Statistical Analysis of Global Connectivity and Activity Distributions in Cellular Networks

Adrián López García de Lomana[1], Qasim K. Beg[2,3], G. de Fabritiis[1] and Jordi Villà-Freixa[1,*]

[1]Computational Biochemistry and Biophysics Laboratory, Research Unit on Biomedical Informatics, IMIM / Universitat Pompeu Fabra, c/ Dr. Aiguader 88, 08003, Barcelona, Spain.
[2]Department of Pathology, 3550 Terrace St., University of Pittsburgh, Pittsburgh, PA 15261, USA.
[3]Current Address: Department of Biomedical Engineering, 44 Cummington St., Boston University, Boston, MA 02215, USA.
* Corresponding author, jordi.villa@upf.edu.

Abstract

Various molecular interaction networks have been claimed to follow power-law decay for their global connectivity distribution. It has been proposed that there may be underlying generative models that explain this heavy-tailed behavior by self-reinforcement processes such as classical or hierarchical scale-free network models. Here we analyze a comprehensive data set of protein-protein and transcriptional regulatory interaction networks in yeast, an *E. coli* metabolic network, and gene activity profiles for different metabolic states in both organisms. We show that in all cases the networks have a heavy-tailed distribution, but most of them present significant differences from a power-law model according to a stringent statistical test. Those few data sets that have a statistically significant fit with a power-law model follow other distributions equally well. Thus, while our analysis supports that both global connectivity interaction networks and activity distributions are heavy-tailed, they are not generally described by any specific distribution model, leaving space for further inferences on generative models.

Key words: cellular networks, fat-tailed distributions, maximum likelihood estimation, hypothesis test.

1 Introduction

Over the last few years it has been extensively argued that the connectivity distributions of various molecular interaction networks follow power-law distributions that are best approximated by classical (Barabási and Albert, 1999) or hierarchical (Ravasz et al., 2002) scale-free network models (reviewed in Refs. (Albert, 2005; Barabási, 2004)). Although this does not imply that power-law distributions do in fact describe the observed degree distributions, it has been used as motivation for developing generative models that yield power laws. Moreover, power-law models also appear to characterize the distribution of activity levels. For example, both calculated metabolic flux values (Almaas et al., 2004), and measured gene expression values (Ueda et al., 2004) seem to follow such a distribution, which is in agreement with theoretical predictions for load distribution on scale-free networks (Almaas et al., 2004; Goh et al., 2001).

Using a statistical framework similar to the one employed in this paper, a recent work (Edwards et al., 2007) revisited several enhanced datasets of foraging patterns in wild animals and concluded that the distributions previously classified as power-law were better described by a Gamma distribution. The previous spurious results were attributed in part to the limited magnitude of the dataset (also typical of biological datasets) and to the graphical method used to assess the character of the distributions. It has been suggested (Khanin and Wit, 2006; Stumpf et al., 2007; Tanaka et al., 2005) that the charactheristic observation that biological networks follow a power-law distribution may have been reached due to the same methodological shortcomings. Specifically, it has been argued that analysing relatively small cellular networks, having only a few hundred to a few thousands data elements using the commonly employed frequency-degree or intensity plots, does not have sufficient power to differentiate among various network models having heavy-tailed distributions, and that the use of rank-degree (intensity) plots proves superior for this purpose (Khanin and Wit, 2006; Stumpf et al., 2007; Tanaka et al., 2005). Furthermore, a rigorous quantification of goodness-of-fit is also required to establish relatedness to various network models, and the quality of the underlying data set (i.e., the quality of network reconstruction) is critical for proper analysis.

1.1 The Models and the MLE Analytical Framework

Assessment of the best model explaining a given data distribution has typically been done using simple linear regression methods. These methods are suitable for normal distribution functions, but not for highly skewed distribution functions. Essentially, the problem arises from the fact that skewed distributions are characterized by the scale of the tail, which forms most of the support for the distribution but that is barely populated (i.e. contains less than 10% of the data points). Because of this, simple least square fits of the probability density distribution computed via histogram methods are very poor estimators of the distribution parameters (see (Tanaka et al., 2005) for a review of possible problems) due to the noisy poor sampling of the tails.

Therefore, it has been argued that density plots should not be used as a base for the fitting of these types of data, as several more reliable methods are available. A simple and better strategy is to use rank-plots as commonly used in engineering and economics (Tanaka et al., 2005). Logarithmic binning (Albert et al., 1999) has also been used as a more robust alternative, but it has been reported that this procedure fails to retrieve the value of the exponent as the slope of the graph for power-law distributions (Tanaka et al., 2005; Goldstein et al., 2004; Clauset et al., 2007). Finally, one could apply a logarithmic transformation to the data and fit the corresponding distribution function (De Fabritiis et al., 2003) thus avoiding the problem of skewed data from the outset. In this case the appropriate transformation of the probability distribution has to be performed (for instance a normal distribution for log-transformed, log-normally distributed data), a procedure which could be cumbersome for some distribution functions.

Below we focus on the discrimination of different models (see Section 2.2) for several types of molecular interaction and expression data sets by using probability-based techniques to perform comparative analyses. In order to have a good mathematical representation of the probability distribution we use cumulative distribution functions (cdf) that are directly related to rank-plots (Tanaka et al., 2005). In addition, instead of graphical-based estimation methods we utilize maximum likelihood estimation (MLE) (see Section 2), which is not dependent on the graphical representation and allows us to perform a statistical test of the proposed models (Goldstein et al., 2004; Stumpf and Ingram, 2005; Hoogenboom et al., 2006; Clauset et al., 2007).

Based on these considerations, here we reexamine both the global topological connectivity distributions and absolute gene and protein expression value distributions in cellular networks, with a focus on the most completely reconstructed, highest quality data sets. We have tested the empirical distributions against several plausible models with heavy-tailed distribution. Note also that generative models in the context of biological networks have been proposed for power-law (Barabási and Albert, 1999; Qian et al., 2001; Rzhetsky and Gomez, 2001; Bhan et al., 2002; Pastor-Satorras et al., 2003) and broad-scale (Amaral et al., 2000) distributions but not for the rest of the tested distributions. We examine the extent to which several types of distributions, some with plausible underlying generative models, can provide a similar probabilistic framework for the experimental data being analyzed. We show that for the analyzed molecular interaction distributions, only the high-throughput protein-protein interaction data set and the out-degree distribution of the transcriptional regulatory network significantly fitted the power-law model. In the case of the distributions of global gene and protein expression values, only two mid-log cultures of *E. coli* showed a statistically significant fit. At any rate, all experimental data sets having significant fits for the power-law distribution, showed equally good fits for other distributions which leads to inconclusive results for the support of a single specific distribution. We discuss the implications of this finding for the uniqueness of generative models solely from topology and activity data distributions.

2 Materials and Methods

For model classification, we have basically used the methods developed and implemented in (Clauset et al., 2007). However, we provide here sufficient details about the mathematical foundations to make the analytical procedure clear.

2.1 Probability Based Estimation Method

The likelihood of a given probability distribution $p_{\mathbf{v}}(x_i)$ depending on parameters $\mathbf{v}$ and describing a given dataset of independent data $(x_1, ..., x_N)$, is defined as $l(\mathbf{v}|x_1, ..., x_N) = \prod_i^N p_{\mathbf{v}}(x_i)$. Our datasets are built of the (discrete) number of edges in interaction networks or the (continuous) amount of expression of the nodes in activity networks. Therefore, from a group of edges or nodes $(x_1, ..., x_N)$, we obtain the probability associated with each value $(p_{\mathbf{v}}(x_1), ..., p_{\mathbf{v}}(x_N))$.

Here we use different model distributions with their corresponding set of parameters $\mathbf{v}$. In each case a maximum likelihood estimation (MLE) based on the log-likelihood function, $L(\mathbf{v}|x) = \log\, l(\mathbf{v}|x)$, is used to compute the set of parameters $\mathbf{v}_{opt}$ which maximizes L. A special case represents the parameter x_{min} which is the lower bound of the power-law distribution (see Section *SI* 2). For its calculation we followed the alternative procedures developed and implemented at (Clauset et al., 2007). Once the vector $\mathbf{v}_{\text{opt}}$ has been obtained, we move on to find if there are differences between the empirical data and the model. An intuitive approach would be to perform a Kolmogorov-Smirnov (KS) test using the empirical distribution set and the estimated model distribution (Goldstein et al., 2004). However the two distributions are not independent, which is one of the required premises of the KS test, because the model was estimated from the same data with which we want to perform the test. Instead, a Monte Carlo protocol can be used to avoid such direct comparison. As described in (Clauset et al., 2007), the KS statistic D is calculated for many Monte Carlo generated data sets from the estimated distribution. Our p-value will be simply the fraction of times that D for the empirical data is larger than D for the generated data sets (Clauset et al., 2007). We define the null hypothesis H_0 as, no statistical difference between the data and the model. In our case, the confidence value of $p = 0.1$ will in fact be more appropriate than $p = 0.05$ for the statistical test, following (Clauset et al., 2007; Mayo and Cox, 2006). The rationale is that we are looking for differences in only one tail of the distribution, we look for p-values that are larger than that associated with the experimental fit.

2.2 Distribution Functions Estimation

Both continuous and discrete model distributions are used depending on the nature of the data analyzed in each case. As described in Section 2.2.1, we have tested seven models of probability distribution against the studied data. In turn, five model distributions were tested against the continuous data sets: power-law, log-normal, exponential, Weibull and

broad-scale. It is worth noting that only some of the distributions tested (e.g., power-law and broad-scale) have been put forward as distributions agreeing with plausible generative models in the biological context (Barabási and Albert, 1999; Amaral et al., 2000). The intention of this work is to expand the application of the methods to other *a priori* suitable distributions to give a more general scope to the study.

2.2.1 Distribution Functions

Power-law model: A random variable X is said to follow a power-law distribution with index α when the scaling law of $P[X > x]$ is power-law, such as $P[X > x] = 1 - P[X \leq x] \approx cx^{-\alpha}$ for x$\rightarrow \infty$. We used the Pareto distribution for the continuous data: probability density function (pdf): $p(x) = \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$, cumulative distribution function (cdf): $P[X \leq x] = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}$ and for the discrete case we used Zipf's law with the following terms, pdf: $p(k) = \frac{k^{-\alpha}}{\zeta\,(\alpha,k_{min})}$, cdf: $P[K \leq k] = \frac{\zeta\,(\alpha,k)}{\zeta\,(\alpha,k_{min})}$. The Hurwitz zeta function is defined as $\zeta\,(\alpha,k) = \sum_{n=0}^{\infty}(n+k)^{-\alpha}$. The reader is referred to (Goldstein et al., 2004) and (Clauset et al., 2007) for further explanations. Specifically, for the determination of x_{min} (or k_{min}), we have used the methods described in (Clauset et al., 2007).

Log-normal model: A random variable X is log-normally distributed if $Y = log(X)$ follows a normal distribution. We have used the following distributions for the continuous case: pdf: $p(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$, cdf: $P[X \leq x] = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{ln(x)-\mu}{\sigma\sqrt{2}})$. Numerical approximations of (Clauset et al., 2007) have been used for the discrete models.

Poisson model: We have used the discrete Poisson distribution to model the discrete data sets using the following probability mass function (pmf): $f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$.

Yule model: Again this is a discrete distribution that we have only used for discrete data sets. We used the following forms, pdf: $p(k) = \rho\mathbf{B}(k,\rho+1)$ and cdf: $P[K \leq k] = 1 - k\mathbf{B}(k,\rho+1)$, where $\rho > 0$ and $\mathbf{B}$ is the Beta function.

Exponential model: Exponential models have no biological relevance in the framework of our study. However, we have included it in some cases for methodological comparison as a representative model of a non fat-tailed distribution. For a random variable X distributed exponentially we have used the following forms: pdf: $p(x) = \lambda e^{-\lambda x}$, cdf: $P[X \leq x] = 1 - e^{-\lambda x}$ for the continuous data sets; we have taken the pmf: $f(k,\lambda) = (1 - e^{-\lambda})e^{\lambda k_{min}}\ e^{\lambda k}$ as the discrete model.

Weibull model: Also referred to as stretched exponential, we have used pdf: $p(x) = (k/\lambda)(x/\lambda)^{k-1}e^{-(x/\lambda)^k}$ and cdf: $P[X \leq x] = 1 - e^{-(x/\lambda)^k}$ for the continuous data sets. For the discrete data we have used the code available from (Clauset et al., 2007) that uses the Nakagawa-Osaki (Nakagawa and Osaki, 1975) method for the discretization of the Weibull distribution. pmf: $f(k;q,\beta) = q^{k^\beta}$.

Broad-scale model: Broad-scale distribution functions, also referred to here as power-law with cut-off or power-law plus exponential, identify a class of distributions that are characterized by a scaling law $p(x) \approx ce^{-\lambda x}x^{-\alpha}$ for $x \rightarrow \infty$. Note that the broad-scale

is very similar to the Gamma distribution. Both contain a shape and a scale parameter. The Gamma distribution in addition contains a normalizing constant. For the Gamma distribution, the pdf would be $p(x) = \frac{\beta^\alpha}{\Gamma(x)} e^{-\beta x} x^{\alpha-1}$, being $\Gamma(x) = (\alpha - 1)!$ for $\alpha > 0$. The broad-scale distribution is a type of nested function, which implies a specific statistical treatment in Section 2.3. In the current case the broad-scale consists of a power-law behavior and after a given threshold it has an exponential decay. Numerical routines available from (Clauset et al., 2007) have been used to calculate the corresponding pdf and cdf.

2.3 Distribution Comparison

To compare the feasibility of the different models with respect to the power-law distribution, we applied a likelihood ratio test, which compares the fits of two given competing distributions. Note that we are not comparing a model against the empirical distribution but two different distributions with each other. Strictly speaking in this case we compare different models with each other to indirectly assess how well the distributions explain the data. We have again applied the methods developed in (Clauset et al., 2007) and (Vuong, 1989) to evaluate the normalized log-likelihood ratio (NLLR). If the NLLR has a positive value the power-law model is supported, whereas the alternative distribution offers a better fit if NLLR has a negative value. In order to determine significant positive or negative values, we calculated the associated p-value. Given the different experimental design with respect to the one described in Section 2.1, we take here as confidence value the most commonly used $p = 0.05$ (Mayo and Cox, 2006). We search for differences on both sides of the distribution and therefore we use $p = 0.05$ as level of significance.

The broad-scale distribution requires a special comment, as it is a nested function containing both power-law and exponential terms. In fact, if we compare the power-law distribution fit with the broad-scale distribution fit, the NLLR test will always be zero or negative, favoring the broad-scale, as it reproduces the power-law behavior plus an additional term. To solve this problem, we used a correction of the p-value calculated for the log-likelihood ratio (LLR) as described in (Vuong, 1989) and (Clauset et al., 2007).

2.4 Data Sets Used

Molecular interaction maps were obtained from different sources: *S. cerevisiae* protein-protein interaction networks (Reguly et al., 2006), *S. cerevisiae* transcriptional regulatory network (MacIsaac et al., 2006), *E. coli* metabolic networks (Feist et al., 2007). Yeast expression values were taken from (Ghaemmaghami et al., 2003). Global mRNA expression data for *E. coli* cells and protein expression data for *S. cerevisiae* cells were obtained from (Beg et al., 2007).

3 Results

3.1 Global Topological Organization of Molecular Interaction Networks

The aim of this paper is to test by means of a probabilistic approach the ability of heavy-tailed distribution functions to explain the experimental data distributions for given datasets. Due to the nature of the experimental data being considered here, we have differentiated between global interaction networks and global activity networks. Global interaction measurements refer to the number of interactions of a given molecule. Thus, measurements are counted as natural numbers and the studied models will be discrete. On the other hand, global activity interactions are defined in terms of cellular expression, measurements are expressed using positive real numbers and the applied models will be continuous.

3.1.1 Global Interaction Networks

To assess the degree distribution of an undirected homotypic molecular interaction network we used data obtained from baker's yeast, *S. cerevisiae*, in (Reguly et al., 2006) that, to date, is the most comprehensive information for a species-specific protein-protein interaction (PPI) network. We analyzed two data sets built using two different methodologies. The first data set was created from the combination of five different studies based on experimental high-throughput techniques, which we refer to as HTP data (Reguly et al., 2006). The second data set was curated manually from the literature (LC data set), and is assumed to be more accurate than the former (Reguly et al., 2006). After the removal of redundant and self-interactions (Reguly et al., 2006), we obtained a data set of 11,571 interactions between 4,474 proteins for the HTP data and 8,165 interactions between 2,689 proteins for the LC data. Table 1 shows that after fitting the parameters using an MLE, only the HTP data set provides a statistically sufficient goodness-of-fit, based on the KS test (see *Materials and Methods*). Whether or not the high rate of false-discovery interactions from yeast-two hybrid experiments in the HTP data set leads to the acceptance of the null hypothesis for the power-law model is an open issue for discussion (Huang et al., 2007). In fact, for the same network, if the data is retrieved manually from the literature (LC data set), the power-law model is rejected. Furthermore, for the case of HTP data, exponential and Poisson distributions behave significantly worse than the power-law model while for the LC data set, log-normal, Yule, Weibull and broad-scale are better models than the power-law. Fig. 1**a** shows the degree distribution of the LC network with its corresponding best fitted power-law model. An equivalent plot for the HTP data set is shown in Fig. *SI* 1**a**.

We also analyzed the degree distribution of a directed heterotypic molecular interaction network, the transcriptional regulatory (TR) network of *S. cerevisiae* (MacIsaac et al., 2006), the edges of which represent interactions between transcription factors (TF) and gene regulatory regions, distinguishing out-degree from in-degree interactions. In total the data includes 99 TFs and 1,851 TF-regulated genes connected through 3,394 links. Table 1 clearly indicates that the degree distribution of out-degree interactions (TF→gene) provide

a statistically significant fit for the power-law model. However, all other tested distributions except the Poisson fit the empirical data equally well which prevents us from establishing unequivocally the power-law distribution as the statistical model to describe the data. This result should be taken with caution as the sample size ($N = 99$) is really on the limit to be considered too small (Hart and Clark, 1999). For the case of in-degree interactions (number of TF affecting a given gene) the range of connectivities is too small to provide conclusive results, although it appears that all other tested models apart from the Poisson model offer significantly better fits than the power-law. A graphical representation of the out-degree distribution with their best-fitted power-law model is shown in Fig. 1**b**. Fig. *SI* 1**c** shows the corresponding plot for the in-degree distribution.

Finally, we assessed the recent high quality reconstruction of the *E. coli* metabolic network (Feist et al., 2007), in which substrates are connected to each other through edges that represent the actual metabolic reactions (Jeong et al., 2000) with a total of 2,381 reactions, of which 304 are exchange, 553 reversible and 1,524 irreversible reactions. We have studied separately the in-degree distribution with 1,657 nodes and 3,050 links and the out-degree distribution with 1,656 metabolites and 2,788 links. Both networks behave very similarly: both display significant differences with the power-law model. While the log-normal model fits the data equally as well as the power-law model, all other tested models behave significantly worse. The empirical distributions with their corresponding best-fitted power-law models are represented graphically in Fig. *SI* 1**e**,**f**.

3.1.2 Clustering Coefficients

To gain further insight into the topological organization of molecular interaction networks, we have analyzed the $C(k)$ *vs.* k relationships for the HTP and LC reconstructed PPI networks of yeast (Reguly et al., 2006), the TR regulatory network of yeast (MacIsaac et al., 2006), and the metabolic network of *E. coli* (Feist et al., 2007). Fig. *SI* 2 shows how the dependence of the clustering coefficient in the three types of molecular interaction networks is not uniform. While the *E. coli* metabolic network displays a distribution close to $C(k) = k^{-1}$ (Fig. *SI* 2**d**), as previously described (Ravasz et al., 2002), the distribution observed in the other two networks is significantly less organized.

3.2 Global Activity Level Distributions in Molecular Interaction Networks

Next we assessed the distribution of activity levels achieved on the underlying network topology. First, we examined single time point snapshots for mRNA and protein expression in *E. coli* and in *S. cerevisiae*, respectively.

We used global gene expression data from *E. coli* cells grown in mid-log phase batch cultures with single carbon source media, using acetate, galactose, glucose, glycerol and maltose as individual carbon sources (Beg et al., 2007). Signals for 3,977 probes were examined for the expression level of the corresponding genes based on their hybridization intensity (Table

2). Activity distribution of acetate growing cells significantly fits the power-law model, in fact representing the best fit with the data in all cases analyzed in this study. Interestingly though the broad-scale model gives us a significantly much better fit. Maltose is another carbon source where gene expression distribution of *E. coli* shows significant fit with the power-law model. However, three other distributions, log-normal, Weibull and broad-scale provide significantly better fits than the power-law model. Of the two cases, in which the power-law model displays no significant differences from the empirical data but other models are also plausible, the classification of power-law character remains uncertain. None of the other mid-log cultures provide a significant fit with the power-law distribution. For glucose, galactose and glycerol, log-normal, Weibull and broad-scale are significantly better models than the power-law, while exponential is significanlty worse model. Worth noting is the graphical comparison of Fig. 1 **c** and 1**d**. The acetate distribution displays no statistical difference with the power-law model while the galactose distribution does differ from it. This important difference in the nature of the empirical data is not revealed from the inspection of the graphical representations but uniquely from the statistical test.

We also examined the transcriptome state of *E. coli* cells grown in chemostat cultures at different dilution rates, representing various steady-state growth rates at different culture densities (Vázquez et al., 2008). We find that the general pattern for gene expression distribution is largely invariant in all different growth media and at different growth rates, but for the steady-state cultures none of the empirical data sets fit the power-law model significantly (Table *SI* 1 and Fig. *SI* 4). Curiously the broad-scale model outperforms for all steady-state cases. Also note that the log-normal and Weibull distributions display superior fits for dilutions 0.25 and 0.4. In all the cases, the exponential distribution fits the empirical data significantly worse than the power-law distribution.

Protein expression values in mid-log phase of *S. cerevisiae* cell cultures from 3,868 different strains expressing a single GFP or TAP tagged protein in a rich growth media (Ghaemmaghami et al., 2003) displayed a slightly different expression mode (Fig. *SI* 5), suggesting difference in mRNA and protein expression value distributions under these growth conditions. Yet again the empirical dataset shows significant differences with respect to the power-law model (Table *SI* 2). Comparing models, the power-law distribution fits significantly better than the exponential, although similarly to the log-normal and Weibull distributions. However the broad-scale model offers a significantly better fit that the power-law model.

Finally, to assess the transcriptome state of *E. coli* cells under the most complex growth conditions, we have examined the dynamical microarray profile of *E. coli* cells grown in batch culture in a mixed-substrate medium (Beg et al., 2007). At all sampled time points, the distribution of expression of the transcriptome was differing significantly from the power-law distribution (Table *SI* 3). In this mRNA expression data set the log-normal, Weibull and broad-scale models show better statistical fit with the experimental data at all time points, while the exponential distribution is also better for some of the cases. A graphical representation of the best fitting power-law distribution against the empirical data is shown in Fig. *SI* 6.

4 Discussion

The structure and activity of intracellular molecular interaction networks represents a basic tenet of life. Thus it is of great importance to correctly identify their true structural and functional properties at the global, intermediate and local levels. Heavy-tailed probability distributions have been used in the literature to explain the topological features of complex biological networks. In this work we have analyzed to what extent, among others, three well known heavy-tailed distributions (power-law, broad-scale and log-normal) can be uniquely used to represent the observed experimental data for protein interaction, transcription regulation, metabolic pathways and expression experiments. Our results demonstrate that the large-scale topology of the molecular interaction networks and the global mRNA and protein expression distributions examined here do not strictly follow power-law distributions. Moreover, none of the three heavy-tailed models tested had a universal agreement with the empirical data even when using the highest quality data sets available. Distributions are evidently heavy-tailed and for this type of data MLE analyses prove superior to graphical methods for assessing different tested distributions. However, we could not assess any of the studied models with statistical reliability. Our results could be explained by several non-exclusive arguments. First, our analyses could be affected by a biased acquisition of the empirical data, a systematic error that would affect, for instance, proteins which have been studied more extensively because they have a high biomedical interest. Also, such simple mathematical frameworks as considered here may not represent entirely the biological mechanisms at work, but also be the convolution of the physicochemical constraints of the cell (Beg et al., 2007). Finally, population-level evolutionary processes, such as activation of a foraging program upon extracellular substrate exhaustion, clearly affect the function of cellular networks (Beg et al., 2007) and are likely to influence the nature of the underlying molecular interactions as well.

Additional problems can have their origin in data acquisition. In some cases, the experimental data might be incomplete or noisy, specially for the HTP data which derives partially from yeast two-hybrid experiments. There has been several efforts (Huang et al., 2007; Scholtens et al., 2008) to infer statistically the actual PPI network from such experiments and consequently our results of the analysis refer to the data, not the actual network. Indeed, the analysis of the reconstructed network from literature, in a manually curated way, considered to be more correct, yields different results. The inference of the actual network is out of the scope of the current work.

We should also note that the topology of cellular networks can be viewed at various levels of complexity. For example, in the metabolic representation considered here each metabolite that participates in an interconversion reaction is considered equal (Jeong et al., 2000). However, in alternative representations molecules that only act as donor or acceptor (e.g., ATP or ADP) or that do not contribute a carbon or nitrogen atom to a metabolic reaction can be excluded (Arita, 2004), or they can be pre-classified according to their perceived biochemical role (Tanaka, 2005). Similarly, in TR networks, genes and their protein products are usually defined as common nodes with regulatory links among them being mediated

by the binding of TFs to the promoter regions of the genes (Shen-Orr et al., 2002). At other times, however, genes and their protein products are considered as separate nodes (Yeger-Lotem et al., 2004). While certain properties are unaffected by alternative network representations, others are affected (Arita, 2004; Tanaka, 2005) potentially complicating the interpretation of subsequent analytical results.

Finally, our analyses also reveal highly similar, but dynamically regulated global mRNA and protein expression profiles, in which expression values are significantly variable. Thus, while the global function of cellular networks is expected to be greatly influenced by their underlying topology (Almaas et al., 2004; Ueda et al., 2004; Goh et al., 2001), mRNA or protein expression data reflect the dynamic physicochemical state of the cell, likely necessitating an explicit dynamical model for establishing a relationship between topology and expression.

Acknowledgments

We thank R.V. Solé, R. Albert, P. Delicado, P. Rué, M. Dies and S. Laurie for comments at different stages of the work. We also thank the constructive comments from two anonymous reviewers. A.L.G. de L. acknowledges Generalitat de Catalunya for FI and BE grants. G.D.F. acknowledges the Ramón y Cajal granting scheme. This research was supported in part by EC funded FP6 STREP projects BioBridge (contract number FP6-037909), QosCosGrid (contract number FP6-033883), and VPH NoE (contract number FP7-223920).

Disclosure Statement

No competing financial interests exist.

References

R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118:4947–4957, 2005.

R. Albert, H. Jeong, and A-L Barabási. Diameter of the World-Wide Web. *Nature*, 401:130, 1999.

E. Almaas, B. Kovács, Z. N. Oltvai, and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427:839–843, 2004.

L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small world networks. *PNAS*, 97:11149–11152, 2000.

M. Arita. The metabolic world of *Escherichia coli* is not small. *PNAS*, 101:1542–1547, 2004.

A.-L. Barabási. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–13, 2004.

A.-L. Barabási and R. Albert. Emergence of Scalling in Random Networks. *Science*, 286:509–512, 1999.

Q. K. Beg, A. Vázquez, J. Ernst, Bar-Joseph Z. de Menezes, M. A., A.-L. Barabási, and Z. N. Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *PNAS*, 104:1266312668, 2007.

A. Bhan, D. J. Galas, and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18:1486–1493, 2002.

A Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. 2007. URL http://arxiv.org/abs/0706.1062.

G. De Fabritiis, F. Pammolli, and M. Riccaboni. On size and growth of business firms. *Physica A*, 324:38–44, 2003.

A. M. Edwards, R. A. Phillips, N. W. Watkins, M. P. Freeman, E. J. Murphy, V. Afanasyev, S. V. Buldyrev, M. G. E. da Luz, E. P. Raposo, H. E. Stanley, and G. M. Viswanathan. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449:1044–1048, 2007.

A.M. Feist, C. S. Henry, J. L. Reed, A. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. . Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, 3:121, 2007.

S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E.K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425:737–741, 2003.

K.I. Goh, B. Kahng, and D. Kim. Universal Behavior of Load Distribution in Scale-Free Networks. *PRL*, 87(27):278701(4), 2001.

M.L. Goldstein, S.A. Morris, and G.G. Yen. Problems with fitting to the power-law distribution. *Eur. Phys. J. B*, 41:255–258, 2004.

R. A. Hart and D. H. Clark. Does Size Matter? Exploring the Small Sample Properties of Maximum Likelihood Estimation. *Annual Meeting of the Southern Political Science Association*, 1999.

J.P. Hoogenboom, W.K. den Otter, and H.L. Offerhaus. Accurate and unbiased estimation of power-law exponents from single-emitter blinking data. *The Journal of Chemical Physics*, 125:204713, 2006.

H. Huang, B. M. Jedynak, and J. S. Bader. Where Have All the Interactions Gone? Estimating the Coverage of Two-Hybrid Protein Interaction Maps. *PLoS Computational Biology*, 3:e214, 2007.

H. Jeong, R. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407: 651–654, 2000.

R. Khanin and E. Wit. How scale-free are biological networks? *J. Comput. Biol.*, 13:810–818, 2006.

K.D. MacIsaac, T. Wang, B. Gordon, D.K. Gifford, G.D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006.

D.G. Mayo and D. R. Cox. Optimality: The second Erich L. Lehmann Symposium. *Institute of Mathematical Statistics, Bethesda, Maryland*, pages 77–97, 2006.

T. Nakagawa and S. Osaki. The Discrete Weibull Distribution. *IEEE Transactions on Reliability*, 24:300–301, 1975.

R. Pastor-Satorras, E. Smith, and R. Solé. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.*, 222: 199–210, 2003.

J. Qian, N. M. Luscombe, and M. Gerstein. Protein Family and Fold Occurrence in Genomes: Power-law Behavior and Evolutionary Model. *J. Mol. Biol*, 313:673–681, 2001.

E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297:1551–1555, 2002.

T. Reguly, A. Breitkreitz, L. Boucher, B.-J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *Journal of Biology*, 5:11, 2006.

A. Rzhetsky and S. M. Gomez. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, 17:988–996, 2001.

D. Scholtens, T. Chiang, W. Huber, and R. Gentleman. Estimating node degree in bait-prey graphs. *Bioinformatics*, 24:218–224, 2008.

S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31:64–68, 2002.

M. P. Stumpf and P. J. Ingram. Probability models for degree distributions of protein interaction networks. *Europhys. Lett.*, 71(1): 152–158, 2005.

M. P. Stumpf, W. P. Kelly, T. Thorne, and C. Wiuf. Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol. Evol.*, 22(7):366–373, 2007.

R. Tanaka. Scale-Rich Metabolic Networks. *PRL*, 94:168101, 2005.

R. Tanaka, T.-M. Yi, and J. Doyle. Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579:5140–5144, 2005.

H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino. Universality and flexibility in gene expression from bacteria to human. *PNAS*, 101:3765–3769, 2004.

A. Vázquez, Q. K. Beg, M. A. de Menezes, E. Jason, Z. Bar-Joseph, A.-L. Barabási, L. G. Boros, and Z. N. Oltvai. Impact of the solvent capacity constraint on *E. coli* metabolism. *BMC Systems Biology*, 2:7, 2008.

Q. H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57:307–333, 1989.

E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, 101:5934–5939, 2004.

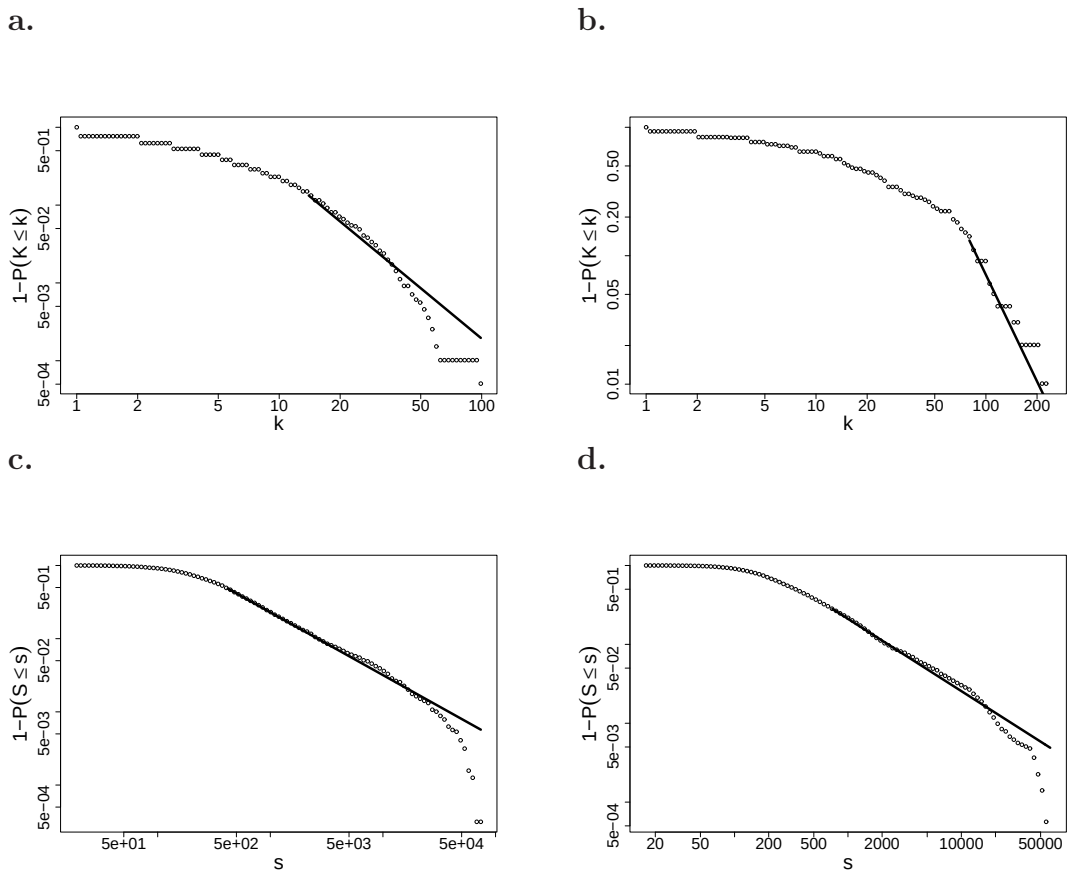

Figure 1: The connectivity degree distribution of (**a.**) LC derived *S. cerevisiae* protein-protein interaction; (**b.**) out-degree distribution of the transcriptional regulatory network of *S. cerevisiae* and (**c., d.**) intensity distribution of mRNA expression values in *E. coli* cells from mid-log (OD 0.2) cultures with (**c.**) acetate and (**d.**) galactose as single carbon source. Points represent the empirical distribution, lines represent the most likely power-law models.

	Power-Law	Log-Normal		Poisson		Yule		Exponential		Weibull		PL + Exp.		
data set	p	NLLR	p	NLLR	p	NLLR	p	NLLR	p	NLLR	p	LLR	p	diagnosis
HTP	**0.76**	-0.67	0.50	3.62	**0.00**	-1.00	0.32	2.71	**0.01**	-0.25	0.80	-0.74	0.22	ambiguous
LC	0.01	-2.14	**0.03**	4.30	**0.00**	-4.29	**0.00**	-0.52	0.60	-2.16	**0.03**	-6.86	**0.00**	reject
TR In	0.00	-4.49	**0.00**	2.08	**0.04**	-7.98	**0.00**	-2.86	**0.00**	-4.64	**0.00**	-26.40	**0.00**	reject
TR Out	**0.75**	-0.23	0.82	2.22	**0.03**	-0.24	0.81	0.25	0.80	-0.23	0.82	-0.10	0.65	ambiguous
Met In	0.04	-1.95	0.05	2.24	**0.02**	2.94	**0.00**	4.00	**0.00**	5.66	**0.00**	0.00	1.00	reject
Met Out	0.00	-1.11	0.27	2.62	**0.01**	4.59	**0.00**	4.68	**0.00**	3.03	**0.00**	0.00	1.00	reject

Table 1: Summary of probabilistic analyses for representative examples of the discrete data sets. The Supplementary Information includes the further results discussed in the text. The second column shows the p-value associated with the differences between the data and the power-law model. The next six columns correspond to the log-likelihood ratio tests comparing the power-law model with other plausible models. Positive values support the power-law model. The normalized log-likelihood ratio (NLLR) is used for non-nested functions while the raw log-likelihood ratio (LLR) is used for the power-law with exponential cutoff model. p-values here are associated with the differences between the two models. Significant p-values are denoted in bold.

	Power-Law	Log-Normal		Exponential		Weibull		PL + Exp.		
data set	p	NLLR	p	NLLR	p	NLLR	p	LLR	p	diagnosis
Acetate	**0.28**	-1.86	0.06	14.49	**0.00**	-1.89	0.06	-10.61	**0.00**	ambiguous
Galactose	0.00	-2.19	**0.02**	10.84	**0.00**	-2.26	**0.02**	-10.91	**0.00**	reject
Glucose	0.05	-2.40	**0.01**	12.16	**0.00**	-2.48	**0.01**	-13.31	**0.00**	reject
Glycerol	0.04	-2.68	**0.01**	11.38	**0.00**	-2.77	**0.00**	-14.58	**0.00**	reject
Maltose	**0.21**	-2.26	**0.02**	12.90	**0.00**	-2.32	**0.02**	-12.33	**0.00**	ambiguous

Table 2: Summary of probabilistic analyses for representative examples of the continuous data sets. The Supplementary Information includes the other results discussed in the text. Interpretation of the table as for Table 1.