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

November 26, 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[6]) 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.

2 Quick primer on mathematically modeling gene regulation

To successfully make the transition between thinking of the landscape as a helpful metaphor to thinking of it as a specific, quantitative object, it is crucial that we are on the same page regarding the mathematical models that underlie the various landscape constructions. Because the landscape is really a reflection of epigenetics, it is completely determined by a system's gene regulatory dynamics: the interactions between genes, proteins, mRNA, and metabolites that control the numbers of each of these species inside a cell.

Let us make a stronger claim: it is *impossible*, at least at present, to really discuss the landscape in a meaningful way without invoking the various approaches used to mathematical model gene regulation. This is because landscapes are generally *model-dependent*: even given the same data, we might come to different conclusions about the landscape if we used different mathematical models of the underlying system.

2.1 Gene regulation is inherently stochastic

Imagine two identical cells, which have the same numbers of every type of protein. After waiting a minute or an hour, their protein numbers will almost certainly be different. This is because gene regulation is *stochastic*: identically prepared systems will evolve differently, because the physical binding, diffusion, and conformational change events underlying gene regulation are sufficiently complicated as to be essentially random.

It is tremendously important that this simple fact is taken into account in models of gene regulation. While many authors try getting away with studying gene regulation using deterministic models (because this is usually hard enough), it is often the case that including stochasticity yields qualitatively new and different behavior—for example, noise-induced transitions, oscillations, and the effective destabilization of cell states that were thought to be stable.

Here is another strong claim: *one cannot sensibly talk about the landscape without taking the stochasticity of gene regulation into account.* If gene regulation was *not* stochastic, we would be forced to return to the earlier conception of development and cell state transitions in which one imagined a cell developing in a completely predetermined way.

In any case, we would like a mathematical modeling framework that allows for random behavior. The canonical way to treat gene regulatory systems that does this, while taking other facts (like the discreteness of species numbers) into account, is to use the formalism associated with the *chemical master equation*.

2.2 The chemical master equation

From a mathematical modeler's perspective, gene regulation consists of just three things: a list of chemical species (proteins, mRNA, genes, metabolites...), a list of all the ways those species interact, and rate parameters quantifying how much or how frequently those species interact in the ways listed. For example, a simple model of an unregulated gene producing protein might look like

$$\begin{aligned} g &\xrightarrow{k_m} g + m \\ m &\xrightarrow{k_p} m + p \\ m &\xrightarrow{d_m} \varnothing \\ p &\xrightarrow{d_p} \varnothing \end{aligned}$$

where g is the number of active gene sites, m is the number of mRNA molecules, p is the number of protein molecules, and k_m, k_p, d_m, and d_p are rate parameters (corresponding to the transcription rate, translation rate, mRNA degradation rate, and protein degradation rate respectively).

In words, our list of reactions says that only four things ever happen: (i) mRNA is transcribed, (ii) mRNA is translated to produce protein, (iii) mRNA degrades, and (iv) protein degrades. There are three chemical species, four chemical reactions, and four rate parameters. Obviously, real biology is more complicated than this, but these models can be made as complicated or as simple as one would like; as long as one knows what reactions should happen, it is easy to write them down.

The stochastic dynamics of a system like this are governed by the *chemical master equation* (CME), which allows you to calculate (among other things) how the numbers of mRNA and protein in this system are likely to change with time. In principle, if a given biological system can be written down like the one above (as a list of species, reactions, and reaction parameters), the CME says *everything about its dynamics that there is to say.*

For the sake of completeness, the CME corresponding to the list of reactions above reads

$$\begin{aligned}\frac{\partial P(m,p,t)}{\partial t} =& k_m\left[P(m-1,p,t) - P(m,p,t)\right] \\ &+ k_p\left[P(m,p-1,t) - P(m,p,t)\right] \\ &+ d_m\left[(m+1)P(m+1,p,t) - mP(m,p,t)\right] \\ &+ d_p\left[(p+1)P(m,p+1,t) - pP(m,p,t)\right] \ ,\end{aligned} \tag{1}$$

where $P(m,p,t)$ is the probability that the system has m mRNA and p proteins at time t. Eq. 1 might look complicated, but it is actually fairly straightforward to write down the CME corresponding to a given list of reactions once one knows the rules. Fortunately, in practice, one does not have to literally *write down* the CME in order to work with it indirectly (for example, to simulate it using the stochastic simulation algorithm).

For a good primer on the CME and how to implement it, see Fox and Munsky's recent tutorial[2]; because this is well-trodden ground, we will not explain more about the CME here.

2.3 The chemical Langevin equation

What we *will* explain, because it is of crucial importance for landscape models, is that the CME is almost always too difficult to work with directly. Instead, it must usually be approximated such that the approximate mathematical model can be simulated on a computer in a reasonable amount of time.

One of the most commonly used approximations to the CME is the *chemical Langevin equation* (CLE), which assumes that the species in the model are sufficiently abundant that their concentrations can be reasonably approximated as continuous variables. This is not always true, of course, so mixed models—models that treat abundant species using the CLE, and species with low numbers using the CME—can be used if necessary.

If $x = (x_1, x_2, ..., x_N)$ is a vector containing the concentrations of each of the N species in the model, the CLE reads

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

where $f(x)$ and $g(x)$ are functions determined in a canonical way from the model's list of reactions (for the precise prescription, see Gillespie's seminal paper on the CLE[3]), and $\eta(t)$ is a Gaussian white noise term. Sometimes $\mu(x)$ is written instead of $f(x)$, and $\sigma(x)$ is written instead of $g(x)$; we will avoid that notation here because it suggests that the average behavior of $x(t)$ is completely determined by $\mu(x)$, which is not true in general.

Just like how one might (naively, at least) simulate a first order differential equation

$$\dot{x} = f(x) \tag{3}$$

by taking discrete time steps according to

$$x_{n+1} = x_n + f(x_n)\Delta t\ , \tag{4}$$

the way one interprets Eq. 2 is to take discrete time steps according to

$$x_{n+1} = x_n + f(x_n)\Delta t + g(x_n)\sqrt{\Delta t}\ r \tag{5}$$

which is identical to Eq. 4, except for the extra term which involves drawing a random number $r \sim \mathcal{N}(0, 1)$. Unlike in the case of first order differential equations, there exist no alternative simulation methods which are more efficient; Eq. 5 is just *the* way to interpret the CLE, a conclusion justified by Gillespie's formal derivation of the CLE from the CME[3].

2.4 Phenomenological continuous models

The biochemistry associated with most gene regulatory systems is not known in excruciating detail; indeed, in most cases, it is not known whether interacting genes even interact directly or indirectly. In these cases, it can be hard to write down a reasonable list of reactions that describes what happens in the system, which means that modeling the system using the CME may not be practical. Since the CME cannot be used, the CLE cannot be used either, because the functions $f(x)$ and $g(x)$ are completely determined by the list of reactions.

Fortunately, there is an out: one can take a phenomenological tact, and write down equations that *look* like Eq. 2 and seem to reasonably describe the system in an empirical sense, but that may not be, strictly speaking, derivable from some list of reactions. Phenomenology—the art of describing things well enough, without forcing yourself to construct a detailed microscopic model—is a fact of life for mathematical modelers working in biology, given that the experimental information one has to work with is typically sparse.

2.5 The technical meaning of gene expression 'noise'

There is a wonderful conceptual advantage to thinking about CLE models of gene expression dynamics: they provide a specific meaning to gene expression 'noise', which can be otherwise hard to define precisely. In Eq. 2, $g(x)$ is the noise term—that's it. If $g(x)$ is constant, we call the system's noise *additive*; if $g(x)$ depends on x (which, for real biological systems, is always true[3]), then we call the noise *multiplicative* or *state-dependent*.

Technically, the $g(x)$ prescribed by the CME represents *intrinsic noise*: the noise in a biological system due to the pseudorandomness of the binding, diffusion, and conformational change events that underlie gene regulation dynamics. However, extrinsic noise (external variation, perhaps due to random gene regulatory influences from a cell's environment) can be in principle included in $g(x)$ in a phenomenological way (i.e. part of $g(x)$ is determined by the model's list of reactions, and there is an additional term representing extrinsic noise that is put in by hand).

If noise is included at all (i.e. if a stochastic differential equation model is used instead of a deterministic one), it is often modeled as additive. It seems to us that this is done for simplicity. Though it is sometimes reasonable to model noise as additive (for example, if one's system remains fairly close to equilibrium), it is generally not. Just like how using a stochastic model instead of a deterministic one can yield qualitatively new different behavior, so too does multiplicative noise yield qualitative behavior which is simply different from additive noise. Some specific examples will be discussed in the next section.

3 How to think about the landscape

For most biologists—especially experimental biologists and physicians—it is more helpful to know how to *think* about the landscape than it is to know how to perform technical landscape-related calculations. The intent of this section is to explore how to think about the landscape, partly by addressing common points of confusion.

In our experience, different researchers think about the landscape differently. But some ways are more appropriate than others, for reasons we will explain.

3.1 The standard conceptual picture

Picture a three-dimensional landscape with hills, valleys, and no other distinguishing features. In the standard metaphor, one imagines dropping a ball somewhere on this landscape, with the ball's starting position corresponding somehow to our cell's initial epigenetic state (what exactly this means will be discussed in more detail later). Next, the ball will move, mostly influenced by the contours of the landscape (moving quickly down a steep hill, slowly up a steep hill, and medium speed along level ground), but also influenced by intrinsic gene expression noise, which causes the ball to jostle around randomly. In this metaphor, one might imagine intrinsic noise as wind that blows in random directions and with random intensity at different moments in time.

Eventually, the ball will settle down into a valley. Because valleys are epigenetic states that are hard to escape, we identify them with cell types or subtypes. Because it is harder to (for example) change the epigenetic state of a mature cardiomyocyte than an iPSC, more mature cell types correspond to deeper valleys, while less mature cell types correspond to shallower valleys. Moreover, if two cell types are epigenetically ‘close’ to each other (we will discuss what this means later), then their corresponding landscape valleys are also close to each other.

Once the ball falls into a valley—that is, a cell becomes a specific cell type or subtype—it does not necessarily remain there forever, especially if that valley is relatively shallow. Because intrinsic noise causes the ball to move around somewhat randomly, there is a chance that it randomly receives enough of a push to be knocked out of the valley. From there, the ball may roll back into that valley, or perhaps roll into a completely different valley.

A ball can also be knocked out of a valley when an experimenter applies some external perturbation that changes the cell’s epigenetic state (by using a drug, for example). This can either be interpreted as the ball moving, or the landscape being reshaped so that the ball is more *likely* to move; which interpretation is more appropriate will be discussed in an upcoming section.

3.2 Problems with the standard conceptual picture

There is no denying that the standard conceptual picture of the landscape is *helpful*; however, we would like to caution the reader that this way of thinking is often misleading, and that it must not be taken too seriously. In this section, we discuss several cases in which the standard metaphor might lead to a serious qualitative misunderstanding of how a stochastic biological system will behave.

To be clear, some of these issues are because it is easy to misinterpret the landscape by non-experts, rather than experts not being aware of these issues.

1. Cells do not move according to the contours of the landscape.

In Waddington’s famous illustration, one sees a ball rolling on a surface with a number of forking valleys. It is implicit in this picture that the ball will move according to these valleys. Indeed, in many qualitative discussions of the landscape, one is lead to imagine that the ball will roll according to the slopes of the landscape.

But this is just not true—not even approximately true. It is well known to landscape experts that, in addition to the ‘force’ on the ball created by the shape of the landscape itself, there is what some refer to as a ‘curl force’ which also contributes to determining its motion. It is this additional ‘force’ which creates effects like hysteresis: the shortest path from point A to point B may not be the shortest path from point B to point A. The existence of this extra influence on a cell’s motion is not a profound fact about real biology, but an artifact of trying to use a simple idea like the landscape to capture more complicated dynamics.

This means that there are really three influences on a cell’s motion through the landscape: the landscape, intrinsic noise, and this extra ‘curl force’ (which is a catch-all for all non-noise influences that the landscape’s shape does not account for).

2. Intrinsic noise is state-dependent in general.

In many discussions of the landscape, intrinsic gene expression noise is suspiciously absent, or its importance is downplayed. But landscape dynamics *do not even make sense* without noise. If there were no noise, the a ball would stay in the first valley it rolled into forever; noise-induced transitions are a fundamental feature of developmental and reprogramming biology.

When noise *is* included in discussions of the landscape, it is likened to 'wind', as we described in the previous section. But this is a poor analogy, because it leads one to liken intrinsic noise to thermal noise: something that behaves the same no matter where you are in the landscape. But we have known at least since Gillespie's seminal paper[3] on the chemical Langevin equation that noise is *always* state-dependent, except for a completely trivial system. This means that the 'wind' changes in strength depending a cell's epigenetic state (i.e. where the ball is on the landscape).

But there is an even more important point to make: noise is not a feature separate from the landscape, but a *key determinant of it.* Because noise is state-dependent, somehow modifying intrinsic noise inside a cell does not just change the motion of the ball, but fundamentally alters the shape of the landscape.

State-dependence of noise in general: [3]

1. Increasing noise doesn't necessarily make jumping between attractors more likely.

Increasing noise makes jumping between attractors more likely for symmetric additive noise. If noise is state-dependent, sometimes jumping increases, sometimes it doesn't. In general, noise is state-dependent, so this counterexample is far from pathological: one can expect complicated behavior in most realistic situations!

2. Ball's path on landscape doesn't necessarily match transition path.

True for symmetric additive noise (in one dimension?). If noise is state-dependent, generally not true. Sometimes there can even be pretty significant deviations between the two.

This is why it is important to clearly distinguish between global landscapes (which provide global relative stability information) and local landscapes (which provide local transition path information): in general, both kinds of information are not compatible.

Just like physical motion due to a potential does not in general correspond to geodesics on that potential's surface, the shortest path between two points in our stochastic system does not in general correspond to geodesics on the landscape.

3.3 What factors determine the landscape?

- noise matters
- all kinetic parameters (transcription/translation rates, decay rates, binding constants...)

- sometimes, strictly speaking, landscape only depends on a combination of kinetic parameters: for example, for a birth-death process, increasing the transcription rate and decay rate in a compensatory way does not change the steady state probability landscape

A noise parameter is not something that is added separately; rather, as Gillespie famously taught us in [CITATION CHEMICAL LANGEVIN EQN], noise is an omnipresent, and a straightforward consequence of all of the other kinetic parameters.

- reverse engineering: needs steady states/attractors for each cell type/subtype that you care about

3.4 Did the landscape change, or did the ball move?

Suppose that our cell is nestled comfortably in a deep valley, and that there is not enough intrinsic noise for the cell to move out of this valley in any reasonable length of time. We can perturb the cell by doing any number of things: for example, we can apply a drug, or adjust what the cell is being fed. If we manage to successfully kick the cell out of this deep valley via an external perturbation, how should we think about what we have done?

In one view, the landscape *did not change*, and the effect of our perturbation was simply to change the cell's epigenetic state (perhaps by adjusting the levels of important proteins up or down). In another view, the effect of our perturbation was to *remodel the landscape*, and the movement of our ball is just a consequence of that valley being destabilized. Which view is more correct?

3.5 Is there just one landscape, or are there many?

Some researchers have speculated that it is possible to define *the* landscape for a given biological system: a complicated surface that, if sufficiently analyzed, could in principle answer all questions regarding that system's gene expression dynamics. In particular, it should be able to answer the following questions:

1. What cell type is this iPSC most likely to eventually become? How much more likely is it to become cell type A than cell type B?

2. Which way is the ball most likely to roll through the landscape? In other words, what will its *transition path* through the landscape look like?

It turns out that, in general, it is not possible to answer both questions with the same landscape. The reason is somewhat mundane: answering questions about the probability a cell *eventually* becomes one cell type or another corresponds to asking about the steady state probability distribution, while answering questions about cell state transitions that happen in a *finite* amount of time corresponds to asking about finite time transition probability distributions. In general, these distributions can look very different.

To make this clearer, let us consider a specific example: imagine the one-dimensional system with five attractors ...[FINISH EXAMPLE LATER, OR MAYBE PUT IN A DIFFERENT SECTION]

With that, we have made the point that there are 'global' landscapes, which contain global relative stability information ("Which cell type is my iPSC *eventually* most likely to become?”), and ‘local’ landscapes, which contain local cell state transition information (“Which way is the cell most likely to go in the next minute or hour?”). Are these landscapes unique?

Neither global nor local landscapes are unique. Both kinds of landscapes can be freely ‘stretched’ or ‘compressed’[SEE FIG] so long as the *relative* height of each point relative to all others remains the same.

In light of the previous fact, it is probably not surprising that landscapes do not uniquely determine the underlying dynamics either—in other words, it is possible for different systems to have the same landscape. As a specific example, consider a 1D system governed by a potential V with additive noise, so that

$$\dot{x} = -V'(x) + \sigma\eta(t) \ .$$

The steady state distribution is easy to calculate exactly in this case (if one makes the reasonable and biological relevant assumption that P_{ss} and its derivatives vanish at infinity; see APPENDIX for more details); it is

$$P_{ss}(x) = N \exp\left(-\frac{V(x)}{\sigma^2/2}\right)$$

where N is a normalization constant. But notice that the system with dynamics governed by

$$\dot{x} = -kV'(x) + \sqrt{k}\sigma\eta(t) \ .$$

for some constant $k > 0$ will have the exact same distribution, since the factors of k will cancel.

Put differently, making the hills and valleys of our potential bigger or smaller will not change the landscape, so long as the noise experienced by the cell is scaled in a similar way.

4 Landscape properties

Having discussed the basics of mathematical models of gene regulation, as well as how to think about landscapes qualitatively, in the section we hone in on the specifics of what information we would like a landscape to contain. We will discuss how, in general, the different kinds of information we might like a landscape to contain are not compatible; this will lead us to think of different landscapes as answers to different questions, rather than imagining that there is one ‘true’ landscape for a given system.

4.1 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).

4.2 Global vs local landscapes

It's true that, given enough time, a cell is overwhelmingly likely to go *way* over there. But what if we mainly care about where it will go in the next minute, or the next hour?

As you might imagine, where the cell is more or less likely to go strongly depends on the timescale we are interested in. For example, consider the system depicted in FIGURE: there are five attractors, each one deeper than the previous one. If a cell starts in the leftmost attractor, it might spend its first few minutes around the first two attractors. But, in the next hour, it is overwhelmingly likely to reach the third attractor. In the next 10 hours, it might be overwhelmingly likely to hit the fourth attractor. And so on.

Things a landscape can capture (but does not necessarily capture)

- Can offer estimate of probability of transition between two stable cell types (not true for local quasipotential, except if quasipotential is defined wrt one of the stable cell types in question) - Can say something about the path of transitions between two cell types (not true for global quasipotential/not strictly true for Wangs P_ss) - Make the point that it might not be possible for ANY landscape to incorporate accurate transition path information and steady state relative occupancy information at the same time; the landscape you want to construct may just depend on the question you are asking.

- Two definitions given, since there are (broadly speaking) two types of landscapes: local and global - Local landscape: provides accurate transition path information (but not relative stability information), satisfies certain other properties - Global landscape: provides accurate relative stability information (but not transition path information), satisfies certain other properties - Will describe below, for each landscape construction, how they satisfy our definition.

4.3 Mathematical definition

This section is intended for landscape experts comfortable with some of the mathematics. Feel free to skip this section: all of its important qualitative content will be explained in the next section through a series of examples.
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 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.

4.4 Intuition for mathematical definition

5 Different landscapes

5.1 Simplest model: Markov chain

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

$$\phi_G(i) := -\log(P_{ss}(i)) \tag{6}$$

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

$$\phi_L(i) := -\log(1 + p_{ji}) \ . \tag{7}$$

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

5.2 Steady-state probability landscape

Wang 2008 paper (first paper discussing this?): [7]

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

$$dx_i = f_i(x_1, ..., x_N)dt + g_i(x_1, ..., x_N)dW_i \ , \tag{8}$$

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.

[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[5], 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.

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

$$0 = \sum_{i=1}^{N} -\frac{\partial}{\partial x_i}\left[f_i(x)P(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x_i^2}\left[g_i(x)^2 P(x,t)\right] \ . \tag{9}$$

In other words, in practice, Eq. 9 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.

5.3 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.

$$\phi^{QP}(x; x_0) := \inf_{T>0} \inf_{\text{paths} X(t)} \int_0^T L(X(t), \dot{X}(t)) \ dt \tag{10}$$

5.4 Vector decomposition landscapes

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

$$dx_i = f_i(x)dt + \sigma dW_i \ , \tag{11}$$

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

$$\sum_{i=1}^{N} \frac{\partial U^{\text{norm}}}{\partial x_i}(x) \left(f_i(x) + \frac{\partial U^{\text{norm}}}{\partial x_i}(x) \right) = 0 \ . \tag{12}$$

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.

$$\mathbf{f}(x) \neq -\nabla U(x) \tag{13}$$

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

$$\mathbf{f}(x) = -\nabla U(x) + \mathbf{F}_R(x) \tag{14}$$

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. 9 and substitute in the WKB ansatz

$$P_{ss}(x) = \exp\left[-\frac{\phi(x)}{(\sigma^2/2)} + \phi_0(x) + \phi_1(x)(\sigma^2/2) + \cdots \right] \tag{15}$$

to find the equation

$$0 = \sum_{i=1}^{N} \ \frac{2}{\sigma^2} \frac{\partial \phi(x)}{\partial x_i} \left(f_i(x) + \frac{\partial \phi(x)}{\partial x_i} \right) - \frac{\partial f_i(x)}{\partial x_i} - \frac{\partial^2 \phi(x)}{\partial x_i^2} \ , \tag{16}$$

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

$$0 = \sum_{i=1}^{N} \ \frac{2}{\sigma^2} \frac{\partial \phi(x)}{\partial x_i} \left(f_i(x) + \frac{\partial \phi(x)}{\partial x_i} \right) \ , \tag{17}$$

which is the same as Eq. 12 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.

6 Beyond the landscape: modeling cell state transitions

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 in an amount of time T. What path through epigenetic state space is most likely to take the cell from state A to state B in time T?

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, the entire promise of the landscape is that its hills and valleys should indicate to us, hopefully in a visually intuitive way, which paths are 'best' for achieving a given differentiation protocol.

Ironically, landscapes are not really the best way to answer this question. Here's why: as we discussed earlier, global landscapes contain no information about transition paths. Local landscapes *do* contain information about transition paths, but in a roundabout way which depends on the Lagrangian formalism we are about to describe. It does not really make sense to use the Lagrangian formalism to construct a local landscape associated with a transition of interest, then use that landscape to extract information about the transition; instead, one can just use the Lagrangian formalism directly.

6.1 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

$$\frac{\partial L}{\partial x_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) \tag{18}$$

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.

6.2 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. 8, it turns out that the Lagrangian

$$L = \sum_{i=1}^{N} \frac{\left[\dot{x}_i - f_i(x)\right]^2}{2 g_i(x)^2} \tag{19}$$

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.

6.3 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. 19 is derivable from the CLE (Eq. 2); 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.

7 Applications

7.1 Landscapes in reprogramming

7.2 Landscapes in developmental biology

7.3 Landscapes in cancer biology

8 Pitfalls, issues, and directions for future research

Baez stochastic mechanics[1]

Weber stochastic path integrals[9]

Wang's Waddington landscape transition path stuff[8]

8.1 Dimensional reduction

8.2 Sparseness of available data

8.3 Computational burden

9 Conclusion

References

[1] J. C. Baez and J. Biamonte. Quantum Techniques for Stochastic Mechanics. *ArXiv e-prints*, September 2012.

[2] Z. Fox and B. Munsky. Stochasticity or Noise in Biochemical Reactions. *ArXiv e-prints*, August 2017.

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

[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] 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.

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

[7] 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.

[8] J. Wang, K. Zhang, L. Xu, and E. Wang. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, 108(20):8257–8262, 2011.

[9] Markus F. Weber and Erwin Frey. Master equations and the theory of stochastic path integrals. 2016.

[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.