Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Minimization of transcriptional temporal noise and scale invariance in the yeast genome

R.C. Ferreira[1]*, F. Bosco[1]*, P.B. Paiva[2] and M.R.S. Briones[1]

[1]Departamento de Microbiologia, Imunologia e Parasitologia,
[2]Departamento de Informática em Saúde,
Universidade Federal de São Paulo, São Paulo, SP, Brasil
*These authors contributed equally to this study.
Corresponding author: M.R.S. Briones
E-mail: marcelo@ecb.epm.br

Genet. Mol. Res. 6 (2): 397-414 (2007)
Received February 6, 2007
Accepted April 16, 2007
Published June 29, 2007

ABSTRACT. The analysis of transcriptional temporal noise could be an interesting means to study gene expression dynamics and stochasticity in eukaryotes. To study the statistical distributions of temporal noise in the eukaryotic model system *Saccharomyces cerevisiae*, we analyzed microarray data corresponding to one cell cycle for 6200 genes. We found that the temporal noise follows a lognormal distribution with scale invariance at the genome, chromosomal and sub-chromosomal levels. Correlation of temporal noise with the codon adaptation index suggests that at least 70% of all protein-coding genes are a noise minimization core of the genome. Accordingly, a mathematical model of individual gene expression dynamics was proposed, using an operator theoretical approach, which reveals strict conditions for noise variability and a possible global noise minimization/optimization strategy at the genome level. Our model and data show that minimal noise does not correspond to genes obeying a strictly deterministic dynamics. The natural strategy of minimization consists in equating the mean of the absolute value of the relative variation of the expression level (α) with noise (η). We hypoth-

esize that the temporal noise pattern is an emergent property of the genome and shows how the dynamics of gene expression could be related to chromosomal organization.

Key words: Transcription, Noise, Scale invariance, Yeast genome

INTRODUCTION

Genomes consist of interdependent informational units (genes) arranged sequentially in linear structures (chromosomes), which compose a larger ensemble. Cellular and organismal levels of life are, therefore, emergent properties of this ensemble. At the root of this emergence is transcription and translation, the information flow from genes to proteins leading to higher order structures with increasing complexity (Adami et al., 2000). Transcription and translation could be viewed as an information transmission channel affected by noise. Accordingly, several studies have addressed transcriptional and translational noise and stochastic mechanisms of gene expression either in prokaryotes or eukaryotes (Elowitz et al., 2002; Blake et al., 2003). Microarray data have been used to investigate the large-scale organization of gene expression and have revealed the complex networks of gene activity (Tong, 2004; Davierwala et al., 2005; Pan et al., 2005). Genome-wide analyses have demonstrated a gene expression dynamics, conserved from *Escherichia coli* to *Homo sapiens*, in which gene expression changes are proportional to their initial expression level (Ueda et al., 2004). The initial expression level, however, varies from cell to cell. This variability is inevitable in biological systems and can result in very different synthesis rates of a specific protein in genetically identical cells living in essentially identical environments (Elowitz et al., 2002; Ozbudak et al., 2002; Blake et al., 2003). The coefficient of this variation is designated noise. This type of noise could be described as populational noise since it corresponds to the cell-to-cell stochastic variation in gene expression levels. Yeast adaptively minimizes noise during the expression of most of its genes. Noise in protein synthesis (translation noise) is minimized in essential genes and in genes encoding protein complex subunits (Fraser et al., 2004). It is of fundamental importance to describe the basic mechanisms of noise and to address the central question of how cells deal with noise. In other words, how is the noise optimized so that gene expression balances the necessary flexibility for adaptive adjustments with its conserved, evolutionarily constrained, mechanisms of control?

Besides the studies that aim to describe mechanistic details of how noise affects gene expression, it is relevant to note that the gene expression noise could be an emergent property of the genome or at least could strengthen the concept of a cell as a self-organizing network (Nicolis and Prigogine, 1977; Kauffman, 1993; Bar-Yam, 2004). Self-organization is a dynamical and adaptive process where systems, generally open systems, acquire and maintain internal organization structures themselves and where complexity increases without external guidance or management (Adami, 2002). Self-organizing systems are widely characterized in physics, chemistry (self-assembly), economics, anthropology (self-organizing behavior), and mathematics (cellular automata). Because biology deals with scales that vary from the sub-cellular to the

ecosystem level, the self-organization concept evidently plays a central role in the description of biological phenomena. According to the theories of hypercycles and autocatalytic networks, the origin of life itself is a product of self-organizing chemical systems (Eigen, 1971; Ycas, 1999). Biological systems exhibit increase in order, autonomy, adaptability, feedback controls, and far-from-equilibrium dynamics, which are signatures of self-organization (De Wolf and Holvoet, 2005). Several, self-organizing, complex behaviors include the formation of lipid bilayer membranes, spontaneous folding of proteins, morphogenesis, and animal social structures. Typically, self-organizing systems display emergent properties.

The concepts of self-organization and emergence are often used incorrectly as synonyms (De Wolf and Holvoet, 2005). The connection between emergence and self-organization is a very intriguing scientific problem with deep implications for genome biology. Emergence is the formation of complex patterns from more elemental parts, or behaviors, constituents of a system (Anderson, 1972; Bar-Yam, 2004). To be termed emergent, a phenomenon should usually be unpredictable from a sheer lower level description of the elemental parts and behaviors. In complex systems, emergence is a central concept, although it is difficult to define and a matter of debate. The identification and characterization of signatures of self-organization and emergence using eukaryotic genome data is a wide open area of research and may be essential to quantitatively understand how the genetic information directs and controls the formation of cells, bodies and higher-order biological behaviors. Gene expression noise, either populational or temporal, could be such a signature.

Here, we studied the temporal fluctuations of gene expression during a single-cell cycle, which we called temporal noise, to discriminate from sheer cell-to-cell stochasticity or populational noise. The statistical distributions and possible correlations of temporal transcriptional noise with gene expression level, codon adaptation (Sharp and Li, 1987), essentiality (Davierwala et al., 2005), and ohnologs (Wolfe and Shields, 1997) by means of whole-genome analysis of the eukaryotic model system *Saccharomyces cerevisiae* were made. We found that the noise follows a lognormal distribution with scale invariance from the genome to the sub-chromosomal level. Also, we identified a noise minimization core in the yeast genome, which encompasses at least 70% of the genome. A mathematical model of gene expression was built, using operator theory, and revealed a possible general genome level strategy for noise minimization/optimization. These results suggest that this noise minimization strategy is a result of self-organization; the noise pattern is an emergent property of the genome whose distribution reveals scale invariance.

MATERIAL AND METHODS

Data source

The *S. cerevisiae* microarray data were obtained from the *Saccharomyces* Cell Cycle Expression Database (http://genomics.stanford.edu) and consists of 17 gene expression measurements (equally time-spaced over one cell cycle in synchronized cells) of 6200 genes of *S. cerevisiae*, strain K3445 (Cho et al., 1998).

Based on whole-genome systematic deletion data in the *Saccharomyces* Genome Database (http://yeastgenome.org), the set of all genes was divided into two subsets of non-essential (4683 genes) and essential (1116 genes) genes. Genes where systematic deletion data are unavailable were excluded from the analysis.

The list of the 554 ohnolog pairs (genes remaining as duplicates after the whole genome duplication) was obtained from the Yeast Gene Order Browser project (http://wolfe.gen.tcd.ie/ygob/) (Byrne and Wolfe, 2005).

The list of genes that are cell cycle regulated was obtained from Spellman et al. (1998).

Calculation of noise

The arithmetic mean expression level (sample mean) $\bar{E}$ and the corresponding sample standard deviation *S* were calculated for each ORF. The transcriptional noise of the expression signal $E_j(t)$ of gene *j* (Elowitz et al., 2002) is defined by

$$\eta_j = \frac{S_j}{\bar{E}_j} \qquad \text{(Eq. 1)}$$

This operational definition of temporal noise, as the relative fluctuation of the expression level, is quite convenient due to the stochastic aspect of expression dynamics. Moreover, the intensive character of the variable η allows the clarification of its importance to expression dynamics. This definition of noise contemplates all possible contributions to transcriptional variability, from the influence of the feedback-based mechanisms of transcriptional regulation to intrinsic fluctuations related to the chemical reactions leading to mRNA synthesis. This definition also allows for theoretical predictions based on the knowledge of the equilibrium density of the expression level for each gene.

Another relevant parameter to be considered is the mean of the absolute value of the relative variation of the expression level, defined as:

$$\bar{\alpha}_j = \frac{1}{n-1}\sum_{l=1}^{n}\frac{\left|E_{l+1}^{(j)} - E_l^{(j)}\right|}{E_l^{(j)}} \qquad \text{(Eq. 2)}$$

where *j* is for each gene, *l* runs over the set of experimental measurements for each gene, and *n* is the total number of measurements *per* gene in the experiment. The relation between this parameter and the noise level is important in order to clarify the issue of transcriptional noise minimization.

RESULTS AND DISCUSSION

Genome-wide analysis

It is a widespread idea that the important issue of gene expression organization in the genome should be focused on the identification of gene networks related to biological functions. Therefore, the identification of gene subsets with special properties is appropriate to describe how genetic information is organized at the genome level in terms of dynamical observables. Here, we analyze the distribution of temporal transcriptional noise and the expression level at three different scales: the genome scale, the chromosomal scale and the sub-chromosomal scale, as determined by the centromere position (chromosome arms). The main goal is to identify the signature of scale invariance of relevant statistical observables at the three scales de-

fined above, which is obviously important for the organization of biological information of dynamical origin. Because we are looking at dynamical signatures we consider the role of temporal fluctuations in gene expression and, therefore, cell-to-cell variation, or populational noise, has to be smeared out so that the population variability does not mask the true temporal fluctuations (Figure 1). In other words, the microarray data and analysis we used were thus selected so that it could be said unequivocally that the observables that we considered are typical (statistically expected) of an individual randomly picked from that population.

At the genome scale both mean expression level (Figure 2A) and noise level (Figure 2B) follow a well-fitted lognormal distribution. Apart from the reasonable argument that the distribution of chemicals in the cell should assume different forms over the cell cycle, it is also reasonable to expect the emergence of lognormal (non-Gaussian) statistics in many cell processes. Due to chemical cascade effects which enhance the propagation of multiplicative fluctuations, the observed long-tail distribution of noise and mean expression reinforces the hypothesis of the lognormal distribution ubiquity in biological systems (Furusawa et al., 2005).

The distribution of the mean expression level has an algebraic tail with exponent $\gamma \sim 2$, whose universal character has been observed elsewhere (Ueda et al., 2004). The distribution of the noise level has an algebraic tail with exponent $\beta \sim 3$, also possibly universal.

Since $\bar{E}$ and η are lognormally distributed, the corresponding normally distributed variables are $\varepsilon = \log(\bar{E})$ and $\xi = \log(\eta)$. The mean values $\bar{\varepsilon}$ and $\bar{\xi}$ for the whole genome are $\bar{\varepsilon}$ = 5.529 (1.147) and $\bar{\xi}$ = -1.421 (0.425). Down to the chromosomal scale, we observed that the mean values of ε and ξ for each of the 16 chromosomes are all very close to the corresponding mean values for the whole genome (Table 1). The means of the individual chromosomes have a very small dispersion (<5%) around the mean value for the whole genome, $\bar{\varepsilon}_{chrm}$ = 5.485(0.281) and $\bar{\xi}_{chrm}$ = -1.408(0.057). Therefore, the geometric means of expression level and the noise level are invariant when rescaling the genome to the chromosome level.

Down to the subdivision of the chromosomes into left and right arms, we observed the same property. The $\bar{\varepsilon}$ and $\bar{\xi}$ for both arms are very close to the values for each chromosome (Table 2 and Figure 3). Therefore, the values of $\bar{\varepsilon}$ and $\bar{\xi}$ (which are related to the medians of $\bar{E}$ and η) are conserved from the genome to the sub-chromosomal scale. The conservation of $\bar{\varepsilon}$ and $\bar{\xi}$ over these two scales implies the possible existence of symmetry principles ruling the distribution of genes in each of the subsets (the chromosome and their corresponding arms). The question about the existence of biological relevant subsets (networks) of genes at even lower scales, obeying the same scaling property, remains to be further investigated. Based on the evidence presented, we conjecture that the location of the centromere is not randomly chosen, but is constrained to the physical distribution of $\bar{E}$ and η over the corresponding chromosome. Although very suggestive, the above results should be verified by new experiments. More accurate measurements of $\bar{E}$ and η as well as the complete identification of all coding regions of the genome should be done to corroborate the observed evidence of scale invariance of $\bar{\varepsilon}$ and $\bar{\xi}$.

The essentiality problem

Knowing the expression level and noise behaviors at the genome level, we considered the subsets of essential (Davierwala et al., 2005) and non-essential genes to determine whether they follow the same statistical and scaling properties of the whole genome.

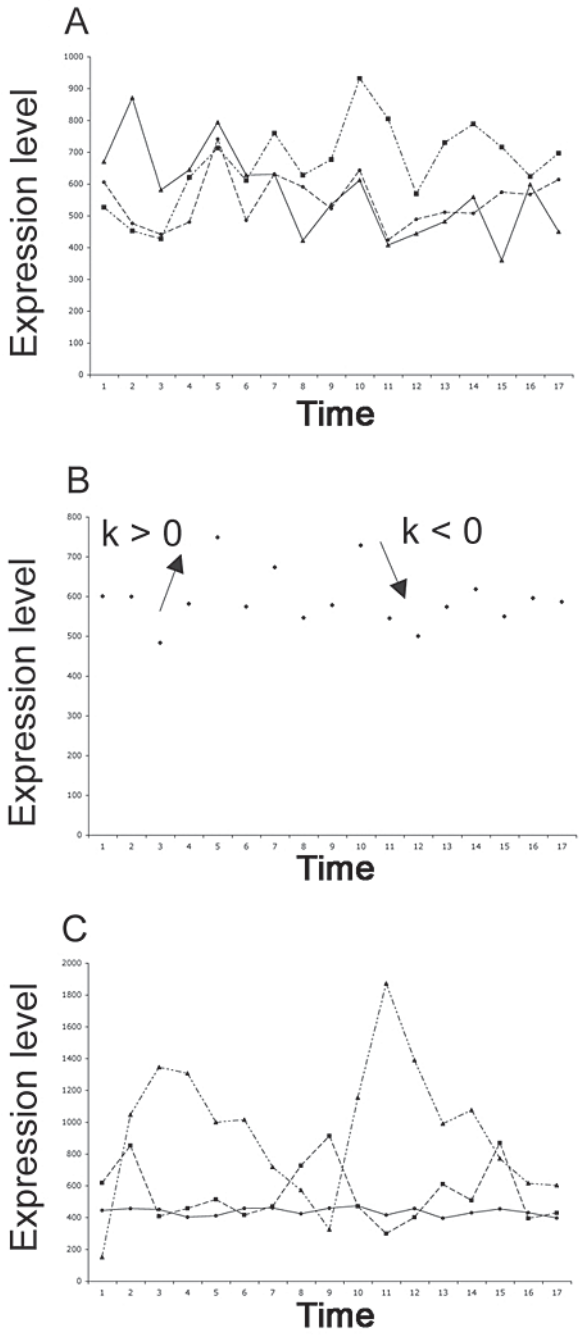

Figure 1. Types of noise. In **A**, we depict the general stochastic cell-to-cell variation of expression levels of a single gene in synchronized cells, or the populational noise. The symbols (triangles, squares and circles) indicate the different, individual cells in a hypothetical sample. In **B**, the temporal noise raw data, as considered in the present study, are an average of the synchronized population in **A** which smears out cell-to-cell variation, or populational noise, so that the population variability does not mask the true temporal fluctuations and allows saying unequivocally that the observables considered are typical (statistically expected) of an individual randomly picked from that population. K indicates the random variable of the temporal fluctuation (see Results and Discussion section Equation 3). In **C**, subtypes of temporal noise of three different genes, where triangles represent a typical cell cycle-regulated gene, squares a high temporal noise gene and circles a low temporal noise gene.

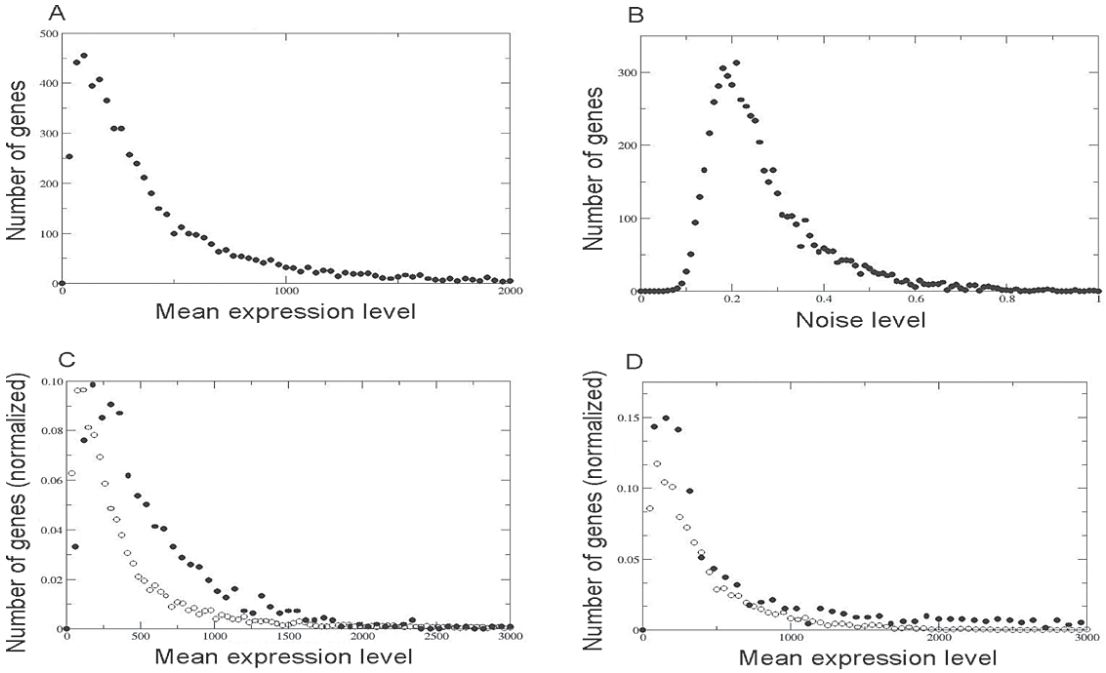

Figure 2. A. Genome-wide probability distribution for the mean expression level. **B.** Genome-wide probability distribution for the temporal noise level. **C.** Probability distribution of the mean expression level for the essential (full circles) and non-essential (empty circles) genes. **D.** Probability distribution of the mean expression level for the ohnologs (full circles) and non-ohnologs (empty circles) genes.

Table 1. Mean temporal noise level and mean expression level for the whole genome and for each chromosome.

	Noise (Ln)		Expression level (Ln)	
	Mean	SD	Mean	SD
Genome	-1.421	0.425	5.529	1.147
Chrm 1	-1.230	0.423	5.363	1.134
Chrm 2	-1.372	0.418	5.651	1.141
Chrm 3	-1.358	0.398	5.329	1.162
Chrm 4	-1.436	0.404	5.635	1.113
Chrm 5	-1.415	0.410	5.511	1.153
Chrm 6	-1.426	0.433	5.229	2.818
Chrm 7	-1.492	0.443	5.116	1.117
Chrm 8	-1.448	0.445	5.128	1.170
Chrm 9	-1.429	0.441	4.915	1.176
Chrm 10	-1.383	0.407	5.564	1.136
Chrm 11	-1.437	0.407	5.907	1.054
Chrm 12	-1.416	0.428	5.819	1.133
Chrm 13	-1.401	0.415	5.817	1.056
Chrm 14	-1.434	0.420	5.625	2.964
Chrm 15	-1.430	0.440	5.561	1.095
Chrm 16	-1.422	0.446	5.596	1.085

SD = standard deviation.

Table 2. Mean temporal noise level and mean expression level for both chromosome arms from each chromosome.

Chromosome	Chromosome arm	Noise (Ln)		Expression level (Ln)	
		Mean	SD	Mean	SD
1	L	-1.232	0.427	5.427	1.088
	R	-1.227	0.422	5.255	1.215
2	L	-1.359	0.426	5.864	1.032
	R	-1.378	0.416	5.569	0.172
3	L	-1.334	0.446	5.331	1.182
	R	-1.374	0.366	5.328	1.154
4	L	-1.407	0.454	5.604	1.166
	R	-1.450	0.379	5.649	1.089
5	L	-1.335	0.396	5.023	1.241
	R	-1.449	0.412	5.368	1.101
6	L	-1.450	0.418	4.962	1.249
	R	-1.397	0.468	5.092	1.059
7	L	-1.493	0.445	5.075	1.127
	R	-1.490	0.441	5.156	1.107
8	L	-1.417	0.488	4.857	1.321
	R	-1.455	0.436	5.188	1.128
9	L	-1.412	0.435	4.961	1.217
	R	-1.502	0.462	4.717	0.967
10	L	-1.360	0.402	5.383	1.176
	R	-1.416	0.412	5.819	1.028
11	L	-1.419	0.392	5.959	1.064
	R	-1.474	0.437	5.794	1.026
12	L	-1.363	0.388	5.644	1.131
	R	-1.424	0.434	5.846	1.133
13	L	-1.426	0.430	6.043	1.032
	R	-1.390	0.409	5.721	1.053
14	L	-1.425	0.424	5.657	1.079
	R	-1.473	0.373	5.477	1.065
15	L	-1.379	0.447	5.488	1.061
	R	-1.451	0.436	5.591	1.108
16	L	-1.422	0.457	5.581	1.107
	R	-1.423	0.433	5.618	1.056

L stands for the left arm, and *R* for the right arm. SD = standard deviation.

The probability density for the mean expression level of both essential and non-essential genes is fitted by lognormal distributions with the same tail given by the exponent $\gamma \sim 2$ (Figure 2C). The noise level of the same subsets also follows a lognormal distribution. Therefore, the subdivision of the genome into essential and non-essential genes preserves the basic statistical properties of the mean expression level and noise level of the whole genome. Nevertheless, the scale invariance observed at the chromosomal and sub-chromosomal levels is apparently not respected by the subdivision of the genome into essential ($\overline{\varepsilon}_{ess}$ = 5.932(0.936), $\overline{\xi}_{ess}$ = -1.564(0.394)) and non-essential genes ($\overline{\varepsilon}_{ness}$ = 5.437(1.148), $\overline{\xi}_{ness}$ = -1.398(0.427)). The

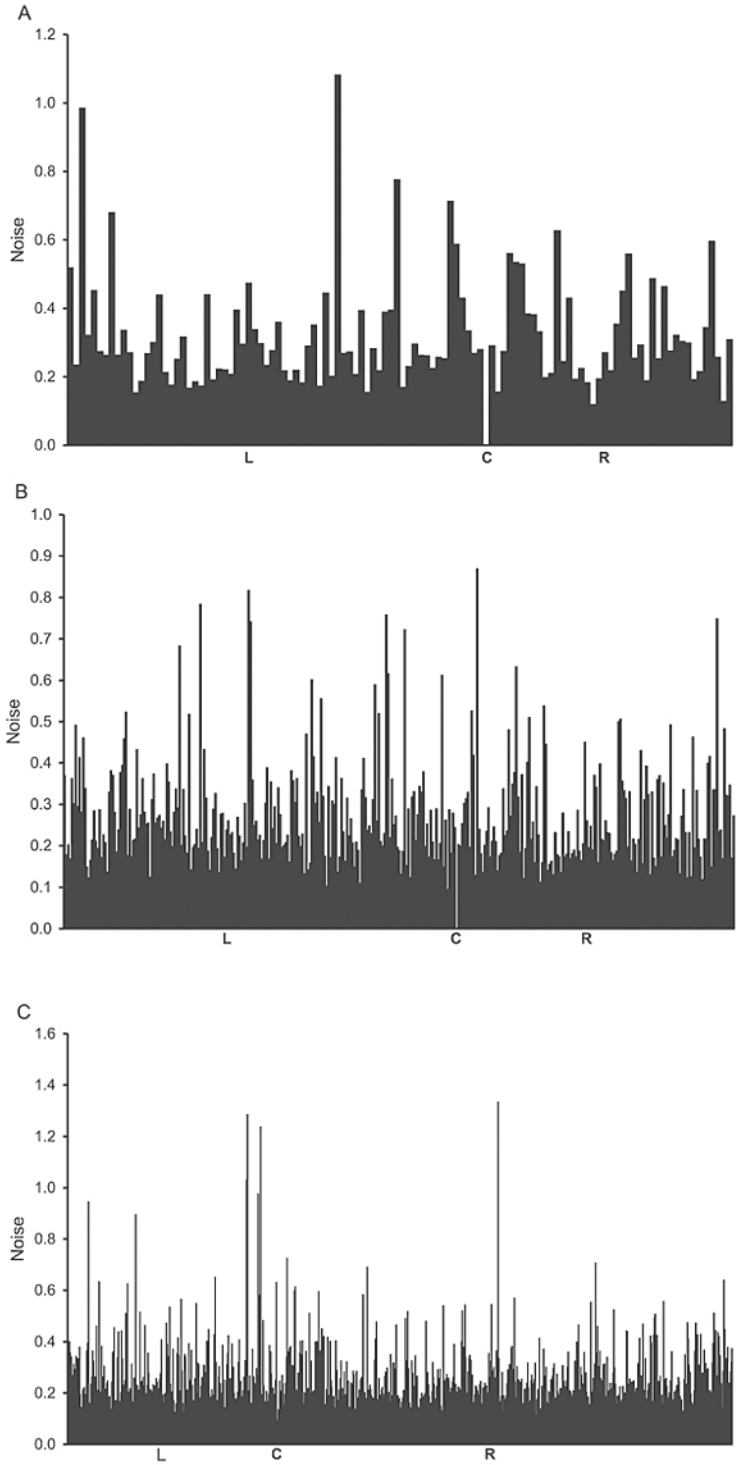

Figure 3. Chromosomal map of temporal noise. **A.** Chromosome 1 (230,208 bp). **B.** Chromosome 10 (745,745 bp). **C.** Chromosome 4 (1,531,918 bp). *C* indicates the centromere, and *L* is for the left arms and *R* for the right arms of chromosomes.

difference observed between the two gene sets is probably due to sample size effect. Note that the mean of the set of non-essential genes is very close to the genome mean in opposition to the set of essential genes.

The mean noise of essential genes (0.23) is lower than that of non-essential genes (0.27). This inequality is inverted if we consider the mean value of the gene expression for these gene subgroups (565 for essentials and 430 for non-essential genes). However, this distinction is not statistically significant (these two sets are almost surely indistinguishable). Individual noise levels of essential and non-essential genes are manifested differently and it is quite easy to identify in the data examples of the inequality $\eta_{ess} > \eta_{ness}$ with high statistical significance. This indicates that the inequality $\eta_{ess} < \eta_{ness}$ is only true in terms of mean values. Because the same holds for the chromosomal and sub-chromosomal scales, we hypothesize that a clear separation of essential and non-essential genes by noise minimization, and likely by mean expression maximization, is perhaps the case for genes belonging to specific networks (lower scale) of the genome, related to specific biological functions.

Ohnologs

Molecular evidence suggests that *S. cerevisiae* is a result of whole-genome duplication that occurred after the divergence of genus *Saccharomyces* from genus *Kluyveromyces*, approximately 10^8 years ago (Wolfe and Shields, 1997), which left approximately 17% of the *S. cerevisiae* genes as duplicates (ohnologs). The probability density of the mean expression level of the ohnolog and non-ohnolog subsets is well fitted by lognormal distributions (Figure 2D), although the tail has an exponent $\gamma = 1.38 \pm 0.15$ in the ohnolog subset, and $\gamma = 2.67 \pm 0.14$ in the non-ohnolog subset. The scale invariance observed at the chromosomal and sub-chromosomal levels is apparently not respected by the subdivision of the genome into ohnolog ($\bar{\varepsilon}_{ohn} = 5.799(1.277)$, $\bar{\xi}_{ohn} = -1.288(0.432)$) and non-ohnolog genes ($\bar{\varepsilon}_{nohn} = 5.465(1.104)$, $\bar{\xi}_{nohn} = -1.451(0.418)$). Therefore, the division of the genome into ohnolog and non-ohnolog subsets does not preserve the statistical properties of the whole genome, and does not follow the scaling properties described in the previous section. Therefore, it is not wrong to conclude that the natural principles involved in the selection of the genes that retained ohnologs versus the ones that lost their ohnologs are different from those involved in the assignment of essentiality.

Correlation between noise and codon adaptation index

The codon adaptation index (CAI) (Sharp and Li, 1987) was originally conceived to be a genetic parameter relating gene activity and protein abundance. This index is a source of fundamental biological information despite methodological considerations (Kliman et al., 2003; Jansen et al., 2003; Drummond et al., 2005).

Here, we explore the relation of CAI with transcriptional noise keeping in mind the importance of translation for transcriptional regulation (Fraser et al., 2004) and, therefore, for possible basic mechanisms of transcriptional noise minimization. Using the CAI to sort the genome, we observed that the genes are separated into subsets of constant CAI, with varying sizes (from ~10 to ~70 genes each) in the interval $0.100 < CAI < 0.185$, which encompasses ~70% of the genome (Figure 4A). Within each of these groups, we observe an internal order characterized by the relation $\bar{\alpha} = \eta$ (Figure 4B,C,D and E). Inside each group of genes, $\bar{\alpha}$ and

η follow the same kind of order. The range of variation is approximately the same with the minimal noise value close to the minimal noise value observed for the whole genome. The pattern found is very regular especially for CAI ~0.118. Outside of the interval 0.100 < CAI < 0.185, the grouping pattern is not obvious because of fast CAI variation. In this domain, the relation $\bar{\alpha} \approx \eta$ is only observed in isolated genes, for example, the genes encoding enzymes of the glycolytic pathway such as *TDH2* (glyceraldehyde-3-phosphate dehydrogenase), *CIT3* (citrate synthase), *PDA1* (pyruvate dehydrogenase) and *LPD1* (dihydrolipoamide dehydrogenase). Therefore, it is likely that the property $\bar{\alpha} \approx \eta$ should be valid for at least 70% of the genome; this would include, for example, the DNA integrity network (Pan et al., 2005). The identification of possible relations between the observed pattern organization and biological function is beyond the scope of the present study and is left for further investigation. Nevertheless, the significance of the property $\bar{\alpha} \approx \eta$ can be elucidated by a simple theoretical model of gene expression.

Modeling individual gene expression

In order to model the time variation of the expression of a single gene (E_j (t)) and to interpret the experimental results in view of theoretical predictions, we consider the expression

$$\frac{dE_j}{dt} = k_j E_j \qquad \text{(Eq. 3)}$$

where k_j is a random variable (normally distributed) that fluctuates around the mean value $\bar{k}$ = 0. This assumption is based on the estimated distribution of k_j obtained directly from the experimental data. Positive (negative) values of k_j describe an increase (decrease) of the expression level E_j (positive defined) whose logarithm varies randomly according to the same distribution of k_j. Therefore, one expects the random variable E_j to follow a lognormal distribution (Kaneko, 2003; Furusawa et al., 2005).

Equation 3 takes into consideration all causes that affect the expression level. It is tantamount to mean field-like approximation to the expression of gene *j*.

To facilitate the analyses of the model and comparison with the experiment, it is convenient to consider the discrete time version of Equation 3 with the parameter $\alpha_j = |k_j|$. In the first approximation the stochastic process $k_j(t)$ may be restricted to random jumps in time of the variable α_j between the values $+\bar{\alpha}_j$ and $-\bar{\alpha}_j$. Therefore, Equation 3 becomes:

$$\left| E_{n+1}^{(j)} - E_n^{(j)} \right| = \bar{\alpha}_j E_n^{(j)} \qquad \text{(Eq. 4)}$$

where n = 0, 1, 2, ..., (the discrete time variable). This may be written in the form of two discrete transformations

$$E_{n+1}^{(j)} = (1 + \bar{\alpha}_j) E_n^{(j)} \qquad \text{(Eq. 5)}$$

$$E_{n+1}^{(j)} = (1 - \bar{\alpha}_j) E_n^{(j)} \qquad \text{(Eq. 6)}$$

where $0 \leq E_n^{(j)} < \infty$.

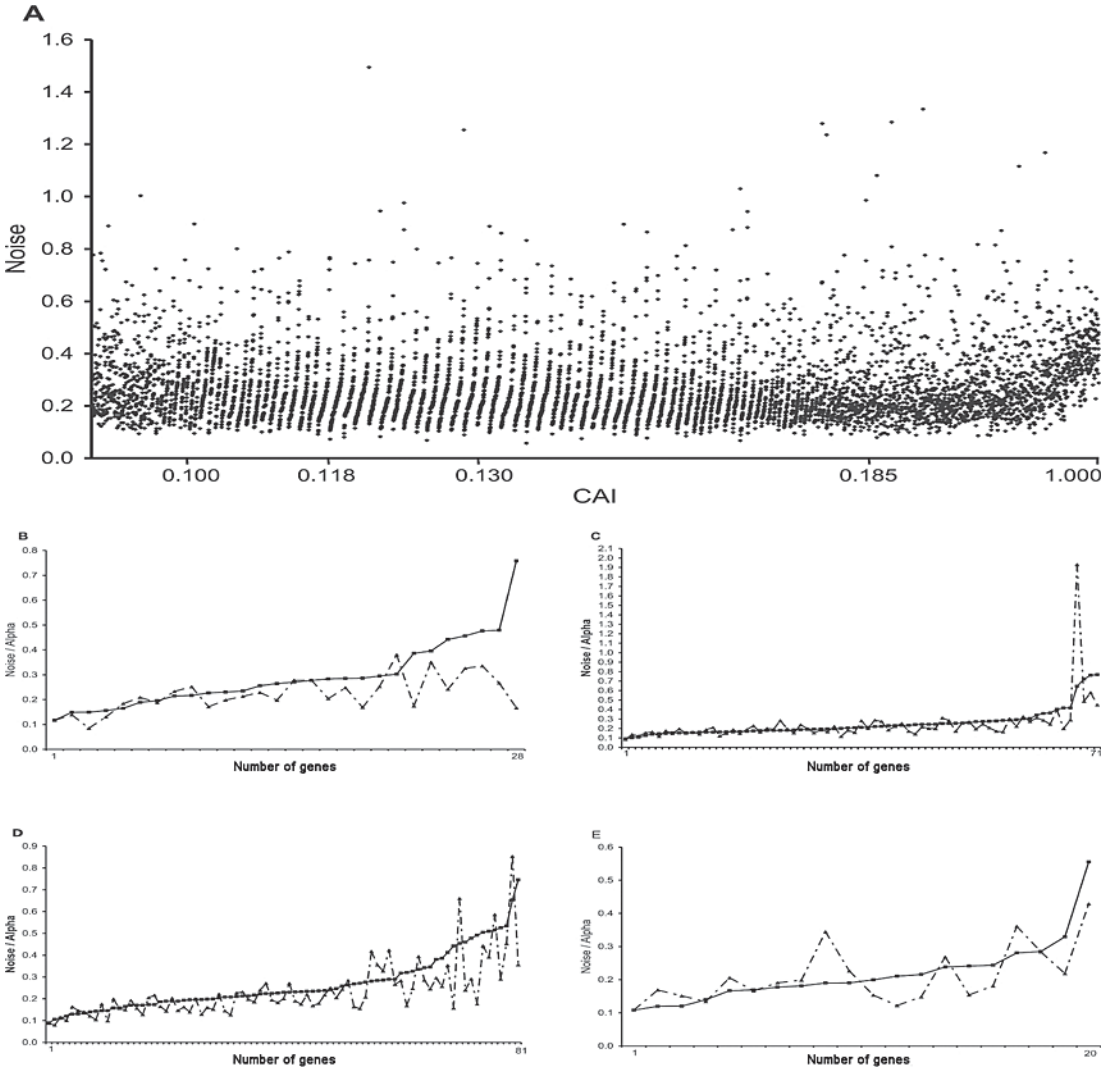

Figure 4. A. Correlation between temporal noise and codon adaptation index (CAI). The ORFs were sorted first by increasing noise level and second by increasing CAI value. A closer view of the pattern showing the noise level (squares) and the corresponding alpha (triangles) for each gene in a CAI interval is shown in **B** for CAI 0.100, **C** for CAI 0.118, **D** for CAI 0.130 and **E** for CAI 0.181.

According to Equations 5 and 6, the time evolution of gene expression is given by the time series $E_n^{(j)}$. At each discrete time step, Equation 5 or 6 is randomly chosen to update the value of the expression level $E_n^{(j)}$ to the value $E_{n+1}^{(j)}$.

It is easily observed from the experimental data that the values of $E_n^{(j)}$ typically vary between a minimal value ($E_m^{(j)}$) and a maximal value ($E_M^{(j)}$) during the cell cycle, characteristic of each gene. This suggests that the model (Equations 5 and 6) should be restricted to the interval ($E_m^{(j)}$, $E_M^{(j)}$). Therefore, depending on the domain of definition and the value of $\bar{\alpha}_j$, the model has three different cases to be considered:

$$(i)\ \frac{E_M^{(j)}}{E_m^{(j)}} > \frac{1+\bar{\alpha}_j}{1-\bar{\alpha}_j}.$$

In this case, the two linear transformations have a common domain in the interval

$$\left(\frac{E_m^{(j)}}{1-\bar{\alpha}_j} < E_n^{(j)} < \frac{E_M^{(j)}}{1+\bar{\alpha}_j} \right),$$

where the dynamics is explicitly probabilistic with equal *a priori* probabilities. Out of the common domain the dynamics is deterministic.

$$(ii)\ \frac{E_M^{(j)}}{E_m^{(j)}} < \frac{1+\bar{\alpha}_j}{1-\bar{\alpha}_j}.$$

In the interval

$$\left(\frac{E_M^{(j)}}{1+\bar{\alpha}_j} < E_n^{(j)} < \frac{E_m^{(j)}}{1-\bar{\alpha}_j} \right),$$

both transformations are not defined and this case, therefore, represents the absence of gene activity.

$$(iii)\ \frac{E_M^{(j)}}{E_m^{(j)}} = \frac{1+\bar{\alpha}_j}{1-\bar{\alpha}_j}.$$

This case defines a sharp border between activity and no activity of the gene. Here, gene expression is fully deterministic.

In the present analyses, it is assumed that the time variation of the expression level does not have to be the same for the same gene in different cells. At each time (t) the expression level of gene j may vary across the cell population. Nevertheless, the statistical properties of gene expression in time should be the same for gene j in any cell of the population (provided that all the individuals are subjected to the same environmental conditions). Therefore, that the probability density function of the variable $E_n^{(j)}$ should be the relevant dynamical signature of gene j.

It is important to stress that there is no proof of the existence of a stable probability density for $E_n^{(j)}$, for any gene of the genome. As a first step we assume its existence and eventually reformulate the hypothesis of statistical stability depending on the model predictions.

The above considerations imply that the model should be studied in terms of the time evolution of probability density. For each time step, there is a probability distribution of $E_n^{(j)}$, characteristic of the population of cells. The gene expression dynamics is given by the time

evolution of this probability distribution, which is fully described by the corresponding density evolution operator. In the present case, the appropriate operator is known as the Perron-Frobenius operator (Lasota and Mackey, 1998).

In formal terms, we write $\rho_n^{(j)}(E)$ (the density function at time *n* for gene *j*), *U* (the Perron-Frobenius operator for the model), such that $U\rho_n^{(j)} = \rho_{n+1}^{(j)}$. If statistical equilibrium exists, the equilibrium density is the solution of $U\rho_{eq}^{(j)} = \rho_{eq}^{(j)}$. If $\rho_{eq}^{(i)} = \rho_{eq}^{(j)}$ for two different genes, statistical equivalence of the two genes should be considered as far as gene expression is concerned.

Comparing the model with experimental data

This approach allows the analytical deduction of the border of gene expression considered in case *iii*. Since we calculated the values of η and $\bar{\alpha}$ of each gene from the experimental data, we analyzed the limit case and determined the curve η(α) from the model. The result is presented in Figure 5A.

By definition, the noise level is given by

$$\eta(\alpha) = \sqrt{\frac{\bar{E}^2(\alpha)}{\bar{E}^2(\alpha)} - 1} \qquad \text{(Eq. 7)}$$

As in case *iii*, the model is fully deterministic, the corresponding Perron-Frobenius operator can be easily obtained, and it gives the equilibrium density

$$\rho_{eq}^{(j)}(E) = \frac{1}{\log\left[\frac{1+\bar{\alpha}_{(j)}}{1 - \bar{\alpha}_{(j)}}\right]}\frac{1}{E} \qquad \text{(Eq. 8)}$$

With the help of Equation 8 one obtains

$$\eta(\alpha) = \sqrt{\frac{1}{2\alpha}\log\frac{1+\alpha}{1 - \alpha} - 1} \qquad \text{(Eq. 9)}$$

The curve described by Equation 9 is presented in Figure 5A along with the experimental data. The curve of Equation 9 separates the diagram η X $\bar{\alpha}$ into two regions: the left side of the diagram (referred to case *i*) and the right side of the diagram (case *ii*) where no points should be found. The experimental points that appeared in this region correspond to a) cell cycle-regulated genes such as *CLN1* (G1 cyclin 1), and *YOX1* (homeodomain-containing transcriptional repressor), b) ohnolog genes such as *KRR1* (essential nucleolar protein required for the synthesis of 18S rRNA and for the assembly of 40S ribosomal subunit), *SOL2* (role in tRNA export), and *CLN1* (G1 cyclin 1) and c) dubious and hypothetical ORFs such as SaYJL195c and YBR090c. Approximately 25% of the points consist of *a* and *b*. The remaining points likely correspond to data strongly affected by experimental error.

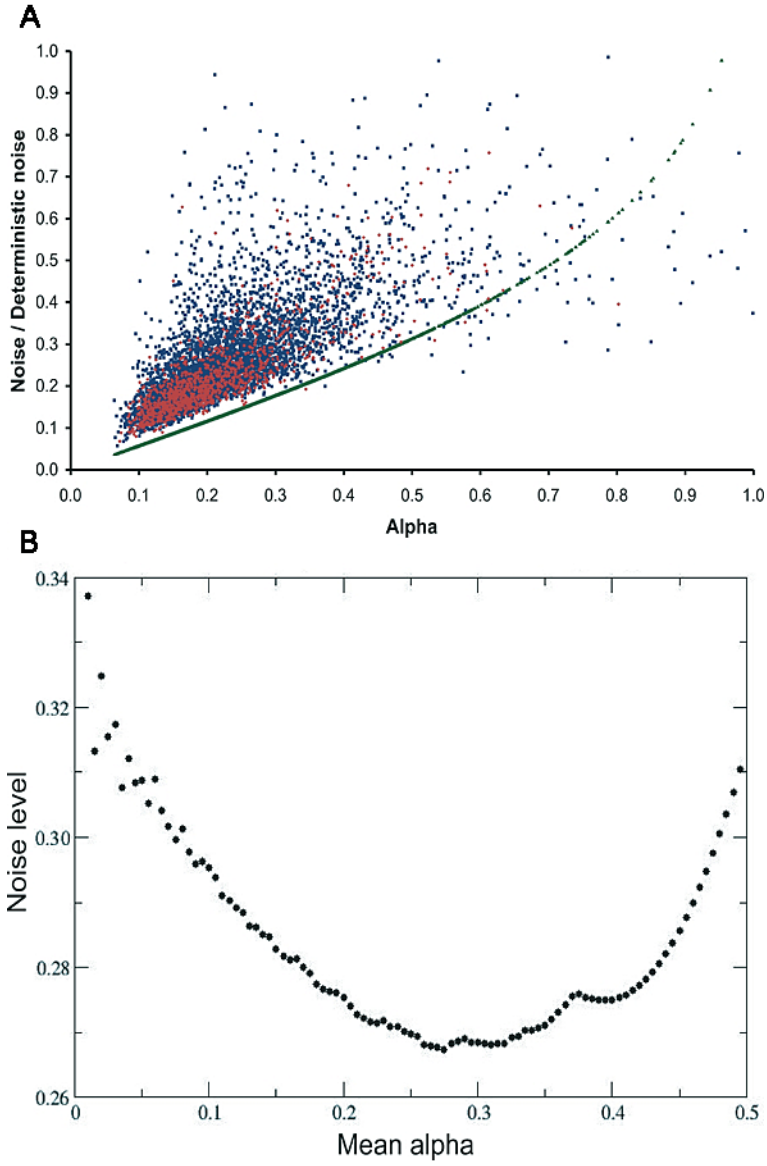

Figure 5. A. Diagram of $\bar{\alpha}$ X η. Essential (red) and non-essential (blue) genes are highlighted. The green curve refers to the case of deterministic dynamics and represents a boundary of gene activity. **B.** Noise minimization as a function of α for the ORF with a given minimal (E_m) and a maximal (E_M) expression level.

Equation 9 also indicates a restricted criterium for transcription noise minimization, namely: for any group of genes with similar values of $\bar{\alpha}$, minimal transcriptional noise is attained by those genes having the most deterministic gene expression dynamics.

Figure 5A also shows that the large majority of genes follow the expression dynamics described in case *i*. In this case, to obtain analytical results is far more difficult than in the case of fully deterministic dynamics, and its study is left for future investigation.

Nevertheless, the numerical simulations give relevant information about the organization of transcriptional noise at the genome scale. As described in the previous subsection, at least 70% of the genome (0.100 < CAI < 0.185) follows a clear pattern of groups of approximately 60 genes with regular variation of the noise level. As previously explained, $\bar{\alpha} \sim \eta$ holds for genes belonging to this group.

To investigate this property, we used the model to determine numerically the relation between $\bar{\alpha}$ and η for a fixed interval of the expression level variation. The result of the numerical simulations is presented in Figure 5B, and it clearly shows that noise is minimized for $\bar{\alpha} \sim \eta$. A similar curve is obtained for different intervals of the expression level variation.

Therefore, the property $\bar{\alpha} \sim \eta$ can be interpreted as a strategy used by ~70% of the genome to globally minimize the transcriptional noise and consequently minimize the deleterious effect of the stochastic component of gene expression. In these terms, the property $\bar{\alpha} \sim \eta$ may be seen as a natural strategy of coping with noise. It is important to stress that, as presented in our manuscript, we consider the role of temporal fluctuations in gene expression, and therefore, the populational noise, cell-to-cell variation or population variability does not mask the true temporal fluctuations. Although some may object to the use of microarray data for noise studies, we argue that the microarray data we used allow the unequivocal statement that the observables that we considered are typical (statistically expected) of an individual randomly picked from that population.

Our data suggest that there is a difference, at least in part, in the noise level of essential and non-essential genes, which may indicate the existence of an organizational order constrained by the maintenance of noise constancy and expression level. We hypothesize that genetic events that substantially affect this noise are drastic for cell viability, thus setting the boundaries of biological variability space upon which Darwinian selection will act (Kauffman, 1993). These would involve scale invariance and symmetry.

The analysis described here suggests a relation involving scale invariance and statistical equivalence with structural and functional organization of genetic information in terms of transcriptional noise and mean expression level. Scale invariance and statistical equivalence are properties related to the structural organization of the genome (the division of the genome into chromosomes and that of the chromosome arms). The existence of similar groups of genes at lower scales remains to be determined. Essentiality and ohnology are properties related to the functional organization of the genome. Following this principle the organization in terms of essential and non-essential genes would be, at least in part, structural, but the organization of the genome in terms of ohnolog and non-ohnolog genes would follow a strictly functional principle.

It can be suggested that the temporal transcriptional noise could be used as a binary classifier for gene function (a prospective task), using the well-known receiver operating characteristic curve (ROC), in contrast to the investigative analysis proposed in the present study. The problem of using transcriptional (population or temporal) noise as classifier for gene function encounters many difficulties and two of them certainly are: 1) the fact that transcriptional noise is a quantitative observable of the system and gene function is a qualitative property and 2) how to efficiently deal with massive amounts of data, typical in microarray experiments. To manage these difficulties, the use of ROC, which is comparable to supervised analysis, should be compared with the results of clustering methods, more often currently used. As an example, the use of the Super-paramagnetic Clustering (unsupervised) Method to identify cell cycle-regulated genes, using the same data we considered in our study, was reported by Getz et al.

(2000), showing good results when compared to the original works of Eisen et al. (1998), and Spellman et al. (1998). Nevertheless, the central goal of our study was the identification of dynamical signatures that provide evidence of the temporal organization of transcriptional information, its multiscale properties and its relation to special groups of genes of biological interest, such as essential genes, ohnologs and genes that have the same codon adaptation index. We conclude that in this perspective the mechanism of temporal transcriptional noise minimization, identified with the property $\bar{\alpha} \sim \eta$, is not related to functional or structural aspects. It would be an emergent property of dynamical nature related to the global network architecture of the genome.

ACKNOWLEDGMENTS

We thank Dr. Ronald W. Davis and Dr. Steve Oliver for excellent suggestions. R.C. Ferreira received a graduate fellowship from CNPq (Brazil). F. Bosco and M.R.S. Briones are supported by grants from FAPESP (Brazil) and M.R.S. Briones by an International Research Scholar grant from the Howard Hughes Medical Institute (USA).

REFERENCES

Adami C (2002). What is complexity? *Bioessays* 24: 1085-1094.

Adami C, Ofria C and Collier TC (2000). Evolution of biological complexity. *Proc. Natl. Acad. Sci. USA* 97: 4463-4468.

Anderson PW (1972). More is different. *Science* 177: 393-396.

Bar-Yam Y (2004). A mathematical theory of strong emergence using multiscale variety. *Complexity* 9: 15-24.

Blake WJ, Kaern M, Cantor CR and Collins JJ (2003). Noise in eukaryotic gene expression. *Nature* 422: 633-637.

Byrne KP and Wolfe KH (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15: 1456-1461.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2: 65-73.

Davierwala AP, Haynes J, Li Z, Brost RL, et al. (2005). The synthetic genetic interaction spectrum of essential genes. *Nat. Genet.* 37: 1147-1152.

De Wolf T and Holvoet T (2005). Emergence versus self-organisation: different concepts but promising when combined. In: Engineering self organising systems: methodologies and applications, lecture notes in computer science (Brueckner S, Serugendo GDM, Karageorgos A and Nagpal R, eds.). Lecture Notes in Computer Science, Springer-Verlag, New York, 1-15.

Drummond DA, Bloom JD, Adami C, Wilke CO, et al. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* 102: 14338-14343.

Eigen M (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58: 465-523.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.

Elowitz MB, Levine AJ, Siggia ED and Swain PS (2002). Stochastic gene expression in a single cell. *Science* 297: 1183-1186.

Fraser HB, Hirsh AE, Giaever G, Kumm J, et al. (2004). Noise minimization in eukaryotic gene expression. *Plos. Biol.* 2: 137.

Furusawa C, Suzuki D, Kashiwagi A, Yomo T, et al. (2005). Ubiquity of log normal distribution in intracellular reaction dynamics. *Biophysics* 1: 25-31.

Getz G, Levine E, Domany E, Zhang MQ (2000). Super-paramagnetic clustering of yeast gene expression profiles. *Physica* 279: 457-464.

Jansen R, Bussemaker HJ and Gerstein M (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in

yeast using a variety of models. *Nucleic Acids Res.* 31: 2242-2251.
Kaneko K (2003). Recursiveness, switching, and fluctuations in a replicating catalytic network. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 68: 031909.
Kauffman SA (1993). The origins of order: self-organization and selection in evolution. Oxford University Press, New York.
Kliman RM, Irving N and Santiago M (2003). Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* 57: 98-109.
Lasota A and Mackey MC (1998). Chaos fractals and noise: stochastic aspects of dynamics. Springer-Verlag, New York.
Nicolis G and Prigogine I (1977). Self-organization in non-equilibrium systems. Wiley-Interscience, New York.
Ozbudak EM, Thattai M, Kurtser I, Grossman AD, et al. (2002). Regulation of noise in the expression of a single gene. *Nat. Genet.* 31: 69-73.
Pan X, Ye P, Yuan DS, Wang X, et al. (2005). A DNA integrity network in the yeast *Saccaromyces cerevisiae*. *Cell* 124: 1069-1081.
Sharp PM and Li WH (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15: 1281-1295.
Spellman PT, Sherlock G, Zhang MQ, Iyer VR, et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273-3297.
Tong AH, Lesage G, Bader GD, Ding H, et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303: 808-813.
Ueda HR, Hayashi S, Matsuyama S, Yomo T, et al. (2004). Universality and flexibility in gene expression from bacteria to human. *Proc. Natl. Acad. Sci. USA* 101: 3765-3769.
Wolfe KH and Shields DC (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708-713.
Ycas M (1999). Codons and hypercycles. *Orig. Life Evol. Biosph.* 29: 95-108.