Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

C. R. Biologies 326 (2003) 1097–1101

COMPTES RENDUS BIOLOGIES

Molecular biology and genetics

Zipf's law and human transcriptomes: an explanation with an evolutionary model

Osamu Ogasawara, Shoko Kawamoto, Kousaku Okubo [*]

Division of Gene Expression analysis, The Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima 411-8540, Shizuoka, Japan

Received 16 September 2003; accepted 23 September 2003

Presented by François Gros

Abstract

Detailed analysis of human gene expression data reveals several patterns of relationship between transcript frequency and abundance rank. In muscle and liver, organs composed primarily of a homogeneous population of differentiated cells, they obey Zipf's law. In cell lines, epithelial tissue and compiled transcriptome data, only high-rankers deviate from it. We propose an evolutionary process model during which expression level changes stochastically proportionally to its intensity, providing a novel interpretation of transcriptome data and of evolutionary constraints on gene expression. ***To cite this article: O. Ogasawara et al., C. R. Biologies 326 (2003).***

Résumé

Loi de Zipf et transcriptomes humains : explication avec un modèle évolutif. L'analyse détaillée de données d'expression des gènes humains révèle plusieurs types de relations entre la fréquence des transcrits et leur rang d'abondance. Dans le muscle et le foie, organes composés principalement d'une population de cellules différenciées, elles obéissent à la loi de Zipf. Dans des lignées cellulaires, le tissu épithélial et des compilations de données de transcriptomes, seuls les transcrits des premiers rangs en dévient. Nous proposons un modèle de processus évolutif lors duquel le niveau d'expression change de manière stochastique proportionnellement à son intensité, permettant une nouvelle interprétation des données du transcriptome et des contraintes évolutives sur l'expression génique. ***Pour citer cet article : O. Ogasawara et al., C. R. Biologies 326 (2003).***

Keywords: abundance; expression; frequency; transcriptome; Zipf's law

Mots-clés : abondance ; expression ; fréquence ; loi de Zipf ; transcriptome

1. Relationship between transcript frequency and abundance

In the genetics–linguistics analogy, a transcriptome is a text in which a life plan is 'expressed' with a genomic vocabulary. By analyzing SAGE tag [1] and 3′ EST [2] data, we found that the human transcriptome follows the statistical constraint, characteristic for natural language, known as Zipf's law [3]. In a corpus of texts, Zipf's law dictates that the frequency of each word, f, and its abundance rank, r ($r = 1$ for the most frequent word, $r = 2$ for the second most frequent word, and so on) are related according to the power-

[*] Corresponding author.
E-mail address: kousaku@genomatrix.com (K. Okubo).

1631-0691/$ – see front matter © 2003 Published by Elsevier SAS on behalf of Académie des sciences.
doi:10.1016/j.crvi.2003.09.031

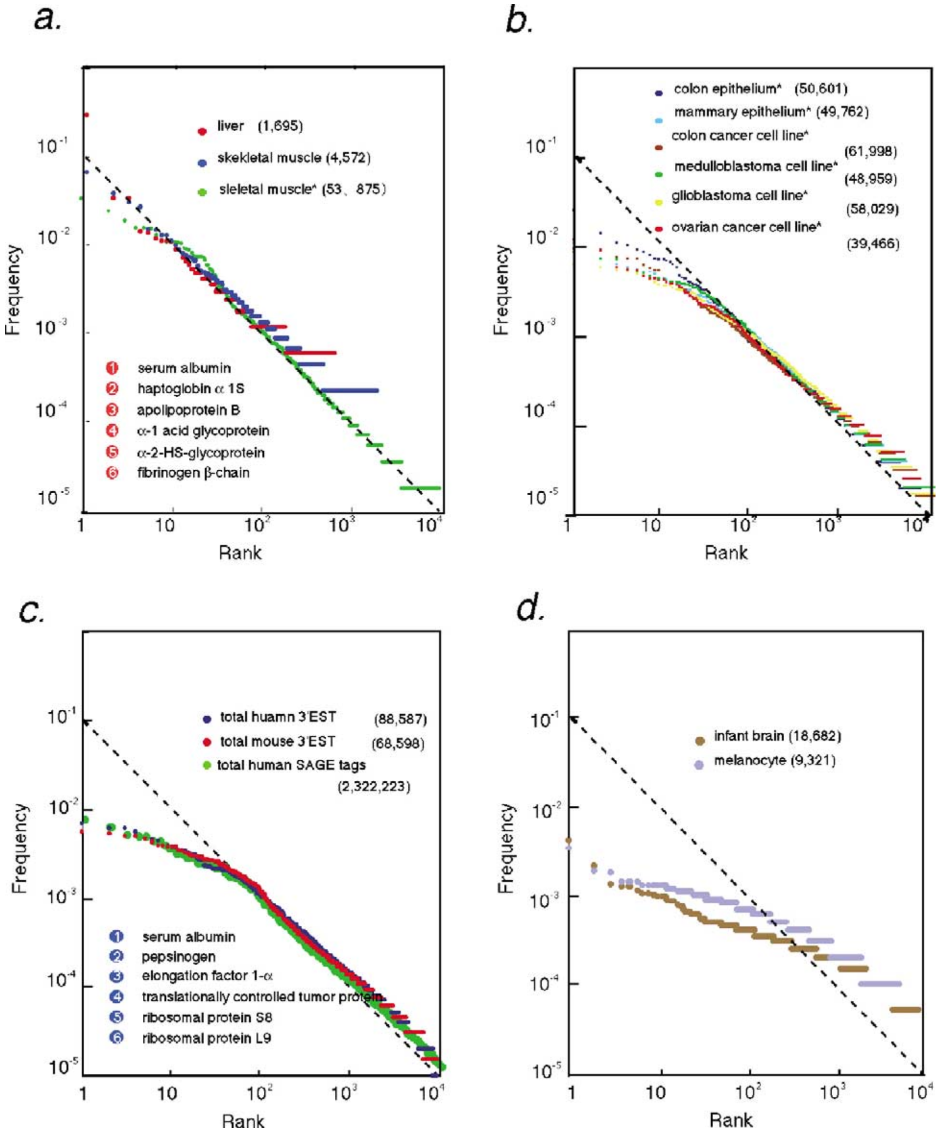

Fig. 1. Log frequency log rank plot (Zipf's plot) of transcriptome data. The frequency of occurrence (f) of each transcript in 3′ EST and SAGE tag (∗) collections representing various transcriptomes were plotted against the abundance rank (r). The broken line represents $f = 0.1/r$. (**a**) Organs with homogeneous populations of differentiated cells. For example, the most abundant transcript $(r = 1)$ in liver, albumin, occurred about 12% in EST data for liver. Gene names for $r = 1$–6 in liver are given. (**b**) Cell lines and complex tissues. (**c**) Compiled data from 51 human EST sets, 31 mouse EST sets, and 64 SAGE tag sets. Gene names for $r = 1$–6 in compiled human transcriptome (3′ EST) are given. (**d**) Occurrence of 3′ EST in normalized libraries. The total tag occurrence for each data set is given in parentheses. The frequency data were obtained from http://bodymap.ims.u-tokyo.ac.jp/datasets/index.html (3′ EST) and ftp://ncbi.nlm.nih.gov/pub/sage/ (SAGE). The data for liver are combined data for two human liver libraries. The frequencies of total SAGE tags are obtained from re-analysis of all available human SAGE tags. Clustering 3′ ESTs for two representative normalized libraries in dbEST, 1N1B and 2NbHM, generated the data for normalized libraries.

law $f \propto r^{-k}$, with $k \approx 1$ for all languages [4]. In a double logarithmic axis plot, such a relation is represented by a linear dependence of f as a function of r with a slope of -1. In muscle and liver, the organs composed primarily of a homogeneous population of differentiated cells, the frequency, f, of each transcript and its abundance rank, r, distributed very close to the line $f = 0.1/r$ (Fig. 1a). In other sources, such as cell lines and epithelial tissue, only high-rankers ($r < 100$), which comprise less than 1% of the transcript variety, deviated from this trend (Fig. 1b). Reduction of tissue-specific transcripts in cell lines and their dilution in complex cell populations explains such deviation, at least in part. Compiling data for different transcriptomes affected the plot similarly (Fig. 1c). Interestingly, the plot appeared universal to the compiled transcriptomes at enough multiplicities, regardless of the data sources. In normalized libraries [5], the Zipf-like structure was lost completely (Fig. 1d).

2. An evolutionary model of expression level variation

In transcript abundance, the slope ($|k|$) is very close to 1, similarly to the classical example in natural language, which suggests the underlying stochastic process that robustly dictates the value of the slope as well as the linearity in log–log plot. We know that the abundance of each transcript is dictated by the DNA sequence on the genome known as cis-elements and the mutations in these elements cause alteration in its abundance. Accordingly it is natural to assume that the abundance distribution results from an evolutionary process. We propose here an evolutional model that explains linearity in log–log plot and the slope ($= 1$) with only simple assumptions as follows: (1) the mutations in cis-elements cause a stochastic change in the expression level that is proportional to the original level of expression; (2) at the initial state, the genome contains from a small number of genes with an arbitrary distribution of gene expression. The gene number gradually increases over generations by gaining new genes by gene duplication [6].

The first assumption can be written as the relation $f_i(t+1) = \lambda_i(t) f_i(t)$, where $f_i(t)$ is the expression level of gene i at generation t, and $\lambda_i(t)$ is a random variable extracted from a time-independent distribution $\pi(\lambda)$.

After T generations, we obtain:

$$\log f_i(T) = \log f_i(0) + \sum_{j=0}^{T} \log \lambda_i(j) \qquad (1)$$

If the random variable $\lambda_i(t)$ is independent and identically distributed with finite mean and variance, the central limit theorem says that the distribution of $\sum_{j=0}^{T} \log F_i(j)$ converges to a normal distribution. For sufficiently large T, $f_i(T)$ is well approximated by the lognormal distribution which is extremely similar in shape to power-law distribution.

As generations pass, the deviation of the distribution becomes large, and expression level of some genes becomes very small. If they become lower than the minimal level, then the genes are regarded as disappeared.

Because we cannot determine the initial distribution $f_i(0)$ of formula (1), *a priori*, the model (1) should be extended slightly. At the initial state, the genome contains a small number of genes with an arbitrary distribution of gene expression. The genome grows over generations by gaining new genes by gene duplication. While the number of gene (N) increases, the total number of mRNA molecules in a cell (W) also increases, according to the following formula:

$$\frac{\Delta N}{\Delta W} = \frac{N(t+1) - N(t)}{\sum f_i(t+1) - \sum f_i(t)} = K$$

where K is a constant positive value smaller than 1.

By simulation of this process, we show that transcriptome distribution converges to Zipf's law (Fig. 2). The evolutionary model is essentially the analogy of the city size growth model, proposed by Blank and Solomon (2000) [7]. In their city size model, the population growth (ΔW) is the cause of the increase in the number of cities (ΔN). In our evolutionary model, increase in total mRNA (ΔW) is driven by the growth of gene number (ΔN).

3. Interpretation of transcriptome data

Zipf's law predicts various important numerical features of the transcriptome. According to this law, 50% of transcripts in a differentiated cell represent only 83 mRNA species. Because the accumulated sum of f for all transcripts is 1, the predicted number of different transcripts in a cell is restricted to 12 367. These values are in accordance with the classical view of an average transcriptome where less than 100 genes are responsible for 50% of the cellular mRNA content and about 10 000 mRNA species comprise the rest [8]. Assuming that the plot for a compiled transcriptome (Fig. 1c) remains unchanged after further com-

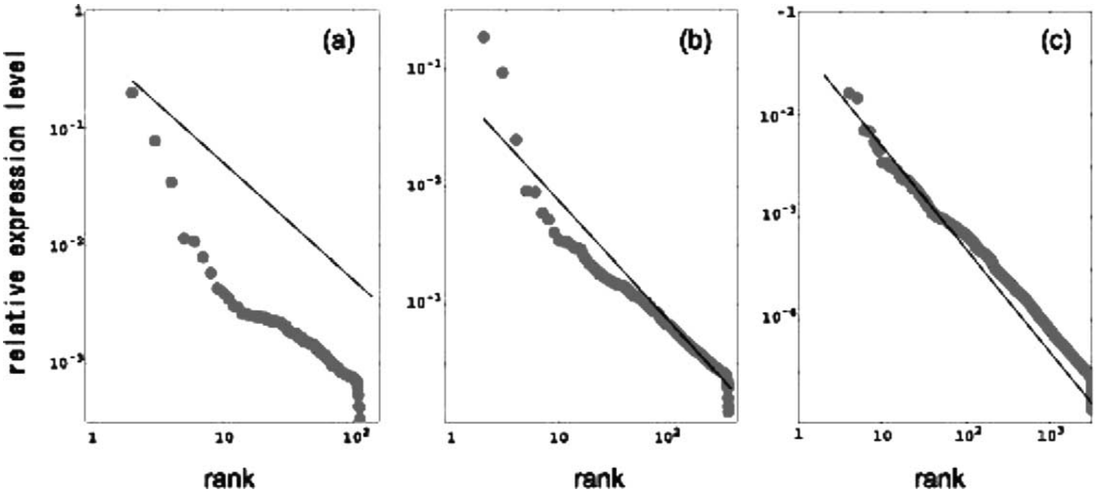

Fig. 2. Simulated emergrence of Zipf-like distribution in transcriptome with evolutionary model. (**a**) Distribution after 50 generations, (**b**) after 100 generations, (**c**) after 200 generations. The continuous line is the Zipf's law function ($f = a/r$). The distribution converged to Zipf's law within a few hundred generations. The number of genes at initial state was 50, and the number was increased exponentially (2% par generation). But the initial state and the exponent parameter values have no substantial influence on the result of the distribution. Other parameters were set as $K = 0.001$, $\pi(\lambda)$ was set as the normal distribution $N(\mu, \sigma^2)$ with $\mu = 1.0$, $\sigma^2 = 0.05$. Again, the resultant distribution was not sensitive to the exact value of the parameters.

pilation of different transcriptomes ($\sum_{r=1}^{10\,000} f = 0.70$ for SAGE tags), the predicted total number of transcripts in a whole body is approximately 210 000. The further benefit of the discovery of such a constraint is that it may eventually assist researchers in designing experiments with tissue mRNA and interpreting results from them. In transcript profiling by DNA microarray, for example, the relation of f and r can be considered as the relation between sensitivity of transcript detection and the number of hybridizing signals.

Applicability of Zipf's law has been reported not only for natural languages but also in several social, economical and behavioral phenomena, such as city size and income [3], and clicks in World Wide Web surfing [9]. Although there have been many controversies, in essence, three kinds of mathematical models have survived to explain such phenomena: The random network model [10] for WWW, Mandelbrot's optimization model [11] for linguistical Zipf's law and a multiplicative model applied to city size [7]. Recently, power-law distributions were also found in wide variety of genomic properties, such as populations of protein families [6,12–14], protein folds [6,14], pseudogenes [14] and protein domains [15,16], although the slope is not always close to 1 in these cases.

After this work was submitted, two papers were published reporting that Zipf's law holds in a wide variety of eukaryotes, supporting our evolutionary model [17,18]. In one of these papers, Furusawa and Kaneko [18] applied a random network model in mRNA production to explain Zipf's distribution, in which they assumed that transcriptional regulation can be modeled as an analogy of a catalytic reaction.

References

[1] V.E. Velculescu, S.L. Madden, L. Zhang, A.E. Lash, J. Yu, C. Rago, A. Lal, C.J. Wang, G.A. Beaudry, K.M. Ciriello, B.P. Cook, M.R. Dufault, A.T. Ferguson, Y. Gao, T.C. He, H. Hermeking, S.K. Hiraldo, P.M. Hwang, M.A. Lopez, H.F. Luderer, B. Mathews, J.M. Petroziello, K. Polyak, L. Zawel, W. Zhang, X. Zhang, W. Zhou, F.G. Haluska, J. Jen, S. Sukumar, G.M. Landes, G.J. Riggins, B. Vogelstein, K.W. Kinzler, Analysis of human transcriptomes, Nature Genet. 23 (1999) 387–388.

[2] T. Hishiki, S. Kawamoto, S. Morishita, K. Okubo, BodyMap: a human and mouse gene expression database, Nucleic Acids Res. 28 (2000) 136–138.

[3] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, UK, 1949.

[4] J.L. Casti, Bell curves and monkey languages, Complexity 1 (1995) 12–15.

[5] L.D. Hillier, G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish,

M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M. Marra, Generation and analysis of 280 000 human expressed sequence tags, Genome Res. 6 (1996) 806–828.

[6] J. Qian, N.M. Luscombe, M. Gerstein, Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model, J. Mol. Biol. 313 (2001) 673–681.

[7] A. Blank, S. Solomon, Power laws in cities population, financimarkets and internet sites (scaling in systems with a variable number of components), Physica A 287 (2000) 279–288.

[8] B. Lewin, Genes VI 659–60, Oxford University Press, New York, 1997.

[9] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, R.M. Lukose, Strong regularities in World Wide Web surfing, Science 280 (1998) 95–97.

[10] A.L. Barabási, R. Albert, H. Jeong, Mean-field theory for scale free random networks, Physica A 272 (1999) 173–187.

[11] B. Mandelbrot, in: Symposium on Applications of Communication Theory, 1953, pp. 486–502.

[12] M. Gerstein, A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure, J. Mol. Biol. 274 (1997) 562–576.

[13] M.A. Huynen, E. van Nimwegen, The frequency distribution of gene family sizes in complete genomes, Mol. Biol. Evol. 15 (1998) 583–589.

[14] N.M. Luscombe, J. Qian, Z. Zhang, T. Johnson, M. Gerstein, The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties, Genome Biol. 3 (2002) 1–7.

[15] G.P. Karev, Y.I. Wolf, A.Y. Rzhetsky, F.S. Berezovskaya, E.V. Koonin, Birth and death of protein domains: a simple model of evolution explains power law behavior, BMC Evol. Biol. 2 (2002) 218–223.

[16] E.V. Koonin, Y.I. Wolf, G.P. Karev, The structure of the protein universe and genome evolution, Nature 420 (2002) 218–223.

[17] V.A. Kuznetsov, G.D. Knott, R.F. Bonner, General statistics of stochastic process of gene expression in eukaryotic cells, Genetics 161 (2002) 1321–1332.

[18] C. Furusawa, K. Kaneko, Zipf's law in gene expression, Phys. Rev. Lett. 90 (2003) 088102 Epub.