Stochastic Physics of the Single Cell: Ergodicity, Prior Probability, and Bayesian Inference

---

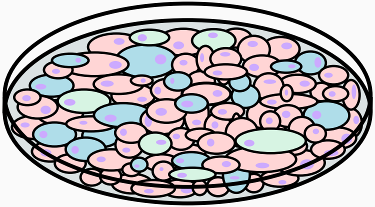
Erin Angelini

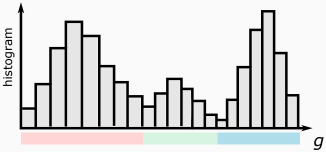
January 17, 2023
2023 Stochastic Physics in Biology GRC

Department of Applied Mathematics
University of Washington

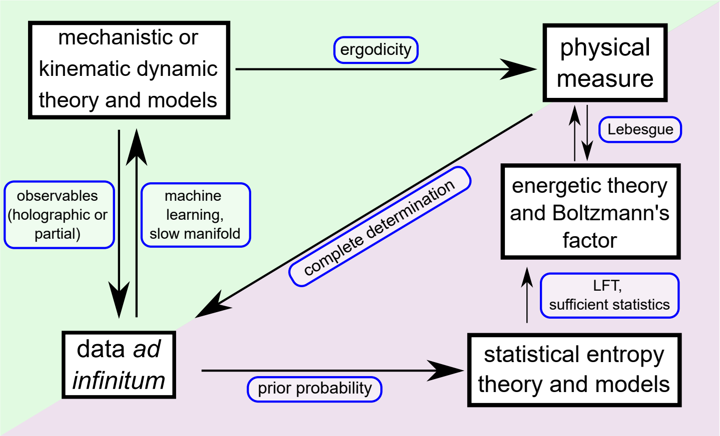
Statistical treatment of single cells *in vitro*

How do we mathematically describe phenotypic heterogeneity? What is the “correct” statistical framework for single-cell data?





Statistics & Gibbs' theory: stochastic thermodynamics for data



mechanistic or kinematic dynamic theory and models

ergodicity

physical measure

observables (holographic or partial)

machine learning, slow manifold

Lebesgue

energetic theory and Boltzmann's factor

complete determination

LFT, sufficient statistics

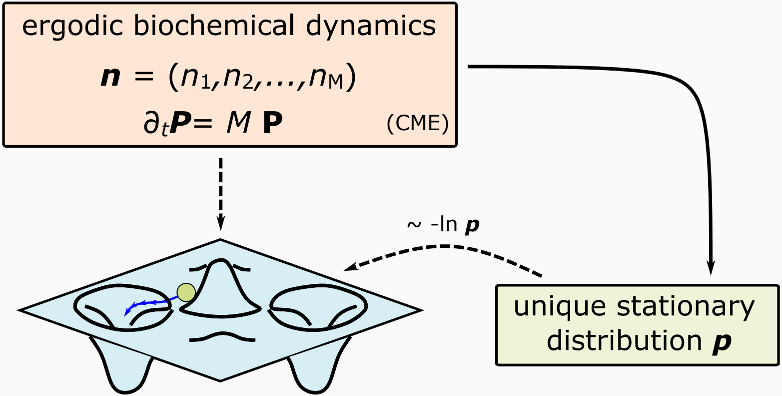
data *ad infinitum*

prior probability

statistical entropy theory and models

E. Angelini and H. Qian. "Statistical Analysis of Random Motion and Energetic Behavior of Counting: Gibbs' Theory Revisited" (2023). Manuscript under review.

Ergodicity and the idealized “eternal cell”



ergodic biochemical dynamics

$$\boldsymbol{n} = (n_1, n_2, \ldots, n_\mathrm{M})$$

$$\partial_t \boldsymbol{P} = M\ \mathbf{P} \qquad \text{(CME)}$$

$\sim$ -ln $\boldsymbol{p}$

unique stationary distribution $\boldsymbol{p}$

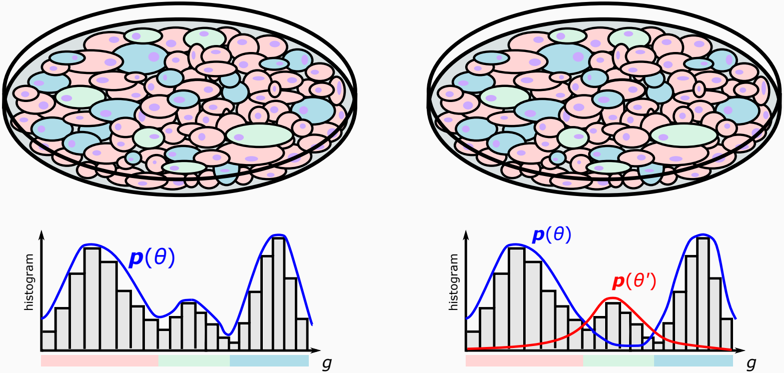
non-equilibrium
steady-state landscape

Wang *et al.* (2008)
PNAS

Zhou *et al.* (2012)
J. R. Soc. Interface

Genomic variability: separation of timescales

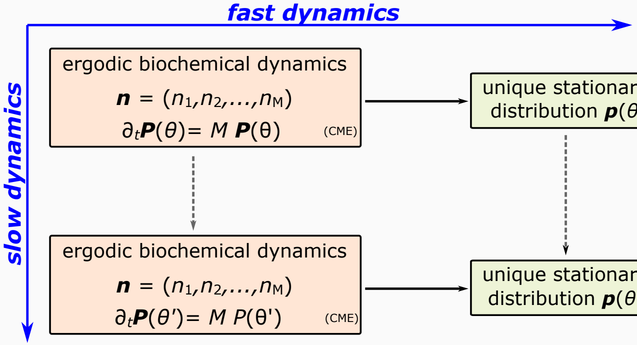
θ: (high-dimensional) parameter, encodes “genomic state” of a cell.



genomically identical -vs- genomically non-identical

Genetic variability: separation of timescales

There is a natural separation of timescales when considering the phenotypic/biochemical state vs. the genomic state of a single cell:



fast dynamics

slow dynamics

ergodic biochemical dynamics

$\boldsymbol{n} = (n_1, n_2, \ldots, n_M)$

$\partial_t \boldsymbol{P}(\theta) = M\, \boldsymbol{P}(\theta)$ (CME)

unique stationary distribution $\boldsymbol{p}(\theta)$

ergodic biochemical dynamics

$\boldsymbol{n} = (n_1, n_2, \ldots, n_M)$

$\partial_t \boldsymbol{P}(\theta') = M\, P(\theta')$ (CME)

unique stationary distribution $\boldsymbol{p}(\theta')$

Bayesian inference for genomic parameters

- $X(t) \in \mathcal{D}_X$: fast-varying variable
    - e.g., expression levels of a biomarker
- $\Theta(t) \in \mathcal{D}_\Theta$: slow-varying variable
    - e.g., genetic/genomic parameter
    - assume fixed: quasi steady state
- $f_{X,\Theta}(x, \theta)$: joint stationary probability density
- $p(x|\theta)$: stationary probability of X given fixed $\Theta(t) = \theta$
- $f_\Theta(\theta)$: marginal probability density (prior)

Bayesian inference for two-timing ergodic dynamics

Suppose we have n i.i.d. measurements $\mathsf{x} = (x_1, x_2, \ldots, x_n)$ on the fast variable X, generated by the process for some fixed $\Theta(t) = \theta^*$ (unknown). We update the posterior density for Θ via Bayes rule:

$$f_\Theta(\theta|\mathsf{x}) = \frac{\Pr(\mathsf{x}|\theta) f_\theta(\theta)}{\int_{\mathcal{D}_\Theta} \Pr(\mathsf{x}|\theta) f_\Theta(\theta) \mathrm{d}\theta} = \frac{f_\theta(\theta)}{Z(\mathsf{x})} \prod_{i=1}^{n} p(x_i|\theta)$$

Bayesian inference for two-timing ergodic dynamics

Suppose we have n i.i.d. measurements $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ on the fast variable X, generated by the process for some fixed $\Theta(t) = \theta^*$ (unknown). We update the posterior density for Θ via Bayes rule:

$$
\begin{aligned}
f_\Theta(\theta|\mathbf{x}) &= \frac{\Pr(\mathbf{x}|\theta) f_\theta(\theta)}{\int_{\mathcal{D}_\Theta} \Pr(\mathbf{x}|\theta) f_\Theta(\theta) \mathrm{d}\theta} = \frac{f_\theta(\theta)}{Z(\mathbf{x})} \prod_{i=1}^{n} p(x_i|\theta) \\
&= \frac{f_\Theta(\theta)}{Z(\mathbf{x})} \exp\left[\sum_{i=1}^{n} \ln p(x_i|\theta)\right]
\end{aligned}
$$

Bayesian inference for two-timing ergodic dynamics

Suppose we have n i.i.d. measurements $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ on the fast variable X, generated by the process for some fixed $\Theta(t) = \theta^*$ (unknown). We update the posterior density for Θ via Bayes rule:

$$\begin{aligned} f_\Theta(\theta|\mathbf{x}) &= \frac{\Pr(\mathbf{x}|\theta) f_\theta(\theta)}{\int_{\mathcal{D}_\Theta} \Pr(\mathbf{x}|\theta) f_\Theta(\theta) \mathrm{d}\theta} = \frac{f_\theta(\theta)}{Z(\mathbf{x})} \prod_{i=1}^{n} p(x_i|\theta) \\ &= \frac{f_\Theta(\theta)}{Z(\mathbf{x})} \exp\left[\sum_{i=1}^{n} \ln p(x_i|\theta)\right] \\ &\approx \frac{f_\Theta(\theta)}{Z(\mathbf{x})} \exp\left[n \sum_{j=1}^{m} \nu_j(\mathbf{x}; \theta^*) \ln p(\xi_j|\theta)\right] \end{aligned}$$

$\{\xi_0, \xi_1, \ldots, \xi_m\}$ partitions state space $\mathcal{D}_X \subseteq \mathbb{R}$ into m bins

$\nu_j(\mathbf{x}; \theta^*)$ = relative frequency of data points x_i in interval $(\xi_{j-1}, \xi_j]$

Bayesian inference for two-timing ergodic dynamics

Suppose we have n i.i.d. measurements $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ on the fast variable X, generated by the process for some fixed $\Theta(t) = \theta^*$ (unknown). We update the posterior density for Θ via Bayes rule:

$$\begin{aligned} f_\Theta(\theta|\mathbf{x}) &= \frac{\Pr(\mathbf{x}|\theta) f_\theta(\theta)}{\int_{\mathcal{D}_\Theta} \Pr(\mathbf{x}|\theta) f_\Theta(\theta) \mathrm{d}\theta} = \frac{f_\theta(\theta)}{Z(\mathbf{x})} \prod_{i=1}^{n} p(x_i|\theta) \\ &= \frac{f_\Theta(\theta)}{Z(\mathbf{x})} \exp\left[\sum_{i=1}^{n} \ln p(x_i|\theta)\right] \\ &\approx \frac{f_\Theta(\theta)}{Z(\mathbf{x})} \exp\left[n \sum_{j=1}^{m} \nu_j(\mathbf{x}; \theta^*) \ln p(\xi_j|\theta)\right] \\ &\sim \frac{f_\Theta(\theta)}{Z(\theta^*)} \exp\left[n \int_{\mathcal{D}_X} p(y|\theta^*) \ln p(y|\theta) \mathrm{d}y\right] \end{aligned}$$

as $m, n \to \infty$. In particular, as $n \to \infty$, $f_\Theta \to$ Dirac δ centered at θ^*!

Summary

Stochastic physics/mathematics for cellular heterogeneity

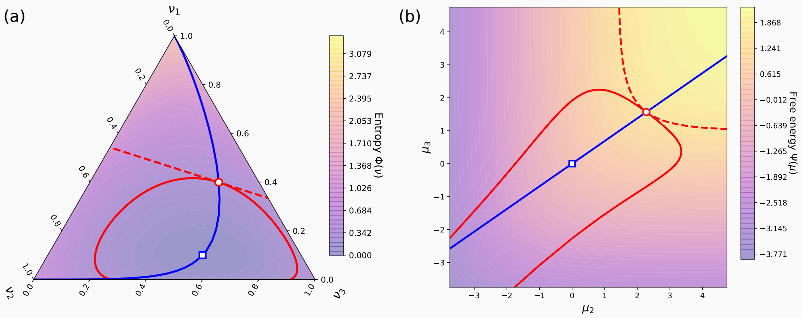
- biochemical dynamics $\xrightarrow{\textit{ergodicity}}$ unique stationary probability for a single cell
  - directly related to non-equilibrium steady-state landscape for cellular differentiation
- genetic vs. non-genetic variations: slow v. fast time scale
  - statistical inference (Bayesian) of genetic parameter from measurements of a biomarker

Questions?

This work is supported by NIH grant R01GM135396 (PI: Dr. Sui Huang).

$$\begin{tikzcd}
\Phi(\boldsymbol{\nu}) \arrow[r, leftrightarrow, "\mathbf{x} = \boldsymbol{\nu}\mathbf{G}"] \arrow[d, leftrightarrow, "\mathrm{LFT}"] & \Phi_{\boldsymbol{g}}(\mathbf{x}) \arrow[r, "\text{minimization}"] \arrow[d, leftrightarrow, "\mathrm{LFT}"] & \phi(\mathbf{x}') \arrow[d, leftrightarrow, "\mathrm{LFT}"'] \\
\Psi(\boldsymbol{\mu}) \arrow[r, leftrightarrow, "\boldsymbol{\mu} = \mathbf{G}\mathbf{y}"] & \Psi_{\boldsymbol{g}}(\mathbf{y}) \arrow[r, "\text{projection}"] & \psi(\mathbf{y}')
\end{tikzcd}$$

$$\underbrace{\hphantom{\Psi(\boldsymbol{\mu}) \longleftrightarrow \Psi_{\boldsymbol{g}}(\mathbf{y})}}_{\text{holographic rep.}} \quad \underbrace{\hphantom{\Psi_{\boldsymbol{g}}(\mathbf{y}) \longrightarrow \psi(\mathbf{y}')}}_{\text{coarse-graining}}$$


(a)
ν_1
ν_2
ν_3
0.0
0.2
0.4
0.6
0.8
1.0
Entropy $\Phi(\nu)$
3.079
2.737
2.395
2.053
1.710
1.368
1.026
0.684
0.342
0.000
(b)
μ_2
μ_3
−3
−2
−1
0
1
2
3
4
Free energy $\Psi(\mu)$
1.868
1.241
0.615
−0.012
−0.639
−1.265
−1.892
−2.518
−3.145
−3.771

References i

E. Angelini and H. Qian.
Statistical analysis of random motion and energetic behavior of counting: Gibbs' theory revisited.
Manuscript submitted for publication, 2023.

J. Wang, L. Xu, and E. Wang.
Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations.
PNAS, 105(34):12271–12276, 2008.

J. X. Zhou, M. Aliyu, E. Aurell, and S. Huang.
Quasi-potential landscape in complex multi-stable systems.
Journal of the Royal Society Interface, 9(77):3539–3553, 2012.