Review

Challenges for modeling global gene regulatory networks during development: Insights from Drosophila

Bartek Wilczynski a, b, Eileen E.M. Furlong a, * a European Molecular Biology Laboratory (EMBL), D-69117 Heidelberg, Germany

b Institute of Informatics, University of Warsaw, Warsaw, Poland

A R T I C L E   I N F O

Article history:
Received for publication 21 July 2009
Revised 14 October 2009
Accepted 21 October 2009
Available online 27 October 2009
doi:10.1016/j.ydbio.2009.10.032

Keywords:
Transcriptional networks
Gene regulatory networks
Developmental networks
Modeling
Drosophila
CRMs

A B S T R A C T

Development is regulated by dynamic patterns of gene expression, which are orchestrated through the action of complex gene regulatory networks (GRNs). Substantial progress has been made in modeling transcriptional regulation in recent years, including qualitative “coarse-grain” models operating at the gene level to very “fine-grain” quantitative models operating at the biophysical “transcription factor-DNA level.” Recent advances in genome-wide studies have revealed an enormous increase in the size and complexity of GRNs. Even relatively simple developmental processes can involve hundreds of regulatory molecules, with extensive interconnectivity and cooperative regulation. This leads to an explosion in the number of regulatory functions, effectively impeding Boolean-based qualitative modeling approaches. At the same time, the lack of information on the biophysical properties for the majority of transcription factors within a global network restricts quantitative approaches. In this review, we explore the current challenges in moving from modeling medium scale well-characterized networks to more poorly characterized global networks. We suggest to integrate coarse- and find-grain approaches to model gene regulatory networks in cis. We focus on two very well-studied examples from Drosophila, which likely represent typical developmental regulatory modules across metazoans.

© 2009 Elsevier Inc. All rights reserved.

Introduction

Gene regulatory networks (GRNs) have been used to describe and model regulatory mechanisms since the late 1960s (Britten and Davidson, 1969; Kauffman, 1969, 1993; Thomas, 1973). As more and more data on the molecular mechanisms of regulatory interactions have become available, GRNs have gained importance and accuracy in modeling regulatory processes across all biological systems and species (Brown et al., 2002; Davidson, 2006; Davidson and Erwin, 2006; Davidson and Levine, 2008; Levine and Davidson, 2005). Particularly in the field of development, GRNs have proved to be a very useful tool to describe and explain complex dependencies between key developmental transcription factors (TFs) and their target genes (Davidson et al., 2002).

GRNs are typically described as network models where the dependencies between genes are depicted by a directed graph, whose nodes represent genes and edges lead from a regulator (often a TF) to its targets. The edges ideally represent dependencies at the transcriptional level, which are mediated by TF binding to a regulatory region in the vicinity of the target gene. In order for the

* This article was accepted in 2009 to celebrate the 50th anniversary of Developmental Biology.
* Corresponding author. Fax: +49 6221 387166.
E-mail address: furlong@embl.de (E.E.M. Furlong).

GRN model to be accurate, each edge should be experimentally verified, i.e. have evidence for direct regulation, including binding of the regulator to a cis-element leading to the expression of the target gene. Typically this is obtained from in vivo transgenic reporter assays combined with a mutation of the putative TF binding sites, for example (Inoue et al., 2005; Ririe et al., 2008). Therefore, three pieces of information are needed to generate an accurate GRN model; the spatio-temporal expression pattern of the TFs, information on which cis-regulatory modules (CRMs) they bind to and regulate and a causal link between the TF activity and the target genes’ expression. The latter is usually obtained from genetic perturbation experiments, where the effect of removing a TF’s activity on other regulators’ expression is assessed using in situ hybridization or expression profiling (Imai et al., 2006). Give the laborious task in verifying all edges and nodes in a GRN, many developmental networks are at a small to medium size, having therefore a tradeoff between accuracy versus completeness (for a more detailed discussion on this issue, see Levine and Davidson, 2005).

A good example illustrating the success of GRN approaches to development is the extensively studied model of endomesoderm development in the sea urchin (Davidson et al., 2002), where the network is mature enough to provide a causal explanation of cell fate decisions (Oliveri et al., 2008). Efforts are ongoing to build comprehensive gene regulatory networks in a number of other systems, including the development of the endomesoderm in Xenopus (Koide et al., 2005; Loose and Patient, 2004), vulva...
development in Caenorhabditis elegans (Inoue et al., 2005; Ririe et al., 2008), segmentation along the Drosophila embryo’s anterior–posterior (A-P) axis (Nasiadka et al., 2002; Scott and Carroll, 1987) and dorsal–ventral (D-V) axis (Stathopoulos and Levine, 2005; Stathopoulos et al., 2002), in addition to a number of well-studied smaller GRNs directing development of the Drosophila eye (Silver and Rebay, 2005), eggshell patterning (Lembong et al., 2009), bristle formation (Ghysen and Thomas, 2003) and cell fate specification in the mesoderm (Bonn and Furlong, in press; Furlong, 2004; Sink, 2006).

The A-P segmentation network is perhaps the best-studied transcriptional network in Drosophila development and involves a large number of TFs including: Bicoid (Bcd), Hunchback (Hb), Knirps (kni), Giant (Gt) and Krüppel (Kr), which are present in gradients along the A-P axis. The combinatorial action of activators and repressors functioning over different concentration thresholds leads to very specific patterns of pair-rule gene expression (reviewed in Nasiadka et al., 2002). For example, the regulation of *even skipped* (*eve*), which is expressed in a stereotypic pattern consisting of 7 stripes in the early embryo, has been extensively studied (Small et al., 1992, 1993; Stanojevic et al., 1991). Fig. 1A shows a schematic of the GRN underlying *eve* expression, including only its immediate regulators. Regulatory connections between the regulators themselves are omitted for clarity. Even in this simplified view it is clear that although the expression patterns, relative concentrations and regulatory effects of all regulators are known (Janssens et al., 2006), it is a significant challenge to accurately predict the expression pattern of *eve* as a function of these inputs.

Similarly complex is the network governing the subdivision and specification of the Drosophila mesoderm into different tissue primordia, including three muscle types (somatic, visceral and heart) as well as the fat body and gonadal mesoderm. The early mesoderm network is an interconnected subnetwork of the dorsal–ventral patterning system (Hong et al., 2008b). Once the presumptive mesoderm is established, its subsequent development is orchestrated by the bHLH TF Twist (Baylies and Bate, 1996; Sandmann et al., 2007), its downstream targets: *tinman* (Yin et al., 1997), *Mef2* (Junion et al., 2005; Nguyen and Xu, 1998; Sandmann et al., 2006) and a number of other TFs (Azpiazu and Frasch, 1993; Jakobsen et al., 2007). Fig. 2A

---

![Diagram A](image1.png)  
**Fig. 1.** (A) *even-skipped* gene regulatory network including inputs known to be involved in regulating the expression patterns of *eve* in the blastoderm embryo and in the cardiogenic region. Arrows represent verified direct transcriptional input. The expression patterns of *eve* in the blastoderm embryo (7 stripes) and later in the heart (14 segmental groups of cells on either side of the embryo) are shown schematically below the *eve* gene. Regulatory connections between inputs, as well as the neuronal enhancer (McDonald et al., 2003), are omitted for clarity. Recent evidence also suggests an involvement of *cad* and *tll* in the formation of stripe 2 (Janssens et al., 2006; MacArthur et al., 2009); however, we have omitted these factors as their direct regulatory input has not been verified. (B) The *cis*-regulatory *eve* network describes the same data; however, individual enhancers (or *cis*-regulatory modules (CRMs)) are indicated as colored boxes with their names in labels. The expression patterns of each CRM, shown in the embryo diagrams below, are color-coded accordingly. To make the connections easier to track, colored circles are placed beside each regulator indicating the enhancers it regulates. The labels stand for (from left to right) minimal autoregulatory sequence (MAS) (Jiang et al., 1991a), Stripe 3 + 7 enhancer (Small et al., 1993), Stripe 4 + 6 (Fujikoa et al., 1999), Stripe 1 (Fujikoa et al., 1999), Stripe 5 (Fujikoa et al., 1999), mesoderm-heart enhancer (Halfon, 2000) and proximal promoter element. Note that the order and location of CRMs was chosen to facilitate the clarity of the drawing and does not reflect their genomic location relative to the transcriptional start site of *eve*. prd = paired; twi = twist; tin = tinman; mad/med = Mothers against dpp/medea; pnt = pointed; dTCF = T cell factor/pangolin; gt = giant; Kr = Krüppel; Stat92E = Signal-transducer and activator of transcription protein at 92E; Bcd = Bicoid; hb = hunchback; D = Dichaete; kni = knirps; ttk = tramtrack. 

---

![Diagram B](image2.png)
CRMs are typically relatively short stretches of DNA (~300–500 bp), which integrate multiple inputs from different TFs to give rise to a distinct output of spatio-temporal expression (Arnosti, 2003; Davidson, 2006; Kirchhamer et al., 1996; Small et al., 1993; Visel et al., 2009a). CRMs therefore act as integration platforms that contain the sequence information to perform a defined regulatory function.

An important feature of CRMs is the modularity of their activity, allowing CRM function to be assessed independently of each other (Visel et al., 2009a; Yin et al., 1997). Although this facilitates experimental studies on CRM activity, it is still not clear how to define the exact boundaries of CRMs. In the past, this was typically deduced empirically by defining the minimal region showing activity. More recently, the boundaries of CRMs are often estimated using sequence conservation. While such de novo CRM boundaries in a living cell.

Despite the detailed level of knowledge underlying the networks depicted in Figs. 1A and 2A, the topology itself is not sufficient to make a predictive model of the target gene’s spatio-temporal pattern of expression. The critical missing component is the regulatory function
that integrates multiple inputs into a coherent single output. Recent years have seen significant effort in developing mathematical models to address this. Modeling approaches based at the gene level have mainly been applied to the GRNs controlling the A-P (Albert and Othmer, 2003; Meinhardt, 1986; Perkins et al., 2006; Sánchez and Thiery, 2001) and D-V (Sanchez et al., 1997) patterning of the Drosophila embryo. These methods typically use Boolean or non-linear Boolean-like functions (De Jong, 2002; Thomas and D’Ari, 1990) aimed at providing standardized representations of regulatory functions and are based on a combination of genetic interaction data and direct cis-regulatory information.

Although successful in predicting relatively simple regulatory connections, these methods face an inherent complexity problem when trying to integrate large numbers of inputs. Namely, the number of possible functions for a given gene increases super-exponentially with an increase in the number of inputs (Edwards and Glass, 2000). The regulation of eve provides a good illustration of how fast that growth is. This gene has 14 regulatory inputs (Fig. 1A, excluding the neuronal enhancer), which can potentially give rise to 2^{16384} different Boolean regulatory functions, far more than the estimated number of atoms in the universe. Selecting a correct function from such an astronomical number of possibilities is difficult, as often multiple functions can agree with the experimental data. Assuming simple regulatory rules, such as “regulatory functions are additive” and “regulator input can only be strictly positive or negative” (Ma’ayan et al., 2008), would reduce the number of possible functions substantially. However, these assumptions do not reflect the biological regulatory complexity of most developmental systems.

While Boolean regulatory functions are appealing for their simplicity of interpretation, a number of alternatives have also been proposed (for a more detailed overview, see Bentaboule, 2009; Gerard and Willadsen, 2009; Kim et al., 2009). Differential equations, for example, have been used to model cell cycle regulation (Csikasz-Nagy et al., 2006), and in particular piecewise-linear systems are able capture Boolean-like behavior of GRNs (Snoussi, 1989). Kinetic-like models, using Hill function, were used to generate a non-linear output from additive inputs of different TFs to model canalization of gene expression patterns in the Drosophila blastoderm embryo (Manu et al., 2009). Linear additive models have been applied to large-scale studies of regulatory interactions in yeast (Buisson et al., 2001) because of their simplicity. Another relatively young branch of approaches is based on a probabilistic framework, where both the uncertainty of measurement and stochasticity of the process can be handled naturally by the model (Segal et al., 2003a; Shmulevich et al., 2002). While this approach is very promising, there is currently little understanding of the role of stochastic events during development, and these are very difficult to quantitate experimentally. Modeling regulatory functions for genes with multiple inputs is difficult for all the approaches mentioned: they either face a combinatorial explosion of possible models (Boolean, piecewise-linear) or require multiple parameters, which are difficult to obtain experimentally (kinetic, probabilistic).

Modeling GRNs at the cis-regulatory level rather than at the gene level can at least in part alleviate this complexity issue and is essential in cases where TFs provide both negative and positive inputs to a gene’s expression, for example the input of Hunchback to eve expression (Figs. 1A and B). At first glance, including explicit information about cis-elements into GRNs might seem to complicate the model topology, as the number of CRMs may be larger than the number of genes by as much as an order of magnitude. However, including explicit cis-elements information clarifies which inputs act through the same CRM and therefore can function combinatorially and effectively lowers the number of inputs into one regulatory element. This therefore reduces the complexity of regulatory functions for a gene into multiple simpler functions that model each cis-element separately.

**Viewing GRNs in cis**

Using the current knowledge of CRMs involved in a gene’s regulation, the framework of GRNs can be extended to include multiple separate “input points” corresponding to distinct CRMs, as suggested by Bolouri and Davidson (2002). This has been used in the graphical representation of GRNs in Xenopus, C. elegans and sea urchin for single genes, e.g. Gac (Koide et al., 2005), egl17 (Inoue et al., 2005), otx (Yuh et al., 2004) and blimp1 (Liv and Davidson, 2006); however, this has not been scaled to all known CRMs within a network. Here we illustrate the advantage of having individual CRM inputs explicit in the GRN model using two examples from Drosophila. Fig. 1B presents how this approach can be applied to better understand the regulation of a single gene eve and Fig. 2B presents a small cis-regulatory network (CRN) consisting of twist, tinman and mef2. The data represent a synthesis of the currently available information on the regulation of these genes to the best of our knowledge.

The eve locus contains 9 CRMs (Fig. 1B, we omitted the neuronal CRM (McDonald et al., 2003) for clarity), 5 of which act at the same stage of development to produce striped expression in different spatial domains of the early blastoderm embryo, while the mhe-CRM regulates expression during cardiac specification at later stages of development. Similarly, the tinman locus contains 4 CRMs (Fig. 2B) that drive expression throughout the unspecified mesoderm, head mesoderm, dorsal mesoderm and cardioblasts (Yin et al., 1997). In the eve and tinman loci, the maximal number of inputs into a single CRM is significantly smaller than the total number of regulators for a target gene, e.g. 5 inputs into the eve-mhe-CRM versus 14 inputs to the eve gene (compare Fig. 1A with Fig. 1B).

This cis-level presentation of GRNs also has the advantage of making a number of regulatory properties readily apparent. First, it highlights which TFs co-occupy the same CRM, indicating combinatorial and perhaps cooperative regulation, versus TFs that bind to separate CRMs, thereby having a more independent contribution to the gene’s expression. For example, Twist and Tinman regulate Mef2 expression through the same CRM, suggesting that these TFs may act synergistically to trigger early expression, while Mef2 auto-regulates itself using a separate CRM responsible for maintaining Mef2 expression at later stages of development (Fig. 2B). Second, it reveals missing TFs in the regulatory network. For example, the tin-A CRM in the tinman locus drives specific expression in the head mesoderm, although the identity of the TFs regulating this expression remains unknown (Yin et al., 1997) (indicated by a question mark in Fig. 2B). Third, it emphasizes the ability of a TF to provide multiple inputs into a single target gene’s expression. For example, Hb contributes to the activity of four CRMs within the eve locus (Fig. 1B). It is essential to take each of these regulatory inputs into account to accurately predict the effect of loss-of-function mutations in the Hb gene. While the view depicted in Fig. 1A suggests that all eve stripes are lost in Hb mutants, the cis-regulatory view predicts that stripe 5 should remain unaffected at stage 5, whereas the remaining eve stripes should be disrupted. The situation is, in fact, more complex as Hb has positive regulatory activity on some CRMs (e.g. str2) and negative regulatory effects on other CRMs (e.g. str3 and 7 and str4 and 6). These dual regulatory inputs from the same TF are readily apparent in the cis-diagram (Fig. 1B) but cannot be accurately represented in simple GRN models that assume strictly negative or positive regulatory interactions (Fig. 1A).

Finally, representing GRNs in cis facilitates modeling mutants that affect regulatory connections (edges) rather than genes (nodes). There is growing evidence that cis-mutations are associated with developmental defects and diseases in higher Eukaryotes, for example
the Human Gene Mutation Database lists currently more than 1500 cis-regulatory mutations associated with human disease states (Stenson et al., 2009). Moreover, cis-regulatory mutants may lead to more subtle phenotypes. For example, a mutation in the binding site for Hb in the str2 CRM will only affect eve expression in the second stripe, in contrast, loss-of-function mutations in Hb itself will disrupt four stripes of eve expression. Modeling GRNs with explicit cis-regulatory connectivity may provide a more comprehensive framework to understand how cis-regulatory mutations lead to specific phenotypes.

An important property of cis-regulatory networks is that they still can be visually represented as GRNs with single CRMs by “collapsing” the modular CRM architecture (Bolouri and Davidson, 2002). This is especially important with large networks where simultaneous display of all CRMs will make the visualization of the whole network overly complicated.

**Modeling regulatory activity at the CRM level**

A number of model-based approaches have been used to predict gene expression in a variety of systems, including probabilistic (Segal et al., 2003b; Tavazoie et al., 1999), thermodynamic (Bintu et al., 2005; Gertz et al., 2009; Granek and Clarke, 2005; Janssens et al., 2006; Segal et al., 2008; Zinzen and Papatsenko, 2007; Zinzen et al., 2006) and reaction-diffusion approaches (Eldar et al., 2002; Jaeger et al., 2004b; von Dassow et al., 2000). While many were conducted in yeast and therefore focused on modeling levels of gene expression (Beer and Tavazoie, 2004; Gertz et al., 2005; Granek and Clarke, 2005; Segal et al., 2003b; Tavazoie et al., 1999), a number of studies have modeled spatio-temporal expression during metazoan development. Notably, Beer and Tavazoie (2004) showed that their probabilistic method for predicting temporal expression patterns from sequence features in yeast can also be applied to a subset of genes expressed during C. elegans development.

Given the comprehensive knowledge of the Drosophila A-P and D-V networks gained from mutagenesis analyses, biochemical studies and genetic dissection, these systems have become the subject of choice for quantitative modeling to predict spatio-temporal CRM activity during metazoan development. For example, a fractional site occupancy approach was used to model ventral neurogenic ectodermal (VNE) CRM activity (Zinzen et al., 2006). A pioneering study by Janssens et al. (2006) goes a step further and uses a thermodynamic model of TF binding to the eve stripe 2 enhancer, which is predictive with regard to quantitative expression along the A-P axis, both in wild-type and mutant embryos. More recently, Segal et al. (2008) applied a thermodynamic model to 44, somewhat heterogeneous, CRMs within the segmentation network (including the eve stripe 2 enhancer), leading to quite accurate predictions of gap gene expression.

The results of these studies indicate that while we have a general understanding of the roles of TF-DNA and TF-TF interactions in transcriptional regulation, there are still many physical parameters that need to be estimated and that are difficult to assess experimentally (such as the cooperativity between TFs). This may explain why the success of generalizing these models to a wider range of CRMs has been limited to CRMs with relatively simple expression patterns (Levine, 2008; Segal et al., 2008). These models currently have not been extended from predicting CRM activity to forming an integrative model to predict the complete expression pattern of the associated gene. For example, in the eve locus, this would require combining the regulatory model of each of the stripe enhancers to predict the gene’s expression at stage 5 (Fig. 1B).

The extent of assumptions that need to be taken into account for most modeling approaches restrict their use to well-characterized systems of limited size. In all cases, these approaches require precise measurements of the relative concentrations of the TFs involved (Janssens et al., 2005). In addition, detailed knowledge of a number of inherent properties of the TFs is required or assumed, including their affinity to various sequence motifs (Bintu et al., 2005; Eldar et al., 2002; Granek and Clarke, 2005; Jaeger et al., 2004b; Janssens et al., 2006; Segal et al., 2008; von Dassow et al., 2000; Zinzen and Papatsenko, 2007; Zinzen et al., 2006), and in some cases their rates of diffusion, transcription, translation and degradation (Eldar et al., 2002; Jaeger et al., 2004a; von Dassow et al., 2000). It is currently not clear how well these approaches will perform with more complex CRMs and whether they are scalable to more global networks. Perhaps the greater challenge is from the experimental side to extrapolate precise measurements of the concentrations of all TFs, obtain good estimates of their DNA binding specificities and an understanding of how these are modulated by chromatin context and cooperative effects of other TFs. Even in a very “mature” model organism like Drosophila, the DNA binding specificities for only a fraction of predicted TFs are known, and there is still debate over the exact number of TFs in the genome.

**Scaling up: moving from highly characterized subnetworks to poorly characterized global networks**

Our detailed knowledge of the regulatory function governing the expression of eve is the result of the combined effort of many labs during the past two decades (Arnosti et al., 1996; Fujikawa et al., 1999; Gray et al., 1994; Janssens et al., 2006; McDonald et al., 2003; Ohtsuki and Levine, 1998; Perkins et al., 2006; Sackerson et al., 1999; Small et al., 1992, 1993, 1996; Stanoevici et al., 1991). Similarly, uncovering the modular structure of the endo16 locus in the sea urchin has taken many years of detailed and methodical experimentation (Yuh et al., 2001; Yuh and Davidson, 1996). Collectively, these studies have lead to much of our understanding of the basic principles of transcriptional regulation in developmental systems. Complementary studies in recent years examining transcriptional networks at a global level have revealed that eukaryotic cells express thousands of genes at any given stage (Arbeitman et al., 2002; Chintapalli et al., 2007; Furlong et al., 2001; Manak et al., 2006) and that TFs occupy thousands of binding sites (Li et al., 2008; MacArthur et al., 2009; Sandmann et al., 2007; Zeitlinger et al., 2007). Given this explosion in regulatory information, how can we expand the detailed models of well-defined networks, (Janssens et al., 2006) to understand cis-regulation in more global, highly interconnected networks?

A number of studies are taking steps to move from small-scale GRNs to larger scale networks. For example, expression studies examining the effect of knocking down one of 76 TFs on the expression of each of the other 75 factors by in situ hybridization generated a comprehensive regulatory network of early development in the chordate Ciona intestinalis (Imai et al., 2006). Similarly, an initial framework for the network regulating T-lymphocyte specification was generated using global expression-profiling experiments in wild-type and mutant contexts (Georgescu et al., 2008). Although both studies are significant steps towards understanding global regulatory activity (Lemaire, 2006), it is important to note that the authors describe these networks as “preliminary” since the connectivity between genes (edges) may represent indirect or direct regulation. In yeast, the effect of systematically deleting 263 TFs on gene expression was used in combination with regulatory epistasis and motif analysis to construct a functional regulatory network (Hu et al., 2007). While this provides an attractive approach to infer global regulatory networks in single cell systems, expression-profiling analysis of mutants has inherent problems in a multicellular developmental context. In the cases where TFs are essential for cell fate specification, maintenance of cell fate or cell viability expression profiling of loss-of-function mutants can yield misleading results as differential expression of genes may sometimes reflect an absence of the cell type of interest rather than transcriptional changes within
that cell type. Also, the signal can be skewed for genes that are expressed in multiple cell types.

For developmental regulatory networks, the verification of direct transcriptional regulation in cis remains the bottleneck in converting expression data into large-scale CRMs. The integration of expression-profiling data with TF occupancy data for the same TFs can reveal direct functional association between TFs and their target genes (Jakobsen et al., 2007; Li et al., 2008; Sandmann et al., 2006, 2007). However, this does not provide any information on which CRM is required or on the regulatory activity involved. Moreover, the activity of TFs serving to fine-tune or boost the levels of a gene’s expression will often be missed due to the insensitivity of the approaches applied but can be more easily uncovered by examining the activity of specific CRMs in isolation. A good example is the regulation of the eye gene by Tinman (Liu et al., 2009), where there is a 30% reduction in the levels of the eye gene’s expression in tinman mutant embryos, a decrease that is not readily observable by in situ hybridization. However, the activity of the Tinman-bound eye CRM is almost completely abolished in tinman mutant embryos. These more subtle affects may represent the majority of regulatory inputs and are therefore important for understanding the overall robustness of a gene’s expression.

Understanding an entire cis-regulatory system will require multiple levels of information, including (a) a comprehensive map of the location of all CRMs regulating all nodes within the CRM network (network components), (b) a detailed knowledge of the TF occupancy on these CRMs (inputs), (c) the spatio-temporal expression pattern driven by each CRM (output) and (d) an understanding of the regulatory logic leading to CRM activity (regulatory functions).

Recent years have seen an explosive growth in the identification of CRMs at genome-wide scales using both experimental and computational approaches. Experimentally, chromatin immunoprecipitation (ChIP) followed by either high-density tiling array hybridization (ChIP-on-chip) (Lee et al., 2002) or deep sequencing (ChIP-seq) (Johnson et al., 2007) is currently the method of choice to provide an unbiased, systematic global view of TF binding. Performing genome-wide ChIP experiments on a number of tissue-specific TFs provides a comprehensive map of the location of all regulatory regions involved in that tissues development (Jakobsen et al., 2007; Li et al., 2008; Sandmann et al., 2006, 2007). An alternative approach is to use dissected tissue to perform ChIP on general co-factors that are recruited to either activate or repress many CRMs (Visel et al., 2009b) or to use chromatin marks as indicators of transcriptional cis-elements (Heintzman et al., 2007). Computational methods, which are largely based on clustering TF motifs (Rajewsky et al., 2002) or to use chromatin marks as indicators of transcriptional activity (Emberly et al., 2003; Kheradpour et al., 2007), are complementary approaches that work with looking at evolutionary conservation (Emberly et al., 2003; Kheradpour et al., 2007). These approaches may impinge on a set of CRMs, but provides no information on when these regulators actually occupy the enhancers.

ChIP experiments on many TFs expressed in the same cell population provide crucial information on CRM input at the level of their combinatorial binding properties (Zinzen et al., in press) and when performed in a developmental time-course provide essential insights on the temporal occupancy of TFs in vivo (Jakobsen et al., 2007; Sandmann et al., 2006). Computational analysis of the sequence content within CRMs can complement ChIP data to identify additional regulators that may impinge on a set of CRMs, but provides no information on when these regulators actually occupy the enhancers.

Global ChIP-on-chip studies have revealed that TFs bind to thousands of regions in the genome, even in a relatively “simple” model organism like Drosophila (Jakobsen et al., 2007; Li et al., 2008; MacArthur et al., 2009; Sandmann et al., 2006, 2007; Zeitlinger et al., 2007), which is an order of magnitude higher than previously anticipated based on estimates of the number of target genes. The current challenge is to assess the contribution of a TF binding event to the regulatory function (output) of the CRM. Further studies are required to distinguish between TF binding events that are key mediators or “switchers” of a CRM’s activity, for example Twist binding to the CRM of the T48 gene (Sandmann et al., 2006), versus “modulators” acting to fine-tune CRM output (Hong et al., 2008a) or apparent non-functional “spurious” binding events (Li et al., 2008).

Drosophila melanogaster has been a particularly useful model organism to examine CRM output and assess the contribution of individual TFs in vivo. This system has a number of advantages for cis-regulatory analysis including (i) stable integration of transgenes into the genome, eliminating complications of assessing CRM activity in mosaic embryos; (ii) single-copy integration of CRMs, rather than tandem arrays as seen in some organisms, eliminates artificial inter-CRM cooperative or repressive interactions; (iii) the ability to collect tightly staged populations of embryos allows temporal aspects of developmental networks to be readily dissected; (iv) the recent development of the phiC31-integrase system (Bischof et al., 2007) allows all transgenes (therefore all CRM-reporter constructs) to be integrated into the exact same location of the genome, standardizing any potential positioning effects on CRM activity; and (v) experimental ease, which together allows for both a qualitative and quantitative analysis of CRM activity in vivo at any stage or in any cell type of the Drosophila embryo.

The activities of more than 700 Drosophila CRMs have been studied in vivo using transgenic reporter lines (Halfon et al., 2008), providing a detailed view of their spatio-temporal CRM activity. At the same time, mutational analysis of particular binding sites in a subset of CRMs have broadened our understanding of their regulatory logic; in particular, studies of genes such as eve (Fujioka et al., 1999; McDonald et al., 2003; Sackerson et al., 1999; Small et al., 1992, 1996; Stanoevich et al., 1991) and Mef2 (Cripps et al., 1998, 1999; Gajewski et al., 1997, 1998; Lovato et al., 2005; Nguyen and Xu, 1998) have demonstrated how complex expression patterns are generated through multiple distinct modular CRMs, each of which is responsible for a part of the total expression pattern (Figs. 1B and 2B). However, scaling up the dissection of the functional rules governing CRM activity (CRM logic) to 1000s of CRMs in vivo remains an essential, yet daunting challenge.

The binding of TFs to DNA and the action of TF complexes assembled on enhancers are physical processes constrained by biophysical properties of the molecules involved, including DNA molecules. Yet, the genetic rules governing how sequence-specificity of TFs gives rise to CRM output have so far remained elusive. We have recently demonstrated that high-resolution data on combinatorial occupancy is sufficient to predict CRM spatio-temporal activity in vivo (Zinzen et al., in press). Although this approach does not provide the regulatory logic of each CRM, it can pinpoint which TFs are most important for defining a particular pattern of expression pattern and demonstrates that CRM activity can be represented as a function of the temporal occupancy of multiple TFs. In the coming years, good estimates of DNA binding specificities for the majority of Drosophila TFs should become available using approaches like ChIP (Harbison et al., 2004), SELEX (Klug and Famulok, 1994) and DNA–protein arrays (Berger et al., 2006). Moreover, the spatio-temporal expression patterns of the majority of TFs will be characterized throughout embryonic development. These data, together with knowledge about the chromatin landscape, should facilitate occupancy models of a large number of CRMs, which can be integrated into more complex schemes (reviewed recently in BentabouleLeon and Davidson, 2009) describing the relations between TF occupancy and transcription rates of target genes. These models would be greatly enhanced by information on the relative concentrations of TFs in a cell at any given stage of development, which is currently lacking for the vast majority of regulators.
Conclusions and outlook

Given the current progress in the identification of CRMs, using both experimental and computational methods, the coming years should yield complete maps of the positions of the vast majority of regulatory sequences important for development in many species such as Drosophila. ChIP studies will provide an unprecedented view of combinatorial and temporal occupancy of CRMs at a genome-wide scale, an essential pre-requisite to understanding how regulatory activity drives gene expression within a network. One of the central challenges in developmental biology is to uncover general regulatory rules governing CRM activity, which can be used to model more global network activity. Modeling GRNs in cis should be a suitable approach to bridge the gap between genome-wide data, which can provide information on the binding properties of cis-regulatory elements and qualitative models that describe their functional connections at the gene level. While this should alleviate some of the classical complexity problems present in current Boolean approaches to simulate GRNs, it highlights the computational challenge to identify regulatory rules for individual CRM activity and for integrating the activity of multiple enhancers into a unified model, capturing the expression of the gene. Global transcriptional networks may require more accurate modeling approaches, finding better ways to estimate the thermodynamic parameters of current models as well as adding new components to their formulation.

In the discussion here, we have focused on the function of TFs themselves in developing models for GRNs. However, this information alone is unlikely to be enough to model gene expression in higher eukaryotes. Factors such as nucleosome positioning (Lam et al., 2008), chromatin status (Heintzman et al., 2007), CRM–promoter specific interactions (Deato and Tjian, 2007) and RNA PolII initiation versus elongation rates (Fuda et al., 2009) are all crucial parameters that affect the status of gene expression. A recent study in yeast incorporated sequence preferences for nucleosomes and transcription factors to model gene expression (Raveh-Sadka et al., 2009). A key bottleneck for integrating information like nucleosome positioning and chromatin remodeling in developmental networks is the lack of experimental data on these events in a tissue-specific and developmental stage-specific manner. Currently, these data are derived from either tissue culture cells or an average view from all cells in the embryo. New strategies are required to obtain these data in a cell type specific manner within the context of a developing embryo. Development is also driven by inductive cues from signaling cascades, which provide spatial and temporal information for cellular transitions. As the end-point of most signaling cascades is the activation of a transcription factor, it should be possible to identify their input on CRMs and use this information as a starting point to integrate signaling and transcriptional cascades. What is currently much more challenging is to integrate post-translational modification that affect the activity of TFs, switching them from an inactive to active state or promoting specific protein–protein interactions. Recent advances in mass spectrometry approaches such as SILAC (Andersen et al., 2005) and selected reaction monitoring (Gstaiger and Aebersold, 2009) should help to bridge this gap in the coming years.

While we are still quite far from generating predictive models for global developmental networks, we should see substantial progress in this direction in the coming years. Having such models in hand is essential for understanding metazoan development and predicting developmental defects associated with genetic perturbations.

Acknowledgments

We are extremely grateful to Denis Thieffry and John Reinitz for very useful comments and discussion. We are indebted to Ciaran Behan for artwork presented in Figs. 1 and 2. We thank all members of the Furlong lab for critically reading the manuscript. This work was supported by Deutsche Forschungsgemeinschaft grant FU 750/1–1 awarded to EF.

References


