape. For example, when $f_r = 1E-3$, (panel **c**), successful transitions most likely go through the Hi/Hi state (3.2% populated at steady state), though this requires a large barrier crossing. Pathway percentages are superimposed on the landscapes

dynamics onto two system coordinates. In this case, projecting onto the protein *a* and protein *b* copy numbers loses information about the sixteen promoter configurations, obscuring the fact that barrier-crossing transitions can occur faster than some within-basin transitions. Plotting a time trajectory of brute force SSA simulations for this network supports the findings from the MSM: the dynamics shows frequent transitions within subnetworks, and less-frequent transitions between subnetworks, indicating the same hierarchy of system dynamics as was revealed by the 4- and 16-state MSMs (Fig. 8e).

Transition path decomposition reveals nonequilibrium dynamics

Mapping the most probable paths forward and backward between macrostate "1" (promoter configuration: $A_{01}B_{00}$) and macrostate "11" (promoter configuration: $A_{00}B_{01}$) revealed that a number of alternative transition paths are accessible to the network, and the paths typically transit between three and five intermediate macrostates. The decomposition shows three paths with significant (i.e., >15%) probability and 12 distinct paths with >1% probability (for both forward and backward transitions, Additional file 1: Tables S3-S4). The pathway decomposition also reveals a great deal of irreversibility in the forward and reverse transition paths, which is a hallmark of nonequilibrium dynamical systems. For example, the most probable forward and reverse paths both transit three intermediates, but have only one intermediate (macrostate 5) in common (Fig. 8c and Additional file 1: Tables S6-S7). Thus, the complete process of transitioning away from macrostate 1, through macrostate 11, and returning to 1 maps a dynamic cycle.

Discussion

Our application of the MSM method to representative GRN motifs yielded dynamic insights with potential biological significance. Decomposition of transition pathways revealed that stochastic state-transitions between phenotypic states can occur via multiple alternative routes. Preference of the network to transition with higher likelihood through one particular pathway depended on the stability of intermediate macrostates, in a manner not directly intuitive from the steady-state probability landscape. The existence of "spurious attractors", or metastable intermediates that act as trap states to hinder stem cell reprogramming, has been discussed previously [11] as a general explanation for the existence of partially reprogrammed cells. By analogy, MSMs constructed in protein folding studies predict an ensemble of folding pathways, as well as the existence of misfolded trap states that reduce folding speed [54]. Our results

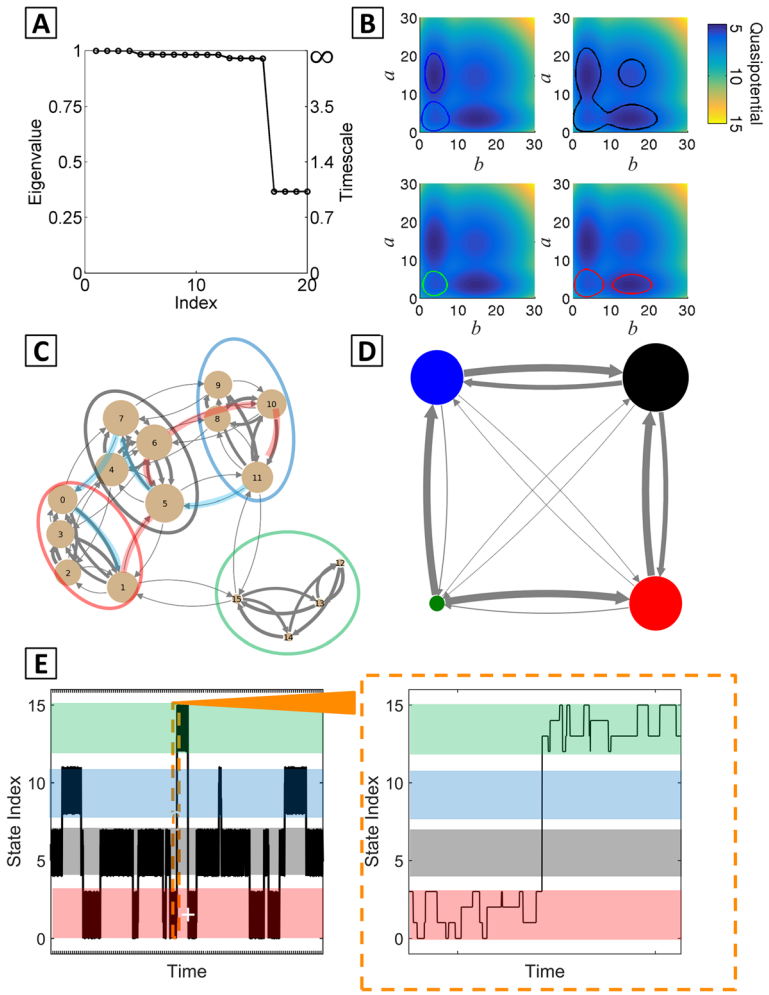

Fig. 8 Hierarchical dynamics revealed by MSM analysis of the MISA network in the slow DNA-binding/unbinding parameter regime. All network parameters listed in Additional file 1: Table S1. **a** Eigenvalue spectrum of $\mathbf{T}(\tau)$, $\tau = 1$, showing 16 dominant eigenvalues. **b** 4-macrostate MSM: 70% probability contours superimposed onto the quasipotential surface. In this parameter regime, separate attractors in the landscape are kinetically linked in the same subnetwork (see text). **c** 16-macrostate MSM showing 4 highly connected subnetworks (*colored ovals*). Each macrostate corresponds to a particular promoter binding-configuration (see numbering scheme in Additional file 1: Table S5). A pair of representative transition paths through the network are highlighted. *Red path*: most probable forward transition path from macrostate 1 to macrostate 11. *Blue path*: most probable reverse path from 11 to 1. **d** State transition graph for the 4-macrostate MSM. **e** Brute force SSA simulation of the MISA network over time. Trajectory is plotted according to the 16-macrostate (promoter configuration) indexing as in panel C and Additional file 1: Table S5. Colored panels reflect the four subnetworks/C = 4 macrostates. *Orange inset*: zoomed in trajectory segment, showing a switching event between the *red* and *green* subnetworks

suggest that multiple partially reprogrammed cell types could be accessible from a single initial cell state. Successful phenotype-transitions can occur predominantly through high-potential (unstable)—and thus difficult to observe experimentally—intermediate cell types. In future applications to specific gene GRNs, the MSM approach could predict a complex map of cell-reprogramming pathways, and thus potentially suggest combinations of targets towards improved safety and efficiency of reprogramming protocols. In synthetic biology applications, the method could be potentially used to optimize biochemical parameters in the design of synthetic gene circuits. For example, it may be desirable to realize synthetic switches with a very crisp on/off macrostate partitioning (i.e., lacking spurious intermediate states) to give a highly digital response.

Our study revealed that the two-gene MISA network can exhibit complex dynamic phenomena, involving a large number of metastable macrostates (up to 16), cycles and hierarchical dynamics, which can be conveniently visualized using the MSM. The quasipotential landscape has been used recently as a means of visualizing global dynamics and assessing locations and relative stabilities of phenotypic states of interest, in a manner

that is quantitative (deriving strictly from underlying gene regulatory interactions), rather than qualitative or metaphorical (as was the case for the original Waddington epigenetic landscape) [21]. However, our study highlights the potential difficulty of interpreting global network dynamics based solely on the steady-state landscape, which is often projected onto one or two degrees of freedom. We found that phenotypically identical cell states—that is, network states marked by identical patterns of protein expression, inhabiting the same position in the projected landscape—can be separated by kinetic barriers, experiencing slow inter-conversion due to slow timescales for update to the epigenetic state (or promoter binding occupancy). Conversely, phenotypically distinct states marked by different levels of protein expression can be kinetically linked, experiencing relatively rapid inter-conversion. This type of stochastic inter-conversion is thought to occur in embryonic stem cells—for example, fluctuations in expression of the Nanog gene have been proposed to play a role in maintaining pluripotency [66, 67]. The hierarchical dynamics revealed by our study supports the idea that the phenotype of a cell could be more appropriately defined by dynamic patterns of regulator or marker expression levels [67], rather than on single-timepoint levels alone. This was seen in the 16-state MSM for the MISA network, where a given expression pattern (e.g., the Lo/Lo peak) comprised multiple macrostates from separate dynamic subnetworks.

Complex, high-dimensional dynamical systems call for systematic methods of coarse-graining (or dimensionality reduction), for analysis of mechanisms and extraction of information that can be compared with experimental results. In the field of Molecular Dynamics, the complexity of, e.g., macromolecular conformational changes—involving thousands of atomic degrees of freedom and multiple dynamic intermediates—has driven the development of automated methods for prediction and analysis of essential system dynamics from simulations [68, 69]. In that field, coarse-graining has been achieved based on a variety of so-called geometric (structural) or, alternatively, kinetic clustering methods [70, 71]. Noe, et al. [71], discussed that geometric (or structure-based) coarse-graining methods can fail to produce an accurate description of system dynamics when structurally similar molecular conformations are separated by large energy barriers or, conversely, when dissimilar structures are connected by fast transitions, as they found in a study of polypeptide folding dynamics. In such cases, kinetic (i.e., separation-of-timescale-based) coarse-graining methods such as the MSM approach are more appropriate. Our application of the MSMs to GRNs demonstrates how similar complex dynamic phenomena can manifest at the "network"-scale.

The challenge of solving the CME due to the curse-of-dimensionality is well known. The MSM approach is related to other projection-based model reduction methods that aim to reduce the computational burden of solving the CME directly by projecting the rate (or transition) matrix onto a smaller subspace or aggregated state-space with fewer degrees of freedom. Such approaches include the Finite State Projection algorithm [31], and methods based on Krylov subspaces [33, 72, 73], sparse-gridding [74], and separation-of-timescales [34, 74, 75] (related timescale-separation-based reduction methods have also been developed to analyze complex ODE models of biochemical networks, e.g., [76, 77]). The MSM is distinct from other timescale-based model reductions in that, rather than partitioning the system into categories of slow versus fast reactions [78] or species [34], or basing categories on physical intuition [75], it systematically groups microstates in such a way that maximizes metastability of aggregated states [40]. The practical benefit of this approach is its capacity to describe a system compactly in terms of long-lived, perhaps experimentally observable, states. Another important distinction between the MSM approach and other CME model reduction methods is that its primary end-goal is *not* to solve the CME per se. Rather, the emphasis in studies employing MSMs has generally been on gaining mechanistic, physical, or experimentally-relevant insights to complex system dynamics [79–81]. As such, the approach does not optimally balance the tradeoff between computational expense versus quantitative accuracy of the solution, as other methods have done explicitly [82]. Instead, the method can be considered to balance the tradeoff between accuracy and "human-interpretability", where decreasing the number of macrostates preserved in the MSM coarse-graining tends to favor the latter over the former.

A potential drawback of the workflow presented in this paper is that it requires an enumeration of the system state-space in order to construct the biochemical rate matrix **K**. Networks of increased complexity or molecular copy numbers will lead to prohibitively large matrix sizes. Here, we restricted our study to model systems with a relatively small number of reachable microstates (i.e., $\sim 10^4$ microstates permitted tractable computations on desktop computers with MATLAB [45]). However, it is important to point out that in typical applications of the MSM framework in Molecular Dynamics, the computational complexity of the coarse-graining procedure is largely decoupled from the full dimensionality of the system state-space, because it is often applied as part of a suite of tools for post-processing atomistic simulation data. An advantage of the MSM approach is its use of the stochastic transition

matrix $\mathbf{T}(\tau)$ (rather than $\mathbf{K}$), which can be estimated from simulations by sampling transition counts between designated regions of state-space in trajectories of length τ [47]. Systems of increased complexity/dimensionality are generally more accessible to simulations, because the size of the state-space is automatically restricted to those states visited within finite-length simulations. Furthermore, in macromolecular systems with high-dimensional configuration spaces, clustering algorithms have been applied in order to obtain a tractable partitioning of state-space, prior to application of the MSM coarse-graining [47]. Typically, a large number of sampled configurations (10^4-10^7) is lumped into a more tractable number of 'microstates' (10^2-10^4), and the MSM framework subsequently identifies ~ tens of metastable macrostates. A recent study of G-protein-coupled receptor activation showcased the high complexity of systems that can be analyzed by MSMs: 250,000 sampled molecular structures were projected to coarse-grained MSMs with either 3000 or 10 states [83]. Based on these previous studies in Molecular Dynamics, we anticipate that the MSM framework will likewise prove useful in analysis of highly complex biochemical networks, particularly when coupled with stochastic simulations and thus bypassing the need for enumerating the CME. In ongoing work (Tse, et al., in preparation), we find that the MSM approach interfaces well with SSA simulations of biochemical network dynamics, combined with enhanced sampling techniques [84–86]. We anticipate that the approach could also potentially interface with other numerical approximation techniques that have been developed in recent years for reduction of the CME.

A potential challenge for the application of the PCCA + −based spectral clustering method to biochemical networks is that, as open systems, biochemical networks generally do not obey detailed balance. This means that the stochastic transition matrices do not have the property of irreversibility, which was originally taken to be a requirement for application of the PCCA algorithm [48]. However, later work by Roblitz et al. [49] found that the PCCA+ method also delivers an optimal clustering for irreversible systems. In this study, we found that the PCCA+ method could determine appropriate clusters in GRNs, and could furthermore uncover non-equilibrium cycles, as seen in the irreversibility (distinct forward and backward) of transition paths in the 16-state system. Newer methods of MSM building, which are specifically designed to treat nonequilibrium dynamical systems, have appeared recently [87]. It may prove fruitful to explore these alternative methods in order to identify the most appropriate, general MSM framework for application to various biochemical networks. On a separate note, another possible area for future study could be the relationship between the MSM framework, specifically its estimation of switching times in multistable networks, to the results from other theoretical approaches to GRNs, such as Large Deviation Theory [88] or Wentzel-Kramers-Brillouin theory [89].

Conclusions

In this work, we present a method for analyzing multistability and global state-switching dynamics in gene networks modeled by stochastic chemical kinetics, using the MSM framework. We found that the approach is able to: (1) identify the number and identities of long-lived phenotypic-states, or network "macrostates", (2) predict the steady-state probabilities of all macrostates along with probabilities of transitioning to other macrostates on a given timescale, and (3) decompose global dynamics into a set of dominant transition pathways and their associated relative probabilities, linking two system states of interest. Because the method is based on the discrete-space, stochastic transition matrix, it correctly identified stochastic multistability where a continuum model failed to find multiple steady states. The quantitative accuracy of the dynamics propagated by the coarse-grained MSM was highest in a parameter regime with slow DNA-binding and unbinding kinetics, indicating that in GRNs the assumption of memory-less hopping among a small number of macrostates is most valid in this regime. By projecting dynamics encompassing a large state-space onto a tractable number of macrostates, the MSMs revealed complex dynamic phenomena in GRNs, including hierarchical dynamics, nonequilibrium cycles, and alternative possible routes for phenotypic state-transitions. The ability to unravel these processes using the MSM framework can shed light on regulatory mechanisms that govern cell phenotype stability, and inform experimental reprogramming strategies. The MSM provides an intuitive representation of complex biological dynamics operating over multiple timescales, which in turn can provide the key to decoding biological mechanisms. Overall, our results demonstrate that the MSM framework—which has been generally applied thus far in the context of molecular dynamics via atomistic simulations—can be a useful tool for visualization and analysis of complex, multistable dynamics in gene networks, and in biochemical reaction networks more generally.

Additional files

Additional file 1: Supporting Information. Description of data: Explains models and algorithms used as well as supporting tables and figures. (PDF 5439 kb)

Additional file 2: Contains the scripts used to produce the figures and tables. (ZIP 76 kb)

Abbreviations

CME: Chemical master equation; ETS: Exclusive toggle switch; GRN: Gene regulatory network; GRNs: Gene regulatory networks; MISA: Mutual inhibition/Self-activation; MSM: Markov state model; ODE: Ordinary Differential Equation; PCCA+: Robust Perron Cluster Analysis; SSA: Stochastic simulation algorithm

Acknowledgements

We thank Jun Allard for helpful discussions.

Funding

We acknowledge financial support from the UC Irvine Henry Samueli School of Engineering.

Availability of data and materials

The datasets supporting the conclusions of this article are included in the Additional file 2.

Authors' contributions

BC and ER designed and performed research. MT contributed to data analysis and manuscript preparation. RS contributed to data analysis. BC and ER wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 27 September 2016 Accepted: 13 January 2017
Published online: 06 February 2017

References

1. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics. 1998;149(4):1633–48.
2. Xiong W, Ferrell JE. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. Nature. 2003;426(6965):460–5.
3. Zhou JX, Huang S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends Genet. 2011;27(2):55–62.
4. Lu M, Jolly MK, Gomoto R, Huang B, Onuchic J, Ben-Jacob E. Tristability in Cancer-Associated MicroRNA-TF Chimera Toggle Switch. J Phys Chem B. 2013;117(42):13164–74.
5. Feng H, Wang J. A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes. Sci Rep. 2012;2:550.
6. Zhang B, Wolynes PG. Stem cell differentiation as a many-body problem. Proc Natl Acad Sci. 2014;111(28):10185–90.
7. Wang P, Song C, Zhang H, Wu Z, Tian X-J, Xing J. Epigenetic state network approach for describing cell phenotypic transitions. Interface Focus. 2014; 4(3):20130068.
8. Hong T, Xing J, Li L, Tyson JJ. A mathematical model for the reciprocal differentiation of T helper 17 cells and induced regulatory T cells. PLoS Comput Biol. 2011;7(7):e1002122.
9. Graf T, Enver T. Forcing cells to change lineages. Nature. 2009;462(7273): 587–94.
10. Huang S. The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? Bioessays. 2012;34(2):149–57.
11. Lang AH, Li H, Collins JJ, Mehta P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput Biol. 2014;10(8):e1003734.
12. Elowitz MB. Stochastic gene expression in a single cell. Science. 2002; 297(5584):1183–6.
13. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. Nat Genet. 2002; 31(1):69–73.
14. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. Cell. 2005;123(6):1025–36.
15. Balaban NQ. Bacterial persistence as a phenotypic switch. Science. 2004; 305(5690):1622–5.
16. Acar M, Mettetal JT, van Oudenaarden A. Stochastic switching as a survival strategy in fluctuating environments. Nat Genet. 2008;40(4):471–5.
17. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, McDermott U, Azizian N, Zou L, Fischbach MA, et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell. 2010;141(1):69–80.
18. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453(7194):544–7.
19. Dietrich JE, Hiiragi T. Stochastic patterning in the mouse pre-implantation embryo. Development. 2007;134(23):4219–31.
20. Yuan L, Chan GC, Beeler D, Janes L, Spokes KC, Dharaneeswaran H, Mojiri A, et al. A role of stochastic phenotype switching in generating mosaic endothelial cell heterogeneity. Nat Commun. 2016;7:10160.
21. Waddington CH. The Strategy of the Genes. London: Allen & Unwin; 1957.
22. Wang J, Zhang K, Xu L, Wang E. Quantifying the Waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci. 2011;108(20):8257–62.
23. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008;9(10):770–80.
24. Kepler TB, Elston TC. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys J. 2001;81(6): 3116–36.
25. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci. 2008;105(45):17256–61.
26. Mackey MC, Tyran-Kamińska M, Yvinec R. Dynamic behavior of stochastic gene expression models in the presence of bursting. SIAM J Appl Math. 2013;73(5):1830–52.
27. Jiao F, Sun Q, Tang M, Yu J, Zheng B. Distribution modes and their corresponding parameter regions in stochastic gene transcription. SIAM J Appl Math. 2015;75(6):2396–420.
28. Schultz D, Onuchic JN, Wolynes PG. Understanding stochastic simulations of the smallest genetic networks. J Chem Phys. 2007;126(24):245102.
29. Ramos AF, Innocentini GCP, Hornos JEM. Exact time-dependent solutions for a self-regulating gene. Phys Rev E. 2011;83(6):062902.
30. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–61.
31. Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys. 2006;124(4):044104.
32. Cao Y, Liang J. Optimal enumeration of state space of finitely buffered stochastic molecular networks and exact computation of steady state landscape probability. BMC Syst Biol. 2008;2(1):30.
33. Wolf V, Goel R, Mateescu M, Henzinger TA. Solving the chemical master equation using sliding windows. BMC Syst Biol. 2010;4(1):42.
34. Pahlajani CD, Atzberger PJ, Khammash M. Stochastic reduction method for biological chemical kinetics using time-scale separation. J Theor Biol. 2011; 272(1):96–112.
35. Sidje RB, Vo HD. Solving the chemical master equation by a fast adaptive finite state projection based on the stochastic simulation algorithm. Math Biosci. 2015;269:10–6.
36. Huang S, Guo YP, May G, Enver T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev Biol. 2007;305(2):695–713.
37. Ma R, Wang J, Hou Z, Liu H. Small-number effects: a third stable state in a genetic bistable toggle switch. Phys Rev Lett. 2012;109(24):248107.
38. Cao Y, Lu H-M, Liang J. Probability landscape of heritable and robust epigenetic state of lysogeny in phage lambda. Proc Natl Acad Sci. 2010; 107(43):18445–50.
39. Munsky B, Fox Z, Neuert G. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. Methods. 2015;85:12–21.
40. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010;52(1):99–105.
41. Chodera JD, Noé F. Markov state models of biomolecular conformational dynamics. Curr Opin Struct Biol. 2014;25:135–44.

42. Bowman GR, Huang X, Pande VS. Network models for molecular kinetics and their initial applications to human health. Cell Res. 2010;20(6):622–30.
43. Sanft KR, Wu S, Roh M, Fu J, Lim RK, Petzold LR. StochKit2: software for discrete stochastic simulation of biochemical systems with events. Bioinformatics. 2011;27(17):2457–8.
44. van Kampen NG. Stochastic processes in physics and chemistry. Amsterdam; Boston; London: Elsevier; 2007.
45. The MathWorks. MATLAB Release. Natick: Massachusetts;2015a.
46. Scherer MK, Trendelkamp-Schroer B, Paul F, Perez-Hernandez G, Hoffmann M, Plattner N, Wehmeyer C, Prinz J, Noé F. PyEMMA 2: a software package for estimation, validation, and analysis of Markov Models. J Chem Theory Comput. 2015;11(11):5525–42.
47. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov models of molecular kinetics: generation and validation. J Chem Phys. 2011;134(17):174105.
48. Deuflhard P, Huisinga W, Fischer A, Schütte C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Linear Algebra Its Appl. 2000;315(1–3):39–59.
49. Röblitz S, Weber M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. Adv Data Anal Classif. 2013;7(2):147–79.
50. Buchete NV, Hummer G. Coarse master equations for peptide folding dynamics †. J Phys Chem B. 2008;112(19):6057–69.
51. Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. J Chem Theory Comput. 2011;7(10):3412–9.
52. W. E and Vanden-Eijnden E. Towards a Theory of Transition Paths. J Stat Phys. 2006;123(3):503–523.
53. Metzner P, Schütte C, Vanden-Eijnden E. Transition path theory for Markov jump processes. Multiscale Model Simul. 2009;7(3):1192–219.
54. Noe F, Schutte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci. 2009;106(45):19011–6.
55. Schultz D, Walczak AM, Onuchic JN, Wolynes PG. Extinction and resurrection in gene networks. Proc Natl Acad Sci. 2008;105(49):19165–70.
56. Morelli MJ, Tănase-Nicola S, Allen RJ, ten Wolde PR. Reaction coordinates for the flipping of genetic switches. Biophys J. 2008;94(9):3413–23.
57. Huang S. Reprogramming cell fates: reconciling rarity with robustness. Bioessays. 2009;31(5):546–60.
58. Huang S. Hybrid T-helper cells: stabilizing the moderate center in a polarized system. PLoS Biol. 2013;11(8):e1001632.
59. Gardmer T, Cantor C, Collins J. Construction of a genetic toggle switch in Escherichia coli. Nature. 2000;403(6767):339–42.
60. Lipshtat A, Loinger A, Balaban NQ, Biham O. Genetic toggle switch without cooperative binding. Phys Rev Lett. 2006;96(18):188101.
61. Hornos JEM, Schultz D, Innocentini GC, Wang JA, Walczak AM, Onuchic JN, Wolynes PG. Self-regulating gene: An exact solution. Phys Rev E. 2005;72(5): 051907.
62. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS. Markov State Model reveals folding and functional dynamics in ultra-long MD trajectories. J Am Chem Soc. 2011;133(45):18413–9.
63. Tse MJ, Chu BK, Roy M, Read EL. DNA-binding kinetics determines the mechanism of noise-induced switching in gene networks. Biophys J. 2015; 109(8):1746–57.
64. Berezhkovskii A, Hummer G, Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. J Chem Phys. 2009; 130(20):205102.
65. Walczak AM, Onuchic JN, Wolynes PG. Absolute rate theories of epigenetic stability. Proc Natl Acad Sci. 2005;102(52):18926–31.
66. Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, Vrana J, Jones K, Grotewold L, Smith A. Nanog safeguards pluripotency and mediates germline development. Nature. 2007;450(7173):1230–4.
67. Kalmar T, Lim C, Hayward P, Muñoz-Descalzo S, Nichols J, Garcia-Ojalvo J, Arias AM. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol. 2009;7(7):e1000149.
68. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J Chem Phys. 2007;126(15):155101.
69. Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys. 2009;131(12):124101.
70. Deuflhard P, Weber M. Robust Perron cluster analysis in conformation dynamics. Linear Algebra Its Appl. 2005;398:161–84.
71. Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G, Noé F. Identification of slow molecular order parameters for Markov model construction. J Chem Phys. 2013;139(1):015102.
72. Burrage K, Hegland M, Macnamara S, Sidje R. A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems, Proceedings of the Markov 150th Anniversary Conference. 2006.
73. Cao Y, Terebus A, Liang J. Accurate chemical master equation solution using multi-finite buffers. Multiscale Model Simul. 2016;14(2):923–63.
74. Hegland M, Burden C, Santoso L, MacNamara S, Booth H. A solver for the stochastic master equation applied to gene regulatory networks. J Comput Appl Math. 2007;205(2):708–24.
75. Peleš S, Munsky B, Khammash M. Reduction and solution of the chemical master equation using time scale separation and finite state projection. J Chem Phys. 2006;125(20):204104.
76. Anna L, Csikász-Nagy A, Gy Zsély I, Zádor J, Turányi T, Novák B. Time scale and dimension analysis of a budding yeast cell cycle model. BMC Bioinformatics. 2006;7:494.
77. Surovtsova I, Simus N, Lorenz T, Konig A, Sahle S, Kummer U. Accessible methods for the dynamic time-scale decomposition of biochemical systems. Bioinformatics. 2009;25(21):2816–23.
78. Haseltine EL, Rawlings JB. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. J Chem Phys. 2002;117(15):6959.
79. Kuroda Y, Suenaga A, Sato Y, Kosuda S, Taiji M. All-atom molecular dynamics analysis of multi-peptide systems reproduces peptide solubility in line with experimental observations. Sci Rep. 2016;6:19479.
80. Jayachandran G, Vishal V, Pande VS. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. J Chem Phys. 2006;124(16):164902.
81. Singhal N, Snow CD, Pande VS. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. J Chem Phys. 2004;121(1):415.
82. Tapia JJ, Faeder JR, Munsky B. Adaptive coarse-graining for transient and quasi-equilibrium analyses of stochastic gene regulation. 2012. p. 5361–6.
83. Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, Altman RB, Pande VS. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. Nat Chem. 2013;6(1):15–21.
84. Bhatt D, Bahar I. An adaptive weighted ensemble procedure for efficient computation of free energies and first passage rates. J Chem Phys. 2012; 137(10):104101.
85. Zhang BW, Jasnow D, Zuckerman DM. The 'weighted ensemble' path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. J Chem Phys. 2010;132(5):054107.
86. Adelman JL, Grabe M. Simulating rare events using a weighted ensemble-based string method. J Chem Phys. 2013;138(4):044105.
87. Marcus W, Fackeldey K. G-pcca: Spectral clustering for non-reversible markov chains. ZIB Rep. 2015;15(35).
88. Lv C, Li X, Li F, Li T. Constructing the energy landscape for genetic switching system driven by intrinsic noise. PLoS One. 2014;9(2):e88167.
89. Assaf M, Roberts E, Luthey-Schulten Z. Determining the Stability of Genetic Switches: Explicitly Accounting for mRNA Noise. Phys Rev Lett. 2011;106(24):248102.