Reconstructing equations of motion for cell phenotypic transitions: integration of data science and dynamical systems theory

Jianhua Xing¹,²,³,*

¹ Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15232, USA.
² Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15232, USA.
³ UPMC-Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA.

* To whom correspondence should be addressed. Email: xing1@pitt.edu

Abstract

Dynamical systems theory has long been fruitfully applied to describe cellular processes, while a main challenge is lack of quantitative information for constraining models. Advances of quantitative approaches, especially single cell techniques, have accelerated the emergence of a new direction of reconstructing the equations of motion of a cellular system from quantitative single cell data, thus places studies under the framework of dynamical systems theories, as compared to the currently dominant statistics-based approaches. Here I review a selected number of recent studies using live- and fixed- cell data, and provide my perspective on the future development.

Introduction

The Danish astronomer Tycho Brahe spent his entire career recording the positions of celestial bodies. From his large data set, Johannes Kepler discovered a set of empirical laws of planetary motion, and later Issac Newton developed a theoretical framework of classical mechanics. Within the framework the solar system is mathematically represented as a dynamical system, in which a set of equations of motion, specified by the Newton’s equation here, describe how each object moves over time. The equations of motion allow one to predict phenomena such as solar eclipse, and design satellite orbits, etc., beyond the original system Tycho studied. Dynamical systems theory is the foundation of many fields of science and engineering.

A cell is a complex system composed of a large number of molecular species that interact with each other to form a regulatory network. From a biology perspective, a cell receives and processes extracellular and intracellular information then responds accordingly through a complex regulatory network. Extensive efforts have been made to decipher how a regulatory network controls cellular dynamics, especially cell phenotypes (1-4). Shared with the same set of genome, a cell can exist in different cell phenotypes with distinct physical and biological features.

From a physics perspective, a cell is a dynamical system and the temporal evolution of the cell state can be described by a set of dynamical equations. A central step in mathematical
modeling studies of cell regulatory networks is to construct the equation of motion, and the efforts are typically restricted to small networks (1, 2). However, in reality such simple networks are embedded in a much larger densely-connected network, and it is highly nontrivial to reconstruct the dynamics of such a network. Determining the corresponding equations of motion of such a network is the Holy Grail in systems biology for our understanding and manipulation of cellular processes. Knowing the equations of motion will help on addressing some fundamental questions in cell and developmental biology, and is of high significance to biomedical research. Although efforts have been made to perform whole-cell simulations of bacteria(5, 6), it remains a grand challenge to reconstruct similar models for experimental data.

Recent years advances of single cell techniques, both for fixed and live cells, facilitate generation of a large body of single cell data (7-11). Together with advances of machine-learning-based computational approach development, an emergent research direction is to extract dynamical and mechanistic information from “big data” for cellular processes. Here my goal is to provide a brief overview of some recent developments, and my personal perspective.

Formulation of cell phenotypic transitions in the context of dynamical systems theory

Let’s first formulate the problem of dynamics reconstruction from data in a general form. For a dynamical system, let’s assume that a state vector \( \mathbf{z} \) completely specifies the internal cellular state at a given time. Mathematically here completeness of the state representation means that one can write down a set of memory-less Langevin equations that govern the dynamics of individual cells,

\[
d\mathbf{z}/dt = A(\mathbf{z}, \lambda) + \zeta(\mathbf{z}, t),
\]

where \( A \) is a governing vector field containing information on how system components interact and act on each other at a given time, \( \lambda \) refers to controlling parameters and extracellular environmental factors that in general can be time-varying, and \( \zeta \) are white Gaussian noises with zero mean to account for fast dynamics of degrees of freedom not explicitly described. This equation alone does not describe cell proliferation and death. Therefore there are two basic questions for reconstructing the dynamics of a system: what is the appropriate choice of the dynamical variable \( \mathbf{z} \), and how to get the function \( A \) and \( \zeta \)?

For a Newtonian system, \( \mathbf{z} \) refers to the position and momentum of composing objects. For a cellular system, in principle one can use the compartmentized concentrations (or amount) of all intracellular species. In practice, however, one only knows information of a subset of the species. For example, with single cell RNA-seq data one can represent a cell state as a vector in the genome-wide transcriptomic expression space. Alternatively, collective cell features, such as morphological and cell texture features, have been widely used for phenotyping cells(12-18). Accumulating studies demonstrate that cell morphological features faithfully reflect the underlying cell expression states. Barkal et al. showed that cell morphological features define the corresponding regulatory network (19). Yin et al. showed with multiple cell lines that gene regulation lead to switch-like transitions between discrete cell shapes (16). Gordonov et al. developed an experimental-computational framework to characterize cell morphological changes in response to drug treatments (12). Wu et al. demonstrated that cells with different morphology have distinct expression profiles and metastasis potential (20). While in most studies these composite cell features are treated as statistical quantities for clustering cells, recent studies have treated them as dynamical variables to specify cell states (12, 21-25), analogous to using collective modes in condensed matter physics. In all cases one quantifies
measurements, either sequencing reads or images, into a vector representing cell state in a multi-dimensional state space. In later discussions, we will use \( \mathbf{z} \) for a generic cell state, \( \mathbf{x} \) in the expression space, and \( \mathbf{y} \) in the composite cell feature space.

For a Newtonian system, \( \mathbf{A}(\mathbf{z}) \) is given by the Newton’s equations. One prominent example in biology is molecular dynamics simulations of macromolecule by simulating the Newton’s equations or Langevin equations. For a gene regulatory network, \( \mathbf{A}(\mathbf{z}) \) gives quantitative information about gene regulation. For cell feature representation, \( \mathbf{A}(\mathbf{z}) \) gives coupling between different cell features, which reflects the underlying interactions of involved cellular species. To find a numerical approach of reconstructing \( \mathbf{A}(\mathbf{z}) \), let us examine the Fokker-Planck equation corresponding to the above Langevin equations (with the assumption that \( \lambda \) is time-invariant for simplicity),

\[
\frac{\partial n(\mathbf{z}, t)}{\partial t} = \nabla \cdot \left[ -\mathbf{J}(\mathbf{z}) n(\mathbf{z}, t) + D \cdot \nabla n(\mathbf{z}, t) \right] + \int d\mathbf{z}' k(\mathbf{z}, \mathbf{z}') n(\mathbf{z}', t),
\]

where \( n(\mathbf{z}, t) \) is the number density, \( D \) the diffusion constant matrix related to \( \zeta \) through the Einstein relation (26). The last nonlocal term accounts for discrete changes due to cell division, and its diagonal terms also include contribution from cell death. One expects that the function \( k(\mathbf{z}, \mathbf{z}') \) is peaked in restricted regions populated with cells right before and after division. The term \( \mathbf{J} \) is related to the vector field under the Stratonovich interpretation of the stochastic differential equations (26, 27) as,

\[
\mathbf{J}(\mathbf{z}) = \frac{\int d\mathbf{z}' \rho(\mathbf{z}', t) \frac{d\mathbf{z}'(t)}{dt} \delta(\mathbf{z} - \mathbf{z}'(t))}{\int d\mathbf{z}' \rho(\mathbf{z}', t) \delta(\mathbf{z} - \mathbf{z}'(t))} = \mathbf{A}(\mathbf{z}).
\]

Here \( d\mathbf{z}/dt \) is the instant velocity of a cell moves in the state space (excluding cell proliferation and death). The above relation of \( \mathbf{J} \) immediately suggests one algorithm to reconstruct \( \mathbf{A}(\mathbf{z}) \) from measured \( (\mathbf{z}, d\mathbf{z}/dt) \). A limitation of this algorithm is that it requires information of \( d\mathbf{z}/dt \), which can be challenging for fixed cell data (but see below). In practice one typically replaces the Dirac’s \( \delta \) function with a finite volume bin and performed the average over \( d\mathbf{z}/dt \). Notice that \( n(\mathbf{z}, t) \) is expected to be a slowly-varying function over the scale of the bin size, so the evaluation of \( \mathbf{J}(\mathbf{z}) \) is insensitive to the sampling of \( n(\mathbf{z}, t) \). This is an attractive feature as compared to the algorithm discussed below.

Another possible algorithm is to regress the equations from a series of measurements of \( n(\mathbf{z}, t) \) at different time points. This algorithm does not require information of \( d\mathbf{z}/dt \), but is sensitive to the sampling of \( n(\mathbf{z}, t) \). A further technical challenge is that in most single cell experiments one typically measures the cell probability density \( \rho(\mathbf{z}, t) = n(\mathbf{z}, t) / N(t) \) with \( N \) being the total cell number instead of the number (or concentration) density. Then, when \( N(t) \) changes with time, unless one also measure \( N_0 \) to obtain \( n(\mathbf{z}, t) \), the partial differential equation of \( \rho(\mathbf{z}, t) \) contains nonlinear terms. For the convenience of seeing how the nonlinear terms arise, and also for practical applications, let us examine a discretized version of Equation 2. We divide the coordinate space into sufficiently small regions, and one can convert the partial differential equations into a set of master equations,
\[ \frac{dn}{dt} = \sum_{ij} \left( -k_{ij}n_i + k_{ji}n_j \right) + \left( -\alpha n_i + \sum_{ji} \beta_{ji}n_j \right) - d_n n_i \quad (4) \]

or (see also (28)),

\[ \frac{dp_i}{dt} = \sum_{ij} \left( -k_{ij}p_i + k_{ji}p_j \right) + \left( -\alpha p_i + \sum_{ji} \beta_{ji}p_j \right) - d_p p_i + \sum_j \left( a_j - d_j \right) p_j \quad (5) \]

where \( n_i \) and \( p_i \) are the number of cells and the fraction of cells in region \( i \) of the state space, respectively. The terms in the first parenthesis are the flux exchange between two neighboring regions due to cell growth or cellular responses to stimuli. The terms in the second parenthesis are due to jumps between often non-neighboring regions in the state space due to cell division, and a relation \( 2\alpha = \sum \beta \) holds to account for the fact that one mother cell divides into two daughter cells. The third term accounts for change of \( n_i \) from cell death. Notice that the last term of Eqn. 5 is nonlinear in \( p \).

**Extraction of dynamical information from snapshot data**

Given the destructive nature of many single cell techniques, continuous attempts have been made on extracting dynamical information from snapshot data. In this section I will discuss a few recent developments.

A class of methods use the ergodic principles (29). The approaches are based on some basic assumptions: 1) one measures the stationary distribution of the underlying system; 2) the average of the dynamics of one cell over time is equivalent to the average over a population of cells at one time point. One example is the ergodic rate analysis (ERA) method of Kafri et al. (30). They applied ERA to analyze cell size regulation of mammalian cells under the exponential growth condition. Under this condition the total cell number of the population grows exponentially, \( N_t = N_0 \exp(\alpha t) \), and the growth rate were deduced from measuring \( N_t \) at different time points followed by exponential fitting. Then they measured the stationary probability density distribution the two-dimensional state space defined by DNA content and the activity of the anaphase-promoting complex (APC) in its active form using fluorescent staining. To simplify the ERA analysis, they further projected the two-dimensional stationary distribution to an average one-dimensional cell cycle axis describing cell cycle progress through \( G_1 \rightarrow G_2/S \rightarrow G_2 \) then back to \( G_1 \) through cell division (Fig. 1). Divide the cell cycle axis into discrete regions, with \( p_i \) the measured percentage of cells in region \( i \), and \( P_i \) the cumulative cell fraction, both can be obtained from the fixed cell data. Then the flux balance condition leads to

\[ \frac{dN_i}{dt} = \alpha P_i N_i = Flux_0 - Flux_i = 2\alpha N_i - \omega_i p_i N_i, \quad \text{so} \quad \omega_i = \frac{\alpha}{p_i} (2 - \frac{P_i}{P}) \].

The factor 2 in \( Flux_0 \) comes from that one mother cell divides into two daughter cells. From such ERA analyses Kafri et al. identified that cell size feeds back onto cell cycle regulation.
Figure 1 Ergodic rate analyses on the dynamics of cell cycle progression. Adapted from (30).

Weinreb et al. applied similar idea to analyze scRNA-seq data and termed their method population balance analysis (PBA) (31). Here again one assumes that the system is under steady state. Since the dimensionality of the state space is typically high, it is impractical to use regular grids. Instead they used spectral graph theory to represent the multi-dimensional Fokker-Planck equations on the irregular grids formed by the data points. Furthermore, to simplify the determination of the function form of $A$, they assume that it can be written as the gradient of a scalar potential, $A = \nabla \phi$. Rigorously speaking this assumption is not complete for describing the dynamics of a system without detailed balance such as a cellular system. Therefore, Weinreb et al. performed systematic analyses on the limitations of extracting dynamical information from snapshot data under this framework. Furthermore, Klein and coworkers showed combining snapshot data and lineage tracing provides a more reliable strategy for reconstructing the genome-wide transcriptional trajectories (32, 33).

With time series measurements using fixed cells, several studies reconstructed cellular dynamics without the steady-state requirement. Yeo et al. introduced a generative model called PRESCIENT (Potential eneRgy undErlying Single Cell gradIENTs) to learn differentiation landscape from time-series scRNA-seq data (34). They also assume that $A = \nabla \phi$, and use deep learning to determine the function form of $\phi$ by fitting the calculated $n(z,t)$ to the measured ones. With $\phi$ and the governing dynamical equations, one has a generative model and can perform in silico perturbation studies. In another set of applications, one clusters cell based on expression or other cell features obtained through sequencing, immunostaining, imaging, etc, and measures how the numbers (i.e., $n_i$ in Eqn. 4) or fractions (i.e., $p_i$ in Eqn. 5) of individual subpopulations change over time. Then one fits the data with a model of transitions between these discrete states, with the Markovian dynamics often (but not always) assumed. Some examples include spontaneous transitions among cancer cell subpopulations (35, 36), and induced cell phenotypic transitions through intermediate subphenotypes (25, 37-39).
While all the above approaches are based on the measured distributions, Qiu et al. developed a framework to learn a set of high-dimensional analytical vector field functions from sparse, noisy, and discrete scRNA-seq datasets (40). To explain the basic idea, let’s start with the original RNA velocity estimation approach (Fig. 2A), which estimates instant RNA transcriptional rates from the spliced and unspliced RNA reads in an scRNA-seq experiment (41). The conventional RNA velocity method from the original paper exploits the kinetics of RNA transcription, splicing, and degradation with corresponding ordinary differential equations (ODEs) as follows:

\[
\frac{du_i}{dt} = \alpha_i - \beta_i u_i, \quad \frac{ds_i}{dt} = \beta_i u_i - \gamma_i s_i,
\]

where \( u_i \) and \( s_i \) are the copies of unspliced and spliced RNA for a particular gene \( i \) in a cell, respectively; \( \alpha_i, \beta_i, \) and \( \gamma_i \) are the rate constants for transcriptional, splicing, and degradation (see Fig. 2A). Notice that for each gene one can read out \( u_i \) and \( s_i \) directly from scRNA-seq data. By taking \( \beta_i = 1 \) (so all other rate constants are values relative to splicing instead of absolute values), and finding ways to estimate \( \gamma_i \), La Manno et al. showed that one can estimate RNA velocity for this gene, \( ds/dt = u_i - \gamma_i s_i \). Notice that the velocity depends on both \( u_i \) and \( s_i \). A positive velocity means that this gene is under upregulation, while a negative value means downregulation. Therefore for a cell at a given expression state (represented by \( s \), a vector with each component representing the copy number of spliced RNA of a gene), another velocity vector \( v = ds/dt \) gives the direction this cell moves in the expression state space. The splicing-based RNA velocity estimation has several inherent limitation on its accuracy. Recently Qiu et al. generalized RNA velocity estimation to scRNA-seq data with RNA metabolic labeling (40). RNA metabolic labeling treats live cells with certain chemicals to incorporate modified ribonucleosides, such as 5′-Bromouridine (BrU) or 4-thiouridine(4sU), into nascent RNA transcripts. At the sequencing stage, the modified ribonucleosides induce ribonucleoside exchange with certain probabilities that can be captured through sequencing. From the amount of nascent and old (unlabeled) transcripts one can infer genome-wide RNA turnover dynamics (42-46), then one can estimate RNA velocities with improved accuracy over the splicing-based method.

A subtle but important conceptual advance of Qiu et al. is to realize that the rate “constants” \( \alpha_i, \beta_i, \) and \( \gamma_i \) are really functions of concentrations of molecular species in the cell. For example, \( \alpha_i \) is a function of the concentrations of transcription factor proteins that regulate gene \( i \). Therefore, let’s make the following assumptions: 1) the state of a cell can be specified by the current transcriptomic state, say the spliced RNA vector \( s \); 2) the transcriptional regulation is a function of \( s \). The first assumption is made when we only have transcriptomic data, and one can relax it when multi-omics data become available. The second assumption is due to the nature of snapshot data, and relaxing it would require true live-cell time series data discussed below. Then the unspliced mRNA vector \( u = u(s) \) is also a function of \( s \), and one reaches that,

\[
\frac{ds_i}{dt} = \beta_i u_i(s) - \gamma_i s_i = F_i(s),
\]

where the function \( F \) forms a to-be-determined genome-wide, generally nonlinear vector field in the expression state space defined by \( s \), and it contains quantitative gene-gene regulation formation.
Qiu et al. developed a machine learning procedure, called dynamo, to reconstruct the $F$ function from scRNA-seq data (Fig. 2B) (40). Given the scarcity of the data and the high dimensionality of the system, it is numerically challenging to apply Eqn. 3 directly. Instead, from measured $x$ and $dx/dt$, one learns the function form of $F$ that maps $x$ to $dx/dt$ as a regularized optimization problem in an infinite dimensional function space, technically called a Reproducible Kernel Hilbert Space (RKHS). A prominent feature of the procedure is that it learns an analytical and continuous function form of the vector field so one can obtain $F$ for regions unexplored by the sampling data. Therefore, dynamo gives a generative model that allows further in silico studies, which is similar to PRESCIENT but without the need of assuming that the dynamics is governed by the gradient of a scalar function.

The function $F$ contains quantitative information on gene regulation. Qiu et al. have developed a number of analysis tools, including efficient differential geometry analyses exploiting the analytical function form of the transcriptomic vector field, and using least-action path (LAP) analysis to predict optimal transition paths between two distinct phenotypes and key regulatory factors for the transition processes. Qiu et al. applied to a dataset on spontaneous differentiation of primary human CD34+ hematopoietic stem cells (HSPCs), and accurately revealed genome-wide quantitative regulation relations between genes, such as the mutual antagonism between two master regulators GATA1 and SPI1/PU.1, from a single set of high-quality scRNA-seq data. Traditionally considerable efforts are needed to acquire such quantitative information even for a pair of genes (47). Further in silico LAP analyses recovered key reprogramming factors that have been experimentally reported to lead reprogramming between different cell types.

**Extraction of dynamical information from live-cell time series data**

Despite the success of extracting dynamical information from snapshot data, the long-term dynamical information can only be inferred indirectly with a number of assumptions whose validity has not been systematically evaluated. For true long-term dynamics one needs to...
perform live-cell imaging over the same cells over time, which has its unique technical challenges. Here we specifically focus on cell phenotypic transition dynamics of eukaryotic cells, which typically takes days or weeks. Compared to fixed cell studies, there are more constraints on choosing the dynamical variables as observables since the number of measurable degrees of freedom is much more restricted. They should be compatible for live-cell imaging with minimal cell toxicity and perturbation to the cellular dynamics, and be sufficient to reflect the course of the transition process. Suppose that a gene is activated and reaches a plateau at the early stage of the transition, monitoring it alone is less informative on the later stage of the transition. However, cell- and photo-toxicity as well as the number of available fluorescent channels limit the number of molecular species one can monitor for characterizing cell phenotypic transitions using fluorescent-based labeling. For example, studies on the epithelial-to-mesenchymal transition (EMT) lead to a consensus that it is insufficient to use a small number of markers to reflect the progression of the process (48). Collective cell features measured through label-free imaging, either alone or in combination with fluorescent imaging, arise as alternative quantities for live-cell imaging.

While live-cell imaging studies on cell phenotypic transitions are still at an early stage, recent years witness an increasing number of such studies. Here I will only discuss a few of them. Buggenthin et al. showed that a combination of cell morphological features and movement accurately predicts lineage choices of mouse primary hematopoietic progenitors three generations before indication by conventional molecular markers (49). In a series of studies (23, 24), Marshall and coworkers examined multiple types of cells in cell motility and morphology state space. Figure 3A shows the measured transition vectors of mouse embryonic fibroblasts (MEFs) in the state space defined by the two leading two principal components (PCs) of a 205-dimensional cell size, shape, and texture feature space. Flux analyses on the transitions reveal that the forward and backward transitions between two states (see Eqn. 4) satisfy detailed balance (Fig. 3B), which is not expected for a cellular system that is out of thermodynamic equilibrium. With detailed balance one can define a scalar potential to represent the system dynamics, which reveals a multi-dimensional attractor in the state space (Fig. 3C).

Figure 3 Cytomorphological state space dynamics of mouse embryonic fibroblasts (MEFs). (A) Transition vectors in the leading PC space. The vectors are defined as the vector between the measured states of individual cells with four-hour separation. (B) Flux analyses of the transition vectors reveal a transition matrix with unexpected detailed balance between forward and backward states. (C) Quasi-potential defined from the transition matrix. Reproduced from (24) with permission.
Figure 4 Determination of reaction coordinates from single cell trajectories using a revised finite temperature string method. (A) Division of cell state space with Veronoi cells. A, I, and B refer to the initial, intermediate, and final regions of the transition, respectively. (B) Reconstructed parallel EMT transition paths of A549 cells (treated with 4 ng/ml TGF-β) shown in the leading 3-D state space. (C) Individual reaction coordinates overlapped with typical recorded single cell trajectories. (D) Reconstructed quasi-potentials along reaction coordinate from single cell trajectory data of A549 cells (treated with 1 and 4 ng/ml TGF-β). Reproduced from (21).

Wang and co-workers developed an integrated live-cell imaging and image analysis platform for studying cell phenotypic transitions (21, 22, 50). Through representing cells in a composite cell feature space, they obtained single cell trajectories, then applied rate theories to analyze the transition dynamics. Rate theories are an old branch of physics studying how a system transits between two distinct regions (i.e., regions A and B in the schematic diagram of Fig. 4A) in a state space (51). In rate theories, a trajectory is called a reactive one if it leaves region A and ends at region B before returning to A, and all of them form an ensemble of reactive trajectories. There are an infinite number of reactive trajectories with varying probabilities that connect A and B, and they typically concentrate within a “reaction tube” (52, 53). Furthermore, reaction coordinate (RC, denoted as l in Fig. 4) is a key concept in rate theories, and it refers to a one-dimensional manifold that connects A and B. Here I used the same notation as in Fig. 1 since the cell cycle stage there can be viewed as a RC. One way of defining a RC is the centroid of the reaction tube, and one may reconstruct such a RC from the reactive trajectory ensemble.
using a finite temperature string method (21, 52). Figure 4B shows two RCs corresponding to two parallel paths for the EMT process of A549 cells treated with TGF-β. Recorded single cell reactive trajectories fluctuate around these RCs and form the reaction tubes (Fig. 4C). With both the cell states and instant velocities from live-cell imaging trajectories, one can apply Eqn. 3 to reconstruct the vector field $\mathbf{A}$. For simplicity and better numerical convergence, Wang et al. considered the equation projected along each 1-D RC $l_\alpha$, which then reduces to the form,

$$\frac{dl_\alpha}{dt} = - \frac{d\phi_{\alpha}}{dt} + \zeta_\alpha(l_\alpha, t),$$

where $\alpha = 1, 2$ for the two parallel RCs. Here since the equation is one-dimensional, one can always define a scalar quasi-potential $\phi_{\alpha}(l_\alpha)$ for each RC. Numerically, the discretized RCs divide the multi-dimensional state space into Veronoi cells, so Eqn. 3 becomes,

$$\frac{d\phi_{\alpha i}}{dl_\alpha} = - \left( \frac{dl_\alpha}{dt} \right)_{i-th Veronoi cell},$$

with the average over all measured (reactive and non-reactive) trajectories that fall into the specific Veronoi cell. Then the quasi-potential relative to a reference point $(l_0)$ can be obtained through integration,

$$\phi_{\alpha i} - \phi_{\alpha 0} = \int_{l_{\alpha 0}}^{l_{\alpha i}} \frac{d\phi_{\alpha}}{dl_\alpha} dl_\alpha \propto \sum_j \int \frac{d\phi_{\alpha i}}{dl_\alpha}.$$  

The $l$-dependent diffusion “constant” is given by, $D_{\alpha i} = Variance \left( \frac{dl_\alpha}{dt} \right)_{\alpha i}$. The reconstructed quasi-potentials (Fig. 4D) reveal that TGF-β destabilizes the original epithelial attractor and the EMT process proceeds as relaxation to a new attractor. The process is analogous to a molecular system being excited from an electronic ground state potential and relaxing along the potential of an excited state (54). The remnant of the epithelial attractor appears as a plateau in the quasi-potential of one path (Fig. 4D left), and is more apparent with a lower TGF-β concentration. The Markovian assumption was justified with the Chapman-Kolmogorov test. The reconstructed Fokker-Planck equation predicts a stationary distribution that agrees well with experimental results, which is impressive given that the dynamical equation was reconstructed from non-stationary trajectories. With more data one may also follow the procedure of Qiu et al. (40) to obtain an analytical form of the multi-dimensional vector field from live-cell imaging data.

**Challenges and perspectives**

Cell biology studies are in an exciting “big data” era. While statistics-oriented approaches are still dominating, a new trend emerges to place the study within the framework of dynamical systems theories. In many senses it parallels to the history of advancing from pattern mining of Kepler to mechanistic theory development of Newton, though the richness of present-day data is unparalleled. With the early stage of the field, several major challenges remain to be tackled.

First, how to achieve and evaluate completeness of the cell state space description? A cell is a complex system and any existing approach can only measure a subset of the degrees of freedom. Completeness of cell state description is only relative and specific to the cellular process and time scale under consideration. A high-dimensional dynamical system like a cellular system always evolves along a low-dimensional dynamical manifold. Mathematically, completeness requires that the manifold of the full system is diffeomorphic to the corresponding projected manifold in the observable state space, so the former can be parameterized by the observables (or dynamical variables defined by the observables). Incomplete resolution of cell state leads to distorted description of the cellular dynamics and breakdown of the Markovian assumption. The unexpected detailed balance observed in the study of Chang and Marshall (24) might be due to such incomplete state resolution, especially cell cycle stages. Similarly it is unclear whether the observed two EMT paths of Wang et al (21, 22) truly originate from the same initial fixed-point attractor, or are actually from two unresolved distinct initial states corresponding to different cell cycle stages, given that one expects a circular instead of a fixed-
point attractor with cell cycle being resolved, and both G1/S and G2/M arrests have been reported during EMT (55-57). Practically, to increase the dimension of the state space, one may increase the modality of the fixed cell data, and the collective features in live-cell imaging. For the latter one may also expand the state space with Takens’s delay embedding theorem (58). It is also desirable to include cell-cell interactions and extracellular factors explicitly in the governing equations, and generalize the imaging platform to three-dimensional culturing and in vivo conditions.

Second, how to impose physical/biological/dynamical constraints with sparse and noisy data? Several efforts have been made on approximating cellular dynamics as gradient dynamics. However except in the 1D case (e.g. Fig. 4), one should take caution since a scalar potential does not provide a full description for the dynamics of a driven system. The Helmholtz-Hodge decomposition theorem states that a vector field can be decomposed into a curl-free, a divergent-free, and a harmonic part, and the first two terms can be further defined by a scalar and vector potential, respectively. While in many cases a cellular system and a potential system are topologically equivalent in the state space (59, 60), the detailed dynamics can be different. The full vector field as reconstructed in Qiu et al. (40) allows systematic investigation on what can and cannot be described with a gradient approximation for a cellular system. With that, the RKHS formalism they used is mathematically convenient, but lacks transparent mechanistic interpretation. It is also challenging to impose physical and biological constraints, such as the requirement that the effect of a gene on another one should reach saturation at high concentrations. This often-conflicting requirement between mathematical convenience and interpretability parallels to the comparison between molecular orbital theory versus valence bond theory in quantum chemistry. It remains to be explored on how to describe gene regulatory network dynamics using function forms familiar to the field of systems biology, and integrate information from scRNA-seq and other types of data.

Third, how to interpret and extract mechanistic information from high-dimensional vector field functions or the equations of motion? One excitement is that unbiased description of cellular dynamics provides information on how various cellular processes couple to or insulate from each other, a fundamental problem on understanding the “design principles” of biological networks. A cellular process may be affected by other processes evolving on the same or even slower time scale, leading to nongenetic heterogeneity. This is a phenomenon also discussed in other contexts, named dynamic/static disorders in chemistry, and annealed/quenched disorders in physics (61-63). One expects even richer dynamics with the complexity and nonequilibrium nature of a cellular system. We have seen some concepts and approaches from other established fields introduced to the study of cell phenotypic transitions. For example, the concept of RC has been repetitively adopted in various studies, such as the cell cycle axis in Fig. 1 (30), the popular pseudo-time trajectory analyses in the single cell genomics field (64, 65), and the LAP in Qiu et al. (40). The LAP is the zero-noise (zero-temperature) limit of the RC obtained through the finite temperature string method used in Wang et al. (21), and corresponds to the minimum-energy path for a gradient system. We expect to see applications of other approaches such as the Koopman operator (including dynamic mode decomposition) (66), Zwanzig-Mori projection operator (67-69), in studying cell phenotypic transition dynamics.

In this perspective I discussed analyses with snapshot and live-cell data, each of which has strengths and limitations. The duality and equivalence of representing cell states, in either the expression space (x) or cell feature space (y), suggest a strategy to combine the strengths of the two (22). The live-cell imaging platform generates real-time single cell trajectories in the cell
feature space $y(t)$. From the data one can reconstruct the governing equation $\frac{dy}{dt} = G(y) + \xi(t)$. Then by establishing a mapping between the cell state in the cell feature space and expression space, $M: y \rightarrow x$, one can obtain the kinematics description and the governing equation $\frac{dx}{dt} = F(x) + \eta(t)$ in the expression space that contains mechanistic information on gene regulation. The feasibility of the proposal may seem questionable given the seemingly mismatch on the dimensionality of fixed cell data one can acquire and that of the live-cell imaging data. The underlying theoretical consideration is that a high-dimensional dynamical system typically evolves along a low-dimensional manifold due to the constraints imposed by the component interactions. It is exciting to see whether this integrated strategy can lead to a more complete description of cellular dynamics, specifically cell phenotypic transitions.

**Acknowledgments**

This work was partially supported by National Cancer Institute (R37 CA232209), and National Institute of Diabetes and Digestive and Kidney Diseases (R01DK119232) to JX.

1. Alon U (2007) *An introduction to systems biology: design principles of biological circuits* (Chapman & Hall/CRC).
2. Tyson JJ, Chen KC, & Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* 15(2):221-231.
3. Ferrell JE, Jr. (2012) Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr Biol* 22(11):R458-466.
4. Huang S & Kauffman S (2012) ComplexGRN complex GeneComplex GRN Regulatory Networks – from Structure to Biological Observables: Cell Fate DeterminationGene regulation, cell fate determination. *Computational Complexity*, ed Meyers RA (Springer New York), pp 527-560.
5. Macklin DN, et al. (2020) Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science* 369(6502).
6. Karr Jonathan R, et al. (2012) A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* 150(2):389-401.
7. Kharchenko PV (2021) The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods* 18(7):723-732.
8. Teschendorff AE & Feinberg AP (2021) Statistical mechanics meets single-cell biology. *Nature Reviews Genetics* 22:459–476
9. Schier AF (2020) Single-cell biology: beyond the sum of its parts. *Nature Methods* 17(1):17-20.

10. Lähnemann D, *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biology* 21(1):31.

11. Wagner DE & Klein AM (2020) Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* 21(7):410-427.

12. Gordonov S, *et al.* (2016) Time series modeling of live-cell shape dynamics for image-based phenotypic profiling. *Integr Biol (Camb)* 8(1):73-90.

13. Mandal M, *et al.* (2016) Modeling continuum of epithelial mesenchymal transition plasticity. *Integrative Biology* 8(2):167-176.

14. Zhong Q, Busetto AG, Fededa JP, Buhmann JM, & Gerlich DW (2012) Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nat Meth* 9(7):711-713.

15. Held M, *et al.* (2010) CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat Methods* 7(9):747-754.

16. Yin Z, *et al.* (2013) A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. *Nat Cell Biol* 15(7):860-871.

17. Alizadeh E, *et al.* (2020) Cellular morphological features are predictive markers of cancer cell state. *Computers in Biology and Medicine* 126:104044.

18. Tsygankov D, *et al.* (2014) CellGeo: a computational platform for the analysis of shape changes in cells with complex geometries. *J Cell Biol* 204(3):443-460.

19. Bakal C, Aach J, Church G, & Perrimon N (2007) Quantitative morphological signatures define local signaling networks regulating cell morphology. *science* 316(5832):1753-1756.

20. Wu P-H, *et al.* (2020) Single-cell morphology encodes metastatic potential. *Science Advances* 6(4):eaaw6938.
21. Wang W, Poe D, Yang Y, Hyatt T, & Xing J (2022) Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. eLife 11:e74866.
22. Wang W, et al. (2020) Live-cell imaging and analysis reveal cell phenotypic transition dynamics inherently missing in snapshot data. Science Advances 6(36):eaba9319.
23. Kimmel JC, Chang AY, Brack AS, & Marshall WF (2018) Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. PLoS computational biology 14(1):e1005927.
24. Chang AY & Marshall WF (2019) Dynamics of living cells in a cytomorphological state space. Proc Natl Acad Sci USA 116(43):21556-21562.
25. Devaraj V & Bose B (2019) Morphological State Transition Dynamics in EGF-Induced Epithelial to Mesenchymal Transition. J Clin Med 8(7):911.
26. van Kampen NG (2007) Stochastic Processes in Physics and Chemistry (North Holland) 3rd Ed.
27. Xing J (2010) Mapping between dissipative and Hamiltonian systems. Journal of Physics A: Mathematical and Theoretical 43(37):375003.
28. Qian H (2014) Fitness and entropy production in a cell population dynamics with epigenetic phenotype switching. Quantitative Biology 2(1):47-53.
29. Wheeler RJ (2015) Analyzing the dynamics of cell cycle processes from fixed samples through ergodic principles. Molecular Biology of the Cell 26(22):3898-3903.
30. Kafri R, et al. (2013) Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. Nature 494(7438):480-483.
31. Weinreb C, Wolock S, Tusi BK, Socolovsky M, & Klein AM (2018) Fundamental limits on dynamic inference from single-cell snapshots. Proc Natl Acad Sci USA 115(10):E2467-E2476.
32. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, & Klein AM (2020) Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* 367(6479):eaaw3381.

33. Wang S-W, Herriges MJ, Hurley K, Kotton DN, & Klein AM (2022) CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nature Biotechnology*.

34. Yeo GHT, Saksena SD, & Gifford DK (2021) Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications* 12(1):3222.

35. Gupta PB, *et al.* (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146(4):633-644.

36. Hu Y, *et al.* (2013) Tumor-Specific Chromosome Mis-Segregation Controls Cancer Plasticity by Maintaining Tumor Heterogeneity. *PLoS ONE* 8(11):e80898.

37. Zhang J, *et al.* (2014) TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Science signaling* 7(345):ra91.

38. Stumpf PS, *et al.* (2017) Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Syst* 5(3):268-282 e267.

39. Zhou P, Wang S, Li T, & Nie Q (2021) Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature Communications* 12(1):5609.

40. Qiu X, *et al.* (2022) Mapping Transcriptomic Vector Fields of Single Cells. *Cell* 185(4):690-711.

41. La Manno G, *et al.* (2018) RNA velocity of single cells. *Nature*.

42. Battich N, *et al.* (2020) Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* 367(6482):1151.

43. Cao J, Zhou W, Steemers F, Trapnell C, & Shendure J (2020) Sci-fate characterizes the dynamics of gene expression in single cells. *Nature Biotechnology* 38(8):980-988.
44. Erhard F, et al. (2019) scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* 571(7765):419-423.
45. Hendriks G-J, et al. (2019) NASC-seq monitors RNA synthesis in single cells. *Nature Communications* 10(1):3138.
46. Qiu Q, et al. (2020) Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods* 17(10):991-1001.
47. Nerlov C, Querfurth E, Kulessa H, & Graf T (2000) GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* 95(8):2543-2551.
48. Yang J, et al. (2020) Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat Rev Mol Cell Biol* 21:341-352.
49. Buggenthin F, et al. (2017) Prospective identification of hematopoietic lineage choice by deep learning. *Nature Methods* 14(4):403-406.
50. Wang W, et al. (2019) Learn to segment single cells with deep distance estimator and deep cell detector. *Computers in Biology and Medicine* 108:133-141.
51. Hanggi P, Talkner P, & Borkovec M (1990) Reaction-rate theory: 50 years after Kramers. *Rev. Mod. Phys.* 62:254-341.
52. Vanden-Eijnden E & Venturoli M (2009) Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J Chem Phys* 130(19):05B605.
53. E W & Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. *Annu Rev Phys Chem* 61:391-420.
54. Herzberg G (2008) *Molecular Spectra and Molecular Structure* (Read Books).
55. Mejlvang J, et al. (2007) Direct Repression of Cyclin D1 by SIP1 Attenuates Cell Cycle Progression in Cells Undergoing an Epithelial Mesenchymal Transition. *Molecular biology of the cell* 18(11):4615-4624.
56. Grande MT, et al. (2015) Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat Med* 9:989-997.

57. Lovisa S, et al. (2015) Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nat Med* 21(9):998-1009.

58. Takens F (1981) Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics, eds Rand D & Young L (Springer, Berlin, Heidelberg), Vol 898.

59. Rand DA, Raju A, Sáez M, Corson F, & Siggia Eric D (2021) Geometry of gene regulatory dynamics. *Proc Natl Acad Sci USA* 118(38):e2109729118.

60. Smale S (1961) On Gradient Dynamical Systems. *Annals of Mathematics* 74(1):199-206.

61. Frauenfelder H, Sligar SG, & Wolynes PG (1991) The Energy Landscapes and Motions of Proteins. *Science* 254(5038):1598-1603.

62. Zwanzig R (1990) Rate-Processes with Dynamic Disorder. *Acc. Chem. Res.* 23(5):148-152.

63. Fischer KHHJA (1993) *Spin glasses* (Cambridge University Press, Cambridge).

64. Setty M, et al. (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 34(6):637-645.

65. Qiu X, et al. (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* 14(10):979.

66. Brunton SL, Budišić M, Kaiser E, & Kutz JN (2021) Modern Koopman Theory for Dynamical Systems. *arXiv*:2102.12086.

67. Zwanzig R (1961) Memory effects in irreversible thermodynamics. *Phys. Rev.* 124:983-992.

68. Xing J & Kim KS (2011) Application of the projection operator formalism to non-Hamiltonian dynamics. *J Chem Phys* 134(4).
69. Chorin AJ, Hald OH, & Kupferman R (2000) Optimal prediction and the Mori–Zwanzig representation of irreversible processes. *Proc Natl Acad Sci U S A* 97(7):2968-2973.