# Universal Protein Fluctuations in Populations of Microorganisms

Hanna Salman,[1,2] Naama Brenner,[3,4,*] Chih-kuan Tung,[1] Noa Elyahu,[3,4] Elad Stolovicki,[5,4] Lindsay Moore,[5,4] Albert Libchaber,[6] and Erez Braun[5,4]

[1]*Department of Physics and Astronomy, University of Pittsburgh*
[2]*Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[3]*Department of Chemical Engineering, Technion, Haifa 32000, Israel*
[4]*Laboratory of Network Biology, Technion, Haifa 32000, Israel*
[5]*Department of Physics, Technion, Haifa 32000, Israel*
[6]*Center for Physics and Biology, Rockefeller University, New York, New York 10065, USA*

(Received 6 January 2012; published 6 June 2012)

The copy number of any protein fluctuates among cells in a population; characterizing and understanding these fluctuations is a fundamental problem in biophysics. We show here that protein distributions measured under a broad range of biological realizations collapse to a single non-Gaussian curve under scaling by the first two moments. Moreover, in all experiments the variance is found to depend quadratically on the mean, showing that a single degree of freedom determines the entire distribution. Our results imply that protein fluctuations do not reflect any specific molecular or cellular mechanism, and suggest that some buffering process masks these details and induces universality.

DOI: 10.1103/PhysRevLett.108.238105 PACS numbers: 87.18.Tt, 87.10.−e, 87.18.−h, 87.18.Nq

The protein content of a cell is a primary determinant of its phenotype. However, protein copy number is subject to large cell-to-cell variation even among genetically identical cells grown under uniform conditions. This variation has been the subject of intensive research in recent years ([1–7] and references therein). Much of this previous work was devoted to characterizing the stochastic properties of various processes underlying gene expression, such as transcription and translation [8], or different stages of the cell cycle [9], and understanding their effect on protein variation. However, gene expression is generally coupled to all aspects of cell physiology, such as growth [10], metabolism [11], aging [12], division [13,14] and epigenetic processes [15,16], as well as gene location and function [17], all of which were shown to affect protein variation. The emerging picture is of a plethora of correlated mechanisms at different levels of organization; how they integrate to shape the total protein variation in a dividing population remains an open question [11,14].

In this work we addressed this question by a phenomenological approach. We measured distributions of highly expressed proteins in proliferating clonal populations of bacteria and yeast under natural conditions, where gene expression is coupled to other cellular processes. By designing an array of different metabolic and regulatory conditions as well as growth environments, we collected a compendium of measurements which systematically covers the major processes of gene expression and cell division, and compared the measured distributions in a wide range of biological realizations. More specifically, our comparisons included the following. (a) Two archetypical microorganisms, bacteria and yeast, with two well-studied regulatory systems of essential metabolic pathways: the LAC operon in *E. coli* [18] and the GAL system in *S. cerevisiae* [19]. Both systems were studied under environmental conditions in which expression is strongly coupled to metabolism; namely, they control the utilization of an essential sugar (lactose and galactose, respectively) as the sole carbon source. (b) Different metabolic growth conditions: the organisms were grown in chemostats—continuous culture in steady state and transients, as well as in batch cultures. (c) Highly regulated versus constitutive (approximately fixed rate) expression. The regulated LAC and GAL systems were compared to constitutively expressed proteins in both organisms. (d) Different promoter copy numbers: the same regulatory systems were placed on high-copy (HC) and low-copy (LC) number plasmids as well as integrated into the genome in a single copy. (e) Reporter GFP was compared to an essential functional tagged-protein controlled by the same promoter (for experimental details see the supplemental material [20]).

The spectrum of our experiments spans an array of ''control parameters'' $\vec{p}$ which covers many of the essential processes affecting protein content in cells. The two organisms used, *E. coli* and *S. cerevisiae*, are distinct in almost every aspect of their cell biology and life style, from gene regulation and expression to cell division and physical characteristics such as shape and volume. A comparative experiment in which some control parameter was varied will reveal the sensitivity of the distribution to that particular parameter. If there is no sensitivity and the distributions are the same, then they do not convey information about that parameter and the two experiments exhibit universal behavior. Given the differences between

 

the organisms, the various regulatory systems and the different experimental conditions, it was not at all obvious *a priori* that any universality could be found.

Figure 1 shows a collection of distributions measured in such comparative experiments. Despite the clear differences in scale between the distributions they all show common features: all are skewed, unimodal and exhibit extended exponential-like tails. These general features were previously reported in multiple publications, and different mechanisms were proposed to account for them [11,13,21–23]. Some of the distributions displayed in Fig. 1 are very similar to one another: for example, Fig. 1(a) shows two indistinguishable distributions of GFP under the control of a *LacO* promoter on a high-copy number plasmid in bacteria, and under the control of the GAL10 promoter integrated into the genome in yeast. Similarly, Fig. 1(d) depicts identical distributions of a reporter GFP expressed under the GAL10 promoter and an essential metabolic protein tagged with GFP at its *C* terminal, both integrated into the genome in yeast.

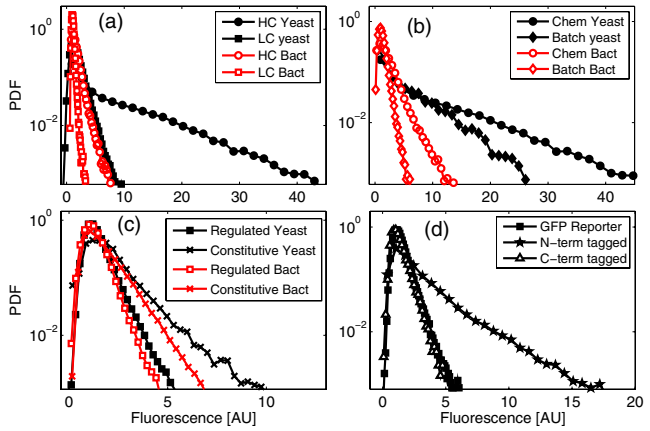


FIG. 1 (color online). Protein distributions. GFP fluorescence distributions in populations of bacteria (red) and yeast (black) measured under different conditions by flow cytometry. (a) High- and low-copy number (HC and LC respectively) regulatory promoters in bacteria and yeast: GFP expressed under the *LacO* promoter on high (circles) and low (squares) copy number plasmids in bacteria, and under the GAL10 promoter on a high-copy plasmid (circles) and integrated into the genome (squares) in yeast. (b) Continuous (chemostats; circles) and batch cultures (diamonds) in bacteria and yeast: GFP expressed from high-copy number plasmids under the *LacO* or GAL10 promoters, respectively. (c) Regulated (squares) and constitutive (x) promoters: In bacteria, *LacO* promoter is compared to ColE1P1 promoter, both on low-copy number plasmids, and in yeast the GAL10 promoter is compared to ADH1, both integrated into the genome. All populations were grown in batch cultures. (d) Reporter GFP under GAL10 promoter (squares) and a functional HIS3-GFP N-terminal (stars) and C-terminal (triangles) tagged, both under the GAL1 promoter. All fluorescence levels in this figure were normalized such that the peak of the distributions appears at 1. The probability density is normalized to unit area. Note the logarithmic *y* axis.

Assuming that these similarities are not coincidental, the possibility arises that there is a universal principle underlying protein distributions in proliferating populations of microorganisms.

To test this possibility, we compared the distributions after normalizing out the obvious differences in absolute scales, which are mostly manifested in their mean and standard deviation. The mean $\mu = \langle x \rangle$ reflects the absolute number of proteins in the cell, the strength of the specific GFP used and its behavior in the different biological contexts. The standard deviation $\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$ is strongly affected by the dynamic range of protein content that also depends on the particulars of the biological system.

Figure 2 shows all the distributions of Fig. 1 on a common $x$ axis. For each distribution the axis was normalized by subtracting its mean and dividing by its standard deviation. Remarkably, all distributions collapsed to a single curve over almost 10 standard deviations in scaled fluorescence ($x$ axis) and more than 3 decades in probability density. This presentation reveals the universality of the protein distribution *shape* within the entire array of our experimental conditions: the distribution $f$ obeys the scaling form

$$f(x\,;\vec{p}) = \varphi\left(\frac{x-\mu(\vec{p})}{\sigma(\vec{p})}\right), \tag{1}$$

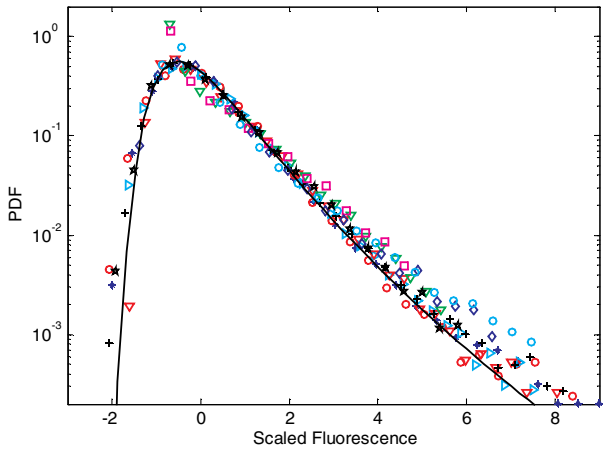


FIG. 2 (color online). (a) All distributions of Fig. 1 are plotted in units scaled by subtracting the mean and dividing by the standard deviation. The symbols represent different experiments as follows: *LacO* promoter on a high-copy plasmid in bacteria (blue star—grown in chemostat, cyan triangle—in batch); the same on a low-copy plasmid in chemostat (red circle); GAL10 promoter on a high-copy plasmid in yeast (green triangle – grown in chemostat, magenta squares—in batch); the same integrated into the genome in chemostat (black cross); ADH1 promoter integrated into the genome in yeast grown in batch (blue diamond); ColE1P1 promoter on a low-copy plasmid in bacteria grown in batch (red triangle); Fused HIS3-GFP under GAL1 promoter in yeast in chemostat (*N* terminal—cyan circles, C-terminal—black pentagram). The black line is the Frechet distribution best fit to the data, Eq. (4) with $k = 0.095$, $m = -7.5$, and $s = 7.09$.

showing that the dependence on the control parameters $\vec{p}$ enters through the mean and standard deviation. Other forms of scaling do not result in such a collapse (supplemental Fig. 1 [20]). Among several well-known skewed distributions we found that the rescaled data can be well fitted by the Frechet distribution, shown by the black curve in Fig. 2, or by a log-normal distribution. Further information on fitting the data is given in supplemental Figs. 3 and 4 [20]. It is emphasized that other fitting functions can possibly describe the data equally well; at this stage these are empirical fittings only.

Normalizing out the first two moments resulted in a universally-shaped distribution with zero mean and unit standard deviation, and discarded information on possible relations between the moments in the original, physical units. Plotting these moments one versus the other, one point for each distribution for both bacteria and yeast [Figs. 3(a) and 3(b)], reveals that the variance defines a curve in the plane with very little scatter, that can be well fitted by a quadratic function $y = Ax^2 + Bx + C$. Figure 3(c) shows a similar relation for many measurements on yeast populations done by fluorescence microscopy [11].

Further support of this relation between moments is found from transient experiments, in which we use the chemostat to switch medium between inducing and repressing conditions of gene expression while still maintaining an exponentially growing culture. Figure 4 depicts the distributions of GFP under the control of LacO promoter in bacteria switched from repressing glucose to inducing lactose medium [Fig. 4(a)] and the GAL10 promoter in a yeast population switched from inducing galactose to repressing glucose medium [Fig. 4(b)]. It is seen that the qualitative features of the distributions are maintained throughout the transient but with a time-varying exponential tail. The insets show that in these experiments, as in the steady states, the variance and mean define a quadratic relation with very little scatter.

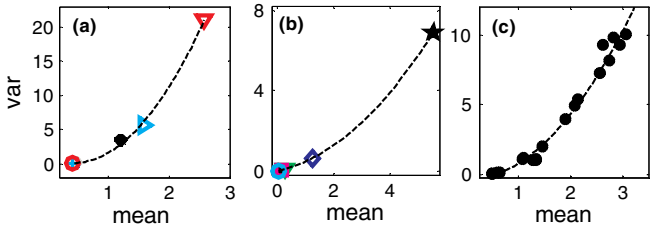


FIG. 3 (color online). Relation between mean and variance. (a) Mean and variance for all experiments in bacteria that appear in Fig. 2 using the same symbols. (b) The same for yeast experiments. (c) Mean and variance for a large collection of experiments on yeast populations, where steady state protein distributions were measured by fluorescence microscopy [11]. GFP was expressed under the control of GAL10 on high or low-copy number plasmids, in different background strains grown in chemostat cultures with different dilution rates and limiting nutrients. Fits: $y = Ax^2 + Bx + C$ with the parameters: (a) $A = 0.46$, $B = -0.39$, $C = 0.1$. (b) $A = 0.17$, $B = 3.1$, $C = -1.54$. (c) $A = 0.167$, $B = -0.31$, $C = -0.015$. The data are presented in three panels since fluorescence is not calibrated and therefore different sets of experiments performed under different conditions contain an arbitrary scale factor.

Finally, we note the quadratic dependence between variance and mean is exhibited also by published genome-wide measurements [17,21,24]. In previous work variation was characterized by the ratio between variance and mean squared ("noise"); this measure is a nonlinear combination of moments and does not provide direct information about the relation between them in the presence of measurement errors. When plotted directly, the data are seen to approximate a quadratic function over a broad dynamic range of measured variables (see supplemental Fig. 5 [20]).

The generality of the universal behavior that we have found remains to be characterized in further experiments and organisms. Clearly it does not necessarily apply to every biological realization; for example, experiments have shown that under some conditions the number of lac permeases in bacteria exhibits a bimodal distribution ([25]; the same group later concluded from a genome-wide study that such distributions are rare [21]). However, an observation of fundamental importance here is the existence of a universality class in biology. The fact that populations of two distinct microorganisms in a broad range of biological contexts exhibit protein distributions that can be scaled by mean and standard deviation to a universal curve is highly significant. The entailed conclusion is that the shape of these distributions cannot convey information on specific biological molecular or cellular

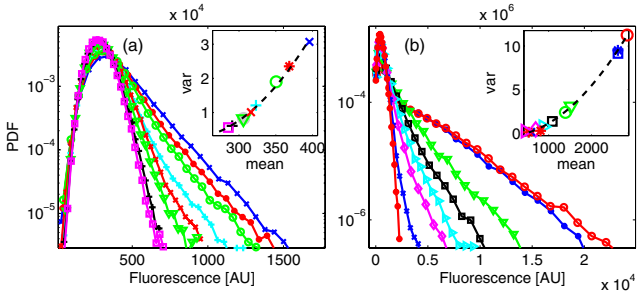


FIG. 4 (color online). Distributions under transient dynamics. Populations of bacteria (a) and yeast (b) were grown in chemostat to steady state and then switched to different medium. Main figures show protein distribution along time; insets show mean and variance throughout the transient. (a) Bacteria were grown to steady state in glucose and then switched to high concentration of lactose. As a result, an overshoot of induction was seen (blue x; broadest distribution) followed by an adaptive response until a steady state in lactose was reached (green triangles). Then the culture was switched again to glucose (last two distributions, black crosses and magenta squares). (b) Yeast cells were grown in chemostat to steady state in galactose (broadest distribution; red circles), then switched to glucose. The tail of the distribution gradually decreased following the switch until a very narrow distribution (red stars) was reached. In both experiments the mean and variance displayed a quadratic relation (insets).

mechanisms related to any of the control parameters covered by our experimental conditions. Together with the observed relation between the variance and mean $\sigma = \sigma(\mu)$, (and regardless of its precise functional form), these results imply that by measuring a single variable, e.g., the mean, the entire protein distribution can be reconstructed:

$$f(x\,;\vec{p}) = \varphi\left(\frac{x-\mu(\vec{p})}{\sigma(\mu(\vec{p}))}\right) = f(x\,;\mu). \tag{2}$$

If protein distributions do not reflect any single dominant molecular or cellular mechanism, they must be the integrated outcome of a large number of stochastic events. The masking of individual stochastic events by an integration of many of them is well known in the case of the central limit theorem. Our data, however, exhibit a universal non-Gaussian skewed exponentially tailed distribution, implying that if a similar masking exists here then some of the conditions of the central limit theorem are not fulfilled. What can one say from the data about the possible candidates of these unfulfilled conditions?

The resemblance of the universal curve to a log-normal distribution immediately raises the possibility of a multiplicative process: if cellular protein content was the product of a large number of independent random variables then its distribution would approximate a log-normal

$$L(x; m, s) = \frac{1}{x\sqrt{2\pi s^2}}\, e^{-(\ln x - m)^2/2s^2} \tag{3}$$

where $m = \langle \ln x \rangle$ and $s^2 = \mathrm{var}(\ln x)$. This distribution, with a shift parameter added, exhibits scaling by the first two moments and a quadratic relation between variance and mean only if the parameter $s$ is kept fixed. However, this is inconsistent with the interpretation of the distribution shape arising universally from a product of random variables, since any change in the variance or number of these variables alters the parameter $s$. We conclude that, despite the apparent similarity of any single measured distribution to a log-normal, the scaling properties of the data set as a whole (Figs. 2 and 3) cannot be explained by a product of many random variables.

A second possibility is that the fluctuations reflect a sum of many random variables which are not independent but rather strongly correlated. This is plausible from a biological point of view, since the different processes that contribute to the protein content of a cell are indeed strongly correlated and reflect different aspects of the same cell's individuality. Moreover considering the protein content as being accumulated over time by many random events, these events are temporally correlated; this can be deduced from single-cell measurements of phenotypic traits along time that typically show a correlation over a few generations [21,26]. These arguments support a picture where protein fluctuations arise as an integration of random variables correlated in time as well as constrained by correlations to other variables.

Non-Gaussian universal distributions with qualitatively similar features to those measured here were observed in complex physical systems where fluctuations in global variables were measured. Examples include turbulent flows, magnetization in spin systems and other equilibrium systems near phase transitions as well as nonequilibrium systems [27,28]. In many of these systems, the universal distributions could be well described by one of three universality classes of extreme value statistics. Recent theoretical work has illustrated a mapping between the extreme values of a set of independent identically distributed random variables and the sum of nonidentically distributed ones [30], showing that they have the same non-Gaussian distribution. This raises also the possibility that the measured fluctuations arise from a sum of random variables that are not identically distributed, and reflect an underlying nonstationary process [31].

Although there is no established theory and the topic is under debate [29], much research has recently been devoted to the understanding of this phenomenon using scaling arguments [27] and models of special cases [30]. Inspired by this line of thought we fitted our data to the GEV distributions and found that it could be best described by the Frechet distribution

$$F(x\,; k, m, s) = \frac{1}{ks}\left(\frac{x-m}{s}\right)^{-(1/k)-1} e^{-([x-m]/s)^{-(1/k)}}. \tag{4}$$

While some individual experiments could be better described by a log-normal or Gamma distribution, the pooled dataset was significantly better fit by the Frechet distribution than any other function we have tried (supplementary Figs. 3 and 4). More importantly, it can be shown that a family of Frechet distributions with fixed shape parameter $k$ exhibits both properties of the data—scaling by the first two moments and quadratic dependence of variance on mean (supplemental analysis [20]). Thus, while it may be possible to choose parameters where the log-normal and Frechet distributions are practically indistinguishable over the finite range of measurements, their scaling and symmetry properties are different and only the latter are consistent with the data. In spite of these consistencies, we still regard the fit to a Frechet distribution as a phenomenological description of the data. In the absence of a theory, it is not possible to exclude at this point that other distributions may describe the data equally well. Recent work has illustrated that much ambiguity can occur when inferring the details of a stochastic process from the phenomenology of its statistical properties [31].

The analogy between a cell population and the above mentioned physical systems is still suggestive at this time. However our results call for understanding of the observed universality and for connecting it with other physical systems exhibiting a similar behavior. The connection is not

straightforward; a population of cells is not a statistical ensemble of separate realizations as they exhibit long-term correlations [11,26] and slow collective modes in gene expression [32]. Searching for such a connection marks a challenging direction for future research at the interface between biology and the physics of complex systems.

This work was supported in part by a US-Israel Binational Science foundation grant and by the Israel Science Foundation (Grant No. 496/10, EB). We are grateful to S. Ruffo for suggesting an extreme value distribution fit to our data.

*Corresponding author.

[1] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, Nat. Rev. Genet. **6**, 451 (2005).
[2] N. Maheshri and E. K. O'Shea, Annu. Rev. Biophys. Biomol. Struct. **36**, 413 (2007).
[3] A. Raj and A. v. Oudenaarden, Cell **135**, 216 (2008).
[4] I. R. Booth, Food Microbiology **78**, 19 (2002).
[5] A. Eldar and M. B. Elowitz, Nature (London) **467**, 167 (2010).
[6] C. J. Davidson and M. G. Surette, Annu. Rev. Genet. **42**, 253 (2008).
[7] S. V. Avery, Nat. Rev. Microbiol. **4**. 577 (2006).
[8] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, Nat. Genet. **31**, 69 (2002).
[9] S. D. Talia, J. M. Skotheim, James M. Bean, Eric D. Siggia, and F. R. Cross, Nature (London) **448**, 947 (2007).
[10] S. Tsuru, J. Ichinose, A. Kashiwagi, B.-W. Ying, K. Kaneko, and T. Yomo, Phys. Biol. **6**, 036015 (2009).
[11] N. Brenner, K. Farkash, and E. Braun, Phys. Biol. **3**, 172 (2006).
[12] R. Bahar, C. H. Hartmann, K. A. Rodriguez, A. D. Denny, R. A. Busuttil, M. E. T. Dolle, R. Brent Calder, G. B. Chisholm, B. H. Pollock, C. A. Klein and Jan Vijg, Nature (London) **441**, 1011 (2006).
[13] T. Friedlander and N. Brenner, Phys. Rev. Lett. **101**, 018104 (2008).
[14] D. Huh and J. Paulsson, Nat. Genet. **43**, 95 (2011).
[15] A. Becskei, B. B. Kaufmann, A. van Oudenaarden, Nat. Genet. **37**, 937 (2005).
[16] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).
[17] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, Nature (London) **441**, 840 (2006).
[18] B. Muller-Hill, *The Lac Operon: a Short History of a Genetic Paradigm* (Walter de Gruyter, Berlin, New York, 1996).
[19] M. Johnston and M. Carlson, in *The Molecular and Cellular Biology of the Yeast Saccaromyces: Gene Expression*, edited by E. W. Jones, J. R. Pringle, and J. R. Broach (Cold Spring Harbor Laboratory Press, 1992), p. 193.
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.108.238105 for experimental methods, analysis of scaling, and supplementary figures.
[21] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Sunney Xie, Science **329**, 533 (2010).
[22] D. Volfson, J. Marciniak, W. J. Blake, N. Ostroff, L. S. Tsimring, and J. Hasty, Nature (London) **439**, 861 (2006).
[23] N. Friedman, L. Cai, and X. S. Xie, Phys. Rev. Lett. **97**, 168302 (2006).
[24] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai, Nat. Genet. **38**, 636 (2006).
[25] P. J. Choi, L. Cai, K. Frieda, and X. Sunney Xie, Science **322**, 442 (2008).
[26] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz, Science **307**, 1962 (2005).
[27] S. T. Bramwell, Nature Phys. **5**, 444 (2009).
[28] M. Clusel and E. Bertin, Int. J. Mod. Phys. B **22**, 3311 (2008).
[29] H. J. Hilhorst, Braz. J. Phys. **39**, 371 (2009).
[30] E. Bertin and M. Clusel, J. Phys. A **39**, 7607 (2006).
[31] D. O'Malley and J. H. Cushman, Phys. Rev. E **82**, 032102 (2010).
[32] E. Stolovicki and E. Braun, PLoS ONE **6**, e20530 (2011).