Population-expression models of immune response

Sean P Stromberg[1,2]**, Rustom Antia**[1] **and Ilya Nemenman**[1,3]

[1] Department of Biology, Emory University, Atlanta, GA 30322, USA
[2] Department of Physics, University of California, Santa Barbara, CA 93106, USA
[3] Department of Physics, Emory University, Atlanta, GA 30322, USA

E-mail: stromberg@physics.ucsb.edu, rantia@emory.edu and ilya.nemenman@emory.edu

Abstract

The immune response to a pathogen has two basic features. The first is the expansion of a few pathogen-specific cells to form a population large enough to control the pathogen. The second is the process of differentiation of cells from an initial naive phenotype to an effector phenotype which controls the pathogen, and subsequently to a memory phenotype that is maintained and responsible for long-term protection. The expansion and the differentiation have been considered largely independently. Changes in cell populations are typically described using ecologically based ordinary differential equation models. In contrast, differentiation of single cells is studied within systems biology and is frequently modeled by considering changes in gene and protein expression in individual cells. Recent advances in experimental systems biology make available for the first time data to allow the coupling of population and high dimensional expression data of immune cells during infections. Here we describe and develop *population-expression* models which integrate these two processes into systems biology on the multicellular level. When translated into mathematical equations, these models result in non-conservative, non-local advection-diffusion equations. We describe situations where the population-expression approach can make correct inference from data while previous modeling approaches based on common simplifying assumptions would fail. We also explore how model reduction techniques can be used to build population-expression models, minimizing the complexity of the model while keeping the essential features of the system. While we consider problems in immunology in this paper, we expect population-expression models to be more broadly applicable.

S Online supplementary data available from stacks.iop.org/PhysBio/10/035010/mmedia

1. Introduction

The central feature of the adaptive immune system is the ability to respond to a broad range of pathogens, including emerging threats never before encountered, without mounting responses to the native tissues of the body [1]. This dynamic is explained by the *clonal selection theory*, which underlies our understanding of immunology. This theory postulates that we begin with a very diverse population of immune cells (lymphocytes), with each lymphocyte having a unique and fixed specificity. Consequently the number of lymphocytes specific for a given pathogen is very small. Following infection

these pathogen-specific lymphocytes undergo rapid division (clonal expansion) and differentiation into effector cells, which are able to control the pathogen. Following clearance of the pathogen some of these lymphocytes differentiate into memory cells, which are maintained for extended periods and account for long-term protection. The clonal selection theory describes the generation of the T cell and B cell responses. In figure 1 we show a schematic of clonal selection for a typical CD8 T cell response to a viral infection.

The enormous changes in population sizes suggested that, as in ecology, ordinary differential equation (ODE) models of the populations would prove useful to understanding the
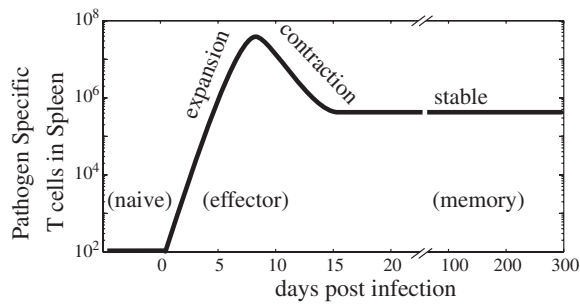
Figure 1. Schematic illustration of a typical CD8 T cell response to an infection. The plot shows the enormous changes in numbers of pathogen-specific CD8 T cells during the course of infection, as well as changes in cell phenotype. The response has three phases which correspond to population expansion, contraction and stability. Differentiation results in changes in the phenotype of cells from naive to effector and memory. Typically this type of a response is described by ordinary differential equations that govern changes in populations of cells having naive, effector and memory phenotypes.

immune response [2–4]. In these models cells are restricted to a few distinct phenotypes with division, death, and transition rates between the phenotypes to describe the dynamics. The models typically ignore how the systems biology on the cellular scale governs the rate laws in the models on the population scale. While such models have proven useful in addressing a number of population level questions, they have their limitations. For the approach to work well, phenotypic states must be well resolved and the transitions between them must be rapid.

Figure 2 presents data capturing the dynamics of T cells obtained via flow cytometry. This figure shows the density of CD8 T cells following a yellow fever vaccination plotted as a function of two surface expressed molecules (CD45RA, a signaling molecule that regulates antigen receptor signaling, and CCR7, a molecule which aids in trafficking of T cells to lymph-nodes) [5]. The population gradually transitions from CD45RA low to high during the contraction and memory phases. This figure illustrates one problem with ODE models of multicellular population dynamics: how does one unambiguously partition data into distinct phenotypes when there is considerable heterogeneity or gradual transitions? This ambiguity gives rise to subjectivity and quantitative

disagreement between labs in the analysis of immunological data [6].

The flow of populations as they differentiate (figure 2) is governed largely by the systems biology of the cells [7–9]. (While the term *systems biology* has been used very broadly, in this paper we adopt the most common usage, referring to models of chemical reaction networks typically within single cells or homogeneous cell cultures [10].) Typical systems biology models consist of ODEs or stochastic differential equations that model reaction rates between chemical species, providing a finer resolution of phenotypic states.

While population models loose accuracy in not considering the chemical scale, systems biology models have contrasting limitations resulting from omission of the population dynamics. Typically, the analysis and parameter estimation of differentiating populations has been performed on time scales where division is negligible [11]. On longer time scales, population dynamics and systems biology are coupled and must be considered together. The expression levels of gene products control cell division and death rates. In their turn, cell division and death rates change the number of cells in various phenotypic states and hence shape the expression profiles of populations. Additionally the process of cell division dilutes expression and can generate spurious correlations between expressed chemicals. Clearly, modeling immune system dynamics requires an integrated approach, combining population dynamics and systems biology.

One way to do this is to conceptualize each flow cytometry data set as samples from a density in a multidimensional cellular configurational space, where each dimension denotes the quantity of a specific chemical. Individual cells would trace out trajectories in this configurational space as they differentiate. Unfortunately *in vivo* single cell longitudinal data is difficult to obtain, and for dividing cells the term longitudinal is undefined. Thus instead of tracking cells over time, one can focus on tracking populations, or distributions of cells in the configurational space. This can be done using partial differential equations (PDEs) and related mathematical concepts, an approach gaining popularity in theoretical immunology [12–14]. We refer to the dynamics of chemical expression (gene, protein, metabolite, etc) in a dynamic population as the *population-expression*, and models of the population-expression as *population-expression models*.
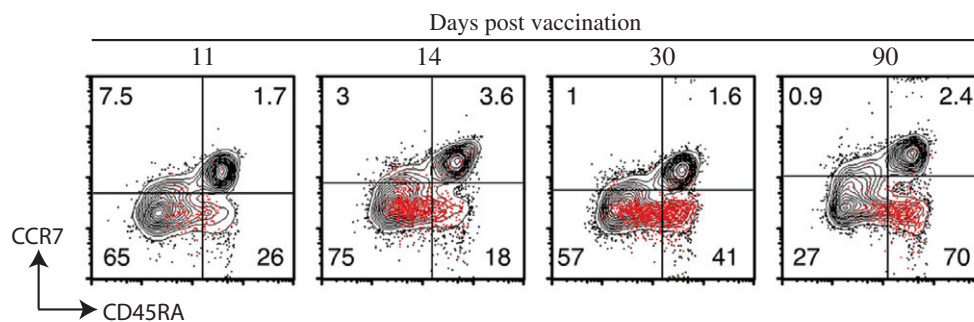
Figure 2. The differentiation of human CD8 T cells following yellow fever vaccination. These flow cytometry plots show the population of antigen specific CD8 T cells (red) responding to the vaccination differentiating from CD45RA negative to positive while expanding and then contracting in number. This transition is associated with the transition from effector memory to central memory. Reproduced with permission from [5] (copyright 2009. The American Association of Immunologists, Inc.).

Such population-expression models circumvent our inability to define clean cellular phenotypic states. They remove the inherent subjectivity in phenotype discrimination [6], and they remove the need to incorporate additional phenotypes to better fit models to data. They integrate the within cell stochastic chemical kinetics into models of the population dynamics. Ultimately, they allow analysis of the diversity of protein expression within populations, how it changes with time, and how the diversity is affected by selection.

The main goal of this paper is to introduce such population-expression modeling, explain utility of the approach in the context of toy models, and discuss the methodological developments needed for practical applications of the ideas. To achieve this, we first introduce a formalism for population-expression models using PDEs, and non-local PDEs. Following this we provide a number of examples of population-expression models, illustrating where ecological based ODE models succeed and fail, how cell division dilutes chemical quantities, where single-cell analyses fail to describe the population, and how we can infer from data which chemicals may be drivers in regulatory networks. We end with a critical look at some of the key problems arising when we confront population-expression models with ever increasing dimensionality of experimental datasets.

2. Population-expression approach: PDE formulation

Instead of predefining a limited number of cell phenotypes, our population-expression approach takes the abundance of cells with different chemical states as the dynamical variable. We denote by $\rho(\vec{A}, t)$ the density of cells at time t, with internal biochemical expressions (internal states) of $\vec{A}$.

To describe how $\rho(\vec{A}, t)$ changes with time, we first consider how a single cell moves in the configurational space of $\vec{A}$ values. Denoting the set of differential equations that describe the changing chemical quantities within a single cell by

$$\frac{\mathrm{d}\vec{A}}{\mathrm{d}t} = \vec{\gamma}(\vec{A}), \tag{1}$$

the abundance ρ flows according to the vector field denoted by $\vec{\gamma}(\vec{A})$. A number of techniques exist to translate from the single cell model to a population model [15]. In the accompanying supplementary materials we provide two such contrasting derivations, one more common to the fluid-dynamics community (based on the divergence theorem), and the other more common to statistical physics and systems biology (based on the chemical master equation). These techniques have identical results, generating an advection equation describing how the density changes according to the vector field $\vec{\gamma}(\vec{A})$:

$$\frac{\partial \rho(\vec{A}, t)}{\partial t} = -\vec{\nabla} \cdot [\vec{\gamma}(\vec{A})\rho(\vec{A}, t)]. \tag{2}$$

The quantity in the square brackets denotes the total flux of cells changing in expression level as they move through the configurational space, and $\vec{\nabla}$ defines the divergence operator,

a vector of partial derivative operators $(\partial/\partial A_1, \partial/\partial A_2, \ldots)$. The formulation is valid for arbitrary dimensionality, and the examples in the following sections use either one or two dimensions for simplicity.

Incorporating population dynamics into these equations can be done with additional terms for cell death and sources of new cells:

$$\frac{\partial \rho(\vec{A}, t)}{\partial t} = -\vec{\nabla} \cdot [\vec{\gamma}(\vec{A})\rho(\vec{A}, t)] - \nu(\vec{A})\rho(\vec{A}, t) + \Gamma(\vec{A}). \tag{3}$$

Here $\nu(\vec{A})$ denotes a cellular death rate that is a function of the chemical concentration and $\Gamma(\vec{A})$ is an influx of new cells entering the system in a chemical state $\vec{A}$.

Cell division can be included by adding nonlocal terms to equation (3). For example, if in a symmetric cell division, all chemicals in the cell are split equally between the two daughters, we have

$$\frac{\partial \rho(\vec{A}, t)}{\partial t} = -\vec{\nabla} \cdot [\vec{\gamma}(\vec{A})\rho(\vec{A}, t)] - \mu(\vec{A})\rho(\vec{A}, t)$$
$$+ 2^d \cdot 2\mu(2\vec{A})\rho(2\vec{A}, t) - \nu(\vec{A})\rho(\vec{A}, t) + \Gamma(\vec{A}). \tag{4}$$

Here $\mu(\vec{A})$ is the division rate, which we assume depends on the cell age and other properties only implicitly through the instantaneous state of the cell, $\vec{A}$. In this equation, cells with chemical quantity $\vec{A}$ are removed from the abundance at $\vec{A}$ as they divide with rate $\mu(\vec{A})$. Separately, each cell dividing at abundance $2\vec{A}$ is adding two cells to the abundance at $\vec{A}$. The factor of 2^d arises from a subtlety of the non-local calculus. Division adds *to* an infinitesimal volume of the space, bounded in each dimension by $(A_i, A_i + \delta A_i)$ where δ is an infinitesimal quantity. The cells however are coming *from* a region with boundaries $(2A_i, 2A_i + 2\delta A_i)$, which is twice the width in each of the d dimensions. This equation can also be modified to describe dilution in asymmetric cell division.

It is frequently the case that non-locality gives rise to integro-differential equations. If we incorporated partitioning noise into our equation, it would generate an integral term as new cells would enter the population at $\vec{A}$ from a range of values centered around $2\vec{A}$. When considering systems with small numbers of molecules, this approach is an important extension. With large numbers of molecules, the relative variation is small, and partitioning noise can be neglected.

Finally we can very naturally incorporate the stochastic fluctuations resulting from the chemical dynamics [16, 17]. This is typically done by constructing a chemical master equation, and expanding in small relative fluctuations [18]. An example derivation is provided in the supplementary material, available at stacks.iop.org/PhysBio/10/035010/mmedia. Expanding the chemical master equation to lowest order gives equation (2). The expansion of the chemical master equation to next highest order results in a nonlocal analogue of the Fokker–Planck equation, which spreads the population in the $\vec{A}$ space due to stochasticity of the intrinsic chemical processes:

$$\frac{\partial \rho(\vec{A}, t)}{\partial t} = -\vec{\nabla} \cdot [\vec{\gamma}(\vec{A})\rho(\vec{A}, t)]$$
$$- \mu(\vec{A})\rho(\vec{A}, t) + 2^d \cdot 2\mu(2\vec{A})\rho(2\vec{A}, t)$$
$$- \nu(\vec{A})\rho(\vec{A}, t) + \Gamma(\vec{A}) + \vec{\nabla} \cdot (\mathbf{D}(\vec{A})\nabla\rho(\vec{A}, t)). \tag{5}$$

Here $\mathbf{D}(\vec{A})$ is a diffusion tensor. The advection dynamics becomes advection-diffusion dynamics with the incorporation of within-cell stochasticity. Such approaches to modeling fluctuations in single cells are now commonplace in molecular systems biology [19], and many efficient simulation and analysis algorithms have been developed [20].

As in systems biology, in population-expression models some state variables may remain discrete. For example the state of transcription factor binding may be best described by a binary *on/off* variable, or compartmental spacial dependence could be incorporated into the model (e.g. lung, spleen, etc). In these cases we typically describe multiple coupled densities $\rho_i(\vec{A}, t)$, with population-expression dynamics, equation (5), for each density and with terms that couple the equations through transitions between the states, such as $\sum_i k_j \rho_j(\vec{A}, t)$. Of possibly high relevance to the current work, cross-sectional flow cytometry samples from cellular populations at different time points have been used to infer parameters of chemical reaction rates $\vec{\gamma}$ [11]. The population-expression approach differs from these analyses by incorporating the effects of proliferation, cell death, and dilution by cell division. We show below that these effects can substantially bias the resulting expression profile of a population.

Note that, for much of this paper, we assume that $\rho(\vec{A}, t)$ can be measured: that the number of samples is large enough so that inference of ρ is not a hard task. This breaks down if $d = \dim \vec{A} \gg 1$. We will discuss this case in section 4. Similarly, we assume that population-expression equations are sufficiently low-dimensional to be numerically solvable. When this is not the case, Monte-Carlo simulations might be needed, and we briefly touch on this topic in the *discussion*.

3. Population-expression approach: examples

In this section, we use the population-expression approach to model simple processes of relevance to different aspects of immune dynamics. The examples illustrate the inadequacy of single cell systems biology (expression) and ecological based ODE (population) modeling approaches.

3.1. Ecological based ODE model failure: slow expression dynamics

Ecological based ordinary differential equation models of phenotypical population dynamics work well only when phenotypes are sharply defined and transitions between them are rapid. This is not always the case. Consider, for example, a transition between phenotypes that occurs when an internal state has changed, but the observables take time to reach their characteristic values for this new state. For example, a good measure of the phenotypic state of a cell may be the binding of transcription factors (TFs) to DNA, which is possible but not easy to measure [21]. On the other hand, we routinely measure expression levels of protein using flow cytometry. These levels are typically controlled by transcription factor binding, but changes in protein expression lag behind changes

in TF binding. Thus the dynamics of switching observed in flow cytometry data may be non-trivial.

Here we model cells having a discrete state denoting transcription factor binding ('off' or 'on'), and a continuous variable A for expression level. Off cells can switch to the on state with the rate k, and the dynamics of A is given by $dA/dt = \gamma_{\text{off/on}}$, where $\gamma_{\text{off/on}}$ depends on the state. Namely, the chemical A has two possible production rates: α_{off}, and α_{on}. In both states there is the same degradation rate β. This kinetics may correspond, for example, to the expression and decay of mRNA or protein if mRNA levels equilibrate quickly in comparison to the protein dynamics. We consider the cells in the two states separately: $\rho_{\text{off}}(A, t)$ is the density of cells in the off state with expression level A, and $\rho_{\text{on}}(A, t)$ are the cells in the on state. The population-expression equations are

$$\frac{\partial \rho_{\text{off}}(A, t)}{\partial t} = -\frac{\partial}{\partial A}[(\alpha_{\text{off}} - \beta A)\rho_{\text{off}}(A, t)]$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial A^2}[(\alpha_{\text{off}} + \beta A)\rho_{\text{off}}(A, t)] - k\rho_{\text{off}}(A, t), \quad (6)$$

$$\frac{\partial \rho_{\text{on}}(A, t)}{\partial t} = -\frac{\partial}{\partial A}[(\alpha_{\text{on}} - \beta A)\rho_{\text{on}}(A, t)]$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial A^2}[(\alpha_{\text{on}} + \beta A)\rho_{\text{on}}(A, t)] + k\rho_{\text{off}}(A, t). \quad (7)$$

Similar models for single cells in equilibrium [22, 23], and even off-equilibrium for simpler cases [24], have been solved exactly. Here we analyze this system numerically in the non-equilibrium context. We solve these equations with a method of lines integration with a finite differencing approximation for A derivatives, and Matlab ODE45 routine for integrating forward in time.

Figure 3 plots numerical solutions of $\rho(A, t) = \rho_{\text{off}}(A, t) + \rho_{\text{on}}(A, t)$, defined by equations (6) and (7), for two contrasting pictures of differentiation. The left panels shows infrequent TF switching with rapid protein expression ($k \ll \beta\langle A \rangle$). In this case the protein concentration in each cell tracks its transcriptional state well, phenotypes are well defined, and an ODE model describing switching between them works well. The right panels in figure 3 represent the case when TF switching is rapid, but change in protein expression is gradual. The initial and final states are identical to the scenario on the left. The gradual protein expression gives a large density of cells with intermediate protein expression on day 15, and no well-resolved phenotypes.

Dashed lines in figure 3 define low and high expressing phenotypes, as is typical in the analysis of flow cytometry data. The number of cells in the low expressing phenotype is shown as a function of time in figure 4 for both scenarios. For rare switching, modeling the system with two phenotypes with population sizes X_1 and X_2, respectively, as

$$\frac{dX_1}{dt} = -kX_1,$$
$$\frac{dX_2}{dt} = +kX_1, \quad (8)$$

produces great fits to the data. In contrast, two-state modeling for the slow protein expression case is inaccurate.

To model the data with ODEs and discrete states, several approaches could be taken. As is common in immunology [25] one could introduce sub-phenotypes, partitioning the cells
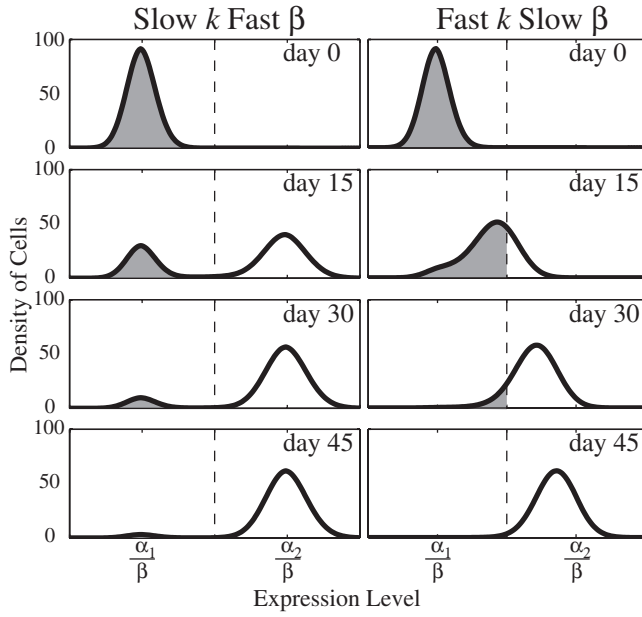
Figure 3. Expression profiles at multiple time points for slow state switching/fast protein production (left column) and fast switching/slow protein production (right column). The dashed line represents the half-way point between the steady state protein concentrations in the two states. ODE models typically count the number of cells above and below a threshold like this one to consider transition rates. The parameters of the two models are such that the steady state values and the overall protein relaxation times are the same. The shaded region corresponds to the number of low-expressing cells plotted in figure 4. In the simulation on the left: $\alpha_1 = 94.5$ copies/day, $\alpha_2 = 190$ copies/day, $\beta = 1.0$ day^{-1}, $k = 0.075$ day^{-1}; while on the right: $\alpha_1 = 5.0$ copies/day, $\alpha_2 = 10.0$ copies/day, $\beta = 0.05$ day^{-1}, $k = 0.2$ day^{-1}.
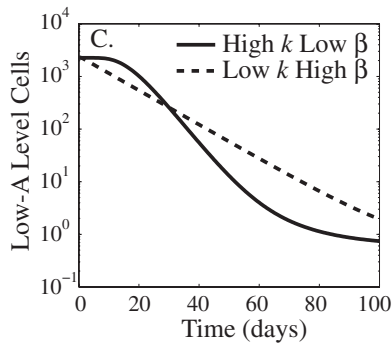
Figure 4. The total number of cells in the low-expression phenotype (as defined by the cutoff in figure 3) for the two scenarios. When the expression level equilibrates rapidly after the transcription factor is bound (dashed line), the system has a single characteristic decay time scale, and the number of cells in the low state can be modeled with a single ODE. When the protein expression dynamics are slow to respond (solid line), the decay of the population in the low state is non-exponential.

into $n > 2$ domains by some predefined thresholding of their expressions, such that

$$\frac{dX_1}{dt} = -k_1 X_1,$$
$$\frac{dX_i}{dt} = -k_i X_i + k_{i-1} X_{i-1}, \qquad i = 2, \ldots, n-1,$$
$$\frac{dX_n}{dt} = +k_{n-1} X_{n-1}. \tag{9}$$

One would then optimize the parameters k_i to produce the best fit to the data. This approach also has its limitations. The steady state distribution given by equations (6) and (7) has a width, while equation (9) has a steady state where all cells are within the X_n partition. Any overlap between the steady state distribution and the X_{n-1} state will not be resolved by such a model. Additionally this method introduces spurious phenotypes having little to do with the underlying biology. Alternatively, one could make the transition rate k a function of time $k(t)$. Like the previous case, this technique describes the data, but provides little insight into the biology of the system.

3.2. Failure of single cell systems biology: cell division

The models presented here are constructed based on chemical number rather than concentration. This gives correspondence with fluorescence experiments and enables accurate estimation of stochastic effects. Upon cell division we must divide the contents of a cell in half (assuming symmetric cell division). This gave us the non-local PDE in equation (4). Such nonlocal partial differential equations are uncommon and most computational tools are ill-equipped to deal with them. The use of finite difference, finite element, and spectral methods in solving these types of equations has been studied in a series of papers [26–28]. For large dimensional systems, Monte-Carlo integration can provide a more efficient numerical solution. In these examples we use finite difference methods.

Dilution of a dye. As a simple example of dilution by division, consider a dye such as CFSE or BrdU. These dyes are used to measure cell division rates in vivo and are frequently used in studying the cellular dynamics of immune responses. These dyes are not produced by the cells and are degraded slowly, yielding $\gamma = 0$. This removes the advection term in equation (4) yielding

$$\frac{\partial \rho(A, t)}{\partial t} = -\mu \rho(A, t) + 4\mu \rho(2A, t). \tag{10}$$

For a dye that initially has a narrow Gaussian distribution in cells, we have the output shown in figure 5. This system has been well described using ODE models [29, 30], with a single ODE for the number of cells in each peak. We note that, for brevity, we are using a model with exponentially distributed division times. For rapidly dividing cells, more detailed models of cell cycle provide greater accuracy [30, 13].

Dilution and homeostasis. For a chemical that is produced in the cell, division can bias the population-expression. Figure 6 shows an example of this effect. Here we have simulated two populations of cells, one not-dividing (solid) and one undergoing homeostatic division (cell death and division rates are equal, dashed curve). These curves are stationary distributions generated by the equation

$$\frac{\partial \rho(A, t)}{\partial t} = -\frac{\partial}{\partial A}[(\alpha - \beta A)\rho(A, t)]$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial A^2}[(\alpha + \beta A)\rho(A, t)] - 2\mu \rho(A, t) + 4\mu \rho(2A, t), \tag{11}$$
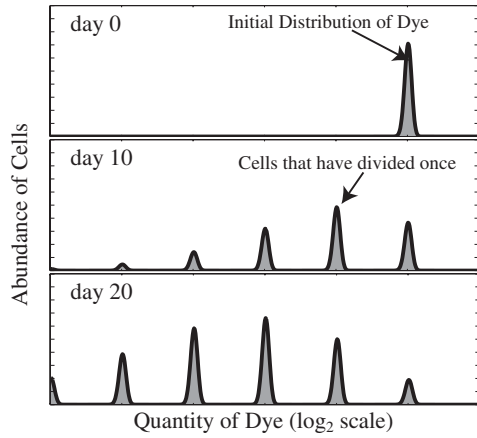
Figure 5. Dilution of dye upon cell division. The rightmost peak is the initial undivided cells. Each peak to the left is cells that have gone through an additional division. Here $\mu = 0.07$ day^{-1}.
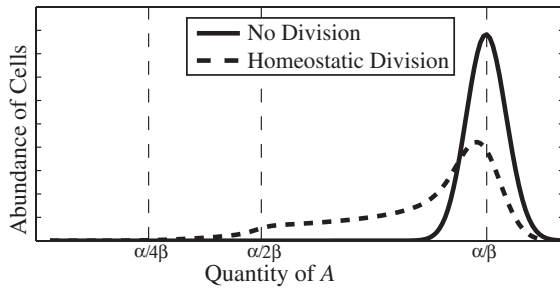

Figure 6. Stable distributions from a non-dividing population and one that is undergoing homeostatic division. Homeostatic division biases the population-expression, reducing the mean and increasing the width. The effects of cell division on the population-expression are greater the more rapid the division rate is. Parameter values were $\alpha = 200$ copies/day, $\beta = 0.4$ day^{-1}, $\mu = 0.07$ day^{-1}.

where we have also included the stochastic effects of the chemical dynamics. To keep the system from growing, we have cell death rate equal to division rate, giving an extra factor of 2 in the second to last term.

Figure 6 shows stable distributions for this system with and without $\mu = 0$. As we can see, cell division biases the distribution, reducing the mean and increasing the width. In general, the more rapid the division, the more exaggerated the effects. If the division rate exceeds the chemical degradation rate β, the stable distribution is very different from what is seen here, and is centered close to $A = 0$.

Statistical deviations resulting from cell division have been studied in previous work [31–33]. This type of noise is typically considered *extrinsic noise* [31]. It is also often approximated as a local and continuous process and incorporated into chemical decay terms [34], a modeling choice which omits many of the effects illustrated in this section.

Dilution and expansion. Figure 7 shows a simulation of a bivariate process where the vertical axis represents a chemical A_1 that is produced by the cell, as in equation (4) (figure 6), and the horizontal axis represents a dye concentration A_2, with
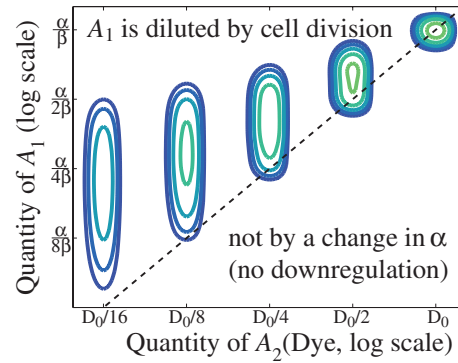

Figure 7. Simulation of a population of cells on day 7 for a two dimensional system with chemical A_1 obeying chemical rate law $\mathrm{d}A_1/\mathrm{d}t = \alpha - \beta A_1$, and A_2 representing a dye. Contours are spaced logarithmically and we have included stochastic effects of A expression. The population initially had A_1 distributed at equilibrium for a non-dividing population. At day 0 the population began dividing which dilutes A_1 and A_2, though A_1 is produced in the cell giving a vertical spread. This simulation has correspondence with a resting population of lymphocytes that is dyed and then stimulated by an infection on day 0 resulting in rapid expansion. Thought there is no change in the production rate α dilution gives a reduction in expression. Population-expression models can discriminate between reduction in expression resulting from a change in chemical dynamics and this simple dilution. Here $\mu = 0.09$ day^{-1}, $\alpha = 43$ copies/day, $\beta = 0.08$ day^{-1}.

dynamics as in equation (10) (figure 5). Here the population is expanding rather than undergoing homeostatic division. The equation describing the dynamics of the system is

$$\frac{\partial \rho(A_1, A_2, t)}{\partial t} = -\frac{\partial}{\partial A_1}[(\alpha - \beta A_1)\rho(A_1, A_2, t)]$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial A_1^2}[(\alpha + \beta A_1)\rho(A_1, A_2, t)]$$
$$- \mu\rho(A_1, A_2, t) + 8\mu\rho(A_1, A_2, t), \tag{12}$$

The simulation considers a system where cells are initially in an equilibrium distribution for a non-dividing population (solid curve in figure 6 for vertical axis, and day 0 density in figure 5 for horizontal axis). Beginning on day 0 in this simulation, the cells are stimulated to divide. This simulation has correspondence with resting lymphocytes that are dyed with CFSE before the system is infected on day zero, initiating rapid lymphocyte division. Contours are drawn with logarithmic spacing. On day 7 shown in figure 7, A_1 is diluted as the population divides. There has been no internal change in the chemical dynamics as is typically considered in down-regulation of a gene-product. The use of population-expression models can help to discriminate between a down regulation where production rate α is decreased and where simple dilution is occurring.

Cell division and spurious correlations. Another effect of cell division is that two chemical quantities that have independent dynamics can have correlations generated by cell division. Cell division will cut both otherwise independent quantities in half simultaneously. The population-expression
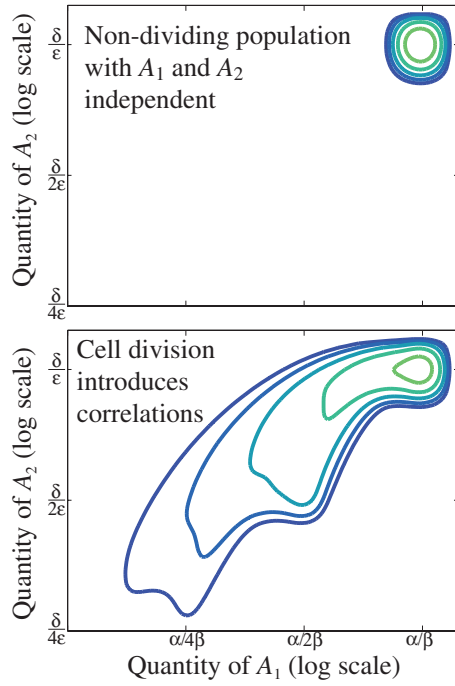
Figure 8. Cell division introduces correlations between otherwise independent gene products. Here A_1 and A_2 have simple rate laws ($dA_1/dt = \alpha - \beta A_1$ and $dA_2/dt = \delta - \epsilon A_2$ with A_2 dynamics faster than A_1). In a non-dividing population, we see that these products are not correlated (top). In dividing cells, both A_1 and A_2 are halved at the same time (cell division) introducing correlations in expression level (bottom). The effects of cell division should be accounted for when analyzing expression data for correlations to avoid spurious conclusions. Numerical values used in this simulation were $\alpha = 200$ copies/day, $\beta = 0.4$ day^{-1}, $\delta = 800$ copies/day, $\epsilon = 1.6$ day^{-1}, and $\mu = 0.07$ day^{-1} for the dividing population.

equation is

$$
\begin{aligned}
\frac{\partial \rho(A_1, A_2, t)}{\partial t} &= -\frac{\partial}{\partial A_1}[(\alpha - \beta A_1)\rho(A_1, A_2, t)] \\
&+ \frac{1}{2}\frac{\partial^2}{\partial A_1^2}[(\alpha + \beta A_1)\rho(A_1, A_2, t)] \\
&- \frac{\partial}{\partial A_2}[(\delta - \epsilon A_2)\rho(A_1, A_2, t)] \\
&+ \frac{1}{2}\frac{\partial^2}{\partial A_2^2}[(\delta + \epsilon A_2)\rho(A_1, A_2, t)] \\
&- 2\mu\rho(A_1, A_2, t) + 8\mu\rho(A_1, A_2, t),
\end{aligned}
\tag{13}
$$

having a similar form to equation (12). However, A_2 also obeys a simple gene-product rate law, and we include an extra factor of 2 in the second to last term for homeostatic division.

A simulation of the equilibrium distribution of equation (13) is shown in figure 8. This is a two-dimensional extension of figure 6. The non-dividing population corresponding to the solid curve in figure 6 is depicted at top, and the correlated spread resulting from cell division shown at bottom. The asymmetry in the distribution is a result of the A_2 dynamics being more rapid than the A_1 dynamics ($\delta > \beta$).

We note that correlated fluctuations in expression levels are frequently used to infer the structure of genetic regulatory

[35, 36], signaling [37], and metabolic networks [38]. Failing to account for the effects of cell division in such an analysis can lead to the incorrect reconstruction of the genetic network. Spurious correlations between gene-products are strongest for pairs where both have slow degradation rates. Correlations in gene-product expression result very naturally from cell division. These correlations are typically grouped with other forms of *extrinsic noise* [31]. Population-expression models allow us to resolve the relative magnitude of different noise sources in extrinsic noise, potentially improving genetic regulatory network reconstruction methods.

3.3. Failure of single cell systems biology: selection bias

Consider now a two gene example with influx and selection. Here there is an initial population of cells localized around $\left(A_1^0, A_2^0\right)$. These cells have chemical dynamics such that at $t = 0$, A_1 begins to rapidly decrease and A_2 begins to gradually increase. The population dynamics that underlies selection in this system arises from changes in the rate of division and death of cells in a manner dependent on the concentrations of A_1 and A_2 within the cell. We set the division rate proportional to A_1 and the death rate proportional to A_2. The system also has a gradual influx of cells $\Gamma(A_1, A_2)$ entering the system around $\left(A_1^0, A_2^0\right)$. Here we do not consider the effects of dilution with cell division.

The system is described, using the vector notation, by

$$
\gamma_1 = \alpha - \beta A_1,
\tag{14}
$$

$$
\gamma_2 = \delta - \epsilon A_2,
\tag{15}
$$

$$
\frac{\partial \rho}{\partial t} = -\vec{\nabla} \cdot [\vec{\gamma}\rho] + \vec{\nabla} \cdot (\mathbf{D}\nabla\rho) + dA_1\rho - dA_2\rho + \Gamma(A_1, A_2),
\tag{16}
$$

where we have used terms introduced in Section 2 and omitted the dependence of terms on the quantities A_1 and A_2 for brevity.

Figure 9 shows the evolution of the density in the A_1, A_2 plane. We see that, by day 10, the initial population has proliferated and progressed along the differentiation pathway. By day 50 we see there has been considerable proliferation and the cells that are furthest along in the differentiation pathway have begun to decay. We also see the effects of the gradual influx of new cells at day 50 where the population now has a tail of recent immigrants that have proliferated. By day 100 the initial population has decayed completely and there is a stable distribution. This stable distribution is maintained by the influx of new cells and not by a lack of cell turnover. There is no change in 'phenotypic state' for the cells in this simulation, meaning cells maintain the same production and degradation values for A_1 and A_2 throughout the simulation.

The average differentiation path of a cell is given by the solid black curve in figure 9. Though the stable distribution is localized, it does not correspond to the phenotypic state described by the A_1, A_2 dynamics. The population dynamics gives a strong bias to the distribution that is not predicted by the chemical dynamics alone. In statistical physics it is common to use the 'fluctuation-dissipation theorem' to estimate model parameters from the equilibrium distribution.
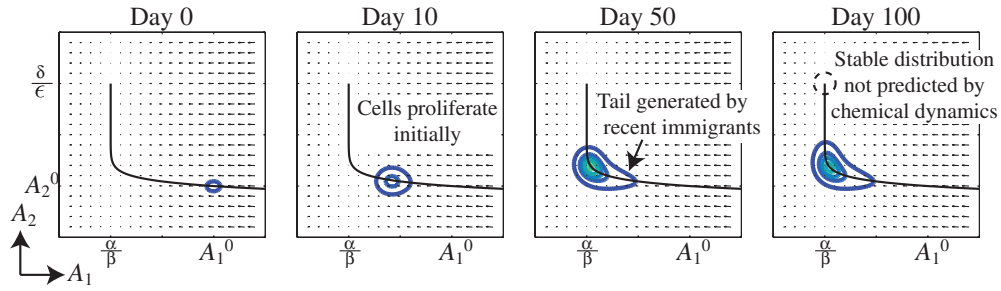
Figure 9. A two dimensional system with selection. Here cells proliferate with rate proportional to A_1 and die with rate proportional to A_2. An initial population localized around (A_1^0, A_2^0) in the lower right (Day 0) and proliferates as they begin to differentiate (Day 10). There is a constant but gradual influx of new cells entering the system around (A_1^0, A_2^0) that can be seen biasing the population by day 50, giving the population a tail of higher A_1 expression. The A_2 dynamics are slower and cells die before they ever reach the steady state predicted by the chemical dynamics (dashed black circle at Day 100). Instead a steady state that is a product of the population dynamics and the differentiation is reached which requires constant influx to maintain. The black curve illustrates the mean trajectory of the cells as predicted by the chemical dynamics. In this simulation $\alpha = 100$ copies/day, $\beta = 1$ day^{-1}, $\delta = 13.33$ copies/day, $\epsilon = 0.07$ day^{-1}, $d = 0.04$ copies^{-1}day^{-1}.

Any such analysis of a dynamic population must also take selective effects into account [39].

Rather than discuss the dynamics of A_1 and A_2 separately we can discuss the differentiation of cells moving along the one dimensional average path (black curve). To do this we introduce the variable a where cells enter the system at $a = 0$ and the differentiation pathway takes them toward $a = 1$; though, as seen in figure 9 Day 100, they may never reach $a = 1$. In this reduced model we also neglect the stochastic effects and the only heterogeneity in the system is due to the influx of new cells.

The one dimensional description is given by

$$\gamma = \epsilon - \epsilon a, \tag{17}$$

$$A_1(a) = \frac{\alpha}{\beta} - \left(\frac{\alpha}{\beta} - A_1^0\right)(1 - a)^{\beta/\epsilon}, \tag{18}$$

$$A_2(a) = A_2^0(1 - a) + \frac{\delta}{\epsilon}a, \tag{19}$$

$$\frac{\partial \rho(a, t)}{\partial t} = -\frac{\partial}{\partial a}[(\eta - \eta a)\rho(a, t)] + dA_1(a)\rho - dA_2(a)\rho + \Gamma(a = 0). \tag{20}$$

In this one-dimensional model we still have a distribution of cells since influx of $a = 0$ cells gives diversity to the system. In the absence of this influx we can describe the population with an ODE model where the population has an internal variable (a zero-dimensional approximation):

$$\frac{da}{dt} = \epsilon - \epsilon a, \tag{21}$$

$$\frac{dX(t)}{dt} = dA_1(a)X(t) - dA_2(a)X(t), \tag{22}$$

where $A_1(a)$ and $A_2(a)$ are described by equation (18) and (19). This approach was recently used to describe the exhaustion of CD8 T cells during a chronic infection where the internal variable corresponded to the level of exhaustion in the population and where thymic influx could be neglected [40].

4. Choosing the right variables

Traditional flow cytometry interrogates large numbers of cells. However the information from a single cell is limited by the spectral overlap of the fluorescent dyes to measuring the concentration of about 15 different molecules. Soon, new techniques such as Cy-TOF [41] (which merges mass-spectrometry with flow cytometry) will allow us to overcome this limitation and obtain simultaneous measurements of the concentration of hundreds of molecules at the single cell level. As the dimensionality increases, the techniques of population-expression modeling become computationally intractable. This necessitates dimensional reduction and identification of 'key players' among the measured molecular expressions. At the same time, even as we measure more and more quantities, some of the key players will still be omitted, forcing us to look for such important missing links.

In the simplest case, the expression dynamics for all chemical species in the system would be determined by a few key regulators, B_μ, i.e. $dA_i/dt = \alpha_i(\vec{B}, A_i) - \beta_i(\vec{B}, A_i) + \eta$, where α and β are the production/degradation functions, and η is the noise term. B_μ can be an individual chemical species, or more likely some function of many of the individual expressions. The goal is to find the minimal set $\vec{B}$ from data, or to understand if the data does not provide sufficient information to do so.

There is no single universal approach for dealing with large-dimensional data that would solve both of these problems in the immunological context. In fact the problems are not unique to immunology, or even to biology. Classic dimensionality reduction techniques include principal components analysis (PCA) [42], independent components analysis (ICA) [43], LASSO regression [44], and other approaches that explicitly identify (locally) linear subspaces spanned by data [45–47]. Many of these would be problematic in immunology since they measure importance by explained variance, which changes depending on the measurement units used. For example in PCA, using the measured brightness or its logarithm as the raw data may give very different results. The problem is solved elegantly with information-theoretic approaches, which are manifestly reparameterization invariant [48].
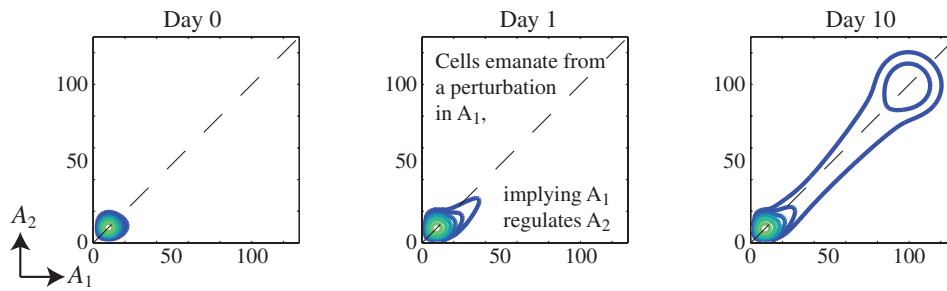
Figure 10. A two dimensional system of coupled stochastic biochemical species, with deterministic dynamics as in equations (23) and (24), with $\alpha_0 = 200$ copies/day, $\alpha_1 = 2000$ copies/day, $\beta = 22$ day^{-1}, $K = 30$ copies, $n = 6$. A_1 exhibits bistability. Since it controls the expression of A_2, the distribution of the latter is also bimodal. Notice the asymmetry of the contour plots of the joint probability distribution. By itself, such asymmetry, as in the central panel, simply signals unequal regulation of the two species. However, time series measurements will notice that the population average of dA_2/dt is correlated with the population average of A_1, but not the other way around. Graphically, this corresponds to the population escaping from the low expression steady state along the A_1 direction first, with A_2 following. This is a signal of the potential causal regulation $A_1 \rightarrow A_2$.

For this and related reasons, some of the most successful dimensionality reduction approaches in quantitative cell biology (and in computational neuroscience) have relied on information-theoretic techniques. For example, finding pairs of genes with high mutual information among their microarray mRNA expression profiles that cannot be explained away by confounding effects of other regulatory interactions uncovers 'minimal' transcriptional regulatory networks in cells as complex as human lymphocytes [35]. Higher order information-theoretic analyses [49] further disambiguate scenarios where simple pairwise interactions do not explain the data and more complex regulatory patterns are needed instead, (e.g. two or more factors regulating expression [50]). Similarly, searching for projections of the combinatorially complex stimulus space that preserve the information about the rate of spiking is one of the most powerful methods for finding receptive fields of neurons from electrophysiology data [51]. All of these approaches are special cases of the rate-distortion framework [48], where a 'small' description of data is sought that nonetheless preserves the information about the variable of relevance [48, 52]. The balance between the amount of information kept and the model size is controlled by the needs of the modeler and the data availability.

These methods should work for the context of immunology, but some changes are needed. First, typical immunology flow cytometry experiments make it hard to assay many different phenotypic or temporal conditions, as is typically used for information-theoretic analyses [35, 51]. This limits the range of variation of the data and can artificially reduce the values of the measured information quantities. Luckily, as demonstrated in [53], having many (tens of) thousands of single cell measurements allows accurate information estimation in these scenarios. However, it is *crucial* for the measurements to be of a *very* high accuracy.

The second distinction of immunological data is that, in the foreseeable future, the number of profiled quantities will be in the hundreds, but not in the thousands, with cell surface molecules being the easiest to profile. This leaves a possibility for missing key regulators in the data sets. As was demonstrated recently [53], information theoretic analysis can detect when such important regulators are missing. This is

done by observing that a missing regulator induces complex statistical dependences among all of its targets that cannot be explained by simple pairwise correlations [49]. While identification of such missing regulators in a semi-automated fashion is possible [50], the smaller dimensionality of the immunological data requires resetting the balance between the precision and the recall.

The third, and the most fundamental, distinction of immunological data is their population-expression nature. As illustrated in figure 8, cell division and death introduces spurious statistical relations among the measured expressions. Distinguishing effects of regulation versus population on the interactions among the measured variables should be possible by measuring the statistics of relations among physically non-interacting variables in experimental data and in numerical simulations.

Since development and differentiation of immune cells is fast and can be tracked in flow cytometry experiments on the scale of days, the data offers an ability to establish causality of regulation [54]. This is in contrast to identification of non-causal, symmetric relations among variables in most systems biology or computational neuroscience data analysis approaches. We illustrate this on the example of two coupled biochemical species obeying the deterministic dynamics

$$\frac{\mathrm{d}A_1}{\mathrm{d}t} = \alpha_0 + \frac{\alpha_1 A_1^n}{K^n + A_1^n} - \beta A_1, \qquad (23)$$

$$\frac{\mathrm{d}A_2}{\mathrm{d}t} = \alpha_0 + \frac{\alpha_1 A_1^n}{K^n + A_1^n} - \beta A_2. \qquad (24)$$

Here A_1 is self-regulating and can have two stable expression levels. A_2 is regulated by A_1 and will also be bimodal, but there is a clear difference between the two variables. Solution of the corresponding Fokker–Planck system is shown in figure 10, illustrating that the dynamics of the transient shapes of the joint probability distribution can signal the causality of regulatory relations.

This could be confirmed experimentally by sorting the cell population at an early time (e.g. day 1) into subpopulations based on expression levels. The contrast between the dynamics of the A_1-high, A_2-low subpopulation and the A_1-low, A_2-high subpopulation would reveal which is the driver of the system.

These sorted subpopulations would be placed into animals where they are recognized by unrelated genetic markers (e.g. Thy1.1) and monitored to see which subpopulation reaches A_1-high, A_2-high more rapidly.

5. Conclusion

Modeling in systems immunology is still in its infancy. Modeling requires identifying the key players and parameters that describe the behavior of interest. Population-expression models provide a tool for interpreting the changing expression profiles of multi-cellular populations that are differentiating while dividing and undergoing selection. They achieve this by connecting the population scale with intracellular systems biology.

The interpretation of immunological data has typically consisted of enumerating cellular phenotypes and describing how the sizes of these populations change over time. In contrast, the interpretation of data with population expression models focuses on the chemical interaction network common to all these phenotypes, and on the dependence of expression levels on division and death rates. One could instead continue adding additional phenotypic states to more accurately describe the data, but this is reminiscent of the 'epicycles on epicycles' used to described the motion of the planets in the Ptolemaic geocentric model of the universe. Looking at the problem differently can yield both simplicity and insight.

A complete view of systems biology would capture population dynamics, within-cell systems biology, and spatial effects. The spatial effects like clustering can occur at different scales. At the within-cell scale for example, clustering of molecules in the cell membrane plays an important role in the detection of antigen (infected cells) by T cells. At the population level, pathogens can be localized to the specific tissues and organs which they infect, while B and T cell responses occur in other sites such as the lymph nodes. Some spatial effects can be easily incorporated into the population-expression framework. The population-expression models are well suited to compartmentalization, where one considers a population-expression equation for different tissues and expression dependent trafficking rates between these compartments. For finer scale spatial effects, the population-expression approach breaks down, as the PDEs assume large numbers of cells in the compartments. In these low density regimes one must instead consider a model which treats cells discretely. In molecular systems biology, master equations and discrete stochastic simulations using Gillespie and related algorithms are very commonly used to describe the discreteness of stochastic changes in the phenotype of individual cells [55] alongside continuous Fokker–Planck and Langevin equation approaches. For methodological purposes, we built the current work around the population-expression analogue of the Fokker–Planck equation. However, it is clearly possible to develop the corresponding master equations and stochastic simulation algorithms, where the number of cells in a certain chemical state would be tracked. Nonlocal transitions due to cell division and related phenomena are not conceptually difficult to implement in such approaches, but the number of

types of possible transitions, and hence the time complexity of a simulation, might grow excessively because of the nonlocality. We leave the development of these simulation algorithms for future publications.

Advances in a field often require the integration of theoretical and experimental approaches. In the past the use of cellular dynamics data, such as flow cytometric data, typically allowed us to enumerate large numbers (millions) of cells but restricted us to making a handful of measurements on each cell, limiting the phenotypic resolution. The extension of traditional flow cytometry to Cy-TOF [41] allows the measurement of hundreds of biochemical species simultaneously at the single cell level. This allows, for the first time, tracking cellular systems biology dynamics and the population dynamics simultaneously and with high accuracy. The aim is to understand interactions among internal states of single cells and the composition of cellular populations, and hence the responses of the populations to infections. In this paper, we touched upon key problems that need to be addressed for such analysis: simultaneous representation of molecular system and population dynamics, including proliferation and cell death, and identification of key components of regulatory networks. We outlined a few ways in which these problems can be tackled computationally, by modifying current analysis approaches and by introducing population-expression modeling.

Acknowledgments

References

[1] Perelson A and Weisbuch G 1997 Immunology physicists *Rev. Mod. Phys.* **69** 1219–67

[2] Nowak M A and May R 2001 *Virus Dynamics: Mathematical Principles of Immunology and Virology* (Oxford: Oxford University Press)

[3] Perelson A S 2002 Modelling viral and immune system dynamics *Nature Rev. Immunol.* **2** 28–36

[4] Antia R, Ganusov V V and Ahmed R 2005 The role of models in understanding CD8+ T-cell memory *Nature Rev. Immunol.* **5** 101–11

[5] Akondy R S *et al* 2009 The yellow fever virus vaccine induces a broad and polyfunctional human memory CD8+ T cell response *J. Immunol.* **183** 7919–30

[6] Maecker H T, McCoy J P and Nussenblatt R 2012 Standardizing immunophenotyping for the human immunology project *Nature Rev. Immunol.* **12** 191–200

[7] Chakraborty A K and Das J 2010 Pairing computation with experimentation: a powerful coupling for understanding T cell signalling *Nature Rev. Immunol.* **10** 59–71

[8] Goldstein B, Faeder J R and Hlavacek W S 2004 Mathematical and computational models of immune-receptor signalling *Nature Rev. Immunol.* **4** 445–56

[9] Goldstein B, Coombs D, Faeder J R and Hlavacek W S 2008 Kinetic proofreading model *Adv. Exp. Med. Biol.* **640** 82–94

[10] Ferrell J Jr 2009 Q&A: systems biology *J. Biol.* **8** 2

[11] Munsky B, Trinh B and Khammash M 2009 Listening to the noise: random fluctuations reveal gene network parameters *Mol. Syst. Biol.* **5**

[12] Yates A, Callard R and Stark J 2004 Combining cytokine signalling with T-bet and GATA-3 regulation in Th1 and Th2 differentiation: a model for cellular decision-making *J. Theor. Biol.* **231** 181–96

[13] Banks H T, Charles F, Jauffret M D, Sutton K L and Thompson W C 2010 Label structured cell proliferation models *Appl. Math. Lett.* **23** 1412–5

[14] Stromberg S and Antia R 2012 On the role of CD8 T cells in the control of persistent infections *Biophys. J.* **103** 1802–10

[15] Metz J A J and Diekmann O 1986 The dynamics of physiologically structured populations *Lect. Notes Biomath.* **68**

[16] Elowitz M, Levine A, Siggia E and Swain P 2002 Stochastic gene expression in a single cell *Science* **297** 1183–6

[17] Raser J and O'Shea E 2004 Control of stochasticity in eukaryotic gene expression *Science* **304** 1811–4

[18] Van Kampen N G 2007 *Stochastic Processes in Physics and Chemistry* 3rd edn (Amsterdam: North-Holland)

[19] Paulsson J 2004 Summing up the noise in gene networks *Nature* **427** 415–8

[20] Munsky B 2012 Modeling cellular variability *Quantitative Biology: From Molecular to Cellular Systems* ed M E Wall (Boca Raton, FL: CRC Press)

[21] Berger M F and Bulyk M L 2009 Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors *Nature Protoc.* **4** 393–411

[22] Iyer-Biswas S, Hayot F and Jayaprakash C 2009 Stochasticity of gene products from transcriptional pulsing *Phys. Rev. E* **79** 031911

[23] Mugler A, Walczak A M and Wiggins C H 2009 Spectral solutions to stochastic models of gene expression with bursts and regulation *Phys. Rev. E* **80** 041921

[24] Ramos A, Innocentini G and Hornos J 2011 Exact time-dependent solutions for a self-regulating gene *Phys. Rev. E* **83** 062902

[25] Saini M, Sinclair C, Marshall D, Tolaini M, Sakaguchi S and Seddon B 2010 Regulation of Zap70 expression during thymocyte development enables temporal separation of CD4 and CD8 repertoire selection at different signaling thresholds *Sci. Signal.* **3** 23

[26] Mantzaris N V, Daoutidis P and Srienc F 2001 Numerical solution of multi-variable cell population balance models: I. Finite difference methods *Comput. Chem. Eng.* **25** 1411–40

[27] Mantzaris N V, Daoutidis P and Srienc F 2001 Numerical solution of multi-variable cell population balance models: II. Spectral methods *Comput. Chem. Eng.* **25** 1441–62

[28] Mantzaris N V, Daoutidis P and Srienc F 2001 Numerical solution of multi-variable cell population balance models: III. Finite element methods *Comput. Chem. Eng.* **25** 1463–81

[29] Revy P, Sospedra M, Barbour B and Trautmann A 2001 Functional antigen-independent synapses formed between T cells and dendritic cells *Nature Immunol.* **2** 925–31

[30] De Boer R, Ganusov V, Milutinović D, Hodgkin P and Perelson A 2006 Estimating lymphocyte division and death rates from CFSE data *Bull. Math. Biol.* **68** 1011–31

[31] Swain P S, Elowitz M B and Siggia E D 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression *Proc. Natl Acad. Sci.* **99** 12795–800

[32] Rosenfeld N, Young J W, Alon U, Swain P S and Elowitz M B 2005 Gene regulation at the single-cell level *Science* **307** 1962–5

[33] Rosenfeld N, Perkins T J, Alon U, Elowitz M B and Swain P S 2006 A fluctuation method to quantify *in vivo* fluorescence data *Biophys. J.* **91** 759–66

[34] Ozbudak E M, Thattai M, Lim H N, Shraiman B I and van Oudenaarden A 2004 Multistability in the lactose utilization network of *Escherichia coli Nature* **427** 737–40

[35] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R and Califano A 2006 ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context *BMC Bioinform.* **7** (Suppl. 1) S7

[36] Stewart-Ornstein J, Weissman J S and El-Samad H 2012 Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae Mol. Cell.* **45** 483–93

[37] Sachs K, Perez O, Pe'er D, Lauffenburger D and Nolan G 2005 Causal protein-signaling networks derived from multiparameter single-cell data *Science* **308** 523–9

[38] Nemenman I, Escola G S, Hlavacek W S, Unkefer P J, Unkefer C J and Wall M E 2007 Reconstruction of metabolic networks from high-throughput metabolite profiling data: in silico analysis of red blood cell metabolism *Ann. New York Acad. Sci.* **1115** 102–15

[39] Lehner B and Kaneko K 2011 Fluctuation and response in biology *Cell Mol. Life Sci.* **68** 1005–10

[40] Johnson P L F, Kochin B F, McAfee M S, Stromnes I M, Regoes R R, Ahmed R, Blattman J N and Antia R 2011 Vaccination alters the balance between protective immunity, exhaustion, escape, and death in chronic infections *J. Virol.* **85** 5565–70

[41] Bendall S C *et al* 2011 Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum *Science* **332** 687–96

[42] Jolliffe I 2002 *Principal Component Analysis (Springer Series in Statistics)* 2nd edn (Berlin: Springer)

[43] Lee T-W 1998 *Independent Component Analysis: Theory and Applications* (Boston, MA: Kluwer)

[44] Tibshirani R 1996 Regression shrinkage and selection via the LASSO *J. R. Stat. Soc.* B **58** 267–88

[45] Blatt M, Wisemann S and Domany E 1997 Data clustering using a model granular magnet *Neural Comput.* **9** 1805–42

[46] Tenenbaum J, de Silva V and Langford J 2000 A global geometric framework for nonlinear dimensionality reduction *Science* **290** 2319–23

[47] Roweis S and Saul L 2000 Nonlinear dimensionality reduction by locally linear embedding *Science* **290** 2323–6

[48] Cover T and Thomas J 2006 *Elements of Information Theory* 2nd edn (New York: Wiley)

[49] Margolin A, Wang K, Califano A and Nemenman I 2010 Multivariate dependence and genetic networks inference *IET Syst. Biol.* **4** 428–40

[50] Wang K *et al* 2009 Genome-wide identification of post-translational modulators of transcription factor activity in human B cells *Nature Biotechnol.* **27** 829–39

[51] Sharpee T, Rust N and Bialek W 2004 Analyzing neural responses to natural signals: maximally informative dimensions *Neural Comput.* **16** 223–50

[52] Tishby N, Pereira F and Bialek W 2000 The information bottleneck method arXiv:physics/0004057

[53] Cheong R, Rhee A, Wang C J, Nemenman I and Levchenko A 2011 Information transduction capacity of noisy biochemical signaling networks *Science* **334** 354–8

[54] Munsky B, Neuert G and van Oudenaarden A 2012 Using gene expression noise to understand gene regulation *Science* **336** 183–7

[55] Gillespie D T 2007 Stochastic simulation of chemical kinetics *Annu. Rev. Phys. Chem.* **58** 35–55