BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 16 2009, pages 2020–2027
doi:10.1093/bioinformatics/btp360

*Gene expression*

# An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems

Chuan Lu and Ross D. King*

Department of Computer Science, Aberystwyth University, Ceredigion SY23 3DB, UK

Received on February 25, 2009; revised on May 20, 2009; accepted on June 5, 2009

Advance Access publication June 17, 2009

Associate Editor: Ivo Hofacker

**ABSTRACT**

**Motivation:** Distribution analysis is one of the most basic forms of statistical analysis. Thanks to improved analytical methods, accurate and extensive quantitative measurements can now be made of the mRNA, protein and metabolite from biological systems. Here, we report a large-scale analysis of the population abundance distributions of the transcriptomes, proteomes and metabolomes from varied biological systems.

**Results:** We compared the observed empirical distributions with a number of distributions: power law, lognormal, loglogistic, loggamma, right Pareto-lognormal (PLN) and double PLN (dPLN). The best-fit for mRNA, protein and metabolite population abundance distributions was found to be the dPLN. This distribution behaves like a lognormal distribution around the centre, and like a power law distribution in the tails. To better understand the cause of this observed distribution, we explored a simple stochastic model based on geometric Brownian motion. The distribution indicates that multiplicative effects are causally dominant in biological systems. We speculate that these effects arise from chemical reactions: the central-limit theorem then explains the central lognormal, and a number of possible mechanisms could explain the long tails: positive feedback, network topology, etc. Many of the components in the central lognormal parts of the empirical distributions are unidentified and/or have unknown function. This indicates that much more biology awaits discovery.

**Contact:** rdk@aber.ac.uk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

A central goal of science is to find patterns in nature. Improved analytical methods now mean that extensive and accurate quantitative measurements can be made of the key classes of dynamic intercellular molecules: mRNAs (Lockhart and Winzeler, 2000; Schulze and Downward, 2001), proteins (Cravatt *et al.*, 2007; Tyers and Mann, 2003) and metabolites (Fiehn, 2002; Kell, 2004). The question that we address in this article is whether, despite the extreme complexity of the specific interactions involved, there is any pattern in the observed *quantities* of these molecules in living systems. The existence of any such pattern would provide valuable insight into living systems, provide constraints on Systems Biology models and aid the parametrization of analysis methods.

One of the the most basic ways of describing complicated systems that involve many components is to use population abundance distributions. These describe the relationship between the abundance of components to the number of components. This type of distribution is perhaps most commonly used in ecology, with the components being species, and with the distribution summarizing how many common and rare species are there. In this article, we investigate the observed population abundance distributions of mRNAs, proteins and metabolites.

There are several reasons which make the existence of pattern(s) in the observed population abundance distributions plausible.

(1) All living systems are both evolutionarily related and homeostatic. This implies that some structure is preserved both between species and during different growth states.

(2) It has been empirically found that there are general patterns in the distribution of connections between components in living systems (Arita, 2005; Barabási and Albert, 1999; Jeong *et al.*, 2001; Tong *et al.*, 2004).

(3) Initial work has been reported on the distribution of the observed population abundances of mRNA species. Ueda *et al.* (2004) and Kuznetsov *et al.* (2002) reported power law (or its variant) distributions, while Konishi (2004) reported a three-parameter lognormal distribution, and Hoyle *et al.* (2002) reported a mixing behaviour of central lognormal with a power law tail. What is significant about these reports is that similar distributions were observed across a wide variety of species and technologies.

### 1.1 Omics measurement technology

An ideal analytical measurement experiment would measure with complete fidelity the quantity of every component in the cell.

Current 'omic' experiments are still far from approaching this ideal due to limitations in: measurement technologies, experimental techniques, preprocessing procedures, etc. In addition, what is observed and modelled are only samples of the underlying distributions.

In discussing omics experiments, we will use the statistical measures of: coverage, bias and accuracy. Arguably the most important of these is coverage, by this we mean what proportion of the total number of components of a system are actually observed. This property applies to both qualitative and quantitative

*To whom correspondence should be addressed.

 

measurement technologies. If the coverage of an omics technology is not complete then the concept of measurement bias is important. By an unbiased measurement technology, we mean one where the measurement accuracy is independent of the attributes of the components (e.g. quantity, size, hydrophobicity, etc.). The most important attribute we are concerned with is quantity. Basic information theory suggests that it is much easier to determine the presence of a molecular species present in large quantities than in small quantities. Thus, most omic measurement technologies are biased, and errors are most probable for low-abundance components. High-abundance components may also be wrongly measured due to machine saturation. As we are concerned with quantitative measurement of cellular components, it is desirable that a measuring technology should make accurate measurements. The concept of measurement bias is also applicable to quantitative measurements.

One important issue relevant to the experimental strategies is most current omic experiments only measure the average value of gene, protein or metabolite expression levels for a population of cells, rather than that of individual cells. This means that, for example, intermediate gene expression levels could result from a mixture of individual cells with high or low states.

The most successful omic technology is probably transcriptomics. A vast amount of biological knowledge has been generated using these techniques (Bertone *et al.*, 2004; Lockhart and Winzeler, 2000; Schulze and Downward, 2001). It is also the omics technology that approaches the ideal most closely. Until recently microarrays have been the dominant transcriptomics technology, but this may be about to change with lower costs for sequencing (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). Given a known genome, it is now reasonable to expect that the coverage of mRNA components is high. However, it is noteworthy, that tiling arrays and direct sequencing have made it clear that many more DNA sequences are transcribed in small quantities than previously expected (Wilhelm *et al.*, 2008), and that standard microarrays do not have a 100% coverage. A recent large-scale comparison of microarrays with other techniques has shown that microarrays have good accuracy (Canales *et al.*, 2006). However, the same study also highlighted that their accuracy was much poorer with low-copy number mRNA transcripts, i.e. they are biased towards high-abundance components. This means that any conclusions that can be made about the distribution of low-abundance mRNAs are less secure than those of high-abundance ones.

A vast amount of research has also gone into proteomic research (Ishihama *et al.*, 2008; Tyers and Mann, 2003). Due to the greater heterogeneity of proteins and their post-processing, the state-of-the-art in proteomics is less advanced than that of transcriptomics (Bantscheff and Schirle, 2007; Cravatt *et al.*, 2007). Within proteomics the 'naming of the parts' is still far from complete even for the model organisms; and it is generally not possible to measure the amount (or even the presence) of all the proteins in a moderately complex biological system. However, there are now datasets available that report on the observation of thousands of proteins, and proteomic technology has greatly improved recently (Ishihama *et al.*, 2008; Newman *et al.*, 2006). The quantification of proteomic observations is also more difficult than that for transcriptomics. Proxies for protein amount may be required, such as ion intensities, the number of peptides observed, the amount of luminosity, etc. (Bantscheff and Schirle, 2007). As with transcriptomics uncertainties can be mitigated against by use of multiple technologies, e.g. two-dimensional gel electrophoresis (2DE) or liquid chromatography (LC) and mass spectrometry (MS).

Metabolomics research is also a large and growing research field (Fiehn, 2002; Kell, 2004; Wishart *et al.*, 2007). Metabolomics faces similar difficulties to proteomics. Metabolites are even more heterogeneous than proteins, which means that different approaches may need to be used for different classes of metabolites, causing bias. However, metabolomics does have the advantages of there being typically far fewer metabolites than proteins (order of 1000 in simple microorganisms), and that many metabolites are shared between species, enabling standard techniques. It is hard to estimate the coverage of metabolomics techniques. For simpler systems this may be high, but for more complicated organisms, especially plants, this coverage is much lower.

Many methods have been developed to preprocess omics data so as to reduce the data limitations resulting from the utilized techniques. For example, there is a lot of research on development of models and algorithms for microarrays that exploit hybridization theory in order to correct the data from noise, non-specified hybridization and saturation effect (Chua *et al.*, 2006; Koltai and Weingarten-Baror, 2008; Marcelino *et al.*, 2006; Wu and Irizarry, 2005). These models for measurement data preprocessing are very important for analysing the distributions for cellular component abundances, but are not the focus of this study. Therefore in this work, we have selected particular datasets that have already been preprocessed to represent relative abundances; and it is assumed that the noise or saturation effect in measurement have been corrected to a certain extent. In our analysis, we were aware of the potential bias resulting from the preprocessing steps in the original study, e.g. removal of non-significant low-abundant components, or imputation of the missing values.

In conclusion, although existing omic technologies are far from ideal, they are now capable of generating high-quality quantitative data on population abundance distributions, and there appears to be little reason to believe that conclusions drawn from existing data are unreliable—except perhaps at the low end of the dynamic range.

## 1.2 The power law and lognormal distributions

The most commonly reported distribution in cellular systems are power law (scale-free) ones (Arita, 2005). Such distributions of node connections have been reported in metabolic networks (Jeong *et al.*, 2000), protein–protein interaction (Jeong *et al.*, 2001), gene interactions (Tong *et al.*, 2004), etc. Although, these findings have been challenged on both empirical (Khanin and Wit, 2006; Stumpf *et al.*, 2005b) and theoretical grounds (Stumpf *et al.*, 2005a). Lognormal distributions have also been reported for node connection distributions, and are ubiquitous (Limpert *et al.*, 2001).

A non-negative random variable $X$ has a power law distribution if its probability density function can be expressed as $f(x)=Cx^{-\gamma}$. A common power law distribution is the Pareto distribution which satisfies $P(X\geq x)=(\frac{x}{k})^{-(\gamma-1)}$ for some $\gamma>1$, $k>0$ and $X\geq k$. Two general classes of argument have been used to explain power laws in biology: it is a result of how the system developed over time (Mitzenmacher, 2004); or it has been selected through natural selection for robustness (Jeong *et al.*, 2000). The most common proposed models for power laws are based around preferential

attachment mechanisms ('the rich get richer'). The first proposed generative mechanism for a power law in biology was that of (Yule, 1925), who used it to explain the distribution of species in genera.

A random variable has a lognormal distribution if the random variable $Y=\log(X)$ has a normal distribution. Lognormal distributions are generally generated by proportionate effect processes. More generally, the central-limit theorem states that the sum of many independent, identically distributed random variables with a finite mean and finite variance converges to a normal distribution asymptotically. Similarly, the product of many positive random variables will approach a lognormal distribution (Limpert *et al.*, 2001). Sinnott (1937) was the first to provide evidence for the importance of the lognormal distribution in genetics in the phenotypes of plants.

The lognormal and power law distributions are closely related: they are both skewed, can be generated by similar multiplicative mechanisms, and may be hard to distinguish empirically. Power law distributions may arise from lognormal distributions from small changes to the generative model: where the sampling time is not uniform, or when a lower boundary is put into effect during a geometric random walk (Mitzenmacher, 2004).

### 1.3 The Pareto-lognormal distribution

Interestingly, there are distributional forms that behave like a lognormal distributions near the centre and like power law distributions in the tails. The double Pareto-lognormal (dPLN) distribution, first introduced by Reed (Reed, 2003; Reed and Jorgensen, 2004), has this property, and has been shown to fit some empirical data better than either the power law or lognormal distributions.

A random variable $X$ follows a dPLN distribution if $\log(X)$ follows the normal-Laplace distribution, which can be represented as a convolution of independent normal and Laplace components. Taking the exponential form of a normal-Laplace random variable results in the dPLN distribution, which can be represented as $X=UQ$ where $U,Q$ are independent, with $U$ lognormally distributed $\log(U)\sim N(\nu,\tau^2)$ and $Q\sim DP(\alpha,\beta)$ following the double Pareto distributions. The density function of the double Pareto distribution can be expressed as

$$f(q)=\begin{cases}\frac{\alpha\beta}{\alpha+\beta}q^{\beta-1}, & \text{for } 0<q\leq 1;\\ \frac{\alpha\beta}{\alpha+\beta}q^{-\alpha-1}, & \text{for } q>1.\end{cases} \quad (1)$$

where $\alpha>0$, $\beta>0$.

The parameters of the dPLN distribution can be used to describe the features of the corresponding probability density function (Reed and Jorgensen, 2004). The parameter $\tau$ is the SD for the lognormal component, as $\tau\to 0$, the distribution tends to a double Pareto distribution. The parameter $\alpha$ and $\beta$ determine the behaviour in the left and right tails, respectively: the smaller the value is, the heavier the corresponding tail is. When both $\alpha,\beta\to\infty$, the distribution tends to be a lognormal distribution; if only $\beta\to\infty$, the distribution has a fatter tail only in the right; if only $\alpha\to\infty$, the distribution has a fatter tail only in the lower end. When $\alpha=\beta$, it is symmetric and bell-shaped, if plotted in a log scale. Therefore, the lognormal, right/left PLN and double Pareto distributionscan be considered as special cases for the dPLN model, which can be regarded as a form of nested model.

An interesting feature of the dPLN distribution is that it can be derived from a simple stochastic model (Reed, 2003). Consider a geometric Brownian motion defined by the Ito stochastic differential equation

$$\mathrm{d}x=\mu X\mathrm{d}t+\sigma X\mathrm{d}w \quad (2)$$

with initial state $X(0)=X_0$ distributed lognormally, $\log(X_0)\sim N(\nu,\tau^2)$. After $T$ time units the state $X(T)$ can be expressed as

$$X(T)=X_0\exp\left(\left(\mu-\frac{\sigma^2}{2}\right)T+\sigma\varepsilon\sqrt{T}\right) \quad (3)$$

where $\varepsilon\sim N(0,1)$ is a standard normal distributed random variable. The distribution for $X(T)$ is then also lognormal and can be expressed as

$$\log X(T)\sim N\left(\nu+\left(\mu-\frac{\sigma^2}{2}\right)T,\tau^2+\sigma^2T\right). \quad (4)$$

Suppose the time $T$ at which the process is ended is an exponentially distributed random variable with density $f_T(t)=\lambda e^{-\lambda t}, t>0$. The distribution of the state $X(T)$ is then a mixture of lognormal random variable (4) with mixing parameter $T$, and this can be proved to be the dPLN distribution (Reed and Jorgensen, 2004).

## 2 APPROACH

In this work, we have extended the investigation of the population abundance distribution for transcriptomic data to other omic datasets using stringent statistical tests. To do this, 22 preprocessed omics datasets have been selected and fitted to a series of highly skewed distribution models. The fitted models have been compared using various statistical criteria such as Bayesian information criteria (BIC). Hypotheses for these models have been further tested using parametric bootstrap goodness-of-fit tests.

The best-fit for these cellular component population abundance distributions was found to be the dPLN. The geometric Brownian motion (GBM) models that generate the fitted dPLN distributions were then explored to seek links of the abundance distribution to the underlying mechanisms in biology. Furthermore, we conducted a functional analysis for the cellular components, in particular to compare the functional categories for the components found in the high-abundance tail with those in the lognormal mode.

For our analysis, we selected a set of datasets from quantitative transcriptomic, proteomic and metabolomic experiments. These datasets were selected for their high quality (high coverage and accuracy, and low bias), and for diversity of organism and technology. They include 10 transcriptomic datasets, 7 proteomic datasets and 5 metabolomic datasets (as summarized in supplementary Table S1). An ID was given to each dataset, starting with 'mRNA-', 'prot-' or 'metab-' depending on the omics technology used, followed by the name of the organism. To distinguish multiple datasets from the same organism, extra words were added to the end of the ID providing information such as

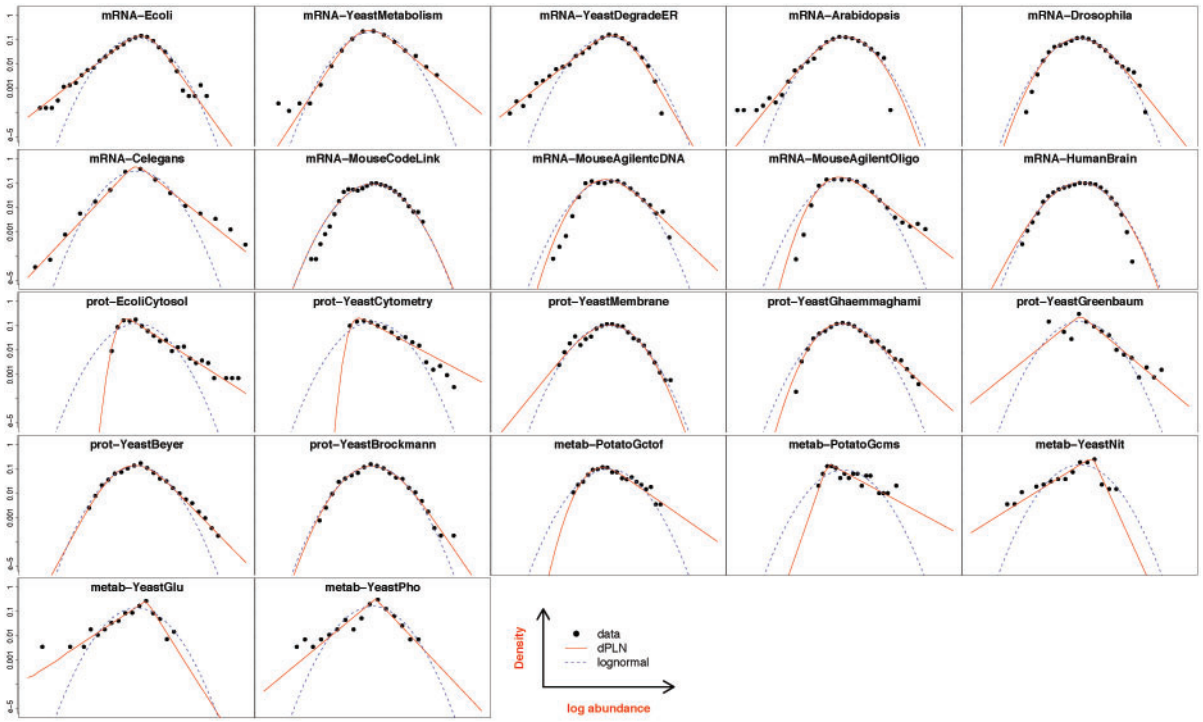


**Fig. 1.** Distributions (log–log histograms) for the omic datasets. The title in each panel is the corresponding dataset ID. The *y*-axis corresponds to the normalized density values and the *x*-axis to the abundance level in log scale. Solid circles display log–log histogram of the abundance levels for the empirical data. The solid line and dashed line represent the fitted density function for dPLN and lognormal distributions, respectively.

measurement techniques, original experimental design or references. The datasets were checked and preprocessed before distributional analysis. Detailed explanation of the dataset preparation is also given in Supplementary Material.

## 3 STATISTICAL DISTRIBUTION TESTS

To estimate the population abundance distributions, we plotted log–log histograms (Fig. 1). These were generated using logarithmic binning for either individual samples or the median samples. Superimposed on the data points are the fitted density functions for lognormal and dPLN distributions. Examination of these indicates that: the data do not follow a power law distribution, and that a lognormal distribution fits the central part of the distribution with a power law distributions in most of the tail(s). The dPLN model seems to fit reasonably well to these distributions of different shapes: some exhibiting heavy tails in both sides (e.g. mRNA-Ecoli, mRNA-Celegans and metab-YeastPho), some not (e.g. mRNA-MouseCodeLink and mRNA-HumanBrain seem to fit well with a lognormal distribution).

To confirm this analysis we carried out goodness-of-fit tests using a set of distributions: Pareto, lognormal, loggamma, loglogistic, right Pareto-lognormal and dPLN (see Supplementary Table S2 for the relevant functional forms and parameters). The composite parametric bootstrap goodness-of-fit tests for all six distributions were done using R scripts. Our distributional goodness-of-fit tests consisted of two basic steps. In the first step, the parameters of the distributions were estimated using maximum likelihood. Then goodness-of-fit tests were conducted to compare the probability models with the experimental data. We applied both the Kolmogorov–Smirnov (K-S) and Anderson–Darling (A-D) tests (Stephens, 1974). The K-S test is based on a statistic that measures the deviation of the observed empirical cumulative distribution function (i.e. the cumulative histogram) from the hypothesized cumulative distribution function (Stephens, 1974).

The A-D test is a modified version of the K-S test, which uses a slightly different statistic defined by $D_{\mathrm{AD}} = \max_i \frac{|\hat{F}(x_{(i)}) - F(x_{(i)})|}{F(x_{(i)})(1-F(x_{(i)}))}$, where $x_{(i)}$ is the $i$-th order statistic of the sample, $\hat{F}(x_{(i)})$ and $F(x_{(i)})$ are the empirical and theoretical cumulative distribution functions, respectively. The A-D test is more sensitive to the difference between the theoretical distribution and the hypothesized distribution at the tails than the K-S test.

Table 1 reports the statistics $D_{\mathrm{AD}}$ of the A-D tests and the $\mathcal{P}$-values estimated from the parametric bootstrap A-D tests (with 1000 bootstrap replicates). The results of parametric bootstrap K-S tests are given in Supplementary Table S3.

The relatively small statistics $D_{\mathrm{AD}}$ or $D_{\mathrm{KS}}$ for the dPLN distribution provide evidence that the dPLN consistently gives the best fit to the abundance data in most of the datasets. It is the only distribution that is not rejected (at significance level of 0.05) by the parametric bootstrap A-D or K-S tests for all examined datasets.

Notice that the number of parameters for the different distribution models vary: two for the power law, lognormal and loglogistic, three for the loggamma and right PLN and four for the dPLN distribution. Therefore, besides model likelihood, $D_{\mathrm{AD}}$ and $D_{\mathrm{KS}}$, information criteria such as the Akaike information criterion (AIC) and BIC have also been calculated to compare the models by taking both the model

**Table 1.** Statistical distribution tests for the omic data

| Distribution | Power law | | Lognormal | | Loglogistic | | Loggamma | | dPLN | | RightPLN | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $D_{AD}$ | $P$-value | $D_{AD}$ | $P$-value | $D_{AD}$ | $P$-value | $D_{AD}$ | $P$-value | $D_{AD}$ | $P$-value | $D_{AD}$ | $P$-value |
| Transcriptomics | | | | | | | | | | | | |
| mRNA-Ecoli | Inf | NA | 27.52 | 0 | *11.11* | 0 | 37.37 | 0 | **0.39** | 0.782 | 28.94 | 0.247 |
| mRNA-YeastMetabolism | Inf | NA | 25.29 | 0 | 13.09 | 0 | 12.73 | 0 | **1.08** | 0.557 | *3.12* | 0.216 |
| mRNA-YeastDegradeER | Inf | NA | 54.20 | 0 | *17.58* | 0 | 75.92 | 0 | **2.71** | 0.343 | 54.14 | 0.869 |
| mRNA-Arabidopsis | Inf | NA | **6.71** | 0 | 8.87 | 0 | 10.02 | 0.001 | *8.33* | 0.525 | 7.01 | 0.784 |
| mRNA-Drosophila | Inf | NA | 11.52 | 0 | 11.92 | 0 | 14.39 | 0 | **10.58** | 0.628 | *10.79* | 0.070 |
| mRNA-Celegans | Inf | NA | 264.09 | 0 | *60.06* | 0 | 225.61 | 0 | **5.73** | 0.047 | 130.51 | 0 |
| mRNA-MouseCodeLink | NA | NA | 16.29 | 0 | 28.50 | 0 | 22.92 | 0 | *16.08* | 0.520 | **15.92** | 0.349 |
| mRNA-MouseAgilentcDNA | Inf | NA | 44.50 | 0 | 54.77 | 0 | **30.34** | 0 | 40.80 | 0.313 | *40.79* | 0 |
| mRNA-MouseAgilentOligo | Inf | NA | 98.00 | 0 | 94.86 | 0 | **23.05** | 0 | *51.47* | 0.958 | 51.50 | 0 |
| mRNA-HumanBrain | NA | NA | *20.22* | 0 | 35.80 | 0 | 28.47 | 0 | **17.77** | 0.992 | 20.55 | 0.680 |
| Proteomics | | | | | | | | | | | | |
| prot-EcoliCytosol | 78.02 | 0 | 40.67 | 0 | 20.93 | 0 | 4.003 | 0 | 1.159 | 0.525 | 1.159 | 0.291 |
| prot-YeastCytometry | Inf | NA | 48.06 | 0 | 35.09 | 0 | 3.28 | 0 | **11.06** | 0.153 | *11.75* | 0 |
| prot-YeastMembrane | 251.3 | 0 | 3.818 | 0 | *2.200* | 0 | 7.051 | 0 | **2.046** | 0.617 | 3.897 | 0.366 |
| prot-YeastGhaemmaghami | NA | NA | 12.00 | 0 | 5.187 | 0 | 4.114 | 0 | *2.098* | 0.678 | **2.060** | 0.301 |
| prot-YeastGreenbaum | Inf | NA | 39.76 | 0 | *33.96* | 0 | 45.98 | 0 | **26.36** | 0.014 | 44.23 | 0.001 |
| prot-YeastBeyer | NA | NA | 10.15 | 0 | **3.543** | 0 | 9.083 | 0 | *5.566* | 0.378 | 6.695 | 0.094 |
| prot-YeastBrockmann | Inf | NA | 5.5214 | 0 | *2.965* | 0 | 5.790 | 0 | **4.589** | 0.530 | 4.685 | 0.324 |
| Metabolomics | | | | | | | | | | | | |
| metab-PotatoGctof | 27.91 | 0 | 2.28 | 0 | 1.83 | 0 | **0.36** | 0.316 | 0.41 | 0.846 | *0.40* | 0.799 |
| metab-PotatoGcms | 2.44 | 0.008 | 2.08 | 0 | 1.86 | 0 | 0.40 | 0.399 | *0.55* | 0.728 | **0.50** | 0.618 |
| metab-YeastNit | Inf | NA | 8.685 | 0 | *5.531* | 0 | 9.902 | 0 | **0.6785** | 0.5766 | 8.989 | 0.204 |
| metab-YeastGlu | Inf | NA | 6.683 | 0 | *3.894* | 0 | 7.940 | 0 | **0.2331** | 0.938 | 6.874 | 0.263 |
| metab-YeastPho | Inf | NA | 7.986 | 0 | *4.019* | 0 | 9.831 | 0 | **1.995** | 0.374 | 8.369 | 0.213 |

The A-D goodness-of-fit tests were conducted on the sample or the median sample (in case of multiple samples) of each dataset. The $\mathcal{P}$-values were obtained via parametric bootstrap method (with 1000 replicates). The statistics $D_{AD}$ and $\mathcal{P}$-values are reported.
The smallest statistics $D_{AD}$ for each dataset have been highlighted in bold, and the second smallest in italic.

fitness and the model complexity into account (see Supplementary Table S4).

The preferred model is that with the lowest AIC/BIC value. As number of data points for each sample is large in this case of omics data, BIC is more appropriate than AIC for model comparison. Table S4 indicate that even when the extra parameters are taken into account the dPLN model still fits the empirical data the best for most of the cases.

Table S6 reports the parameter estimates, bias and standard errors from the bootstrap tests for power law, dPLN/right PLN distribution.

*The dPLN variance components*: for random variable $X \sim$ dPLN$(\nu, \tau^2, \alpha, \beta)$, the variance of $\log(X)$ can be decomposed into three parts: $\sigma^2_{\text{LN}} + \sigma^2_{\text{rP}} + \sigma^2_{\text{lP}} = \tau^2 + (1/\alpha)^2 + (1/\beta)^2$, corresponding to the variance from the central lognormal component, the right-sided and left-sized Pareto component, respectively (Reed and Jorgensen, 2004).

To check the dependancy of the parameters on the omics technologies, the variances estimated using dPLN models have been compared for yeast (see Supplementary Fig. S1). The average $\sigma^2_{\text{LN}}$ and $\sigma^2_{\text{rP}}$ are smaller for transcriptomics and metabolomics data than for proteomics data, although the differences among the three types of omics technoloiges are not really significant according to the ANOVA tests. Note that only two yeast datasets were used here for transcriptomics and three for metabolomics; also there are large variances in the parameters for proteomics data, which seem to originate in the diversity in analysis techniques and data preprocess/integration methods. In the case of *Escherichia coli* $\sigma^2_{\text{LN}}$ and $\sigma^2_{\text{rP}}$ are also larger for proteomics than for transcriptomics. This implies that the omics techniques somehow play a role in the dPLN variance components.

The relationship between the variance components and the number of genes in genome for transcriptomics data has also been checked (see Supplementary Fig. S2). A weak positive correlation ($R^2 = 0.36$) was observed between $\sigma^2_{\text{LN}}$ and the number of genes in genome. This correlation seems to be consistent with, although not as strong as, the one for the variance of log data in Hoyle *et al.* (2002). Moreover, the total variance from the Pareto components in both tails $\sigma^2_{\text{rP}} + \sigma^2_{\text{lP}}$ was observed to be negatively correlated ($R^2 = 0.42$) with the number genes in the genome. One could expect that these correlations would be much stronger if more datasets were used.

## 4 DYNAMIC MODELS LINKING TO ABUNDANCE POPULATION DISTRIBUTIONS

As noted above, the PLN distribution can be generated by a simple stochastic dynamic model: GBM. To explore the biological relevance of this generative approach, and to investigate the

evolution of the dPLN distribution, GBM model simulation was performed to generate random samples which followed the fitted distribution. Two examples of the data simulation with parameters estimated from the median sample are given in Figure S3.

Our proposed GBM model can be viewed as a simplified and abstract description of cellular dynamics. It ignores many details of biochemical networks, such as regulation factors, and the distribution of the average evolving time for various cellular components is unlikely to be as simple as an exponential distribution. On the other hand, as what we are interested in is the population distribution of the general average abundance for the cell, such fluctuations and variations may be cancelled out by averaging, and also might be partially explained by the volatility of the GBM model. For example, positive feedback or autocatalysis in general increases the sensitivity to internal or external signals and hence increase the variance of the fluctuation; and negative feedback or product inhibition dampens the noise and rejects perturbations (Hornung and Barkai, 2008; Stelling *et al.*, 2004).

Similar dynamic models have been applied to explain the pattern in the population abundance in living systems in varied research areas. We review some of the relevant work here and describe the links between these models. Ochiai *et al.* (2004) proposed a constructive approach to a probability model for gene dynamics from a gene expression instantaneous transition probability (ITP) model for individual genes by assuming the gene expression level is described by a stochastic process with Markov property. And the ITP model used was obtained experimentally and fitted well to the yeast ITP data as shown in Ueda *et al.* (2004). This leads to emergence of the Black–Scholes model (originally from economics), which is exactly the same stochastic model as that for GBM given in Equation (2). From this, by using the same assumptions as described in the Section 1, including lognormalilty of the initial distribution of the component-level abundances, and an exponential distributed life time for the components, a dPLN distribution is derived for the population abundance levels—if the evolving time is long enough.

Friedman *et al.* (2006) have also presented a theoretical model to reconcile the time resolved and population measurements for protein concentration, considering that protein production occurs in random bursts with an exponentially distributed number of molecules. The proposed analytical framework of gene expression links stochastic dynamics to population distributions. Starting from the simple kinetic scheme for protein production characterized by two parameters: the mean number of bursts per cell cycle, and the mean number of protein molecules produced per burst, they derived a gamma distribution in the steady state via approximation for the same protein, which fits well to their experimental data.

Paulsson (2004, 2005) provides a review of a generic approach utilizing fluctuation-dissipation theorem (FDT) to stochastic modelling of the gene expression, which can be applied to model concentration levels of metabolites as well (Elf *et al.*, 2003).

Note that the population distributions described by Ochiai *et al.* Friedman *et al.* and Paulsson are ones of a particular chemical species from the population of genetically identical cells, rather than the population distribution for various cellular components within a system which we have analysed in this work. However, it is still interesting that similar theoretical framework can be used to link the cellular dynamics to population distributions, but with different levels of abstractions and different assumptions.

There have also been some other attempts to understand the mechanisms explaining the population distribution for general average over all chemical species within an intracellular network (Furusawa and Kaneko, 2003; Tokita, 2006). These models may explain some phenomena in nature, but they are inconsistent with our empirical observations of omic data.

## 5 FUNCTIONAL ANALYSIS OF THE CELLULAR CONSTITUENTS

To gain insight into the biological meaning of the observed distributions, we compared the molecular species found in the high-abundance tail with those in the lognormal mode. The most abundant 5% of mRNA metabolites/proteins were taken, along with the 10% close to the median (adjacent to the mode) and the functions of these two subsets were compared.

For the mRNA abundance data, their Gene Ontology (The GO consortium, 2000) annotations were used to obtain the annotation classes, which were then mapped to a set (128) of generic GO slim terms (Fig. 2). Some generic terms that are not commonly shared by all the organisms have been excluded from this analysis. Only the annotated gene products were used for further comparison. GO annotations were obtained for the gene products in eight transcriptomic datasets (the two mouse datasets measured with Agilent techniques were excluded) as were the genes from the yeast cytometry proteomic dataset.

Functional differences between the two component subsets (the right tail and the mode) were checked using $\chi^2$-tests for the 128 GO classes and for eight individual datasets. To summarize the comparison of the functional classes over different transcriptomic datasets, the variance-based weighted odds ratios across different datasets were computed (Kenneth and Sander, 1998). The significantly different GO classes, which have an odds ratio of >1.5 or <0.7, were then plotted.

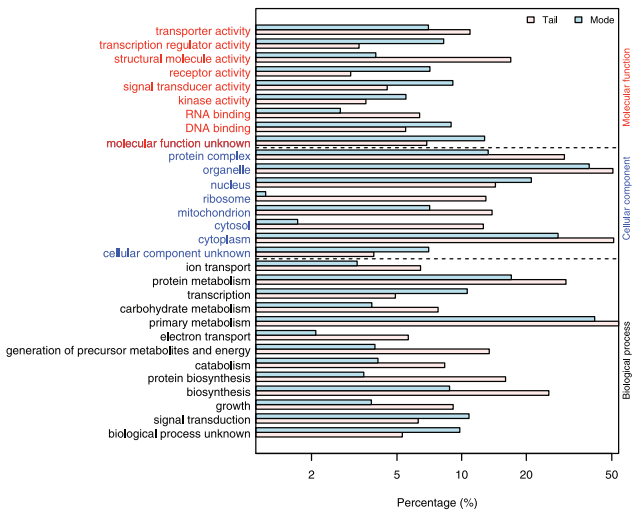


**Fig. 2.** Comparison of GO functional classes for the gene products with the highest expression levels and with modal expression levels across the transcriptomic datasets. The two coloured bars correspond to the percentages of genes that are associated to a particular GO class at the high-end tail and at the lognormal mode (approximately with abundance level of median range), respectively.

The summary results of the analysis for the transcriptomics data using GO are shown in Figure 2. All these class differences are statistically significant and present in a wide range of microarray experiments. The over-represented high-abundance classes are generally the core cellular components: structural components, primary metabolism, ribosome associated, electron transport, etc. The classes that are over-represented in the modal region are: signal transduction, kinases, DNA binding, protein modification, etc. These are the cell's control elements. *These results indicate that cells require a relatively large numbers of different types of low-abundance control elements, and fewer types of high-abundance core elements.*

It is interesting to note that the number of genes related to transcription regulation grows much faster with genome size than the ones for protein biosynthesis (Van Nimwegen, 2003). More complicated biological systems seem to require larger amount of genes to participate in the increasingly elaborate regulatory mechanism, while keep them relatively low abundant. This is consistent with our results.

The results of the analysis for the yeast cytometry proteomics data using GO is shown in Figure S4. One feature consistent between mRNAs, proteins, and metabolites is that the components in the mode are generally less well characterized (many unknown metabolites and gene products of unknown function/process/component) than the high-abundance tail. These largely uncharacterized modal regions in the proteome and metabolome hint at a large amount of undiscovered biology.

For metabolomics, we examined the potato GC-Tof MS profiling data (Catchpole *et al.*, 2005). The most abundant metabolites are amino acids and sugars (see Supplementary Table S5). This is consistent with these being the core products of metabolism. In contrast, the metabolites present in the mode are generally not well characterized. We confirmed these observations by analysis of another GC-MS potato dataset.

## 6 DISCUSSION

*A single distributional form*: the results described in this article show that a dPLN model fits well the observed population abundance distributions in the datasets we have examined. We hypothesize that this will be a general observation for transcriptomic, proteomic and metabolomic data. If this general hypothesis is true then it would be of significant biological and technological interest. In biology, it would require us to seek an explanation for such distributions in the statistical mechanics of living systems. For technology, the expectation of observing a specific distribution would enable omic measurement techniques to be better parameterized, resulting in better empirical measurements.

*The left tail—low-abundance components*: the least well-understood part of population abundance distributions is the low abundance—left tail. This is the hardest part to empirically measure. The biological relevance of the left tail is most important to toxicology, as poisons are compounds that have adverse affects at low dosage.

Transcriptomics has made the greatest progress in measuring the left tail. Recent work using tiling arrays (Bertone *et al.*, 2004) and direct sequencing (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008) have made it clear that many more DNA sequences are transcribed in very small quantities than previously expected. Likewise, it seems reasonable to postulate that cells contain small numbers of very rare chemical compounds (Jaynes, 2005).

However, long left tails were not observed in some of the omic datasets. In some cases the explanation for this is understood, for example, that of prot-YeastCytometry where the data were preprocessed before publication to remove the tail. In other cases, it may be due to the great difficulties in measuring such low-abundance molecules. It is of course also possible that the dPLN model does not generally fit biological systems in the left tail. Time will tell.

*Possible explanations for the common pattern in population abundance within cellular systems*: what mechanisms could explain the observation of a common pattern of population distribution in mRNA, protein and metabolite distributions? First, note that lognormal and power law distributions are generated by multiplicative effects. Probably the most straightforward causative mechanism is that it is the result of the involvement of chemical reactions (Limpert *et al.*, 2001). Chemical kinetics clearly has a central role in determining the concentrations of metabolites, and reflection also makes it clear that it also has a key role in transcription and transcription factor binding, translation and protein binding, i.e. all the processes that control the abundances of mRNAs and proteins.

Given that populations of chemical reactions are expected to produce multiplicative distributions, it is then reasonable to expect from the central-limit theorem that the observed abundances would follow a lognormal distribution, as the product of many independent, positive random variables forms a lognormal distribution (Limpert *et al.*, 2001). As we observe a dPLN distribution and not a lognormal distribution, we conclude that chemical pathways are not fully independent.

It is less clear what could cause the observed long tails. There are a number of plausible explanations: positive feedback, a network-topology effect, a lower boundary effect, etc. Cellular systems are not random chemical soups, they are under cybernetic control (Milo *et al.*, 2002; Monod, 1971). A key element of this is cybernetic control of positive feedback: the rich-get-richer mechanism typical of power laws. An alternative, but related explanation, is that it is the topology of the network that causes the long tails. For example, considering metabolism: if the pathway node distribution has a dPLN distribution then the abundances may also reasonably be expected to have the same distribution given random movement through the network (we have confirmed this through simulation); or if the topology of metabolism is tree-like then this might also produce a dPLN distribution as this produces an exponential distribution of path lengths.

Life is an auto-catalytic process. We therefore hypothesize that a non-living complex chemical mixture would produce a lognormal distribution, and that power law tails is a signature of a living system. This hypothesis could be easily tested by observing what happens to mRNA, protein and metabolite distributions after death, and in complex non-living chemical systems such as those designed to create conditions on the early earth (Rasmussen *et al.*, 2004).

To conclude, we observe a common distribution for the population abundances of mRNAs, proteins and metabolites in biological systems: with behaviour like a lognormal distribution around the centre and power law distributions in the tail(s). The existence of common distribution provides insight into the statistical mechanics of living systems, constraints on systems biology cellular models and aids the parametrization of analysis methods.

## ACKNOWLEDGEMENTS

We thank J. Draper and I. Scott for access to their metabolomics data, and D.B. Kell, S.G. Oliver and T. Dix for their help and advice.

*Funding*: BBSRC MeTRO project; EU project UNICELLSYS (FP7-Health-201142).

*Conflict of Interest*: none declared.

## REFERENCES

Arita,M. (2005) Scale-freeness and biological networks. *J. Biochem.*, **138**, 1–4.

Bantscheff,M. and Schirle,M. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.*, **389**, 1017–1031.

Barabási,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Canales,R. *et al*. (2006) The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Catchpole,G.S. *et al.* (2005) Hierarchical metabolomics demonstrates substantial coompositional similarity between genetically modified and conventional potato crops. *Proc. Natl Acad. Sci.USA*, **102**, 14458–14462.

Chua,Alvin, L. *et al*. (2006) Pareto-gamma statistic reveals global rescaling in transcriptomes of low and high aggressive breast cancer phenotypes. In *Pattern Recognition in Bioinformatics, International workshop, PRIB 2006* Vol. 4146 of *Lecture Notes in Computer Science (LNCS)*,Springer, Berlin/Heidelberg, pp. 49–59.

Cravatt, Benjamin,F. *et al*. (2007) The biological impact of mass-spectrometry-based proteomics. *Nature*, **450**, 991–1000.

Elf,J. *et al*. (2003) Near-critical phenomena in intracellular metabolite pools. *Biophys. J.*, **84**, 154–170.

Fiehn,O. (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Friedman,N. *et al*. (2006) Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.*, **97**, 16830.

Furusawa,C. and Kaneko,K. (2003) Zipf's law in gene expression. *Phys. Rev. Lett.*, **90**, 088102.

Hornung,G. and Barkai,N. (2008) Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput. Biol.*, **4**, e8.

Hoyle,D.C. *et al*. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.

Ishihama,Y. *et al*. (2008) Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics*, **9**, 102, http://www.biomedcentral.com/1471-2164/9/102.

Jaynes,E. (2005) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kell,D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin Microbiol.*, **7**, 296–307.

Kenneth,J.R. and Sander,G. (eds) (1998) *Modern Epidemiology*. 2nd edn. Lippincott Williams & Wilkins, Philadelphia.

Khanin,R. and Wit,E. (2006) How scale-free are biological networks. *J. Computat. Biol.*, **13**, 810–818.

Koltai,H. and Weingarten-Baror,C. (2008) Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic Acids Res.*, **36**, 2395–2405.

Konishi,T. (2004) Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics.*, **5**, 1471–2105.

Kuznetsov,V.A. *et al*. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, **161**, 1321–1332.

Limpert,E. *et al*. (2001) Log-normal distributions across the sciences: keys and clues. *Bioscience*, **51**, 341–352.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Marcelino,L.A. *et al.* (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proc. Natl Acad. Sci. USA*, **103**, 13629–13634.

Milo,R. *et al*. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Mitzenmacher,M. (2004) A brief history of generative models for power law and lognormal distributions. *Internet Math.*, **1**, 226–251.

Monod,J. (1971) *Chance and Necessity*. Alfred A. Knopf, New York.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Newman,J.R.S. *et al*. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, **441**, 840–846.

Ochiai,T. *et al*. (2004) A constructive approach to gene expression dynamics. *Phys. Lett. A*, **330**, 313–321.

Paulsson,J. (2004) Summing up the noise in gene networks. *Nature*, **427**, 415–418.

Paulsson,J. (2005) Models of stochastic gene expression. *Phys. Life Rev.*, **2**, 157–175.

Rasmussen,S. *et al*. (2004) Transitions from nonliving to living matter. *Science*, **303**, 963–965.

Reed,W.J. (2003) The Pareto law of incomes - an explaination and an extention. *Physica A*, **319**, 469–486.

Reed,W.J. and Jorgensen,M.J. (2004) The double Pareto-lognormal distribution - a new parametric model for size distributions. *Com. Stats Theory Methods*, **33**, 1733–1753.

Schulze,A. and Downward,J. (2001) Navigating gene expression using microarrays a technology review. *Nat. Cell Biol.*, **3**, E190–E195.

Sinnott,E.W. (1937) The relation of gene to character in quantitative inheritance. *Proc. Natl Acad. Sci. USA*, **23**, 224–227.

Stelling,J. *et al*. (2004) Robustness of cellular functions. *Cell*, **118**, 675–685.

Stephens,M.A. (1974) EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.*, **69**, 730–737.

Stumpf,M.P.H. *et al*. (2005a) Statistical model selection applied to biological network data. *Proc. Computat. Syst. Biology*, **3**, 65–73.

Stumpf,M.P.H. *et al*. (2005b) Subnets of scale-free networks are not scale-free: Sampling properties of the networks. *Proc. Natl Acad. Sci. USA*, **102**, 4221–4224.

Tokita,K. (2006) Statistical mechanics of relative species abundance. *Ecol. Inform.*, **1**, 315–324.

Tong,A.H. *et al* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Tyers,M. and Mann,M. (2003) From genomics to proteomics. *Nature*, **422**, 193–197.

Ueda,H.R. *et al.* (2004) Universality and flexibility in gene expression from bacteria to human. *Proc. Natl Acad. Sci. USA*, **101**, 3765–3769.

Van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **9**, 479–484.

Wilhelm,B. *et al*. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

Wishart,D. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.

Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, **12**, 882–893.

Yule,G. (1925) A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis. *Philos. Trans. R. Soc. Lond. B*, **213**, 21–87.