Zipf-like behavior in procaryotic protein expression

J. J. Ramsden[1] and J. Vohradský[1,2]
[1] *Biozentrum of the University, 4056 Basel, Switzerland*
[2] *Institute of Microbiology, Czech Academy of Sciences, Videnska 1083, 14220 Prague, Czech Republic*
(Received 22 June 1998)

The relative rates of synthesis p_r of proteins present in various procaryotic organisms have been found to follow the simple canonical law $p_r \sim (r+\rho)^{-1/\theta}$, where r is the rank. The parameter ρ is interpreted as the bias characterizing the mode of control (i.e., the overall preference for positive or negative control) of gene expression. By analogy with thermodynamics, and drawing parallels with the abstract theory of messages, θ is the informational temperature, which characterizes the extent to which the organism's genome is used to produce proteins. The quantity of selective information H (analogous to thermodynamic entropy) was calculated for the distribution of synthesis rates using Shannon's formula. For all the organisms investigated, H was approximately 8 bits/protein. [S1063-651X(98)14512-9]

PACS number(s): 87.15.−v, 87.10.+e

I. INTRODUCTION

Even a simple procaryotic organism is highly complex. Although focusing on characteristics of the individual molecules of an organism has been the dominant approach during the past decade or so, and has generated an impressive collection of data, we are still far from understanding how these components are integrated as a system. Here an alternative approach is adopted, and we start by examining some of the statistical properties of organisms. We describe some macroscopic parameters characterizing expression of the protein repertoire. This repertoire (called the proteome) occupies an intermediate position between the genome or genotype, the organism's ultimate repository of information, and the phenotype or morphology of the organism.

During an organism's life cycle, the DNA genome is being constantly transcribed into RNA, which in turn is translated into protein (polypeptide), the ultimate mediator of phenotype. Many mechanisms serve to regulate protein synthesis [1,2], and at any given moment the rates of synthesis of different proteins vary over many orders of magnitude. Our aim is to explore whether the *distribution* of these rates is a characteristic and interpretable feature.

The technique of two-dimensional gel electrophoresis [3,4] allows individual proteins in crude cell extracts to be separated and makes it possible to determine their apparent rates of synthesis (or abundance, depending on the mode of protein detection [4]), thus providing a global snapshot of expression. Much effort has hitherto been expended in trying to identify the separated proteins and their functions. This identification is often difficult at the current state of knowledge. For example, the functions of about a third of the proteins in *Haemophilus influenzae*, whose entire genome was recently sequenced, are unknown [5]. The analysis presented here does not require the proteins to be identified, but merely ranked according to abundance or rate of synthesis, and is complementary to recently reported work on patterns of expression [6,7].

II. EXPERIMENT

Liquid cultures of *S. coelicolor* J1501 (*hisA1 uraA1 strA1 pgl* SCP1$^-$ SCP2) were grown from seed precultures as described previously [6,7]. As growth proceeded, samples were pulse radiolabeled with ^{35}S-met/cys. The amount of radioactive sulfur incorporated into the protein is therefore proportional to the rate of protein synthesis at the time of labeling. The rate of protein degradation is not considered rapid enough to significantly affect the protein distribution during labeling (which lasts 40 min [7]). The proteins were extracted from the cells and quantified on 2D gels according to molecular weight (M_r) and point of zero charge (pzc). 51 gels were analyzed. Autoradiographs of the gels were scanned digitally, the density scale was converted to disintegrations per minute (dpm)/mm^2, and integrated spot densities I_r were determined [7]. Densities calibrated with known amounts of protein were found to vary linearly with protein amount. The total density was summed, and normalized to unity in order to obtain the p_r from the integrated spot densities, i.e., $p_r = I_r / \Sigma_{r=1}^{R} I_r$, which were then ranked. Each spot corresponded to an individual protein, or if not, then overlapping spots could be deconvoluted [7]. Not all proteins can be detected using this technique: molecular weights below 15 000 and above 90 000 (corresponding to sequence lengths of ~100–700), and points of zero charge less than pH 3 and greater than pH 8 are excluded, and very low expression levels (the tail of the distribution) are invisible. Moreover, the extraction procedure does not efficiently recover integral membrane proteins, which may constitute an appreciable fraction of the total.

The integrated spot densities are proportional to the product of protein quantity and the number of methionines per protein. We have investigated the possible correlation of methionine content with protein size for several genomes whose sequences have been recently published. The correlation is weak: as an example, for *H. influenzae* [5] the Pearson correlation coefficient equals 0.64 for the entire genome, and only 0.43 for our sequence length window of ~100–700 amino acids. The reason for this is that the different amino acids are not distributed at random in proteins. In the case of methionine, small proteins contain proportionately more than larger ones because (a) the start codon with which each sequence begins specifies methionine, therefore every protein should contain at least one methionine (whether it is post-

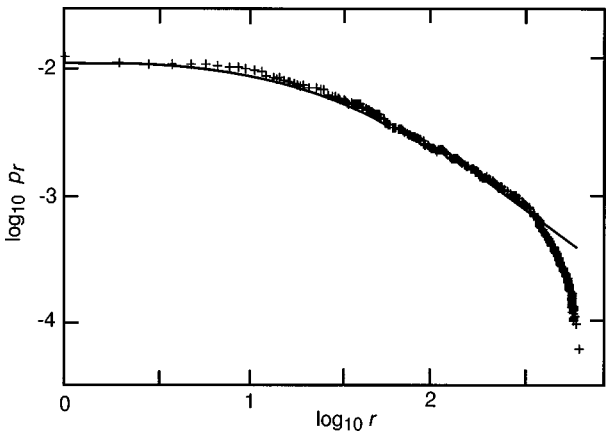

FIG. 1. Plot of $\log_{10}p_r$ vs $\log_{10}r$ for *S. coelicolor* (early growth). Points: data; solid line: Eq. (1) with the parameters given in Table I.

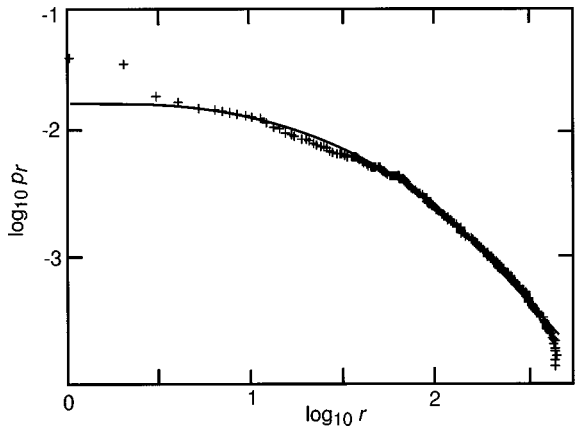

FIG. 2. Plot of $\log_{10}p_r$ vs $\log_{10}r$ for *S. coelicolor* (late growth). Points: data; solid line: Eq. (1) with the parameters given in Table I.

translationally excised depends on the actual protein function), and (b) sulfur is a relatively rare element in most terrestrial environments and hence most organisms have evolved to minimize the use of methionine. Both these effects tend to compensate for the otherwise intrinsic tendency of the number of methionines to increase with protein size. From a biological viewpoint, methionines are distributed not according to protein size, but according to protein function.

We have also directly investigated whether spot density is correlated with protein molecular weight, as it should be were the observed density distribution preponderantly due to methionine content being proportional to M_r. We calibrated our gels with standards of known molecular weight in order to assign an M_r value to each spot, and correlated M_r with spot density. The mean Pearson correlation coefficient for the 51 gels was -0.0067 (the first and third quartiles were -0.0642 and 0.0438, respectively). So we can rather confidently exclude the possibility of this correlation.

III. RESULTS

The ranked distribution of protein synthesis rates follows the so-called simplified canonical law (SCL):

$$p_r = P(r+\rho)^{-1/\theta}, \tag{1}$$

where r is the rank and P, ρ, and θ are the parameters of the distribution. Figures 1 and 2 show two of the distributions plotted as $\log_{10}p_r$ versus $\log_{10}r$. The parameters ρ and θ (Table I) were determined by fitting Eq. (1) to each set of data using a nonlinear least-squares Levenberg-Marquardt algorithm. P is not independent, but is fixed by the requirement that the p_r sum to unity and hence by the state variables R and θ according to [8]

$$P^{-1} = \sum_{r=1}^{R} (r+\rho)^{-1/\theta}. \tag{2}$$

R is the potential number of proteins, i.e., a measure of the size of the proteome. Its value is not very critical if $\theta<1$, because Eq. (2) converges quite rapidly for values of ρ typically encountered. Hence distributions for which $\theta<1$ are called open (i.e., not closed by a fixed value of R) [8,9].

The statistical quality of the fit is good. For 16 gels taken at random from the 51 which were analyzed, containing a total of 6192 spots, the slope and intercept of the best straight line through the calculated $\log_{10}p_r$ plotted against measured $\log_{10}p_r$ were 0.9985 and -0.0027, respectively, and the Pearson correlation coefficient was 0.997 (Fig. 3).

Figures 4 and 5 show data for two other bacteria, *E. coli* and *H. influenzae*; the parameters of the fitted SCL [Eq.(1)] are also given in Table I.

IV. DISCUSSION

The SCL was previously introduced by Mandelbrot in the context of messages, with p_r as the frequency of word usage

TABLE I. Parameters of the expressed protein repertoires of various procaryotic organisms.

Organism	Genome [a]	R [b]	H	$\ln\rho$	θ
E. coli [c]	4.7 Mb	882	9.17	5.17	0.58
H. influenzae [d]	1.8 Mb	244	7.06	3.67	0.54
S. coelicolor [e]	8 Mb	615	8.43	3.52	0.84
S. coelicolor [f]	8 Mb	454	7.75	3.95	0.54
uncertainty [g]			0.15	0.15	0.03

[a]Approximate genome size in millions of bases.

[b]R is here the number of proteins whose synthesis rates or abundances could be estimated.

[c]Proteins were ^{35}S-met labeled. Data provided by R. A. VanBogelen, Parke-Davis Pharmaceutical Research, Ann Arbor, Michigan.

[d]Proteins detected using Coomassie blue G250 staining [21], which detects total protein abundance rather than rate of synthesis [4]. The abundance involves not only gene regulation and protein manufacture (both contributing to the observed rate of synthesis, the quantity detected by pulsed ^{35}S-met labeling), but also the rate of degradation, which is of course also under genetic control. Hence in this case we have a snapshot representing the balance of both production and degradation pathways. Data provided by P. Cash, Department of Medical Microbiology, University of Aberdeen, Scotland.

[e]Proteins were pulse ^{35}S-met labeled. Early growth (12 h).

[f]Proteins were pulse ^{35}S-met labeled. End of growth (72 h).

[g]Estimated from multiple gels. The uncertainties in the parameters for a given gel are about an order of magnitude smaller.

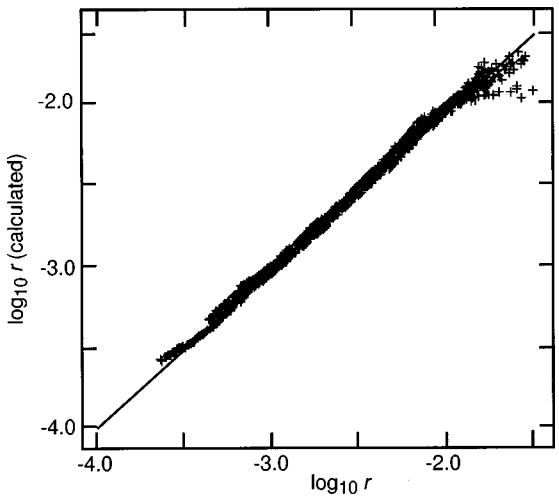

FIG. 3. Plot of $\log_{10}p_r$ calculated from Eq. (1) vs the measured $\log_{10}r$ for 16 *S. coelicolor* gels.

[8], and was shown to be an excellent description of natural languages such as English [9]. This striking statistical feature of languages, i.e., that the probability p_r of the occurrence of a certain word is roughly inversely proportional to its rank r (its position in a list of words arranged in order of decreasing p_r), has been established by extensive empirical studies of Chinese [10] and Indo-European languages [11]. It was proposed by Zipf in the form $p_r \propto 1/r$ [11], but Mandelbrot subsequently showed that the SCL, Eq. (1), fitted the data much better [9] [thus the Zipf law is a special case of Eq. (1) with $\theta=1,\rho=0$].

Furthermore, Mandelbrot established that Eq. (1) is precisely the form to be expected on theoretical grounds if the mean cost ($\Sigma_r p_r \log_2 r$) is to be minimized for a given quantity of information per word [cf. Eq. (3)] and a given potential number of words R [8]. Since the synthesis of a protein by an organism manifestly costs energy, and since it equally manifestly contains information, we propose a linguistic analogy, in which the proteins are the ''words'' and the protein repertoire is the ''vocabulary.'' Protein synthesis is estimated to consume at least a third of the total energy resources of a procaryote [12], and hence it is to be expected that there is strong selection pressure to organize protein synthesis as efficiently as possible, such that the energy expended on the amount of synthesis needed to satisfy the requirements of a given metabolic state is minimized.

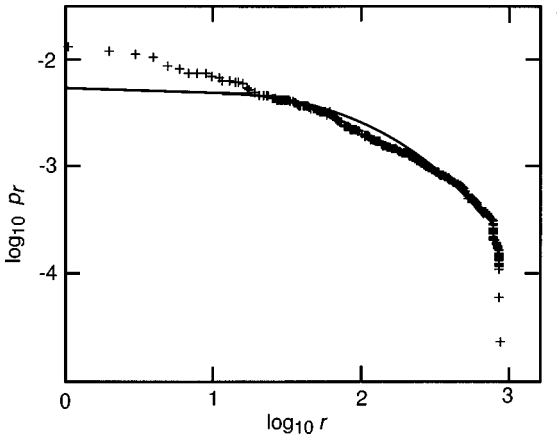

FIG. 4. Plot of $\log_{10}p_r$ vs $\log_{10}r$ for *E. coli*. Points: data; solid line: Eq. (1) with the parameters given in Table I.

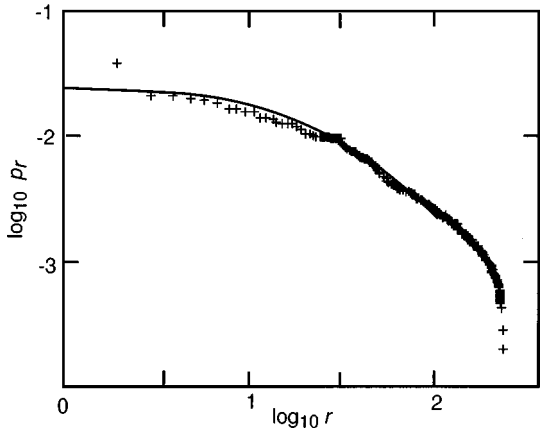

FIG. 5. Plot of $\log_{10}p_r$ vs $\log_{10}r$ for *H. influenzae*. Points: data; solid line: Eq. (1) with the parameters given in Table I.

In all the organisms investigated, the frequency drops more precipitously for very low synthesis rates than predicted by Eq. (1). At present it is not clear whether this is due to some distortion introduced by the procedure used to determine these rates. For example, rare proteins represented by only a few molecules per cell will only need to be synthesized intermittently and may not therefore be registered in our assay.

Regulation: the interpretation of ρ. One of the most intriguing issues in molecular biology is understanding the ways organisms regulate their metabolism. Which sections of DNA get transcribed is the subject of tight control. In procaryotes, a very common mode of control is the operon, which consists of one or more regulator sites (promoter regions) upstream of a gene or group of genes coding the proteins whose expression is to be regulated. Binding of specific regulatory proteins, alone or in combination, to one or several promoter regions can enhance or inhibit expression of many proteins in the same or different operons [13]. [The overall process of information transfer from DNA to polypeptide (protein) is known as gene expression.]

When the enzyme products are practically continuously required by the organism, negative (inhibitory) regulation is preferred, in which the regulatory protein acts to repress gene expression [14,15]; but when the enzyme products are rarely required, the regulatory protein is more likely to be a positive (activator) element. In other words, the process being regulated tends to be either spontaneously active (i.e., biased on) or normally inactive (biased off).

We propose that ρ is a sum parameter representing the overall bias of the organism. A rudimentary analogy would be the Schottky barrier (metal/semiconductor contact), through which the current flowing depends on the position of the Fermi level μ as e^{μ}, i.e., the bigger μ is, the greater the flow. Writing the Fermi-Dirac distribution of energy levels E as $f(E)=e^{-E}(e^{E}+e^{\mu})^{-1}$, by comparison with Eq. (1) we can identify μ with $\ln\rho$. Hence larger values of ρ appear to indicate overall bias towards more positive regulatory systems.

The quantity of selective information H (the analogy of

the thermodynamic entropy) is defined according to [9,16]

$$H = -\sum_{r=1}^{R} p_r \log_2 p_r . \tag{3}$$

H has a value ~ 8 (Table I), i.e., the mean informational content per protein is around 8 bits. This value is much smaller than the information contained in, e.g., the atomic coordinates of a protein, or in the sequence. The latter has been estimated as 2.5–3 bits/residue [17,18] (hence a typical protein will have an information content of thousands of bits), and this is largely shared with the information contained in the structure, i.e., its configurational entropy, equal to its algorithmic complexity [19]. Our value of ~ 8 bits per protein apparently corresponds to the ''macroscopic'' information [20], which is possibly related to the interconnectedness of the regulatory network [2,13], i.e., the mean number of control pathways acting upon a given protein.

θ is the ''informational temperature,'' analogous to the thermodynamic temperature. θ was less than unity for all the procaryotes investigated, that is, the protein repertoire is an open vocabulary in the sense of Mandelbrot [8,9] in common with most natural languages, but not artificial (e.g., Esperanto) or pathological (e.g., Basic English) ones [9]. The smaller θ is, the more strongly is protein expression concentrated on the most frequently occurring proteins.

The analysis carried out here is in close analogy to the reduction in the number of state variables effected in the thermodynamics of inanimate matter. Our primary aim was to obtain a small number of parameters with which an exceedingly complex set of data can be rendered more tractable, without, initially, wishing to specify exactly how these parameters should be interpreted.

By focusing attention on what we hope are key parameters, we should obtain new insights into the way an organism makes use of the resources available in its genome. Our concept provides a bridge between the microscopic, molecular biological characterization of the regulation of expression of individual proteins, and the properties of the ensemble of proteins constituting the whole organism, and could turn out to be very fruitful, for example in the context of questions such as: ''How can the organism use its genome as efficiently as possible? What is the informational value of individual proteins, and the genome as a whole? How many genes are used, and to what extent? How can the organism optimize protein expression for any given physiological or developmental situation? Are the parameters θ and ρ indicative of the metabolic health of the cells?'' We are continuing to investigate these questions.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to X.-M. Li and C.J. Thompson for having provided the 2D gels of *S. coelicolor* for analysis, and R.A. VanBogelen and P. Cash for, respectively, the *E. coli* and *H. influenzae* 2D gel spot intensities. We also thank V.V. Poroikov and B.N. Sobolev for having determined the methionine contents of several genomes, and C.J. Thompson and M.G. Cacace for stimulating discussions.

[1] C. Burks and D. Farmer, Physica D **10**, 157 (1984).
[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *The Molecular Biology of the Cell*, 2nd ed. (Garland Publishing, New York, 1989), Chap. 10.
[3] P. H. O'Farrell, J. Biol. Chem. **250**, 4007 (1975).
[4] R. A. VanBogelen and E. R. Olson, Biotech. Annu. Rev. **1**, 69 (1995).
[5] R. D. Fleischmann *et al.*, Science **269**, 496 (1995).
[6] A. M. Puglia, J. Vohradský, and C. J. Thompson, Mol. Microbiol. **17**, 737 (1995).
[7] J. Vohradský, X.-M. Li, and C. J. Thompson, Electrophoresis **18**, 1418 (1997).
[8] B. Mandelbrot, Publ. Inst. Statist. Univ. Paris **2**, 1 (1952).
[9] B. Mandelbrot, Word **10**, 1 (1954).
[10] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language* (Harvard University Press, Cambridge, MA, 1932).
[11] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949).
[12] L. Gold, Annu. Rev. Biochem. **57**, 199 (1988).
[13] F. J. Neidhart and M. A. Savageau, in *Escherichia Coli and Salmonella*, 2nd ed. edited by F. C. Neidhart (ASM Press, Washington, D.C., 1996), pp. 1310–1324.
[14] M. A. Savageau, Proc. Natl. Acad. Sci. USA **71**, 2453 (1974).
[15] M. A. Savageau, Proc. Natl. Acad. Sci. USA **74**, 5647 (1977).
[16] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[17] H. P. Yockey, J. Theor. Biol. **67**, 345 (1977).
[18] B. J. Strait and T. G. Dewey, Biophys. J. **71**, 148 (1996).
[19] T. G. Dewey, Phys. Rev. E **54**, R39 (1996).
[20] D. S. Chernavsky, Matematika Kibernetika **5**, 3 (1990).
[21] P. Cash, E. Argo, P. R. Langford, and J. S. Kroll, Electrophoresis **18**, 1472 (1997).