Journal of Biotechnology 149 (2010) 98–114

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Gene expression density profiles characterize modes of genomic regulation: Theory and experiment

Hans Binder[a,b,*], Henry Wirth[a,c], Jörg Galle[a,*]

[a] Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstr. 16-18, Germany
[b] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE) of University Leipzig, D-4103 Leipzig, Philipp-Rosenthal-Str. 27, Germany
[c] Helmholtz-Zentrum für Umweltforschung, D-04318 Leipzig, Permoserstr. 15, Germany

ARTICLE INFO

Article history:
Received 26 August 2009
Received in revised form 29 January 2010
Accepted 8 February 2010

Keywords:
Transcriptional regulation
Random genome model
Transcription factor network
Gene expression
Power law distribution

ABSTRACT

Our study addresses modes of genomic regulation and their characterization using the distribution of expression values. A simple model of transcriptional regulation is introduced to characterize the response of the global expression pattern to the changing properties of basal regulatory building blocks. Random genomes are generated which express and bind transcription factors according to the appearance of short motifs of coding and binding sequences. Regulation of transcriptional activity is described using a thermodynamic model. Our model predicts single-peaked distributions of expression values the flanks of which decay according to power laws. The characteristic exponent is inversely related to the product of the connectivity of the network times the regulatory strength of bound transcription factors. Such 'expression spectra' were calculated and analyzed for different model genomes. Information on structural properties and on the interactions of regulatory elements is used to build up a framework of basic characteristics of expression spectra. We analyze examples addressing different biological issues. Peak position and width of the experimental expression spectra vary with the biological context. We demonstrate that the study of the global expression pattern provides valuable information about transcriptional regulation which complements conventional searches for differentially expressed single genes.

1. Introduction

The results of microarray-based gene expression analyses, genome sequencing and other high throughput technologies have given us estimates of the complexity of molecular networks. There are tens of thousands of elements (e.g. genes, RNA-transcripts, proteins, metabolites) and at least as many connections between them subsumed as 'interactome'. How to understand the functional principles of such complex, highly structured and internally interacting systems? One option comprises top-down approaches of reverse engineering which attempt to reconstruct networks from high-dimensional experimental 'omics'-data via data fitting and parameter optimization of appropriate theoretical models. Alternatively, one can pursue a descriptive down-top approach which models the system and its functioning from first principles. Their reasonability should be tested by comparing the predictions of the models with experimental data. This modeling approach intentionally includes simplifications which enable to take aim at particular properties of the studied systems while ignoring less relevant ones.

The present study addresses transcriptional regulation in a simple model genome. This genome randomly generates and binds transcription factors forming a gene regulatory network. Our whole genome view is motivated by the idea that many aspects of gene functioning cannot be understood at the level of single genes but require a systemic approach which considers the manifold of an ensemble of genes, their possible microstates of activity and their mutual interactions. Primarily we are interested in characterizing the response of the global expression pattern to changing properties of basal regulatory building blocks. As starting point we used the Random genome model (RGM) approach introduced by Reil (1999). This model has been utilized to demonstrate emerging robustness against single base mutations and against random changes in initial network states as a consequence of stabilizing selection for a phenotype (Rohlf and Winkler, 2009). Comparable artificial genetic regulatory network models have been used to study a number of dynamic phenomena found in natural genetic networks such as heterochrony, evolution and stability (Banzhaf, 2003). Moreover, topological properties of these network models have been addressed such as the abundance of selected network motifs and subgraph distributions (Banzhaf and Dwight Kuo, 2004; Dwight Kuo et al., 2006).

Transcription in gene regulatory networks is based on biophysical processes which involve interactions on the molecular

* Corresponding authors. Tel.: +49 3419716671; fax: +49 3419716679.
E-mail addresses: binder@izbi.uni-leipzig.de (H. Binder), galle@izbi.uni-leipzig.de (J. Galle).

0168-1656/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2010.02.002

level between DNA, transcription factors and enzymes such as RNA-polymerase. We therefore complement the RGM with a thermodynamic model of transcriptional regulation adapted from Bintu et al. (2005). It allows tuning gene expression by the regulatory action of transcription factors.

Our whole genome approach is also motivated by the technological development of RNA analytics in the 'post-genomic era' which enables to measure global gene expression pattern of a series of organisms in single experiments. Particularly microarray studies have revealed the complex nature of the large-scale organization of gene expression. A common result observed in different analyses is that the distribution of gene expression values seems to exhibit a broad-tail that is characterized by a power law (Furusawa and Kaneko, 2003; Hoyle et al., 2002; Ueda et al., 2004). This power law distribution has been interpreted in terms of general principles such as 'proportional dynamics' (Ueda et al., 2004), the optimization of self-reproduction (Furusawa and Kaneko, 2003) and stochastic, noise-driven dynamics of transcription (Nacher and Akutsu, 2006).

Our model generates a power law distribution of gene expression as well, however with increasing and decreasing tails for transcriptional repression and activation, respectively. We analyzed these spectrum-like distributions for special situations to characterize the properties of the RGM in terms of simple rules which reflect different modes of transcriptional regulation. In the experimental part of this paper we apply these rules to experimental expression spectra which were calculated for a series of microarray measurements taken from public data repositories. These examples comprise different biological issues and samples ranging from embryonic development and cell differentiation to mutants and oncogenic de-regulation in different tissues and organisms. The chosen examples show that characterization of the global expression pattern in terms of the distribution of expression values provides valuable information about transcriptional regulation which complements conventional searches for differentially expressed genes.

2. Theory

2.1. The random genome model (RGM)

Following previous work (Reil, 1999; Rohlf and Winkler, 2009), the construction of the random genome of size L_{genome} comprises the following steps (see Fig. 1 for illustration and Table 1 for definitions): first, a random string of length L_{genome} is generated using four digits [0,1,2,3] for each position. This choice provides correspondence with the ATGC alphabet of real genomes. Secondly, a *promoter sequence* of length L_{prom} is defined to specify the position of the genes in the genome; for example the sequence motif (01010). Thirdly, L_{cod} digits downstream of the promoter sequence are selected that represent the *coding region* of the gene. The L_{reg} digits upstream of the promoter sequence up to the coding region of the preceding gene define the *regulatory region* of the gene.

The random genome thus decomposes into elementary regulatory building blocks or 'genes'. Each of them comprises a regulatory region, a promoter and a coding region in downstream direction. The minimum length of a gene is the sum of L_{prom} and L_{cod}. We exclude overlapping genes assuming that the distance between the start points of two adjacent promoter regions is always larger or equal to the minimum gene length. Our gene definition is in accordance with previous approaches such as the finite state linear model (Schlitt and Brazma, 2006). Each coding region can produce transcripts which subsequently translate into *transcription factors* (TF). The binding motif of the TF is generated from the respective coding motif by counting up the sequence code by one using the rule $i \rightarrow (i+1)\text{mod}[3]$ for each digit of the coding motif, for example

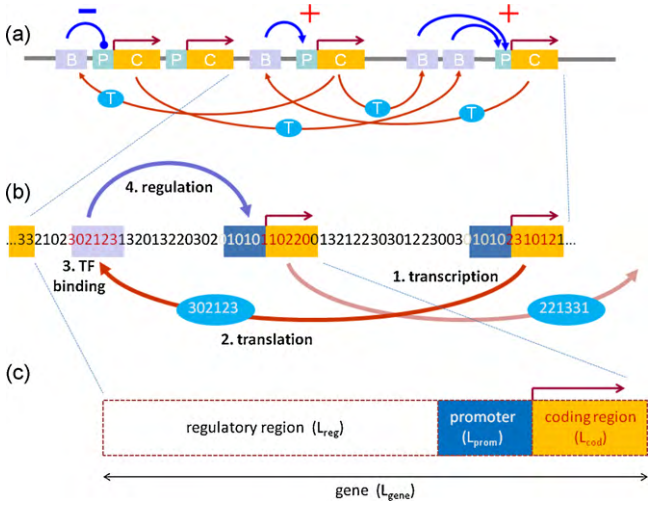

Fig. 1. The random genome model: (a) and (b) the random genome consists of a sequence of 'genes' (see also panel c) the number and position of which is determined by the realization of the promoter motif (P) along a sequence of randomly generated nucleotides $\in$ [0,1,2,3] = [A,C,G,T]. The coding region (C) defines the length and sequence of the transcripts (T) which translate into the transcription factor using the rule $i = (i+1)\_\text{mod}\_3$. This 'translated' sequence motifs bind to identical motifs (B) in the genome. The bound TF regulate the occupancy of the next downstream promoter (alternatively by repression '–' or activation '+') which, in turn, determines the transcriptional rate of the adjacent coding region. (c) Each 'gene' consists of a regulatory region, the promoter and the coding region of length L_{reg}, L_{prom} and L_{cod}, respectively.

$(231012) \rightarrow (302123)$. Each TF binds to motifs of identical sequence in the regulatory regions of the genome. *Bound TF* regulate the occupancy of the nearest downstream promoter by a thermodynamic interaction model (see below). The occupancy is defined as the probability of the promoter to bind RNA-polymerase. It governs transcription of the adjacent coding region and thus the *expression of the respective gene*. We assume direct proportionality between promoter occupancy and gene expression. The cumulative promoter occupancy of all genes consequently determines the total expression level of the artificial genome.

2.2. Statistical properties of the RGM

The generation rules of the RGM use three length parameters (L_{genome}, L_{prom} and L_{cod}). They determine the basic size relations of the random genome: the mean length of the genes and of the regulatory regions, the mean number of genes forming the genomes and the average number of TF which bind to the genes (see Table 1).

The probability that a given position of the genome is a starting point of a promoter sequence, is exclusively defined by the length of the promoter motif. It determines the mean length of the genes (Eq. (1) in Table 1 and Fig. 3a) and the mean number of genes (Eq. (2)). Analogously, the mean separation distance between two binding sites of a TF depends exclusively on the length of the coding region L_{cod} (Eq. (3)) which together with the length of the genome determines the mean number of binding sites per TF in the genome (Eq. (4)).

Fig. 2 shows the network of interacting genes for one realization of the RGM: 'OUT-bound' and 'IN-bound' connections refer to transcription and TF binding, respectively as illustrated in direction of the arrows. Clearly, the number of IN- and OUT-bound edges varies from gene to gene. We derived analytical solutions for three important probability distributions comprising the distribution: (i) of the length of a regulatory region, (ii) of the number of binding sites per TF in the genome (OUT-degree distribution) and (iii) of the number of TF-binding sites within the regulatory region of one gene (IN-degree distribution).

Table 1
The random genome model: glossary of symbols, size and thermodynamic relations and values used.

	Symbols and equations	Values	Eq. no.
Length parameters			
Length of the genome	L_{genome}	10^5	
Length of the promoter	L_{prom}	5	
Length of the coding region	L_{cod}	7	
Size relations			
Mean length of the gene	$\langle L_{gene}\rangle = 4^{L_{prom}}$	~1000	(1)
Mean number of genes in the genome	$\langle N_{gene}\rangle = \frac{L_{genome}}{\langle L_{gene}\rangle}$	~100	(2)
Mean distance between two binding sites of a TF	$\langle L_{bind}\rangle = 4^{L_{cod}}$	~1000	(3)
Mean number of TF-binding sites per gene	$\langle N_{bind}\rangle = \frac{L_{genome}}{\langle L_{bind}\rangle} \approx \frac{L_{genome}}{4^{L_{cod}}}$	~6	(4)
Thermodynamic parameters			
Ratio of the number of RNAP/TF molecules to the number of non-specific binding sites of RNAP and TF, respectively	$r_{RNAP/ns} = \frac{N_{RNAP}}{N_{ns}}$	10^{-4}	
	$r_{TF/ns} = \frac{N_{TF}}{N_{ns}}$	10^{-4}	
Standard free energy: increment of specific binding of RNAP and TF relative to non-specific binding, respectively	ε_{RNAP} in units of −kT	2	
	ε_{TF} in units of −kT	14	
Regulation free energy: increment of the standard free energy of specific RNAP binding induced by bound TF	ε_r in units of −kT	±1	
Maximum value of the binding activity of TF	$X_{TF}^{\max}$	1000	
Thermodynamic relations			
Basal binding activity of RNAP	$X_0 = r_{RNAP/ns} \cdot \exp(\varepsilon_{RNAP})$	~0.001	(5)
Binding activity of TF	$X_{TF} = r_{TF/ns} \cdot \exp(\varepsilon_{TF})$	0…1000	(6)
Regulation factor	$F = \frac{1+X_{TF}\cdot\exp(\varepsilon_r)}{1+X_{TF}}$		(7)
Promoter occupancy (probability that RNAP binds to the promoter)	$\Theta = \frac{F\cdot X_0}{1+F\cdot X_0}$	0…1	(8)
Scaling condition, defines the expression level of each gene	$X_{TF} \equiv X_{TF}^{\max}\cdot\Theta$		(9)

We assume that the genome length largely exceeds the mean gene length and that the mean gene length is large compared to the minimum gene length:

$$L_{genome} \gg \langle L_{gene}\rangle \gg L_{prom} + L_{cod} \tag{10}$$

Condition (10) implies artificial genomes which comprise a large number of genes and ensures that the mean length of the regulatory region of a gene can be approximated by the mean length of a gene, i.e. $\langle L_{reg}\rangle \approx \langle L_{gene}\rangle$ (see also Fig. 3a).

The individual lengths of the regulatory regions and also the numbers of TF-binding sites within each regulatory region vary from gene to gene and depend on the particular realization of the genome. Particularly, the length of the regulatory regions can vary within the limits $L_{genome} \geq L_{reg} \geq 0$. The probability distribution of L_{reg} is approximately given by the conditional probability to find the next promoter motif along the random genome sequence after $L_{reg}+1$ positions (Rohlf and Winkler, 2009):

$$w_{reg}(L_{reg}) \approx 4^{-L_{prom}}\cdot(1-4^{-L_{prom}})^{L_{reg}} = (1-e^{-1/k_0})\cdot\exp\left(\frac{-L_{reg}}{L_0}\right)$$
$$\text{with}\quad L_0 = \frac{-1}{\ln(1-4^{-L_{prom}})^{L_{reg}}} \approx 4^{L_{prom}} = \langle L_{gene}\rangle \tag{11}$$

Eq. (11) defines an exponential decay with L_{reg} where the decay length L_0 is given to a good approximation by the mean gene length (see Fig. 3b).

The *OUT-degree distribution* can be approximated by the probability distribution to find k binding sites of a TF (binding length L_{cod}) in the independent regulatory regions of a number of $\langle N_{gene}\rangle$ genes each of them of average length $\langle L_{reg}\rangle$. It is given by the binomial distribution:

$$w_{out}(k) = w_{out}(\langle L_{reg}\rangle, k) \approx \binom{\langle L_{reg}\rangle}{k}\cdot p^k\cdot(1-p)^{\langle L_{reg}\rangle - k}$$
$$\text{with}\quad p = \frac{\langle N_{gene}\rangle}{\langle L_{bind}\rangle} = \langle N_{gene}\rangle\cdot 4^{-L_{cod}} \tag{12}$$

The mean value of k averaged over $w_{out}(k)$ provides the mean number of binding sites per TF in the genome, $\langle N_{bind}\rangle = p\cdot\langle L_{reg}\rangle$ (Eq. (4)).

Typically one gets $p \ll 1$. This transforms the out-degree distribution into a Poissonian one to a good approximation: $w_{out}(k) \approx \langle N_{bind}\rangle^k\cdot\exp(-\langle N_{bind}\rangle)/k!$ (Rohlf and Winkler, 2009).

The *IN-degree distribution* is calculated as the weighted mean of $w_{out}(L_{reg}, k)$ averaged over all possible lengths L_{reg}

$$w_{in}(k) \approx \sum_{L_{reg}=0}^{L_{genome}} w_{reg}(L_{reg})\cdot w_{out}(L_{reg}, k) \approx (1-e^{-1/k_0})\cdot\exp\left(\frac{-k}{k_0}\right)$$
$$\text{with}\quad k_0 = \frac{1}{\ln(1+\langle N_{bind}\rangle^{-1})} \approx \langle N_{bind}\rangle \tag{13}$$

Eq. (13) defines an exponential decay with a characteristic decay number k_0. For $\langle N_{bind}\rangle \gg 1$ the decay number can be approximated by the mean number of bound TF (see Fig. 3b). The mean value of k averaged over the IN-degree distribution provides the mean number of TF-binding sites per regulatory region. This number has the same value $\langle N_{bind}\rangle$ as the mean number of binding sites per TF within the genome according to the condition of material balance. Hence, $\langle N_{bind}\rangle$ represents a measure of the intrinsic connectivity of the regulatory network. Representative OUT- and the IN-degree

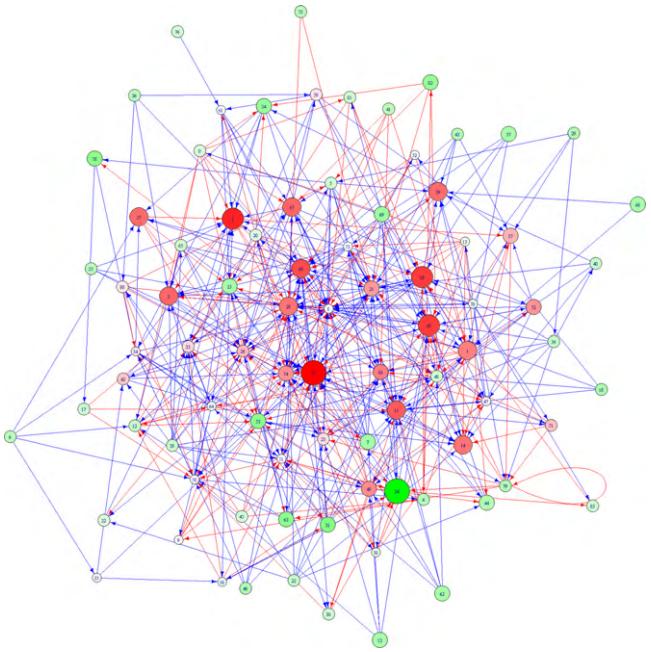

Fig. 2. Network of 84 interacting genes produced by one particular realization of the random genome of size $L_{genome} = 10^5$. The nodes are the genes. The arrows point in direction from transcription to binding of TF. The number of OUT-bound and IN-bound connections varies from gene to gene. Their distribution is given by the OUT-degree and IN-degree distributions, respectively. The IN-bound events regulate the expression of the respective gene using a thermodynamic model (see text). Activating and repressing interactions are colored in red and blue, respectively. They combine for each gene giving rise to activated (red) or repressed (green) promoter activity. The size of the circles scales with the degree of activation or repression relative to the basal expression rate. The nodes are ordered inwards with increasing number of IN-bound connections and thus with increasing degree of transcriptional regulation. Loops refer to auto-regulation.

distributions are shown in Fig. 3c. The dependence of the average number of binding sites per TF in the genome on the coding length is illustrated in Fig. 3d.

The promoter length used in our RGM realization ($L_{prom} = 5$) defines a mean gene length of $\langle L_{gene} \rangle = 10^3$ which is comparable with the mean gene length of prokaryotes (Koonin and Wolf, 2008). The fraction of gene products dedicated to gene regulation in prokaryotes ranges from less than 1% to about 10% (see Pérez-Rueda et al., 2004 and references cited therein) where the percentage scales approximately quadratically with the genome size (Maslov et al., 2009). For total genome lengths of $(2–6) \times 10^6$ one gets $(0.2–6) \times 10^5$ for the size of the respective subnets of genes. This, in turn, covers the length of the RGM used. The size of the RGM in our study consequently falls into the typical range of the TF subnet of prokaryotes.

Increasing complexity during evolution is related to changes of the organization and regulation of the genome and transcriptome such as the massive increase of the amount of non-coding RNA and the larger fragmentation of the genes into introns and exons (Mattick, 2007; Taft et al., 2007). This goes along with the appearance and increasing complexity of 'epigenetic' mechanisms such as chromatin remodeling (Meissner et al., 2008; Mikkelsen et al., 2007; Simon and Kingston, 2009) and also with the strong increase of proteome and interactome sizes (Stumpf et al., 2008). On the other hand, the size of the protein coding sequence per haplotype only weakly varies (Mattick, 2007). Thus, our RGM can be understood as a model of the TF sub-network also in higher organisms.

2.3. Thermodynamics of genomic regulation

Each promoter initiates transcription of the coding motif via specific binding of RNA-polymerase (RNAP). This process is modulated by TF binding in the regulatory region upstream of the promoter. Both processes are taken into account using a thermodynamic model. In agreement with other thermodynamic models for transcription regulation (Bintu et al., 2005; Lassig, 2007; Segal et al., 2008) we postulate that the expression level directed by each promoter is proportional to the promoter occupancy which is defined as the probability that RNAP occupies the promoter sequence. We also assume that the system is in thermodynamic equilibrium, such that each regulatory state is achieved with probability proportional to the Boltzmann distribution. This assumption is justified in cases where the transcription rate is slower than the rate at which transcription factors and polymerase bind and unbind the DNA. Accordingly, stochastic fluctuations of expression on short time scale are assumed leveled out.

Following the thermodynamic model of Bintu et al. (2005) the promoter occupancy Θ is governed by the following factors (for illustration see Fig. 4 and Table 1 for definitions and equations):

(i) The basal RNAP binding activity X_0 (Eq. (5)). It was derived assuming a binding equilibrium of free and bound RNAP molecules where the latter ones distribute between non-specific and specific binding sites. X_0 is governed by the number of available RNAP molecules, the number of non-specific RNAP binding sites and the standard free energy increment upon specific binding of RNAP compared with non-specific binding. Eq. (5) assumes excess of non-specific sites, $N_{ns} \gg N_{RNAP}$ (see Bintu et al., 2005 for details). The RNAP binding affinity defines the basal promoter occupancy in the absence of regulators (see Fig. 4a).

(ii) The binding activity of the transcription factors X_{TF} (Eq. (6)) which refers to the binding equilibrium of free and bound TF where the latter ones distribute between non-specific and specific binding sites in analogy with RNAP binding. X_{TF} is governed by the number of available and thus expressed transcription factors, by the number of the non-specific TF-binding sites and by the standard free energy increment upon specific binding of TFs to the DNA compared with non-specific binding. We assume that TF binding is strongly driven by the free energy ($\varepsilon_{TF} \approx 14$) compared to RNAP binding ($\varepsilon_{RNAP} \approx 2$; see Table 1 and Lassig, 2007).

(iii) The regulation free energy which is specified by the change of the free energy increment of specific RNAP binding induced after binding of TF to the regulatory region. It either activates ($\varepsilon_R > 0$) or represses ($\varepsilon_R < 0$) specific RNAP binding.

The total effect of TF binding on the promoter occupancy depends on the TF-binding activity (ii) and the regulation term (iii) as well. It is considered using the regulation factor approach introduced by Bintu et al. (2005) (see Eqs. (6) and (7) and Fig. 4a). Regulation modifies the recruitment of RNAP to the promoter which is equivalent with the change of the binding activity of RNAP. Repressors and activators are characterized by the sign of the regulation free energy as suggested by Lassig (2007), see (iii) above. In consequence the occupancy of the promoter is increased (regulation factor $F > 1$) or decreased ($F < 1$) compared with the basal level ($F = 1$).

Eq. (7) provides the regulation factor for the special case of single regulators acting on the particular promoter. For the more general case of k mutual independent regulators one gets the total regula-

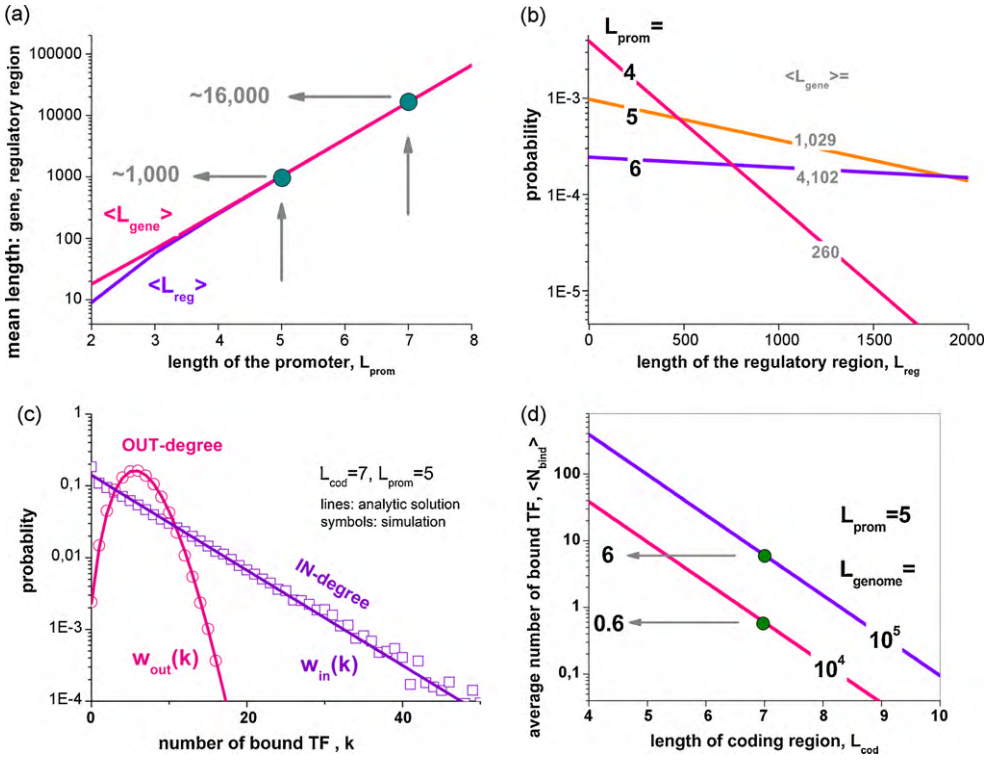

Fig. 3. Statistical properties of the RGM: (a) the mean length of the genes and of their regulatory region increase exponentially with the length of the promoter motif. The arrows indicate the mean gene length for $L_{prom} = 5$ and 7. The difference between $\langle L_{gene}\rangle$ and $\langle L_{reg}\rangle$ can be neglected for $L_{prom} > 3$. (b) Distribution of the length of the regulatory region $w_{reg}(L_{reg})$ in dependence of L_{prom}. (c) IN-degree $w_{in}(k)$ and OUT-degree $w_{out}(k)$ distributions for a genome of size $L_{genome} = 10^5$. $w_{out}(k)$ follows a binominal distribution (Eq. (12)) whereas $w_{in}(k)$ is an exponential decay (Eq. (13)). (d) The mean number of TF-binding sites per gene in the genome is determined by the genome size and the length of the promoter motif (Eq. (4)). $\langle N_{bind}\rangle$ characterizes the connectivity of the genome.

tion factor as the product of the individual ones (Lassig, 2007):

$$F(k) = \prod_{i=1}^{k} F(X_{TF}^i, \varepsilon_r^i) \quad \text{with} \quad F(X_{TF}^i, \varepsilon_r^i) = \left(\frac{1 + X_{TF}^i \exp(\varepsilon_r^i)}{1 + X_{TF}^i}\right) \tag{14}$$

Each individual regulation factor $F(X_{TF}^i, \varepsilon_r^i)$ considers the effect that exerts one bound TF on the promoter occupancy. It is governed by the binding activity and regulation free energy of the ith TF (see Table 1). The total regulation factor $F(k)$ then substitutes the individual one in Eq. (8) to get the regulated promoter occupancy as a function of the number of regulators k:

$$\Theta_{gene}(k) = \frac{F(k) \cdot X_0}{1 + F(k) \cdot X_0} \tag{15}$$

Eqs. (14) and (15) describe the regulation of one particular gene by k arbitrary regulators where each of them is characterized by its individual regulation free energy and TF activity.

Let us assume that the k regulators per promoter split into j repressors and $(k - j)$ activators which act both with the same absolute value of the regulation free energy ε_r. Eqs. (14) and (15) rewrite for this particular case into

$$\Theta_{gene}(k, j) = \frac{F(k, j) \cdot X_0}{1 + F(k, j) \cdot X_0} \quad \text{and}$$

$$F(k, j) = F(\langle X_{TF}\rangle, -\varepsilon_r)^j \cdot F(\langle X_{TF}\rangle, +\varepsilon_r)^{k-j} \tag{16}$$

The mean promoter occupancy averaged over all genes with k regulators is

$$\Theta(k) = \sum_j p(k, j) \cdot \Theta_{gene}(k, j) \tag{17}$$

where $p(k, j)$ is the probability that a gene is regulated by j repressors and $k - j$ activators.

We assume direct proportionality between promoter occupancy and gene expression which is scaled in dimensionless units of *TF activity* X_{TF} (see Eq. (6) in Table 1). The proportionality constant X_{TF}^{max} defines the maximum possible expression referring to maximum promoter occupancy $\Theta = 1$ (Eq. (9)).

It implies steady state of TF expression and degradation in the genome. Note that X_{TF} is proportional to the number of transcribed TF (Eq. (6)).

2.4. Mean expression approximation

The intrinsic structure and connectivity of a particular random genome is characterized by the size relations (Table 1) and by the OUT- and IN-degree distributions. The analytical expressions (Eqs. (12) and (13)) are approximations referring to condition (10). To validate these approximations we compare the analytic IN- and OUT-degree distributions with simulated ones (panel c of Fig. 3). For this purpose random genomes were generated according to the rules given in Section 2.1. The respective IN- and OUT-degree distributions are then calculated and averaged over 1000 independent realizations of the genome. The comparison of analytic and simulated distributions provides good agreement which justifies the used approximations.

For calculating the expression values in a particular genome it is important to recall that the regulation factor is a function of the number of transcribed TF (see Eq. (6)) which, in turn, is determined by the promoter occupancies of the genes that transcribe the respective TF (Eq. (9)). Hence, the regulation of the genes determines their rate of transcription and vice versa.

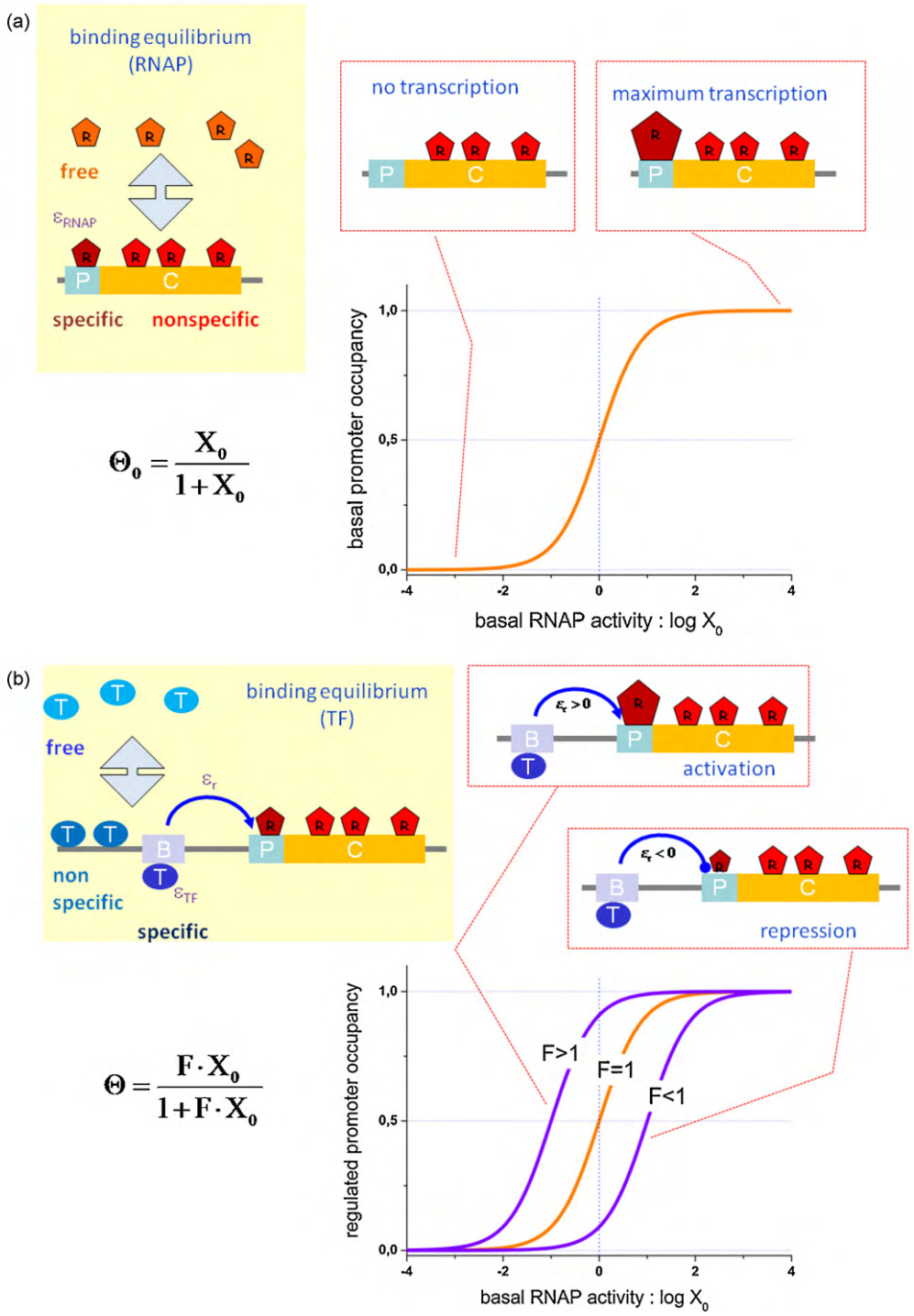

Fig. 4. Thermodynamics of genomic regulation: (a) RNAP (RNA-polymerase, R) distributes between free RNAP in solution and RNAP bound to specific and non-specific binding sites. Specific binding is accompanied with a free energy increment of ε_{RNAP} (in units of thermal energy, kT). The basal promoter occupancy, Θ_0, defines the probability that a RNA-polymerase (R) molecule specifically binds to the promoter (P) and initiates transcription. The basal promoter occupancy switches between 0 and 1 with an inflection point given by the basal RNAP binding activity $X_0 = 1$ (i.e. $\log X_0 = 0$). (b) The promoter occupancy is regulated by transcription factors (T) which bind to TF-binding sites (B) in the regulatory region according to a binding equilibrium of free and bound (specifically and non-specifically) TF molecules (ε_{TF} is the free energy increment). TF binding changes the binding free energy of RNAP to the promoter by ε_r which, in turn, affects the promoter occupancy. In consequence the inflection point of the promoter occupancy shifts towards smaller or larger values of the RNAP binding activity for activators and repressors, respectively.

We applied a *mean expression approximation* to obtain a self-consistent solution of this feedback problem in the regulatory network generated by the RGM: Accordingly, the promoter occupancies of all genes are pooled into one mean occupancy level of the considered genome. This step is equivalent with calculating the weighted sum of the occupancy Eq. (15) over the number of regulators k with the in-degree distribution Eq. (13) as weighting factor:

$$\langle \Theta \rangle = \sum_j w_{in}(k) \cdot \Theta(k) \tag{18}$$

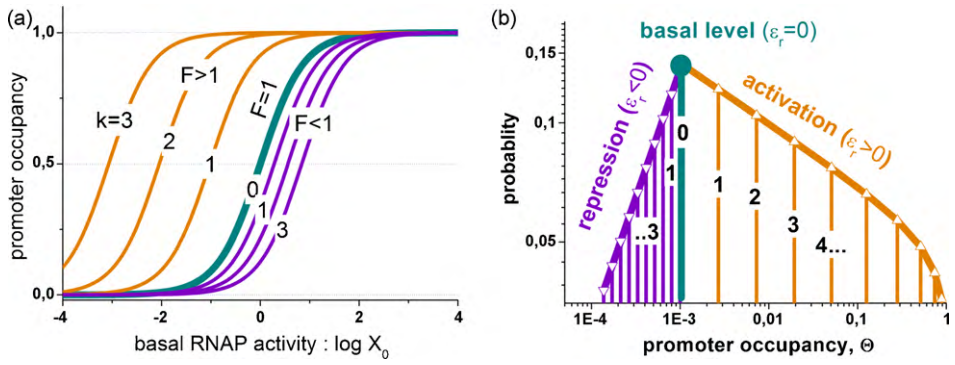

Fig. 5. Promoter occupancy as a function of the basal RNAP binding activity in pure activator and pure repressor systems (panel a) and the corresponding expression profile calculated according Eq. (20) (panel b). Note the asymmetry of activation and repression branches due to the Boltzmann factor of regulation free energies of equal absolute values in Eq. (7).

The mean occupancy of the promoters in the genome transforms into the mean TF-binding activity according to the scaling condition (Eq. (9)):

$$\langle X_{TF}\rangle = X_{TF}^{\max}\cdot\langle\Theta\rangle \tag{19}$$

It equally applies to all TF. The particular value of $\langle X_{TF}\rangle$ is obtained by numerical solution of Eqs. (18) and (19).

To validate the mean expression approximation in terms of the resulting distribution of expression values we generated genomes as described above and calculated expression spectra as averages over 1000 independent realizations. Note that the genes in these simulations are regulated according to the individual expression of each TF. Comparison of simulated and calculated expression distributions justifies the mean expression approximation. A particular example is given below (see part b of Fig. 7).

2.5. Expression spectra

The expression profile of the genome is defined as the probability distribution of the expression values of all of its genes. We will use the term 'expression spectrum' as synonym because the term 'expression profile' is often used in a different, more unspecific meaning in expression analyses to designate a characteristic set of expression values. To obtain the expression spectrum one has to correlate the probability of all possible regulatory states of the genes in the genome with their promoter occupancies. The probability of the regulatory states is given by the IN-degree distribution, $w_{in}(k)$, which has been derived as a function of the number of regulators k (Eq. (13)). On the other hand, k governs the promoter occupancy $\Theta(k)$ and thus the expression. Thus, the expression spectrum is given by $w_{in}(k)$ as a parametric function of $\Theta(k)$ with k as the parameter. In this subsection we discuss two special cases of promoter regulation in order to illustrate the properties of the RGM in terms of the resulting expression spectrum.

2.5.1. Regulation by separate repressors and activators

In this special case all genes are exclusively regulated either by activators or by repressors. The probabilities in Eq. (17) apply with $j=0$ (only activators) and $j=k$ (only repressors), respectively.

Fig. 5 shows the promoter occupancy as a function of the basal RNAP binding activity (panel a), and the corresponding expression profiles (panel b) for activators (orange) and repressors (violet). The superposition of both branches can be assigned to genomes which consist of equal number of genes which are regulated either by activators or repressors.

The expression spectrum can be derived in analytical form for this special case: first, one expresses k as a function of the promoter occupancy Θ by re-arranging Eq. (16) into $k=\ln(F)/\ln(F(X_{TF},\varepsilon_r))$ and then substitutes the regulator strength F as function of the Θ (Eq. (16)). Secondly, insertion into Eq. (13) transforms the exponential decay of the IN-degree distribution into a power law of the form:

$$w_{in}(\Theta)=\left(\frac{X(\Theta)}{X_0}\right)^{-\lambda}\cdot(1-e^{-1/k_0}) \quad \text{with} \quad X(\Theta)\equiv F\cdot X_0=\frac{\Theta}{1-\Theta}$$

$$\text{and} \quad \lambda=(k_0\cdot\ln(F(\langle X_{TF}\rangle,\varepsilon_r)))^{-1} \tag{20}$$

The maximum of the spectra shown in Fig. 5b corresponds to the basal level of expression obtained for spontaneous 'unregulated' binding of RNAP ($\Theta_0=X_0/(1+X_0)$). The exponent λ determines the slope of the flanks of the expression spectrum. Its value is inversely related to the characteristic decay number k_0 and the individual regulation factor $F(\langle X_{TF}\rangle,\varepsilon_r)$. The sign of λ is given by the sign of $\ln F(\langle X_{TF}\rangle,\varepsilon_r)$ which is negative for repressors ($F<1\rightarrow\ln F<0$) and positive for activators ($F>1\rightarrow\ln F>0$) giving rise to a power law of increasing and decaying slope, respectively.

Activation of expression is more efficient than repression in the chosen example, i.e. the absolute value of the slope of the right flank of the spectrum is smaller (see Fig. 6). This property reflects the asymmetry of the Boltzmann factor for exponents of the same absolute value but of opposite sign. Saturation and thus deviation from the power law is observed if the promoter occupancy approaches its maximum value of unity.

The equation for the decay exponent of the expression spectrum can be simplified for the special case of well-connected networks ($\langle N_{bind}\rangle>1$, see Eq. (13)) and strong mean TF-binding activity ($\langle X_{TF}\rangle\gg1$, see Eq. (7)) which is equivalent with large values of promoter occupancy and TF expression (Eq. (19)). Under these conditions λ can be approximated y:

$$\lambda\approx(\langle N_{bind}\rangle\cdot\varepsilon_r)^{-1} \tag{21}$$

Accordingly, the absolute value of λ and thus the steepness of the flanks of the expression spectrum are large for genomes of low internal connectivity characterized by small values of $\langle N_{bind}\rangle$ and/or for weak regulatory effects per bound TF characterized by small absolute values of ε_r.

The impact of the particular model parameters on the expression profiles is demonstrated in Fig. 6. A decrease of X_0, the basal RNAP binding activity, shifts the position of the peak to lower values (Fig. 6a). A decrease of X_{TF}, the TF-binding activity, increases the steepness of the flanks (Fig. 6b). Note that the repressor system is more sensitive for changes of X_{TF}. In the limit of large values of X_{TF} the spectrum becomes symmetric as predicted by Eq. (21). The respective approximation of strong TF-binding activity and expression obviously applies to the right flank for $\log X_{TF}>-1$ but for the left flank only for $\log X_{TF}>+1$. This difference can be simply ratio-

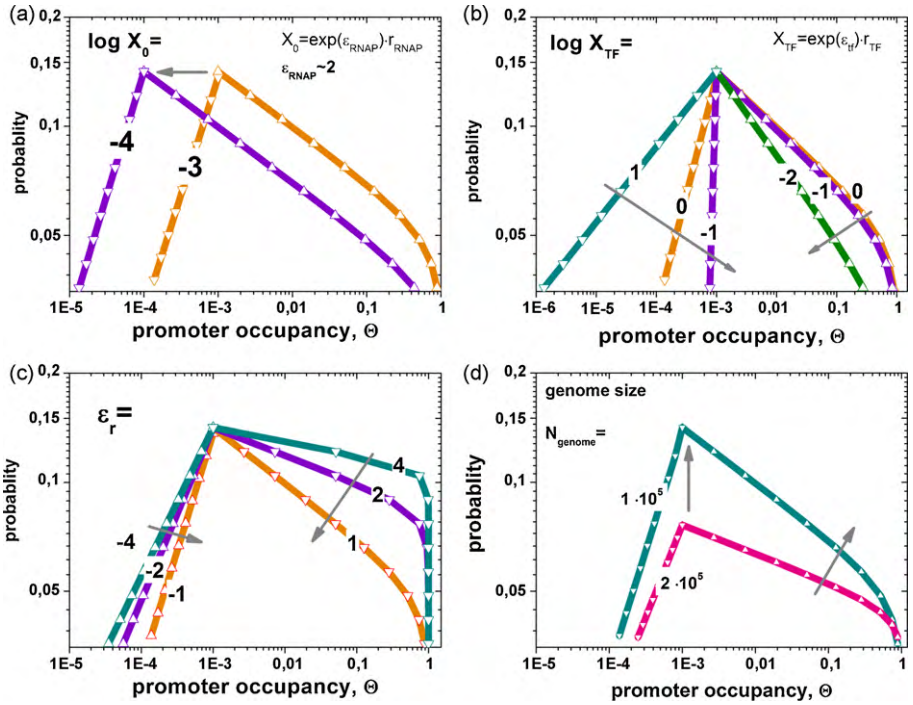

Fig. 6. Changes of the expression spectrum of separated activator and repressor genomes upon variation of the basal RNAP binding activity (panel a), of the TF-binding activity (b), of the regulation free energy which modifies the recruitment of RNAP to the promoter (c) and of the genome size which changes the mean number of bound TF per gene (d). The arrows point in direction of the changes which are induced by the decrease of the value of the respective parameter.

nalized by the fact that the right flank refers to activated genes giving rise to a higher expression level than the left flank due to repressed genes. Panel c of Fig. 6 illustrates that a decrease of the absolute value of ε_r, the regulation free energy, increases the slopes of the flanks. In this case the right activation flank responds more sensitively owing to the asymmetry of the Boltzmann factor as discussed above. The steepness of the slope of the curves decreases with decreasing mean number of TF per gene $\langle N_{bind}\rangle$ (Fig. 6d). This trend is achieved by decreasing the genome size or by increasing the length of the coding region (see Eq. (4)). Panel c and d of Fig. 6 thus illustrate the trends due to changes of ε_r and $\langle N_{bind}\rangle$ predicted by Eq. (21). With respect to invariant values of the decay exponent, the decrease of regulation free energy ε_r can be compensated by the increase of network connectivity $\langle N_{bind}\rangle$ and vice versa.

2.5.2. Regulation by combinations of repressors and activators

In this special case the genes are regulated by combinations of activators and repressors acting independently with identical absolute values of the regulation free energies (i.e. $|\varepsilon_r|$ = const). The respective probability distribution $p(k, j)$ in Eq. (19) becomes the binomial distribution:

$$p(k,j) = \binom{k}{j} \cdot (f_R)^j \cdot (1-f_R)^{k-j} \tag{22}$$

where f_R denotes the fraction of repressors in the genome.

Panel a of Fig. 7 compares the expression spectrum of the binominal mixture of equal numbers of repressors and activators ($f_R = 0.5$) with the spectra of single repressor ($f_R = 1$) and activator ($f_R = 0$) genomes. Note that for a defined number of regulators each possible combination of repressors and activators contributes to the expression spectrum of the binominal mixture (Eq. (22)). In Fig. 7a each thin line refers to a fixed number of regulators (small numbers) and the symbols indicate the probability of the possible combinations. Their envelope refers to the total spectrum (thick lines). In our analytical solution we calculated binominal combinations up to $k = 50$.

The probability density is obtained by binning the individual values of the discrete probability distribution into equally space intervals on the logarithmic scale. Comparing the results of the mean expression approximation with those of explicit simulations of the RGM we found a good agreement. Results are shown in panel b of Fig. 7 for systems with $f_R = 0.5$ and 0.73. The slope of the left repression flank clearly decreases whereas that of the right activation flank shows the opposite tendency with increasing fraction of repressors. This trend generalizes the result discussed above, namely that stronger regulation flattens the slope of the respective flank of the spectrum and vice versa (see Eq. (20)). The linear shape of the flanks in the double-logarithmic plots indicates their power law character in analogy with the single repressor and activator systems discussed above. Eq. (21) can be adapted to the case of combinations of both types of regulators by scaling the mean number of bound TF by the effective fraction of repressors f_R^{eff} and activators f_A^{eff} contributing to regulation:

$$\lambda_{left/right} \approx \left(f_{R/A}^{eff} \cdot \langle N_{bind}\rangle \cdot \varepsilon_r\right)^{-1}$$

$$\text{with} \quad f_{R/A}^{eff} \approx (1 - \langle N_{bind}\rangle \ln(f_{R/A}))^{-1} \tag{23}$$

A decreasing fraction of repressors f_R decreases f_R^{eff} and increases λ for the left flank. Vice versa it increases f_A^{eff} and decreases λ for the right flank. As indicated by the indexing in Eq. (23) the effective fraction $f_{R/A}^{eff}$ consequently refers either to repressors or to activators if one uses the decay constant of the left or right flank, respectively. The approximation given in Eq. (23) assumes that the right and left flanks are determined by the expression of genes that are regulated solely by activators and repressors, respectively.

So far we assumed identical absolute values of the regulation free energy of activators and repressors. More realistic applications require regulation free energies which might vary specifically from regulator to regulator and from gene to gene. As an example panel c of Fig. 7 plots the spectra of binominal mixtures of repressors

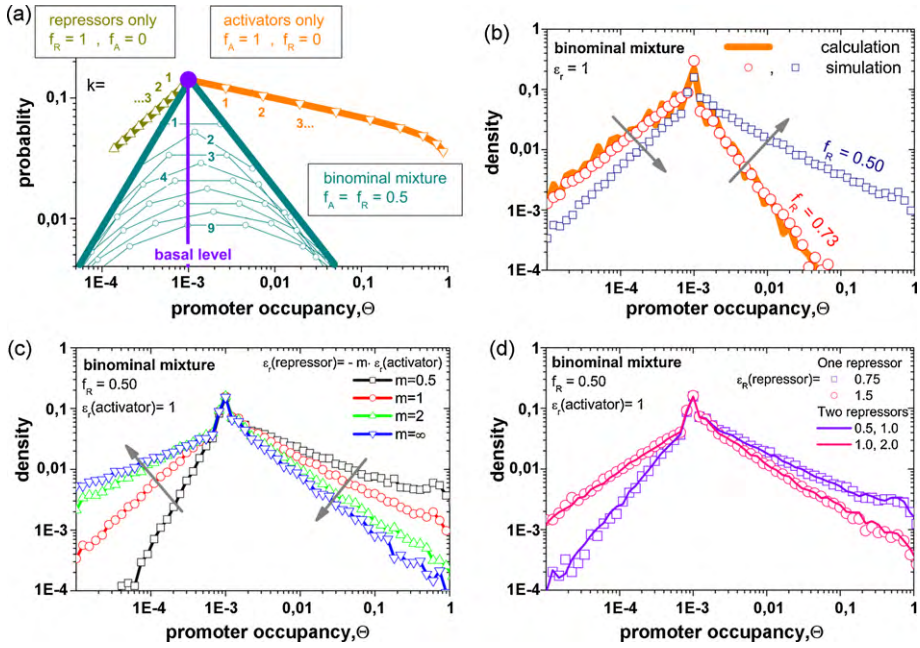

Fig. 7. Expression spectra of binominal mixtures of repressors and activators. (a) Comparison of the spectrum of a genome in which each gene is regulated by equal numbers of repressors and activators on average ($f_R = 0.5$) with genomes in which the genes are regulated solely by one type of regulators. The thin lines refer to all combinations of repressors and activators for particular values of $k = 1 \ldots 9$ which are summed up to provide the expression spectrum of the respective genome (see Eqs. (17) and (22)). (b) Comparison of the simulated spectrum and the respective numerical calculation for a genome which is dominated by repressors ($f_R = 0.73$). Decreasing the fraction of repressors ($f_R = 0.5$) mostly affects the right activator flank (see arrows). (c) Simulated spectra ($f_R = 0.5$) assuming different fold ratios $m = |\varepsilon_r(repressor)/\varepsilon_r(activator)|$ of the regulation free energies of repressors and activators. (d) Simulated spectra of genomes where a single type of activators was combined either with a single type of repressors (symbols) or with two equal-distributed types of repressors (lines). In case that the mean regulation free energy of the repressors in two systems is identical, for example, $0.75 = (0.5 + 1.0)/2$, the spectra of both systems well agree demonstrating that their shape is determined by the effective regulation free energy.

and activators acting with different fold ratios of their regulation free energies, $m = |\varepsilon_r(repressor)/\varepsilon_r(activator)|$. The increase of m increases the steepness of the right and decreases that of the left flank. A further example compares the expression spectra of genomes regulated by one type of repressors with that of genomes regulated by two types of repressors (panel d of Fig. 7). In case the regulation free energy of the former case and the mean regulation free energy of the repressors in the latter system are identical, the resulting spectra of both genomes match each other nearly perfectly.

These examples show that, in general, the decay constant can be assumed to be governed by an effective value ε_r^{eff} which substitutes the individual value ε_r in Eq. (23):

$$\lambda_{left/right} \approx \left(f_{R/A}^{eff} \cdot \langle N_{bind} \rangle \cdot \varepsilon_{R/A}^{eff}\right)^{-1} \quad \text{with} \quad \varepsilon_{R/A}^{eff} \approx \sum_{i \in R/A} w^i \cdot \varepsilon_r^i \qquad (24)$$

The effective free energy can be approximated by the weighted average over the regulation free energy of the regulators for small variations of their values.

In conclusion the RGM approach gives rise to expression spectra the flanks of which decay according to power laws also for more heterogeneous situations such as combinations of different numbers of activators and repressors and/or of different gene-specific regulation free energies. The characteristic power exponents are inversely related to effective values of the fraction of regulators and of the regulation free energy where the left and right flanks characterize repression and activation, respectively. The RGM leaves ample space for more specific interaction models assuming, for example, sequence dependent regulation free energies.

3. Data analysis: extracting expression spectra from microarray intensities

The microarray technology enables to estimate the 'expression degree' of thousands of different transcripts in a given RNA extract in one measurement. The basic principle of microarray experiments relies on the hybridization intensity measurement for an individual probe to infer the transcript abundance specific for a selected gene. The detected intensities are affected by parasitic effects such as non-specific background hybridization and saturation of the probe spots with bound targets which are either not related to the abundance of the transcripts of interest and/or which give rise to a non-linear relation between transcript concentration and intensity. The raw intensities therefore require calibration to obtain appropriate expression measures. Fig. 8 shows the raw intensity distribution (panel a) of a Drosgenome DG-1 GeneChip array and the distribution of expression values after calibration of the raw intensities using the so-called hook method (Binder et al., 2008, 2009; Binder and Preibisch, 2008). This algorithm corrects each probe intensity for its sequence specific non-specific background hybridization, its specific binding affinity and for non-linear saturation effects and summarizes the intensity values of all probes interrogating the same transcript into one expression value, L_S, which is linearly related to the transcript concentration in the studied mRNA extract. The background correction strongly affects small intensity (and expression) values whereas the saturation correction is important in the range of large intensity (and expression) values.

The single-peaked intensity distribution transforms into a double peaked distribution of expression values after calibration (compare panel a and b of Fig. 8). The left maximum can be attributed to so-called 'absent' genes without transcribed mRNA. The respective N-peak is caused by non-specific hybridization of transcripts the sequence of which partly mismatches the respec-

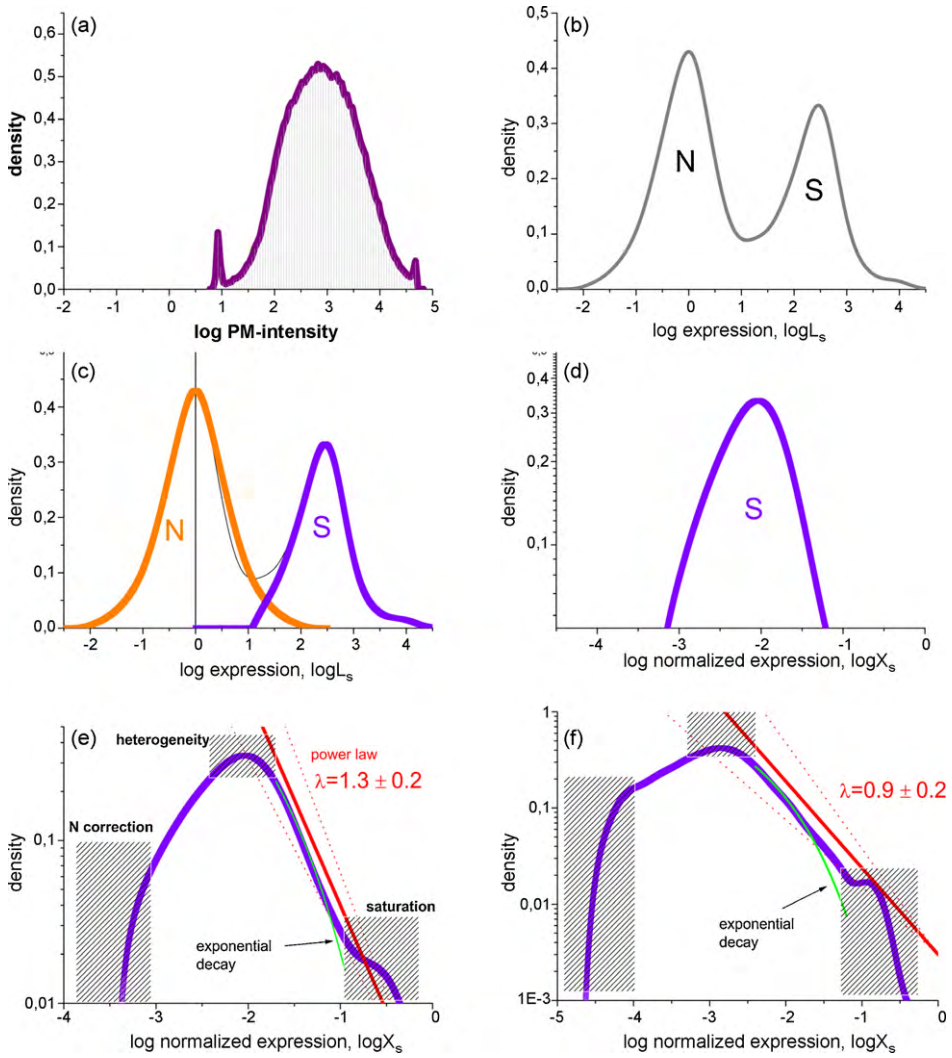

Fig. 8. Transformation of the intensity distribution into the distribution of expression values: (a) distribution of raw intensity values of perfect match (PM) probes of a drosgenome DG1 GeneChip array (data set no. 1; see Table 2). The spikes at the left and right end are due to optical background and optical saturation, respectively. (b) Distribution of expression values after calibration of the intensity values using the hook method (Binder and Preibisch, 2008). (c) The two peaks are attributed to non-specific (N) and specific (S) hybridization of 'absent' and 'present' genes. The fraction of absent genes interrogated by the used chip is $\%N = 63\%$. For spectral decomposition we assume mirror symmetry of the N-peak, reflect its left flank at the ordinate and subtract these values from the total spectrum. (d) The obtained distribution of expression values with logged y-axis and normalized x-axis, $\log X_S = \log(L_S/M)$ (X_S—specific binding strength, L_S—linearized expression in intensity units, M—maximum saturation intensity of the probe spots). (e) The flanks of the density distribution decay linearly over a range of about 1–2 orders of magnitude in the log density-versus-log expression plot which is equivalent with a power law of the form $\sim(X_S)^{-\lambda}$. Exponential decays of the form $\sim\exp(-X_S)$ clearly fail to describe the data (F-test: $p < 0.01$). The hatched areas assign regions prone to measuring artefacts due to the N-correction and saturation and/or to model inconsistencies presumably due to the heterogeneous superposition of expression spectra of different sub-networks (see text). (f) The same as panel (e) but for data set no. 3.

tive probes. Contrarily, the right S-peak originates from genes with 'present' transcripts the probes of which are hybridized specifically. For decomposition of both peaks we assume mirror symmetry of the N-peak with respect to its maximum position, reflect the left flank to the right and subtract the respective values from the total distribution of expression values to obtain the probability density distribution of expression value as shown in panel c and d of Fig. 8.

The expression axis is either scaled in units of signal intensity (L_S) or of specific binding strength of probe/target association ($X_S = L_S/M$, M is the saturation intensity of the probe spots). Both scalings are proportional to the transcript concentration [S], for example, $X_S = [S] \cdot K_S$, where K_S is the mean specific binding constant averaged over all probes of the chip. The estimation of its values requires additional experimental adjustment. We note that the transformation of X_S (and of L_S) into units of promoter occupancy used in the theoretical part requires also normalization with respect to the total amount of extracted mRNA used in the particular hybridization. Both issues are beyond the scope of the present publication. The scaling of the logged expression axis is therefore uncertain except for an additive constant relative to the promoter occupancy. On the other hand, a very similar value of this constant can be assumed for samples which are studied using the same protocol for RNA extraction and preparation and the same GeneChip technology for hybridization which allows the direct comparison of the obtained expression spectra.

Previous analyses report monotonously decaying power law distributions of expression values (Furusawa and Kaneko, 2003; Nacher and Akutsu, 2006; Ueda et al., 2004). Before curve fitting we have however to recall that the expression values and thus also their distributions are prone to artefacts in the limits of large and small abscissa values owing to insufficient correction for effects such as saturation of the probe spots and the overlap with the N-peak. We also emphasize that the obtained experimental density distributions refer to whole genome expression data of heterogeneous samples extracted usually from a mixture of cell types referring to a distribution of phenotypes. They can be interpreted

as the superposition of 'elementary' expression spectra where each of them refers to a subnet of genomic regulation in the sense of the RGM. The superposition of such elementary spectra with varying positions and decays then leads to heterogeneous broadening of the observed total expression spectrum. The affected regions of the spectra around the peak and along the tails have to be excluded from the decay analysis. These restrictions narrow the range available for comparison with theoretical predictions as indicated by the hatched areas in panel e and f of Fig. 8.

The experimental part of this study intends to illustrate that (i), the experimental expression spectra exhibit essentially the same qualitative single-peaked features as predicted by our simple theory, and that (ii), the slopes of the flanks and the position of the peak can vary with the biological context.

Visual inspection of the examples shown in Fig. 8 clearly reveals that the available range of the left and right flanks of the extracted S-peak can be described using a power law over 1–2 orders of magnitude. The adequacy of this choice is confirmed by reduced χ^2-values of 1.5–2 for curve fits to the examples shown (number of data points per decay >50). We also tested exponential laws which however clearly fail to fit the decaying and increasing branch of the spectra compared with the respective power law (p-value: $p < 10^{-2}$; F-test with F-values: $F > 4$; see the examples shown in Fig. 8e and f). Single fits of the power law provide a relative error of less than 2% (95% confidence level) for estimates of the characteristic power constant λ. This uncertainty however increases after considering replicated measurements and taking into account the somewhat arbitrarily chosen range of the flanks of the spectra: As a rule of thumb, the characteristic power constant λ was estimated with a relative accuracy of 10–20% which is an adequate error limit for the qualitative discussion given in the next section.

Previously published distributions of expression values show no distinct maximum in contrast to our data. Detailed inspection however reveals that the published distributions deviate from the power law in negative direction at small expression values and often exhibit saturation behavior (Ueda et al., 2004). Note that we chose an abscissa of logged expression values for the finally obtained expression spectrum where the density refers to equally spaced bins, $\rho^i_{LOG} = n^i/(N \cdot \Delta^i \log L_S)$ (i denotes the bin-index, n^i is the number of genes per bin, N their total number and $\Delta^i \log L_S = \text{const}$ is the bin-width; see panel c and d of Fig. 8). The cited authors used an abscissa which linearly relates to the expression values and applies equally spaced bins $\Delta^i L_S = \text{const}$ to get the density $\rho^i_{LIN} = n^i/(N \cdot \Delta^i L_S)$. The bin-widths of both different scales transform into each other according to $\Delta^i \log L_S = \Delta^i L_S \cdot \partial(\log L_S)/\partial L_S \propto \Delta^i \log L_S / L_S$. For the densities one gets $\rho^i_{LIN} \propto \rho^i_{LOG}/L_S$. Both scales provide different power laws, namely $\rho^i_{LOG} \propto (L_S)^{-\lambda_{LOG}}$ and $\rho^i_{LIN} \propto \rho^i_{LOG}/L_S = (L_S)^{-\lambda_{LIN}}$, where the 'linear' exponent exceeds the 'logged' one by unity:

$$\lambda_{LIN} = \lambda_{LOG} + 1 \tag{25}$$

In other words, power laws decay steeper in linear scale than in logarithmic one. On the other hand, small positive slopes of the left, increasing flank of the expression spectrum with $0 > \lambda_{LOG} > -1$ transform into flat decays with $0 < \lambda_{LIN} < +1$ in linear scale in agreement with the observed saturation behavior at small expression values (Ueda et al., 2004). Hence, logarithmic scale virtually amplifies the left, repression flank of the spectrum compared with linear scale.

In addition to the chosen scale, the obtained distribution of expression values is sensitively affected by the applied preprocessing method which transforms measured probe intensities into expression values. Particularly, the behavior of the distribution at small expression values strongly depends on the applied correction method to remove intensity contributions due to non-specific background hybridization. Global corrections which estimate one common background value for all probes (such as Variance Stabilization Normalization (VSN) and Robust Multiarray Analysis (RMA), see Binder et al., 2009 for a mini review) typically underestimate the non-specific background and therefore overestimate the expression level at the left boundary of the distribution. Moreover, insufficient background correction gives rise to the overestimation of the number of low expressed genes because a certain number of absent probes are counted as expressed ones. The respective expression distributions are therefore imprecise especially at small expression values. Contrarily, probe-specific background correction methods which estimate individual background values for each probe such as MAS5 and gcRMA deliver expression values with much better resolution in the limit of low expression however typically on the expense of larger stochastic noise.

The hook method also applies this probe-specific background correction in combination with a strict criterion for 'absent' probes the specific expression of which is judged as not detectable by the applied microarray technology (Binder and Preibisch, 2008). This detection threshold virtually removes the absent probes from the expression distribution and results in a peaked distribution with a maximum at intermediate expression values which is in qualitative agreement with the distribution predicted by the RGM.

Note that adequate correction requires deconvolution of the total intensity distribution according to $P(I) = N \otimes S$ where $\otimes$ denotes the convolution product of the distributions of the non-specific background and of the specific signal. A deconvolution algorithm by Havilio (2005) provides expression distributions which show a maximum at intermediate expression values in agreement with our results. The discussed method applies however a global background correction with the limitations for small expression values discussed above. In conclusion, the reported shape of the obtained expression distributions at small abscissa values must be judged as a rough estimation. We expect considerable improvement by combining the deconvolution method with the probe-specific background correction applied by the hook method which is currently in development.

4. Expression spectra: experimental examples

We select a series of example data sets to illustrate the properties of expression spectra in the context of different biological issues, treatments, systems and taxa (see Table 2 for an overview).

4.1. Development and cell differentiation

As a first example we analysed time series characterizing developmental and differentiation processes. The *Drosophila* data set comprised three replicated series of 12 consecutive time points of chip measurements starting at 1 h and ending at 12 h post-egg laying of the flies. The left panel of Fig. 9 shows the mean expression spectra averaged over the replicates for six selected developmental stages of the fly embryos. Three phases of transcriptional activity can be distinguished. During the first 5 h of development (until the onset of gastrulation) the position of the expression spectra shifts leftwards indicating the decrease of the mean expression level. Subsequently, it remains stable over about 5 h (phases of post-blastoderm mitosis) before the expression level again increases after about 11 h of development (head involution, dorsal closure, closure of the midgut). In this third phase the slope of the left flank becomes smaller and thus the profile broadens. In terms of our model this indicates an increase of transcriptional repression.

A comparable scenario was observed during the differentiation of murine hematopoietic progenitor cells (FDCP-mix cells). The expression spectra of data set no. 2 averaged over 3 replicates

Table 2
Overview of the example data sets.

No	Experiment	Further remarks	GeneChip array (number of probe sets)	Reference accessibility of chip data (*.cel files)[a]
1	*Drosophila melongaster* developmental time series	Fruit fly embryo development stages L1–L16	Drosgenome DG-1 (~14,000)	(Tomancak et al., 2002) ftp://ftp.fruitfly.org/pub/embryo_tc_array_data/
2	Differentiation time series of hemopoietic stem cells	Growth factor induced differentiation of FDPC-mix cell line into megakaryocyte cells	Mouse genome U74av2 (~12,500)	(Bruno et al., 2004) Personal information
3	Stage specific profiling of LIN-mutants of *C. elegans*	Embryonic and larvae stages L1 and L4 of wild type versus mutants	*C. elegans* genome (~22,500)	(Kirienko and Fay, 2007) GEO: GSE6547
4	Oncogenic pathway signatures	Oncogene activated versus control study on human mammary epithelial cells	Human genome HGU133plus2.0 (~22,500)[b]	(Bild et al., 2006) GEO: GSE3151
5	Human body index transcriptional profiling	Transcriptional profiling of 90 human tissues	Human genome HG133plus2.0 (~22,500)[b]	(Roth et al., 2006) GEO: GSE7307
6	Transcriptional repression of *E. coli* by arginine	Repression of wild type strain and mutant	*E. coli* genome 2.0 (~10,000)	(Caldara et al., 2006) GEO: GSE4724
7	Genome scale changes of the transcriptional oscillator in *Saccharomyces cerevisiae*	Expression profiling of budding yeast in the reductive and respiratory phases	Yeast genome 2.0 array (~5500)[c]	(Li and Klevecz, 2006) GEO: GSE9302
8	Zebrafish embryonic retina (*Danio rerio*)	Wild type retina, microdissected tissue samples	Zebrafish genome (~15,000)	(Leung et al., 2007) GEO: GSE5048
9	Chicken brain (*Gallus gallus*)	Sexually dimorphic expression before gonadal differentiation	Chicken genome (~32,000)	(Lee et al., 2009) GEO: GSE12268
10	Different treatment of *Arabidopsis thaliana* roots	Air and ethanol treatment of wild type and mutants	*A. thaliana* genome ATH1 (~22,500)	(Stepanova et al., 2007) GEO: GSE 7432

[a] GEO abbreviates the web repository Gene Expression Omnibus accessible under http://www.ncbi.nlm.nih.gov/geo/. All shown spectra are averaged over three replicated chip measurements except data set no. 10 for which only two replicates are available.
[b] The HG133plus2.0 array (55,000 probe sets) integrates the probe sets of the HG-U133A chip (22,000) and, in addition, the probe sets of the HG-U133B chip (23,000). In our analysis we mask the probe intensity data taken from the latter chip because most of them are called absent.
[c] The yeast genome array contains probe sets to detect transcripts of the two most commonly studied species of yeast, *S. cerevisiae* (5744 probe sets) and *Schizosaccharomyces pombe* (5021). In our analysis we mask the intensity data of the latter species.

clearly reveal down-regulation of gene expression during the first 4 h of growth factor induced differentiation (right panel of Fig. 9). Subsequently, the profile remains constant until terminal differentiated cells appear after about 24 h (Bruno et al., 2004). At this time point the left flank of the profile starts flattening. Again this change indicates an increased transcriptional repression. This is in agreement with experimental findings that the progenitors co-express several programs of lineage-affiliated gene activity most of which are actively repressed in course of subsequent commitment and differentiation (Bruno et al., 2004).

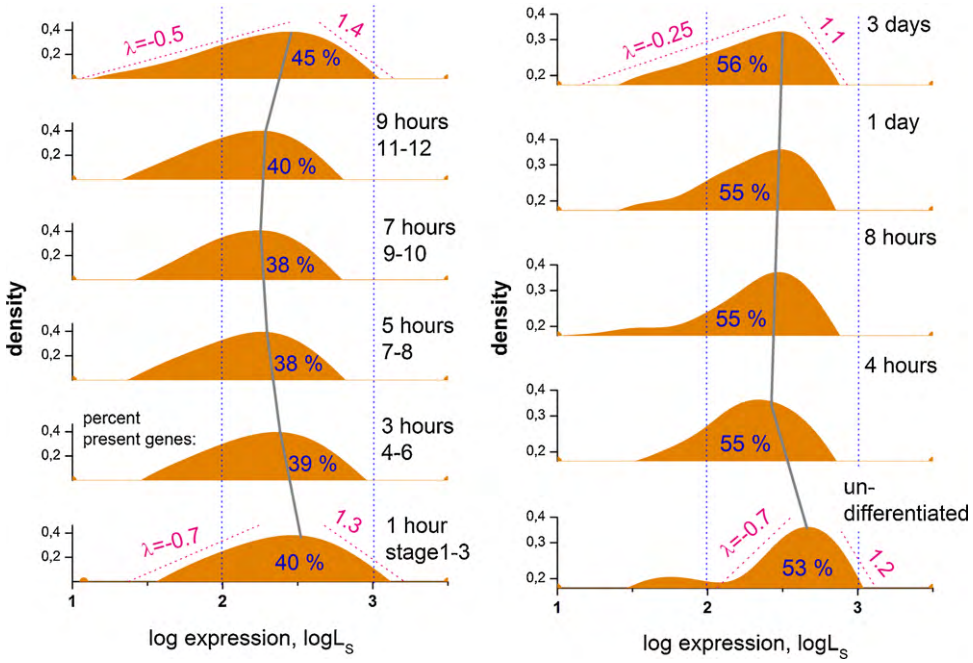

Fig. 9. Expression spectra of the *Drosophila* developmental series (left panel, data set no. 1) and differentiation time series of murine hematopoietic progenitors (right panel, data set no. 2). The shift of the peak position is illustrated by the vertical curves which follow the peak positions of the spectra. The percentages of present genes are given in the figure. The respective standard error is about ±4%. The dotted diagonal lines are power laws $(\sim(L_S)^{-\lambda})$ with the characteristic exponent given in the figure for selected examples. They serve as a guide for the eye to estimate the steepness of the slopes of the increasing and decreasing flanks of the peaks.

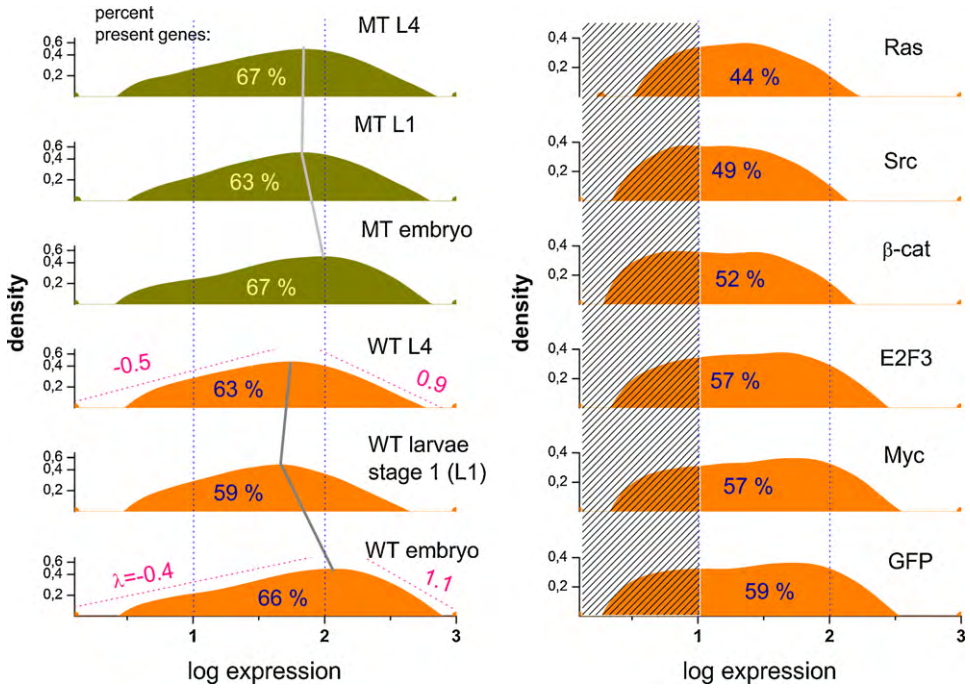

Fig. 10. Stage specific profiling of wild type and lin-35 mutants of *C. elegans* (left panel, data set no. 3). Oncogenic pathway signatures of human mammary epithelial cells (right panel, data set no. 4). The shadowed region refers to small expression values of low resolution (expression <10).

The results obtained by analyzing data set no. 1–2 demonstrate that specific states of development and differentiation correspond to characteristic gene expression profiles. Thereby, early states are characterized by higher mean expression values and late stages by smaller slopes of the left flank of the prolife. The latter can be associated with the occurrence of a larger number of actively repressed genes. Note that transcripts of only ~40% (data set no. 1) and ~55% (dataset no. 2) of all interrogated genes are detectable. These percentages are virtually invariant in the analyzed data sets.

4.2. Mutants and oncogenic de-regulation

In a second series of examples we compared the gene expression profiles of wild type and mutant genomes. In data set no. 3 embryonic and larvae development stages of wild type nematode *C. elegans* were compared with that of lin-35 mutants lacking pocket proteins (left panel of Fig. 9). The expression spectra demonstrate the shift of the mean expression level towards lower values in the course of the development in wild type and mutant worms as well. However the shift was found to be much weaker for the mutated nematodes. This difference obviously reflects the developmental lag due to the malfunction of many genes involved into development. Note that up to 500 genes involved in larvae proliferation, cell cycle regulation and neurological development are repressed in the mutants (Kirienko and Fay, 2007).

Data set no. 4 provides oncogenic pathway signatures of human mammary epithelial cell cultures (Bild et al., 2006). Five selected pathways were permanently activated by adenoviral activation (c-myc, E2F3, β-catenin, c-Src, H-Ras). The expression patterns of the de-regulated cells were found to be highly specific for each activated oncogenic pathway and clearly different from the control GFP-cells (Bild et al., 2006). The studied pathway signatures included about 100–400 differently expressed genes compared with the control. Comparison of the expression spectra showed that oncogenic pathway activation is accompanied by a decreased overall expression level for three of the five studied examples (right panel of Fig. 10). In these cases the spectra shift in the same direction as observed upon differentiation (see right panel of Fig. 9). This result is somewhat surprising because oncogenic transformation was often related to de-differentiation processes.

The results obtained by analyzing data set no. 3 and 4 demonstrate that mutations and selective pathway activations can induce changes of the entire gene expression spectrum. The observed changes reflect de-regulated functions of cell activity.

4.3. Tissue specific and metabolic variability

Data set no. 5 profiles gene expression of ninety distinct human tissue types. We arbitrarily select 10 of them to illustrate tissue-specific heterogeneity of the respective expression spectra (Fig. 11). It has been shown that the variability of expression values within tissues of the same type taken from different individuals was relatively small compared with variability between tissues of different type which enables identifying gene expression differences between tissues (Roth et al., 2006). Application of unsupervised pattern recognition (principal component analysis and hierarchical clustering) to normalized, i.e. relative expression values provides tissue-specific expression characteristics related to organ function (Roth et al., 2006). Our analysis of the gene expression spectra reveals only small differences between the expression spectra of the analyzed tissues indicating comparable transcriptional activity in terms of the RGM.

However, the question remains whether tissue-specific functional modes, for example in the course of de-regulated metabolic activity may change these spectra. The following examples indicate that such changes occur in simple organisms, but appear to depend on whether these changes are environmentally induced or intrinsically regulated.

For example, data set no. 6 provides expression values on gene regulation of arginine biosynthesis in bacteria *E. coli* (left panel of Fig. 12). The involved arg genes are not organized into one single operon, in contrast to what was observed for several other pathways. Instead, the respective genes are scattered over different chromosomal loci. Their expression is repressed by arginine to dif-

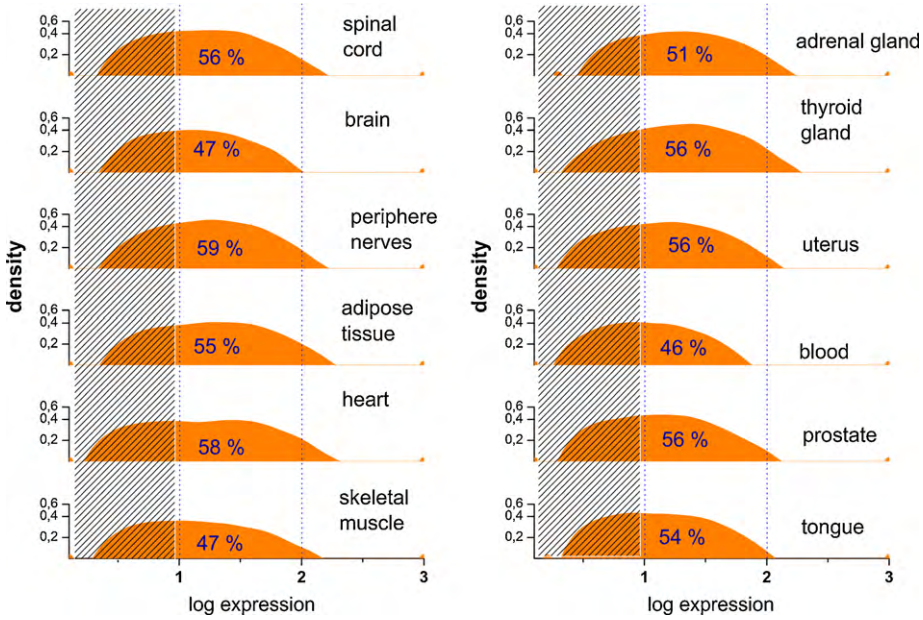

Fig. 11. Human body index profiling of selected tissue types (data set no. 5).

ferent extents in a coordinated fashion through the action of the arginine repressor ArgR (Caldara et al., 2006). The data set compares gene expression of the wild type strain under partial and full repression of arginine biosynthesis induced either by intracellular arginine only (partial repression, WT) or by added extracellular arginine (maximum repression, WT). Additionally, a mutant strain with genetically de-repressed arginine biosynthesis is analyzed under the action of extracellular arginine (maximum repression, mutant). Consistently both, de-repression by reduction of the arginine level and genetic de-repression in the mutant strain cause the shift of the spectra towards smaller expression values. The lower mean expression is paralleled by the flattening of the slope of the right decaying flank. The absolute values of the respective power law exponent decrease (see Fig. 12). The flattening of the right flank suggests the involvement of transcriptional activation in the course of de-repression. To check this hypothesis we calculated the mean expression value of the five arg genes argA, argB. . .argE which are directly controlled by ArgR-repression (see the open triangles in Fig. 12). Their mean expression level roughly agrees with the mean expression level of all genes near the maximum position of the expression spectrum in the repressed sample (WT + arg). De-repression markedly gains the expression level of the arg genes which are now beyond the most strongly expressed genes of the bacteria. We conclude that arg genes are highly activated under normal conditions but reduce their activity under conditions of arginine excess of the wild type strain due to ArgR-repression.

Data set no. 7 provides data on the global gene expression changes during respiratory cycle oscillations of budding yeast. This cycle is characterized by genome wide oscillations of transcription (Klevecz et al., 2004). Different temporal clusters of maximum expressed genes have been identified during the reductive and respiratory phases consisting of about 89% and 12% out of the about 5300 significantly expressed genes in the yeast (Klevecz et al., 2004). This separation in time between oxidative and reductive phases propagates throughout the transcriptome and is coordinated with the initiation of DNA replication. This temporal switch has been explained as an important strategy evolved in cells to prevent oxidative damage to DNA during replication (Klevecz et al., 2004). Our data reveal that the overall expression spectra of the reductive and respiratory phase appear to be invariant (Fig. 12, right part). The expression changes of individual genes are obviously balanced resulting in a relative constant level of total gene activity during the respiratory cycle of yeast.

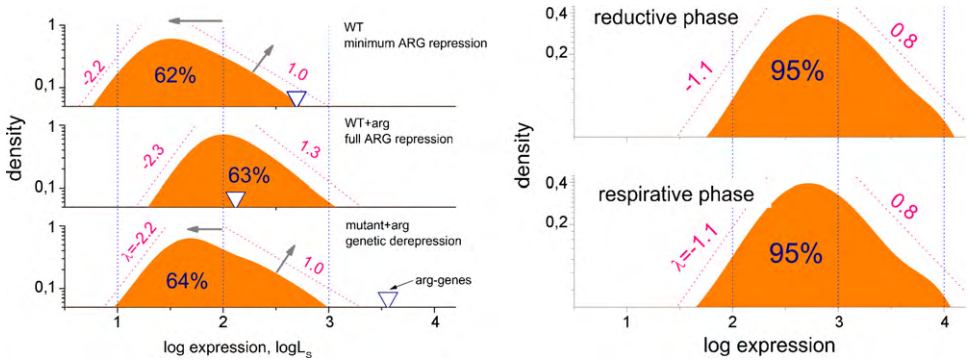

Fig. 12. *E. coli* repression data set no. 6 (left) and *S. cerevisiae* data set no. 7 (right). In the left part the spectra referring to de-repressed arg biosynthesis (WT and mutant + arg) are compared with the conditions of full repression (WT + arg): the arrows illustrate the left-shift of the spectra and the decreased slope of their decaying flank after de-repression. The open triangles indicate the mean expression value averaged over the five arg genes (argA, argB, argC, argD, argE) which are regulated by the ArgR repressor. The expression spectra of the yeast are virtually invariant during the respiratory phase cycle (right part). Note also the exceptional high percentage of present genes in yeast compared with other organisms.

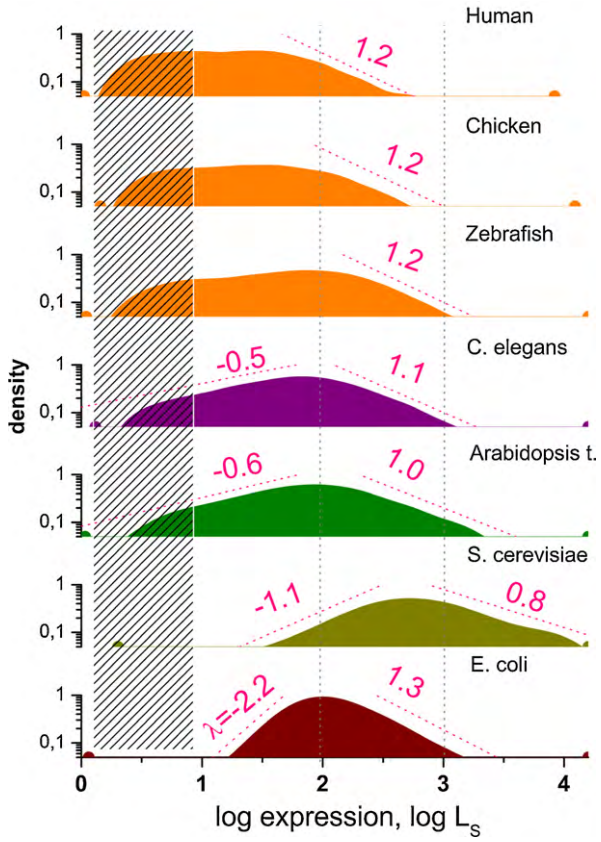

Fig. 13. Expression spectra of different species. Representative expression distributions are taken from data sets 5, 8, 9, 2, 7, 10 and 6 (from bottom to top). The activation flank of the spectra decays according to Zipfs law, i.e. a power with a decay exponent near unity.

4.4. Gene expression from bacteria to human

We selected example data sets which investigate organisms taken from different biological kingdoms and phyla: bacteria (*E. coli*), unicellular eukariotes (*Saccharomyces cerevisiae*) and multicellular eukaryotic animals (*C. elegans*, zebrafish, chicken, human) and plants (*Arabidopsis thaliana*). Fig. 13 directly compares representative expression spectra which have been selected from the examples which used whole genome expression arrays of the GeneChip type of the same generation. Partly, the spectra differ markedly among each other in width (see, e.g. *E. coli* showing a relatively narrow distribution of expression values compared with the other examples) and/or position (see, e.g. *S. cerevisiae* showing a large mean expression level). These different characteristics exceed the variability of the spectra of the same type of organisms for different treatments and individuals (compare with the preceding examples).

The right activation flank decays in all cases with a power law exponent λ near unity which is characteristic for Zipfs law (Zipf, 1949). The exponent is in agreement with previous experimental findings of power laws of expression data in linear scale (Furusawa and Kaneko, 2003; Ueda et al., 2004). The reported decay exponents of about two (Ueda et al., 2004) become unity after rescaling into logged scale according to Eq. (25).

Using the estimated value of the decay exponent of about unity Eq. (24) becomes $f_A^{eff} \cdot \langle N_{bind} \rangle \cdot \varepsilon_A^{eff} \approx 1$. It implies that changes of the network connectivity and/or the particular combination of regulators are balanced by changes of the effective regulation free energy. For example, the connectivity of TF regulatory networks increases for higher organisms (1.2 edges per vertex in *E. coli*; 1.6 in *S. cerevisiae*; more than 2 in mammals) (Goemann et al., 2009; Madan Babu and Teichmann, 2003). This increased connectivity within the genomes of higher organisms is consequently expected to be paralleled by smaller effective regulation free energies and/or by a smaller fraction of activators.

The expression spectra shown in Fig. 13 largely differ in the slope of their left repression flank. The steeper increase of the left flank in the spectra of the unicellular organisms indicates that the impact of repression mechanisms in the regulation of gene activity seems to increase in multicellular eukaryotes compared with bacteria and yeast. Note that the similar trend was observed upon differentiation and development (Fig. 9).

Potapov et al. (2008) showed that simple organisms such as bacteria, yeast and also nematodes avoid significant parallelism of regulatory paths in their gene regulatory networks in contrast to higher level organisms such as mammalians. The RGM predicts the narrowing of the spectrum for such less connective regulatory networks in qualitative agreement with the measured expression spectra. Vice versa, adaptation of genomic regulation to environmental changes may be more heterogeneous and thus flexible for higher organized species (Huang et al., 2005).

5. Summary and conclusions

We presented a statistical thermodynamics model of whole genome transcriptional regulation which combines the RGM approach of gene regulatory network organization with a biophysical description of gene activity. The gene expression of the RGM is governed by transcription and binding of TF and by the ability of bound TF to modulate the recruitment of RNAP by the promoter regions of the genes. The basic size relations and distributions of TF binding of the model genome were presented. Using these distributions together with the respective microstates of promoter occupancy we calculated the expression spectra of different genomes. Essential properties of these spectra were analyzed as a function of different input parameters. For this purpose an analytical solution was derived based on a mean expression approximation. This solution was demonstrated to provide results in good agreement with explicit simulations of the RGM. We found that the expression spectra respond in a characteristic way on different changes of the modes of transcriptional regulation.

Our model predicts a power law distribution of gene activity. Repressors and activators of gene activity give rise to increasing and decreasing tails of the distribution with a maximum in between. This maximum is assigned to the basal transcriptional activity of unregulated promoter states. The decay exponents of the power laws are inversely related to the network connectivity and the average strength of regulation. Hence, the position of the expression spectra and the slopes of their increasing and decreasing flanks provide a simple framework for the interpretation of experimental gene expression spectra. For example, relatively steep tails reflect modes of relatively weak regulation due to weak regulation strengths of bound TF and/or due to low connectivity, i.e. a small number of regulating TF per gene.

Previous studies of experimental expression data based on state-of the art preprocessing methods reported monotonously decaying power law distributions. Applying novel Hook data calibration we demonstrated that the abundance of transcribed mRNA measured with microarrays actually shows a single-peaked distribution in agreement with the shape of the expression spectra predicted by our model. We found that peak position and width of the experimental expression spectra vary with the biological context. Particularly, changes of the spectra were found to occur, e.g. during developmental processes as a consequence of changed metabolisms and in the course of oncogenic transfor-

mations, demonstrating that the spectra describe well-balanced transcriptional modes.

The interpretation of the expression spectra in terms of the RGM approach leads to explanations in agreement with existing knowledge. For example, changes of the spectra during progressive development and cell differentiation can be explained by an increasing fraction of actively repressed genes in accordance with general findings in developmental and stem cell biology (Efroni et al., 2008). A similar trend was also found in multicellular eukaryotes compared with unicellular bacteria and yeast, suggesting that increased complexity in regulation is accompanied by fine-tuned gene repression. Note that alternative mechanisms of repressing gene activity such as DNA-methylation and Polycomb-binding have a high impact on gene regulation during differentiation and development (Meissner et al., 2008; Mikkelsen et al., 2007; Mohn and Schuebeler, 2009; Simon and Kingston, 2009). Such epigenetic effect will be addressed in extended versions of the RGM.

A decay constant of the observed power law of about unity was found to be a universal property in different species. In our model this decay constant characterizes the structure and the expression regulation of the genome in terms of the connectivity and the effective regulation strength of TF. Other approaches explain the power law by the optimization of self-reproduction of metabolic networks in the cells (Furusawa and Kaneko, 2003) or by external and internal noise subjected to the transcription process and the kinetics of transcript degradation (Nacher and Akutsu, 2006). Studying the impact of noise in the context of the dynamics of genomic regulation and the effect on the expression spectra predicted by the RGM is an interesting issue of model extension.

Beside its potential in interpreting experimental expression spectra the RGM represents a suited framework for integrating prespecified selected regulatory units at the genome level to study their properties in the matrix formed by the entire network. This approach can be applied, for example, to investigate the effect of mutations and of pathway activation more in detail. First results presented here suggest that both modifications are capable of changing the expression spectrum of the whole genome.

Acknowledgements

We like to thank Sonja Prohaska and Thimo Rohlf for the fruitful discussion. The work was supported by BMBF grant number 0313836 (Bundesministerium für Bildung und Forschung). HW was kindly supported by Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE). The project LIFE is financially supported by the European Fonds for Regional Development (EFRE) and the State of Saxony (Ministry for Science and the Arts).

References

Banzhaf, W., 2003. Artificial regulatory networks and genetic programming. In: Banzhaf, W., Riolo, R., Worzel, B. (Eds.), Genetic Programming – Theory and Applications. Kluwer Academic, Boston, MA, pp. 43–61.

Banzhaf, W., Dwight Kuo, P., 2004. Network motifs in natural and artificial transcriptional regulatory networks. Journal of Biological Physics and Chemistry 4, 85–92.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.A., Marks, J.R., Dressman, H.K., West, M., Nevins, J.R., 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439, 353–357.

Binder, H., Krohn, K., Preibisch, S., 2008. "Hook" calibration of GeneChip-microarrays: chip characteristics and expression measures. Algorithms for Molecular Biology 3, 11.

Binder, H., Preibisch, S., 2008. "Hook" calibration of GeneChip-microarrays: theory and algorithm. Algorithms for Molecular Biology 3, 12.

Binder, H., Preibisch, S., Berger, H., 2009. Calibration of microarray gene-expression data. In: Grützmann, R., Pilarski, C. (Eds.), Methods in Molecular Biology, vol. 375–407. Humana Press, New York.

Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., Phillips, R., 2005. Transcriptional regulation by the numbers: models. Current Opinion in Genetics & Development 15, 116–124.

Bruno, L., Hoffmann, R., McBlane, F., Brown, J., Gupta, R., Joshi, C., Pearson, S., Seidl, T., Heyworth, C., Enver, T., 2004. Molecular signatures of self-renewal, differentiation, and lineage choice in multipotential hemopoietic progenitor cells in vitro. Molecular and Cellular Biochemistry 24, 741–756.

Caldara, M., Charlier, D., Cunin, R., 2006. The arginine regulon of *Escherichia coli*: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. Microbiology 152, 3343–3354.

Dwight Kuo, P., Banzhaf, W., Leier, A., 2006. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. Biosystems 85, 177–200.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D.G., Buetow, K.H., Gingeras, T.R., Misteli, T., Meshorer, E., 2008. Global transcription in pluripotent embryonic stem cells. Cell Stem Cell 2, 437–447.

Furusawa, C., Kaneko, K., 2003. Zipf's law in gene expression. Physical Review Letters 90, 088102.

Goemann, B., Wingender, E., Potapov, A., 2009. An approach to evaluate the topological significance of motifs and other patterns in regulatory networks. BMC Systems Biology 3, 53.

Havilio, M., 2005. Signal deconvolution based expression-detection and background adjustment for microarray data. Journal of Computational Biology 13, 63–80.

Hoyle, D.C., Rattray, M., Jupp, R., Brass, A., 2002. Making sense of microarray data distributions. Bioinformatics 18, 576–584.

Huang, S., Eichler, G., Bar-Yam, Y., Ingber, D.E., 2005. Cell fates as high-dimensional attractor states of a complex gene regulatory network. Physical Review Letters 94, 128701.

Kirienko, N.V., Fay, D.S., 2007. Transcriptome profiling of the *C. elegans* Rb ortholog reveals diverse developmental roles. Developmental Biology 305, 674–684.

Klevecz, R.R., Bolen, J., Forrest, G., Murray, D.B., 2004. A genomewide oscillation in transcription gates DNA replication and cell cycle. Proceedings of the National Academy of Sciences of United States of America 101, 1200–1205.

Koonin, E.V., Wolf, Y.I., 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research 36, 6688–6719.

Lassig, M., 2007. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 8, S7.

Lee, S.I., Lee, W.K., Shin, J.H., Han, B.K., Moon, S., Cho, S., Park, T., Kim, H., Han, J.Y., 2009. Sexually dimorphic gene expression in the chick brain before gonadal differentiation. Poultry Science 88, 1003–1015.

Leung, Y.F., Ma, P., Dowling, J.E., 2007. Gene expression profiling of zebrafish embryonic retinal pigment epithelium in vivo. Investigative Ophthalmology & Visual Science 48, 881–890.

Li, C.M., Klevecz, R.R., 2006. A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change. Proceedings of the National Academy of Sciences of United States of America 103, 16254–16259.

Madan Babu, M., Teichmann, S.A., 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Research 31, 1234–1244.

Maslov, S., Krishna, S., Pang, T.Y., Sneppen, K., 2009. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proceedings of the National Academy of Sciences of United States of America 106, 9743–9748.

Mattick, J.S., 2007. A new paradigm for developmental biology. Journal of Experimental Biology 210, 1526–1547.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., Gnirke, A., Jaenisch, R., Lander, E.S., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454, 766–770.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., Bernstein, B.E., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560.

Mohn, F., Schuebeler, D., 2009. Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends in Genetics 25, 129–136.

Nacher, J.C., Akutsu, T., 2006. Sensitivity of the power-law exponent in gene expression distribution to mRNA decay rate. Physics Letters A 360, 174–178.

Pérez-Rueda, E., Collado-Vides, J., Segovia, L., 2004. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Computational Biology and Chemistry 28, 341–350.

Potapov, A., Goemann, B., Wingender, E., 2008. The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. BMC Bioinformatics 9, 227.

Reil, T., 1999. Dynamics of gene expression in an artificial genome – implications for biological and artificial ontogeny. In: Proceedings of the 5th European Conference on Artificial Life Springer, pp. 457–466.

Rohlf, T., Winkler, C., 2009. Network structure and dynamics, and emergence of robustness by stabilizing selection in an artificial genome. Advances in Complex Systems 12, 293–310.

Roth, R., Hevezi, P., Lee, J., Willhite, D., Lechner, S., Foster, A., Zlotnik, A., 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 7, 67–80.

Schlitt, T., Brazma, A., 2006. Modelling in molecular biology: describing transcription regulatory networks at different scales. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 483–494.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U., 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature 451, 535–540.

Simon, J.A., Kingston, R.E., 2009. Mechanisms of Polycomb gene silencing: knowns and unknowns. Nature Reviews Molecular Cell Biology 10, 697–708.

Stepanova, A.N., Yun, J., Likhacheva, A.V., Alonso, J.M., 2007. Multilevel interactions between ethylene and auxin in *Arabidopsis* roots. Plant Cell 19, 2169–2185.

Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M., Wiuf, C., 2008. Estimating the size of the human interactome. Proceedings of the National Academy of Sciences of United States of America 105, 6959–6964.

Taft, R.J., Pheasant, M., Mattick, J.S., 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays 29, 288–299.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S., Rubin, G., 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biology 3, 0088.0081.

Ueda, H.R., Hayashi, S., Matsuyama, S., Yomo, T., Hashimoto, S., Kay, S.A., Hogenesch, J.B., Iino, M., 2004. Universality and flexibility in gene expression from bacteria to human. Proceedings of the National Academy of Sciences of United States of America 101, 3765–3769.

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