Abstract

Achieving precise control of stem cell differentiation remains one of the most challenging open problems in biology. Some experimental reasons for this will be discussed elsewhere in this thematic issue; however, an important theoretical reason precise control remains elusive is that it is not clear how best to think about one cell changing into another type of cell. The epigenetic landscape view of differentiation, where one imagines a cell as a ball rolling around a rugged pasture full of hills and valleys, offers one way to think; however, it has proven difficult to make this conceptual picture mathematically precise, partly because there are many inequivalent proposals for accomplishing this. In this review, we discuss how to think about the landscape, its different mathematical formulations, the implications of those formulations, and applications to various areas of biology. We conclude with the practical issues currently preventing the landscape from being a useful, predictive tool for understanding differentiation.

From developmental metaphor to quantitative framework: A review of Waddington landscapes and their applications

John Vastola and William R. Holmes

December 4, 2018

1 Introduction

Changing some skin cells from a patient into heart cells or brain cells—at least, cheaply and in high enough volumes to be clinically useful—has proven to be harder than understanding fundamental properties of nature (like the anomalous magnetic moment of the electron) to ten decimal places, or using astronomical observations to estimate how long ago the universe began. *Why?* Why isn't our encyclopedic knowledge of molecular biology, medicine, mathematics, and computation enough?

There are many, many, many good reasons, some of which will be discussed elsewhere in this thematic issue on iPSC differentiation: difficulties with experimental set-up, culturing, automation, data collection and analysis, and so on. But there are still other reasons, which prevent even the best-designed equipment in the world from being enough to solve the problem on its own. They boil down to the following: how should we *think* about iPSC differentiation? More generally, how should we think about one cell changing into another type of cell?

One popular approach is to take the so-called *epigenetic landscape* (or *Waddington landscape*, after Conrad Hal Waddington, who first proposed the idea[8]) view of cell identity. Usually, the differences between one type of cell and another type of cell in the same organism are not genetic, but *epigenetic*: different proteins are expressed more or less highly, perhaps because the same transcriptional regions are regulated differently, or because of some other difference in (not necessarily transcriptional) regulation. In this view, a well-equipped and sufficiently patient researcher could in principle change any cell into any *other* type of cell, provided that they adjusted the expression levels of each protein accordingly.

The standard way of imagining one cell changing into another type of cell, in this view, is to picture a ball rolling around a rugged collection of hills and valleys. A ball in a valley corresponds to an epigenetic state which is somewhat hard to escape; if you move the ball slightly by making small epigenetic changes, the ball will roll back down into the valley. In other words, balls in valleys represent (possibly mature, possibly not) cell types or subtypes.

Meanwhile, a ball on a steep hill will roll down that hill quickly until it reaches a valley; it corresponds to a cell whose epigenetic state is constantly changing.

This mental picture has some important consequences. First, it suggests that differentiation is always *reversible*, and that any cell can be directly or indirectly reprogrammed into any other type of cell with sufficient effort. This is because there always exists some direct epigenetic path—however difficult to realize in a real laboratory experiment—in the landscape between any two cells types or subtypes. Second, it suggests that there are two ways an experimenter can influence the epigenetic state of a given cell: they can either *move the ball* (change the expression levels of proteins or other relevant species), or *change the landscape* (supply a drug that fundamentally changes some regulation event, causing the hills and valleys to change shape) to make the ball more likely to move where they want.

Unfortunately, though the landscape has been influential as a metaphor and way of thinking about differentiation, it has been somewhat difficult to realize as a precise mathematical object. In other words, given some experimental data or a mathematical model of a biological system, how does one actually construct the landscape corresponding to that system? Part of the problem is that there are many inequivalent proposals for mathematically defining a landscape. Some of these proposals are equivalent *in certain limits*, as discussed by Zhou and Li[11], but they are not equivalent in general. *Why?* Why is it so difficult to define a landscape?

In this review, we will offer an explanation for why there are different landscapes—and why this should be expected, since the kind of landscape one is interested in will change depending on the question one is asking. We will also explain how to think about landscapes, how the various proposed landscapes are similar and different, and how they can be applied to understanding real biological systems. Finally, we conclude with some discussion of where landscape research might go next, given the myriad important theoretical issues that remain to be addressed.

It should be noted that we owe a great intellectual debt to Zhou and Li, whose recent review[11] of different landscape formulations we draw upon liberally. Our approach differs from theirs in that we take a somewhat less mathematical tact, offer a unifying definition of landscapes, and cover a few landscape constructions that they did not.

[WRH: The purpose of this review is not to provide a primer on how to technically construct landscapes; that is best left for the source literature where the highly technical mechanics of doing so are developed. Rather, our intent here is to discuss the main theoretical ideas that come from this body of work and how it alters the fundamental picture of what a landscape is and how it should be conceptualized. As such, this review is structured as follows. We begin by discussing the subtleties of how the qualitative concept of a landscape can be formalized, the different conceptual ways a landscape can be described, and the practical consequences of those subtleties. Next we briefly describe the varying mathematical techniques that have been developed to quantitatively construct landscapes from models of gene expression.The purpose of this more technical section is to provide enough detail to give the reader a sense of what these ideas are useful for and the challenges that must be addressed going forward. We have endeavored to make this section stand alone so that the it is not

required reading for the remainder of the review. Subsequently, we discuss applications where these formalized landscapes have provided new insights and future directions and challenges in this field.]

2 What is a "Landscape" and what influences its structure?

At the most basic level, a landscape is a construct (mathematical or visual) that helps describe the observable states and / or dynamics of a system. While in the 1950s Waddington first conceptualized development in terms of a landscape that cells occupy and dynamically evolve in, the more general concept was theoretically developed to study physical systems well before that. In the physics literature, the most common conceptualization of a landscape is that of an "energy landscape". This is a mathematical function that maps all possible states of a system onto an energy level, typically the Gibbs free energy[CITATION?]. However, one of the fundamental tenants of statistical mechanics (e.g. the Boltzmann distribution) is that the probability of a system being in a particular state is determined by the energy of that state[CITATION standard stat mech book?]; lower energy states are more likely. Thus, an energy landscape can alternatively be viewed as mapping between the states of a system and their probabilities. This is probably the most central property of a landscape: *it should provide useful information about the likelihood (e.g. probability) of observing a system in different possible states*. In the context of cellular differentiation for example, a landscape should identify stable cell states and provide information about the relative stability of those states (e.g. the relative occupancy of those states).

In addition to providing probabilistic information about a system, the "energy" landscape also provides dynamic information. For classical conservative systems, the time evolution of that systems state is known to evolve according to the contours of the underlying landscape (e.g. gradient descent)[CITATION]. Visually, this is often depicted as a ball rolling through the landscape and following a path of steepest descent toward a valley. This is the second type of information one might like to glean from a landscape, system dynamics. In the differentiation example, this type of information could facilitate, for example, predictions of how cells transition from one state to another or the development of directed differentiation protocols.

These are the two basic characteristics one would like a biological landscape to have: (i) probabilistic information about cell states, and (ii) dynamic information about how cells evolve within that landscape. Both of these characteristics are an intrinsic part of Waddington's original conceptualization of the developmental landscape. There are, however, well-known flaws in the analogy between physical landscapes and Waddington's conceptualization of biological landscapes that are of fundamental importance and impact how systems work. These flaws are not merely mathematical details, but rather are the result of fundamental differences (conservative versus non-conservative dynamics) between biological systems and the typical physical systems for which the landscape concept was first developed.

In the coming sections, we will begin by discussing, in a moderate level of mathematical detail, the construction of classical "energy" landscapes and the fundamental assumptions that are required for that approach. We will then discuss how biological systems are fundamentally different and don't obey those assumptions. Recognizing these differences is a necessary first step, as they fundamentally alter how one constructs a landscape, what kind of information can be obtained from that landscape, and even how one should conceptually think about a landscape.

2.1 The classic conservative system landscape

We begin by describing the physics-based motivation for the concept of a landscape. The temporal dynamics of numerous physical systems can be described by a class of systems referred to as deterministic conservative systems, which are by definition systems that do not consume energy; that is, they "conserve" energy. In continuum settings, such systems can always be described by a dynamical system of the form

$$\frac{dx}{dt} = f(x) = -\nabla U, \tag{1}$$

where f describes the dynamics of the system and U is a "potential energy" that the system seeks to minimize. In this setting, U defines the landscape of this system. More specifically, the minima of U describe the stable states of the system and the spatial gradient of U determines the dynamics of the system. In the classical view of a landscape as a system of hills and valleys that a ball rolls around in, U determines where those hills and valleys are and ∇U determines how a ball rolls around within them. [WRH: I would provide a specific example of this with a figure. Could be 1D, but 2D might be more useful with a nice surface plot of U with some gradient paths.]

Dynamical systems and Ordinary Differential Equations have long been used to describe gene expression dynamics. There is however a critical feature of gene expression models that distinguish them from conservative systems (Equ. 1). The equations associated with these models do not describe the physical motion of molecules through space, but rather the change in number of different molecular species over time. One does not expect such dynamics to be conservative *a priori*, and indeed they are not in general. As a result, the dynamics (f) cannot be written in terms of a potential energy (U), and thus the standard physics-based method of translating dynamics (f) into a landscape (U) does not apply. In subsequent sections we will describe the theory and mathematical techniques that have been developed to conceptually and quantitatively construct landscapes that lack this simplifying conservation property.

2.2 Mathematical models underlying landscapes

Landscapes provide an alternative way to describe the dynamics of a system. Thus, a landscape is fundamentally linked to an underlying mathematical model describing the biology (or physics) of the system. In the classical conservative systems context, the underlying

model is a deterministic dynamical system. However, there are alternative ways to mathematically model such systems. Since the concept of a landscape is linked to the underlying framework being used to model a system, we briefly discuss these frameworks.

The most basic, common method of describing gene regulatory systems is via a Chemical Master Equation (CME). This is a continuous time, discrete concentration (e.g. it models the number of molecules rather than continuous concentrations) method of modeling gene expression dynamics. For example, consider the following simple birth-death process that models the production and decay dynamics of unregulated mRNA. From a reaction kinetics perspective, one way to represent this system is

$$g \xrightarrow{k_m} g + m, \qquad m \xrightarrow{d_m} \varnothing$$

where the first reaction represents production of mRNA from a gene that is always on and the second the decay of that mRNA. The CME corresponding to this system is then

$$\frac{dP(m,t)}{dt} = k_m \left[P(m-1,t) - P(m,t)\right] + d_m \left[(m+1)P(m+1,t) - mP(m,t)\right] \ ,$$

where $P(m,t)$ is the probability that there are exactly m mRNA molecules at time t. While this is only a simple example, it can be extended easily to account for the (in)activation of genes, the production of protein from mRNA, and the crosstalk between different genes.

Gillespie demonstrated [2] that this CME can be rigorously translated into a stochastic differential equation (SDE) referred to as the chemical Langevin equation (CLE)

$$\dot{x} = f(x) + g(x)\ \eta(t). \tag{2}$$

This is a continuous time, continuous concentration based model where $x = (x_1, x_2, ..., x_N)$ is a vector containing the concentrations of each of the N species in the model, $f(x)$ describes the deterministic dynamics of the system, $g(x)$ the stochastic influences, and η is a standardized Gaussian noise term[JJV: Not sure if this is sufficiently well-motivated. It isn't said why one should use SDEs here.]. While in Gillespie's formulation the functions f, g are fully determined by the underlying reaction kinetics, in many cases SDEs of this form are phenomenologically constructed using combinations of linear and Hill functions for f and simplified noise kinetics such as additive ($g(x) = \sigma = \text{const.}$)[CITATIONS], or less commonly multiplicative ($g(x) = \sigma x$)[CITE HOLMES EMBRYO PAPER], for g. SDEs such as this typically, or their deterministic counterpart (where $g = 0$) typically form the basis landscape constructions.

While they are not the focus of this review, Markov models are another common way to model system dynamics. These are discrete time, discrete state models of cell dynamics that typically take one of two forms. In the first, a Markov process models the actual gene expression / protein dynamics within a cell. [WRH: Maybe briefly discuss some of Munsky's stuff here. No more than 2 or 3 sentences.]

[WRH: A second way of doing this is to define a bunch of cell states and generate a Markov process for transitions among them. Some brief introduction to this is useful here (non-technical), since this is where data science approaches link in.]

2.3 The critical role of stochasticity in determining biological landscapes

One thing that should be clear from the prior section is that models of gene expression dynamics are intrinsically stochastic. While that stochasticity may sometimes be neglected, numerous studies **(REFS)** have shown that noise in these processes is significant and that it can have profound effects. This is however somewhat at odds with the qualitative conceptualization of landscapes that has become pervasive in recent decades. In that view, the landscape is typically viewed as a deterministic surface and that stochastic dynamics and external perturbations drive fluctuations of a ball (i.e. cell state) on that surface. As will be shown in the more technical discussion in this review, stochasticity intrinsically shapes the landscape itself. Consider the simple example of a conservative dynamical system from above, but in this case suppose that conservative system is stochastic

$$\dot{x} = f(x) + g(x)\ \eta(t) = -\nabla U + g(x)\eta(t). \tag{3}$$

[WRH: John. I would take the example from above, where a basic deterministic landscape was constructed, and construct it again here with varying levels of additive noise and varying levels of multiplicative noise. So you could have 3 panels. (a) No noise. (b) Multiple levels of additive noise. (c) Multiple levels of multiplicative noise. No need to write down a FP equation or anything. Just solve the think. I would add a short appendix where you manually solve the three conservative landscape problems. I think this is a good use of appendix, which can just be referenced.] As this example shows, the landscape is intrinsically a stochastic construct and noise characteristics fundamentally shape that landscape. This is particularly relevant given recent observations that noise in gene regulation can be regulated independently of mean expression dynamics [5, 7] and that specific ways to modulate gene expression noise are being identified[3, 1].

3 Constructing a landscape

Having discussed the basics of mathematical models of gene regulation, as well as how to think about landscapes qualitatively, we provide a more technical discussion of what a landscape is, the various ways of constructing one, and the ways in which those different landscapes differ. We will discuss how, in general, the different kinds of information we might like to gain from a landscape are incompatible; necessitating the use of different types of landscapes to address different questions. The purpose of this discussion is two fold. First, to provide the interested reader with a primer on the different techniques in this field, though we defer many mathematical details to the literature from which these methods originated. Second, to use this more in depth discussion to demonstrate how subtle choices made in the construction of landscapes have consequences on the type of information those landscapes contain and their interpretations.

3.1 Generic properties of all landscapes

Let us begin with the *most basic* requirement of any mathematical construction we would like to call a landscape: it must associate each of the system's possible states with a 'height' (a real number). Mathematically, this means that a landscape must be a function $L : S \to \mathbb{R}$, where S is the state space of the system. This property allows us to compare any two possible states; however, how we interpret this comparison will change depending on the kind of landscape we are interested in, as we will describe below.

On top of this, there are several properties that we would *like* to be true for our construction. We would like there to be one valley for each cell type in our model; we would like the landscape to be bounded from below (i.e. there are no states which are completely impossible to escape); and we would like the landscape to be continuous for a continuous state space S. Furthermore, we would like our construction to involve few or no arbitrary choices, and we would like it to be agnostic to the underlying kind of model—in other words, we should be able to construct a landscape regardless of whether our system is modeled using the CME, CLE, some combination, or something else entirely.

We note that, since the state space S is not continuous in general, we should not require that a landscape be continuous; in fact, we will give an example in the next section of a commonly used discrete landscape (which we think should legitimately be considered a landscape according to the definition we are about to give).

3.2 Global vs local landscapes

[WRH: I removed the text you had in this section. I didn't have an issue with what was there, it was just incomplete and in bullet point form. So I left this blank so that you can start with a clean slate on writing it.]

3.3 Mathematical definition

[WRH: Can we combine this with the previous sub-section?]

Here we provide a simple but technical definition for what a landscape is. Note that this is not a definition of "the landscape". As we have discuss previously and will formalize later, there are multiple was to construct a landscape for a given system; it is not unique. Instead, we provide a set of mathematical properties that a landscape should have. We note that while this formalization does place some of the subsequent discussion on firmer mathematical ground, it is not necessary for this review and we have endeavored to make all important points in this review accessible without this more technical formalism.

Definition. A *global landscape* is a function $L : X \to \mathbb{R}$ satisfying:

- Let $x_1, x_2 \in X$. If $P_{ss}(x_1) \geq P_{ss}(x_2)$, then $L(x_1) \leq L(x_2)$.

Definition. A *local landscape* is a function $L : X \to \mathbb{R}$, a distinguished point $x_0 \in X$, and a (possibly degenerate) time interval $[T_-, T_+)$ such that:

- Let $x_1, x_2 \in X$. If $P(x_0 \to x_1, T) \geq P(x_0 \to x_2, T)$ for all $T \in [T_-, T_+)$, then $L(x_1) \leq L(x_2)$.

Note that a local landscape becomes a global landscape in the limit taking $T_-, T_+ \to \infty$.

Note also that the definition of a local landscape requires two pieces of additional information: a base point, and a time interval. This is because we want local landscapes to contain information about the cell state transitions, and the likelihood of a cell state transition to any particular cell state depends on the starting point and the time scale of interest. Since the likelihood of a transition will change as these choices change, so too should the local landscape associated with the transition.

3.4 Intuition for mathematical definition

[WRH: Not sure what you are going for here, but will wait until you write it.]

3.5 Simplest model: Markov chain

[WRH: Since most of this article revolves more around continuous systems, would it make sense to start with the Wang, Huang, etc. landscapes and put the discrete approach last? I realize mathematically this is simpler and might make sense as the starting point in that sense, but it is not the primary thing we discuss here.]

Consider a discrete time Markov chain: a list of M vertices X together with transition probabilities p_{ij} for each $i, j = 1, ..., M$ (all satisfying $0 \leq p_{ij} \leq 1$). In a given time step, the state either changes or it doesn't, so we have

$$p_{i1} + p_{i2} + \cdots + p_{iM} = 1$$

for each $i = 1, ..., M$.

The structure of a discrete time (or continuous time) Markov chain can be visualized using a directed graph, as in [FIGURE]. For our purposes, each node might represent a cell type or subtype, which means the model's state space is drastically reduced (to just a finite number of states) from the full state spaces considered in most landscape models.

This is the landscape most people compute in practice, although we have rarely seen it be explicitly identified as a distinct kind of landscape model. The motivation for using a landscape like this is the following: most often, we care about the relative stability of attractors/cell types, and transitions between attractors/cell types, but not about the relative stability of/transitions between any two *arbitrary* states. Because many theoretical landscapes assign a 'height' to *every* possible state, these landscapes contain much more information than is actually desired. We might want something coarser—and in particular, we might want something coarse enough that we can actually construct it from the generally sparse experimental data currently available.

To define a global landscape on a Markov chain, we can define

$$\phi_G(i) := -\log(P_{ss}(i)) \tag{4}$$

for $i = 1, ..., M$. To define a (single time step) local landscape (with base point j) on a Markov chain, we can define

$$\phi_L(i) := -\log(1 + p_{ji}) \ . \tag{5}$$

It is easy to check that this satisfies our earlier definition of a local landscape.

3.6 Steady-state probability landscape

Wang 2008 paper (first paper discussing this?): [9]

This is the landscape most people are thinking about when they talk about "the landscape". It is defined as

$$\phi(x) := -\log P_{ss}(x)$$

for all x in the state space, where P_{ss} is our system's steady state probability distribution (if it exists[1]).

Usually this landscape is discussed in the context of continuous regime dynamics, i.e. in systems where the absolute numbers of all species in the model are large enough that their concentrations can be approximated as continuous variables. In this case, if there are N species, and the state of the system can be described by a state vector $x = (x_1, x_2, ..., x_N)$ (where x_i is the concentration of the ith species), the system evolves stochastically in time via a Langevin equation

$$dx_i = f_i(x_1, ..., x_N)dt + g_i(x_1, ..., x_N)dW_i \ , \tag{6}$$

where the functions $f_1, ..., f_N$ control the 'deterministic' part of the dynamics, the functions $g_1, ..., g_N$ control the 'noise' part of the dynamics, and each W_i is a Weiner process. Given these dynamics, the time-dependent probability distribution $P(x, t)$ is governed by the Fokker-Planck equation

$$\frac{\partial P(x,t)}{\partial t} = \sum_{i=1}^{N} -\frac{\partial}{\partial x_i}\left[f_i(x)P(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x_i^2}\left[g_i(x)^2 P(x,t)\right] \ .$$

The steady-state probability distribution $P_{ss}(x)$ can be recovered from taking $\frac{\partial P(x,t)}{\partial t} \to 0$, so that it is the solution of the equation

$$0 = \sum_{i=1}^{N} -\frac{\partial}{\partial x_i}\left[f_i(x)P(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x_i^2}\left[g_i(x)^2 P(x,t)\right] \ . \tag{7}$$

[1]It is easy to come up with pathological systems for which P_{ss} does not exist. For example, one can imagine a system with oscillations that produce limit cycles[6], or with dynamics that do not 'dissipate' sufficiently quickly at the boundary of the domain (see pg. 4 of W. Huang et al.[4]). Oscillations are common in biological systems [CITATION? cell cycle, circadian rhythms, etc], but one can get around this restriction in practice by (for example) time-averaging over a period of oscillation. CME-derived dynamics are usually well-behaved, so the second issue is not usually a problem.

In other words, in practice, Eq. 7 is the equation that must be solved to compute the steady-state probability landscape.

At this point, we should make a few comments. First, the steady-state probability landscape is not *only* defined for stochastic systems governed by Langevin equations. Steady-state probability distributions exist in fairly general circumstances, including when systems have discrete state spaces. For example, one can easily talk about P_{ss} for stochastic dynamics governed by the CME, for mixed systems (whose dynamics are partially governed by the CME, and partially by Langevin equations), and for discrete or continuous-time Markov chains. This is nice, because we would like to be able to talk sensibly about a landscape regardless of the kind of mathematical model we choose to use.

Next, the local maxima of P_{ss} will correspond to the local minima of ϕ (i.e. the attractors that represent cell types or subtypes). If each g_i is constant (so that noise is additive), then there will be a one-to-one correspondence between the local maxima of P_{ss} and the solutions to $f_1(x) = f_2(x) = \cdots = f_N(x) = 0$ with $\frac{\partial f_i}{\partial x_j}(x) < 0$ for all $i, j = 1, ..., N$. However, *in general*, it can happen that the two do not correspond. For example, it can happen that there is a state y that satisfies the aforementioned conditions, and that we might expect is an attractor/local maximum of P_{ss}, but that is not, because that attractor is destabilized by excess noise. Put differently, the state-dependence of the g_i can be very important for determining which states are attractors and which are not; this serves as yet another reason that we should take the state-dependence of noise seriously.

3.7 Local and global quasipotential landscapes

Big 2016 review, good discussion of local/global quasipotentials: [11]

Often discussions of the landscape are centered around the *global landscapes* that we defined earlier: landscapes that can capture information related to the relative stability of any two states, but that do not in general contain information about transitions between states. This is perhaps an unfortunate state of affairs, given that one of the original motivations for thinking about the landscape was as a way to visualize cell state transitions.

$$\phi^{QP}(x; x_0) := \inf_{T>0} \inf_{\text{paths} X(t)} \int_0^T L(X(t), \dot{X}(t))\ dt \tag{8}$$

3.8 Vector decomposition landscapes

Suppose that we are interested in a system whose dynamics are governed by a Langevin equation

$$dx_i = f_i(x)dt + \sigma dW_i\ , \tag{9}$$

where $\sigma > 0$ is *constant*. We can describe this system as having symmetric additive noise (i.e. each species has the same constant noise σ, rather than there being different σ_i for each species); while almost all biological systems do *not* satisfy this requirement, symmetric additive noise may be a reasonable approximation in cases where noise is relatively small and not qualitatively important to dynamics.

For such a system, Zhou, Aliyu, Aurell, and Huang defined a landscape[10] U^{norm} via the equation

$$\sum_{i=1}^{N} \frac{\partial U^{\text{norm}}}{\partial x_i}(x) \left(f_i(x) + \frac{\partial U^{\text{norm}}}{\partial x_i}(x) \right) = 0 \ . \tag{10}$$

There are two reasonable ways to motivate this definition. The first way goes according to the following argument from the original paper.

The force vector $\mathbf{f} = (f_1, f_2, ..., f_N)$ which determines much of a cell's dynamics is generally *not* the gradient of a potential, i.e.

$$\mathbf{f}(x) \neq -\nabla U(x) \tag{11}$$

for some function U. It would be nice if this *were* true, as it is for many differential equations from physics, because it would reduce the problem of understanding the dynamics due to the N functions $f_1, ..., f_N$ to the problem of understanding the single function U.

The best we can do is write

$$\mathbf{f}(x) = -\nabla U(x) + \mathbf{F}_R(x) \tag{12}$$

for some remainder force $\mathbf{F}_R$ which will depend on how U is chosen. To make sure that U is as 'independent' from $\mathbf{F}_R$ as possible, we can restrict ourselves to thinking about ∇U and $\mathbf{F}_R$ that are always perpendicular, so that

$$\begin{aligned} 0 &= \nabla U(x) \cdot \mathbf{F}_R(x) \\ &= \sum_{i=1}^{N} \frac{\partial U}{\partial x_i}(x)(F_R)_i(x) \\ &= \sum_{i=1}^{N} \frac{\partial U}{\partial x_i}(x) \left(f_i(x) + \frac{\partial U}{\partial x_i}(x) \right) \ , \end{aligned}$$

which is the condition for U^{norm} written above.

Alternatively, U^{norm} can be thought of as an approximation to the steady state probability landscape ϕ which is only valid in the small symmetric additive noise limit. Start with Eq. 7 and substitute in the WKB ansatz

$$P_{ss}(x) = \exp\left[-\frac{\phi(x)}{(\sigma^2/2)} + \phi_0(x) + \phi_1(x)(\sigma^2/2) + \cdots \right] \tag{13}$$

to find the equation

$$0 = \sum_{i=1}^{N} \ \frac{2}{\sigma^2} \frac{\partial \phi(x)}{\partial x_i} \left(f_i(x) + \frac{\partial \phi(x)}{\partial x_i} \right) - \frac{\partial f_i(x)}{\partial x_i} - \frac{\partial^2 \phi(x)}{\partial x_i^2} \ , \tag{14}$$

which is valid in the limit where σ is sufficiently small that the higher-order terms $\phi_0, \phi_1, ...$ can be neglected.

For sufficiently small σ, the term proportional to $1/\sigma^2$ dominates, leading to the equation

$$0 = \sum_{i=1}^{N} \frac{2}{\sigma^2} \frac{\partial \phi(x)}{\partial x_i} \left(f_i(x) + \frac{\partial \phi(x)}{\partial x_i} \right) , \tag{15}$$

which is the same as Eq. 10 after dividing out the factor of $2/\sigma^2$.

There is an important downside to using this landscape: it does not take state-dependent (or even asymmetric additive) noise into account, and so one cannot sensibly apply it in situations where the state-dependence of noise is qualitatively important. To be frank, given that this landscape is known to be an approximation to the steady state probability landscape which is *only* valid in the small symmetric additive noise limit, the steady state probability landscape should be used instead of this one whenever possible.

Even if the system of interest can be reasonably well-described as having symmetric additive noise, one may as well just solve the Fokker-Planck equation to get a sense of how that noise affects the landscape.

4 Beyond the landscape: modeling cell state transitions

Landscapes are not the only way to quantify or represent properties of an underlying biological system. As discussed earlier, there are many types of insight or information one might like to gain about a system. One particularly important aspect to understand about a system is how cells transition between different states. From a landscape perspective, one can think of this as trying to quantify the paths of transitions between two states within a landscape. Suppose that a cell is currently in state A, and that we would like to perturb it so that it makes the transition to state B. What path through epigenetic state space is most likely to take the cell from state A to state B?

This is one of the fundamental questions faced by biologists interested in controlling differentiation—for example, to turn a skin cell into a heart cell. If we know the most likely path, then we know which ways we should be trying to 'push' the cell through epigenetic state space, which can help us determine the kinds of chemical perturbations that are most likely to facilitate a given differentiation protocol's successful completion. Incidentally, it seems to us that this question is also the most salient motivation for these same biologists to think about landscapes. After all, one of the salient potential uses of a landscape representation of a system is to is to facilitate the development of effective differentiation protocols.

While landscapes may provide an intuitive way to think about this issue and certain landscapes may contain intrinsic information about these paths, Landscapes themselves are not the most useful way to quantify and analyze these transition paths. Instead, the related formalism of minimal action paths provides a more direct way to do so. Here we discuss this formal way of quantifying transition paths mathematically. While this is a distinct way of thinking about system dynamics, we do note that it is not mutually exclusive of the landscape concept. Rather, landscapes and minimal action paths are intimately linked.

4.1 Motivation: least action paths in physics

In physics, how a (mechanical) system will transition from one state to another state in a fixed amount of time is addressed by the *least action principle*. Given a conservative system with total kinetic energy T and total potential energy U, we can define a function called the *Lagrangian* by $L := T - V$. It turns out that the path $x(t)$ that a system takes through state space (satisfying the boundary conditions $x(0) = A$ and $x(T) = B$) is such that the integral

$$S[x(t)] := \int_0^T L(x, \dot{x}, t)\ dt$$

is at an extremum (local maximum or minimum). In practice, this often means that the path is such that S is as *small as possible*.

A branch of math called the *calculus of variations* says that S will be at an extremum if and only if the Euler-Lagrange equations

$$\frac{\partial L}{\partial x_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) \tag{16}$$

are satisfied for each $i = 1, 2, ..., N$. To summarize, to find a transition path for a mechanical system, you (i) write down the system's Lagrangian, (ii) use that Lagrangian to write down corresponding Euler-Lagrange equations, and (iii) solve those equations for $x(t)$ subject to the original boundary conditions ($x(0) = A$, $x(T) = B$). Of course, there are fancy tricks one can incorporate (like taking advantage of the fact that certain quantities might be conserved), but that is the general idea.

4.2 Least action paths in biology

It turns out that there is an analogous procedure for what might be called *stochastic mechanics*: the study of the most likely transition path taken by a stochastic system between two points in state space in a given amount of time. Given Langevin dynamics governed by Eq. 6, it turns out that the Lagrangian

$$L = \sum_{i=1}^{N} \frac{\left[\dot{x}_i - f_i(x)\right]^2}{2g_i(x)^2} \tag{17}$$

determines the most probable transition path.

Incidentally, this Lagrangian has recently been the subject of much confusion in literature related to stochastic dynamics. While this Lagrangian was known to be the correct one in the case of additive noise, many authors [CITATIONS] suggested that it should be modified in the case of multiplicative/state-dependent noise. For readers interested in the gory mathematical details of why this is not so, Cugliandolo and Lecomte[CITATION] wrote an excellent paper regarding the subtleties of the improper ways that many authors were making changes of variables to derive the results. They discuss the Lagrangian for the general α interpretation of SDEs, but we are only interested in the Ito interpreted result ($\alpha = 0$), because Gillespie[CITATION] showed us this is the natural way to interpret SDEs originally derived from the CME.

4.3 Least action paths beyond the continuous regime

The Lagrangian we described in the previous section answers the following question: given a stochastic gene regulatory network whose concentrations can all be well approximated as continuous, which path is a cell most likely to take from state A to state B to achieve a transition in time T? Notice the requirement that the system's concentrations be approximately continuous. This is almost always not true for every species in a system, since (for example) important transcription factors are often present in low copy numbers. What if this requirement does not hold for every species in the system, or even *any* species in the system?

Eq. 17 is derivable from the CLE (Eq. 3); specifically, it is derivable from the path integral associated with transition probabilities in the CLE framework. In principle, one can consider a path integral formulation of transition probabilities in the CME framework to derive an action appropriate for species with low copy numbers. However, one might not be able to associate this action with a Lagrangian in the usual sense, since the validity of considering a Lagrangian and Euler-Lagrange equations *in lieu of* the full action is predicated upon the state space being continuous. More specifically, the derivation of the Euler-Lagrange equations depend on our ability to make infinitesimal perturbations of an arbitrary path through state space; clearly, this is not possible for a discrete state space.

A Solving the 1D Fokker-Planck equation

For Langevin dynamics that proceed according to

$$\dot{x} = f(x) + g(x)\eta(t) \ ,$$

the 1D steady state Fokker-Planck equation for $P_{ss}(x)$ reads

$$0 = -\frac{d}{dx}\left[f(x)P_{ss}(x)\right] + \frac{1}{2}\frac{d^2}{dx^2}\left[g(x)^2 P_{ss}(x)\right] \ .$$

Integrating, we have

$$C = f(x)P_{ss}(x) - \frac{1}{2}\frac{d}{dx}\left[g(x)^2 P_{ss}(x)\right]$$

for some constant C. But we need $P_{ss}(x)$ and its derivative to go to zero as x goes to infinity; hence, we can see that the right hand side must approach zero for large x, which means that C must be zero.

Now we have

$$f(x)P_{ss}(x) - g(x)g'(x)P_{ss}(x) - \frac{1}{2}g(x)^2 P'_{ss}(x) = 0 \ ,$$

or equivalently,

$$P'_{ss}(x) + \left[\frac{g(x)g'(x) - f(x)}{g(x)^2/2}\right] P_{ss}(x) = 0 \ .$$

Define the function

$$\boxed{V(x) := \int \frac{g(x)g'(x) - f(x)}{g(x)^2/2} \ dx} \ .$$

Now our equation is

$$P'_{ss}(x) + V'(x)P_{ss}(x) = 0 \ ,$$

and its solution is clearly

$$\boxed{P_{ss}(x) = Ne^{-V(x)} = \frac{e^{-V(x)}}{\int_0^\infty e^{-V(x)} \ dx}}$$

where we are using N as shorthand for the normalization constant, and we normalize over $[0, \infty)$ since concentrations can only be nonnegative.

B Second Appendix

References

[1] Roy D. Dar, Nina N. Hosmane, Michelle R. Arkin, Robert F. Siliciano, and Leor S. Weinberger. Screening for noise in gene expression identifies drug synergies. *Science*, 344(6190):1392–1396, 2014.

[2] Daniel T. Gillespie. Chemical Langevin equation. *Journal of Chemical Physics*, 113(1):297–306, 2000.

[3] Maike M.K. Hansen, Winnie Y. Wen, Elena Ingerman, Brandon S. Razooky, Cassandra E. Thompson, Roy D. Dar, Charles W. Chin, Michael L. Simpson, and Leor S. Weinberger. A post-transcriptional feedback mechanism for noise suppression and fate stabilization. *Cell*, 173(7):1609 – 1621.e15, 2018.

[4] Wen Huang, Min Ji, Zhenxin Liu, and Yingfei Yi. Steady states of fokker–planck equations: I. existence. *Journal of Dynamics and Differential Equations*, 27(3):721–742, Dec 2015.

[5] Christopher Rackauckas, Thomas Schilling, and Qing Nie. Mean-independent noise control of cell fates via intermediate states. *iScience*, 3:11–20, 2018.

[6] M. San Miguel and S. Chaturvedi. Limit cycles and detailed balance in fokker-planck equations. *Zeitschrift für Physik B Condensed Matter*, 40(1):167–174, Mar 1980.

[7] Julian Sosnik, Likun Zheng, Christopher V Rackauckas, Michelle Digman, Enrico Gratton, Qing Nie, and Thomas F Schilling. Noise modulation in retinoic acid signaling sharpens segmental boundaries of gene expression in the embryonic zebrafish hindbrain. *Elife*, 5:e14034, 2016.

[8] C H Waddington. *The strategy of the genes.* 1957.

[9] J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. *Proceedings of the National Academy of Sciences*, 105(34):12271–12276, 2008.

[10] Joseph Xu Zhou, D. S.M. Aliyu, Erik Aurell, and Sui Huang. Quasi-potential landscape in complex multi-stable systems. *Journal of the Royal Society Interface*, 9(77):3539–3553, 2012.

[11] Peijie Zhou and Tiejun Li. Construction of the landscape for multi-stable systems: Potential landscape, quasi-potential, A-type integral and beyond. *Journal of Chemical Physics*, 144(9), 2016.