

Distribution Associated with Stochastic Processes of Gene Expression in a Single Eukaryotic Cell

Vladimir A. Kuznetsov

Laboratory of Integrative and Medical Biophysics, NICHD, NIH, Bethesda, MD 20892, USA
Email: vk28u@nih.gov

Received 1 August 2001 and in revised form 18 September 2001

The ability to simultaneously measure mRNA abundance for large number of genes has revolutionized biological research by allowing statistical analysis of global gene-expression data. Large-scale gene-expression data sets have been analyzed in order to identify the probability distributions of gene-expression levels (or transcript copy numbers) in eukaryotic cells. Determining such function(s) may provide a theoretical basis for accurately counting all expressed genes in a given cell and for understanding gene-expression control. Using the gene-expression libraries derived from yeast cells and from different human cell tissues we found that all observed gene-expression levels data appear to follow a Pareto-like skewed frequency distribution with parameters dependent of the size of the libraries. We produced the skewed probability function, called the binomial differential distribution, that accounts for many rarely transcribed genes in a single cell. We also developed a novel method for estimating and removing major experimental errors and redundancies from the Serial Analysis Gene Expression (SAGE) data sets. We successfully applied this method to the yeast transcriptome. A “basal” random transcription mechanism for all protein-coding genes in every eukaryotic cell type is predicted.

Keywords and phrases: gene expression, stochastic processes, Pareto-like distributions, binomial differential distribution, number of genes, single cell.

1. INTRODUCTION

Cells must adjust genome expression to accommodate changes in their environment, and in outside signals. Gene expression within a cell is a complex process involving chromatin remodeling, selective transcription of DNA into mRNA, mRNA export from the nucleus to cytoplasm where it is translated into proteins. The expression level of any protein-coding gene is generally measured by the number of associated mRNA transcripts (messenger RNA abundance) present in a sample from many thousands of cells. While the mRNA abundance in a cell at a given moment does not guarantee the precise prediction of amounts of subsequently produced protein, mRNAs, sampled from same-type cell population, nevertheless serve as important indicators that a certain proteins are being produced.

The complete gene expression profile for a given cell is the list of all expressed genes, together with each gene's expression level defined as the number of cytoplasmic mRNA transcripts in the cell [1, 2, 3]. However, gene-expression profiling technologies (e.g., serial analysis of gene expression (SAGE) [4, 5, 6], cDNA, GeneChips methods [7, 8, 9]) currently can only measure the gene-expression levels for a fraction of all expressed genes based on sampling transcripts

found in many thousands of cells (i.e., not a single cell). These methods to determine certain short tags on a transcript, one can then count the numbers of transcripts carrying the same tag. Many genes, in particular those expressed at low levels, cannot be unambiguously detected due to the limited sampling of transcripts and experimental errors. However, many of these lower level transcripts may be essential for determining normal and pathological cell phenotypes.

The expression levels of genes in such assays range typically between 0.1 to 500 transcript per yeast cell [4, 5, 6, 7], and between 0.1 to 30,000 transcripts per human cell [1, 5, 6]; a large proportion of these genes had less than 1 transcript per cell [1, 4, 5, 6, 7, 8, 9]. Such gene-expression data has skewed long tail frequency distributions [10], which are also often observed in physiological processes [11] and in DNA-related phenomena [12, 13, 14], as well as in many self-organizing systems with strong stochastic components [15, 16]. It becomes increasingly evident that stochastic processes within signaling pathways and crosstalk between different pathways need to be considered to fully understand basic processes of gene expression [17, 18, 19, 20]. In particular, a large body of evidence indicates that gene transcription is a discrete process by which many individual protein-coding genes exist in an off state, but can stochastically switch to the

on state [19, 20, 21, 22, 23]; the production of mRNAs occurs in sporadic pulses skewed around the average [19, 22]. Such statistical knowledge bears upon the fundamental biological problems of cell regulation, adaptation, and development.

Statistically gene-expression behavior can be characterized by the gene-expression level probability function (GELPF). The GELPF is a function that for each possible gene-expression level value takes on the probability of that value occurring for a given gene. For a cell or cell population, this function specifies the proportions of expressed genes which have 1, 2, and so forth, transcripts present. Given histograms of gene-expression level values, we can model the underlining “population” probability functions. General features of gene-expression patterns were elucidated more than 25 years ago through RNA-DNA hybridization measurements [1]. However, mathematical models of the underlying true distribution of gene expression levels have not been previously identified due to undersampling and non-reliable detection of many low abundance genes, as well as sequencing errors and complications of tag-gene matching. The goal of this study is to develop such a model for eukaryotic cells.

2. DATA BASES, METHODS, AND SOFTWARE

2.1. Data bases

There are several useful methodologies that allow global quantitative measure of RNAs captured from cells of interest. All these techniques make and then use DNA sequences complementary to less stable mRNA molecules.

cDNA library method counts the number of sequences having the matching or overlapping sequences. Such cDNA expression sequence tags (ESTs) consisting of similar or overlapping ~ 500 nucleotide cDNA sequences, called UniGene clusters [2, 3], are used to group observed sequences into clusters representing presumed genes or ESTs on sequence homology. The UniGene clusters can be used to “tag” genes expressed in specific cell types. The occurrence frequencies of each UniGene in cDNA library might serve as estimators of the gene-expression levels in the cell population from which the cDNA library was constructed.

The SAGE methodology is based on isolating distinct 10-nucleotide DNA sequences called *SAGE tags* from 3' end regions of individual transcripts and concatenating the tags serially into long DNA molecules [4, 6]. Cloning and sequencing of such molecules allows the identification and enumeration of cellular mRNA transcripts. Since the genome organization in yeast (*Saccharomyces cerevisiae*) is relatively simple, and since almost all yeast genes are known, we have analyzed with a large yeast SAGE database (www.sagenet.org, <http://genome-www.stanford.edu/Saccharomyces>) [4]. We analyzed three SAGE libraries for yeast cells in log phase, S-phase-arrested, and G2/M phase-arrested states separately and pooled. SAGE and cDNA libraries for various human cell lines and cell tissues were downloaded from CGAP (www.ncbi.nlm.nih.gov/CGAP; www.ncbi.nlm.nih.gov/SAGE) databases. Some of these libraries are characterized in Table 1.

DNA chip technology can measure the expression of

thousands of genes simultaneously [24] onto $\sim 1 \text{ cm}^2$ square onto which a cDNA mixture derived from a cell population can hybridized to form labeled complimentary spots. For example, in Affymetrix GeneChips, each gene is represented on the high density oligonucleotide arrays by ~ 20 unique 25mer oligonucleotide probes, that match the sequence of the gene (perfect match oligo's) and 20 oligonucleotide probes that are identical but differ by one base (mismatch oligo's). The gene-expression levels in a cell sample estimated by the mean of the differences in the hybridization signals of matched and mismatched probes labeled message hybridized with the probes. The mean difference value over 20 paired signals is a measure of the expression level of that gene. Based on this estimate, the computed score of the gene-expression levels can sometimes be negative; therefore, database could be additionally scaled for positive values. In this paper, data sets of oligonucleotide arrays containing probes for $\sim 6,200$ yeast open reading frames (ORFs) and genes (GeneChip[®] Ye6100 arrays, Affymetrix, Santa Clara, CA) [7, 9] have been downloaded, scaled, and analyzed.

2.2. Goodness of fit analysis method, sampling and software

Let the data points $(m, g(m))$ for values $m = 1, \dots, J$ form an empirical relative frequency distribution g . Note that $\sum_{m=1}^J g(m) = 1$. We adjusted the vector of parameters \bar{a} (a_1, a_2, \dots, a_v) in the model probability function $f(m; \bar{a})$ to fit the histogram points $(m, g(m))$ by maximizing the similarity between f using the modified Akaike's Information Criteria [25] which we will call the Model Selection Criteria (MSC):

$$\Psi = \log \left(\frac{\sum_{m=1}^J (g(m) - E(g))^2}{\sum_{m=1}^J (g(m) - f(m; \bar{a}))^2} \right) - 2 \frac{v}{J}, \quad (1)$$

where J is the maximum observed value of m , v is the number of unknown parameters of the model f , and $E(\cdot)$ is the mean value of the observed data. The most appropriate model will be that with largest Ψ . The Ψ is independent of the scaling of the data points. The Ψ -criterion ranges between excellent (11,8), very good (8,6), and satisfactory (6,4).

Assuming each identified transcript is selected at random, we used Monte-Carlo sub-sampling of transcripts in a library without replacement to generate sub-libraries. Sub-libraries were used in order to generate same-size libraries for comparison and to construct the growth curve for distinct true tags or genes of a given library (see below).

Parameters in both differential and algebraic models were estimated using the Marquardt-Levenberg iterative curve fitting algorithm in MLAB mathematical modeling software (Civilized Software, Inc., www.civilized.com) coupled with Monte-Carlo refinements of random sampling fluctuations. We also used an additional standard goodness of fit MLAB criteria (sum of squares for deviations, a residual analysis, the Wilcoxon 2-sample rank-order test, etc.). Symbolic differentiation and subsampling were performed using MLAB. Monte-Carlo experiments and numerical analysis were also performed in MS Digital Visual Fortran. Data-mining tools of

TABLE 1: Fitting of the GDP-model to the empirical frequency distributions for cDNA and SAGE libraries of human cell tissues and SAGE libraries of yeast cells. $k \pm SE$, $b \pm SE$ are the estimated parameters. Ψ (model selection criterion) is the goodness of fit criterion. p_1 is the fraction of distinct tags represented by one copy. Unilib identifiers (www.ncbi.nlm.nih.gov/UniLib): 2427 (choriocarcinoma, cDNA library (Life Technology method)), 2892a (LNCaP, prostate cancer cell line, SAGE libraries), 166 (normal colon, SAGE), 2892b (the prostate cancer cell line library 2892a after one-year upgrading), 161 (pooled normal brain tissues, SAGE), and 154 (normal brain cells, > 95% white matter, SAGE). Three yeast SAGE libraries and a pool of these libraries are characterized in [4].

Sample	M	N	M/N	p_1	J	J/M	$k \pm SE$	$b \pm SE$	Ψ
Lib. 2427	10087	3586	2.81	0.54	246	0.029	1.88 ± 0.04	1.34 ± 0.05	7.1
Lib. 2892a	6313	3531	1.79	0.81	78	0.012	1.05 ± 0.015	-0.48 ± 0.008	10
Lib. 166	14616	5383	2.72	0.70	462	0.032	1.28 ± 0.01	0.015 ± 0.02	10
Lib. 2892b	22637	9348	2.42	0.74	221	0.010	1.08 ± 0.03	-0.28 ± 0.01	11
Lib. 161	49334	15182	3.25	0.59	832	0.017	1.44 ± 0.01	0.57 ± 0.007	8.3
Lib. 154	81516	19137	4.26	0.53	1598	0.020	1.25 ± 0.012	0.57 ± 0.016	7.1
Yeast, G2/M	19527	5303	3.68	0.67	519	0.027	0.96 ± 0.006	-0.195 ± 0.006	8.8
Yeast, S-phase	19871	5785	3.44	0.67	561	0.028	0.98 ± 0.004	-0.197 ± 0.004	9.8
Yeast, log-phase	20096	5324	3.78	0.66	636	0.032	0.97 ± 0.004	-0.173 ± 0.004	9.3
Yeast, total	59494	11329	5.25	0.62	1716	0.029	0.94 ± 0.008	-0.108 ± 0.008	7.7

the Cancer Research Anatomy Project including X-profiling, SAGE/map [26], have been also used.

3. EMPIRICAL SKEWED HISTOGRAMS AND PARETO-LIKE STATISTICS

We define a *library* as a list of cDNA's sequenced tags that match mRNAs together with the number of occurrences of each specific tag observed in a cell sample. The size of a library, M , is the total number of tags observed in the library. Let $n(m, M)$ denote the number of distinct tags, which have expression level m (tags) in the library of size M . Let J denote the maximum observed expression level of tags in the library. Let $N = \sum_{m=1}^J n(m, M)$; N is the number of distinct tags in the library. The points $(m, n(m, M)/N)$ for $m = 1, \dots, J$ form the histogram corresponding to the empirical relative frequency distribution $g(m)$. Note that $\sum_{m=1}^J n(m, M)/N = 1$.

Note that due to experimental errors, the observed values of m and n might only approximately reflect the transcripts numbers (or gene expression level) for a given gene and the number of genes represented by m transcripts, respectively. The observed values M and N also only approximately reflect the total number of mRNA transcripts and the number of different transcripts in a library, respectively (see below).

The histogram of the proportions of distinct tags (the 10 bp tags of SAGE libraries or the expression sequence tags (ESTs) of cDNA libraries) represented by one, two, and so forth, tags is the empirical relative frequency distribution of tags which reflects the gene expression levels in a given cell sample. This is a size-frequency form of the probability distribution which represents an estimate of the GELPF for the corresponding cell sample (cf. Figures 1a and 2a). We found that such histograms, constructed for all analyzed yeast and human gene-expression libraries, exhibited remarkably similar, monotonically-skewed shapes (see also Figures 1a and 2a)

with a greater abundance of rarer transcripts and more gaps among the higher-occurrence level values.

Several classes of skewed probability functions (Poisson, exponential, logarithmic series, power law Pareto-like [27]) were fit to empirical gene expression level histograms for various libraries. The best fit (by our criteria) was obtained using the discrete Pareto-like probability function [10]:

$$f(m) := \Pr(X = m) = \frac{1}{z} \frac{1}{(m+b)^{k+1}}, \quad (2)$$

where the random variable X is the expression level for a randomly chosen distinct tag (representing a gene). The function value $f(m)$ is the probability that a randomly chosen distinct tag is represented by m tags (representing an expression level). The argument m denotes a possible value of X . The function f involves two unknown parameters, k , and b , where $k > 0$, and $b > -1$; z is the generalized Riemann Zeta-function value: $z = \sum_{j=1}^J 1/(j+b)^{k+1}$.

Note our model involves the sample-dependent quantity $J = J(M)$. We call equation (2) the Generalized Discrete Pareto (GDP) model. The parameter k characterizes the skewness of the probability function; the parameter b characterizes the deviation of the GDP distribution from a simple power law (with $b = 0$, see for example, dotted line on Figure 1a).

The GDP model with $b \neq 0$ provides the best fit to almost all empirical histograms we studied. In the log-log plot forms, the empirical distributions for larger human SAGE and all cDNA libraries show systematic deviations from a straight line (cf. Figure 2a). In SAGE libraries with a library size less than $\sim 40,000$ tags, the GDP-model at $b = 0$ fits well (see Figures 1a and 2a).

Let $R_e(m) = (\sum_{j=1}^m j \cdot n(j, M)/N)/M$. This is the cumulative fraction of the total number of tags for genes represented by m or fewer tags in a given library. The corresponding theoretical cumulative fraction function $R(m)$ is calculated

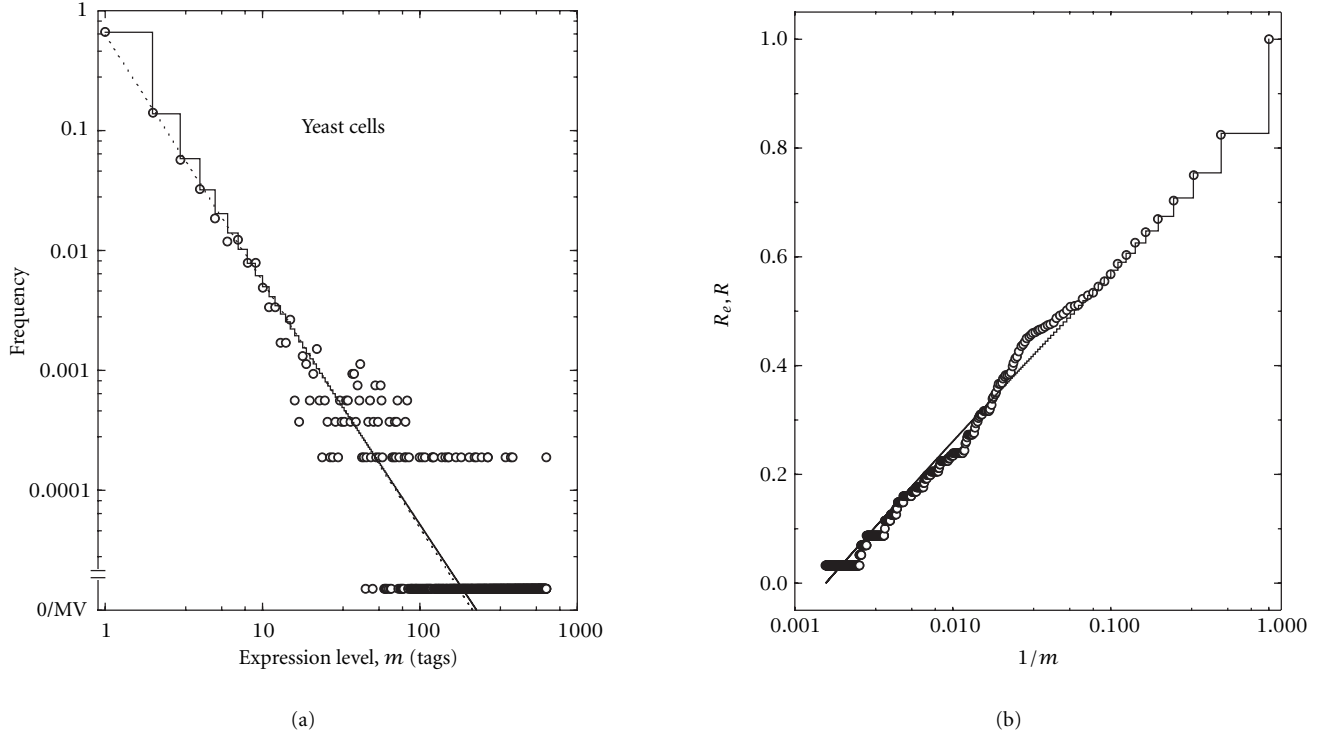


FIGURE 1: Fitting the empirical relative frequency distributions of gene expression levels. (a) Log-log plot. \circ : frequency of expression levels for the log-phase yeast cell growth library of size 20,096 tags; solid step line: best-fit generalized discrete Pareto (GDP) model with parameters $k = 0.974 \pm 0.004$, $b = -0.173 \pm 0.004$; dotted direct line links (for guidance) the best-fit values for $m = 1, 2, \dots$ at $k = 1.03 \pm 0.005$ and $b = 0$. MV is a missing value. (b) Cumulative fraction plot for SAGE tags computed from the empirical histogram and the GDP distribution shown on figure (a).

in terms of the fitted probability function model f as follows:

$$R(m) = \frac{\sum_{j=1}^m j \cdot f(j)}{\sum_{j=1}^J j \cdot f(j)}. \quad (3)$$

Note that $\sum_{j=1}^J j \cdot f(j) = E(X)$. For goodness of fit assessment over the entire range of gene-expression levels, we plot the cumulative fraction of the total number of transcripts in a given library on the ordinate versus the reciprocal of expression level $1/m$, on the abscissa. Using cumulative data reduces the apparent “noise” in the histogram data. The plot of the cumulative function $R(m)$ versus $1/m$ in Figure 2b and Table 1 confirm that the underlying fit GDP model fits over the entire range of expression levels, even when the number of cloned sequences in the library was greater than 80,000. By our criteria, the GDP model has priority in comparison to more complex models, for example, a mixture distribution logarithmic series and exponential distributions. For example, for library sizes greater than 40,000, the values of the MSC-criterion for the latter model were regularly ~ 20 –40% less than for the GDP model.

Note that, given the number of distinct tags, N , and the best-fit parameters, we sample the values of m at random based on the function $f(m)$ $2N$ times (once for each gene), then we count the occurrence numbers of calculated values m

in the intervals $(0, 1]$, $(1, 2]$, \dots and construct the frequency histogram for a given value N and corresponding random value M (see Figure 4b). This Monte-Carlo procedure was used in order to estimate the variability of any expression levels associated with N distinct tags in a given library and to estimate the maximum gene-expression level. Using this procedure many times with the GDP model, we computed the largest expression level J and the factor s such that $s = J/M$ for each Monte-Carlo experiment. We then averaged these scale-factors to obtain their mean \hat{s} . We did that for our SAGE libraries and found that values of \hat{s} ranged in $[0.012$ – $0.045]$. Similar ranges were observed in empirical ratios J/M (see Table 1).

4. EFFECT OF LIBRARY SIZE ON EMPIRICAL DISTRIBUTIONS

Similarly-sized libraries derived from various human tissues have many similar numbers of expressed genes (see figure legend, Figure 2a and Table 1). They also are characterized by similar empirical relative frequency distributions of gene-expression levels with nearly equivalent parameters in their best-fit probability function models (cf. the prostate cancer cell line (library 2892a) and sub-sampled normal brain cell tissue library 154 in Figure 2a). Although the yeast genome

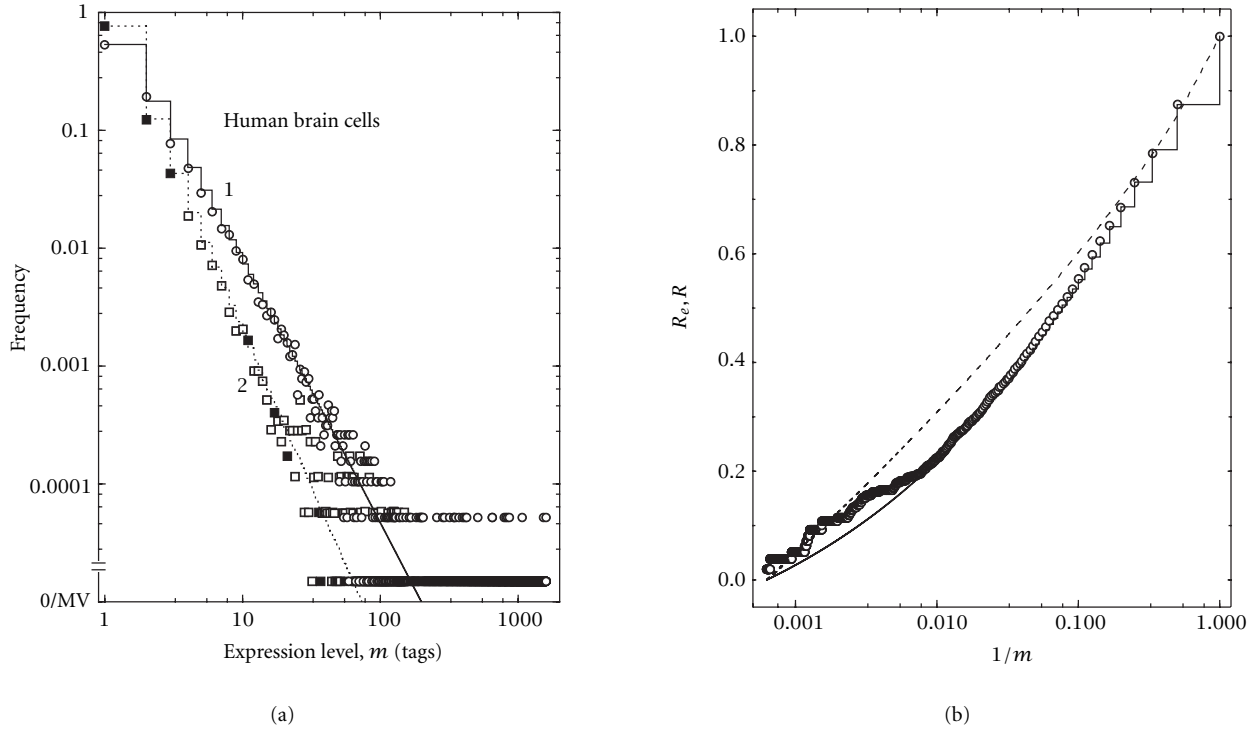


FIGURE 2: Fitting the empirical relative frequency distributions of human gene-expression levels. (a) Log-log plot. 1, \circ : frequency of expression levels for human normal brain cell library 154 of size 81,516 tags; solid line: best-fit GDP model for \circ data; 2, empty square and black square: the average frequency of expression levels for 10 sub-libraries of size 6313 tags taken at random without replacement from library 154 and represented in an average by 3497 distinct tags; dashed line: best-fit GDP model for these data with parameters $k = 1.62 \pm 0.07$, $b = 0.01 \pm 0.004$; black square indicates a frequency value which is significantly different (at $p < 0.05$) from corresponding frequency value in human prostate cancer library 2892a with size 6313 tags represented by 3531 distinct tags (library 2892a data is not presented). MV is a missing value. (b) Cumulative fraction plot for SAGE tags computed from the empirical histogram (\circ) and corresponding GDP model (solid line), shown on figure (a), and from the best-fit regular discrete Pareto model ($k > 0, b = 0$; linked with dashed line for guidance).

is less complex, yeast libraries show similar relationships (see Table 1).

However, as library size increases, the fraction of low abundance distinct tags becomes smaller (see parameter p_1 , Table 1), and the shape of the probability distribution function changes systematically (b becomes bigger, see Table 1; Figure 2a). We also found that the value of the maximum observed gene-expression level in the sample, J , was linearly correlated with the library size M (Table 1).

Thus, we might assume that all human cells and yeast cells have a common GELPF. However, a single fixed GDP model (where parameters are constants) cannot describe all empirical frequency distributions independent of library size, since the probability function changes as the number of transcripts in a library becomes larger.

Interestingly, in self-similar (fractal) systems, described by a power law or Pareto-like distributions, the parameter(s) are independent of the size of the system [16], but not in our case. Moreover, such models, including the GDP model, predict an unlimited increase in the number of species as the sample size approaches infinity, whereas the number of expressed genes is a finite number. The problems of library size dependence of

the GDP model parameters and the incorrect infinite limit for the number of genes as $M \rightarrow \infty$ are both solved by introducing a new statistical distribution model. This new model also explains the GELPF invariance for many cell types.

5. BINOMIAL DIFFERENTIAL DISTRIBUTION

We assume that (1) the number of expressed genes in cell population is a finite number, N_t , (2) each gene in a given cell population is expressed with a certain probability, and (3) a transcription event for a given gene is statistically independent of such events for other genes. Although transcription events of some genes may in fact be correlated in a given cell, most transcription events in a cell population seem to be random, independent events. This is consistent with observations in [17, 21, 22, 23]. We further assume that tags in a library are chosen at random. Our assumptions are consistent with constructing such libraries by sampling from a hypergeometric distribution [27]. We also take into account that a typical SAGE (and cDNA) library size ($\sim 10^3$ – 10^5 tags) is much smaller than the number of transcripts in a typical cell sample ($> 10^{11}$ transcripts in $> 10^6$ cells). That allows us to

use the multinomial approximation [27, 28] of the hypergeometric distribution and leads to the following GELPF model.

We assume that N_t genes $1, 2, \dots, N_t$ are expressed with M_t associated transcripts in total in the cells of a large cell population. Also assume that these genes are expressed independently with respective probabilities q_1, q_2, \dots, q_{N_t} , where \Pr (a random transcript corresponding to gene i) = q_i .

Let the random variable s_i denote the number of transcripts in a random library of size M . Note $\sum_{i=1}^{N_t} s_i = M$. When $M \ll M_t$, sampling with replacement is an acceptable model of library construction. This follows a multinomial distribution. The joint probability of observing $s_1 = \gamma_1$ mRNA transcripts of gene 1, $s_2 = \gamma_2$ mRNA transcripts of gene 2, $\dots, s_{N_t} = \gamma_{N_t}$ mRNA transcripts in a given library with size M is defined by the probability function $f(\gamma_1, \dots, \gamma_{N_t}; M) := \Pr[s_1 = \gamma_1, \dots, s_{N_t} = \gamma_{N_t}]$, where

$$f(\gamma_1, \dots, \gamma_{N_t}; M) := \frac{M!}{\prod_{j=1}^{N_t} \gamma_j!} \prod_{j=1}^{N_t} q_j^{\gamma_j}. \quad (4)$$

The function f has the unknown parameters q_1, q_2, \dots, q_{N_t} , and N_t , together with the constraints $\sum_{i=1}^{N_t} q_i = 1$ and $\sum_{j=1}^{N_t} \gamma_j = M$.

The marginal probability function $f_i(m; M) := \Pr(s_i = m)$ is the probability that the distinct tag for gene i occurs exactly m times in our library of size M :

$$f_i(m, M) = \frac{M!}{m!(M-m)!} q_i^m (1-q_i)^{M-m}. \quad (5)$$

We can estimate the expected number of distinct genes, $n(m, M)$, which have m transcripts in our library of size M . Let $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. Now,

$$n(m, M) := \sum_{i=1}^{N_t} E(\delta_{s_i, m}) = \sum_{i=1}^{N_t} f_i(m; M). \quad (6)$$

Let the random variable $G_M = \#\{i \mid s_i > 0\}$; G_M is the number of distinct genes represented in the library of size M . We can estimate the expected number of genes $N(M)$ in a given library as $E[G] := N(M)$, where

$$N(M) := \sum_{m=1}^M n(m, M) = \sum_{j=1}^{N_t} \left(1 - (1 - q_j)^M\right). \quad (7)$$

Thus,

$$\begin{aligned} N(1) &= \sum_{j=1}^{N_t} q_j = 1, \\ N(2) &= 2 \left(1 - \sum_{j=1}^{N_t} q_j^2\right), \\ N(3) &= 3 \left(1 - \sum_{j=1}^{N_t} q_j^2 \left(1 - \frac{1}{3} q_j\right)\right), \dots \end{aligned} \quad (8)$$

Finally,

$$N(M) = N_t - n(0, M), \quad (9)$$

where $n(0, M)$ denotes the expected number of distinct genes which escaped detection in the given library; $n(0, M)$ is given as

$$n(0, M) := \sum_{j=1}^{N_t} (1 - q_j)^M. \quad (10)$$

Now, using (5), (6), (7), and (10) we can derive the recursion formulas

$$\begin{aligned} \frac{n(0, M)}{M} - \frac{n(0, M+1)}{M+1} &= \frac{1}{M+1} \frac{n(1, M+1)}{M+1} + \frac{n(0, M)}{(M+1)M}, \\ \frac{n(1, M)}{M} - \frac{n(1, M+1)}{M+1} &= \frac{2}{M} \frac{n(2, M+1)}{M+1}, \\ &\vdots \\ \frac{n(m, M)}{M} - \frac{n(m, M+1)}{M+1} &= \frac{m+1}{M-(m-1)} \frac{n(m+1, M+1)}{M+1} \\ &\quad - \frac{m-1}{M-(m-1)} \frac{n(m, M)}{M}, \end{aligned} \quad (11)$$

where $m \in \{0, 1, \dots, M\}$. Also, $n(m, M) = 0$, if $m > M$. These results allow us to compute $n(m, M)$ for any given values of m and M .

On the other hand, taking into account the mass conservation law

$$\begin{aligned} n(0, M) - n(0, M+1) &= N(M+1) - N(M) \\ &= \frac{n(1, M+1)}{M+1} \end{aligned} \quad (12)$$

and using the initial conditions $N(1) = 1$, $n(1, 1) = 1$, we can obtain an important relationship between N and $n(1, j)$, where $j = 1, \dots, M$, as follows:

$$N(M) = \sum_{j=1}^M \frac{n(1, j)}{j}. \quad (13)$$

Equation (13) shows that the expected number of all genes in a library is determined by the expected numbers of unique species (distinct genes occurred once) for the sample sizes that ranged from 1 to M .

Using (5), (6), (7), (10), and (11), we can rewrite $n(m, M)$ in terms of N and M as follows:

$$\begin{aligned} n(0, M) &= N_t - N(M), \\ n(1, M) &= M(\nabla N(M)), \\ n(2, M) &= -\frac{M(M-1)}{2} (\nabla^2 N(M)), \\ &\vdots \\ n(m, M) &= (-1)^{m+1} \frac{M!}{m!(M-m)!} (\nabla^m N(M)), \end{aligned} \quad (14)$$

where ∇ is the backward difference operator [27]. If $m =$

1 then $\nabla N(M) = N(M) - N(M-1)$, and $n(1, M) = M(\nabla N(M))$. In general, $\nabla^m N(M) := \nabla^{m-1} N(M) - \nabla^{m-1} N(M-1)$.

If $m > 1$ and M is large enough, then we have the ‘‘quasi-steady state’’ relationship

$$n(m+1, M) \approx n(m, M) \frac{m-1}{m+1}. \quad (15)$$

Using this recursive formula with $m > 1$, we obtain

$$n(m+1, M) \approx \frac{n(1, M)}{(m+1)m}. \quad (16)$$

Equations (7) and (16) can be used to estimate the probability p_m that a randomly-chosen gene from $\{1, \dots, N_t\}$ has exactly m transcripts in a given library, that is, $p_m \approx n(m, M)/N$. Then, for large M and $m > 1$ we have

$$p_{m+1} \approx \frac{p_1}{(m+1)m}. \quad (17)$$

The probability function p_m has a skewed form, and is approximated by the power law form ($p_m \sim m^{-2}$; Lotka-Zipf law, <http://linkage.rockefeller.edu/wli/zipf>), which describes many other large-scale, complex phenomena such as income, word occurrence in a text, numbers of citations to journal article, and so forth.

When M is large enough, we can approximate (14) with its continuous analog and obtain the probability function p_m , in terms of M and N as follows:

$$p_m \approx h(m) := (-1)^{m+1} \frac{1}{N} \cdot \frac{M!}{m!(M-m)!} \frac{d^m N}{dM^m}, \quad (18)$$

where $m = 1, 2, \dots$. The function $h(m)$ with the parameters M and N taken as function of M will be called the binomial differential (BD) probability function. Taking $m = 1$ in (18), we obtain a differential equation

$$\frac{dN}{dM} = p_1 \frac{N}{M} \quad (19)$$

with $N(1) = 1$. We call the function N defined by (19) the population ‘‘logarithmic growth’’ (LG) model. Note that (19) could be rewritten in the following explicit form:

$$\frac{dN}{dM} = \sum_{j=1}^{N_t} q_j (1 - q_j)^{M-1}, \quad (20)$$

where the right side is a sum of geometric distribution probabilities of an initial success in a sequence of M trials. However, the values of q_j and N_t are unknown. Using (19) and (20), we can show that p_1 is a monotonically decreasing function of M . We will use the empirical approximation

$$p_1 = \frac{1 + (1/d)^c}{1 + (M/d)^c}, \quad (21)$$

where the c and d are positive constants (see Figure 4a). This function was selected among many possible forms for $p_1(M)$ by fitting the LG model to data points obtained from

many yeast and human SAGE libraries and sub-libraries (not presented). Using an explicit specification of p_1 allows us to fit the BD model to empirical histograms. The parameter $1/c$ roughly characterizes the rate of accumulation of genes (or gene tags), and the parameter d roughly estimates the maximum number of genes (or gene tags), N_t . With the above empirical choice for p_1 , (19) has an exact solution

$$N(M) = \left(M^c \frac{1 + 1/d^c}{1 + (M/d)^c} \right)^{(1+1/d^c)/c} \quad (22)$$

with

$$\lim_{M \rightarrow \infty} N(M) = N_t = (1 + d^c)^{(1+1/d^c)/c}. \quad (23)$$

We now have an explicit, although complicated, expression for the BD probability function

$$h(m) = (-1)^{m+1} \frac{1}{N} \frac{M!}{m!(M-m)!} \cdot \frac{d^m}{dM^m} \left(M^c \frac{1 + 1/d^c}{1 + (M/d)^c} \right)^{(1+1/d^c)/c}. \quad (24)$$

Thus, unlike the fixed GDP models, the BD probability function depends on the number of distinct genes, N , and the library size, M ; it also yields the finite value N_t for the total number of genes as $M \rightarrow \infty$. Equations (19), (20), (21), (22), (23), and (24) will be used below to exclude redundancies present in yeast SAGE libraries and thereby to accurately estimate the GELPF for a single yeast cell.

6. ANALYSIS OF EXPERIMENTAL ERRORS

To identify the correct distribution of gene-expression levels in cell types by fitting the empirical gene-expression levels histograms, we must first eliminate the experimental errors in SAGE libraries so the corresponding histograms will be unbiased. Since almost all yeast protein-coding genes and open reading frames ORFs (an ORF is a DNA sequence which is (potentially) translatable into protein, that is, likely to be a gene) are known, we can obtain the ‘‘true’’ distinct tags and their expression levels in a yeast SAGE library by eliminating the erroneous tags that fail to match known genes/ORFs in the Tag Location database for yeast transcriptome [www.sagenet.org, <http://genome-www.stanford.edu/Saccharomyces>]. This database was generated by Velculescu et al. [4] and currently contains information about $\sim 8,500$ distinct SAGE tags, match $\sim 4,700$ genes/ORFs (of ~ 6200 known genes/ORFs in the yeast genome), together with the chromosome coordinate of each SAGE tag, the strand, and the associated gene name(s) (where relevant) and the chromosome location of genes/ORFs. In our analysis, the 10 bp sequences immediately downstream of the 3' most NlaIII site found within the gene/ORF or within 500 bp genomic adjacent genomic region with 3' NlaIII site have been taken as ‘‘true SAGE tags.’’ Thus, the ‘‘true’’ tags are those tags that mostly match ORFs/genes, but do not match any noncoding regions or opposite (non-translated) strand.

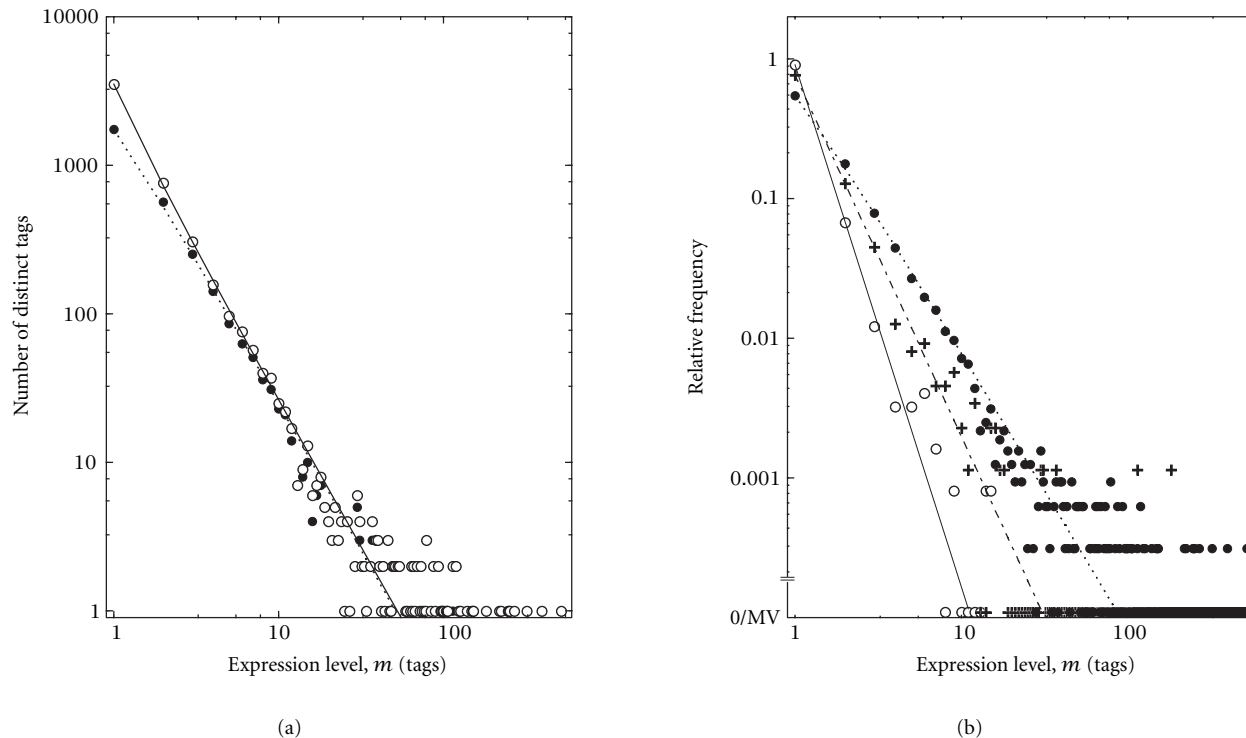


FIGURE 3: Decomposition of the frequency distribution of gene-expression levels for SAGE yeast cells library. (a) Log-log plot. \circ : the numbers of 5,303 distinct tags represented by 19,527 tags in G2/M phase-arrested cells library; dashed line: best-fit GDP models (with $b = -0.195 \pm 0.005$, $k = 0.96 \pm 0.006$) for $+$ -data; \bullet : the numbers of “true” tags of the same library after removing erroneous tags; dotted line: best-fit GDP model (with $b = 0.207 \pm 0.013$; $k = 0.991 \pm 0.011$) for \bullet -data. (b) Frequency distribution of “true” tags and best fit GDP model. \bullet , dotted line: “true” tags; \circ , solid line: “outside” erroneous tags; $+$, discontinue line: “inside” erroneous tags. Fitted probability function values (counted at $m = 1, 2, \dots$) are linked by lines for guidance of the visual presentation of the models.

Figure 3a shows the empirical histogram of the 5,303 distinct tags represented by 19,527 tags in the yeast library derived from G2/M phase-arrested cells, and of the 3,200 “true” distinct tags of the same library after the elimination of 2,103 distinct tags associated with 3,239 tags that match noncoding genomic regions and antisense sequences. Most of these erroneous tags occur with only 1 or 2 copies (Table 2, Figure 3b). These erroneous tags comprise 16.6% of the 19,527 tags in the library and might be considered as a sum of sequencing erroneous (“outside”) tags which do not match yeast genome at all, and a false-positive (“inside”) tags matching the noncoding regions or the opposite strands. The “inside” erroneous tags consist 9.2% of library size (Table 2). Figure 3 and Table 2 show the GDP model at $b = 0$ (simple power law) fitted well a frequency distribution of different classes of erroneous tags, but $b > 0$ in the case of frequency distribution of true tags.

Matches of many distinct tags to the same gene, and one distinct tag to many genes, constitute serious and common problems in correctly identifying genes and properly determining their expression levels [5, 26, 29], particularly in larger SAGE libraries. Such matching confusions are associated with using short-length (10 nucleotide) tags and with

the existence of multiple restriction sites on the 3’ end of sequences [5, 26, 29]. Thus, we have tags with redundancy match the same genes/ORFs as do other tags as well as tags that match several different genes/ORFs. These difficult problems are, obviously, more acute in the case of higher organisms due to the higher complexity of their genome. Figure 4a shows that the difference between the growth curves for “true” distinct tags and for ORFs matched by “true” tags rapidly increases for $M > 10,000$. This difference reflects a rapid increase in the mean number of distinct “true” tags per gene as library size increases. In particular, we observed that 20% of ORFs (596 of 2,936 ORFs) have more than one matching distinct tag in the library of size 19,527 “true” tags for G2/M phase-arrested cells, and that 41% of ORFs (1,817 of 4,439 ORFs) have more than one matching distinct tag in the pooled yeast library of size 49,073 “true” tags.

Importantly, tags that matched only noncoding DNA regions and “redundant” tags apparently have not been correctly discarded in any recent predictions of the number of expressed genes in cell types. Therefore, basing such estimates on uncorrected bigger human SAGE libraries (100,000–600,000 tags) must lead to a significant over-estimation of the number of expressed genes in human cell types (cf. [5]).

TABLE 2: Decomposition of the Pareto-like distribution of G2/M-phase arrested yeast cells SAGE library. Characteristics of distributions of different classes of SAGE tags are as follows: the erroneous tags that fail to match the entire yeast genome sequences “outside” errors, mostly associated with sequencing errors), the erroneous tags that fail to match known ORFs/coding regions or mapping within its 500 bp adjacent downstream genomic regions (“inside” errors), and the “true” tags which contain a fraction of ambiguity matching tags.

Sample	M	N	M/N	p_1	J	$k \pm SE$	$b \pm SE$	Ψ
G2/M	19527	5303	3.68	0.67	519	0.96 ± 0.01	-0.195 ± 0.006	8.8
Outside errors	1447	1234	1.17	0.91	15	2.74 ± 0.02	0.0	10.2
Inside errors	1792	869	2.06	0.76	182	1.59 ± 0.01	0.0	8.5
True Tags	16288	3200	5.09	0.55	519	0.99 ± 0.01	0.21 ± 0.01	7.3
Genes or ORFs	15000	3009	4.99	0.382	288	1.56 ± 0.04	2.17 ± 0.08	7.7

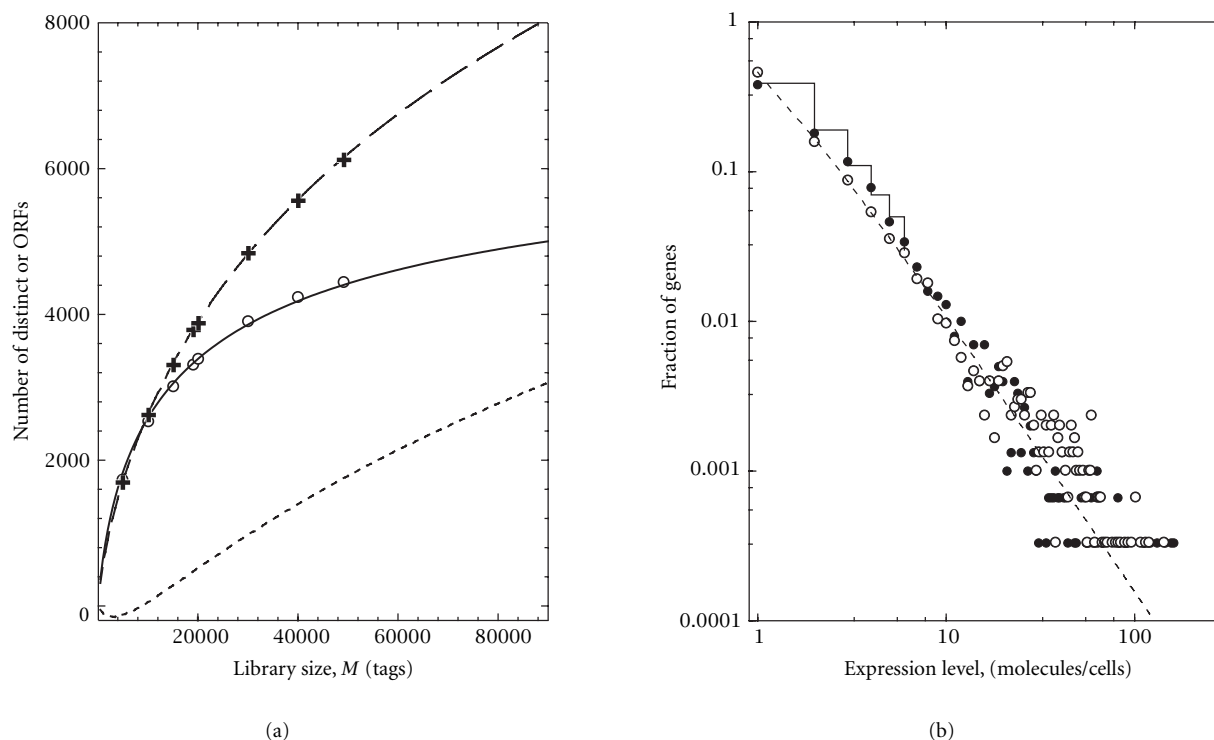


FIGURE 4: Population growth curves and estimates of the GELPF for a single yeast cell. (a) Growth curves. +: the number of “true” distinct tags of sub-libraries from the pooled yeast library of pooled “true” tags; dashed line: best-fit LG model (with $d = 20,000 \pm 1,946$; $c = 0.356 \pm 0.02$) for +-data; o: the number of distinct ORFs found in these sub-libraries; best-fit LG model (with $d = 6,575 \pm 185$, $c = 0.579 \pm 0.01$) for corresponding ORFs data; short-dashed line: the number of redundant “true” tags. (b) Log-log plot. Solid step line: the relative frequencies of ORFs by the BD model for a single yeast cell estimated by SAGE data; • a histogram generated from fitted GDP model for 3,009 ORFs in a single yeast cell, and o: a relative frequency of 3,009 ORFs in a single log-phase yeast cell, estimated by GeneChip data [9]. Dashed line links the fitted GDP model values to o-data points for $m = 1, 2, \dots$ at $k = 0.86 \pm 0.01$, $b = 0.37 \pm 0.003$.

7. ESTIMATING THE NUMBER OF GENES IN YEAST CELLS

A common difficulty encountered with SAGE (and cDNA) methodology is that there is no easy way to determine the numbers of different types of mRNAs expressed in a single cell and in a population of the cells. One approach is to exhaustively oversample until no new transcripts are observed [5]. However, as we showed above, this approach itself, without adequate analysis and filtration of gene-expression data, leads to dramatic accumulation of intrinsic experimental

errors and redundant tags. Using our new species richness estimator ((19) and (21)), it is possible to obtain accurate and robust estimates of the total number of distinct mRNA transcripts expressed in a single cell and in cell type, including those not observed in an available database, using relatively small sample with only partial actual coverage.

We found that the LG model ((19) and (21)) fits both the size-dependent data for “true tags” and for ORFs/genes (Figure 4a). However, in the case of “true” distinct tags (+, Figure 4a) (but where tags-to-gene and tag-to-genes multiple

matches were not considered), the LG model predicts a very large value, $25,103 \pm 2,000$ genes (by (23) with $d = 20,000 \pm 1,946$; $c = 0.356 \pm 0.02$) in the large yeast cell population. When we tabulated the distinct ORFs that correspond to these “true” distinct tags at various sample sizes (\circ , Figure 4a) and fit this M versus N data, the fitted LG model predicts $7,025 \pm 200$ genes/ORFs with $c = 0.579 \pm 0.010$ and $d = 6,580 \pm 190$) in a yeast cell population.

This estimate is ~ 4 – 10% higher than current estimates of the total number of distinct ORFs in the yeast genome ($6,200$ – $6,760$ genes/ORFs) [30, 31]. This difference could be due to the small number of erroneous tags and redundant tags which nevertheless match genes/ORFs and their adjacent genomic regions. Our analysis does not take into account missed ORFs within the yeast genome (in particular, shorter ORFs), and overlapped ORFs. Additionally, about 1–3% of transcripts would be expected to lack an NlaIII anchoring enzyme site and would therefore be missing in the database. Using an estimate of the number of mRNAs per yeast cell ($M_{\text{cell}} = 15,000$ [4]), (23) predicts 3,009 ORFs per cell. This estimate is consistent with the number of genes/ORFs for a single yeast cell in the G2/M phase-arrested state (2,936 ORFs matched by “true” distinct tags in this library) and with a published estimate of ORFs for a single yeast cell in the log-phase of cell growth [4], which also was based on tabulating the distinct ORFs found in the yeast tag location database.

8. ESTIMATING THE GELPF FOR A SINGLE YEAST CELL

First, we used the BD-model (24) with the fitted parameters $c = 0.579 \pm 0.010$ and $d = 6,580 \pm 190$ in $p_1(M)$ to compute values p_1, \dots, p_6 for 3,009 ORFs corresponding to the library size $M_{\text{cell}} = 15,000$. Then we fit the GDP model (2) to these 6 points and extrapolated the fitted GDP model to estimate values of p_m for $m > 6$. This use of the GDP model was necessary because numerical algorithms cannot accurately and reliably compute values of high-order derivatives [32]. However, when we fit the GDP model and the BD model to the same empirical histograms for “true” distinct tags, we observed that the GDP model is a good approximation of the BD model (data not shown). Moreover, both the BD model and the GDP model are power law forms with a similar shape. These observations justify using the GDP model to estimate p_m for larger m . To check the self-consistency of our predictions, we, additionally, estimated the total number of transcripts, M , from the fitted GDP model and noted that the result was 15,000.

Figure 4b shows the predicted GELPF at all possible levels of gene expression for a single yeast cell. The step-function (solid line) represents the relative frequencies estimated by the BD model (step-function, solid line) for low-abundance genes consisting of 85% of $\sim 3,000$ genes/ORFs in a yeast cell. The GELPF was estimated with the use of the GDP model which was fitted to the BD data points and then extrapolated for larger abundance mRNA transcripts. The theoretical histogram (\bullet) in Figure 4b was generated in 3,009 Monte Carlo experiments by sampling from fitted GDP distribution and

than by counting the numbers of genes/ORFs found at a same expression value.

Figure 4b shows that 38% of $\sim 3,000$ expressed genes are represented by a single mRNA copy per cell. Moreover, Figure 4a shows that a given single cell (at $M = 15,000$ transcripts per cell) expresses only 45% of all protein-coding genes; the other 55% of all protein-coding genes are expressed at very low levels (< 1 copy per cell).

We used data obtained by GeneChip technology [9] to construct the empirical histogram of the gene expression levels in untreated log-phase yeast cells (Figure 4b). This histogram was constructed as follows: for each ORF/gene, we converted the scaled hybridization intensity signal value, I , in the yeast GeneChip database [9], to the number of mRNA molecules per single yeast cell by the empirical formula $m = (I - 20)/165$. The conversion shows close agreement with the estimates of transcript numbers per cell for of 16 different yeast genes [8] observed in three different yeast GeneChip data bases [7, 8, 9]. Then summing of m -values in the unit intervals centered at $1, 2, \dots, 143$ (an estimated value of the maximum gene expression level in the log-phase yeast cell estimated for the library) produce gene expression levels for a single yeast cell characterized by the GeneChip.

We then obtained a gene expression levels histogram (\circ , Figure 4b). The entire expression level ranges contained $\sim 3,000$ expressed genes/ORFs representing $\sim 16,000$ transcripts per cell. Figure 4b shows that the frequency distribution for GeneChip data also follows the GDP model ($k = 0.86 \pm 0.001$, $b = 0.37 \pm 0.003$ at $\Psi = 7.4$). Similarly, skewed frequency distributions were also observed in other (untreated) yeast cell GeneChip libraries found in [7, 9].

Thus, the distribution predicted by our analysis of SAGE data and our estimated frequency distribution based on GeneChip data are close to each other (see Figure 4b). A larger fraction of unique transcripts in the case of GeneChip data ($\sim 45\%$ versus 38% in our SAGE data distribution) is expected because the microarray methods are more sensitive in determining, at least, low-abundance genes [7, 8]. A relatively small systematic differences between the tails of the two distributions might be because the hybridization intensity score does not strongly linearly correlate with the target molecule concentration for highly abundant transcripts [7, 33]. Observed deviations between our two gene expression level distributions could also be related to differences in experimental conditions, experimental normalization procedures, and cell types. However, both experimental techniques provide Pareto-like distributions.

9. DISCUSSION

This paper has demonstrated that the empirical histograms of gene expression levels for yeast cells in various cell cycle stages and for all analyzed human cell types, are well described by a “generalized” power law, called the Binomial Differential (BD) distribution. For a given sample size, this skewed distribution is approximated by the GDP model.

We also found that the empirical histograms of gene expression levels change in the same way for many cell types or

cell states as the number of transcripts in a library changes (Figure 2a, Table 1). The skewed form and quantitative similarity of the empirical histograms of gene expression levels for any two same-size libraries, regardless of human cell type, suggest a common underlying GELPF, perhaps due to the action of a common stochastic mechanism for gene expression. This conclusion also applies to the BD model, which assumes that almost all protein-coding genes in a cell are expressed sporadically and independently.

Modeling SAGE experiments in yeast has allowed us to develop a method to estimate the cumulative numbers of expressed protein-coding genes and of erroneous and redundant sequences. After eliminating the erroneous tags and redundant tags, we estimated that $\sim 55\%$ of all yeast protein-coding genes are expressed at very-low levels (< 1 transcript per cell) in a single cell. This 55% estimate is consistent with data in the yeast high-density oligonucleotide array databases (52% [7], and 56% [8]). About 70% of all protein-coding human genes are also estimated to be expressed with < 1 transcript per cell (V. Kuznetsov, 2001, unpublished data). Such low copy numbers in a large cell sample may be due to the action of a random transcription process. Such a random processes have been observed, both spatially and temporally, in a variety of cell systems [17, 18, 19, 20, 21, 22, 23]. In particular, Chelly et al. [23] have detected low abundance transcripts of various tissue-specific genes (genes for anti-Mullerian hormone, beta-globin, aldolase A, factor VIIc, etc.) and in human nonspecific cells, such as fibroblasts, lymphoblasts, hepatoma cells. The existence of a random transcription process implies that all or almost all protein-coding genes in a genome have a small but positive probability to be transcribed in any given cell during a fixed time-interval. Although not all cells of a population would have a copy of a specific transcript at a given moment, we would expect to see all these genes expressed, at least at a low level, in a sufficiently large cell population at any point in time. That is, ergodicity holds.

A random transcription mechanism could provide a basic level of phenotypic diversity in a cell population and thus could facilitate adaptation. This also assumes a “basal” transcription level of almost all genes (including their exons) in a large same-type cell population in global transcriptional response of cells due to internal random perturbations. In normal yeast libraries [9], we observed that only ~ 250 ORFs of ~ 6200 yeast ORFs/genes are not detected. About 100 of these 250 ORFs are classified as questionable ORFs and, additionally, more than 50 other of 250 ORFs are classified as hypothetical protein ORF. Treatment with 6 different damaging factors [9] shows that only ~ 100 yeast ORFs was still not observed using GeneChip technology. However, most genes/ORFs are still represented by a very small number of transcripts.

10. CONCLUSION

We have developed a novel approach for global characterization of large-scale gene-expression data sets. This approach is based on statistical modeling and parametric identification

of size-frequency distributions of the gene expression levels data. This approach allowed us to estimate the number of genes/ORFs at very low expression levels in a single yeast cell as well as to estimate the total number of genes/ORFs in a population of these cells. Similar method might be developed for counting the number of expressed genes in other eukaryotic cell types.

We have found that transcript populations appear to follow a discrete Pareto-like skewed distribution in a number of different human tissues and in yeast cells, suggesting that this distribution can represent a universal statistical characteristic of many eukaryotic cells. It provides new insight into the statistical mechanics of gene expression levels in cells.

Identification of the GELPFs may be important in current attempts to characterize “complete” profiles of gene expression in normal and diseased human cells. We have also analyzed differences of the gene expression level distributions in different cell types and cell states of higher eukaryotic organisms that will be covered in future reports.

It seems, the binomial differential distribution could be applicable for analysis of many other complex large-scale systems (e.g., in business, linguistic, informatics, internet, physics) having a strong stochastic component.

ACKNOWLEDGEMENTS

I wish to thank Robert Bonner and Gary Knott for very useful discussions. I am also grateful to Igor Belyakov, Jay Berzofski, Konstantin Chumakov, Ralph Nossal, Robert Strausberg, Aly Strunnikov, and Tatyana Tatusov and anonymous referees for critical comments of the manuscript. I also thank Victor Velculescu for supplementary materials on analysis of yeast SAGE database.

REFERENCES

- [1] J. O. Bishop, J. G. Morton, M. Rosbash, and M. Richardson, “Three abundance classes in hela cell messenger RNA,” *Nature*, vol. 250, no. 463, pp. 109–204, 1974.
- [2] R. L. Strausberg, K. H. Buetow, M. R. Emmert-Buck, and R. D. Klausner, “The cancer genome anatomy project: building an annotated gene index,” *Trends Genet.*, vol. 16, no. 3, pp. 103–106, 2000.
- [3] M. R. Emmert-Buck, R. L. Strausberg, D. B. Krizman, et al., “Molecular profiling of clinical tissue specimens: feasibility and applications,” *Am. J. Pathol.*, vol. 156, no. 4, pp. 1109–1115, 2000.
- [4] V. E. Velculescu, L. Zhang, W. Zhou, et al., “Characterization of the yeast transcriptome,” *Cell*, vol. 88, no. 2, pp. 243–251, 1997.
- [5] S. L. Velculescu, V. E. Madden, L. Zhang, et al., “Analysis of human transcriptomes,” *Nat. Genet.*, vol. 23, no. 4, pp. 387–388, 1999.
- [6] L. Zhang, W. Zhou, V. E. Velculescu, et al., “Gene expression profiles in normal and cancer cells,” *Science*, vol. 276, no. 5316, pp. 1268–1272, 1997.
- [7] S. A. Jelinsky and L. D. Samson, “Global response of *saccharomyces cerevisiae* to an alkylating agent,” *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 4, pp. 1486–1491, 1999.
- [8] F. C. Holstege, E. G. Jennings, J. J. Wyrick, et al., “Dissecting the regulatory circuitry of a eukaryotic genome,” *Cell*, vol. 95, no. 5, pp. 717–728, 1998.

- [9] S. A. Jelinsky, P. Estep, G. M. Church, and L. D. Samson, "Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes," *Molecular and Cellular Biology*, vol. 20, no. 21, pp. 8157–8167, 2000.
- [10] V. A. Kuznetsov, "Analysis of stochastic processes of gene-expression in a single cell," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, University of Delaware, Baltimore, MD, USA, June 2001.
- [11] J. J. Ramsden and J. Vohradsky, "Zipf-like behavior in prokaryotic protein expression," *Phys. Rev. E.*, vol. 58, no. 6, pp. 7777–7780, 1998.
- [12] M. Y. Borodovsky and S. M. Gusein-Zade, "A general rule for ranged series of codon frequencies in different genomes," *J. Biomol. Struct. Dynam.*, vol. 6, no. 5, pp. 1001–1012, 1989.
- [13] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, et al., "Linguistic features of non-coding DNA sequences," *Phys. Rev. Lett.*, vol. 73, no. 23, pp. 3169–3172, 1994.
- [14] W. Li, "Statistical properties of open reading frames in complete genome sequences," *Computers and Chemistry*, vol. 23, no. 3–4, pp. 283–301, 1999.
- [15] C. Adami, *Introduction to Artificial Life*, Springer-Verlag, New York, 1998.
- [16] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, and M. Simons, "Scaling features of noncoding DNA," *Physica A*, vol. 273, no. 1–2, pp. 1–18, 1999.
- [17] M. S. H. Ko, "Induction mechanism of a single gene molecule: stochastic or deterministic?," *Bioessays*, vol. 14, no. 5, pp. 341–346, 1992.
- [18] I. L. Ross, C. M. Browne, and D. A. Hume, "Transcription of individual genes in eukaryotic cells occurs randomly and infrequently," *Immunol. Cell Biol.*, vol. 72, no. 2, pp. 177–185, 1994.
- [19] H. H. McAdams and A. Arkin, "It's a noisy business! genetic regulation at the nanomolar scale," *Trends in Genetics*, vol. 15, no. 2, pp. 65–69, 1999.
- [20] D. A. Hume, "Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression," *Blood*, vol. 96, no. 7, pp. 2323–2328, 2000.
- [21] M. C. Walters, S. Fiering, J. Eidemiller, W. Magis, M. Groudine, and D. I. K. Martin, "Enhancers increase the probability but not the level of gene expression," *Proc. Natl. Acad. Sci. USA*, vol. 92, no. 15, pp. 7125–7129, 1995.
- [22] S. Newlands, L. K. Levitt, C. S. Robinson, et al., "Transcription occurs in pulses in muscle fibers," *Genes and Dev.*, vol. 12, no. 17, pp. 2748–2758, 1998.
- [23] J. Chelly, J. P. Concordet, J. C. Kaplan, and A. Kahn, "Illegitimate transcription: transcription of any gene in any cell type," *Proc. Natl. Acad. Sci. USA*, vol. 86, no. 8, pp. 2617–2621, 1989.
- [24] L. Wodicka, H. Dong, M. Mittmann, M. H. Ho, and D. J. Lockhart, "Genome-wide expression monitoring in saccharomyces cerevisiae," *Nature Biotechnol.*, vol. 15, no. 13, pp. 1359–1367, 1997.
- [25] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 1999.
- [26] A. E. Lash, C. M. Tolstoshev, L. Wagner, et al., "SAGEmap: a public gene expression resource," *Genome Res.*, vol. 10, no. 7, pp. 1051–1060, 2000.
- [27] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions*, John Wiley & Sons, New York, 2nd edition, 1992.
- [28] V. A. Kuznetsov, "The genes number game in growing sample," *J. Comput. Biol.*, vol. 7, no. 3/4, pp. 642, 2000.
- [29] J.-J. Chen, J. D. Rowley, and S. M. Wang, "Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 1, pp. 349–353, 2000.
- [30] M. Johnson, "The yeast genome: on the road to the gold age," *Current Opinion in Genetics and Development*, vol. 10, no. 6, pp. 617–623, 2000.
- [31] C. R. Cantor and C. L. Smith, *Genomics*, John Wiley & Sons, New York, 1999.
- [32] The Numerical Algorithms Group Limited, "NAG Fortran Library Mark 19 Introductory Guide," Mark 19, July 1999.
- [33] H. Lockhart, D. J. Dong, M. C. Byrne, et al., "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology*, vol. 14, no. 13, pp. 1675–1680, 1996.

Vladimir A. Kuznetsov received the M.S. degree in physics from the Kyrgyz State University, Frunze, USSR, in 1971. From 1977 to 1980 he was a postgraduate student at the Institute of Molecular Biology, USSR Acad. of Sci. (Moscow). He received the Ph.D. degree in physics and mathematics from Moscow University in 1984, and the Sci.D. degree in physics and mathematics from the Science & Technical Union of the USSR Acad. of Sci. in 1992. From 1972 to 1981 he was a researcher at the Research Institute of Oncology and Radiology (Frunze, USSR) and a lecturer at the Mathematical Department of the Kyrgyz State University. From 1981 to 1995 he was a researcher and then a chief of the Laboratory of Mathematical Immunobiophysics at the Institute of Chemical Physics, Russian Acad. of Sci., Moscow and from 1995 he was chief of the same laboratory in the new Institute of Biochemical Physics, Russian Acad. Sci. From 1995–1997 Dr Kuznetsov received research grants from American Cancer Society (International Cancer Research Fellowship; annual grant) and from National Cancer Institute, NIH, USA. He worked at the Laboratory of Molecular Tumor Biology, CBER/FDA, Bethesda, MD, USA, and the Laboratory of the Experimental and Computational Biology, NCI/NIH, where he was engaged in a research of cytokine control of tumor growth and in clinical trails of pediatric patients with HIV-1. From 1998–1999, he was a chief scientist at Civilized Software, Inc., Bethesda, MD, USA. In 1999 he received visiting fellowship research grant (for 6 months) from Engineering and Physical Sciences Research Council, UK. Currently, he is a senior research fellow at the Laboratory of Integrative and Medical Biophysics, NICHD/NIH. His interests include mathematical and computational biology, bioinformatics, statistics, computer science, biophysics, cellular biology, immunology and genetics. Dr Kuznetsov was awarded by P. L. Kapitsa's silver medal "To the author of scientific discovery"; he is a member of the Russian Academy of Natural Sciences.

