9/13/18

**TENTATIVE TITLE:**

From developmental metaphor to quantitative framework:

a review of Waddington landscapes and their applications

**Abstract: (I added a question mark for grammar)**

Waddington’s epigenetic “landscape” was originally a metaphor for conceptualizing the link between gene regulation dynamics and developmental processes. In recent times however it has become a powerful theoretical framework to map the lineage dynamics of cells in a wide array of contexts, developmental and otherwise. In particular, it has become a powerful method for interpreting the dynamics of model systems describing stem cell dynamics and has the potential to help improve our understanding and control of directed cell differentiation. But how does one quantify the “landscape”? What does the term landscape even mean in a precise quantitative sense? And how can that landscape be used to improve our understanding and control of cells’ states**?** Here we review different theoretical formulations / interpretations of the “landscape” concept (e.g. Fokker-Planck, quasi-potential, etc. approaches) as well as methods for charting transition paths within them (e.g. least action paths). We also discuss practical issues that must be addressed if these methods are to become predictive, including reconciling the differences between different landscape interpretations, understanding the sensitivity of these landscapes to different types of stochastic noise, the computational burden of landscape construction in high dimensional systems, the need for effective dimensional reduction, and the need to integrate theory with data in these constructions.

**FIGURES**

**Cartoon depiction of the landscape idea**

- It’d be nice if we didn’t reuse the same image everyone else does; nonetheless, some sort of picture like this is a given.

**Landscape comparison via directed network**

- Each landscape is a node. Arrows display relationships between landscapes. Ex: an arrow between the FW large deviation theory global quasipotential and the Wang P\_ss landscape to indicate the former reduces to the latter in the small noise limit.

**Direct comparison of landscape for (one or more) concrete example(s)**

- Plot landscapes for same system (possibly superimposed, or side by side) to see if there are any qualitative differences

**TABLES**

**Landscape comparison**

- One column: landscape candidate

- Column: Definition of candidate (i.e. phi = - ln(P\_ss) )

- Column: Global relative occupancy info? Y/N

- Column: Local transition path info? Y/N

- Column: Computational cost low/med/high (discussed in more detail in text)

- Column: Notes (i.e. this candidate not defined for discrete systems, this candidate is not mathematically guaranteed to exist in general)

- Column: Sample references

**INTRODUCTION**

**The absolute basics**

* Begin by (briefly) explaining Waddington’s epigenetic landscape metaphor: a cell is like a ball rolling around in a rugged landscape, until it eventually settles down in a valley/stable cell type. Don’t dwell on this too much, since it’s been explained in exhaustive detail elsewhere.
* Provide at least one example from developmental biology (possibly stem cell to TE or ICM), one example from cancer biology (cancer cell types are normally unused attractors), and an example from a different context (direct reprogramming would be helpful to mention).
* Mention nice properties of the metaphor: for example, it is a way to imagine one mature cell type being directly reprogrammed to another mature cell type; people used to think such a thing was impossible.
* Possibly (briefly) mention that the idea of a landscape is useful in other areas: population biology, protein folding, evolutionary biology.

**What is the landscape?**

* Explain how, despite the nice properties of the metaphor, it is not yet clear how to make it quantitatively precise. Several groups have proposed their own constructions of the landscape, each intended to capture different information/be valid in different circumstances. It is not even clear if, in some sense, there exists a ‘correct’ notion of the landscape.
* Nonetheless, all that matters if that the concept is useful. In this paper we will explain various formulations of the landscape, say when each version is useful, and offer many examples.

**The landscape in practice**

* Most theoretical approaches depend somewhat upon the availability of detailed mathematical models. This is, obviously, a major limitation that will eventually need to be overcome if landscapes are to become more useful.
* We will talk more about trying to construct landscapes from multi-omic information (transcriptomics, proteomics, metabolomics, etc) briefly later in this paper. For now, just keep the caveat in mind that some of these ideas are hard to apply directly at present.

**GENERIC PROPERTIES OF THE LANDSCAPE**

**The landscape for equilibrium systems**

* Explain what a potential tells you in an equilibrium system (probability of transitions between states, relative occupancy, first passage time, etc). Explain why V generally doesn’t exist for non-equilibrium systems.
* Not sure where to put this section. Can go somewhere else.

**Modeling a generic biological system**

* A generic biological system consists of a list of reactions (for example, gene produces mRNA, mRNA produces protein) along with rate constants for each reaction. Stochastic dynamics are determined by the chemical master equation, whose dynamics can be approximated via a Langevin equation in some reasonable circumstances.
* A generic system’s state consists of a tuple of numbers: generally the numbers of each protein/mRNA/metabolite/etc in the model. In the continuous dynamics limit, concentrations may be used instead of absolute numbers.
* Assuming the temperature/pressure/other external variables to be roughly constant in many situations of interest, all rate constants are genuine constants. This means that the regulatory network (species and their directed interactions) remains constant in time, so that a landscape should *also remain constant in time*. If the landscape does look like it is changing, that is because there is a missing species, or possibly one is looking in a reduced number of dimensions.
* Generally spatial information is ignored, or treated via compartments. For us, the CME is the ‘fundamental’ level.
* Make the point that, since stochastic information is qualitatively important, landscape candidates must incorporate noise information. Moreover, noise is the primary mechanism by which a cell navigates between stable attractors.

**Properties of a generic landscape**

* Should be agnostic to the underlying kind of model (whether SDE or CME or something else). In other words, we should be able to construct a landscape whether the system is discrete, or continuous, or some combination.
* One minimum/valley per each stable state/attractor/cell type. This is hard to guarantee mathematically in general, but is often borne out in applications.
* Should assign each state in state space a ‘height’. In other words, the landscape is a function f: S -> R.
* Should be bounded from below (i.e. there are no states which are impossible to escape). Along with the above point, suggests we can limit ourselves to functions f: S -> R+ since we can always shift the minimum height to 0.
* Should exist mathematically in fairly general circumstances (Yuan, Tang, Ao claim that P\_ss doesn’t exist in general, limiting its utility as a landscape; is this true for CME-derived systems, or only for ad-hoc systems?)
* Ideally, should involve few or none arbitrary choices; otherwise, those arbitrary choices shouldn’t effect qualitative characteristics
* Need not be continuous; can be discrete (simplest example: Markov chain model)

**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 Wang’s 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.

**DIFFERENT LANDSCAPES**

**Simplest model: Markov chain**

* Roughly agrees with the potential (possibly the potential integrated over some region of state space) for equilibrium systems.
* Each stable cell type is a node. Each possible (reasonable) directed transition between stable cell types is represented by an arrow, and has an associated rate. (Some transitions may be so rare that we just suppose they are impossible and don’t draw the arrow)
* Generally steady state exists: landscape height is the probability of occupancy at steady state.
* This kind of landscape indicates the likelihood of transitions between stable cell types, as well as relative occupancy at steady state. (for an even simpler model, just assign each stable cell type with the ss probability of occupancy; in this case, no transition information is included)
* Problem: this model is particularly coarse. Reasonable to generate from empirical data, but probably hard to generate from DE/PDE models.
* Most popular kind of landscape used in practice, although this discrete kind of model is rarely identified as distinct from the other types of landscapes. I think it’s good to point it out explicitly.

**Steady-state Fokker-Planck landscape**

* Agrees with the potential for equilibrium systems.
* Should note that this landscape formulation does not actually depend on the system being expressed in Langevin/Fokker-Planck form; can be a CME steady state, only difference is that landscape heights are assigned to a discrete rather than continuous state space.
* Say something about interpretation. Does tell you: relative occupancies. Possibly does not tell you: transition path information. (I am waffling on this because I think it does say something about transition paths in the small noise limit; see Zhou and Li 2015, 10.1063/1.4943096)
* Actually calculating in a high-dimensional (d > ~3) system is a nightmare. Can use Monte Carlo methods, Gaussian ansatz (see Peng et al 2017, 10.1186/s12918-017-0429-x), and various mean field approximations (like “proteomic mean field theory” by Walczak et al 2005, 10.1529/biophysj.104.050666).

**Freidlin-Wentzell/stochastic path integral-derived local and global quasipotentials**

* Agrees with the potential for equilibrium systems
* Briefly say that CME/Fokker-Planck dynamics are equivalent to a stochastic path integral. From this path integral one can derive an effective Lagrangian/Hamiltonian (sometimes called the Onsager-Machlup function in the continuous limit).
* Can use action/Lagrangian/Hamiltonian to get accurate information about most likely transition path, which can be used to build local quasipotential. Local quasipotential says how likely it is to go from one fixed state to any other state (i.e. it involves the arbitrary choice of a reference point). Global quasipotential, obtained from suitably gluing together local quasipotentials, says something about global relative occupancy, but detailed transition information is lost. (This is partially due to transitivity not generally being true. In other words, if P(1->2) > P(2->1) and P(2->3) > P(3->2), not necessarily true that P(1->3) > P(3->1). See Zhou and Li 2015, 10.1063/1.4943096)
* Global quasipotential agrees with Wang landscape in small noise limit. See Zhou and Li 2015, 10.1063/1.4943096
* Mention something about Lagrangian ‘ambiguity’ and justify our choice.

**A-type integral and SDE decomposition**

* Apparently the math is somewhat controversial. Zhou and Li (2015, 10.1063/1.4943096) claim it doesn’t exist in general. Yuan et al (2016, 10.1063/1.4964681) disagree. Zhou and Li disagree with their disagreement (2016, 10.1063/1.4964682).
* In any case, agrees with previous two quasipotentials in certain circumstances, subject to some disclaimers. See Zhou and Li (2015, 10.1063/1.4943096)

**Vector decomposition landscapes (Zhou et al 2012 RSIF, 10.1098/rsif.2012.0434)**

* Agrees with the potential for equilibrium systems.
* Huang’s approach assumes additive noise and is purely deterministic. In principle, this does not matter, since noise information (via the Gillespie prescription) is derived from deterministic dynamics information for fully specified systems. But still…
* Not sure to what extent agrees with other approaches (like the SS prob landscape).

**Possibly more landscapes go here**

**APPLICATIONS**

**General comments on how landscapes have been applied**

**Landscapes in developmental bio**

**Landscapes in reprogramming bio**

* Lang et al 2014 (10.1371/journal.pcbi.1003734)

**Landscapes in cancer bio**

* All the work on EMT…
  + Li et al 2016 (10.1039/c6cp03174a)

**THE LANDSCAPE IN PRACTICE**

**General comments on applying landscapes yourself**

* Good point you made: one should keep in mind that constructing the landscape is not always the end goal. Often, the landscape is intended as a means to understand a certain question (ex: How does the typical cell move through state space as it transitions from a fibroblast to a cardiac cell?) Some questions might be better off answered without the landscape, like transition path-related issues (which can be addressed directly using the stochastic path integral formalism).

**Parameter/reaction inference issues**

**Poor time/space resolution of multi-omic measurements**

**Dimensional reduction**

**Computation burden of landscape construction**

* Curse of dimensionality: solving for P\_ss, for example, in 1D is not bad. But solving for P\_ss in 10D is exponentially harder! How does one get around this? Probably no way without dirty numerical tricks (as above, Monte Carlo, or mean field theory).

**CONCLUSION**

* The epigenetic landscape is an influential idea with various quantitative implementations; not clear if there is a ‘best’ one, and it seems that different choices are more or less useful in different situations.
* We covered these landscapes: \_\_, \_\_, \_\_, \_\_. We identified known connections between them.
* People have used landscapes in dev bio, reprogramming, cancer bio, etc.
* Still much work to be done to make landscape construction numerically realistic for larger (species # > 10) systems.
* In some cases may be prudent to not use a landscape at all.