Quantitative Biology 2020, 8(2): 172–176
https://doi.org/10.1007/s40484-020-0196-3

PROTOCOL AND TUTORIAL

Counting single cells and computing their heterogeneity: from phenotypic frequencies to mean value of a quantitative biomarker

Hong Qian, Yu-Chen Cheng*

Department of Applied Mathematics, University of Washington, Seattle, WA 98195-3925, USA
* Correspondence: yuchench@uw.edu

Received December 6, 2019; Revised January 13, 2020; Accepted January 15, 2020

This tutorial presents a mathematical theory that relates the probability of sample frequencies, of *M* phenotypes in an isogenic population of *N* cells, to the probability distribution of the sample mean of a quantitative biomarker, when the *N* is very large. An analogue to the statistical mechanics of canonical ensemble is discussed.

Keywords: large deviation principle; chemical kinetics; Boltzmann's law; variational Bayesian method; maximum entropy principle

INTRODUCTION

Statistical analyses of data and stochastic models of mechanisms are two very different, but complementary approaches in biological research. While the former obtains a quantitative representation of high-throughput measurements [1], the latter can provide "laws of nature" through limit theorems [2], widely called *emergent phenomenon*. A case in point is the theory of phase transition [3] which shows that a nonlinear stochastic dynamical system with bistability and cusp catastrophe, in the limit of time $t\to\infty$ followed by system's size $V\to\infty$, necessarily exhibits a discontinuous transition [4]. Another example is the recent work [5] which demonstrates that Gibbsian equilibrium chemical thermodynamics can be reformulated as a limit theorem in a mesoscopic chemical kinetic system, with N species and M reversible stochastic elementary reactions, as the system's size becoming macroscopic.

With the rise of single-cell biology, one naturally is interested in the limiting behavior of the phenotypic frequencies among a population of cells, usually based on one, or several biomarkers. In this case, there is actually a very powerful mathematical result that is widely known to probabilists and statistical physicists. In this tutorial, we give an introduction of this theory and discuss its broader implications.

CHARACTERIZING HETEROGENEITY IN SINGLE CELLS

Asymptotic probaility distribution for sample frequencies of cellular phenotypes

To study phenotypic heterogeneity, let a population of N isogenic cells as independent and identically distributed (i.i.d.) realizations of random events from a set $\Omega=\{1,2,...,M\}$: There are totally M possible phenotypes. Among the N cells, let n_k denotes the random number of cells in the k^{th} state: $n_1+n_2+\cdots+n_M=N$. By *phenotypic frequency*, we mean $f_k^{(N)}\equiv n_k/N$.

Let p_k denote the probability of a cell in the k^{th} state. Then the probability distribution for the observed frequency $\vec{f}=(f_1,...,f_M)$ being $\mathrm{x}=(x_1,...,x_M)$ follows a multinomial distribution

$$\Pr\left\{\vec{f}^{(N)}=\mathrm{x}\right\}=\frac{N!}{(Nx_1)!(Nx_2)!\cdots(Nx_M)!}p_1^{Nx_1}p_2^{Nx_2}...p_M^{Nx_M}. \tag{1}$$

Since usually N is very large in a high-throughput single-cell experiment, one can safely approximate Eq. (1) using Stirling's formula and obtain:

$$\ln \Pr\left\{\overrightarrow{f}^{(N)}=\mathrm{x}\right\}$$

$$=\ln\left(\frac{N!}{(Nx_1)!(Nx_2)!\cdots(Nx_M)!}p_1^{Nx_1}p_2^{Nx_2}\ldots p_M^{Nx_M}\right)$$

$$\simeq -N\left[\sum_{k=1}^{M}x_k\ln\left(\frac{x_k}{p_k}\right)\right]. \tag{2}$$

Therefore, one has the asymptotic limit

$$\varphi(\mathrm{x})=-\lim_{N\to\infty}\frac{1}{N}\ln\Pr\left\{\overrightarrow{f}^{(N)}=\mathrm{x}\right\}=\sum_{k=1}^{M}x_k\ln\left(\frac{x_k}{p_k}\right). \tag{3}$$

In the theory of large deviations of probability, this is known as Sanov's theorem [6]. Since $\varphi(\mathrm{x})>0$ except when $\mathrm{x}=(p_1,\ldots,p_M)$, in the limit of $N\to\infty$, the probability of $\overrightarrow{f}_k^{(N)}\neq p_k$ is zero, and the probability of $\overrightarrow{f}_k^{(N)}=p_k$ is one. The frequency yields the probability for an infinitely large number of i.i.d. samples. Furthermore, Eq. (2) shows that $(p_1,p_2,\ldots,p_M)$ are the most probable sample frequencies for a finite but large N.

Asymptotic distribution for the mean value of a biomarker

Eqs. (1) and (2) give the probability for the frequencies within the N cells distributed among the M phenotypic states. We now consider a specific biomarker $\boldsymbol{g}$, which is assumed to be a well defined real-valued function of the phenotype of a cell: $\boldsymbol{g}=g_k$ when a cell is in the k^{th} state.

It is very clear that if one knows the frequencies $\overrightarrow{f}^{(N)}$, then the mean value for $\boldsymbol{g}$ over the entire population of the N cells is determined:

$$\overline{\boldsymbol{g}}^{(N)}=\frac{n_1g_1+n_2g_2+\cdots+n_Mg_M}{N}=\sum_{k=1}^{M}f_k^{(N)}g_k; \tag{4}$$

since the frequencies $f_k^{(N)}$ are random, so is $\overline{\boldsymbol{g}}^{(N)}$. Then when $N\to\infty$, one expects the $\overline{\boldsymbol{g}}^{(N)}$ approaching to the expected value $\mathbb{E}[\boldsymbol{g}]$. This is easy to show:

$$\lim_{N\to\infty}\overline{\boldsymbol{g}}^{(N)}=\sum_{k=1}^{M}\left(\lim_{N\to\infty}f_k^{(N)}\right)g_k=\sum_{k=1}^{m}p_kg_k=\mathbb{E}[\boldsymbol{g}]. \tag{5}$$

What is the probability distribution for $\overline{\boldsymbol{g}}^{(N)}$ when N is very large but not infinite? One can calculate this:

$$\Pr\left\{\overline{\boldsymbol{g}}^{(N)}=y\right\}=\sum_{\{\mathrm{x}:\sum_{k=1}^{M}x_kg_k=y\}}\Pr\left\{\overrightarrow{f}^{(N)}=\mathrm{x}\right\}$$

$$\simeq\exp\left(-N\inf_{\{\mathrm{x}:\sum_{k=1}^{M}x_kg_k=y\}}\varphi(\mathrm{x})\right)$$

$$\text{as } N\to\infty. \tag{6}$$

We obtain Eq. (6) because among the many sets of x that give the same value y, each has a probability of $e^{-N\varphi(\mathrm{x})}$. Therefore, as $N\to\infty$, only the set with the smallest $\varphi(\mathrm{x})$ matters. Eq. (6) indicates that for very large N, the probability distribution for the mean value of the biomarker $\overline{\boldsymbol{g}}^{(N)}$ has the form $e^{-N\psi(y)}$, in which

$$\psi(y)\equiv-\lim_{N\to\infty}\frac{1}{N}\ln\Pr\left\{\overline{\boldsymbol{g}}^{(N)}=y\right\}=\inf_{\{\mathrm{x}:\sum_{k=1}^{M}x_kg_k=y\}}\varphi(\mathrm{x}). \tag{7}$$

In the theory of large deviations of probability, this result is known as contraction principle [6]. $\psi(y)$ and $\varphi(\mathrm{x})$ are called a level-1 and a level-2 large deviations rate functions, respectively.

From phenotypic frequencies to biomarker mean values

The right-hand-side of Eq. (7) can be further carried out; this is a problem of constrained minimization using multivariate calculus:

$$\begin{cases}\min\limits_{\mathrm{x}}\left\{\left(\sum\limits_{k=1}^{M}x_k\ln\dfrac{x_k}{p_k}\right)\right\},\\ \sum\limits_{k=1}^{M}x_kg_k=y,\\ \sum\limits_{k=1}^{M}x_k=1.\end{cases} \tag{8}$$

Introducing Langrage multipliers for Eq. (8),

$$\mathscr{L}(\mathrm{x},\beta,\varphi)=\left(\sum_{k=1}^{M}x_k\ln\frac{x_k}{p_k}\right)+\beta\left(\sum_{k=1}^{M}x_kg_k-y\right)$$

$$+\lambda\left(\sum_{k=1}^{M}x_k-1\right). \tag{9}$$

Then we can find $\mathrm{x}^*=(x_1^*,x_2^*,\ldots,x_M^*)$, β^*, and λ^* as the solution of

$$\frac{\partial \mathscr{L}(\mathrm{x},\beta,\lambda)}{\partial x_j}=\frac{\partial \mathscr{L}(\mathrm{x},\beta,\lambda)}{\partial \beta}=\frac{\partial \mathscr{L}(\mathrm{x},\beta,\lambda)}{\partial \lambda}=0. \tag{10}$$

That is,

$$x_k^*=\frac{p_k e^{-\beta^* g_k}}{\sum_{j=1}^M p_j e^{-\beta^* g_j}}, \tag{11a}$$

$$y=\sum_{j=1}^M x_j^* g_j=-\left[\frac{\partial}{\partial \beta}\ln\sum_{j=1}^M p_j e^{-\beta_{g_j}}\right]_{\beta=\beta^*}, \tag{11b}$$

in which β^* is a function of y through Eq. (11b), which gives the function implicitly. We therefore obtain

$$\begin{aligned}\psi(y)=\varphi(\mathrm{x}^*)&=\sum_{k=1}^M x_k^*\ln\left(\frac{e^{-\beta^* g_k}}{\sum_{j=1}^M p_j e^{-\beta^* g_j}}\right)\\&=-\beta^*\sum_{k=1}^M x_k^* g_k-\ln\sum_{j=1}^M p_j e^{-\beta^* g_j}\\&=-\beta^* y+\beta^* F(\beta^*),\end{aligned} \tag{12}$$

where

$$F(\beta)=-\frac{1}{\beta}\ln Z(\beta),\ Z(\beta)\equiv\sum_{j=1}^M p_j e^{-\beta^* g_j}, \tag{13}$$

and $\beta^*(y)$ solves $\mathrm{d}[\beta F(\beta)]/\mathrm{d}\beta=y$.

The above computation tells us that if one knows the values of a biomarker for all the M states of a cell, $g_1,g_2,...,g_M$, together with a prior knowledge of $p_1,p_2,...,p_M$, one should construct the $Z(\beta)$ function and calculate the $F(\beta)$ given in Eq. (13). Then the probability distribution for the mean value of the biomarker is going to be:

$$\Pr\left\{\overline{\boldsymbol{g}}^{(N)}=y\right\}\propto\left[e^{-N\beta\{F(\beta)-y\}}\right]_{\beta=\beta^*(y)}. \tag{14}$$

It also tells us that if one observes the mean biomarker value being $\hat{y}$, then the most probable phenotypic frequencies will have a posterior form that deviates from its prior $\{p_k\}$:

$$f_k\{\text{conditioned on }\overline{\boldsymbol{g}}=\hat{y}\}=\left[\frac{p_k e^{-\beta g_k}}{Z(\beta)}\right]_{\beta=\beta^*(\hat{y})}. \tag{15}$$

Both Eqs. (14) and (15) suggest that the functional relationship $y=\mathrm{d}[\beta F(\beta)]/\mathrm{d}\beta$, between the mean value of the biomarker $y=\overline{\boldsymbol{g}}$ and the Lagrangian multiplier β, or its inverse form $\beta=\beta^*(y)$, are very fundamental to the probabilistic problem, in the limit of infinite sample size $N\to\infty$.

BEYOND AN i.i.d. POPULATION

We derived the expression in Eq. (3) based on the assumption of a population of N cells that are i.i.d. samples of a single M-state random individual with probability $\{p_k\}$. When there are cell-cell interactions among the individuals within a population, the mathematics immediately becomes much more involved.

Two types of research go beyond an i.i.d. population in the stochastic modeling; they were originally motivated, respectively, by chemical kinetics in solution [7] and Ising model for ferromagnetism of solid [8,9]. In chemical kinetics, rapid spatial movement of all "individual molecules" in an aqueous solution leads to the assumption that every individual collides with every other invidivual, and certain "reactions" can occur randomly. The Gibbs function in chemical thermodynamics is precisely like the $\varphi(\mathrm{x})$ function in Eq. (3), for complex chemical reaction systems in equilibrium [5]. Actually there is a general equation, first discovered by [10], whose solution can provide $\varphi(\mathrm{x})$ for non-i.i.d. populations. For $J=M-1$ reversible unimolecular reactions among M species, $X_1,X_2,...,X_M$, with concentrations $\mathrm{x}=(x_1,x_2,...,x_M)$ and arbitrary non-negative functions $R_{\pm j}(x)$ being the rates of the j^{th} reaction between species j and the species M,

$$X_1\underset{R_{+1}}{\overset{R_{-1}}{\rightleftharpoons}}X_M,\ X_2\underset{R_{+2}}{\overset{R_{-2}}{\rightleftharpoons}}X_M,\cdots,X_{M-1}\underset{R_{+J}}{\overset{R_{-J}}{\rightleftharpoons}}X_M, \tag{16}$$

the equation reads

$$\begin{aligned}&\sum_{j=1}^J R_{+j}(\mathrm{x})[1-e^{\partial\varphi/\partial x_j-\partial\varphi/\partial x_M}]+R_{-j}(\mathrm{x})\\&\times[1-e^{-\partial\varphi/\partial x_j+\partial\varphi/\partial x_M}]=0.\end{aligned} \tag{17}$$

If $R_{+j}(\mathrm{x})=q_j x_M$ and $R_{-j}(\mathrm{x})=r_j x_j$, then the solution to Eq. (17) recovers the Eq. (3),

$$\varphi(\mathrm{x})=\sum_{m=1}^M x_m\ln\left(\frac{x_m}{p_m}\right), \tag{18}$$

in which the p's are functions of q's and r's,

$$p_m=\frac{\frac{q_m}{r_m}}{\frac{q_1}{r_1}+\cdots+\frac{q_{M-1}}{r_{M-1}}+1}. \tag{19}$$

The particular set of $R_{\pm j}(x)$ represents chemical reactions in an ideal solution. A reader who had a course on freshman chemistry might recognize Eq. (18) as

$$G(\mathrm{x})=\sum_{m=1}^M x_m\mu_m,\ \mu_m(x_m)=\mu_m^o+RT\ln x_m,$$

where μ_m is the chemical potential of m^{th} specie with mole fraction x_m (not molarity) in an ideal solution, and $\mu_m^o = -RT\ln p_m$. Then $\Delta\mu_{ij}^o = \mu_i^o - \mu_j^o = -RT\ln(p_i/p_j)$, where (p_i/p_j) is the equilibrium constant between species i and j [11]. Apart from the RT, the Gibbs energy function is a consequence of statistical counting, which has very little to do with the energy of the atoms in the molecules [12].

In the second type, Ising model and alike, "individual atoms" are located at fixed lattice points in a solid, each one only interacts with its neighbours. The limit of $N\to\infty$ of such an interacting particle system is known as *hydrodynamic limit* of the stochastic model.

Cell-cell interactions in a tissue or in a culture medium can have both types: When an interaction is mediated by rapidly diffusing small molecular factors, one can safely assume the interaction is between every two individual cells in a population. If an interaction between nearby cells is mediated by slowly diffusing molecules, or due to direct contacts via mechanical interactions, gap junctions, or synapses, then a lattice model is more appropriate. Combining these two types of mathematical descriptions leads to the "reaction diffusion" paradigm [13] which serves the foundation for describing living phenomena [14].

DISCUSSION

Statistical mechanics and Boltzmann's law

A reader who had a course on statistical mechanics [15] will certainly recognize $Z(\beta)$, $F(\beta)$, and β^{-1} in Eq. (13) as partition function, Helmholtz free energy, and temperature, if one identifies g_k as the energy of the k^{th} state of a mechanical system. Eq. (12) then shows that $F(\beta) = y + \beta^{-1}\psi(y)$ where $-\psi(y)$ should be identified as "entropy" of the mechanical system with energy y; and it is related to $F(\beta)$ through a Legendre transform. Most textbooks on statistical mechanics do not tell its readers, however, the clear mathematical logic of all these formulae. But actually, Boltzmann's 1877 paper [16], by counting the molecules with different kinetic energy in an ideal gas, had proceeded exactly the steps we took and derived the celebrated Boltzmann's law, in the form in Eq. (11a).

Variational Bayesian method

The $F(\beta)$ obtained in Eq. (13) has a very important property: For any, arbitrary, *normalized* distribution $\{z_k\}$,

$$\sum_{k=1}^{M} z_k \ln\left(\frac{z_k}{p_k e^{-\beta g_k}}\right) \geqslant -\ln\left(\sum_{k=1}^{M} p_k e^{-\beta g_k}\right) = \beta F(\beta). \tag{20}$$

In the variational Bayesian method for inference [17], one often knows a target, posterior distribution $p_k e^{-\beta g_k}$ but computing its normalization factor is expensive. Eq. (20) shows that to obtain the target distribution, one can simply minimizes the left-hand-side of Eq. (20) among a set of possible $\{z_k\}$. This same idea had also been used by Gibbs in his variational method [18]: The free energy $F(\beta)$ of an equilibrium state is the minumum among all others through a virtual change of state.

Maximum entropy principle

The constrained optimization in Eq. (8) leading to distribu- tion in Eq. (11a) has also become the foundation of *maximum entropy principle* (MEP) championed by Jaynes [19], which has played a productive role in data science. The axiomatic nature of MEP [20] and the role of conditional probability [21] have been elucidated.

The fundamental premises behind the large deviations principle (LDP) and the MEP are very different: Entropy, as a large deviation rate function, is used in the former to find the rare event that is the most probable, which is the only possible event in the limit: For an arbitrary set of n real values $\{\varphi_i\}$,

$$(e^{-N\varphi_1} + e^{-N\varphi_2} + \cdots + e^{-N\varphi_n}) \sim e^{-N\varphi^*} \text{ as } N\to\infty, \tag{21}$$

where $\varphi^* = \min\{\varphi_1, \ldots, \varphi_n\}$. This is the same idea in choosing only the term with the largest eigenvalue among the terms in a linear eigenvalue decomposition, in the limit of infinite time or system's size. In MEP, however, entropy function is used as a measure for "unbias". Actually, according to LDP, the x^* in (11a) is not a probability distribution, it is the most probable frequency among N i.i.d. samples. In MEP, it is interpreted as the least biased probability distribution with maximum uncertainty.

ACKNOWLEDGEMENTS

We thank Ivana Bozic, Ken Dill, Hao Ge, Liu Hong, Matt Lorig and Wenning Wang for many helpful discussions. H. Q. is partially supported by NIH grant R01GM109964 (PI: Sui Huang) and the Olga Jung Wan Endowed Professorship.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Hong Qian and Yu-Chen Cheng declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Pence, C. H. (2011) "Describing our whole experience": the statistical philosophies of W. F. R. Weldon and Karl Pearson. Stud.

Hist. Philos. Biol. Biomed. Sci., 42, 475–485
2. Chibbaro, S., Rondoni, L. and Vulpiani, A. (2014) Reductionism, Emergence and Levels of Reality. New York: Springer
3. Anderson, P. W. (1972) More is different. Science, 177, 393–396
4. Qian, H., Ao, P., Tu, Y. and Wang, J. (2016) A framework towards understanding mesoscopic phenomena: Emergent unpredictability, symmetry breaking and dynamics across scales. Chem. Phys. Lett., 665, 153–161
5. Ge, H. and Qian, H. (2016) Mesoscopic kinetic basis of macroscopic chemical thermodynamics: A mathematical theory. Phys. Rev. E, 94, 052150
6. Dembo, A. and Zeitouni, O. (1998) Large Deviations Techniques and Applications, 2nd ed. New York: Springer
7. Kurtz, T. G. (1972) The relationship between stochastic and deterministic models for chemical reactions. J. Chem. Phys., 57, 2976–2978
8. Liggett, T. M. (1985) Interacting Particle Systems. New York: Springer-Verlag
9. Derrida, B. (1998) An exactly soluble nonequilibrium system: The asymmetric simple exclusion process. Phys. Rep., 301, 65–83
10. Gang, H. (1986) Lyapunov function and stationary probability distribution. Zeit. Physik B: Cond. Matt. 65, 103–106
11. Chang, R. and Goldsby, K. A. (2012) Chemistry, 11th ed. New York: McGraw-Hill
12. Qian, H. (2019) Stochastic population kinetics and its underlying mathematicothermodynamics. In: The Dynamics of Biological Systems, Bianchi, A., Hillen, T., Lewis, M., Yi, Y. eds., pp. 149–188. Springer: New York
13. Murray, J. D. (2011) Mathematical Biology II: Spatial Models and Biomedical Applications, 3rd ed. New York: Springer
14. von Bertalanffy, L. (1950) The theory of open systems in physics and biology. Science, 111, 23–29
15. Huang, K. (1963) Statistical Mechanics. New York: John Wiley & Sons
16. Sharp, K. and Matschinsky, F. (2015) Translation of Ludwig Boltzmann's paper "On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium". Entropy (Basel), 17, 1971–2009
17. Ghahramani, Z. (2001) An introduction to hidden Markov models and Bayesian networks. Int. J. Pattern Recognit. Artif. Intell., 15, 9–42
18. Pauli, W. (1973) Pauli Lectures on Physics: Thermodynamics and the Kinetic Theory of Gas. Cambridge: The MIT Press
19. Jaynes, E. T. (2003) Probability Theory: The Logic of Science. London: Cambridge University Press
20. Shore, J. E. and Johnson, R. W. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inf. Theory, 26, 26–37
21. van Campenhout, J. M. and Cover, T. M. (1981) Maximum entropy and conditional probability. IEEE Trans. Inf. Theory, 27, 483–489

