Supporting information for: Quantifying the Landscape for Development and Cancer from a Core Cancer Stem Cell Circuit

Chunhe Li[1], Jin Wang[1,2,*]
1 Department of Chemistry and Physics, State University of New York at Stony Brook, Stony Brook, NY, USA
2 State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin, China
∗ E-mail: jin.wang.1@stonybrook.edu

Self Consistent Mean Field Approximation

The time evolution the dynamical systems are governed by the diffusion equations. Given the system state $P(X_1, X_2, ..., X_n, t)$, where $X_1, X_2, ..., X_n$ represent the concentration or populations of molecules or species, we expected to have N-coupled differential equations, which are difficult to solve. Following a self consistent mean field approach [1–3], we split the probability into the products of individual ones: $P(X_1, X_2, ..., X_n, t) \sim \prod_i^n P(X_i, t)$ and solve the probability self-consistently. This can effectively reduce the dimensionality from M^N to $M \times N$, and thus make the computation of the problem tractable.

However, for the multi-dimensional system, it is still hard to solve diffusion equations directly. We start from moment equations and simply assume specific probability distribution based on physical argument, i.e. we give some specific connections between moments. In principle, once we know all moments, we can acquire the probability distribution. For instance, the Poisson distribution has only one parameter, so we may calculate all other moments from the first moment, the mean. In this work, we use gaussian distribution as approximation, which means we need two moments, mean and variance.

When the diffusion coefficient D is small, the moment equations can be approximated to [4, 5]:

$$\dot{\bar{\mathbf{x}}}(t) = F[\bar{\mathbf{x}}(\mathbf{t})] \tag{1}$$

$$\dot{\sigma}(t) = \sigma(t)\mathbf{A^T}(\mathbf{t}) + \mathbf{A}(t)\sigma(\mathbf{t}) + 2\mathbf{D}[\bar{\mathbf{x}}(\mathbf{t})]. \tag{2}$$

Here, $\mathbf{x}$, $\sigma(t)$ and $\mathbf{A}(t)$ are vectors and tensors, and $\mathbf{A^T}(t)$ is the transpose of $\mathbf{A}(t)$. The matrix elements of A is $A_{ij} = \frac{\partial F_i[X(t)]}{\partial x_j(t)}$. According to this equation, we can solve $\mathbf{x}(t)$ and $\sigma(t)$. Here, we consider only diagonal elements of $\sigma(t)$ from mean field splitting approximation. Therefore, the evolution of probabilistic distribution for each variable could be acquired using the mean and variance based on gaussian approximation:

$$P(x,t) = \frac{1}{\sqrt{2\pi\sigma(t)}} \exp - \frac{[x - \bar{x}(t)]^2}{2\sigma(t)} \tag{3}$$

The probability obtained above corresponds to one fixed point or basin of attraction. If the system has multistability, then there are several probabilistic distributions localized at every basin of attraction, with different variations. Therefore, the total probability is the weighted sum of all these probability distributions. The weighting factors (w_1, w_2) are the size of the basin, representing the relative size of different basins of attractions. For example, for a bistable system, the probability distribution takes the form: $P(x,t) = w_1 P^a(x) + w_2 P^b(x)$, here $w_1 + w_2 = 1$. Here, we determine the weights w_i by giving a large number of random initial conditions for ODEs to find solution, and then collect the statistics for different solution. For example, for a bistable system, if 10% initial condition goes to the first steady state, and 90% initial condition goes to the second steady state, then the weight w_1 for the first basin is 0.1 and w_2 for the second basin is 0.9. By giving large number of random different initial conditions for ODEs, we can solve the equations at a fixed parameter set. By collecting the statistics of the solution,

we can determine if the system is monostable or bistable or multi-stable at current parameter region. In this work, the solution of 32 ODEs giving multiple (100000) different random initial conditions produce the multi-stability (bistable or tri-stable). For tri-stability, the probability distribution takes the form: $P(x,t) = w_1P^a(x) + w_2P^b(x) + w_3P^c(x)$, $w_1 + w_2 + w_3 = 1$. For 32 dimensional system, we can acquire 32 dimensional probability distribution. To exhibit the results in a 2-dimensional space, we integrated out the other 30 variables and left two variables AKT and RB.

Finally, once we have the total probability, we can construct the potential landscape by the relationship with the steady state probability: $U(x) = -lnP_{ss}(x)$. In the gene regulatory network system, every parameter or link contributes to the structure and dynamics of the system, which is encoded in the total probability distribution, or the underlying potential landscape.

For nonequilibrium gene regulatory systems, the driving force F can not be written as the gradient of potential U, like the equilibrium case. In general, F can be decomposed into a gradient of the potential and a curl flux force linking the steady state flux $\mathbf{J_{ss}}$ and the steady state probability P_{ss} [2,6]($\mathbf{F} = +\mathbf{D}/P_{ss} \cdot \frac{\partial}{\partial \mathbf{x}} P_{ss} + \mathbf{J_{ss}}(\mathbf{x})/P_{ss} = -D\frac{\partial}{\partial \mathbf{x}} U + \mathbf{J_{ss}}(\mathbf{x})/P_{ss}$). P_{ss} denotes steady state probability and potential U is defined as $U = -lnP_{ss}$. The probability flux vector $\mathbf{J}$ of the system in concentration or gene expression level space $\mathbf{x}$ is defined as [4]: $\mathbf{J}(\mathbf{x},t) = \mathbf{F}P - \mathbf{D} \cdot \frac{\partial}{\partial \mathbf{x}} P$.

Kinetic Path from Path Integral

In the cells, there exist intrinsic noise from statistical fluctuations of the finite number of molecules, and external noise from highly dynamical and inhomogeneous environments. Both of them can be significant to the dynamics of the system [7, 8]. Therefore, one needs to study the cellular network dynamics in fluctuating conditions in order to model the cellular inner and outer environments realistically. A network of chemical reactions in fluctuating environments can be addressed by:$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) + \zeta$, where $\mathbf{x} = (x_1(t), x_2(t), ..., x_6(t))$ represents the vector of protein concentrations or gene expression levels. $\mathbf{F}(\mathbf{x})$ is the vector for the driving force of chemical reactions. ζ is Gaussian noise term whose autocorrelation function is $< \zeta_i(\mathbf{x},t)\zeta_j(\mathbf{x},0) >= 2D\delta(t)$, and D is diffusion coefficient matrix.

The dynamics for the probability of starting from an initial configuration $\mathbf{x}_{initial}$ at t=0 and ending at an final configuration $\mathbf{x}_{final}$ at time t, in terms of the Onsager-Machlup functional, can be formulated [9,10] as: $P(\mathbf{x}_{final}, t, \mathbf{x}_{initial}, 0) = \int \mathbf{Dx}\ \exp[-\int dt(\frac{1}{2}\nabla \cdot \mathbf{F}(\mathbf{x}) + \frac{1}{4}(d\mathbf{x}/dt - \mathbf{F}(\mathbf{x})) \cdot \frac{1}{\mathbf{D}(\mathbf{x})} \cdot (d\mathbf{x}/dt - \mathbf{F}(\mathbf{x})))] = \int \mathbf{Dx}\ \exp[-S(\mathbf{x})] = \int \mathbf{Dx}\ \exp[-\int L(\mathbf{x}(t))dt]$. Here, $\mathbf{D}(\mathbf{x})$ is the diffusion coefficient matrix. The integral over $\mathbf{Dx}$ denotes the sum over all possible paths from the state $\mathbf{x_{initial}}$ at time $t = 0$ to the state $\mathbf{x_{final}}$ at time t. The exponent factor gives the weight of each path. Thus, the probability of network dynamics from the initial state $\mathbf{x_{initial}}$ to the final state $\mathbf{x_{final}}$ is equal to the sum of all possible paths with different weights. $S(\mathbf{x})$ is the transition action and $L(\mathbf{x}(\mathbf{t}))$ is the Lagrangian or the weight for each path.

The path integrals can be approximated with a set of dominant paths, since each path is exponentially weighted, and the other subleading path contributions are often small and can be neglected. So, the dominant path with the optimal weight can be acquired through minimizing the transition action S or Lagrangian. In our case, we identify the dominant paths as the biological paths.

Hamilton-Jacobian (HJ) Framework for Path Integral.

From our path integral formalism, we can evaluate the weights of the kinetic paths. The most probable trajectory can be acquired when the action $S(x)$ is minimized directly. The Lagrangian is written as:

$$L(\mathbf{x}) = \frac{1}{4D}\dot{\mathbf{x}}^2 + V(\mathbf{x}) - \frac{1}{2D}\mathbf{F}(\mathbf{x}) \cdot \dot{\mathbf{x}} \tag{4}$$

and thus the generalized momentum can be written out as: $\mathbf{P}(\mathbf{x}) = \frac{\partial L}{\partial x} = \frac{1}{2D}(\dot{\mathbf{x}} - \mathbf{F}(\mathbf{x}))$. In the kinetic system, the Hamiltonian of the system has the form:

$$H(\mathbf{x}) = -L(\mathbf{x}) + \mathbf{P}(\mathbf{x}) \cdot \dot{\mathbf{x}} = E_{eff} \tag{5}$$

According to the above equation, we can obtain $\frac{1}{4D}\dot{\mathbf{x}}^2 - V(\mathbf{x}) = E_{eff}$ and $|\dot{\mathbf{x}}| = \sqrt{4D(E_{eff} + V(\mathbf{x}))}$. After substituting Eq. S2 into the action, we can obtain $S(\mathbf{x}) = \int(\mathbf{P}(\mathbf{x}) \cdot \dot{\mathbf{x}} - H(\mathbf{x}))dt$. We can see that the action characterizing the weights of the paths depends on the values of the Hamiltonian. Specific values of the Hamiltonian correspond to specific values of the final time T. For a fixed Hamiltonian, a corresponding optimal path exits when minimizing the action $S(\mathbf{x})$.

From the least action principle, if the Hamiltonian of the system is constant, the variation of the action, for given initial and final coordinates and initial and final time, is zero. Allowing a variation of the final time T and leaving the initial and the final coordinates fixed, we have $\delta S = -H\delta t$. For a constant Hamiltonian, $\delta S = -E\delta t$. We define $S_0 = \int \mathbf{P}(\mathbf{x}) \cdot \dot{\mathbf{x}}dt$, since $S(\mathbf{x}) = \int(\mathbf{P}(\mathbf{x}) \cdot \dot{\mathbf{x}} - H(\mathbf{x}))$. We find $\delta S_0 = 0$. Thus, the action S_0 is minimized with respect to all the paths satisfying the constant Hamiltonian and passing through the final point at any instant.

For multidimensional questions, the action depends not only the initial and final coordinates but also on the initial and final time. In the HJ framework, we can transform the formulations into a different representation in x space: $S_0 = S_{HJ}(\mathbf{x}) = \int \sum_i \frac{1}{2D}(\dot{\mathbf{x}}_\mathbf{i} - \mathbf{F}_\mathbf{i})dx_i = \int \sum_i p_i(\mathbf{x})dx_i$. Here p_i is the associated momentum. Now the action only depends on the initial and final coordinates. This action can be further simplified and is equivalent to a line integral along a particular one dimensional path l so that $S_{HJ}(\mathbf{x}) = \int \sum_i p_i(\mathbf{x})dx_i = \int p_l dl$ where $p_l = \sqrt{(E_{eff} + V(\mathbf{x}))/D} - \frac{1}{2D}F_l$. This switch from the time-dependent to the Hamiltonian-dependent HJ description [9–11]. The dominant path connection given initial and final states is obtained by minimizing the action in the HJ representation $S_{HJ} = \int_{x_i}^{x_f}(\sqrt{(E_{eff} + V(\mathbf{x}))/D} - \frac{1}{2D}F_l)dl$, where dl is an infinitesimal displacement along the path trajectory. E_{eff} is a free parameter that determines the total time elapsed during the transition.

In the current work, for simplification we chose $E_{eff} = -V_{min}(x)$, which is the effective potential by minimizing $V(\mathbf{x})$, and corresponding to the longest kinetic time. Finally, the optimal paths were obtained by minimizing the discrete target function:

$$S_{HJ} = \sum_n^{N-1}(\sqrt{(E_{eff} + V(n))/D} - \frac{1}{2D}F_l(n))\Delta l_{n,n+1} + \lambda P \tag{6}$$

where

$$\begin{aligned}
P &= \sum_i^{N-1}(\Delta l_{i,i+1} - <\Delta l>)^2 \\
(\Delta l)^2_{n,n+1} &= \sum_i(\mathbf{x}_i(n+1) - \mathbf{x}_i(n))^2 \\
F_l(n) &= \sum_i \mathbf{F}_\mathbf{i}(\mathbf{x}(n))(\mathbf{x}_\mathbf{i}(n+1) - \mathbf{x}_i(n))/\Delta l_{n,n+1} \\
V(n) &= \sum_i(\frac{1}{4D}\mathbf{F}^2(\mathbf{x}_\mathbf{i}) + \frac{1}{2}\sum_j \frac{\partial \mathbf{F}_\mathbf{j}(\mathbf{x}_\mathbf{i})}{\partial \mathbf{x}_\mathbf{j}})
\end{aligned} \tag{7}$$

Here, $\Delta l_{n,n+1}$ is the Euclidean measure of the nth elementary path step, and P is a penalty function, which keeps all the length elements close to their average and becomes irrelevant in the continuum limit. The minimization of the discrete HJ effective action was performed by applying a simulated annealing algorithm or the conjugate gradient algorithm. In this study, we chose the discrete steps n as 20, and the diffusion coefficient is chosen as 0.01.

Determination of Parameter Values for CSC Network

About the determination of parameter values, we use the following criteria:

1, We chose parameter values according to some previous work [3, 10, 12] and used Hill cooperative binding expressions to represent the regulations. Here, S represents the threshold of the explicitly sigmoidal functions, i.e. the strength of the regulations, and n is the Hill coefficient determining the steepness of the sigmoidal function. Previous studies explored a two gene system, where the region to produce bi-stability is given as: threshold S=0.5-1.5, Hill coefficient n=4-8. Here, we choose parameter values as: S=0.5, n=4 (tetramer binding). For the degradation k, activation a and repression strength b, we assume they are uniform for different genes for simplicity. This is because so far for cancer stem cell network there is no such information about regulation strength — or the magnitude of activation and repression parameters — which should come from the detailed biochemistry reactions involved in cancer stem cell system.

2. Our purpose here is to explore the switching between stem cell state, cancer stem cell state, normal state and cancer state (4 key stable states). Therefore, we chose those parameters that can satisfy some biological constraints, including steady state emergence and multi-stability.

3. In global sensitivity analysis section, we obtained some results by changing individual regulation strengths from any gene i to gene j (Fig. S2). These results show the relative robustness of current parameter choices.

References

1. Sasai M, Wolynes P (2003) Stochastic gene expression as a many-body problem. Proc Natl Acad Sci USA 100: 2374-2379.
2. Wang J, Li C, Wang EK (2010) Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. Proc Natl Acad Sci USA 107: 8195-8200.
3. Li C, Wang J (2013) Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths. PLoS Comput Biol 9: e1003165.
4. Van Kampen NG (1992) Stochastic Processes in Chemistry and Physics. Amsterdam: North Holland, 1 edition, 120-127 pp.
5. Hu G (1994) Stochastic Forces and Nonlinear Systems. Shanghai: Shanghai Scientific and Technological Education Press, 68-74 pp.
6. Wang J, Xu L, Wang EK (2008) Potential landscape and flux framework of non-equilibrium networks: Robustness, dissipation and coherence of biochemical oscillations. Proc Natl Acad Sci USA 105: 12271-12276.
7. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA 99: 12795-12800.
8. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6: 451-64.
9. Wang J, Zhang K, Wang EK (2010) Kinetic paths, time scale, and underlying landscapes: A path integral framework to study global natures of nonequilibrium systems and networks. J Chem Phys 133: 125103:1-13.
10. Wang J, Zhang K, Xu L, Wang EK (2011) Quantifying the waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci USA 108: 8257-8262.

11. Faccioli P, Sega M, Pederiva F, Orland H (2006) Dominant pathways in protein folding. Phys Rev Lett 97: 1-4.

12. Huang S, Guo Y, May G, Enver T (2007) Bifurcation dynamics of cell fate decision lineage-commitment in bipotent progenitor cells. Dev Biol 305: 695-713.

Table S1. Names of 6 genes in the cancer and developmental network and the corresponding function description

Gene	Functions
P53=1	Tumor suppressor gene
miR200=2	Tumor suppressor micro-RNA
miR145=3	Tumor suppressor micro-RNA
ZEB=4	Oncogene, promoting EMT transition
OCT4=5	Stem cell marker gene
MDM2=6	Oncoegene

Table S2. Activation link names in sensitivity analysis and the corresponding regulations they represent. The order numbers for causal and target genes are shown, which are corresponding to the gene name in Table S1.

Link Name	Causal Genes	Target Genes
A1	1	1
A2	1	2
A3	5	2
A4	1	3
A5	4	4
A6	5	5
A7	1	6

Table S3. Repression link names in the sensitivity analysis and the corresponding regulations they represent. The order numbers for causal and target genes are shown, which are corresponding to the gene name in Table S1.

Link Name	Causal Genes	Target Genes
R1	6	1
R2	4	2
R3	4	3
R4	5	3
R5	2	4
R6	3	4
R7	1	5
R8	3	5
R9	3	6

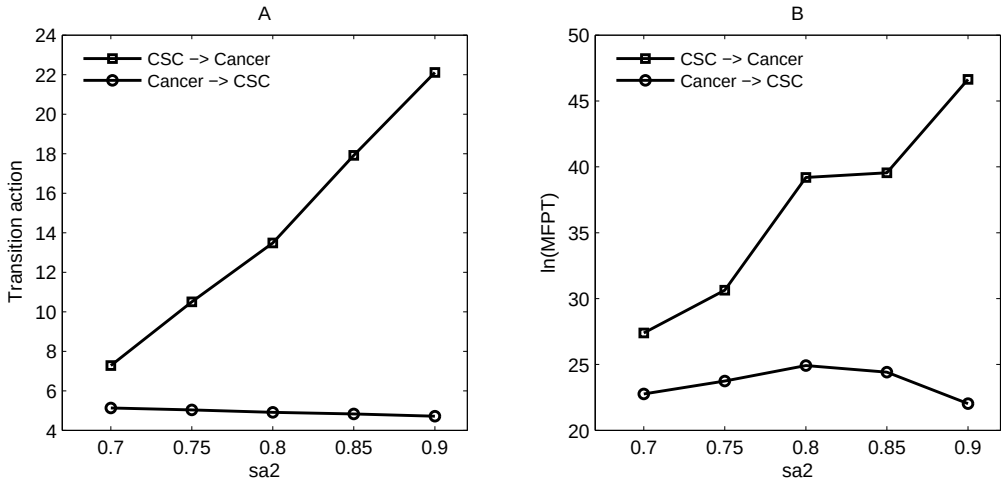

Figure S1. Transition action for the dominant path and the mean first passage time between CSC attractor and cancer attractor. As $sa2$ (self-activation of ZEB/OCT4) increases, for cancer differentiation process (transition from CSC to cancer attractor), both transition action and MFPT increase. This behavior arises because the activation of ZEB makes CSC state more stable, thus leading to longer time for the transition from CSC to cancer. This suggests, unexpectedly, that activating ZEB might provide a way to delay the recurrence process of cancer from CSC.

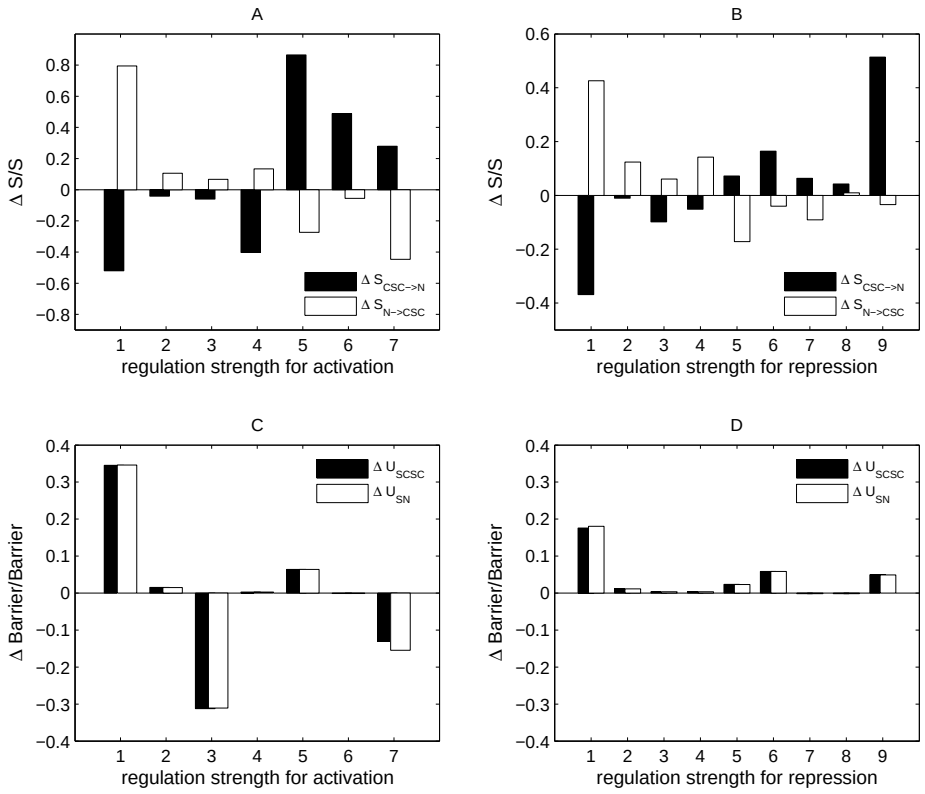

Figure S2. Global sensitivity analysis for regulation strengthes (7 activation links and 9 repression links) in terms of transition action S (A, B) and potential barriers (C, D). Parameters are specified as: $a = 0.5, b = 0.8, sa = sa1 = sa2 = 0.5$. $A1, ..., A7$ represent 7 activation links (Table S2), and $R1, ..., R9$ represent 9 repression links (Table S3). We can see that the key activation regulations include $A1(P53- > P53), A4(P53- > miR145), A5(ZEB- > ZEB), A6(OCT4- > OCT4), A7(P53- > MDM2)$ (Fig. S2A), and the key repression links include $R1(MDM2 - |P53), R6(miR145 - |ZEB), R9(miR145 - |MDM2)$ (Fig. S2B). These results confirm the role of the P53-MDM2 negative feedback loop $(A1, A7, R1, R9)$ on cancer formation, and the role of microRNAs on cancer and development. It has been suggested that MDM2 dysregulation caused by downregulation of miR-143 and miR-145 contributes to epithelial cancer development.

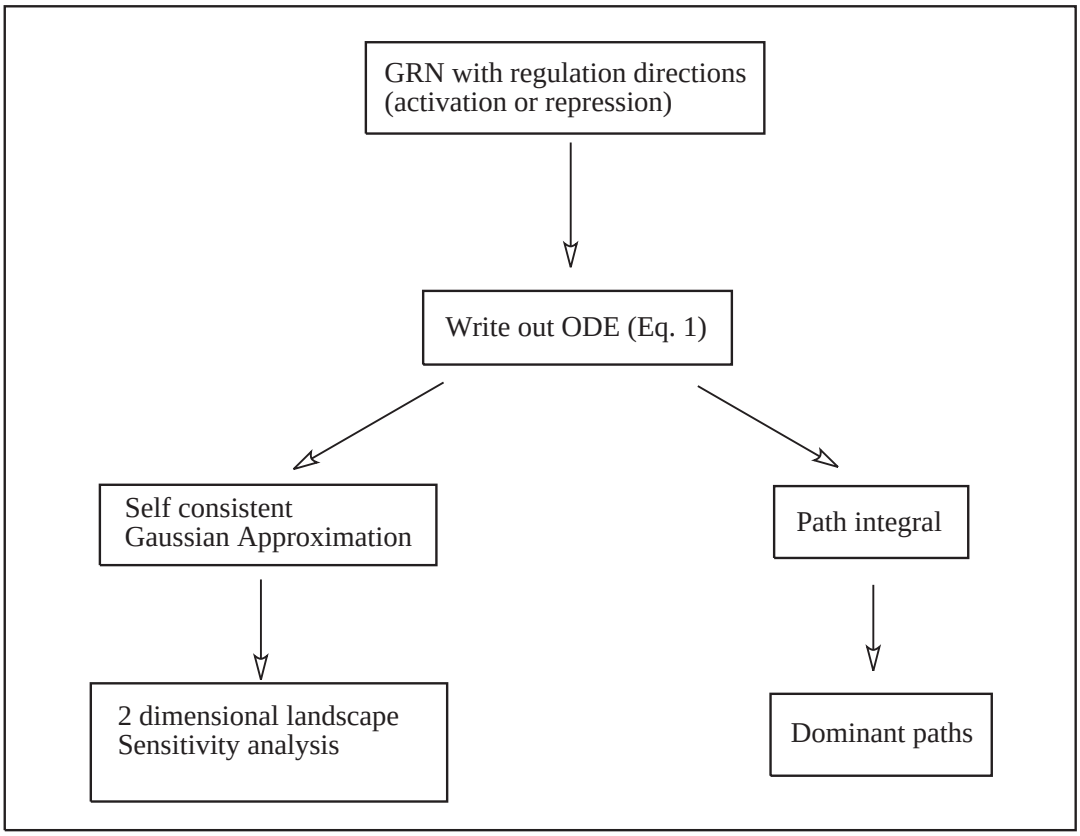

Figure S3. A flowchart for methods employed. GRN represents the gene regulatory network.