Large deviations theory: Data, single-cell biomarkers, and Fisher's fundamental theorem of natural selection

Hong Qian

Department of Applied Mathematics
University of Washington

Genomics and the physical chemistry of biomolecules are two pillars of molecular biology. One of the key branches of the latter is *chemical thermodynamics*, which, curiously, is quite marginalized in the current research on computational molecular biology (CMB). I shall revisit this subject, but no starting from physics nor chemistry, but rather through a result from probability, beyond the law of large numbers and central limit theorem, call *large deviations theory*. I shall show how our theory can be applied to Fisher's FTNS as well as to analyzing data from single cells.

Prologue: The two pillars of molecular biology

- Genetics, genomics, and *bioinformatics*:

  It is about “information”;

- Physical chemistry, *molecular dynamics (MD) and structural biology (SB)*:

  It is abut “molecules, their states, and the processes cause their changes”.

CHEMISTRY

10th EDITION

Reinforced Binding

CHANG

Thermochemistry 228

Chemical Kinetics 556

Things like …

$$G = H - TS, \quad \frac{\partial G}{\partial c_i} = \mu_i$$

$$\Delta\mu = \Delta\mu^o + kT\ln\frac{[C][D]}{[A][B]}$$

$$\Delta\mu^o = -kT\ln K_{\mathrm{eq}}$$

(1)

How to represent (*e.g.*, describe) a biochemical or a biological systems, not just MD and SB: Its states and its changes?

(i) classifying biochemical (or biological) individuals into populations of kinetic species;

(ii) counting the number of individuals in each and every "pure kinetic species";

(iii) representing changes in terms of "stochastic elementary processes".

A stochastic elementary process

$$A + 2B + C + E \xrightarrow{r_j} X + 3B + E$$

$$\boldsymbol{N} = (N_A, N_B, \dots, N_Z)$$

$$P\{\boldsymbol{N}(t + dt) = \boldsymbol{n} + \Delta\boldsymbol{n} | \, \boldsymbol{N}(t) = \boldsymbol{n}\}$$

$$= \begin{cases} r_j(\boldsymbol{n})dt + o(dt), if \ \Delta\boldsymbol{n} = \ \boldsymbol{\nu}_j \\ 1 - rdt + o(dt), if \ \Delta\boldsymbol{n} = \boldsymbol{0} \\ 0, \qquad\qquad\qquad\qquad otherwise. \end{cases}$$

A stochastic elementary process

$$A + 2B + C + E \xrightarrow{r_j} X + 3B + E$$

$$\boldsymbol{N} = (N_A, N_B, \ldots, N_Z)$$

instantaneous rate function

$$P\{\boldsymbol{N}(t+dt) = \boldsymbol{n} + \Delta\boldsymbol{n} | \boldsymbol{N}(t) = \boldsymbol{n}\}$$

$$= \begin{cases} r_j(\boldsymbol{n})dt + o(dt), if\ \Delta\boldsymbol{n} = \boldsymbol{\nu}_j \\ 1 - rdt + o(dt), if\ \Delta\boldsymbol{n} = \boldsymbol{0} \\ 0, \qquad otherwise. \end{cases}$$

stoichiometric coefficients

An essential feature of a *stochastic elementary reaction* with *pure kinetic species*

$$P\{\mathbf{T} \geq t\} = e^{-rt}$$

$$\frac{P\{\mathbf{T} \geq t+\tau\}}{P\{\mathbf{T} \geq \tau\}} = \frac{e^{-r(t+\tau)}}{e^{-t\tau}} = e^{-rt}$$

$$-\frac{d \ln P\{\mathbf{T} \geq t\}}{dt} = r$$

The rate for the next event within multiple *stochastic elementary processes* among *pure kinetic species*

$$P\{\mathbf{T}_j \geq t\} = e^{-r_j t} \quad (1 \leq j \leq M)$$

$$\mathbf{T}_* = \min\{\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_M\},$$

$$P\{\mathbf{T}_* \geq t\} = e^{-r_* t},$$

$$r_* = r_1 + r_2 + \cdots + r_M.$$

A theorem on the rate for complex kinetic species that contain heterogeneous subpopulations

$$P\{\mathbf{T} \geq t\} = \int_0^\infty e^{-r(s)t} f(s)ds$$

$$-\frac{d \ln P\{\mathbf{T} \geq t\}}{dt} = \frac{\int_0^\infty r(s)e^{-r(s)t} f(s)ds}{\int_0^\infty e^{-r(s)t} f(s)ds} = \bar{r}(t)$$

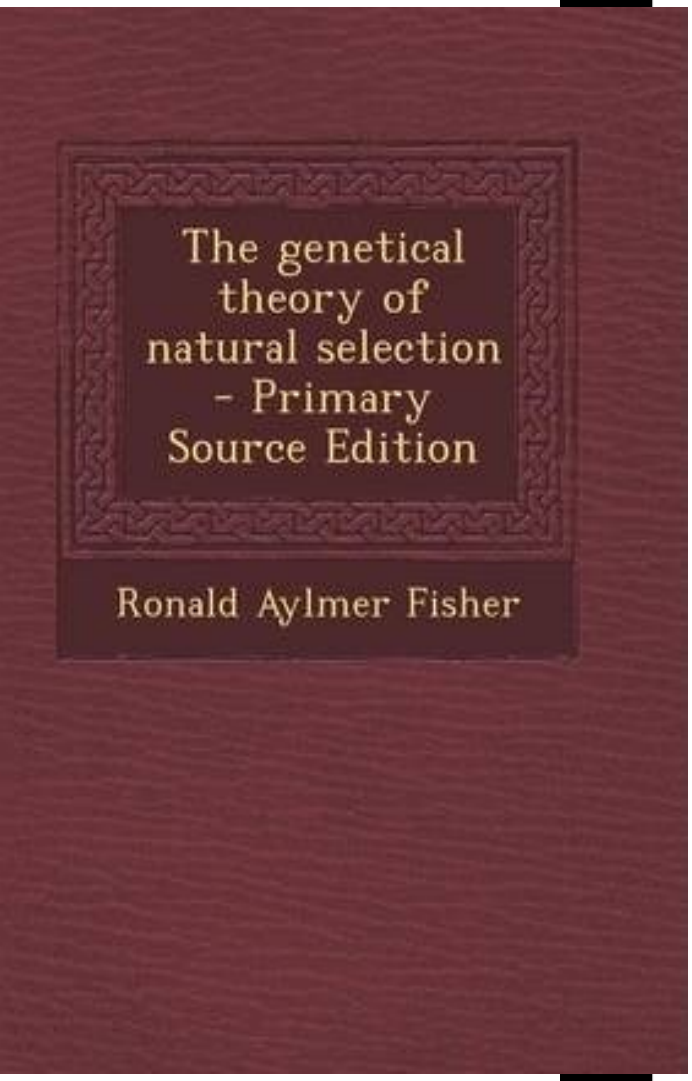
$$\frac{d\bar{r}(t)}{dt} = -\overline{[r(s) - \bar{r}(t)]^2} \leq 0.$$

The instantaneous rate for a non-elementary process with complex kinetic species:

$$\bar{r}[\mathbf{N}(t), X(t), t]$$

$$\frac{d\bar{r}(t)}{dt} = \frac{\partial \bar{r}}{\partial \mathbf{N}}\left(\frac{d\mathbf{N}}{dt}\right) + \frac{\partial \bar{r}}{\partial X}\left(\frac{dX}{dt}\right) + \frac{\partial \bar{r}}{\partial t}$$

$$\frac{\partial \bar{r}}{\partial t} \leq 0\,!$$



and, taking all factors into consideration, the total increase in fitness,

$$\Sigma(\alpha dp) = \Sigma(pq\alpha\alpha)dt = Wdt.$$

If therefore the time element dt is positive, the total change of fitness Wdt is also positive, and indeed the rate of increase in fitness due to all changes in gene ratio is exactly equal to the genetic variance of fitness W which the population exhibits. We may consequently state the fundamental theorem of Natural Selection in the form:

The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.

The rigour of the demonstration requires that the terms employed should be used strictly as defined; the ease of its interpretation may be increased by appropriate conventions of measurement. For example, the ratio $p : q$ should strictly be evaluated at any instant by the enumeration, not necessarily of the census population, but of all individuals having reproductive value, weighted according to the reproductive value of each.

Since the theorem is exact only for idealized populations, in which fortuitous fluctuations in genetic composition have been excluded, it is important to obtain an estimate of the magnitude of the effect of these fluctuations, or in other words to obtain a standard error appropriate to the calculated, or expected, rate of increase in fitness. It will be sufficient for this purpose to consider the special case of a population mating and reproducing at random. It is easy to see that if such chance fluctuations cause a difference δp between the actual value of p obtained in any generation and that expected, the variance of δp will be

$$\frac{pq}{2n},$$

where n represents the number breeding in each generation, and $2n$

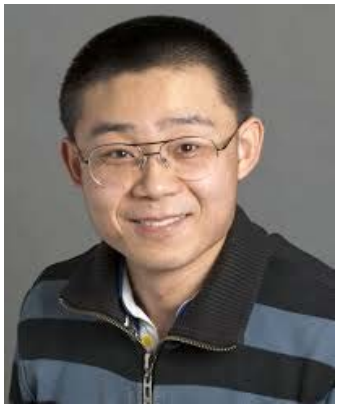
Counting, it is everywhere!

- Newtonian particles in fluid mechanics, Lagrangian vs. Eulerian representations;
- second quantization in quantum field theory;
- atomic numbers in molecules, chemical species;
- biological organisms.

Lu and Qian, *arXiv:*2009.12644 (2020)

(2)
Thermodynamics of Small Systems

Joint work with Prof. Zhiyue Lu, UNC.



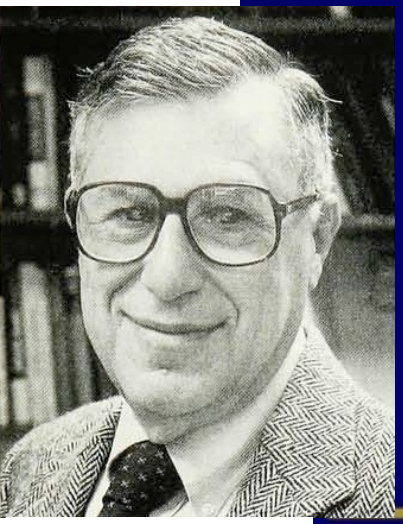
What are?

Fundamental thermodynamic relation

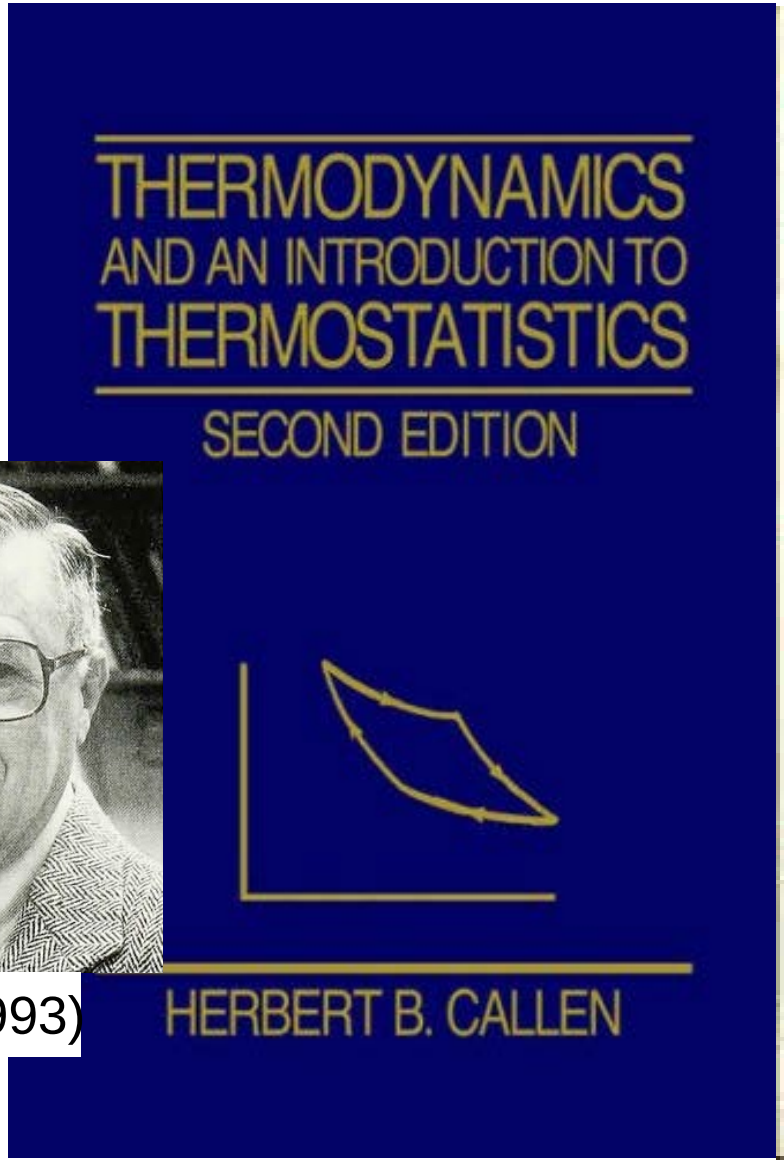
Gibbs–Duhem equation

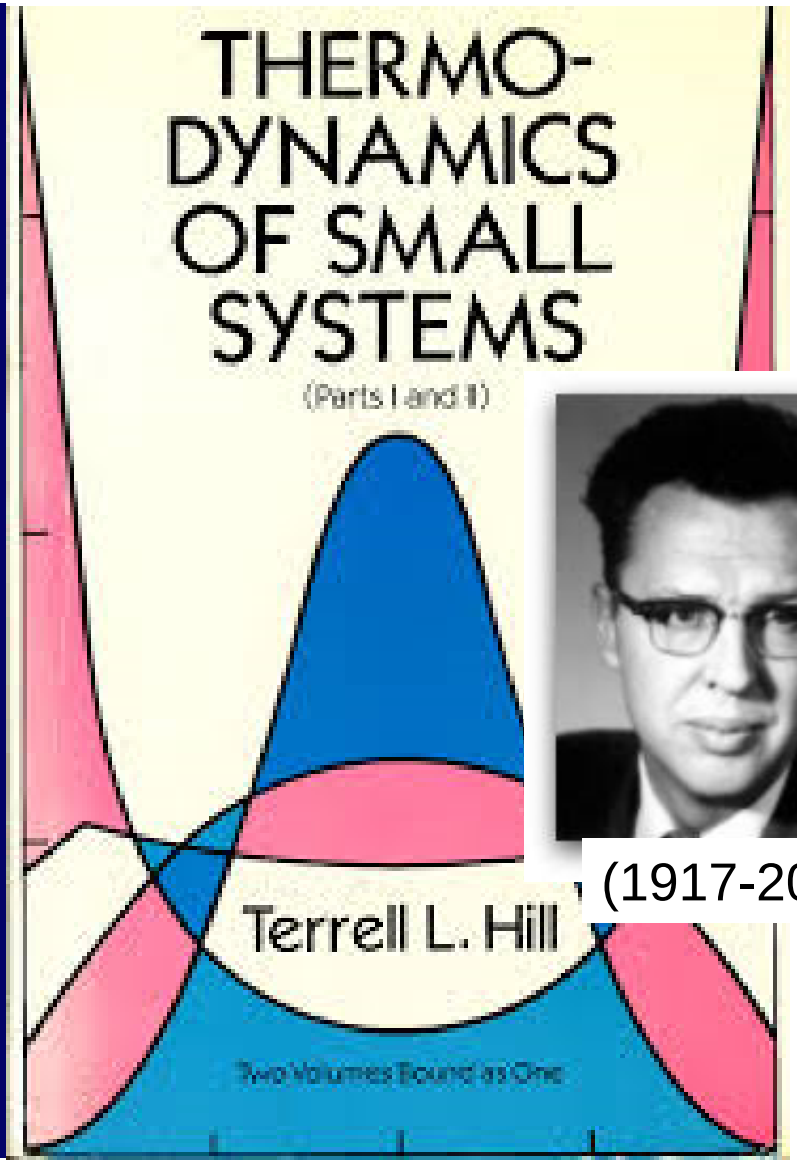
Current understanding of the foundation of *thermodynamics*

- The existence of an *entropy function* or *entropy functional*.
- The *entropy* is a statistical concept; it is an Eulerian *homogeneous function* of all extensive variables with order 1.
- It is a universal result, as a limit, for *macroscopic* large systems.



(1919-1993)





(1917-2014)

How can a *small system* have universal behavior?

The world is stochastic.

The repeated measurements are *not* a way to obtaining truth via eliminating uncertainty. Variation (heterogeneity) *is* a part of the truth!

On the border of this stochastic world, three major landmarks:

LLN: *law of large numbers,*
CLT: *central limit theorem,*
LDP: *large deviations principle.*

Law of Large Numbers

$$\lim_{M \to \infty} \frac{X_1 + X_2 + \cdots + X_M}{M} = \mathbb{E}[X]$$

$$\lim_{M \to \infty} \frac{m_k}{M} = \mathbb{P}_k \; (k = 1,2, \ldots, K)$$

La Nur

stationary samples

expected value

$$\lim_{M\to\infty} \frac{X_1 + X_2 + \cdots + X_M}{M} = \mathbb{E}[X]$$

sample frequency

$$\lim_{M\to\infty} \frac{m_k}{M} = \mathbb{P}_k \, (k = 1,2, \ldots, K)$$

probability

- *statistical concepts*
- *probabilistic concepts*

Central Limit Theorem

$$\lim_{M\to\infty} \frac{X_1 + X_2 + \cdots + X_M}{\sqrt{M}} = \infty$$

$$\lim_{M\to\infty} \left\{ \frac{X_1 + X_2 + \cdots + X_M}{\sqrt{M}} - \sqrt{M}\,\mathbb{E}[X] \right\}$$

$$= \mathcal{N}(0, \sigma^2)$$

$\boxed{\infty}$

Central Limit Theorem

$$\lim_{M\to\infty} \frac{X_1 + X_2 + \cdots + X_M}{\sqrt{M}} = \infty$$

$$\lim_{M\to\infty} \left\{ \frac{X_1 + X_2 + \cdots + X_M}{\sqrt{M}} - \sqrt{M}\,\mathbb{E}[X] \right\}$$

$$= \mathcal{N}(0, \sigma^2)$$

normal random variable

variance of X

Large Deviations Principle

$$f_{\bar{X}_M}(x;M) \to \delta(x - x^*), x^* = \mathbb{E}[X]$$

the premise

$$f_{\bar{X}_M}(x;M) \sim e^{-M\varphi(x)},$$

the existence

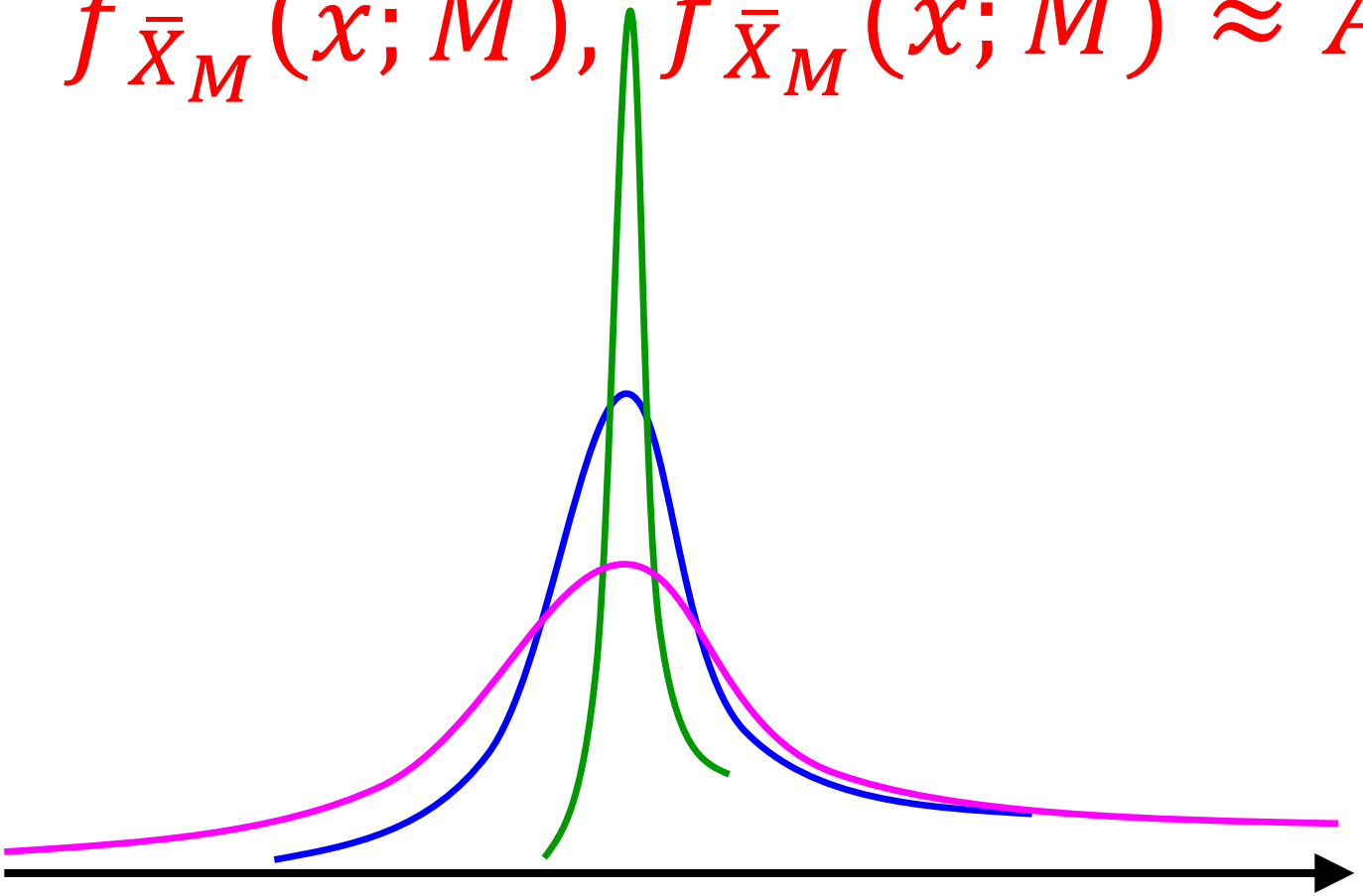
$$\min_{x \in \mathfrak{S}} \varphi(x) = \varphi(x^*) = 0,$$

$$\lim_{M \to \infty} -\frac{1}{M} \ln \mathbb{P}\{X_M \in \mathcal{A} \subset \mathfrak{S}\} = \min_{x \in \mathcal{A}} \varphi(x).$$

its property

When the M tends to infinite …

$$f_{\bar{X}_M}(x;M),\ f_{\bar{X}_M}(x;M) \approx A\, e^{-M\varphi(x)}$$

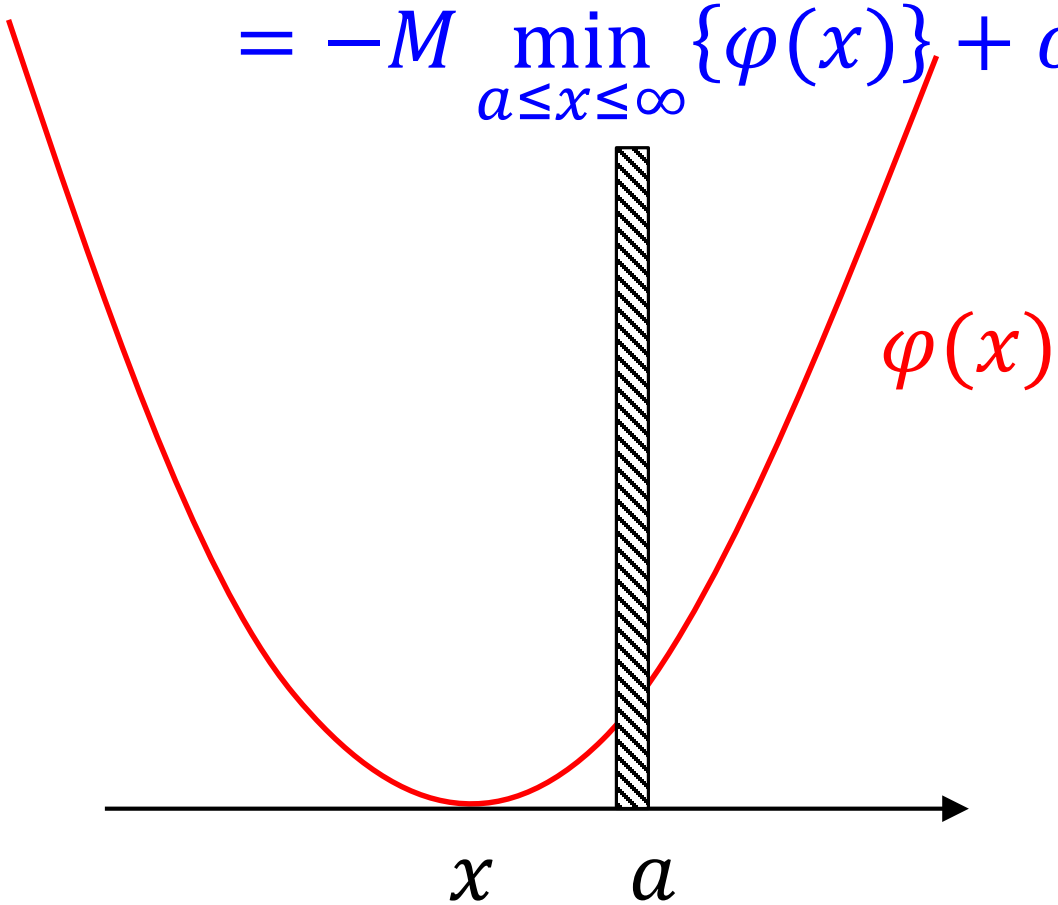


sample mean value x

rate function φ as entropy

$$\ln \Pr\{\bar{X}_M \geq a\} = \ln \int_a^\infty A e^{-M\varphi(x)} \mathrm{d}x$$

$$= -M \min_{a \leq x \leq \infty} \{\varphi(x)\} + o(M)$$



Cramér's theorem for arbitrary distribution $p_X(x)$

Harald Cramér (1893-1985)

Cumulant Generating Function and Legendre-Fenchel transform

$$\psi(\beta) = \ln \int p_X(x) e^{-\beta x} dx$$

$$\boxed{\psi(\boldsymbol{\beta}) = \sup_{\boldsymbol{x}} \{-\boldsymbol{\beta} \cdot \boldsymbol{x} + \eta(\boldsymbol{x})\}}$$

$$\boxed{-\varphi(x) = \inf_{\beta} \{\boldsymbol{\beta} \cdot \boldsymbol{x} + \psi(\boldsymbol{\beta})\}}$$

Massieu-Guggenheim (free) entropy ψ and Gibbs entropy η

$$\psi(\beta) = \ln \int p_X(x) e^{-\beta x} dx$$

Massieu-Guggenheim entropy

$$\boxed{-\varphi(x) = \inf_{\beta}\{\boldsymbol{\beta} \cdot \boldsymbol{x} + \psi(\boldsymbol{\beta})\}}$$

$$\eta(x) = -\varphi(x) = \frac{d[\psi(\beta)/\beta]}{d[1/\beta]}$$

Gibbs entropy

Legendre-Fenchel Transform and Free Entropy-Gibbs Entropy Duality

$$\eta(\boldsymbol{y}) = \inf_{\boldsymbol{\beta}}\{\boldsymbol{\beta}\cdot\boldsymbol{y} + \psi(\boldsymbol{\beta})\}$$

$$d\eta(\boldsymbol{y}) = \boldsymbol{\beta}\cdot d\boldsymbol{y} = \beta_1 dy_1 + \beta_2 dy_2 + \cdots$$

fundamental thermodynamic relation

Hill-Gibbs-Duhem equation

$$\psi(\boldsymbol{\beta}) = \sup_{\boldsymbol{y}}\{-\boldsymbol{\beta}\cdot\boldsymbol{y} + \eta(\boldsymbol{y})\}$$

$$d\psi(\boldsymbol{\beta}) = -\boldsymbol{y}\cdot d\boldsymbol{\beta} = -y_1 d\beta_1 - y_2 d\beta_2 + \cdots$$

T. L. Hill's Nanothermodynamics

$$S(U,V,N) = \left(\frac{1}{T}\right)U + \left(\frac{p}{T}\right)V - \left(\frac{\mu}{T}\right)N - \left(\frac{1}{T}\right)\mathcal{E}$$

extensive quantities

sub-extensive

$$\mathrm{d}S(U,V,N) = \left(\frac{1}{T}\right)\mathrm{d}U + \left(\frac{p}{T}\right)\mathrm{d}V - \left(\frac{\mu}{T}\right)\mathrm{d}N$$

$$\mathrm{d}\mathcal{E}(T,p,\mu) = -S\mathrm{d}T + V\mathrm{d}p - N\mathrm{d}\mu$$

In the <u>*Large System Limit*</u>:
Entropy is an order 1 homogeneous function of all extensive variables

$$\eta(\boldsymbol{y}) = \boldsymbol{y} \cdot \nabla_y \eta(\boldsymbol{y}),$$

$$\boldsymbol{\beta}(\boldsymbol{y}) = \nabla_y \eta(\boldsymbol{y}),$$

$$\psi(\boldsymbol{\beta}) = \sup_{\boldsymbol{y}} \{-\boldsymbol{\beta} \cdot \boldsymbol{y} + \eta(\boldsymbol{y})\}$$

$$= 0\,!$$

In summary, not just one entropy and one limit, but three entropies and two limits!

Three Entropies and Two Limits

Boltzmann entropy

Massieu-Guggenheim entropy

Gibbs entropy

$$\ln \Omega(\mathbf{y}) \longrightarrow \left\{ \psi(\boldsymbol{\beta}) \longleftrightarrow \eta(\mathbf{y}) \right\}$$

$$\longrightarrow \left\{ \frac{\psi(\boldsymbol{\beta})}{\mathbf{y}} = 0, \ \frac{\eta(\mathbf{y})}{\mathbf{y}} = \frac{\mathrm{d}\eta}{\mathrm{d}\mathbf{y}} = \boldsymbol{\beta} \right\}$$

big data limit

large system limit

Three Entropies and Two Limits

convex functions

$$\ln \Omega(\mathbf{y}) \longrightarrow \left\{ \psi(\boldsymbol{\beta}) \longleftrightarrow \eta(\mathbf{y}) \right\}$$

$$\longrightarrow \left\{ \frac{\psi(\boldsymbol{\beta})}{\mathbf{y}} = 0, \ \frac{\eta(\mathbf{y})}{\mathbf{y}} = \frac{\mathrm{d}\eta}{\mathrm{d}\mathbf{y}} = \boldsymbol{\beta} \right\}$$

subextensive function

homogeneous function

From here to things like …

$$G = H - TS, \quad \frac{\partial G}{\partial c_i} = \mu_i$$

$$\Delta\mu = \Delta\mu^o + kT\ln\frac{[C][D]}{[A][B]}$$

$$\Delta\mu^o = -kT\ln K_{\mathrm{eq}}$$

Stochastic biochemical kinetic description dictates a *macroscopic, deterministic biochemical kinetics* (as LLN) as well as a *biochemical thermodynamics* (as LDP)!

PHYSICAL REVIEW E **94**, 052150 (2016)

Mesoscopic kinetic basis of macroscopic chemical thermodynamics: A mathematical theory

Hao Ge[1,2,*] and Hong Qian[3,†]

[1]*Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing 100871, People's Republic of China*
[2]*Biodynamic Optical Imaging Center (BIOPIC), Peking University, Beijing 100871, People's Republic of China*
[3]*Department of Applied Mathematics, University of Washington, Seattle, Washington 98195-3925, USA*
(Received 11 April 2016; revised manuscript received 11 October 2016; published 30 November 2016)

Gibbs' macroscopic chemical thermodynamics is one of the most important theories in chemistry. Generalizing it to mesoscaled nonequilibrium systems is essential to biophysics. The nonequilibrium stochastic thermodynamics of chemical reaction kinetics suggested a free energy balance equation $dF^{(\mathrm{meso})}/dt = E_{\mathrm{in}} - e_p$ in which the free energy input rate E_{in} and dissipation rate e_p are both non-negative, and $E_{\mathrm{in}} \leqslant e_p$. We prove that in the macroscopic limit by merely allowing the molecular numbers to be infinite, the generalized mesoscopic free energy $F^{(\mathrm{meso})}$ converges to φ^{ss}, the large deviation rate function for the stationary distributions. This generalized macroscopic free energy φ^{ss} now satisfies a balance equation $d\varphi^{\mathrm{ss}}(\mathbf{x})/dt = \mathrm{cmf}(\mathbf{x}) - \sigma(\mathbf{x})$, in which $\mathbf{x}$ represents chemical concentration. The chemical motive force $\mathrm{cmf}(\mathbf{x})$ and entropy production rate $\sigma(\mathbf{x})$ are both non-negative, and $\mathrm{cmf}(\mathbf{x}) \leqslant \sigma(\mathbf{x})$. The balance equation is valid generally in isothermal driven systems and is different from mechanical energy conservation and the first law; it is actually an unknown form of the second law. Consequences of the emergent thermodynamic quantities and equalities are further discussed. The emergent "law" is independent of underlying kinetic details. Our theory provides an example showing how a macroscopic law emerges from a level below.

J Stat Phys (2017) 166:190–209
DOI 10.1007/s10955-016-1678-6

Mathematical Formalism of Nonequilibrium Thermodynamics for Nonlinear Chemical Reaction Systems with General Rate Law

Hao Ge[1] · Hong Qian[2]

Conclusions

(1) The thermodynamic structure presented in the present work, while assumes a probability distribution ***a priori***, does not require the concept of ***equilibrium*** in connection to detailed balance in stochastic dynamics, nor ***ergodicity***. Therefore, it is applicable to measurements on biomarkers from isogenic single living cells. Of course, if a large system consists of many statistically identical but independent smaller parts, then the entire argument based on i.i.d. measurements can be applied to a single measurement of extensive variables of the large system as a whole

Conclusions – cont.

(2) The present result augments the current understanding of the nature of thermodynamic behavior, which so far has been focused on large systems. We now see there is actually a ***large measurements limit*** that generates a different kind of emergent order, a duality symmetry, for any small stochastic systems. This symmetry is lost, however, in the large systems limit.

Qian and Cheng, ***Quant. Biol***. **8**, 172 (2020)

(3)
Application to Single-cell Biology: counting frequencies for phenotypic heterogeneity and mean value of a quantitative biomarker

Joint work with Mr. Yu-Chen Cheng , UW.

Consider total M isogenic cells,

assuming there are K phenotypic states (clusters), with m_1, m_2, …, m_K number of cells within different phenotypes,

Let g_k be the value of a single-cell biomarker when the cell is in the phenotype k.

The standard LLN for mean value and the Borel's LLN for frequencies

$$\lim_{M\to\infty} \frac{g_1 + g_2 + \cdots + g_M}{M} = \mathbb{E}[g]$$

$$\lim_{M\to\infty} \frac{m_k}{M} = \mathbb{P}_k \ (k = 1,2,\ldots,K)$$

$$\boxed{\Pr\{m_1 = x_1, \cdots, m_K = x_K\} \sim e^{-MI(\mathbf{x})}}$$

$$I(\mathbf{x}) = \sum_{k=1}^{K} x_k \log \frac{x_k}{p_k}$$

$$\bar{g}^{(M)} = \frac{m_1 g_1 + \cdots + m_K g_K}{M}$$

$$= \sum_{k=1}^{K} x_k g_k \rightarrow \mathbb{E}[g]$$

Contraction Principle

$$\boxed{\Pr\{\bar{g}^{(M)} = y\} \sim e^{-M\varphi(y)}}$$

$$\varphi(y) = \inf_{\left\{\mathbf{x}:\ \sum_{k=1}^{M} x_k g_k = y\right\}} I(\mathbf{x})$$

An Optimization Problem

$$\begin{cases} \min\limits_{\mathbf{x}} \sum\limits_{k=1}^{M} x_k \log \dfrac{x_k}{p_k}, \\ \sum\limits_{k=1}^{M} x_k g_k = y, \\ \sum\limits_{k=1}^{M} x_k = 1. \end{cases}$$

Using the method of Lagrangian multiplier

$$\begin{cases} \varphi = -\beta y - \log \sum_{k=1}^{M} p_k e^{-\beta g_k} \\ y = -\frac{d}{d\beta} \log \sum_{k=1}^{M} p_k e^{-\beta g_k} \end{cases}$$

Using the method of Lagrangian multiplier

$$\begin{cases} \varphi = -\beta y - \log \sum_{k=1}^{M} p_k e^{-\beta g_k} \\ y = -\dfrac{d}{d\beta} \log \sum_{k=1}^{M} p_k e^{-\beta g_k} \end{cases}$$

cumulant generating function

$$\boxed{\varphi(y) = -\min_{\beta}\{\boldsymbol{\beta} \cdot \boldsymbol{y} + \psi(\boldsymbol{\beta})\}}$$

Cramér's theorem for arbitrary distribution $p_X(x)$

Harald Cramér (1893-1985)

The Shannon entropy is to Borel's LLN what the Gibbs entropy is to standard LLN for mean value!

$$\lim_{M\to\infty} \frac{g_1 + g_2 + \cdots + g_M}{M} = \mathbb{E}[g]$$

the Gibbs entropy

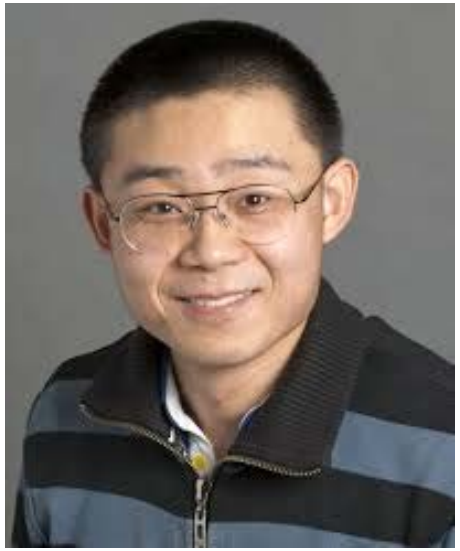
$$\lim_{M\to\infty} \frac{m_k}{M} \quad (k = 1,2,\ldots,K)$$

the Shannon entropy

Tentative Summary

- Applying the theory of large deviations to the statistical analysis of single cells, there could be a “thermodynamic behavior” in the date;
- There is a relation at the fundamental level between Waddington’s single cell phenotypic landscape and Gibbsian thermodynamic;
- Large deviations theory offers nonlinear statistical dependency beyond Gaussian fluctuations.

Acknowledgements



Prof. Zhiyue Lu
UNC, Chapel Hill

Mr. Yu-Chen Cheng
Univ. of Wash.

Thank you!