Regulatory Gene Networks in Embryonic Stem Cell Differentiation

Steven Zwick
Applied Physics
Harvard University

(Dated: May 15, 2015)

Motivated by the questions of how cells choose when to differentiate and what type of cell to become, we review a model of gene expression for large regulatory gene networks. By treating these networks analogously to magnetic systems, we rigorously define the epigenetic landscape that governs the dynamics of these cells. Under this view, cell types become attractors in a multidimensional gene expression phase space with differentiation resulting from transitions between these attractors. We also discuss how the predictions of this model fits with data and the implications of this approach for future work in cellular decision-making.

I. INTRODUCTION

During the development of multicellular organisms, progenitor cells must choose their fate and differentiate accordingly into one of the many different types of adult cells. These decisions are governed by complex, stochastic networks of DNA-binding proteins (transcription factors) that reorganize the cell's gene expression during differentiation. This transition has been classically viewed in terms of Waddington's "epigenetic landscape," in which the transition of cells through the multidimensional gene expression space is akin to marbles rolling to the bottom of a rugged valley [1]. More rigorously, we can view different cell states as robust attractors in a high dimension potential landscape, with differentiation manifesting as transitions between these attractors. Therefore, knowledge of the underlying network structure enables one, using analytical approaches from statistical mechanics, to construct a phase space of gene expression with different attractors corresponding to the different states of the cell.

A standard and intuitive method to simulate such molecular networks is to directly model the chemical reactions involved using a Monte Carlo algorithm [2]. However, this type of simulation can become computationally impractical for realistically large networks, in which case a higher level approach may be more suitable for study of these networks. Here, we review a model proposed by Zhang and Wolynes [3] that treats gene networks analogously to magnetic systems and uses methods from statistical mechanics to characterize the attractors in the system. They explicitly model the synthesis, degradation, and DNA binding of proteins in a gene network as a birth-death process on a lattice. They then apply this model to the regulatory transcriptional network of embryonic stem cells (ESCs), which early in development must choose when to leave a pluripotent state and which cell type to differentiate towards.

II. MODEL

Following Zhang and Wolynes [3], we start by describing the master equation for stochastic gene expression in the regulatory network. Suppose the network in question has M genes that transcribe M different transcription factors. A schematic for our model given a simple two gene network is given in Figure 1. We assume that each gene's expression is determined by the binding of these transcription factors to adjacent regulatory regions of DNA. Assuming that these transcription factors bind independently of each other to unique regulatory DNA elements, from trivial combinatorics we have $N = 2^M$ possible DNA occupation states, in which each transcription factor is either bound or unbound. If we label each occupancy state with an integer from $[1, N]$, we can write the joint probability of a single gene having DNA occupancy state j with n proteins in the system at time t as $P_j(n,t)$. More conveniently, we can write these probabilities as a N-component probability vector for all the possible occupancy states of a single gene,

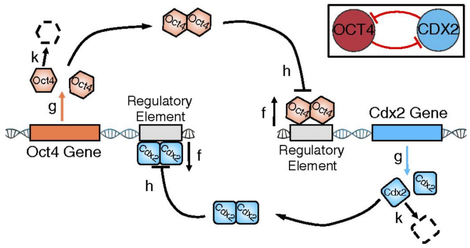

FIG. 1. Gene regulation model for network of two mutually repressing genes Oct4 and Cdx2, from Zhang and Wolynes [3].

$$\mathbf{P}(n,t) = \begin{pmatrix} P_1(n,t) \\ P_2(n,t) \\ ... \\ P_N(n,t) \end{pmatrix}. \tag{1}$$

As a first order approximation for the expression of the network of M genes, we can assume that the probabilities for each gene's DNA occupancy are independent, referred to as the self-consistent proteomic field approximation[3]. While this approximation is crude, it will enable us to solve the master equation exactly and should not affect the broken symmetries in the network. Under this assumption, if we denote the occupancy configuration probabilities for gene m as $\mathbf{P}_m(n_m,t)$, the DNA occupancy probability for the entire gene network is just the tensor product of single gene probabilities,

$$\mathbf{P}(n_1, n_2, ..., n_M, t) = \mathbf{P_1}(n_1, t) \otimes \mathbf{P_2}(n_2, t) ... \otimes \mathbf{P_M}(n_M, t). \tag{2}$$

The dynamics of the gene expression in our network can therefore be described by the following master equation for the DNA occupancy probabilities,

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{P}(n,t) = \mathbf{G}\{\mathbf{P}(n-1,t) - \mathbf{P}(n,t)\} \\ +\mathbf{K}\{(n+1)\mathbf{P}(n+1,t) - n\mathbf{P}(n,t)\} \\ +\mathbf{W}\mathbf{P}(n,t). \end{aligned} \tag{3}$$

The first term describes the change in DNA occupancy probability due to the synthesis of a new transcription factor, with transition matrix $\mathbf{G}$ which is a diagonal matrix of the protein translation rates g_j, which are determined by the DNA occupancy configuration j and the structure of the gene network. A simple rule is that the protein translation is “on” $(g_j = g_{on})$ if the occupancy configuration j includes an activator and no inhibitors for that gene, and translation is “off” $(g_j = g_{off})$ if there are no bound activators or if there are bound inhibitors. The second term accounts for exponential degradation of the transcription factors, such that $\mathbf{K} = k\mathbb{I}$ where k is the protein degradation rate. The final term has a non-diagonal transition matrix $\mathbf{W}$ that describes how the DNA occupancy states change given the network structure, and is defined as follows. We assume that the transcription factors dimerize before binding to DNA, as is the case for most transcription factors [citation]. If a transition from occupancy state j to state i requires the binding of a homodimer of transcription factor m, then $W_{ij} = hn_m(n_m - 1)/2$ where n_m is the number of transcription factors m and h is the DNA binding rate. If the transition requires the binding of a heterodimer of transcription factors l and m, then $W_{ij} = hn_l n_m$. If the transition requires a transcription factor m to unbind from the DNA, then $W_{ij} = j$ where j is the DNA unbinding rate. All other entries of the transition matrix $\mathbf{W}$ are 0.

It is worth discussing a key feature of this model before solving the master equation. Models of protein translation often assume that the timescales of transcription factors binding to DNA or unbinding from DNA are much faster than the protein translation timescale. Under this assumption, it is valid to assume that the ensemble of DNA occupancy configurations reaches a quasi-equilibrium, with the probabilities of each configuration reaching a steady state value and the translation rate depending solely on some function of the transcription factor concentrations n_m. However, experimental evidence has demonstrated that this assumption is not valid for gene expression in eukaryotes, in which the chromatin architecture and and the wrapping of DNA around histones can slow the process of DNA binding [5,6]. In this model, the DNA binding and unbinding rate parameters h and j can give insight into how these epigenetic modifications can alter gene expression and cellular decision-making.

A. Steady States

We can now solve this master equation by treating the problem as a birth-death process on a lattice and using a path integral approach. Following Doi [7] and Pelilti [8], for each gene m we can define creation and annihilation operators $a_m^\dagger|n_m\rangle = |n_m + 1\rangle$ and $a_m|n\rangle = n_m|n_m - 1\rangle$ respectively and a state vector $|\psi_m\rangle = \sum_{n=0}^{\infty} \mathbf{P}_m(n_m, t)|n_m\rangle$. The master equation for a single gene m can then be written in the following ”second-quantized” form,

$$\partial_t|\psi_m(t)\rangle = \mathbf{\Omega}_m|\psi_m(t)\rangle, \tag{4}$$

with non-Hermitian, ”Hamiltonian”-like operator

$$\mathbf{\Omega}_m \equiv \mathbf{G}(a_m^\dagger - 1) + \mathbf{K}(a_m - a_m^\dagger a_m) + \mathbf{W}^\dagger. \tag{5}$$

Here, $\mathbf{G}$ and $\mathbf{K}$ are the same as in equation (3) and $\mathbf{W}^\dagger$ is obtained by rewriting terms $hn_m(n_m - 1)/2$ and $hn_l n_m$ in operator form as $h(a_m^\dagger a_m)^2/2$ and $h(a_l^\dagger a_l)(a_m^\dagger a_m)$.

We can easily generalize this approach for the whole network of m genes under the self-consistent proteomic field approximation using equation (2), through which we obtain

$$|\Psi_m(t)\rangle = \prod_{m=1}^{M} |\psi_m(t)\rangle, \tag{6}$$

$$\partial_t|\Psi(t)\rangle = \mathbf{\Omega}|\Psi(t)\rangle, \tag{7}$$

$$\mathbf{\Omega} = \sum_{m=1}^{M} \mathbf{\Omega}_m. \tag{8}$$

Using equation (7), we can calculate the transition probability $P(\mathbf{n_f}, \tau|\mathbf{n_i}, 0)$ of finding protein concentrations $\mathbf{n_f} = (n_f^1, ..., n_f^M)^\top$ at time $t = \tau$ given initial protein concentrations $\mathbf{n_i} = (n_i^1, ..., n_i^M)^\top$ at $t = 0$, where n_i^m and n_f^m are the initial and final concentrations of transcription factor m:

$$P(\mathbf{n_f}, \tau|\mathbf{n_i}, 0) = \langle \mathbf{n}_f| \exp(\mathbf{\Omega}\tau)|\mathbf{n_i}\rangle. \tag{9}$$

Following Zhang et al. [9], we can derive a path integral representation for the transition probability using a resolution of identity . For the sake of brevity, we refer the reader to the Supplementary Information of Zhang and Wolynes [3] for the precise definitions and derivations of quantities. Following this approach, we write the transition probability as a path integral

$$\begin{aligned} P(\mathbf{n_f}, \tau|\mathbf{n_i}, 0) \propto \int \prod_m \mathcal{D}\mathbf{x} \prod_m \mathcal{D}\mathbf{c} \\ \times \exp\left(- \int dt \sum_m [\mathbf{p}_m \dot{\mathbf{q}}_{m=1}^M - \mathcal{H}(\mathbf{q}_m, \mathbf{p}_m)] \right), \end{aligned} \tag{10}$$

with
$\mathbf{q}_m = (c_1 x_1, ..., c_N x_N, (c_N - c_1)/2, ..., (c_N - c_{N-1})/2)^\top$, and $\mathbf{p}_m = (p_1^x, ..., p_N^x, p_1^c, ..., p_{N-1}^c)^\top$. Here, coordinates c_j and x_j are the probability and average protein number of DNA occupation configuration j for gene m, and p_j^c and p_j^x are the corresponding conjugate momenta. The Hamiltonian $\mathcal{H}$ is defined directly from our master equation [3] and can be seen via equation (10) to be separable over the set of genes in the network.

Following this result, the steady state solutions of the master equation are attractors of the deterministic dynamics that extremize the action of our path integral:

$$\frac{d\mathbf{q}_m}{dt} = \left.\frac{\partial \mathcal{H}}{\partial \mathbf{p}_m}\right|_{\mathbf{p}_m = 0}. \tag{11}$$

Finally, from the steady state solutions for $\mathbf{q}_m$ and $\mathbf{p}_m$, we can calculate the steady state probability distribution $P_m(n)$ and mean x_m of the concentration of transcription factor m by simply averaging over all DNA occupancy configurations,

$$\begin{aligned} P_m(n) &= \sum_{j=1}^{N} c_j \frac{(x_j)^n e^{-x_j}}{n!}, \\ x_m &= \sum_{j=1}^{N} c_j x_j. \end{aligned} \tag{12}$$

B. Transition Paths

Once the steady states of the cell network have been found, the next question is how the cell switches between these states. Now that our master equation is in Hamiltonian form in equation (10), the most probable transition paths between two steady states is simply determined by the standard Hamiltonian equations for evolution of a system

$$\frac{d\mathbf{q}_m}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_m}, \frac{d\mathbf{p}_m}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}_m}, \tag{13}$$

with the two steady states of interest as boundary conditions. Once we have determined the most probable transition path we can estimate the transition rate k between these states as $k \propto e^S$ where S is the transition action over the path:

$$S = \int dt \sum_m \mathbf{p}_m \dot{\mathbf{q}}_m. \tag{14}$$

C. Adiabatic Limit

The "adiabatic limit" of the model refers to the regime in which DNA binding and unbinding occurs on a much faster time-scale than translation. In terms of our model parameters, this is the regime $h, f \gg g_j$ for DNA occupancy configuration j. In this limit, we can easily calculate the relevant quantities by assuming that the occupancy configurations reach equilibrium, in which case we can define an effective protein translation rate over all possible configurations,

$$\bar{g} = \sum_{j=1}^{N} c_j g_j. \tag{15}$$

This reduces the Hamiltonian to

$$\mathcal{H} = \bar{g}[e^p - 1] + kx[e^{-p} - 1], \tag{16}$$

from which we can calculate the new steady state solutions from the deterministic dynamics of equation (11), which is now

$$\frac{dx}{dt} = \bar{g} - kx. \tag{17}$$

Lastly, the Hamiltonian equations from (13) governing the most probable transition path between steady states are now

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial \mathcal{H}}{\partial p} = \bar{g}e^p - kxe^{-p}, \\ \frac{dp}{dt} &= -\frac{\partial \mathcal{H}}{\partial x} = -\frac{\partial \bar{g}}{\partial x}[e^p - 1] - k[e^{-p} - 1]. \end{aligned} \tag{18}$$

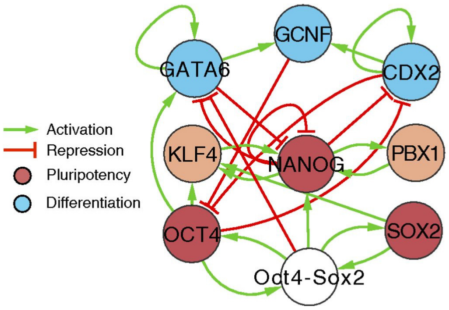

FIG. 2. Gene regulatory network for embryonic stem cell model in Zhang and Wolynes [3]. Each node represents a transcription factor except for Oct4-Sox2, a heterodimer of two transcription factors. Arrows indicate that a transcription factor binds to a regulatory DNA element for a gene encoding another transcription factor. Green arrows show that the transcription factor increases expression of a gene, whereas red arrows indicate that the transcription factor represses gene expression. The color of the transcription factor corresponds to the state of the cell that the protein is experimentally found in.

III. EMBRYONIC STEM CELL DIFFERENTIATION

By connecting steady states of the model to states of an ESC cell, we can make predictions about differentiation that can be tested experimentally. Figure 2 depicts the simplified regulatory network for mouse embryonic stem cells (mESCs) used by Zhang and Wolynes

[3]. This network governs the decision of the mESC to leave the pluripotent state and differentiate into different cell types. Simulation results of the Zhang and Wolynes model for this network [3] are shown in Figure 3, with the steady state solutions given in 3(A) and the most probable transition path between two of the states described in 3(B). The model has five steady state solutions, corresponding with the gene expression of four different cell types: differentiated cells (DC), trophectoderm (TE), primitive endoderm (PE), and stem cells (SC1 and SC2). Interestingly, the model predicts two different steady states for the stem cell state, characterized by a difference in Nanog and Pbx2 expression, agreeing with experimentally measured heterogeneity in Nanog expression in stem cell populations [10]. In figure 3(B), we can see that the SC2 state is an intermediary state in the most probable transition path from the SC1 to PE states. This agrees with experimental observations that Nanog downregulation is necessary for differentiation of mESCs [11].

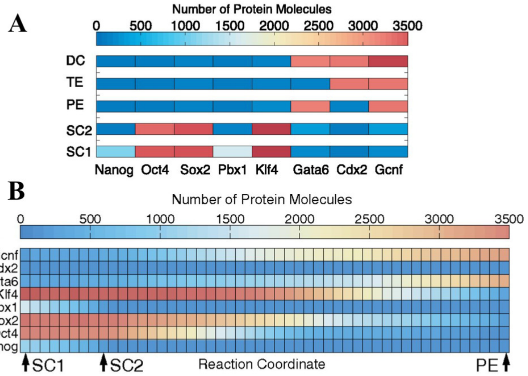

FIG. 3. Simulation results of model from Zhang and Wolynes [3]. (A) Steady state solutions (five) for the mESC network, linked to experimentally observed cell types. Each steady state is defined by the concentrations of the transcription factors along the x-axis. (B) Most probable transition pathway between SC1 (stem cell) state and PE (primitive endoderm) steady states. The x-axis represents the progression along the path and each row corresponds with a different transcription factor.

Past experimental results suggest that pluripotency genes Oct4 and Sox2 are differentially regulated in response to differentiation signals in mESCs *in vitro* [11], with high Oct4 and low Sox2 in cells adopting the mesendoderm fate and the opposite in cells differentiating into neural ectodermal cells. Preliminary experimental results in human ESCs, which have a very similar regulatory gene network to mESCs, also support this conclusion (Figure 4). It is no surprise that Zhang and Wolynes's results do not predict this result since their network (Figure 2) does not include any differential regulation of Oct4 and Sox2, leading to no steady states in which Oct4 and Sox2 expression are significantly different (Figure 3A). However, this model offers a promising approach to identify the transcription factors that are critical for the differentiation, such as the downregulation of Nanog is necessary in Figure 3B. Further work implementing this model to other gene networks may give insight into which genes are the "master" regulatory factors for a particular cell fate choice and which are merely responsive.

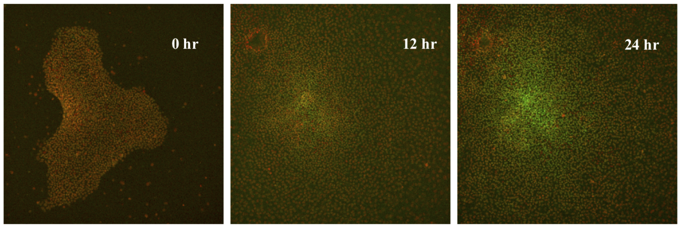

FIG. 4. Time lapse microscopy images of human ESCs with fluorescent reporters tagged to copies of Oct4 (red) and Sox2 (green) after receiving mesoendoderm differentiation signals for 0, 12, and 24 hours respectively from left to right (Unpublished).

IV. CONCLUSION

In this paper, we discussed the view of cell differentiation as transitions through an epigenetic phase space, in which cell types are steady state attractors. The model proposed by Zhang and Wolynes [3] allowed us to quantitatively characterize these attractors and the transitions between them using the formalism for magnetic systems from statistical mechanics. Even with broad assumptions, this approach produces results that agree with experiments and offer insight into how differentiation arises from the structure of the network. However, there remain many open questions to be addressed through such models. Why does the differentiation of cells appear to be an irreversible process experimentally? What role does DNA-binding kinetics (in adiabatic or non-adiabatic regime) have in the differentiation between two different cell fates? Such general questions should not depend on the specific microscopic details of the problem and invite the possibility of discovery through approaches from statistical mechanics like those discussed here.

[1] Waddington CH (1957) *The Strategy of the Genes* (Allen & Unwin, London).

[2] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340-2361.

[3] Zhang B, Wolynes PG (2014) Stem cell differentiation as a many-body problem. PNAS 111:10185-10190.

[4] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340-2361.

[5] Feng H, Wang J (2012) A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes. Sci Rep 2:550.

[6] Li C, Wang J (2013) Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. J R Soc Interface 10(89):20130787.

[7] Doi M (1976) Second quantization representation for classical many-particle system. J Phys A: Math Gen 9:1465-1477.

[8] Peliti L (1985) Path integral approach to birth-death processes on a lattice. J Phys 46: 1469-1483.

[9] Zhang K, Sasai M, Wang J (2013) Eddy current and coupled landscapes for non- adiabatic and nonequilibrium complex system dynamics. Proc Natl Acad Sci USA 110(37):14930-14935.

[10] Chambers I, et al. (2007) Nanog safeguards pluripotency and mediates germline development. Nature 450(7173):1230-1234.

[11] Thomson M, et al. (2011) Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell 145(6):875-89.