Zipf's Law in Gene Expression

Chikara Furusawa
Center for Developmental Biology, The Institute of Physical and Chemical Research (RIKEN), Kobe 650-0047, Japan

Kunihiko Kaneko
Department of Pure and Applied Sciences, University of Tokyo, Komaba, Meguro-ku, Tokyo 153-8902, Japan
(Received 27 September 2002; published 26 February 2003)

Using data from gene expression databases on various organisms and tissues, including yeast, nematodes, human normal and cancer tissues, and embryonic stem cells, we found that the abundances of expressed genes exhibit a power-law distribution with an exponent close to -1; i.e., they obey Zipf's law. Furthermore, by simulations of a simple model with an intracellular reaction network, we found that Zipf's law of chemical abundance is a universal feature of cells where such a network optimizes the efficiency and faithfulness of self-reproduction. These findings provide novel insights into the nature of the organization of reaction dynamics in living cells.

DOI: 10.1103/PhysRevLett.90.088102 PACS numbers: 87.17.Aa, 87.80.Vt, 89.75.Fb

In a cell, an enormous number of organized chemical reactions are required to maintain its living state. Although enumeration of detailed cellular processes and the construction of complicated models is important for a complete description of cellular behavior, it is also necessary to search for universal laws with regard to the intracellular reactions common to all living systems and then to unravel the logic of life leading to such universal features. For example, scale-free networks have recently been discussed as a universal property of some biochemical reaction networks [1]. These studies, however, focused only on the properties of the network topologies, while the reaction dynamics were not discussed. Here, we report a universal property of the reaction dynamics within cells, namely, a power-law distribution of the abundance of expressed genes with an exponent close to -1, i.e., Zipf's law [2]. By using an abstract model of a cell with simple reaction dynamics, we show that this power-law behavior in the chemical abundances generally appears when the reaction dynamics leads to a faithful and efficient self-reproduction of a cell. These findings provide insights into the nature of the organization of complex reaction dynamics in living cells.

In order to investigate possible universal properties of the reaction dynamics, we examined the distributions of the abundances of expressed genes (that are approximately equal to the abundances of the corresponding proteins) in six organisms and more than 40 tissues based on data publicly available from SAGE (serial analysis of gene expression) databases [3–5]. SAGE allows the number of copies of any given mRNA to be quantitatively evaluated by determining the abundances of the short sequence tags which uniquely identify it [6].

In Fig. 1, we show the rank-ordered frequency distributions of the expressed genes, where the ordinate indicates the frequency of the observed sequence tags (i.e., the population ratio of the corresponding mRNA to the total mRNA), and the abscissa shows the rank determined from this frequency. As shown, the distributions follow a power law with an exponent close to -1 (Zipf's law). We observed this power-law distribution for all the available samples, including 18 human normal tissues, human cancer tissues, mouse (including embryonic stem cells), rat, nematode (*C. elegans*), and yeast (*S. cerevisiae*) cells. All the data over 40 samples (except for two plant data) show the power-law distributions with the exponent in the range from -1– -0.86. Even though there are some factors which may bias the results of the SAGE experiments, such as sequencing errors and nonuniqueness of tag sequences, it seems rather unlikely that the distribution is an artifact of the experimental procedure.

The abundance of each protein is the result of a complex network of chemical reactions that is influenced by possibly a large number of factors including other proteins and genes. Then, why is Zipf's law universally observed, and what class of reaction dynamics will show the observed power-law distribution?

In order to investigate the above questions, we adopt a simple model of cellular dynamics that captures only its basic features. It consists of intracellular catalytic reaction networks that transform nutrient chemicals into proteins. By studying a class of simple models with these features, we clarify the conditions under which the reaction dynamics leads to a power-law distribution of the chemical abundances.

Consider a cell consisting of a variety of chemicals. The internal state of the cell can be represented by a set of numbers $(n_1, n_2, \ldots, n_k)$, where n_i is the number of molecules of the chemical species i with i ranging from $i = 1$ to k. For the internal chemical reaction dynamics, we chose a catalytic network among these k chemical species, where each reaction from some chemical i to some other chemical j is assumed to be catalyzed by a third chemical ℓ; i.e., $(i + \ell \rightarrow j + \ell)$ [7]. The rate of increase of n_j (and decrease of n_i) through this reaction is given by $\epsilon n_i n_\ell / N^2$, where ϵ is the coefficient for the chemical

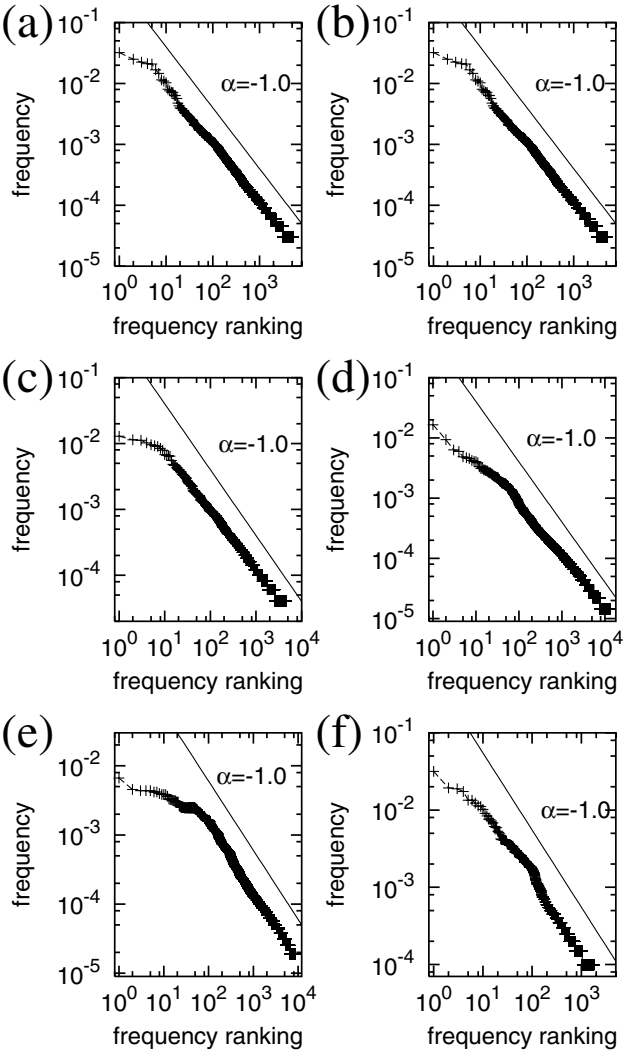

FIG. 1. Rank-ordered frequency distributions of expressed genes. (a) Human liver, (b) kidney, (c) human colorectal cancer, (d) mouse embryonic stem cells, (e) *C. elegans*, and (f) yeast (*S. cerevisiae*). The exponent of the power law is in the range from -1–-0.86 for all the samples inspected, except for two plant data (seedlings of *Arabidopsis* and the trunk of *Pinus taeda*), whose exponents are approximately -0.63.

reaction. For simplicity all the reaction coefficients were chosen to be equal [8], and the connection paths of this catalytic network were chosen randomly such that the probability of any two chemicals i and j to be connected is given by the connection rate ρ [9].

Some resources (nutrients) are supplied from the environment by diffusion through the membrane (with a diffusion coefficient D) [10], to ensure the growth of a cell. The nutrient chemicals have no catalytic activity in order to prevent the occurrence of catalytic reactions in the environment. Through the calaytic reactions, these nutrients are transformed into other chemicals. Some of these chemicals may penetrate [8] the membrane and diffuse out while others will not. With the synthesis of the unpenetrable chemicals that do not diffuse out, the total number of chemicals $N = \sum_i n_i$ in a cell can increase, and accordingly the cell volume will increase. We study how this cell growth is sustained by dividing a cell into two when the volume is larger than some threshold. For simplicity the division is assumed to occur when the total number of molecules $N = \sum_i n_i$ in a cell exceeds a given threshold $N_{\max}$. Chosen randomly, the mother cell's molecules are evenly split among the two daughter cells.

In our simulations, we randomly pick up a pair of molecules in a cell and transform them according to the reaction network. In the same way, diffusion through the membrane is also computed by randomly choosing molecules inside and outside the cell. In the case with $N \gg k$ (i.e., continuous limit), the reaction dynamics is represented by the following rate equation:

$$dn_i/dt = \sum_{j,\ell} \mathrm{Con}(j, i, \ell)\epsilon n_j n_\ell / N^2 - \sum_{j',\ell'} \mathrm{Con}(i, j', \ell')\epsilon n_i n_{\ell'} / N^2 + D\sigma_i(\overline{n}_i/V - n_i/N),$$

where $\mathrm{Con}(i, j, \ell)$ is 1 if there is a reaction $i + \ell \rightarrow j + \ell$, and 0 otherwise, whereas σ_i takes 1 if the chemical i is penetrable, and 0 otherwise. The third term describes the transport of chemicals through the membrane, where $\overline{n}_i$ is a constant, representing the number of the ith chemical species in the environment and V denotes the volume of the environment in units of the initial cell size. The number $\overline{n}_i$ is nonzero only for the nutrient chemicals.

If the total number of molecules $N_{\max}$ is larger than the number of chemical species k, the population ratios $\{n_i/N\}$ are generally fixed, since the daughter cells inherit the chemical compositions of their mother cells. For $k > N_{\max}$ [11], the population ratios do not settle down and can change from generation to generation. In both cases, depending on the diffusion coefficient D, the reaction dynamics can be classified into the three classes [12].

First, there is a critical value $D = D_c$ beyond which the cell cannot grow continuously. When $D > D_c$, the flow of nutrients from the environment is so fast that the internal reactions transforming them into chemicals sustaining "metabolism" cannot keep up. In this case all the molecules in the cell will finally be substituted by the nutrient chemicals and the cell stops growing since the nutrients alone cannot catalyze any reactions to generate unpenetrable chemicals. Continuous cellular growth and successive divisions are possible only for $D \leq D_c$. When the diffusion coefficient D is sufficiently small, the internal reactions progress faster than the flow of nutrients from the environment, and all the existing chemical species have small numbers of approximately the same level. A stable reaction network organization is obtained only at the intermediate diffusion coefficient below D_c, where some chemical species have a much larger number of molecules than others.

The rank-ordered number distributions of chemical species in our model are plotted in Fig. 2, where the ordinate indicates the number of molecules n_i and the

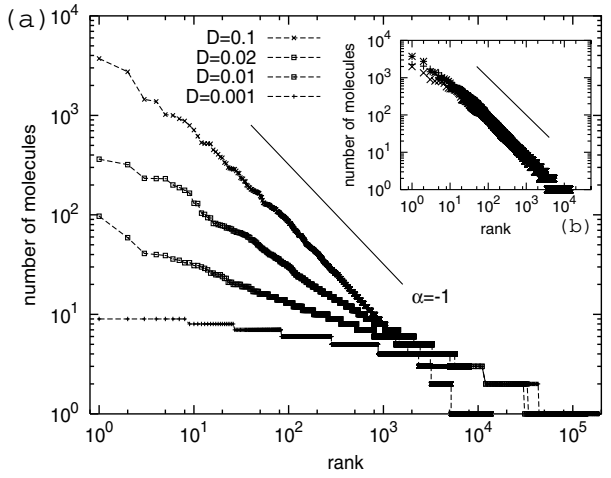

FIG. 2. Rank-ordered number distributions of chemical species. (a) Distributions with different diffusion coefficients D are overlaid. The parameters were set as $k = 5 \times 10^6$, $N_{\max} = 5 \times 10^5$, and $\rho = 0.022$. 30% of the chemical species are penetrating the membrane, and the others are not. Within the penetrable chemicals, ten chemical species are continuously supplied to the environment, as nutrients. In this figure, the numbers of nutrient chemicals in a cell are not plotted. With these parameters, D_c is approximately 0.1. (b) Distributions at the critical points with a different total number of chemicals k are overlaid. The numbers of chemicals were set as $k = 5 \times 10^4$, $k = 5 \times 10^5$, and $k = 5 \times 10^6$, respectively. Other parameters were set the same as those in (a).

abscissa shows the rank determined by n_i. As shown in the figure, the slope in the rank-ordered number distribution increases with an increase of the diffusion coefficient D. We found that at the critical point $D = D_c$, the distribution converges to a power law with an exponent -1.

The power-law distribution is maintained by a hierarchical organization of catalytic reactions, where the synthesis of higher ranking chemicals is catalyzed by lower ranking chemicals. For example, major chemical species (with, e.g., $n_i > 1000$) are directly synthesized from nutrients and catalyzed by chemicals that are slightly less abundant (e.g., $n_i \sim 200$). The latter chemicals are mostly synthesized from nutrients (or other major chemicals) and catalyzed by chemicals that are much less abundant. In turn these chemicals are catalyzed by chemicals that are even less abundant, and this hierarchy of catalytic reactions continues until it reaches the minor chemical species (with, e.g., $n_i < 5$) [13].

Based on this catalytic hierarchy, the observed exponent -1 can be explained using a mean-field approximation. First, we replace the concentration n_i/N of each chemical i, except the nutrient chemicals, by a single average concentration (mean-field) x, while the concentrations of nutrient chemicals S is given by the average concentration $S = 1 - k^* x$, where k^* is the number of non-nutrient chemical species. From this mean-field equation, we obtain $S = \frac{DS_0}{D+\epsilon\rho}$ with $S_0 = \sum_j \overline{n}_j/V$. With linear stability analysis, the solution with $S \neq 1$ is stable if $D < \frac{\epsilon\rho}{S_0 - 1} \equiv D_c$. Indeed, this critical value does not differ much from numerical observation.

Next, we study how the concentrations of non-nutrient chemicals differentiate. Suppose that chemicals $\{i_0\}$ are synthesized directly from nutrients through catalyzation by chemicals j. As the next step of the mean-field approximation we assume the concentrations of the chemicals $\{i_0\}$ are larger than the others. Now we represent the dynamics by two mean-field concentrations. The concentration of $\{i_0\}$ chemicals, x_0, and the concentration of the others, x_1, are represented by the following equations:

$$dx_0/dt = \epsilon x S + \epsilon\rho k^*(x^2 - x x_0) - x_0 D(S_0 - S),$$
$$dx_1/dt = \epsilon\rho k^*(x^2 - x x_1) - x_1 D(S_0 - S),$$

where the average concentration of non-nutrient chemicals x is given by $x = \rho x_0 + (1 - \rho)x_1$. The last terms of the above equations represent the dilution effect by the increase of the cell volume. The solution with $x_0 \neq x_1$ satisfies $x_0 \approx x_1/\rho$ at the critical point D_c. Since the fraction of the $\{i_0\}$ chemicals among the non-nutrient chemicals is ρ, the relative abundance of the chemicals $\{i_0\}$ is inversely proportional to this fraction. Similarly, one can compute the relative abundances of the chemicals of the next layer synthesized from i_0. At $D \approx D_c$, this hierarchy of the catalytic network is continued. Chemicals at a given layer of the hierarchy are synthesized from the nutrients catalyzed by the layer one step down in the hierarchy. The abundance of chemical species in a given layer is $1/\rho$ times larger than chemicals in the layer one step down. Then, in the same way as this hierarchical organization of chemicals, the increase of chemical abundances and the decrease of the number of chemical species are given by factors of $1/\rho$ and ρ, respectively. This is the reason for the emergence of a power law with an exponent -1 in the rank-ordered distribution [14].

In general, as the flow of nutrients from the environment increases, the hierarchical catalyzation network pops up from random reaction networks. This hierarchy continues until it covers all chemicals, at $D \to D_c - 0$. Hence, the emergence of a power-law distribution of chemical abundances near the critical point does not rely on the details of our model, such as the network configuration or the kinetic rules of the reactions. Instead it is a universal property of a cell with an intracellular reaction network to grow, by taking in nutrients, at the critical state, as has been confirmed from a variety of models.

There are two reasons to assume that such a critical state of the reaction dynamics is adopted in existing cellular systems. First, as shown in Fig. 3, the growth speed of a cell is maximal at $D = D_c$. This suggests that a cell whose reaction dynamics are in the critical state should be selected by natural selection. Second, at the critical point, the similarity of chemical compositions between the mother and daughter cell is maximal as shown in Fig. 3. Indeed, for $k > N$, the chemical compositions differ significantly from generation to generation

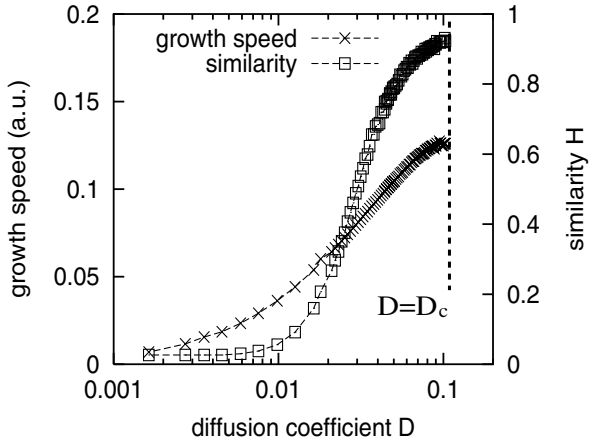

FIG. 3. The growth speed of a cell and the similarity between the chemical compositions of the mother and daughter cells, plotted as a function of the diffusion coefficient D. The growth speed is measured as the inverse of the time for a cell to divide. The degree of similarity between two different states m (mother) and d (daughter) is measured as the scalar product of k-dimensional vectors $H(\mathbf{n}_m, \mathbf{n}_d) = (\mathbf{n}_m/|\mathbf{n}_m|) \cdot (\mathbf{n}_d/|\mathbf{n}_d|)$, where $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ represents the chemical composition of a cell and $|\mathbf{n}|$ is the norm of $\mathbf{n}$ [15]. Both the growth speed and the similarity are averaged over 500 cell divisions. Note that the case $H = 1$ indicates an identical chemical composition between the mother and daughter cells.

when $D \ll D_c$. When $D \approx D_c$, several semistable states with distinct chemical compositions appear. Daughter cells in the semistable states inherit chemical compositions that are nearly identical to their mother cells over many generations, until fluctuations in molecule numbers induce a transition to another semistable state. Hence the most faithful transfer of the information determining a cell's intracellular state is at the critical state. (Inheritance of chemical compositions is also discussed in [15] in connection with the origin of reproducing cells.) In this state, cells of specific chemical compositions are reproduced and can also "evolve" into other states. For these reasons, it is natural to conclude that evolution favors a critical state [16] for the reaction dynamics.

Last, we investigated the relationship between the abundance of a chemical species and the number of reaction paths connected with it. By comparing the SAGE data and the protein-protein interaction data in yeast (*S. cerevisiae*) [17,18], we found that there is a significant negative correlation between the abundance of any given mRNA and the number of protein-protein interaction links that the corresponding protein takes part in ($p < 0.01$; determined by randomization test). In our model simulations, this negative correlation between the abundance of chemical species and the number of possible catalytic paths of the chemical is also found. In this sense, chemicals minor in abundance can play a relatively important role in the control of the behavior of a cell [19]. In the future it will be important to study this kind of interplay in the context of evolution since the evolution of reaction networks has been discussed only in the context of network topology [1].

We thank T. Yomo and L. M. Jakt for stimulating discussions and F. H. Willeboordse and A. Ponzi for a critical reading of the manuscript. This work was supported by 11CE2006.

Note added.—After submission of this Letter, the authors were informed that another illustration of Zipf's law in gene expression data is given in Ref. [20].

[1] H. Jeong *et al.*, Nature (London) **407**, 651 (2000); H. Jeong, S. P. Mason, and A.-L. Barabási, Nature (London) **411**, 41 (2001).
[2] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).
[3] A. E. Lash *et al.*, Genome Res. **10**, 1051 (2000).
[4] V. E. Velculescu *et al.*, Cell **88**, 243 (1997); SAGE data are available from http://www.sagenet.org/
[5] S. J. Jones *et al.*, Genome Res. **11**, 1346 (2001); data are available from http://elegans.bcgsc.bc.ca/SAGE/
[6] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, Science **270**, 484 (1995).
[7] The results reported here do not rely on the specific reaction form we adopted. For example, by using bimolecular reactions, i.e., $(i + \ell \rightarrow j + m)$, without explicit catalyzation, the power-law distribution of chemical abundance appears in the same manner as presented.
[8] Even if the reaction coefficient and diffusion coefficient of penetrating chemicals are not identical but distributed, the results reported here are obtained.
[9] The results reported here do not depend on the distribution of path connectivity. The Zipf's law of abundance is still valid for a scale-free network [1] also, i.e., a network with power-law connectivity distribution.
[10] The results reported here are robust against the change of the number of nutrient chemical species.
[11] Note that, in the case $k > N_{\max}$, the number of some chemical species n_i is 0, while a subpopulation of chemical species sustains the intracellular dynamics.
[12] These three classes of intracellular dynamics also appear when changing the connection rate ρ. There is a critical value $\rho = \rho_c$, where in the case $\rho < \rho_c$ the cell stops growing. The power-law distribution of chemical abundances with an exponent -1 appears at $\rho = \rho_c$.
[13] In the case depicted in Fig. 2, a hierarchical organization of catalytic reactions with 5–6 layers is observed at the critical point.
[14] Within a given layer, a further hierarchy exists, which again leads to the Zipf rank distribution. For details, see C. Furusawa and K. Kaneko (to be published).
[15] D. Segré, B. Danfna, and D. Lancet, Proc. Natl. Acad. Sci. U.S.A. **97**, 4112 (2000).
[16] P. Bak and K. Sneppen, Phys. Rev. Lett. **71**, 4083 (1993); P. Bak, *How Nature Works* (Springer, New York, 1996).
[17] P. Uewtz *et al.*, Nature (London) **403**, 623 (2000).
[18] I. Xenarios *et al.*, Nucleic Acids Res. **28**, 289 (2000).
[19] K. Kaneko and T. Yomo, J. Theor. Biol. **214**, 563 (2002).
[20] V. A. Kuznetsov *et al.*, Genetics **161**, 1321 (2002).