EMERGENCE OF SPARSITY AND MOTIFS IN GENE REGULATORY NETWORKS*

Marcin Zagórski

Institute of Physics, Jagiellonian University
Reymonta 4, 30-059 Kraków, Poland
marcin.zagorski@uj.edu.pl

(Received December 9, 2011)

We consider a simple model of gene regulatory dynamics derived from the statistical framework describing the binding of transcription factors to DNA. We show that the networks representing essential interactions in gene regulation have a minimal connectivity compatible with a given function. We discuss statistical properties using Monte Carlo sampling. We show that functional networks have a specific motifs statistics. In the case where the regulatory networks are to exhibit multistability, we find a high frequency of gene pairs that are mutually inhibitory and self-activating. In contrast, networks having periodic gene expression patterns (mimicking for instance the cell cycle) have a high frequency of bifan-like motifs involving four genes with at least one activating and one inhibitory interaction.

DOI:10.5506/APhysPolBSupp.5.171
PACS numbers: 87.16.Yc, 87.18.Cf, 87.17.Aa

1. Introduction

After billions of years of evolution Earth's life is a very diverse phenomenon, yet all the living organisms are made of simple building blocks called cells. The single cell is a device designed to interpret internal or external signals in order to enhance its survival prospects. We focus here on gene regulatory networks (GRN), the set of interactions between genes. These interactions along with the gene expression machinery allow all living cells to control their gene expression patterns. In the last decade, our knowledge how any given gene can affect another's expression has been significantly extended through various experiments. For example, small gene

* Presented at the 8th European Conference on Mathematical and Theoretical Biology, Minisymposium "Modeling of Collective Phenomena in Biological Systems", Kraków, Poland, June 28–July 2, 2011.

networks have been constructed to implement simple functions *in vivo* [1,2], and much larger sets of interactions have been derived from a number of organisms [3, 4, 5]. Therefore, it has been possible to show that several subgraphs of interactions ("motifs") arise more frequently than might be expected [6,7,8,9].

Here we ask whether motifs can emerge in network architecture due to functional constraints (output patterns) imposed on GRNs. This question along with framework and results presented further in the article have been already extendedly discussed by Burda, Krzywicki, Martin and myself [10]. With respect to this last reference, the following work contains summary of results presented by myself during Mini-symposium in Kraków with additional discussion of possible questions that can be explored within our framework.

The structure of the article is following. First, we briefly describe the model structure. Second, the process of Markov chain Monte Carlo (MCMC) sampling is described. Third, we investigate the statistical properties of the obtained ensemble of networks. Particularly, we define the concept of "essential" interactions and quantify the sparsity of observed GRNs. Furthermore, we analyze two classes of GRNs having different functional capabilities (multistability *vs.* time periodic behavior). Interestingly, we find very different motifs for these two types; in Alon's [11,12] terminology, the first type leads to mutually inhibitory pairs acting as bistable switches, while the second type leads to bifan, diamond and four point cycle motifs. Last, we conclude by summing up all the findings and discuss possible ideas that might be investigated using our framework.

2. The model

2.1. General framework

Compared to the well-known model of Boolean networks (see [13] and references therein) in which a given gene can be either *on* or *off*, here we allow gene expression to have intermediate values. We consider a system of N genes where each gene produces corresponding transcription factor (TF). Particularly, for i-th gene its normalized expression level S_i is a continuous variable ranging from 0 to 1, where zero means no production of TF and one corresponds to maximal production rate. Since we have N genes we can define a vector variable $\boldsymbol{S} = (S_1, S_2, \ldots, S_N)$ which we call a phenotype. In our approach, we assume that every gene can be influenced by any of N types of TFs. As a result we obtain a $N \times N$ weight matrix $\boldsymbol{W}$, where a given entry W_{ij} corresponds to the strength of interaction between i-th gene and j-th TF. Hereafter we refer to $\boldsymbol{W}$ as the genotype and a formula to determine the values of W_{ij} will be given in the following subsection.

To find gene expression pattern $\boldsymbol{S}(t)$ at any given time t, we propose a deterministic dynamics described by a map $\boldsymbol{S}(t+1) = G\left(\boldsymbol{S}(t), \boldsymbol{W}\right)$, where we call initial phenotype $\boldsymbol{S}(0)$. In the context of discrete dynamical systems, G is the global transition function and $\boldsymbol{S}(t)$ denotes the configuration of the system at time t. This discrete dynamics can be represented by a sequence of steps leading to an attractor which is either a cycle or a fixed point.

If we had only one target phenotype, as it is in [14], we would sample the space of all genotypes leading from $\boldsymbol{S}^{(\text{initial})}$ to the "vicinity" of some fixed $\boldsymbol{S}^{(\text{target})}$. In order to quantify how close a given phenotype is to the target one, we define a fitness function

$$F(\boldsymbol{S}) = \exp\left(-fD\left(\boldsymbol{S}, \boldsymbol{S}^{(\text{target})}\right)\right), \tag{1}$$

where $D(\boldsymbol{S}, \boldsymbol{S}') = \sum_i \mid S_i - S_i' \mid$ is the difference of expression levels for each gene, and $f \in \mathbb{R}$ is a control parameter.

Summary of the main model ingredients is given in Table I. In the case of multiple target phenotypes, we can still use Eq. (1), but the "total" distance used to calculate fitness should be a sum of distances from n fixed point phenotypes and corresponding target phenotypes; for cycling behavior target phenotypes are consecutive steps of the imposed cycle.

TABLE I

Short summary of model ingredients: (i) S_i is a continuous variable corresponding to expression level of gene i; (ii) W_{ij} represents the strength of interaction between gene i and TF j; (iii) $\boldsymbol{S}^{(\text{target})}$ corresponds to cell's function; (iv) using fitness function we sample space of genotypes leading from $\boldsymbol{S}^{(\text{initial})}$ to the "vicinity" of $\boldsymbol{S}^{(\text{target})}$.

Phenotype	$\boldsymbol{S} = (S_1, S_2, \ldots, S_N)$ $S_i \in [0,1]$
Genotype	$\boldsymbol{W}$ is matrix $N \times N$
Dynamics	$\boldsymbol{S}(t+1) = G\left(\boldsymbol{S}(t), \boldsymbol{W}\right)$
$\underbrace{\boldsymbol{S}(0)}_{\boldsymbol{S}^{(\text{initial})}} \xrightarrow{\boldsymbol{W}} \boldsymbol{S}(1) \xrightarrow{\boldsymbol{W}} \ldots \xrightarrow{\boldsymbol{W}}$	$\underbrace{\boldsymbol{S}(t) \xrightarrow{\boldsymbol{W}} \boldsymbol{S}(t+1)}_{\text{fixed point phenotype}}$
Fitness	$F(\boldsymbol{S}) = \exp\left(-f \sum_i \left\lvert S_i - S_i^{(\text{target})} \right\rvert\right)$

2.2. Microscopic interactions

In order to determine the strength of interaction between TF and DNA strand, we represent each TF as well as each binding site by a character string of length L with characters belonging to a 4 letter alphabet. Following the standard practice [15], we assume that the free energy of one TF molecule bound to its target site is, up to an additive constant, equal εd_{ij}, where ε is the single mismatch energy and d_{ij} is a number of mismatches between i-th binding site and j-th TF (see Fig. 1). Furthermore, one can define the "interaction strengths" W_{ij} via Boltzmann factor

$$W_{ij} = e^{-\varepsilon d_{ij}} \,, \tag{2}$$

with normalizing constant set to 1 (*cf.* [16]). In the case of n_j TFs of j-th type one can derive [16, 17] the probability p_{ij} that precisely one of them is bound to the binding site of i-th gene

$$p_{ij} = \frac{1}{1 + 1/(W_{ij} n_j)} = \frac{1}{1 + \exp\left(\varepsilon d_{ij} - \ln(n S_j)\right)} \,, \tag{3}$$

which dependence is known to physicists as Fermi function. In the above formula, for the sake of simplicity, we assume that $n_j = n S_j$, where n is a model parameter representing the number of TFs. For the current work we use $n = 1000$, $L = 12$ and $\varepsilon = 2$, though we have checked that for biologically relevant parameters the model findings are qualitatively the same [14, 18].

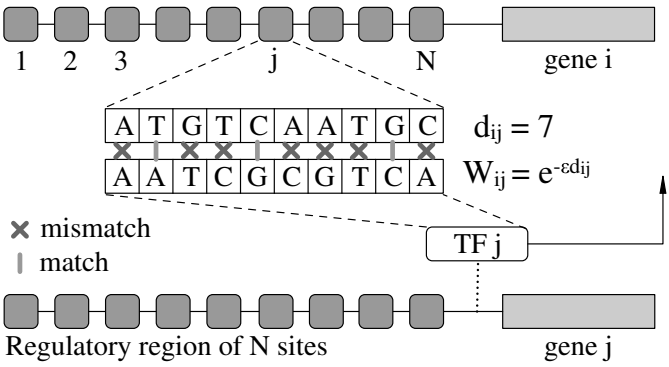

Fig. 1. Schematic representation of the regulatory region of gene i with N binding sites. Represented is the interaction W_{ij} mediated by the binding of TF j to the j-th site of that region. Here, the string representing the TF is of length $L = 10$, number of mismatches (not complementary regions) $d_{ij} = 7$, and the resulting W_{ij} is calculated according to Eq. (2).

To keep the framework simple, we assume that gene i transcription is "on" whenever at least one TF is bound within its regulatory region and otherwise it is "off". For inhibitory interaction the TF of type j bound to its binding site is assumed to stop the transcription. The (normalized) mean expression level of a gene is then identified with the probability that transcription is "on". Hence, for gene i with both activators and repressors we have

$$S_i(t+1) = \left[1 - \prod_j \left(1 - p_{ij}(t)\right)\right] \prod_{j'} \left(1 - p_{ij'}(t)\right) , \tag{4}$$

where j runs over activating interactions and j' over inhibitory interactions, and $p_{ij}(t)$ is given by Eq. (3) with S_j replaced by $S_j(t)$. In the above, just like in many other modeling frameworks, we use discrete time [13,19,20,21].

3. Mutation-selection balance

As already mentioned, we constrain GRNs to exhibit two types of behavior: *(i)* multistability, where a gene regulatory network has 2, 3, or more fixed points (steady state expression patterns); *(ii)* cycling behavior, where phenotype follows a cyclic trajectory in the space of gene expression patterns. In the first case we start the system in one of these fixed points and check whether after a long time it stays close enough to its starting point. In the second case, we start with one of the patterns in the target cycle and check if it stays close to the periodic trajectory. Particularly, for the steady state behavior, we impose up to four fixed points that consist of $N/2$ levels at 0 and $N/2$ at 1, and furthermore that are taken to be orthogonal. For the case where one enforces a target cycle, we use the toy sequence [21] for the yeast cell-division cycle (for details see [10]).

Since the space of viable GRNs is only a tiny fraction of the space of all regulatory networks, we need to introduce some effective sampling method. In particular, we use Markov chain Monte Carlo with the Metropolis rule to explore this ensemble. The procedure is following, we start with random genotype that is we draw all the characters representing gene's regulatory regions and TF molecules randomly, and calculate the corresponding weight matrix $\boldsymbol{W}$. Next, with each step we apply a point mutation to characters representing DNA binding sites (alternatively we change the character of interaction from activatory to inhibitory or vice-versa), recalculate $\boldsymbol{W}$, and according to Eq. (4) the associated fixed point phenotypes or cycling expression patterns $\boldsymbol{S}$. Afterwards, having $\boldsymbol{S}$ we compute fitness of the genotype and accept or reject the attempted move according to Metropolis acceptance probability. This way by applying mutation-selection balance we obtain, after some initial period, an ensemble of viable genotypes constrained to have particular function.

4. Results

4.1. Sparsity of the essential interactions

Having obtained different ensembles of viable genotypes one would like to know which of the interactions between DNA regulatory regions and TFs are essential for network function. In order to get this information, we remove one of the interactions from the genotype and check if the gene expression pattern corresponding to this modified genotype is still close to the target phenotype. If the removal of interaction from GRN leads to loss of its functional capabilities we refer to this interaction as *essential*. Furthermore, the set of all essential interactions for a given genotype defines *essential network* for that GRN.

In previous work [14] on a simpler model with only one fixed point and no allowance for inhibitory interactions, we found that the great majority of genotypes had just one essential interaction per gene. In the case of multistability, as we impose more fixed points, the mean number of essential interactions grows only slightly, with a mean of 1.2, 1.5, 1.9 for 2, 3, and 4 fixed points respectively (for $N = 16$). Moreover, one gets analogous results by forcing the expression vector to cycle through given patterns.

Qualitatively, the observed sparsity can be easily understood: in our framework all the viable networks are subject to mutation-selection balance, that is on the one hand system prefers to increase its fitness (selection criterion), and on the other hand the random mutations try to "keep" the whole genotype maximally random. In terms of entropy, every additional essential interaction has typically a high entropic cost: there are a few strings that have low mismatch values and many that have high mismatch values. Hence, observed essential networks are as sparse as possible to maintain their functional capabilities.

4.2. Motifs emerge from function

To obtain insights into network structure, one can search for network motifs [11,22], that is subgraphs which are overrepresented in a given GRN compared to the randomized version of the network. In particular, we use randomization proposed by Maslov and Sneppen [23]: the links are interchanged, so that both the in- and out-degrees of network nodes remain unchanged. As a result, we find six types of motifs (defined in Fig. 2) which are almost not present in the randomized GRNs. For instance motif "a" is found on average 0.706(16), 2.358(39), 2.984(4) times per GRN with 2, 3, and 4 fixed points respectively, and in the case of randomized networks its frequency is only 0.002(1). Moreover, motif (a) is not present at all in the case of GRNs with periodic expression patterns where the remaining five motifs ((b) to (f)) play leading roles. The frequencies of these latter motifs

for cyclic case are following: 5.451(41) for (b), 5.170(40) for (c), 4.533(42) for (d), 6.676(22) for (e) and 2.296(29) for (f); in the case of randomized ensemble the average number of these motifs per GRN is about two orders of magnitude smaller. Additionally, motifs (b) to (f) are not present at all in the multistable case.

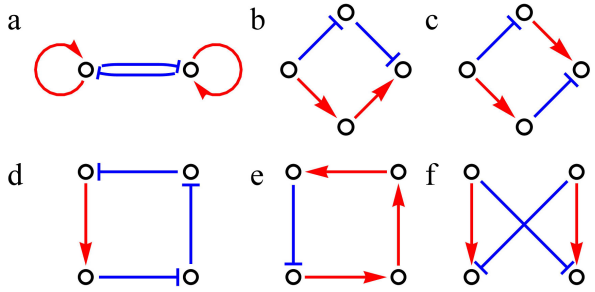

Fig. 2. The most important motifs found for our two classes of functional constraints. Case of many stable fixed points: (a) double negative feedback loop with auto regulation. Case of time periodic gene expression: (b), (c) incoherent diamond, (d), (e) frustrated four-node loop, (f) incoherent bifan.

On qualitative level we see that different motifs arise for our two classes of functional capabilities. In the case of multistability, a single motif (a) with two genes which are mutually inhibitory and self-activating is extremely important. Such a pair of genes can act as a bistable switch that fixes itself in the considered target pattern, and then regulate other genes in a downstream effect. Interestingly, this simple motif is found in a number of biological gene networks, for instance in the genetic switch between lysogeny and lysis of the phage λ [24].

For the case with time periodic output patterns, instead of the previous motif we have a few four gene motifs that are strongly over-represented. In the nomenclature of Alon [11,12], motifs (b) and (c) are *incoherent diamonds*, while motif (f) is the *incoherent bifan*; the others, motifs (d) and (e), involve a regulatory loop, and in fact these loops are "frustrated" (they have an odd number of inhibitory interactions). Again, some of these motifs have been found in biological gene networks [12] with the bifan motif being perhaps the most prominent. Additionally, it is worth to notice that the presence of "frustrated" loops (negative circuits) is expected for the systems that exhibit time periodic behavior [25].

5. Discussion and conclusions

We have considered a relatively simple model of GRNs in which molecular information is used to obtain a matrix of interactions between DNA binding sites and TF molecules. Using our MCMC sampling procedure we

produce many regulatory networks and study statistical properties of the obtained ensemble of GRNs. By introducing the concept of "essential interactions" we are able to quantify the sparsity level of GRNs. As a result, we find that regulatory networks are as sparse as possible being compatible with a given function. This feature qualitatively agrees with biological networks, since the sparsity of interactions is also found in experimental studies of simple organisms [26, 27].

Within our model generated networks are evolvable and a given target expression pattern can be realized through different topologies. Having such framework, we can ask whether functional constraints shape the network structure. Particularly, we consider two classes of constraints which resemble two types of biological processes: *(i)* different stable gene expression patterns can be interpreted as different types of cells during cell development, *(ii)* cyclic gene expression is characteristic for cell cycle, where different genes are excited/inhibited during different stages of cell division process. In the case of multistability we observe two node motif that works as a bistable switch between situations with one gene being "on" and the other being "off". In the case of target phenotypes being periodic in time the bistable switch is not present, and four node motifs like bifan, diamond and "frustrated" loop appear and are highly overrepresented. Hence, we can conclude that different classes of motifs are observed for different types of functional capabilities of GRN. This result is very striking if we realize that no motif structures are incorporated inside our framework on any level. Instead, motifs emerge from purely random background due to imposed functional patterns and selection pressure.

A very gratifying point is that analyzed motifs are also found in biological networks. Interestingly, this result is obtained with assumption of TFs binding independently to DNA, which is not always the case in biological systems. Therefore, we have also checked what happens if one modifies Eq. (4) to incorporate explicitly term giving the probability that two different types of TFs bind to DNA in cooperative manner [28]. By increasing the strength of cooperation we are able to obtain GRNs which are more dense (have more interactions), thus frequencies of various motifs are also generally higher than in the case with no cooperation. However, if we compare motif statistics with randomized networks, the only strongly overrepresented motifs are the same as depicted in Fig. 2. Therefore, we can draw conclusion that network motifs are determined by functional constraints rather than cooperativity effects.

Last but not least, one can explore other interesting options within proposed framework. One possibility is to study different kinds of output patterns, which approach corresponds to imposing different functional constraints on GRN evolution. Particularly, it is still to be checked whether the

most abundant motif in experimental networks, namely feed-forward loop (FFL), can be obtained in our model by constraining GRN to perform a given function. Another option is to explore dynamics stability [29] of regulatory networks in order to see which of the obtained network topologies are most robust to fluctuations in gene expression or node removal (gene knockout). Finally, one might think of constraining the system to have not only certain functional capabilities, but also to fulfill other types of criteria. This way it might be possible to get GRNs closer and closer to biological regulatory networks in a way similar to what was done in [30] for metabolic networks.

The project operated within the Foundation for Polish Science International Ph.D. Projects Programme cofinanced by the European Regional Development Fund, agreement no. MPD/2009/6. This work was supported by the Polish Ministry of Science and Higher Education Grant No. N N202 229137 (2009–2012). M.Z. is grateful to the LPT for hospitality.

REFERENCES

[1] M. Elowitz, S. Leibler, *Nature* **403**, 335 (2000).
[2] T. Gardner, C. Cantor, J. Collins, *Nature* **403**, 339 (2000).
[3] M. Herrgard, M. Covert, B. Palsson, *Curr. Opin. Biotechnol.* **15**, 70 (2004).
[4] H. Salgado *et al.*, *Nucl. Acids Res.* **34**, D394 (2006).
[5] Z. Hu, P. Killion, V. Iyer, *Nat. Genet.* **39**, 683 (2007).
[6] S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nat. Genet.* **31**, 64 (2002).
[7] H. Ma *et al.*, *Nucl. Acids Res.* **32**, 6643 (2004).
[8] T. Lee *et al.*, *Science* **298**, 799 (2002).
[9] J. Zhu *et al.*, *Nat. Genet.* **40**, 854 (2008).
[10] Z. Burda, A. Krzywicki, O.C. Martin, M. Zagorski, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17263 (2011).
[11] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC, Boca Raton, FL, 2007.
[12] U. Alon, *Nat. Rev. Genet.* **8**, 450 (2007).
[13] S. Kauffman, *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford 1993.
[14] Z. Burda, A. Krzywicki, O.C. Martin, M. Zagorski, *Phys. Rev.* **E82**, 011908 (2010).
[15] P. von Hippel, O. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1608 (1986).
[16] U. Gerland, J. Moroz, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12015 (2002).

[17] M. Lässig, *MC Bioinf.* **8**, S7 (2007).
[18] M. Zagórski, *Acta Phys. Pol. B Proc. Suppl.* **4**, 2 (2011).
[19] A. Wagner, *Evolution (Lawrence, Kans.)* **50**, 1008 (1996).
[20] S. Bornholdt, T. Rohlf, *Phys. Rev. Lett.* **84**, 6114 (2000).
[21] F. Li *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4781 (2004).
[22] R. Milo *et al.*, *Science* **298**, 824 (2002).
[23] S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
[24] M. Ptashne, *A Genetic Switch: Phage λ Revisited*, Cold Harbor Spring Laboratory Press, NY, 2004.
[25] R. Thomas, *Int. J. Dev. Biol.* **42**, 479 (1998).
[26] D. Thieffry, A. Huerta, E. Perez-Rueda, J. Collado-Vides, *Bio Essays* **20**, 433 (1998).
[27] G. Balazsi, A. Heath, L. Shi, M. Gennaro, *Mol. Syst. Biol.* **4**, 225 (2008).
[28] N.E. Buchler, U. Gerland, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5136 (2003).
[29] F. Ghanbarnejad, K. Klemm, *Phys. Rev. Lett.* **107**, 188701 (2011).
[30] A. Samal, O.C. Martin, *PLoS ONE* **6(7)**, e22295 (2011).