Chromatin architecture and the orchestration of gene expression: cell systems to explore epigenetic gene control

Brink, M.C.

Publication date
2009

Document Version
Final published version

Chapter 6

Towards hierarchical eukaryotic gene network representations: a perspective

Maartje C. Brink, Roeland Merks, Pernette J. Verschure & Roel van Driel

In preparation

Gene network representations in eukaryotes

A gene network consists of a set of genes that, directly or indirectly, affect each other's activity. The combination of regulatory proteins and non-coding RNAs (ncRNAs) in the cell are the key determinant of the functional state of the network, thereby defining, in principle, the properties of the cell. Hence, understanding gene networks is essential to achieving a basic understanding of cell function. Currently, the representation of gene networks is often in the form of graphs, which can have a simple (Friedman *et al.*, 2000) or a more sophisticated format (such as CellDesigner, Cytoscape and BioTapestry; Kitano *et al.*, 2005; Longabaugh *et al.*, 2005; Shannon *et al.*, 2003). These representations are oversimplifications, ignoring most of the molecular complexity of the system, such as system dynamics and the multiple processes implicated by the edges of the graph. These edges include, even in the most simple gene-gene interaction mechanism; transcription, transcript maturation and packaging, nuclear export, translation of RNA to protein, posttranslational modification, nuclear import and assembly of an active transcription initiation complex. In this sense, gene networks are fundamentally different from metabolic networks, which mostly refer to metabolites, enzyme-catalyzed chemical conversions and protein-protein interactions (Schilling *et al.*, 1999).

Here we address the question of how eukaryotic gene network models should be shaped and which components and processes may be incorporated, including those acting at the epigenetic level of gene regulation. The rapidly increasing knowledge of genome-wide gene regulation creates a promising basis to expand gene network models in order to present a comprehensive system, allowing an in-depth understanding of gene regulation. Here, we explore what information should be incorporated in eukaryotic gene network models to make them more realistic and informative, augmenting their predictive value and insight into the logic of gene networks. Amongst others we indicate what components and variables constitute a eukaryotic gene network and how epigenetic regulatory interactions can be integrated.

Limitations of present representations of eukaryotic genetic networks

At present, most gene network representations depict genes as nodes in classical graphs. However, such representations ignore many aspects of eukaryotic gene regulation. In particular, the edges connecting the nodes are often ill-defined and may represent a wide variety of interactions, including highly indirect interactions involving for instance signal transduction pathways or processing through multi-step metabolic pathways (Schilling *et al.*, 1999). Some types of gene network representations compensate for this to some extent. For instance, the BioTapestry format annotates nodes by using different colours, signifying the type of interaction (Longabaugh *et al.*, 2005). A major challenge is to develop a gene network format that incorporates all essential aspects of the system. A key component in gene regulation that has been ignored in all eukaryotic gene network models so far is the epigenetic regulation of

gene expression, despite its impact on gene expression (Jenuwein and Allis, 2001; Turner, 2007). Additionally, the role of regulatory ncRNAs, probably being equally important to gene control at all levels as proteins, should be incorporated (Goodrich and Kugel, 2006). Inevitably, new concepts have to be developed for eukaryotic gene network models to incorporate these aspects.

Two hierarchical levels of gene regulation

Gene regulation in eukaryotes comprises at least two levels of gene regulation, *i.e.* the classic gene level, acting on individual genes, and the epigenetic level, controlling genomic loci that often contain multiple genes (Bruggeman *et al.*, 2008; de Wit and van Steensel, 2008). Gene-level and epigenetic regulation are both under control of transcription factors (TFs) and possibly ncRNAs (Mattick and Makunin, 2006), defined here as factors that recognize specific DNA sequences. These factors decide, often in a combinatorial way, exactly where in the genome a change in gene activity and/or chromatin state is induced. Subsequently, they initiate the assembly of a chromatin-associated protein or protein-ncRNA complex to execute transcription.

Epigenetic gene regulation changes the chromatin state of defined genomic loci. Locus switching may for instance be accompanied by a transition between facultative heterochromatin and euchromatin (Goetze *et al.*, 2007). Changes in the chromatin state are potentially controlled by locus control region (LCR)-like genomic sequence elements, which are different from promoters and enhancers (Noordermeer *et al.*, 2008). Epigenetic loci range from tens of kbs (*e.g.* the β-globin locus) up to several Mbs (*e.g.* the MHC locus). Locus boundaries are defined by so called boundary elements or insulators (Wei *et al.*, 2005). Epigenetic switching of chromatin states is induced by (i) changes in DNA methylation, (ii) various (but not all) histone modifications, and (iii) the exchange of canonical histones by variant histones (Bird, 2002; Henikoff *et al.*, 2004; Turner, 2007). The epigenetically controlled chromatin state is stably transmitted during cell division, *e.g.* in imprinting phenomena where one of the parental alleles is expressed while the other is repressed (Bartolomei *et al.*, 1991; Edwards and Ferguson-Smith, 2007). Epigenetic regulation is thought to play a key role in cell differentiation (Bird, 2002; Ng and Gurdon, 2008). An example of the role of epigenetics in deciding cell fate is the presence of 'bivalent domains' in the developmental regulatory genes in pluripotent ES cells (Azuara *et al.*, 2006; Bernstein *et al.*, 2006). A single bivalent domain in the undifferentiated cell contains both activating and repressive histone modifications, only one of which remains upon cell differentiation (Bernstein *et al.*, 2006). Epigenetic mechanisms are an integral part of gene regulatory systems in eukaryotes and should therefore be incorporated in gene networks.

How can we integrate the epigenetic regulatory level in gene networks?

To incorporate epigenetic regulation into eukaryotic gene networks, we propose a

simple set of rules. The epigenetically controlled genomic locus is demarcated at both sides by an insulator or boundary element. The locus can occur in two epigenetic states.

1) The non-permissive state. In the epigenetically silent state all genes of the locus are inactive, irrespective of the presence or absence of TFs that may or may not bind to the regulatory sequences of the constituent genes.

2) The permissive state. In the epigenetically permissive state the genes of a locus can be switched on and off individually by the binding of TFs to their promoter and enhancer sequences.

A well-studied example of epigenetic regulation is the β-globin locus. In non-haemopoietic cells, the β-globin locus is epigenetically non-permissive, blocking transcription of all of its genes. On the other hand, in haemopoietic cells the locus is in a permissive state, allowing regulation of its individual genes (Brown *et al.*, 2001; Higgs *et al.*, 2006). More examples of epigenetically controlled multi-gene loci are known, indicating a potentially important role for gene clustering (Sproul *et al.*, 2005). Gene clusters may contain functionally related genes or genes that are not related in an obvious way except that they have to be available for transcription under the same physiological condition.

Consequently, we identify three possible states for any gene, characterized by the state of the locus and the state of the individual gene. The three states are (epigenetic state/ genetic state): off/off, on/off, on/on. In the epigenetic 'off' state, all genes are switched off, independent of the presence of relevant TFs. On the other hand, when the epigenetic locus is 'on', the individual genes can be activated or inactivated, depending on the availability of the required TFs. This simple set of rules allows the formal, hierarchical integration of the epigenetic control level in gene regulatory networks.

What are the relevant components and variables for eukaryotic gene networks?
Gene networks in higher eukaryotes can be defined by the components and variables described below, accompanying genome-wide detection techniques are listed in table 1.

System components

1. Genes: defined as genomic sequences coding for proteins or for functional ncRNAs.
2. Transcripton factors (TFs): proteins and regulatory ncRNAs that recognize genomic sequence elements that play a role in gene-level or epigenetic regulation.
3. TF binding sites: key regulatory sequences at the epigenetic level and the individual gene level. At the gene level, TF binding sites include promoters, enhancers and enhancer blockers. At the epigenetic level they include locus control regions and boundary elements.
4. Epigenetic units: genomic domains (often gene clusters) that can be

switched between a permissive and a non-permissive state. Presently, only a limited number of epigenetic units have been thoroughly characterized, the β-globin cluster being a paradigm. The difficulty in identifying epigenetic units is that our knowledge of epigenetics is still insufficient to recognize them on the basis of genome-wide measured epigenetic markers. However, this will change rapidly with the development of highly efficient ChIP-seq technologies (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007).

5. Boundary elements, insulators and enhancer blockers: boundary elements mark the two borders of epigenetic units. So far, only a limited number of boundary elements (or insulators) has been identified (van der Vlag *et al.*, 2000). Enhancer blockers act as regulatory elements that control enhancer-promoter interations. Insulator is a term used for both of these two types of elements, as we often cannot discriminate between their function at present.

System variables

1. Activity of genes: the activity level of a gene is often represented by a Boolean parameter, *i.e.* on or off, whereas in reality gene activity is a continuous variable. Currently, we have no simple and accurate method for quantifying gene activity. Most networks up to now are based on microarray data, which measure levels of RNA, rather than rates of RNA synthesis (Harbers and Carninci, 2005).
2. Activity state of TFs: the activity state of TFs depends on regulation by post-translational modifications of the TF and/or translocation between the nucleus and the cytoplasm.
3. Occupation of TF binding sites: this variable controls the activity state of genes and epigenetic units. It should be noted that gene activity and the epigenetic chromatin state often rely on the combined binding of multiple TFs to regulatory regions in the genome.
4. on/off state of epigenetic units: the epigenetic state is likely to be related to chromatin structure and can be inferred from the presence of histone modifications, DNA methylation and the incorporation of variant histones (Bird, 2002; Henikoff *et al.*, 2004; Turner, 2007). Not all histone modifications relate to the epigenetic state, some modifications are pertinent to the individual gene state. For instance, H3K9 di- and trimethylation generally mark epigenetically non-permissive loci, whereas H3K4 trimethylation and acetylation of histones H3 and H4 are components of regulatory systems that act on individual genes (Turner, 2007).
5. Activity state of boundary elements/enhancer blockers: the activity of boundary elements and enhancer blockers may be regulated as well. An example for the differential activity of enhancer blockers is best illustrated by the well-described Igf2/H19 imprinting locus (Schoenherr *et al.*, 2003). In

this locus, the LCR functions as an enhancer blocker to either the Igf2 or the H19 gene, dependent on its DNA methylation status and subsequent binding of the CCCTC-binding factor (CTCF) protein (Bell and Felsenfeld, 2000; Hark *et al.*, 2000; Szabo *et al.*, 2000). Similarly, binding by factors such as CTCF and non-coding RNAs potentially activate boundary elements (Guelen *et al.*, 2008; Rinn *et al.*, 2007).

These network components define the eukaryotic gene networks, whereas the values of the variables define the state of the network. Switching between network states may be achieved by changing intracellular or extracellular cues, such as cell-cycle signals or hormones, respectively.

Additional levels of gene control

The systematic enumeration of all components and variables does not necessarily add up to a comprehensive eukaryotic gene network. Additional factors and regulatory levels may exist that have an impact on gene regulation. For instance, the spatial folding of the genome inside the cell nucleus (at nuclear, chromosomal and subchromosomal level) may have implications for gene activity (reviewed by Goetze *et al.*, 2007) possibly resulting in chromatin-chromatin interactions at different genomic length scales. For instance, multiple transcription factor-mediated chromatin interactions have been identified at the β-globin locus by 3C and 5C (Chromosome Conformation Capture (Carbon Copy)) techniques between the LCR and the β-globin genes (Dostie *et al.*, 2006; Drissen *et al.*, 2004; Vakoc *et al.*, 2005). Another example is the possible formation of transcription factories, bringing together genes that are far apart on the linear genome (Osborne *et al.*, 2004). The functional significance of these and other chromatin-chromatin interactions remains to be clarified.

The black box

By their very nature, gene-gene interactions are indirect and involve many steps in a variety of cellular subsystems. The most direct type of gene-gene interaction is that of a gene coding for a TF, which binds to a regulatory sequence of another gene. Such an interaction is often depicted by a simple edge in a graph, but in reality includes many steps. These include pre-mRNA synthesis, RNA processing, RNA transport to the cytoplasm, interaction with ribosomes, protein synthesis, protein import into the nucleus, (frequently) posttranslational modification and binding to its regulatory site in a promoter or enhancer. Often, gene-gene interactions are much more complex, involving signal transduction or metabolic networks. These multi-step processes are not stated explicitly in gene networks and are treated as black boxes. The use of black boxes is acceptable if all components and interactions inside the black box are constitutively present and active, *i.e.* when their quantity and activity is not regulated. Obviously, this is almost never the case in gene-gene interactions. Currently, in most

Table 1. Overview of commonly used, mostly high-throughput techniques used to determine factors critical to genome-wide regulation.

system component or variable	genome-scale technique	reference
Genomic sequence elements, *e.g.* genes, TF binding sites, insulators	high throughput sequencing, computational comparative sequence analysis	reviewed in Mardis, 2008
Binding sites for TFs and regulatory proteins	Dam-ID	van Steensel and Henikoff, 2000
Regulatory sites on the basis of nucleosome depletion	DNAseI sensitivity and FAIRE	Crawford *et al.*, 2006a; Crawford *et al.*, 2006b; Giresi *et al.*, 2007; Lee *et al.*, 2004; van Steensel and Henikoff, 2000
TF proteins and their activity state	proteomics	Conrotto and Souchelnytskyi, 2008; Graham *et al.*, 2005; Patterson and Aebersold, 2003
TF binding kinetics	light microscopy: FLIP, FRAP, FRET	Hoogstraten *et al.*, 2002; Houtsmuller *et al.*, 1999; Matyus, 1992; Phair and Misteli, 2000
Occupation of TF protein-binding sites	ChIP-chip, ChIP-seq	Barski *et al.*, 2007; Blat and Kleckner, 1999; Mikkelsen *et al.*, 2007; Ren *et al.*, 2000
Binding of regulatory RNAs	no methods yet	
on/off state of genes	micro-array, RNA-seq	reviewed in Rando, 2007; Wang *et al.*, 2009
epigenetic state of chromatin	ChIP-Seq and quantitative ChIP methods: SACO (Serial Analysis of Chromatin Occupancy) and G-MAT (Genome-wide MApping Technique), bisulfite sequencing, meDIP	reviewed in Callinan and Feinberg, 2006; Rando, 2007; Schones and Zhao, 2008
chromatin compaction	4C and 5C techniques	Dostie *et al.*, 2006; Simonis *et al.*, 2006

cases it is not possible to integrate all relevant steps in gene-gene interactions. This problem should be addressed in any type of gene network. Solving this problem is essentially equivalent to integrating metabolic and signal transduction networks, as well as cell-cell signaling systems, in gene-gene networks. Difficult as this may be, there is no way around it.

Modeling epigenetic units in eukaryotic gene networks

Once epigenetic regulation units have been incorporated into a computational representation of genetic regulatory networks, the next step is to study the dynamics and to analyze its structure. In the absence of detailed descriptions of hierarchical regulation levels in real systems, a first approach is to study theoretical models. Continuous and stochastic modeling approaches have yielded detailed insights into the workings of small natural gene networks (Chabot *et al.*, 2007; Suel *et al.*, 2006; van Hoek and Hogeweg, 2006; van Zon *et al.*, 2007) and synthetic gene networks (Elowitz and Leibler, 2000; Kobayashi *et al.*, 2004; Stricker *et al.*, 2008). However, for larger natural networks this approach becomes unfeasible, because the required parameters (*e.g.*transcription rates, TF-DNA binding constants, TF-TF binding constants) are currently unknown at the large scale. Boolean approaches only consider 'on' or 'off' states of genes and can give us better insight into the switching logic of gene networks. Boolean approaches were introduced in the pioneering work of Kauffman, Thomas and Glass in the late sixties and early seventies (Glass and Kauffman, 1973; Kauffman, 1969; Thomas, 1973). Related approaches are still widely used today (for recent examples see Istrail and Davidson, 2005; Li *et al.*, 2006; Mendoza *et al.*, 1999; Shmulevich *et al.*, 2005). These 'logical' gene network analyses represent genetic regulatory circuits using networks of genes and regulatory units that assume an 'on' or an 'off' state, while the edges represent TFs produced by the other genes in the networks. Boolean functions (*e.g.* 'AND' or 'OR') determine whether the TFs act as an activator or as an inhibitor of gene expression. For example, assuming two TFs can bind to each promoter, the 'AND' function would indicate that both TFs are required for transcription to proceed, while an 'OR' function would indicate that either one of them would suffice. Other Boolean functions express more complicated inhibitory and activating interactions between the TFs and the promoter.

Although hierarchical regulation levels are currently lacking from datasets of biological regulatory networks, we can already explore the role of hierarchical regulation in dynamic models of random regulatory networks. Figure 1 illustrates how we might incorporate epigenetic regulation into Kauffman's Boolean network formalism. Figure 1A shows a standard, randomly connected, 'flat' Boolean network of 8 nodes and two TFs regulating each node, with Boolean functions chosen at random. After randomly putting each of the nodes in an 'on' or 'off' state, we iteratively update the state of each node in parallel. In this example, after an initial transition the network alternates between two expression patterns, called a state cycle.

This is typical behavior for Boolean networks: the system can reach one out several steady states or state cycles, depending on its initial pattern. These steady states or state cycles, called *attractors*, are often considered models of cell types, while the initial transitory phases—or the transition to a different attractor after an external perturbation—might resemble differentiation (Kauffman, 1969; Shmulevich *et al.*, 2005).

In panel B we add a hierarchical level of description. We insert an insulator into the network, creating two regions of hierarchical regulation. For simplicity, we assume the region assumes a 'permissive' state (*i.e.* the state in which it is accessible for

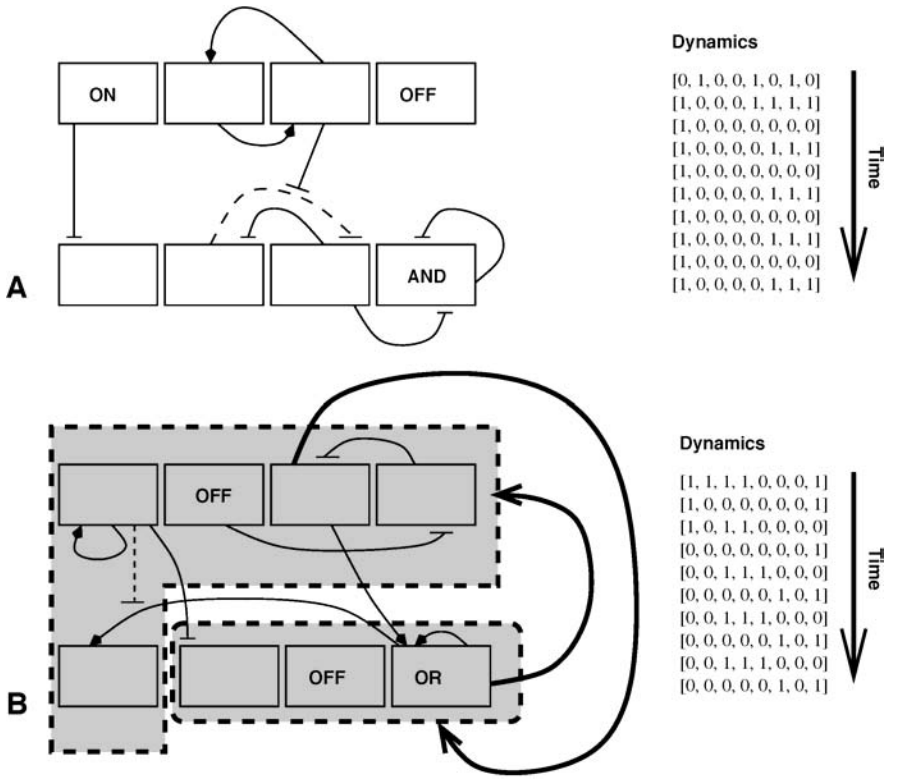

Figure 1. Boolean networks, with 8 genes receiving input from two transcription factors. Non-functional inputs were removed to improve graphical representation. Activating interactions shown with arrow heads, inhibitory interactions shown with flat arrow heads. Boolean functions indicate combinatory dependence on inputs; OR: either input, or both, can be present for their effect on transcription; AND: both inputs must be present; OFF: gene is never transcribed; ON: gene is always transcribed. Dynamics with synchroneous updating shown on the right with random initial conditions and time pointing downwards. A) Standard Boolean network. B) Boolean network with two hierarchically regulated regions shown as grey regions, regulated by genes 3 and 8 respectively. Regions are in the permissive state if the regulator is switched on.

transcription) if its regulator (chosen at random) is in the 'on' state. Here, after an initial transition state, the two subnetworks alternate between a permissive and a closed, non-permissive state.

We are currently systematically exploring the behaviour of larger Boolean networks with several insulators. Interestingly, our first results indicate that the hierarchical regulation level tends to reduce both the number and length of the attractors found in Boolean networks. Thus we may speculate that, apart from the perhaps obvious role of hierarchical regulation in compartmentalizing functionality required in only a subset of cell types, hierarchical regulation might stabilize and structure regulatory networks. The insights gained using these theoretical explorations will help to interpret the role of hierarchical regulation in biological networks once detailed datasets of eukaryotic networks, including the hierarchical regulation units, have become available.

Acknowledgements

We thank dr. Jana Rudolph for valuable discussions. This work was supported by the Netherlands Organisation for Scientific Research (NWO; project number VIDI2003/03921/ALW/016.041.311).