Chapter 2

Gene Regulatory Networks

"Gene regulatory network" (GRN) is shorthand for the system of regulatory genes and their encoded interactions that determines the genetic functions to be expressed in cells of each spatial domain in the organism, at every stage of development. This includes the expression of regulatory genes (i.e., genes encoding transcription factors), genes that encode intercellular signaling functions, and genes that participate in downstream differentiation and morphogenesis functions. Since everything the cells do depends on the genes they express, the GRN control system essentially operates the developmental process. By linking the expression of every gene to its upstream transcriptional regulators, GRNs determine the coexpression of particular cellular functions exclusively in different cells. During development, the spatial expression of transcriptional regulatory genes is the driving force for the formation of the discrete structures of organisms, such as organs, body parts, and cell types. Therefore the most important function of GRNs, that is most directly causal for development of the specific body plan, is organizing the spatial allocation of

Genomic Control Process
http://dx.doi.org/10.1016/B978-0-12-404729-7.00002-2

regulatory gene expression. This determines the developmental fate and ultimately the differentiated cellular activity of the descendants of embryonic cells giving rise to each portion of the organism. GRNs are thus the "brains" of each act of the developmental process.

GRN theory emerges from diverse insights gained over the last quarter century regarding mechanisms of transcriptional control in animal development. This includes understanding of the expression of the regulatory apparatus itself, the structure and function of the DNA sequences that process regulatory inputs, the dimensions of the regulatory system operative in given contexts, and experimentally acquired knowledge of GRN structure/function relationships. We begin this chapter with a brief introduction to GRN theory, and then go on to discuss particular foundational principles of this concept. These include the Boolean nature of spatial gene expression in development, the combinatorial utilization of the transcriptional regulatory apparatus, the sequence-specific control of gene transcription by *cis*-regulatory modules (CRM), and the mechanisms of choice among alternative modules. Finally, this chapter concludes with a discussion of transcriptional dynamics and regulatory cascade behavior, which provides a first principles approach to thinking quantitatively about these regulatory processes. We revisit this dynamic treatment in the context of a discussion of GRN models in Chapter 6.

1. Introductory Overview of Developmental GRNs

1.1 GRN function and how GRNs are encoded in the genome

Every bilaterian animal consists of tissues, organs, and cell types, the function of which depends ultimately and completely on what genes each cell expresses. Yet as we all know, each of these cells contains an identical genome, with a few specialized exceptions. The relation between DNA sequence and gene expression activity is therefore not linear, in that DNA sequence per se does not directly predict the developmental process in the way that it does predict the proteins that can be synthesized. Instead, genomic control of development is encoded in a complex program which is active throughout the process and which functions to define the different parts of the organism by sequential specification. This program is distributed in many parts of the genome, those referred to in Chapter 1 as the regulatory genome.

GRNs provide the fundamental control mechanism directing developmental process. As briefly summarized in Chapter 1 (see also below), gene expression is regulated sequence-specifically by the interaction of transcription factors with *cis*-regulatory DNA modules. Thus, the control operations which assign diverse cellular functions are those determining when and where transcription factor encoding genes will be expressed. By encoding the *cis*-regulatory inputs of every regulatory gene, GRNs specify the interactions among regulatory genes that are responsible for the expression of particular sets of transcription factors. These transcription factors in turn also control cohorts of genes encoding many other kinds of protein, here referred to as effector genes, that is, differentiation genes and morphogenesis genes. Cells manifest their fates in development by the programmed activation of distinct suites of effector genes, directly determining their biological properties, the final specific readout of developmental GRNs. Thus ultimately the expression of all genes in the genome is linked by interactions within GRNs.

Regulatory genes have the special feature that they play dual roles in the GRN, in that their expression is at once the output of the upstream regulatory genes which provide their transcriptional inputs, and at the same time they provide inputs to other target genes within the same network. Thus, the set of transcription factors present in a given time and place determines the new set of transcription factors to be expressed, which then in turn establishes the expression of another regulatory condition. The continuous changes of regulatory gene expression in developmental time can be regarded as the major driver of developmental progression. Development is powered by changes in states of regulatory gene activity and as a consequence of these changes, new cell fates are established in the construction of the

body plan. Development is ultimately controlled by GRNs, and these constitute the primary machinery of control in Metazoa.

The genomic components of developmental GRNs are on the one hand the genes encoding transcription factors, signaling components, and effector functions, and on the other hand, the CRMs controlling the expression of all of these genes. "GRNs" denote the physical and functional relationships among regulatory genes. The term "network" in this context is not metaphorical but literal: the network structure grows directly out of two fundamental physical facts of regulatory life in animal systems. First, the CRMs which control spatial developmental gene expression function combinatorially in that they always require qualitatively multiple inputs, as we discuss below (i.e., inputs encoded by several different regulatory genes). Second, the outputs of regulatory genes, the targets of the transcription factors that they encode, are always qualitatively multiple as well (i.e., each regulatory gene has multiple targets). Therefore, each regulatory gene operates at the node of a network. Unlike certain other kinds of "interaction networks" such as protein:protein networks, these are directed, oriented networks in which information flows in only one direction, from transcription factor to *cis*-regulatory target sites and from regulatory genes to transcription factor production. Therefore GRNs are intrinsically hierarchical. Much of the ultimate functional import of GRNs in development and in evolution devolves from the hierarchical structure of these networks. The main point to be retained in this introductory discussion is that all of the general characteristics of developmental GRNs have as their physical basis the way gene regulation works in multicellular animal systems.

1.2 GRN hierarchy and modular organization

The basic elements of GRNs are the genes themselves, the modular *cis*-regulatory control systems that regulate gene expression at every level of the genetic hierarchy, and the causal directional relationships between regulatory and target genes. Regulatory interactions are denoted by statements such as "Gene 1 directly activates Gene 2", meaning that the transcription factor encoded by Gene 1 binds to and functions as an activator at a *cis*-regulatory target site in a regulatory module controlling the expression of Gene 2. We refer to this relationship as a "regulatory linkage" from Gene 1 to Gene 2 (in graph theory what we term "linkages" are referred to as "edges"). GRN linkages function in a strictly unidirectional manner, and irreversibility is a definitive property of animal development at every stage and phase, distinguishing development from other processes such as physiological response (Amit et al., 2009). The fundamental mechanistic basis for this irreversibility is the unidirectionality of the regulatory interaction between an upstream transcription factor and the *cis*-regulatory sequences of its target gene. This is intrinsically a one-way street. Even though the transcription factor–target site interaction is a reversible equilibrium reaction, the upstream gene controls the downstream gene.

As we consider more extensively in Chapter 6, the function of developmental GRNs emerge not just from the individual CRMs at its nodes, but from the structural features of the assemblages of genes that execute most of the developmental operations that GRNs perform. Individual CRMs set the conditions upon which each given gene is expressed or repressed, accounting causally (and in a causal sense entirely) for the activity of each gene. However, individual events of development never depend only on single genes, but always on expression of multiple regulatory genes and eventually on the expression of very large numbers of downstream effector genes. This requires that appropriate linkages among these genes be encoded in the GRN. Furthermore, there are many canonical developmental "jobs" that, as we shall see in the next three chapters of this book, every developmental process entails, each of which requires participation of multiple genes "wired" together and often including both activators and repressors. Examples of subprocesses or jobs that constitute the developmental process as a whole include formation of boundaries between spatial domains of gene expression; lockdown or stabilization of states of gene expression originally installed by transient inputs; reception of signals and organization of downstream regulatory

consequences; mediation of cell fate choice mechanisms; interpretation of initial inputs; and so forth. Such functions are executed by GRN "subcircuits". These subcircuits consist of small sets of genes (typically 3–8) which together execute a particular "job" according to a specific architecture or topology of their regulatory linkages. An important insight that has emerged from structure/function studies of such subcircuits is that the type of developmental function executed by a network subcircuit is determined by the subcircuit architecture, and that similar developmental functions are often executed by subcircuits of similar architecture, regardless of the particular genes they are composed of. We address subcircuit structure/function relationships in detail in Chapter 6. Suffice it to say here that the architecture, and hence function, of every GRN subcircuit is hardwired in the regulatory genome. A given developmental GRN will include several separate subcircuits joined by encoded regulatory linkages. Thus, considered from the perspective of the structural elements that perform its overall control functions, the developmental GRN has a modular character.

The two salient features of developmental GRN structure are its internal subcircuit composition and its strongly hierarchical organization. Both features directly reflect the intrinsic nature of the developmental process. Development from egg to adult body plan requires many phases, each phase depending on the preceding phase, and these phases occupy different successive spatial and temporal domains. Since each of these phases is controlled by a regional GRN, as a whole the GRN has de facto a deep hierarchical structure. That is, the earliest phases at the top of the GRN hierarchy give rise to the subdivisions of embryonic space controlled by the next hierarchical level. Subsequently, embryonic precursor domains undergo further subdivisions in the development of the body parts of the adult form, an increasingly complex process controlled by additional levels of GRN hierarchy. The GRN hierarchy terminates with the allocation of differentiation gene batteries and morphogenesis gene cassettes. The expression of these effector genes is directly controlled by the developmental GRNs, but since their gene products do not have gene regulatory functionality, these genes provide no further input into the regulatory circuitry. Therefore, they define the downstream periphery of the network. In contrast, at the upper levels of hierarchy, the network consists essentially of cross-regulatory circuitry.

1.3 Models of GRN topology

GRNs are statically encoded in the linear DNA sequence and this code is invariant in every cell of the organism throughout life. In every phase of development, given portions of the overall GRN interactions are utilized to direct the progressive creation of local regulatory states. Our problem is how to convey the network of regulatory interactions in play in any given phase of development in a meaningful way. The key objective is to represent graphically the origin and destination of every input into the *cis*-regulatory control loci of each gene in the system, and the destinations of the outputs of each gene, if it encodes a transcription factor. Network models that graphically represent genes as nodes and display regulatory interactions as input/output linkages are referred to as topological network models. In this book we deal essentially with one standard representation of topological models, although we also discuss mathematical models of different kinds in Chapter 6. Our standard mode of presentation of topological GRN models is by use of the BioTapestry computational and graphical platform (Longabaugh et al., 2005, 2009). In these BioTapestry models, individual spatial domains or phases of development are depicted in separate frames (for example, see Fig. 3.5I). The frames can be used to indicate the linkages which are in operation at each point in time and in each spatial domain in order to activate or repress target genes. All participating regulatory genes in each frame are linked by validated interactions, such that the inputs into each relevant *cis*-regulatory module and the outputs of the genes controlling these modules are explicit. This presentation contrasts with the standard epistasis diagrams traditionally used in genetic presentations which do not distinguish between direct and indirect interactions. BioTapestry models are genomically centered in that they explicitly represent the driver regulatory inputs into the genes occupying each node of the network.

That is, these models indicate the existence of specific binding sites for input transcription factors in the CRMs where the inputs terminate. In addition, these models clearly distinguish regulatory genes, signaling interactions, effector genes and off the DNA functions. Developmental GRN models are focused on control of spatial gene expression and do not in general include either ubiquitous biochemical functions or functions that affect transcript, levels only to a minor extent. For these reasons, except where they are shown to directly cause spatially specific gene expression, neither miRNAs nor epigenetic chromatin modifications are included (see Chapter 1).

The developmental GRN models with which we are here concerned are based on standards of evidence implied by the foregoing definitions. Direct causal evidence is required to demonstrate the existence of a functional GRN linkage. *Trans*-perturbations, in which a transcription factor input is removed from the system (by genetic mutation, treatment with morpholino antisense oligonucleotides or siRNAs, for instance) can reveal the requirement of this factor in the control of given target genes. *Cis*-regulatory analyses, including experimental assessment of the requirement of target sites for given factors within the regulatory sequences of the target gene, distinguishes whether observed *trans*-regulatory effects are direct or indirect. Additional evidence of many kinds can be useful for making this distinction (for example, Oliveri et al., 2008; Peter and Davidson, 2011). Many networks can be found in the literature constructed solely on the basis of statistical analysis of gene expression information, such as clustering of transcriptome data, or genome-wide physical interaction data such as ChIP-seq observations. But such network models are not considered in this volume, because the contribution of the proposed linkages to gene regulation and to the causality of the developmental process has not been experimentally assessed.

2. Boolean Spatial Output

Animal bodies display a fundamentally Boolean character at every level of organization. The body parts of which they are composed are discrete and their structure and function as well as location within the body are deterministically programmed and spatially bounded. Heads are not partly legs, eyes do not blend into mouths, and the pelvis does not grade into the spine. Similarly, at a microscopic level, differentiated cell types are discrete. Neurons, muscle cells, keratinocytes, and gland cells each express discrete sets of effector genes which determine their respective functions, and their spatial location within each body part is also discretely determined in development. This discrete organization of the body plan is the outcome of a sequential series of regulatory definitions of space in the developing organism. In molecular terms, such definitions are achieved by the spatially discrete expression of sets of regulatory genes. From the point at which the cells of an embryo begin to express their genes, the patterns of expression are discrete and Boolean. That is, cells in every domain of the embryo express some genes that are not expressed in other domains. The expression of genes in some cells but not in others requires that their regulators be present only in some but not in other cells. How this occurs right from the beginning of embryogenesis is taken up in the following chapter.

Thus the fundamental explanatory challenge is to understand how the genomic code that underlies the developmental process gives rise to spatially Boolean patterns of gene expression. Illustrations of such patterns are now available for every developmental system that has been studied at the molecular level, including both embryogenesis and subsequent processes of body part formation. As an illustration, in Fig. 2.1(A) we show a collection of in situ hybridization images in developing sea urchin embryos, which at this stage consist of several hundred cells differentially expressing their genomes in geometrically organized patterns. These patterns indicate the domains of the embryo in which the gene is expressed and also the domains in which it is not expressed. Most genes in Fig. 2.1(A) are regulatory genes, and the absence of detectable in situ hybridization signal in sea urchin embryos indicates a level of transcripts too low to produce a functional level of transcription factor (see below for quantitative basis). There is nothing particular to the sea urchin embryos shown in Fig. 2.1(A). Similar

Figure 2.1

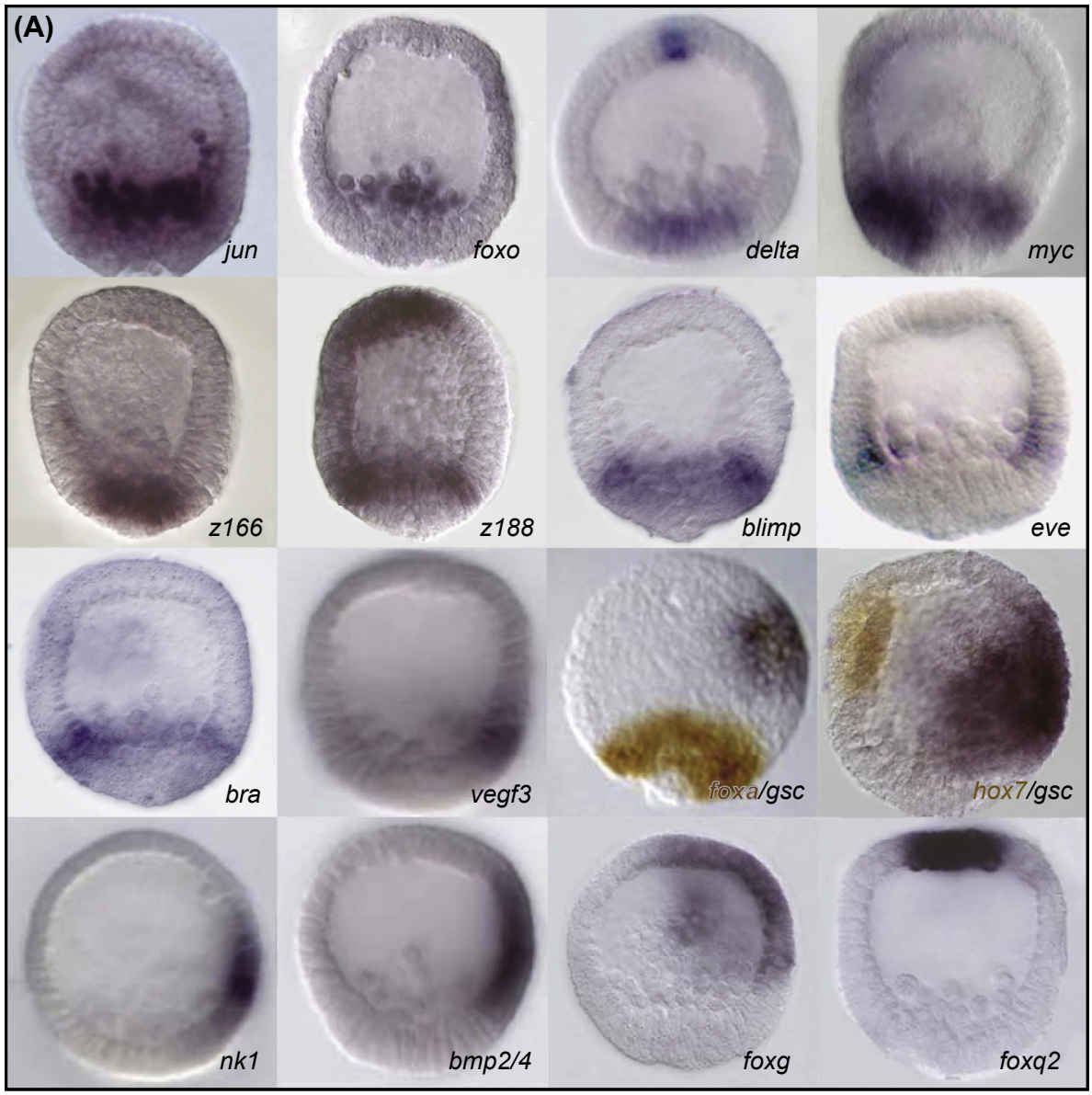
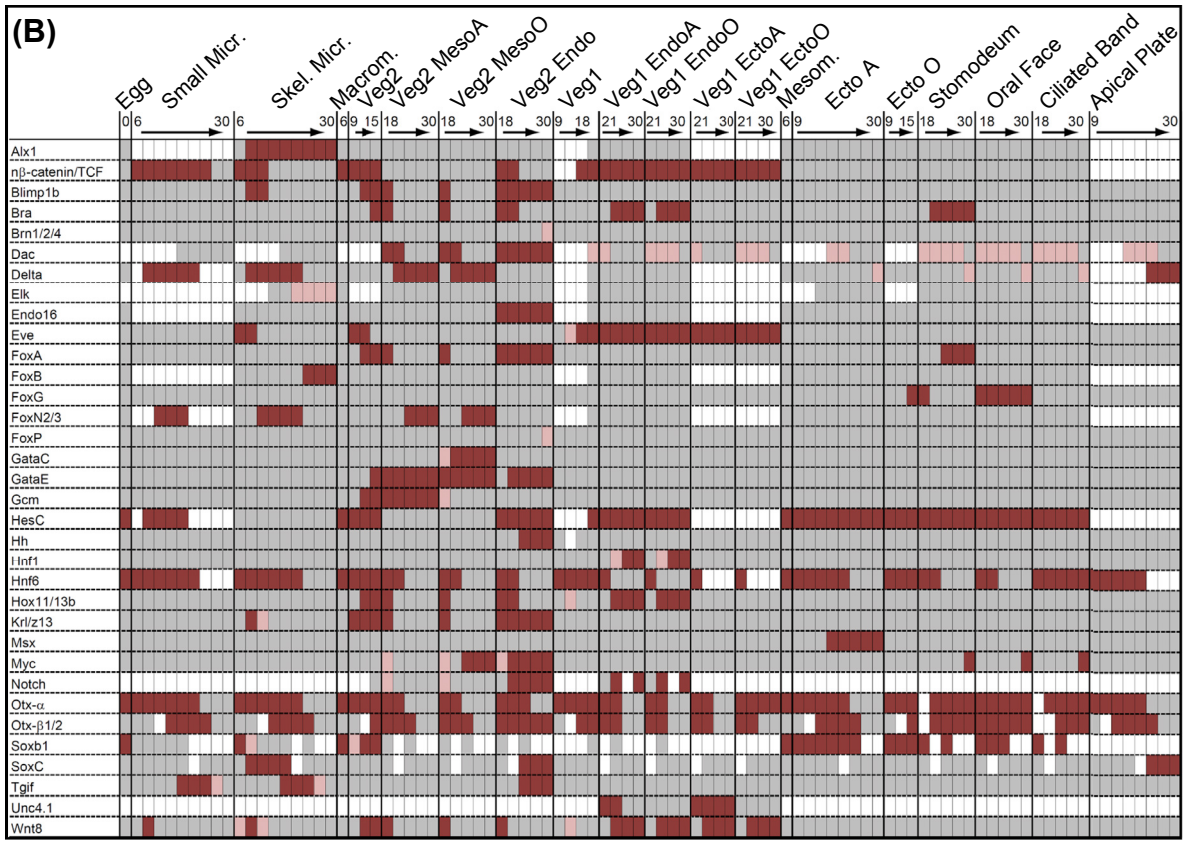


Figure 2.1 Boolean spatial patterns in development. (A), Typical examples of discrete spatial gene expression patterns in sea urchin (*Strongylocentrotus purpuratus*) embryos visualized by whole mount in situ hybridization. Genes shown all encode transcription factors or signaling ligands. All embryos shown are lateral views at 24 h of development. (B), Combined digitalized display of expression patterns of regulatory and signaling genes in the regulatory state territories of this embryo up to gastrulation at 30 h after fertilization. Dark red: active expression; gray: no detectable expression; white: no data. Genes are listed alphabetically on left. Each cell in the vertical columns of the chart represents a 3 h interval and the spatial domains of the embryo are shown across the top of the diagram; micr, micromeres; skel, skeletogenic; MesoA and MesoO, aboral and oral non-skeletogenic mesoderm respectively; EndoA and EndoO, aboral and oral endoderm; EctoA and EctoO, aboral and oral ectoderm; Mesom, mesomeres. *(From Peter and Davidson (2011)).*

Figure 2.1



demonstrations of discrete spatial expression patterns fill thousands of pages of literature on mouse, worm, frog, and fly development. Where expression of virtually all of the regulatory genes involved in a specific phase of a developmental process has been studied, the patterns of gene expression can be formalized in a global Boolean matrix, as shown in Fig. 2.1(B). This chart includes all regulatory genes specifically expressed in the endoderm and mesoderm of the developing sea urchin embryo up to gastrulation, when this embryo (*Strongylocentrotus purpuratus*) consists of approximately 600 cells. Temporal and spatial expression information is summarized across the whole period from cleavage to gastrulation for dozens of genes encoding transcription factors. Just as in any Boolean system, there is as much information content in the cells of the matrix showing absence of expression as there is in cells showing presence of expression. Thus, a mechanistic explanation of the expression pattern must provide causes for both states. The essential difference between Fig. 2.1(A) and Fig. 2.1(B) is that in the latter the developmental output of the system has been converted to a digital form which enables the computational processing of large-scale spatial expression data.

Underlying the Boolean spatial expression of genes in development are the discrete functions of the array of spatially active enhancers that service each gene. Just as genes are expressed in only a fraction of the cells of an embryo or body part, each of their enhancers is expressed in only a fraction of the cells expressing each gene. A classic example was shown in Fig. 1.3A2 which demonstrated the modularity of the genomic regulatory system that controls the expression of the *even skipped* gene in the *Drosophila* embryo. This gene is expressed in seven discrete stripes, but each of the enhancers controlling this function specifically causes expression in only one or two of the stripes. Thus these regulatory modules display

canonical Boolean behavior, as illustrated for the *eve* stripe 3+7 module in Fig. 1.3A2. This is the general rule for developmentally active enhancers. The elemental unit in animal gene regulation is the CRM and the Boolean spatial output of developmentally active CRMs is fundamentally the source of the Boolean character of animal body plan organization.

3. Regulatory States

Regulatory states are the collective sum of specifically coexpressed transcriptional regulators in a given place at a given time. Unlike individual transcription factors, which are used repeatedly in many developmental contexts, the regulatory state uniquely defines a developmental condition and phase. Regulatory states determine all developmentally specific functions, and later in development cell type specific functions. For this reason, considering GRN output in terms of regulatory states provides the key to the causal relation between genome and developmental function.

Returning to Fig. 2.1(B), each column shows the regulatory genes expressed at a given time in a given domain of the embryo, that is, each column as a whole specifies the particular regulatory state for that domain and that time. Moving to the right within each domain in this figure we can perceive the progressive change in regulatory state with time. Comparison of regulatory states among different domains allows the identification of spatially specific regulatory signatures. Here we consider the output of regulatory genes at the mRNA level as a proxy for the transcription factors that they will give rise to. As just noted, a significant or functional level of transcription factor protein can be generated from the levels of regulatory gene transcripts which are detectable by in situ hybridization (about 10 copies of message per cell). Therefore, assessment of regulatory gene transcript matrices as in Fig. 2.1(B) is directly informative of both regulatory state per se and of downstream transcriptional functions.

Among the genes which are expressed under the control of a given regulatory state are the regulatory genes which will compose the next regulatory state. Regulatory states are generated by the networks of interacting regulatory genes. Regulatory states are therefore both the output and the input to GRNs and they represent the individual active states the network can assume.

4. Regulation in *Cis*

Controlled transcription of animal genes in development and physiology depends on the specific dedicated DNA sequences for which we use the term "*cis*-regulatory modules" (Chapter 1). In general, the role of CRMs is to recruit specific sets of transcription factors which in turn control the expression of the genes to which the CRMs are dedicated. The specificity of CRM function devolves from their content of the particular DNA sequence elements that individual transcription factors recognize and physically bind to. Within a short stretch of DNA sequence, CRMs contain target sites for multiple transcription factors, and combinatoriality in the function of inputs in CRMs is a fundamental principle of gene regulation in animal cells. Often the transcription factors binding in a given CRM interact with one another cooperatively (see below) or with third-party cofactors which do not themselves bind DNA but are recruited to specific CRMs by the transcription factors bound to their specific target sites.

4.1 CRM genomics

An initial concept immediately illuminates the meaning of the word "module" in the term CRM. This concept highlights the difference between a transcription factor-binding site located within a CRM and

anywhere else in the genome; by elemental probability such sites should and do appear randomly many thousands of times per genome. For example, a six base pair sequence would occur in a random DNA sequence on average once every 4000 bp or almost a million times in a mammalian-sized genome. The difference between the target site(s) within the CRM and the much larger number of these sites occurring randomly is not necessarily in the recruitment of their cognate transcription factor, as shown by many ChIP-seq results, but rather in their regulatory function. The organization of CRMs ensures that multiple functions can be deployed when the CRM is fully loaded. These functions depend on (1) integration of multiple diverse transcription factor inputs; (2) recruitment of cofactors which in turn determine the transcriptional activity of the gene, including particular coactivators (such as p300), or corepressors (such as Groucho) or nucleosome modification enzymes (see Chapter 1); (3) genomic looping which brings the CRM and its associated protein complex into immediate contact with the basal transcription apparatus at the start site of the gene; (4) topological positioning of the CRM and the gene in specific intranuclear compartments where transcription is favored or disfavored. In the absence of this multiplicity of events, regulatory function cannot occur. This is why random orphan-binding sites do not contribute to specific regulation of gene expression. While individual orphan sites occur frequently, clusters of particular binding sites within a few hundred base pairs of sequence are uniquely improbable. The functional regulatory apparatus of the genome is thereby uniquely encoded in its CRMs.

At the next level of genomic regulatory organization, a dominant feature of animal gene control systems is the multiplicity of CRMs per gene. As discussed briefly in Chapter 1, there are on average about 5–10 CRMs per gene. Some genes have many more CRMs than this, like for example the *Ubx* and *string* genes in *Drosophila*, and the *mrf4*, *myf5, sox2*, and *otx2* genes in mouse (Davidson, 2006). In general, each of the multiple CRMs that service a given gene contains a unique set of binding sites for transcription factors that in concert are present at particular times and places in development. Therefore most CRMs act only in particular developmental contexts. This is the genomic basis for the polyfunctionality of many genes. Thus, typically, if a gene is expressed in many domains of a developing organism, its control system will be mosaic in that the same gene will be controlled by a separate CRM for each phase of its expression. A great example discussed in Chapter 5 is the aforementioned *mrf4* and *myf5 cis*-regulatory system, which operates in muscles all over the body, while expression in each particular part is driven by different individual CRMs (Hadchouel et al., 2003; see Chapter 5). The essential basic principle is that developmental gene regulation is a modular process and the structural basis underlying this fundamental feature is the modularity of the *cis*-regulatory apparatus.

4.2 CRM logic functions

CRMs have the potential to affect the basal transcription apparatus in multiple ways. Functions mediated by CRMs ultimately control activation, repression, and also rate of expression. The many different types of biochemical interactions between CRM-bound transcription factors and cofactors on the one hand and the transcription complex on the other, lie beyond the scope of this chapter (for reviews see Ma, 2011; Marsman and Horsfield, 2012; Spitz and Furlong, 2012). Most of the functions of the basal transcription apparatus, which consists of an enormous complex of over 50 polypeptides, are not specific to particular genes but instead are generally deployed for gene transcription. The answers to the question why any particular gene is expressed in a specific developmental context therefore usually do not lie within the structure of the basal transcription apparatus. Rather, developmental specificity resides in the encoded combination of transcription factor target sites that constitute the working sequences of CRMs.

Active CRMs bind many different sequence-specific transcription factors, as well studied cases demonstrate (Thanos and Maniatis, 1995; Yuh et al., 1998, 2001; Swanson et al., 2010). These factors contribute in various ways to the function of the CRM and interference with the binding of any of them affects the regulatory output. Factors which perform positive spatial and temporal control of gene

expression are commonly referred to as "drivers". In the case of *endo16*, a gene which encodes a secreted protein of the midgut in the sea urchin embryo, only three out of thirteen sequence-specific transcription factors bound to the *endo16* CRMs, function as drivers. The other proteins have a variety of functions including quantitative amplification of the regulatory output, spatial repression, internal communication among different modules constituting the overall *cis*-regulatory system and regulation of the use of driver inputs (Davidson, 2006). From an informational point of view, the determinants of developmental gene expression are the drivers which animate each CRM plus any repressors that restrict CRM activity from given domains.

It is useful to consider CRMs in terms of their input and output functions. The inputs are defined by the ambient transcription factors which convey spatial regulatory information to the CRM (or in other words by the transcription factors which are expressed in a non-ubiquitous manner). The regulatory output of a CRM, which it conveys to the basal transcriptional apparatus when it is in play, is the result of informational processing of these inputs. CRMs mandate logic functions by which these driver inputs will be combinatorially processed (Istrail and Davidson, 2005). For example, a gene responds to input A, which is expressed in a given set of cells, and also to input B, expressed in a second set of cells. If the CRM utilizes these inputs by AND logic, the consequence is that it will produce a positive output only in the unique spatial subset of cells where A and B are both present, or else it produces no output at all. Thus the structure/function relations encoded in the information processing system in this CRM require that it runs only when both A and B are present simultaneously. Or else, the same inputs might be utilized by OR logic, which would result in a positive CRM output wherever either A or B are expressed. Other CRMs respond to specific repressors, which if present execute NOT logic functions. Often it is the case that whenever such repressive inputs are received by the CRM, the CRM is prevented from producing a positive output, irrespective of the presence of activating inputs. In life, given CRMs may process multiple inputs with the use of multiple logic functions. The consequence is that the output of genes driven by such information processing systems is different from any one of their individual inputs in time and/or space, and in development this property of CRMs lies at the heart of their ability to generate novel patterns of gene expression. CRM logic functions can be conveniently modeled using Boolean logic operators and this provides a way of predictively treating very large developmental GRNs, as we discuss in Chapter 6.

4.3 Flexibility and constraint in CRM sequence

CRMs may consist of densely packed functionally important sequences. Recent studies in which the significance of virtually every nucleotide has been assessed for several known enhancers demonstrate that the majority of sites may contribute to CRM function (Kwasnieski et al., 2012; Melnikov et al., 2012), although this is not always the case (Patwardhan et al., 2012). A detailed analysis of the *Drosophila sparkling* enhancer demonstrated that sequences between the known driver target sites were also essential for CRM spatial activity (Swanson et al., 2010). Mutation of these sequences showed that most of the *sparkling* enhancer contains critical regulatory information (Fig. 2.2(A)). The significance of such observations is to remind us that CRMs are far more complex machines than simply target sites for one or two driver transcription factors. When we consider the complexity of interactions in enhancers such as *sparkling*, it is clear that we have only begun to assess the actual information processing functions of CRMs. Integrating over the many thousands of developmentally active CRMs that exist in animal genomes (see Chapter 1) directly implies the informational complexity of the developmental process.

Despite the dense informational content of CRM sequence, comparison of orthologous enhancers with conserved function among more and less closely related species reveals an amazing flexibility at the sequence level. When *eve* stripe 2 regulatory sequences were isolated from the genomes of even distantly related flies, associated with reporter gene expression constructs and introduced into *Drosophila melanogaster*, these constructs generated accurate stripe 2 expression (Fig. 2.2(B1,2);

Figure 2.2

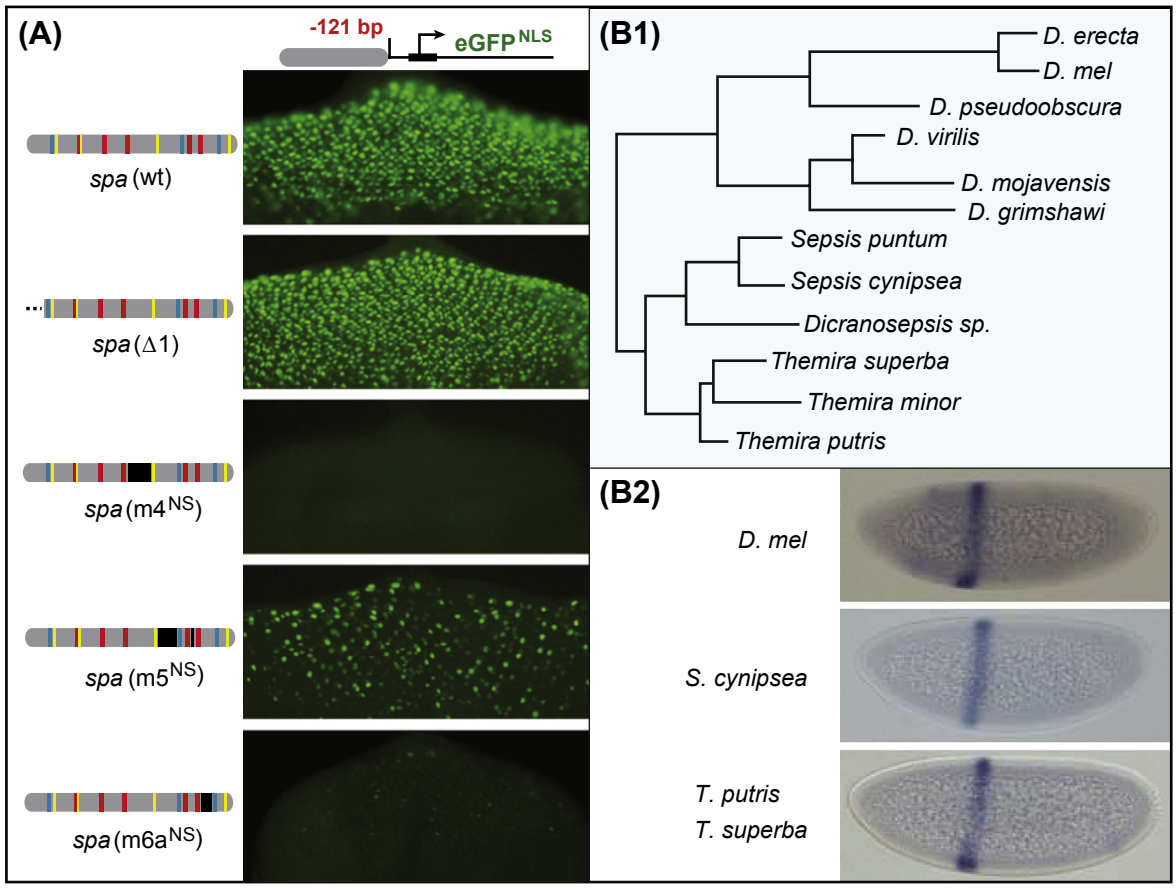
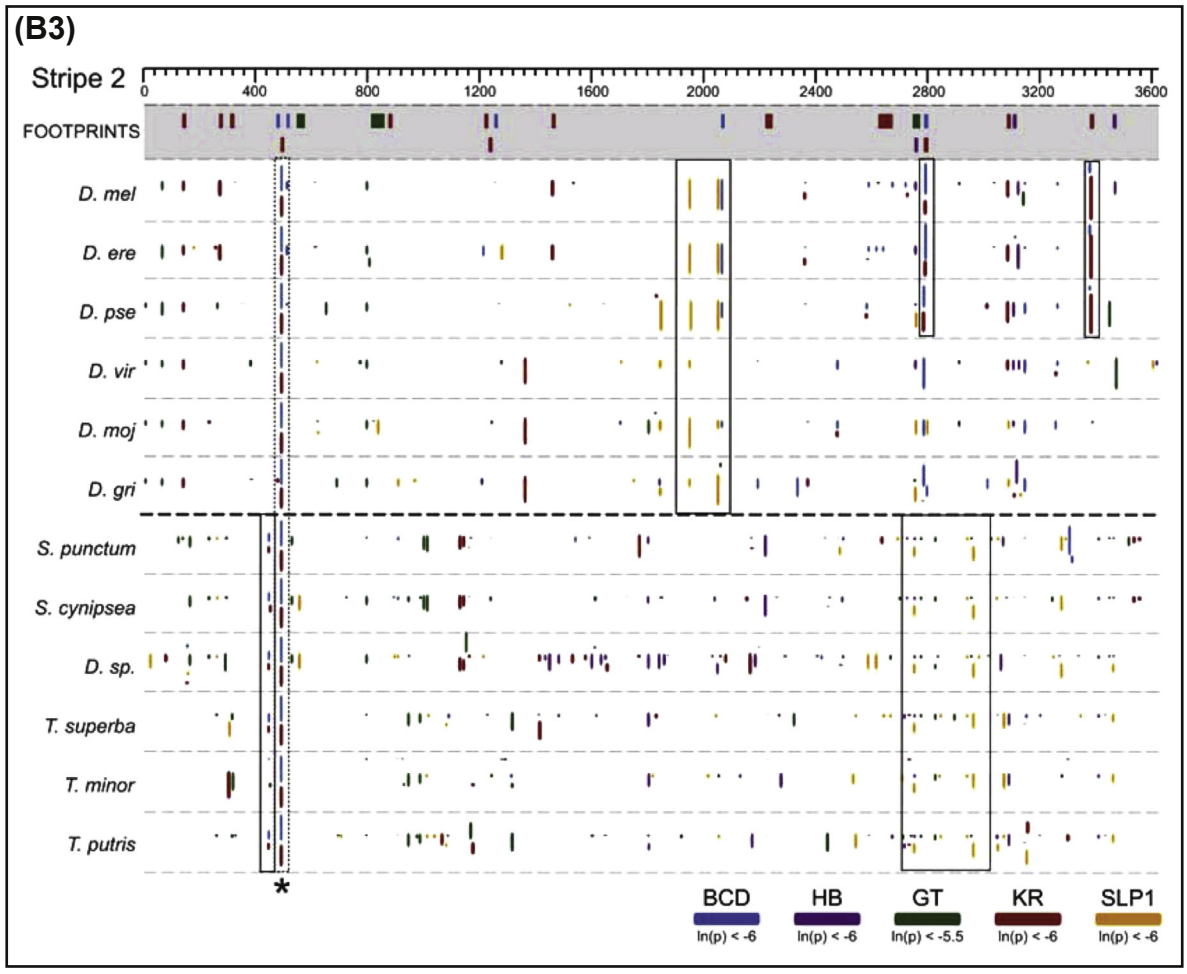


Figure 2.2 Sequence requirements for conservation of *cis*-regulatory function. (A), *Drosophila sparkling* enhancer of the *pax2* gene, active in cone cells of the eye *(from Swanson et al. (2010).)* Target sites are color coded in diagrams of the 362 bp enhancer to left. This enhancer is known to respond to Notch (Su(H), red), EGFR via Ets (yellow), and the runt family factor Lozenge (blue). Here in a construct diagrammed at top the enhancer is placed only 121 bp from the GFP reporter transcriptional start site, and expression of constructs in which regions other than the Su(H), Ets and Lz sites were systematically deleted is shown; wt, wild type. The 5′ sequence deleted in Δ1 is required only for interaction with the promoter from its normal distal position but not when located only 121bp away as in this construct; deletions $m4^{NS}$ and $m6a^{NS}$ abolish expression, and $m5^{NS}$ severely weakens expression indicating the necessity of additional inputs. (B), Constraint and flexibility in sequence requirements for *eve* stripe 2 enhancer function, demonstrated by evolutionary comparisons in dipteran flies *(from Hare et al. (2008b).)* (B1), Phylogenetic tree displaying relationships of flies included in following analyses; note comparisons are within Drosophilidae and from *Drosophila melanogaster* (*D. mel)* to the distant clades Sepsidae and Themidiridae. (B2), Demonstration by transgenesis in *D. mel* that *eve* stripe two enhancers isolated from genomes of sepsid and themirid flies respond to the *D. mel* regulatory state exactly as does the *D. mel* enhancer, producing a normal stripe 2 pattern. (B3), Sequence organization of stripe 2 enhancers from the species included in (B1). Site identities are color coded as at bottom of figure; the only site arrangements conserved either inside or outside of the Drosophilidae are closely adjacent boxed site pairs. The figure demonstrates that it is the presence of sites for all requisite

Figure 2.2



factors that is required for similar function. Neither similar site arrangement nor inter-site sequence need be conserved. (C), Interclade transgenesis of an extremely conserved *cis*-regulatory module of a *soxb* gene *(from Royo et al. (2011).)* (C1), top tier, almost identical expression of enhancer constructs from orthologous human, amphioxus and sea urchin *soxb* genes in developing zebrafish brain; e, eye; h, hindbrain; m, midbrain; s, spinal cord; t, telencephalon. Bottom tier, from left to right, endogenous expression of sea urchin (*S. purpuratus*) *soxb2* gene in neurogenic ciliated band, oral perimeter, and midgut wall of feeding larva, detected by WMISH; GFP expression of sea urchin (Sp), human (HS), and *Nematostella* (Nv) enhancers (CNR) in individual neurons of these regions in sea urchin larvae of same stage. (D), Organizational requirements for spatial function in the *sparkling* enhancer (color coding as in (A); *from Swanson et al. (2010))*. (D1), Interactions within module. Green arrows indicate short-range, i.e., position dependent, intra-modular interactions among diverse factors bound at indicated sites. These interactions are necessary for expression in cones ("patterning circuitry"). (D2), Alteration of relative positions of target sites changes spatial specificity of enhancer, even though same set of sites is present (i.e., same "combinatorial code"); alternative cell types are cones and photoreceptors (R1 and R6).

Figure 2.2

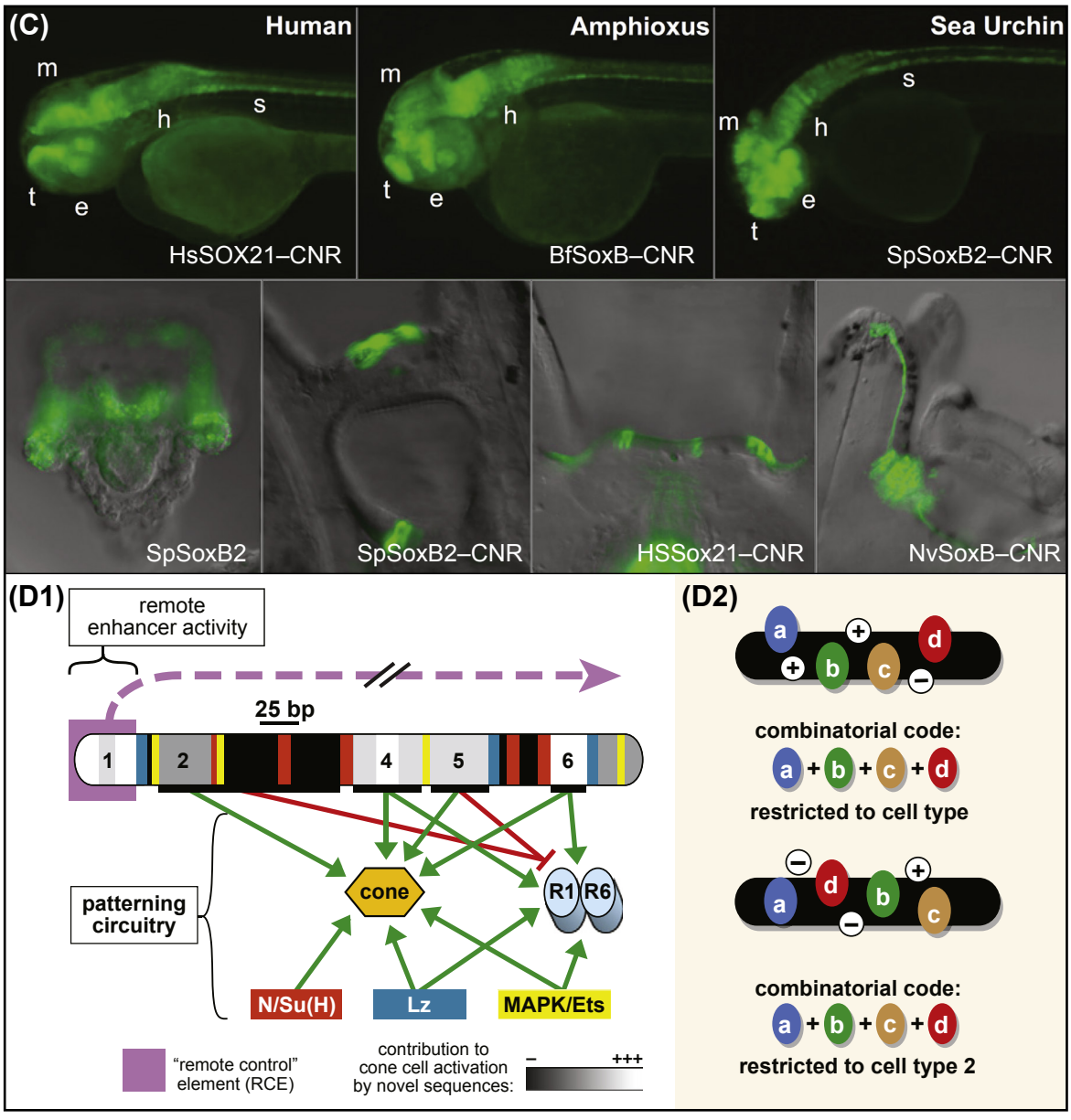


Figure 2.2 (continued...)

Hare et al., 2008a,b). Thus the function of these enhancers is perfectly conserved. Yet as shown in Fig. 2.2(B3), the arrangement of the well characterized target sites in the *eve* stripe 2 enhancers is clearly different. Most of those features that are conserved appear to consist of target sites for two different factors which directly abut one another or even overlap. In general, the binding sites for all the known crucial drivers of the *eve* stripe 2 enhancer are present in the CRMs of every species. However, their relative positions and their location within the enhancer have changed since divergence, and the sequences of these CRMs have turned over extensively, including both the

driver sites and the sequences separating them. This suggests that with the exception of the closely contiguous conserved site pairs, what is important for function is essentially just the presence of the driver sites, and not their relative positions. A similar import derives from an experimental study on a *Drosophila sog* CRM (Liberman and Stathopoulos, 2009). Comparison of the regulatory architecture of this CRM among *Drosophila* species again displays significant variation of the organization of target sites for the transcription factors that spatially regulate *sog*. Nonetheless when functionally tested in *D. melanogaster* embryos, these various CRM architectures are all capable of generating similar lateral stripes of gene expression. Furthermore, experiments using synthetic *cis*-regulatory elements demonstrated that various arrangements of the necessary target sites suffice for driving gene expression specifically in the lateral stripes.

Observations on the naturally occurring flexibility of CRMs of similar spatial function led to the proposition of the "billboard model" of CRM structure/function relations (Kulkarni and Arnosti, 2003; Arnosti and Kulkarni, 2005). According to this idea, the essential information within CRMs is comprised by the identity of the target sites within it, but the relative organization of these target sites is not usually significant. However, as pointed also out by these authors, some types of CRMs are known that are organized in an entirely different manner (Kulkarni and Arnosti, 2003; Arnosti and Kulkarni, 2005). The canonical example is the *interferon-β* "enhanceosome", which consists of tightly apposed sites for six different transcription factors and which operates only when all six factors are in place. Here the architecture of the binding sites is required to enable the essential cooperative interactions among bound transcription factors, and almost every nucleotide in this 44-bp-long enhancer is required for function (Melnikov et al., 2012). In vertebrates, that small fraction of developmentally active CRMs in which the nucleotide sequence is conserved across the whole length of the CRM, from fish to mammals, may belong to this CRM structural class (Siepel et al., 2005; Woolfe et al., 2005; Katzman et al., 2007; Elgar and Vavouri, 2008). An extreme example illustrated in Fig. 2.2(C), is an enhancer for a *soxb* class regulatory gene, which at the sequence level is conserved from *Nematostella* (starlet sea anemone) to human genomes (Royo et al., 2011). Remarkably, when tested by transgenesis in sea urchins, zebrafish, and mice, this CRM functions almost identically, irrespective of the genome of origin, promoting expression in developing or mature neurons. Thus, despite the vast evolutionary distance, these CRMs are capable of responding to similar regulatory states expressed in neurogenic tissues.

An alternative example of structure/function relations in a developmental CRM is provided by the *sparkling* enhancer referred to above. This enhancer is very poorly conserved in sequence, even within the Drosophilidae; yet it contains tightly packed regulatory sequence for many factors. As shown in the diagram of Fig. 2.2(D1), multiple factor interactions are required for the cell type specific expression that it mediates. When the arrangement of the target sites is altered experimentally, the spatial function of the CRM changes, so that it promotes expression in a different cell type (Fig. 2.2(D2)). The implication is that functional tightly organized *cis*-regulatory architecture can turn over rapidly in evolution, and sequence conservation is not a required indication of functionality in the regulatory genome.

5. Module Choice

As we have seen in Chapter 1, animal genes are controlled by multiple CRMs located in different regions within and surrounding the gene, and frequently these genes utilize alternative promoters as well. With the advent of elegant methods for cross-linking chromatin in its native configuration in living cells ("chromatin capture"; 3C, hiC), it has now become clear that the mechanism of transcriptional control by distantly located CRMs involves the formation of specific loops which bring the CRM and the promoter into direct contact. Loops are not an intrinsic structural property of the genome sequence but rather their formation is determined by physiological or developmental context. Thus given loops have been shown to be cell type specific and their formation to be tightly correlated with transcriptional activity. In some circumstances there are multiple enhancers looping to one promoter or individual enhancers may service an array of multiple promoters, with

the formation of alternative loops (Guo et al., 2012). A number of important issues devolve, among which are: what is the mechanism by which these loops form; what controls the specificity of loop formation during development; and what determines the choice of enhancers and promoters participating in any given loop.

5.1 General looping mechanisms

The physical stability of transcriptional looping configurations is dependent on special protein complexes. A major component of these is the polyfunctional multiprotein complex known as cohesin. Cohesin forms rings which assist in maintaining the physical contiguity of enhancer and promoter components (see Fig. 2.3(A) for several such configurations; Ong and Corces, 2011). Cohesin often interacts with the mediator protein complex, an enormous polypeptide entity associated with a large fraction of promoters in yeast and animals. Furthermore, cohesin interacts with CTCF, which was initially discovered as a major component of insulator complexes, where it specifically binds the CCCTC sequence motif. CTCF has now been associated with many chromatin looping functions because of its capacity to interact homotypically once bound to DNA (Bell et al., 1999; Yang and Corces, 2012). Some loops are facilitated by the interaction between insulator CTCF and CTCF bound to sites within or near enhancers, which assist in generating productive transcriptional loops (Krivega and Dean, 2012). Numerous observations have shown that

Figure 2.3

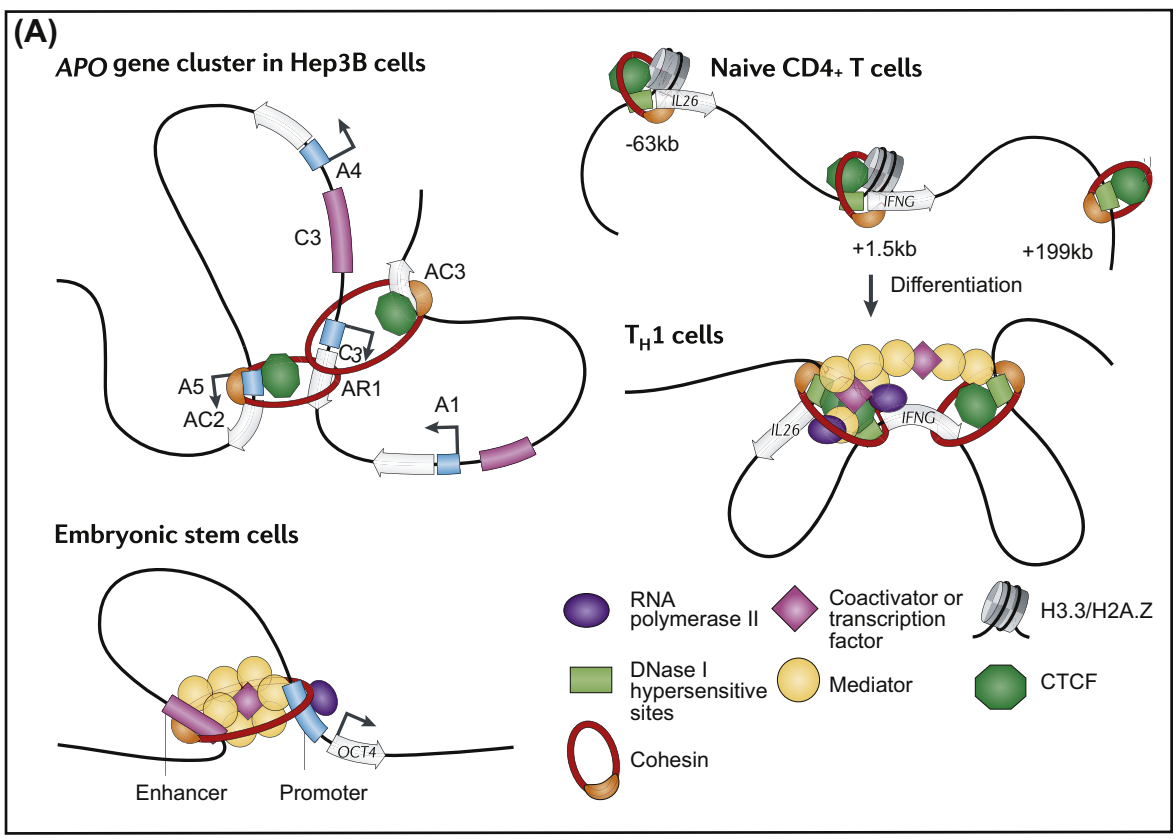
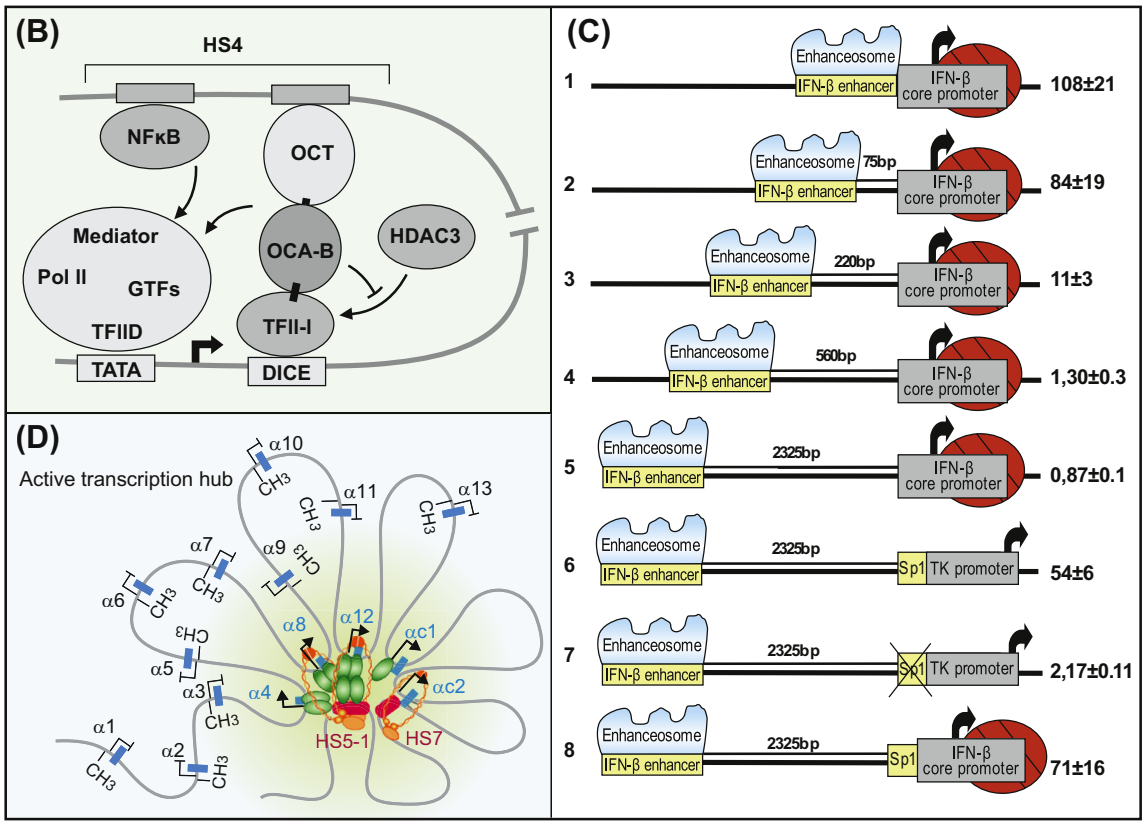


Figure 2.3 Looping interactions between distant enhancers and promoters. (A), Specific examples of cohesin loop stabilization structures *(from Ong and Corces (2011))*. Components are shown in key at bottom of figure; of these, transcription factors and CTCF factor interact with DNA sequence

Figure 2.3



specifically. Promoter sequences are shown as blue cylinders. In the APO gene, abbreviations are: C3 denotes an enhancer, which is present in the same loop as the C3, A4, and A5 promoters; A1 promoter is present in a different loop. AC3, AC2, and AR1 are insulator sites. (B), The transcription factors NFκB and OCT interact with the enhancer HS4. NFκB in addition interacts with Mediator complex while OCT interacts with the cofactor OCA-B, through which it is tethered to the initiation site by TFII-I. The result is the direct physical interaction between proteins bound to the immunoglobulin heavy chain enhancer and proteins bound to the promoter of this gene, while the intervening DNA sequence (break) forms a loop *(from Ren et al. (2011))*. (C), Synthetic experiments demonstrating that transcriptional induction of the *interferon-β* gene enhancer requires the transcription factor Sp1 to be bound near the promoter when the enhancer is distantly located though not if proximity makes looping unnecessary *(from Nolis et al. (2009))*. The experiment shows decreasing levels of inductive gene expression with increasing distance between promoter and enhancer sequences in the absence of SP1 binding sites (Exp. 1–5). But even at the normal distance of 2325bp, presence of an SP1 site restores activity of either heterologous (Exp. 6) or endogenous promoters (Exp. 8) to levels comparable to the construct where enhancer and promoter are juxtaposed (Exp. 1). Numbers at right indicate fold induction of gene expression upon addition of stimulatory reagent to which the enhanceosome responds. (D), Enhancer-promoter looping in the protocadherin gene cluster *(from Guo et al. (2012))*. Promoters are indicated as short blue bars. Promoters in contact with enhancers (HS5-1 and HS7, red) cause transcription of protocadherin gene segments. Those active here are α4, α12, α8, αc1, and αc2, which have looped to the enhancers forming a transcription hub held together by CTCF (green)/cohesin (red circles) complexes.

CTCF, cohesin, and mediator are present in the transcriptional complexes detected by chromatin capture, and that they are required for loop stability and for continued transcription. Thus removal of CTCF or knockdown of cohesin disrupts the chromatin loop and reduces gene transcription (Mishiro et al., 2009; Guo et al., 2012). But these generally utilized structural facilitators of looping are present in all cells and thus cannot explain the differences in chromatin loops observed when different cell types are compared.

5.2 Specificity of looping

The multiple CRMs controlling expression of a given gene generally respond to distinct sets of transcription factors, endowing the gene with the capacity to be expressed in various developmental contexts. This is particularly true of the classes of genes controlling development, those encoding signaling molecules and transcription factors. For the gene to be regulated in each particular context by the appropriate enhancer requires the formation of a loop that brings this enhancer into immediate contact with a promoter. What mechanisms ensure specificity of interaction between promoter and the particular CRM, thus regulating enhancer choice? A number of recent studies demonstrate that transcriptional loops occur only when the enhancers are loaded with their specific transcription factors. Thus the sequence specificity of loop formation with respect to which enhancer is to be engaged depends directly on the binding of these transcription factors just as do the other functions of the enhancer. A particularly clear case concerns the regulation of the interaction between an immunoglobulin heavy chain enhancer and promoter (Ren et al., 2011). In B cells, Oct2 is a sequence-specific activator of this enhancer and the same factor participates in looping of the enhancer to the promoter, as shown in Figure 2.3B (Ren et al., 2011). The loop is formed by an interaction between the Oct2 cofactor Oca-b and the basal transcription factor TFII-1, which binds in the promoter region. Additionally there is an interaction between another transcription factor bound in the enhancer, NFκB, and the mediator/PolII/TFIID complex at the transcription start site. Thus the sequence-specific binding of Oct2 and NFκB to their *cis*-regulatory target sites in the CRM directs the enhancer/promoter interaction. Another illuminating case concerns the *interferon-β* (*ifn-β*) enhancer discussed above (Nolis et al., 2009). Here the loop is formed by the interaction between transcription factors of the enhanceosome and the transcription factor Sp1, which binds directly upstream of the promoter. The necessity of Sp1 for interaction between the enhanceosome and the promoter is demonstrated in the experiments shown in Fig. 2.3(C) (Nolis et al., 2009). In the absence of the Sp1 site, the enhanceosome is capable of activating gene expression only when located immediately adjacent to the promoter. When located distantly, a productive looping interaction occurs only in the presence of the promoter-proximal Sp1. Specific transcription factors and their cofactors have been shown to be required for enhancer/promoter looping in many other circumstances. Many clear examples of such transcription factor-dependent mechanisms of enhancer/promoter interactions have appeared in the literature. Among these are: class-switch recombination, in which enhancer-bound Pax5 and its cofactor Ptip are required for long-range interactions (Schwab et al., 2011); inflammatory gene expression in response to interleukin-1 (IL-1), in which transcriptional loop formation is cooperatively regulated by sequence-specific transcription factors Lef1, RelA, and c-Jun (Yun et al., 2009); the *β-globin* gene cluster in which a distant locus control element loops to the promoter of the gene in a process requiring binding of the transcription factors Klf1 and Gata-1, as well as the cofactor Fog (Drissen et al., 2004; Vakoc et al., 2005); *opsin* loci, where the transcription factors Crx, Nrl, and Nr2e3 bind specifically in rods and cone cells and are required for the formation of transcriptional loops (Peng and Chen, 2011; see also the examples in Fig. 2.3(A)).

Looking at the process from the perspective of the actively loaded CRM, the looping mechanism must be able to define the promoter sequence to which the CRM should loop. This includes recognition of a promoter sequence from which transcription may be initiated, or choice among multiple promoter sequences. Recognition of promoter sequences as opposed to other sequences is mediated by the general factors of the transcription complex, which bind at promoter sequences and may include mediator and cohesin tethers. However, there are many examples in which given promoters must be distinguished from other promoters to

ensure specific enhancer/promoter interaction. These cases include genes with multiple promoters, each serviced by different CRMs, and genes in which the relevant CRM is located at a great distance from the promoter to which it must loop, often over an array of other intervening promoters. Specific promoter proximal tethering sequences have been identified in *Drosophila* that mediate particular interactions to distantly located CRMs. In such cases, specific *cis*-regulatory sequences located directly at the promoter serve as target sites for tethering factors, the binding of which is required for looping with distant CRMs (examples are reviewed in Davidson, 2006; Ho et al., 2011). An interesting contrast to the use of individual promoter proximal transcription factors (as in the Sp1 case above) is the absence of any promoter-specific proximal tethering sequences where choice amongst multiple promoters is required to be stochastic. This mechanism is utilized in *α-protocadherin* gene clusters, where two enhancers loop to and activate a large number of alternative promoters (Fig. 2.3(D); Guo et al., 2012).

In summary, looping is generally facilitated by a particular set of looping proteins that bind in most promoters such as mediator/cohesion, and by CTCF which binds specific DNA sequences located in appropriately positioned flanking regions of the gene. But the specificity of transcriptional looping depends on sequence-specific transcription factors bound to the CRMs and often, but not always, specific promoter proximal transcription factors as well. Thus, the potential for the formation of particular loops is directly encoded in the genomic sequence, and its context-specific realization occurs according to the developmental regulatory state.

6. Transcriptional Dynamics

GRNs consist mainly of interactions between transcription factors, the products of regulatory genes, and target sites in the CRMs controlling other regulatory genes. In development, these networks can be large and complex, and many individual interactions must be taken into account. It soon becomes impossible to think clearly about these interaction systems without considering their dynamics. This subject includes many often encountered conundrums: What is the time interval between successive gene activations in regulatory gene cascades? How do the real time kinetics of transcriptional systems in different animals living at different temperatures compare? How long does it take for the cell to make given amounts of mRNAs and what do you have to know to compute that? If we measure the approximate number of molecules of a regulatory gene mRNA in a cell, how do we determine if that number will suffice to generate enough transcription factor protein to affect downstream genes? What is the relation between transcription factor concentration and effective occupancy of a *cis*-regulatory target site, and how can we think about this occupancy with respect to the rate of transcription of the target gene? These are matters arising continuously in network regulatory molecular biology. Different mathematical approaches have been used to treat such problems, some more and some less closely tied to the actual nature of the physical processes of transcription control. Here we briefly summarize an elemental set of approaches to transcriptional dynamics, using as a guide the principle that we would like each step in the analysis to be related transparently to the basic processes of gene expression. We have tried to steer between Scylla and Charybdis; between on the one hand overly abstract formulations that behave mathematically as desired but do not directly and literally represent the real processes occurring in the cell, and on the other overly detailed mechanistic biochemical models which quickly become useless because they include too many microscopic constants which are hard to establish for most animal systems.

6.1 Transcriptional initiation and the life and death of mRNA

Productive gene transcription is biochemically an immensely complex process. Transcription is mediated by an assembly of dozens of proteins, only a few of which are well understood. The overall reason for the biochemical complexity is the number of diverse jobs that must be done for transcription to occur: the

polymerase must be recruited to the transcriptional start site; it must be released from this site; if it is paused in situ it must be released from its tether; the polymerase complex must ratchet along the DNA; the DNA must be rendered accessible and transiently melted so that one strand can be copied; etc. Nonetheless, two precepts enable a practical treatment of the rate-limiting features of the process that are relevant to GRN dynamics. First, in processes such as embryonic development, the microscopic stochasticity of transcriptional processes and their bumpy instant by instant local rates in individual cells affect only insignificantly the accumulation dynamics of given transcripts over relatively long periods of real time. Thus we can deal in average rates and processes. Below we illustrate the fact that for a typical gene expressed in a field of similarly functioning cells, such cell by cell stochasticity is of little real consequence anyway. Secondly, one major dynamic variable among the multiple microscopic processes involved in transcription largely determines the rate of output of transcriptional systems, and that is the rate of initiation. This is defined as the frequency of the events by which the productive traverse of the gene by the RNA polymerase complex begins at the transcriptional start site of the gene. The rate of initiation per gene over time, which we shall term ***I*** (molecules of transcript initiated/minute) is a variable dependent probabilistically on the concentration of the relevant transcription factors, as we discuss in the next section.

I varies between 0 and the maximum initiation rate, $\boldsymbol{I_{max}}$, which is a temperature-adjusted constant for each animal system: if we imagine a maximally loaded gene being transcribed at the maximum rate, the polymerases will be tightly packed, abutting one another as they traverse the gene (Fig. 2.4(A)), and the maximum rate of initiation is simply the inverse of the amount of time it takes for the first polymerase to move far enough down the gene so another polymerase can load on and initiate a new transcription event. The time required to clear the transcriptional start site depends on how fast the polymerase transits along the DNA, average values for which have been measured in many animal systems (Davidson, 1986; Bolouri and Davidson, 2003). The polymerase transit rate is among the basic numerical "facts of life", and it may seem surprisingly slow. For example, in sea urchin embryos at 15 °C the polymerase transit rate is only 6–9 base pairs per second, and since a polymerase occupies about 100 base pairs, it takes 11 s or more for the initial polymerase to move out of the way so that another may move in. Thus $\boldsymbol{I_{max}}$ is no more than about 5.5 molecules of RNA initiated per minute and gene, on average. For animals which operate at different temperatures, comparison of polymerase transit rates showed that this parameter roughly obeys the same "Q10" rule as do most chemical reactions, i.e., it, and $\boldsymbol{I_{max}}$, increase about twofold for each 10 °C rise in temperature. Thus for example in mice, the transit rate is ~30 base pairs per second.

Since the polymerase transit rate is more or less constant for all genes in a given animal at a given temperature, the initiation rate averaged over time directly predicts the average length of DNA separating adjacent polymerases and nascent transcripts on the gene. As a vast amount of direct and indirect measurement demonstrates, in a typical developing embryo such as a sea urchin or frog embryo, about 90% of the genes are being transcribed at rates far below $\boldsymbol{I_{max}}$, and the expected spaced appearance of transcription arrays on such genes can be seen in Figs 2.4(B)1,2,3. An important conceptual point is that the rate of completion of transcription and release of finished transcripts is about equal to the rate of initiation (except in some special circumstances where there is attrition and fewer polymerases complete than initiate transcription). Thus, given continuing initiation, the length of the gene does not affect the ongoing rate of transcript release although at the low polymerase transit rates several hours may elapse between initiation and release of the first transcript. This interval is merely a time lag (dependent on gene length) and the completion and release rate in molecules per minute and gene is generally equal to ***I***.

The primary RNA transcript is processed in a biochemically complex chain of events that includes intron removal, splicing, capping, polyadenylation, and interaction with proteins that chaperone and transit the mature mRNA out of the nucleus. But here again, in terms of dynamics, we are looking only at a lag, since in most (though not all) circumstances processing is 100% efficient. Thus in molecules per minute and gene, the rate of pre-mRNA synthesis is also the rate of mRNA entrance into the cytoplasm, again equal to ***I***. The average processing time lag is 20–30 min in 15 °C sea urchin embryos, shorter at higher temperatures, following the Q10 rule.

Figure 2.4

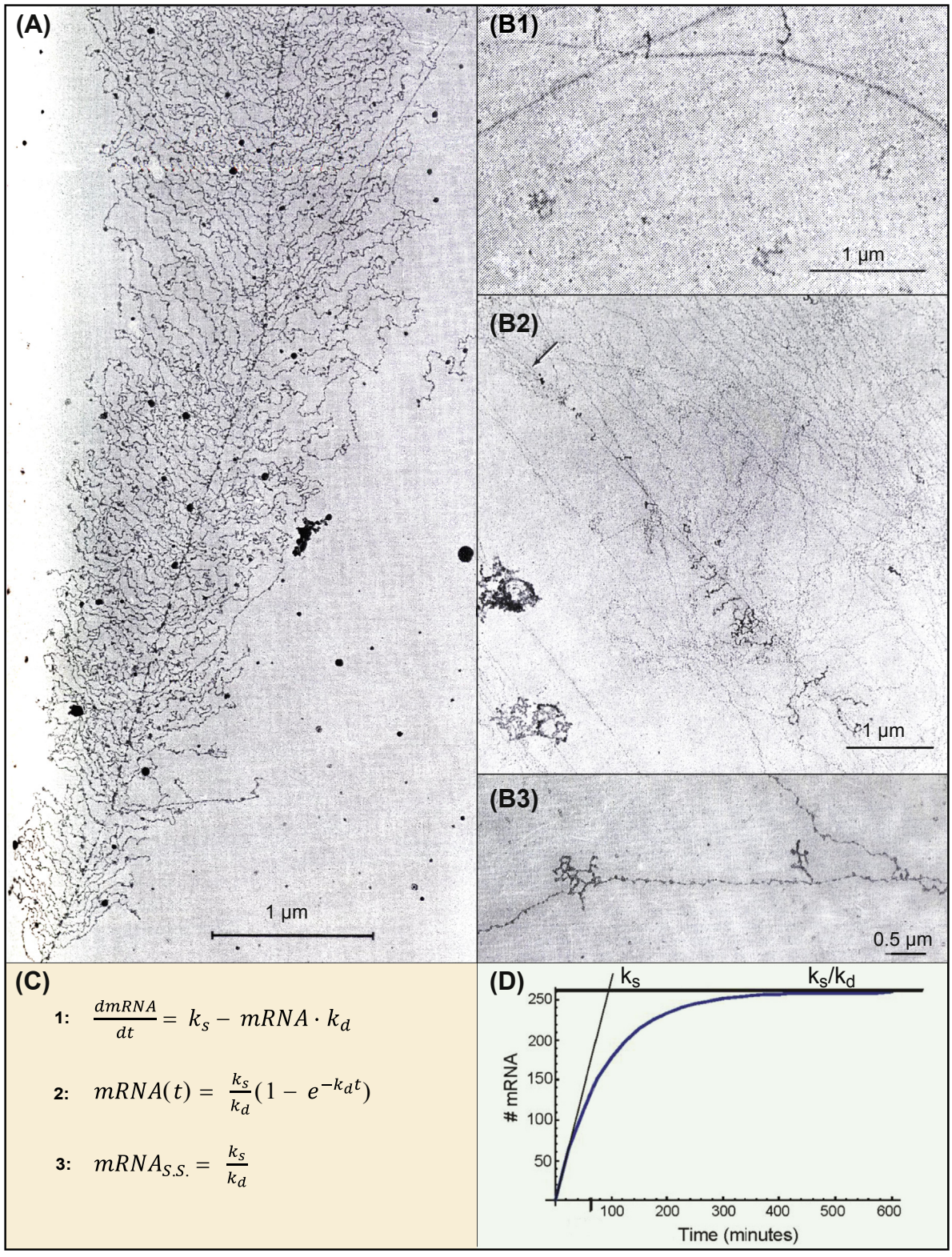

(A)
1 μm
(B1)
1 μm
(B2)
1 μm
(B3)
0.5 μm
(C)
1: $\frac{dmRNA}{dt} = k_s - mRNA \cdot k_d$
2: $mRNA(t) = \frac{k_s}{k_d}(1 - e^{-k_d t})$
3: $mRNA_{S.S.} = \frac{k_s}{k_d}$
(D)
ks
ks/kd
mRNA
250
200
150
100
50
100
200
300
400
500
600
Time (minutes)

Figure 2.4 Transcription dynamics. (A) and (B), "Miller chromosome spreads", electron micrographic displays of individual transcription units with their nascent transcripts, after treatment with weak detergent; the technology was developed by the late Oscar Miller. (A), Densely packed transcription unit from *Triturus* (newt) oocyte lampbrush chromosome. Calculations show that this transcription unit is being transcribed at near the maximum possible rate at ambient temperature (Davidson, 1986), and transcripts are spaced only about 100 nucleotides apart. Some polymerase molecules can be seen on the chromosomal DNA at the base of individual nascent transcripts. The initiation point for this gene is just beyond the lower left corner of the image *(from Miller and Bakken (1972))*. (B), Miller spreads of genes transcribed at much lower rates. These are far more typical (see text), and in the majority of transcription units visualized in these tissues only single nascent transcripts are observed by this method at any one time (Davidson, 1986). Because the intervals between successive initiations are long, in these cases several minutes, the nascent transcripts are much farther apart. (B1, B2), Sea urchin embryo (*Strongylocentrotus purpuratus*) transcription units. Scale bars represent 1 μm (i.e., ~3000 bp of DNA). The transcription unit in (B2) is transcribed with a slightly higher initiation rate than that in (B1), but compare (A); the arrow marks the point of initiation *(from Busby and Bakken (1979))*. (B3), Rabbit embryo transcription unit, also displaying infrequent initiations. Scale bar is 0.5 μm *(from Cotton et al. (1980))*. (C), Fundamental equations for mRNA synthesis dynamics, as discussed in text. The O.D.E. in Eqn (1) gives the rate of mRNA synthesis; Eqn (2) gives the solution of Eqn (1); and Eqn (3) defines the steady state amount of mRNA ($mRNA_{s.s.}$): parameters are k_s, the mRNA synthesis rate constant (number of newly synthesized mRNA molecules appearing in cytoplasm per minute); k_d, mRNA decay rate constant (fraction of mRNA pool decaying per minute, i.e., t^{-1}). (D), Graphical evaluation of Eqn (2), illustrating the eyeball estimation of these same constants: k_s is the initial slope, and $mRNA_{s.s.}$ is k_s/k_d, the plateau value, thus knowing k_s, k_d can be obtained. In this measurement, $k_s = 3$ molecules/min and $k_d = 0.012\,min^{-1}$, i.e., the half-life of the mRNA $ln2/k_d$ (tick on abscissa), is 60 min. *(From Ben-Tabou de-Leon and Davidson (2009).)*

A most useful quantitative relationship for those working with transcriptional processes or cascades thereof, is that which relates mRNA synthesis rate, mRNA accumulation over time, and mRNA turnover rate. This is Eqn (1) in Fig. 2.4(C), which we will consider to represent events on an average per cell basis. It is straightforward to think about dynamic transcriptional problems as processes that can be represented literally by differential equations, and Eqn (1) illustrates this perfectly. Equation (1) states the obvious: the rate of change in quantity of an mRNA per unit time ***(dmRNA/dt)***, is given by the rate of its synthesis (***k_s***) minus the rate of its decay by turnover (***k_d·mRNA***); that is to say by the difference between the rate of entry of the mRNA into the cytoplasm and the rate of its disappearance. Each of the three terms in Eqn (1) has the same units, molecules/min. Let us dwell momentarily on each of these terms: If the differential equation is solved, the left side becomes the absolute number of mRNA molecules present at each given point in time (this solution is Eqn (2) of Fig. 2.4(C), a plot of which is seen in Fig. 2.4(D)). Considering that there are two of each gene in the diploid cell we see that ***k_s*** is just $2 \times \boldsymbol{I}$, which directly relates the flow rate of the mRNA into the cytoplasm to the rate of transcriptional initiation per gene encoding the message. In most circumstances decay of any individual molecule in a pool of molecules occurs independently of how long it has been present, i.e., all mRNAs have an equal probability of being targets of the decay mechanism (nuclease attack) whether they have just arrived or are many hours old. However, the mRNA cannot be subject to these decay processes until it is present in the cytoplasm. The probability of decay (***k_d***) is expressed as the fraction of the mRNA pool that is likely to decay per time interval, for example 1% per minute (here $\boldsymbol{k_d} = 0.01\,min^{-1}$). Thus the net number of molecules decaying per minute is the number of mRNA molecules present in the pool at a given time multiplied by the probability of decay and the decay term would be written ***mRNA*** (molecules) × **0.01** per minute; the whole term is again molecules/min.

Here is why Eqns (1) and (2) are so useful. It has become experimentally straightforward to measure quantitatively the time course of accumulation of any specific mRNA or set of mRNAs by QPCR or other means. Consider the situation at early times after a gene is activated: the amount of mRNA yet made is small and so the decay term is insignificant, and thus the initial slope of the output of Eqn (2) directly gives the transcription initiation rate, a valuable parameter. When the decay term and synthesis term balance out and become equal, the left side of Eqn (1) becomes 0 since there is now no net change. The mRNA accumulation process for that gene is said to have attained steady state. This is seen graphically in Fig. 2.4(D) as the ultimate plateau. From Eqn (1), when $dmRNA/dt = 0$ we have Eqn (3), which defines the amount of mRNA (molecules) at steady state as k_s/k_d. This is again very useful for it provides a way of estimating either one of these constants if the other and the easily measured steady state quantity are known. The decay rate constant k_d is not always simple to measure directly, but it is easily extracted by estimation of the mRNA half-life ($t_{1/2}$). Imagine that the gene has turned off and whatever mRNA is present is decaying. From Eqn (1), when $k_s = 0$ we can derive k_d from the time required for 50% of the mRNA to decay: $k_d = ln2/t_{1/2}$.

The logical simplicity of Eqn (1) allows us to relate directly measurements of mRNA at given times to the dynamic parameters of transcriptional synthesis and mRNA turnover, and to define computationally the mRNA half-life and steady state. Thus we are enabled to rationalize the amounts of message present through real time in the quantitative terms of mRNA life and death processes.

6.2 Productive *cis*-regulatory occupancy by transcription factors in the nucleus

In a GRN, cascades of transcriptional events occur in which given regulatory genes are transcribed, producing transcription factors that activate other regulatory genes immediately downstream in the network, and by a similar process these in turn activate further regulatory genes. To evaluate such cascade dynamics, we require a means by which the real time intervals between the steps in the cascade can be computationally resolved. The steps of this computation are as follows: (1) consideration of the probability of *cis*-regulatory occupancy in terms of transcription factor concentration and of the intrinsic affinity of the factor for its specific target site sequence; (2) consideration of the nonspecific interaction of transcription factors with DNA sequences as well as the specific interactions at their target sites; (3) generalization, for occupancy of a *cis*-regulatory module by more than one transcription factor; (4) relation between occupancy of the *cis*-regulatory module and the initiation rate of the gene it controls; (5) computation of the accumulation of the transcription factor protein based on the computed initiation rate; (6) computation of (1)–(5) for the next step of the cascade. By this means we may arrive at what turns out to be an extremely useful parameter in considering network dynamics: the step time. This is the time period that elapses for a given system running at a given temperature between the transcriptional activation of a gene encoding an upstream transcription factor and the transcriptional activation of its target gene in a transcriptional cascade.

Transcription factor—DNA interaction

In life, sequence-specific interaction of transcription factors with their *cis*-regulatory target sites is considered productive if such interaction contributes to regulatory function, such as transcriptional initiation, or repressor-dependent transcriptional silencing. Sequence-specific DNA–protein interactions have been studied in vitro for decades, although conditions in the animal cell nucleus are in many ways dissimilar from the conditions controlling interactions of these same factors with oligonucleotides bearing their target sites in vitro. The main purpose of most current in vitro studies is to determine the target site sequence preferences ("position weight matrices, PWMs") for given factors. Recent high-throughput methods have provided enormous, statistically supported databases that provide PWMs for a large number of known transcription factors (Berger and Bulyk, 2009; Stormo and Zhao, 2010; Christensen et al., 2011; Robasky

and Bulyk, 2011; Zhao and Stormo, 2011). These databases prove to be of very great value in qualitative analyses of regulatory DNA sequence, since target site specificity is an intrinsic property of the amino acid sequence of the DNA-binding domain of the transcription factor, and similar DNA sequences are bound by individual factors in vivo and in vitro. Exceptions may occur if protein–protein interactions affect the binding behavior of given factors in ways that cannot be accessed easily in purified in vitro systems. In order to calculate the real life occupancy of a DNA-binding site within a nucleus, we have to be concerned with the concentration of the factor as well as the qualitative characteristics of the binding interaction. We shall treat the amount of binding as the fractional occupancy, $\boldsymbol{Y}$, where $\boldsymbol{Y}=1$ denotes complete occupancy of a site, or of required multiple sites, and $\boldsymbol{Y}=0$ denotes no binding or zero occupancy.

Considering network dynamics for developing systems in terms of real amounts and real times, an initial project is to try to understand what range of occupancies is being deployed in vivo. For any given genes, are the initiation rates often near maximal, indicating high values of $\boldsymbol{Y}$, or on the other hand are they usually expressed at modest or low levels? Not surprisingly the answer is different depending on the biological context. This and the next three chapters concern GRNs for various phases of embryonic development, and the vast majority of the regulatory genes at the nodes of these networks are expressed at relatively low levels. One exception is the syncytial fourteenth cleavage cycle *Drosophila* embryo, where it appears that prominent regulatory genes are being transcribed near maximum possible rates, suggesting a close to maximal occupancy of the relevant CRMs. Similarly, in Chapter 5 we deal with cell type specification, including terminal differentiation, and in that context some downstream differentiation genes also run near maximum rates of initiation. Here we focus on more typical developmental GRN dynamics. Many kinds of evidence converge on the conclusion that developmentally active regulatory genes are typically transcribed at relatively low rates of initiation. This is demonstrated in the sea urchin embryo as an example, where many relevant measurements are available (Davidson, 1986; Bolouri and Davidson, 2003). Direct time course measurements of transcript accumulation revealed a "default" mRNA half-life of about 3–5 h. This conclusion is substantiated in dozens of individual regulatory gene mRNA time courses that include decay phases following turn off of transcription (Materna et al., 2010). The same data show directly the low synthesis rates of regulatory gene transcripts. In addition, extensive quantitative transcriptome data confirm the low levels of regulatory gene mRNA transcripts (Tu et al., 2014). These data demonstrate that most active regulatory genes in this embryo are represented by only about 10–50 molecules of mRNA per cell; these mRNA's are generated at rates within a factor of two of about 100 molecules/h-embryo (Peter et al., 2012). Considering the number of cells in the respective domains expressing these genes, this means the initiation rates per gene and minute are for virtually all regulatory genes being expressed in the embryo only a few percent of the $\boldsymbol{I_{max}}$ calculated above, i.e., per active gene initiation occurs no more frequently than once every several minutes. The implication of these low $\boldsymbol{I}$ values is that there are low average levels of occupancy of the relevant CRMs by the driver factors controlling the expression of these genes. The significance of this for occupancy calculations is that we require a treatment that deals appropriately with intermittent low occupancy rather than preferentially with saturation occupancies.

In Fig. 2.5, Eqns (1) and (2), we see the equilibrium constant for transcription factor–DNA target site interaction from several vantage points. In a bimolecular reaction between transcription factor $\boldsymbol{A}$ and DNA target site $\boldsymbol{D}$, equilibrium is the point at which the rate of formation of the product, here factor $\boldsymbol{A}$-DNA site complex, or $\boldsymbol{AD}$, is equal to its rate of dissociation. Equation (1) shows the bimolecular reaction kinetics. Here $\boldsymbol{k_{as}}$ is the rate constant for formation of the complex; given this constant, the actual rate of formation of course depends on the concentrations of factor $\boldsymbol{A}$ and site $\boldsymbol{D}$. Once formed, the tightness of binding (affinity) determines the rate of complex dissociation, $\boldsymbol{k_{ds}}$, an intrinsic chemical feature of each DNA–protein sequence interaction, i.e., the tighter the binding (the more energy released) the longer the complex will last. For a given $\boldsymbol{k_{ds}}$, the actual amount of complex dissociating per unit time in Eqn (1) depends on the amount of complex there is. The factor finds its target site by diffusion, hence the rate at which this occurs, $\boldsymbol{k_{as}}$, depends essentially on the rate of diffusion. Since most transcription factors fall within a range of a factor of two or so in protein mass, on which the diffusion rate depends, $\boldsymbol{k_{as}}$ is about the same for almost all transcription factors. However, $\boldsymbol{k_{ds}}$ reflects the intrinsic energetics of the factor–DNA interaction, and since transcription

Figure 2.5

(A)

1: $\frac{dAD}{dt} = k_{as}A \cdot D - k_{ds}AD$

At equilibrium: $\frac{dAD}{dt} = 0$ and now

2: $\frac{AD}{A \cdot D} = \frac{k_{as}}{k_{ds}} = K_{eq}\ [\mathrm{M}^{-1}]$

3: $(S)K_{eq} = e^{-\Delta G/RT}$

4: $Y = \frac{e^{-\Delta G/RT} \cdot A}{1 + e^{-\Delta G/RT} \cdot A} = \frac{K_{eq} \cdot A}{1 + K_{eq} \cdot A}$

5: $Y_{AB} = \frac{K_{eqA} \cdot A \cdot K_{eqB} \cdot B \cdot K_q}{1 + A \cdot K_{eqA} + B \cdot K_{eqB} + K_{eqA} \cdot A \cdot K_{eqB} \cdot B \cdot K_q}$

6: $K_r = \frac{K_{eq}(specific)}{K_{eq}(non-specific)}$

7: $Y_{AB} = \frac{A \cdot K_{rA} \cdot B \cdot K_{rB} \cdot K_q}{D_N^2 + A \cdot D_N + B \cdot D_N + A \cdot K_{rA} \cdot D_N + B \cdot K_{rB} \cdot D_N + A \cdot K_{rA} \cdot B \cdot K_{rB} \cdot K_q}$

8: $I = I_{max} \cdot (1 - e^{-k_b \cdot Y_{AB}\ \ I_{max}})$

Parameters and Symbols

$mRNA$	molecules of mRNA
k_s	synthesis rate constant
k_d	mRNA decay rate constant
A, B	concentration of transcription factors A and B
D	concentration of specific target sites
AD, BD	concentration of [factor:DNA] complex
k_{as}	association rate constant
k_{ds}	dissociation rate constant
K_{eq}	equilibrium constant
S	standard medium activity coefficient
T	temperature
ΔG	change in free energy per mole
R	natural gas constant
Y	*cis*–regulatory occupancy
Y_{AB}	*cis*–regulatory occupancy for two factors, A and B
K_q	cooperativity constant
K_r	relative equilibrium constant
D_N	concentration of non-specific DNA target sites
I	initiation rate
k_b	initiation efficiency factor
k_T	translation rate constant
k_{dP}	protein decay rate constant
P	molecules of protein per cell

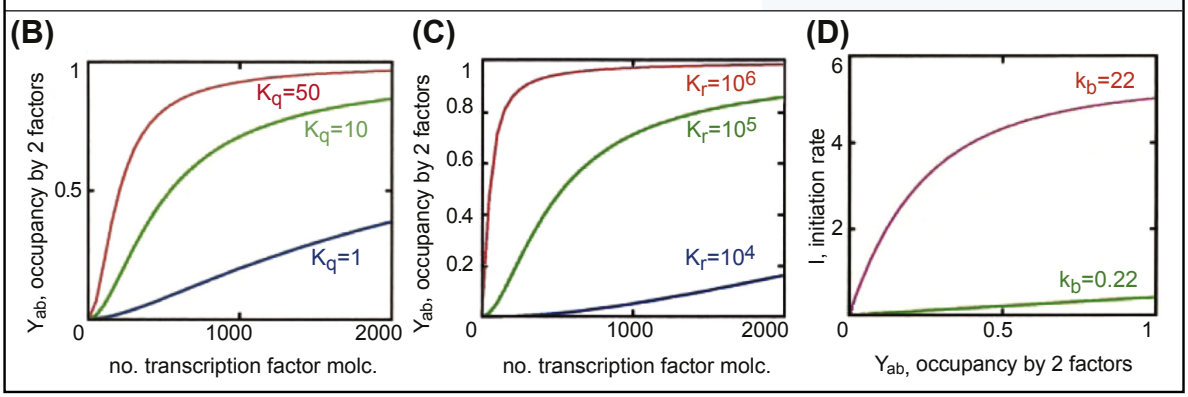


Figure 2.5 Kinetic treatment of DNA-protein interaction, *cis*-regulatory occupancy, and transcriptional initiation. (A), Equations for quantitation of DNA–protein interactions and transcriptional initiation; see list of terms at right for designations of symbols. These equations have been reviewed earlier (Emerson et al., 1985; Bolouri and Davidson, 2003; Ben-Tabou de-Leon and Davidson, 2009; Phillips et al., 2009). Equations (1–3), Definitions of equilibrium constant for formation of complex between transcription factor at concentration A and its specific DNA target site, at concentration D (a convenient use of concentration is to consider the number of molecules of factor per nuclear volume). Equations (1), O.D.E. giving rate of formation of the complex AD in a second order reaction. When this rate is 0 the reaction is in a state of equilibrium. Eq. (2), The equilibrium constant K_{eq} defined kinetically from Eqn (1) at $dAD/dt=0$, in terms of the association and dissociation rate constants k_{as} and k_{ds}; and stoichiometrically as the ratio of the molecules of complex to molecules of site D and factor A at equilibrium. Eq. (3), Thermodynamic definition of K_{eq} in terms of molar free energy change ΔG occurring in the reaction $A+D\leftrightarrow AD$; note that for this reaction the units of K_{eq} are M^{-1} while the right side of Eqn (3) is unit-less. This is because of an implicit term S [M], the activity coefficient, which accounts for the activity of the relatively high ionic strength of the solution in which these reactions take place, and is conventionally set at 1. Eq. (4), Fractional occupancy, Y, of a single target site when its factor is present at concentration A, and it binds the site with the equilibrium constant K_{eqA}; Y is the fraction of total site in complex, or the occupancy. Eq. (5), Double occupancy, Y_{AB}, for a *cis*-regulatory system with sites for factor A and also factor B when these factors are present at concentrations A and B, respectively; K_q, cooperativity constant for interaction between factor A and factor B (see text for discussion of this equation). Eq. (6), Definition of relative equilibrium constant K_r, i.e., ratio of K_{eq} for specific to K_{eq} for nonspecific DNA–protein interaction. Eq. (7), Y_{AB}, computed using K_r rather than K_{eq} for both A and B factors, and including K_q terms for cooperative interactions between A and B. Eq. (8), Relation between I and Y_{AB}; the higher the value of Y_{AB} the closer I will be to the maximum rate of initiation I_{max} (Eqns (7) and (8) from Bolouri and Davidson (2003)). (B–D), Simulations using typical sea urchin parameters *(from Bolouri and Davidson (2003))*. (B), Effect of cooperativity on Y_{AB}. (C), Effect of K_r on Y_{AB}. (D), Effect of initiation activation efficiency of given factors, k_b, on I, initiations per minute and gene.

factors vary over orders of magnitude in how strongly they bind their DNA target sites, $\boldsymbol{k_{ds}}$ varies likewise. The equilibrium constant $\boldsymbol{K_{eq}}$ is defined in Figure 2.5 Eqn (2) for the condition when $\boldsymbol{dAD/dt=0}$ in Eqn (1). We see that the equilibrium constant is the ratio of the two rate constants, $\boldsymbol{k_{as}/k_{ds}}$. Variations in the equilibrium constants for different transcription factor–target site interactions, depend almost entirely on $\boldsymbol{k_{ds}}$ since the values of $\boldsymbol{k_{as}}$ are usually similar. $\boldsymbol{K_{eq}}$ also gives the stoichiometry of the ratio of complex to unbound factor and site: thus from Eqn (2) we also see that the more factor there is the more complex will be formed, and the higher the equilibrium constant the more complex will form per amount of factor. This is the essential behavior which basically explains gene regulation: binding of a transcription factor to its target site is controlled by factor concentration, that is, by the activity of the upstream regulatory gene encoding the factor.

Occupancy

A classic approach is to treat occupancy as the physical manifestation of the probability that the specific sites in a regulatory DNA sequence will be in the productive bound state. The probability of productive occupancy of a CRM consisting of multiple binding sites is given by the likelihood with which the productive state is obtained compared to that of all possible states of binding. Thus the probability is calculated by normalizing the likelihood of occurrence of the desired bound state under given conditions to the sum of the likelihoods of occurrence of all the possible states of binding of the sites on the sequence. Here the normalization factor is the sum of the probabilities of all sites remaining unbound, of all combinations of

some sites bound and others not, plus the probability of the functional bound state. Transcription factor interaction with DNA is an energetic transaction, since the amino acid side chains of the DNA recognition domain interact chemically with residues of the nucleotides constituting the DNA target site. It follows that the probability of a particular state of transcription factor occupancy for a given DNA sequence, is actually the ratio of the particular interaction energy of the productive bound complex to the sum of all the interaction energies of all the complexes the sequence could generate ("Boltzmann partition function"; for modern derivation and discussion in a biophysical context see Phillips et al., 2009). It is worth spending a few lines to explore the parameters and behavior of this representation of *cis*-regulatory occupancy as a probabilistic function.

We need first to be clear about what quantitatively controls individual transcription factor–DNA target site interactions: intuitively, this has to depend on the amount of the factor in the nucleus and on its innate affinity for its target site(s). The first is conveyed by the factor concentration (molar), and the second by the equilibrium constant for this particular factor–site interaction. Since the equilibrium constant value depends on k_{ds}, which depends on the energetic exchange at binding, K_{eq} can also be defined in terms of energy released (per mole) upon binding: for reactions in ionic solutions we can evaluate the equilibrium constant as in Eqn (3) of Fig. 2.5(A) (see note in caption). We can now use the principle of the Boltzmann partition function to see how occupancy, Y, will change for the maximally simple system of a single site and a single-factor species as a function of its concentration, A: this is given in Eqn (4) of Fig. 2.5(A). Here the likelihood of complex formation is calculated by the ratio of bound state to bound plus unbound state. Thus, the energetic definition of the equilibrium constant for factor A multiplied by its concentration is used to compute the bound state, and in the denominator the same term appears plus that for the alternative that the site is not bound (for naked DNA site, ΔG is 0, hence the 1 in the denominator, see Eqn (3)). We can see that at low A concentrations, Y will rise linearly with A, and with increasing concentrations of A the value of Y in Eqn (4) approximates 1.

Now, to become a step more realistic, consider a system where a second factor at concentration B must also bind its site in the sequence to generate an occupancy configuration required for function. By the same algorithm as used in Eqn (4), we obtain Eqn (5) (Ackers et al., 1983). This equation includes an additional term, K_q which captures the possibility that the two factors A and B interact cooperatively on the DNA, contributing a further energetic term that additionally stabilizes the complex. The definition of K_q could be extended to include any mechanism by which binding of one factor energetically facilitates the binding of a second factor, for example by introducing torsion of the DNA, or by increasing the accessibility of binding sites for the second factor so that more complex is formed for a given concentration of factor.

One further step is required to transit closer to the real regulatory world of the animal cell nucleus. It was realized early on that because of the enormous amount of genomic DNA in animal cell nuclei (relative to bacteria, let alone phage), and because all transcription factors display some affinity for any DNA sequence, factors not bound to specific target sites will in general be transiently bound nonspecifically to the DNA. This must be taken into account when computing specific DNA–protein interactions (Von Hippel et al., 1974; Lin and Riggs, 1975; Emerson et al., 1985). The nonspecific binding of transcription factors to DNA is due to the general property of this class of proteins that they include basic domains which interact transiently with the acid phosphate bridges of the DNA helix as the proteins swivel along the genome. The average $t_{1/2}$ for these nonspecific interactions is very brief, on the order of one or a few milliseconds (e.g., Emerson et al., 1985; Calzone et al., 1988), but on the other hand there is a very large number of sites since every base pair in internucleosomal sequence throughout the genome begins a new binding site for nonspecific interactions. So what is really required for in vivo considerations of site occupancy is the relative equilibrium constant, K_r, as defined in Eqn (6) of Fig. 2.5(A). Exhaustive studies of DNA–protein interactions in complex sea urchin *cis*-regulatory systems revealed that all those displaying $K_r > \sim 5 \times 10^4$ as measured in vitro will have regulatory function in vivo (Calzone et al., 1988; Kirchhamer and Davidson, 1996; Yuh et al., 1998, 2001). Consistent with such K_r values, recent measurements of transcription factor/DNA interaction dynamics in mammalian cells indicate that these complexes last typically from a few seconds to several minutes, as might be expected from comparison with in vitro

binding results, i.e., taking into account the effects of high endogenous salt concentration in the nucleus (e.g., Poorey et al., 2013).

Incorporation of K_r into Eqn (5) of Fig. 2.5(A) results in Eqn (7) (Bolouri and Davidson, 2003). Plots of occupancy as a function of factor concentration and of cooperative interaction between the bound factors using Eqn (7) are also shown in Fig. 2.5. Note that cooperativity, which is probably the rule rather than the exception in loaded CRMs, has a strong effect on levels of occupancy (Fig. 2.5(B)). More importantly, significant occupancy is attained even at relatively low numbers of transcription factor molecules, given a very typical K_r of 10^5 (Fig. 2.5(C)). Here we see also that the computation supports the observation quoted above, namely that $K_r > \sim 5 \times 10^4$ means functional interaction; much less than this requires relatively large numbers of protein molecules to attain any occupancy.

Finally, it remains to relate occupancy to initiation rate, I. The probability of an initiation occurring in a time window might be expected to rise linearly at low occupancies, but to increase less efficiently with increasing occupancy at high levels because of interference by polymerase molecules that have not yet moved out of the way as the system operates closer and closer to I_{max}. The probability of an initiation is (1−the Poisson probability of no initiation). These simple ideas are captured in Eqn (8) of Fig. 2.5(A) (Bolouri and Davidson, 2003), where the rate of initiation as a fraction of I_{max} is related to the value of Y, computed as in Eqn (7); the higher the value of Y the lower the probability of no initiation, and at low Y, the initiation rate is essentially linear with Y. This result is also shown graphically in the simulations of Fig. 2.5(D). We are now equipped to deal with cascades of regulatory gene expression, that is, with network dynamics of the nature of those occurring in developmental GRNs.

6.3 Regulatory cascade dynamics

Familiar issues that arise continuously in working with GRNs include the following: if the transcription factor encoded by Gene 1 directly activates Gene 2, how much time will elapse between the onset of transcription of Gene 1 and the onset of transcription of Gene 2, i.e., the step time? About how many molecules of Gene 1 mRNA are likely to be required to make enough transcription factors to produce sufficient occupancy of the regulatory system of Gene 2 so that this gene can in turn generate enough molecules of its product to affect the next gene downstream? What do gene cascade kinetics look like? How important in a developing embryonic system is control of activity level? Here we apply real time measurements of kinetic constants using the computational apparatus in Figs. 2.4 and 2.5, and direct observations on regulatory gene cascades, in order to arrive at the answers to an interconnected set of important questions. It is relatively easy to measure accumulation time courses for multiple mRNAs simultaneously, and if we know enough to interpret the behavior of regulatory gene cascades, the measured kinetics can provide evidence that is useful in discriminating between alternative network architectures. Our examples will again be taken from the sea urchin embryo because both GRN architecture and the gene expression kinetics it generates are available. However, the conclusions are unlikely to be grossly different for other embryonic systems once a Q10 correction for temperature effects is made.

Before we can deal with these issues the rate of protein synthesis per mRNA needs to be taken into account; this has been measured at 2 molecules/min-mRNA for sea urchin embryos at 15 °C, a constant (Davidson, 1986). Application of Eqn (1) of Fig. 2.6(A), where this constant appears as k_T, now allows us to estimate the kinetics of transcription factor synthesis in terms of the kinetics of regulatory gene mRNA generation, computed as above.

A series of simulations was carried out (Bolouri and Davidson, 2003), using the occupancy, initiation, synthesis, and turnover treatments in Figs. 2.4, 2.5, and 2.6, and applying average values taken from a large number of prior measurements for the relative equilibrium and decay constants called for in these equations. Fig. 2.6(B) shows the kinetics of mRNA synthesis from a gene controlled by two transcription factors present at various concentrations per cell: note the predicted insensitivity of the

Figure 2.6

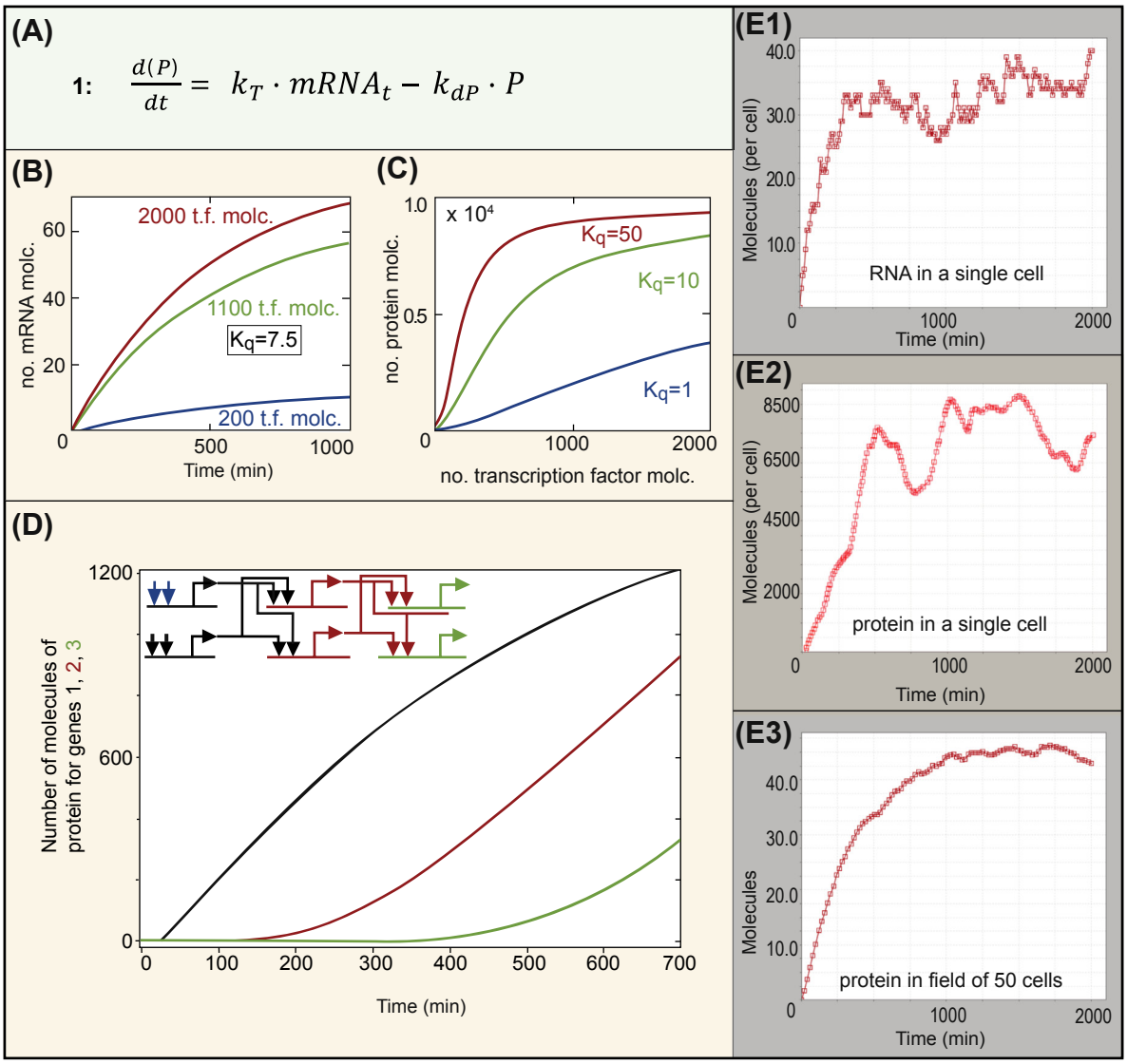


Figure 2.6 Protein output dynamics. (A), O.D.E giving rate of protein synthesis as a function of mRNA concentration (per cell); see Fig. 2.5 for terms. (B), mRNA output from a gene requiring two transcription factors as in Fig. 2.5 Eqn (7), as a function of the level of transcription factors present per cell (the concentrations of the two factors are set to be equal). (C), Protein output using the expression in (A), given 1100 transcription factor molecules, as in (B, green curve), shown for different degrees of cooperativity between the two transcription factors. (D), Gene cascade diagrammed at top, and kinetics with which the protein output of the second (red) and third (green) gene appear following activation of the first (black) gene (B–D from Bolouri and Davidson (2003)). (E), Effect of microscopic stochasticity in the intervals between successive initiations of transcription on mRNA and protein output (unpublished simulations of Bolouri and Davidson). Intervals were distributed around a mean by Poisson statistics and the RNA and protein output dynamics reported. (E1), mRNA output from a single cell. (E2), Protein output from a single cell, computed from the expression in (A). (E3), Protein output from a field of 50 uncoupled cells, all expressing the same gene.

response to very low factor concentrations. In this light it is interesting to see in Fig. 2.6(C) how quickly effective numbers of transcription factor molecules can be generated by a regulatory gene operating on two driver inputs present at intermediate levels, even in the absence of cooperativity ($K_q = 1$). The most important of these predictions is seen in Fig. 2.6(D). Here a cascade of three gene pairs in a tandem causal array is considered, such that double occupancy of the regulatory system of each gene by its immediately upstream transcription factors is required, and the output of the first pair of genes (black) activates the second two genes, the output of which (red) activates the third pair of genes (green). The plot shows, for typical constants including $K_q = 7.5$ and $K_r = 10^5$, that the interval between activation of an upstream Gene 1 and the activation of its downstream target Gene 2 is about 3 h, and this predicts the step time for this 15 °C system. There are two additional very important lessons, combining the import of Figs. 2.6(B) and 2.6(D). First, about 500–1000 transcription factor molecules per cell suffice to trigger activation of a direct target gene in a regulatory cascade (Fig. 2.6(D)), but this level of input factor leads to the accumulation of only <60 mRNA molecules per cell (Fig. 2.6(B)). Note that a similar level of mRNA is also produced if the level of upstream transcription factor is doubled to 2000 molecules, illustrating a profound insensitivity to transcription factor level once this rises beyond the effective threshold indicated in Fig. 2.6(B). This is why typically <60 regulatory gene mRNA molecules are to be found per cell in this system, and why the transcriptional initiation rates of regulatory genes are so leisurely. Second, from the shape of the curves in Fig. 2.6(D), we see that the downstream genes in the cascade are activated long before the gene products of the upstream genes ever attain steady state. Thus as soon as mRNA encoding transcription factors appears, it gets translated, and as soon as even modest levels of this transcription factor have accumulated it is used to activate target genes. A similar result, that is, activation of target genes long in advance of steady state, was obtained in a kinetic study of Ftz target gene transcription in *Drosophila* (Nasiadka et al., 2002). The direct implication is that control of transcription factor concentration at steady state is not relevant for target gene activation in a developing embryo, even though the only general way expression levels can ever be controlled closely is at steady state. Thus it is no surprise that when specific regulatory gene mRNA levels are measured in diverse batches of embryos, they often differ by 30–50% or more (Materna et al., 2010). As opposed to the finely tuned balanced homeostatic systems of physiology, developing embryos are forward drive systems and relatively level insensitive.

The insensitivity of cascade kinetic behavior to factor level in these systems provides a way of estimating the significance of microscopic transcriptional stochasticity. In Fig. 2.6(E) we see the stochasticity of gene output modeled under the assumption that transcriptional initiation occurs stochastically in time according to a Poisson distribution. The gene output is shown for a single cell as mRNA level over time (Fig. 2.6(E1)) and given canonical turnover rates, as protein level over time (Fig. 2.6(E2)). Even though the accumulation curves are noisy, the actual amount of variation will have little effect on cascade kinetics. In a field of 50 similarly functioning cells, even this amount of noise is effectively averaged out (Fig. 2.6(E3)).

The above computation, built using the kinetics of basic molecular processes of gene expression, predicts a step time of 3 h for this sea urchin embryo (Bolouri and Davidson, 2003). A remarkable confirmation that this step time is indeed the metric by which the sea urchin GRN operates was obtained 10 years after these a priori simulations were generated. In a dynamic model analysis of the sea urchin embryo GRN (Peter et al., 2012), which we discuss later in this book (Chapter 6), the 3 h step time was applied across the whole GRN to the large number of known direct regulatory gene inputs, over a 30 h period. The result was to produce a computed sequence of expression patterns extremely close to those observed, while imposition of 2 h or 4 h step times created disastrous nonconcordance with observation. This general result is also in agreement with many individual experimental observations made during study of the sea urchin embryo GRN that repeatedly revealed approximately 3 h intervals between activation of known regulatory genes and of their immediate target genes.

In summary, the quantitative arguments given here provide a basis on which the dynamics of gene expression and regulation can be rationalized, in terms of the rate limiting molecular functions that

actually control the dynamics. A major conclusion that speaks to the general character of regulatory system dynamics in the embryo is that the regulatory state changes progressively in real time, in a way not dependent on steady state levels. We have here ignored the intermediate dynamics of nuclear pre-mRNA processing, since essentially it contributes only a time lag (for this see Bolouri and Davidson, 2003). Nor have we dealt with repression, for the reason that repression is often a multistep process in which the role of the transcription factors that initiate the process is transient. Transcriptional repressors may recruit chromatin repression complexes that maintain the silenced state in time (Chapter 1), and which cannot be encompassed in a literal transcription process model such as we have dealt with here. Both repression and activation are conventionally modeled mathematically by expressions taken from (approximate) kinetic analysis of enzymatic reactions, although mechanistically neither transcriptional activation nor transcriptional repression actually resembles such reactions. These mathematical approaches give the required external behavior, however, and they have enabled very interesting and useful analyses of small regulatory circuits, as we review later in this book (Chapter 6). Here we have structured the discussion so as to preserve as direct as possible a relation between the mathematical terms and the physical processes of gene regulation.

7. Historical Origins and Antecedents of GRN Theory

In this chapter we have taken up the different aspects of regulatory system mechanism which, when combined, produce the powerful synthetic concept that animal development is controlled by GRNs. These mechanistic aspects are now well and deeply supported by multiple kinds of experimental result. Our current concepts grew from separate ideological roots of diverse origin; some are of modern vintage, others descend from insights that go back to the initial period of causal developmental biology. To remind ourselves of the breadth and depth of the conceptual terrain in which developmental GRN theory is embedded is to enrich our appreciation for it.

It is interesting to consider how many different scientific trajectories have contributed to the overall concept of developmental GRNs. First there are the basic concepts that the process of development is encoded in the genome, that genomic information is equivalent in all cells of the embryo, and that therefore the differential readout of this information in different cells requires the existence of a regulatory system which is itself encoded in the genome. Second is the fundamental idea that developmental spatial gene expression differs from place to place and time to time because of combinatorial transcriptional control at the *cis*-regulatory sequences of each individual gene. Third are the concepts of systems biology, in particular the idea that a complex process like development must be the output of a large number of individual control functions and that the framework explanation must include all or most of these in order to arrive at a satisfactory explanation. Furthermore, this explanation cannot be derived by adding up observations on individual corners or beams without knowledge of the architecture of the edifice as a whole. Finally there is what could be described as the biological regulatory theory of GRN structure/function relations per se. This includes the spatial logic outputs of network subcircuits according to their architecture, the significance of the innate hierarchy of developmental GRNs, and the means by which GRNs encompass spatial changes of state by interdomain signaling. We leave for later in this book (Chapter 7) the large additional body of concept that arises when evolution of the body plan is considered in terms of change in GRN architecture.

Each of these areas of concept and advance could of course profitably be the subject of a treatise in the history of modern molecular biology, and what follows are only very brief sketches. Some aspects are only of recent origin: for example, the systems biology arguments are of essentially modern vintage. The antecedents of most of the strains of thought comprising our current synthetic concept lie further back, however. A brief reminder of some of the twentieth century way stations along each of these scientific pathways shows how wide and deep are these antecedents.

7.1 Organism-wide genomic regulatory system required for development

The concept that there is resident in the chromosomes of every cell a genomic program that controls cellular functions during development arose explicitly in the first decades of the twentieth century. A landmark along this pathway was the famous polyspermy experiment of Boveri (Boveri, 1905, 1907; Laubichler and Davidson, 2008). Using a protocol that engendered aneuploidy in early blastomeres of sea urchin embryos, Boveri showed quantitatively that only those blastomeres containing complete chromosome sets are capable of giving rise to morphologically complete larvae. He concluded that every chromosome contains unique genetic determinants and that all chromosomes (i.e., the whole genome) are required to be present in every cell for embryogenesis to be completed: ergo, the genomic program for development. Contemplating Boveri's work together with the accumulating evidence of chromosome behavior in mitosis and meiosis, interspecific nuclear transfer experiments, and much else that was qualitatively understood about development in the era before molecular biology, E.B. Wilson enunciated clearly the proposition that "hereditary" information in the chromosomes encodes the developmental process (to use our modern words; Wilson, 1925). He then followed a similar trail of logic as in the introductory section of Chapter 1 of this book, deducing from the increase of complexity during the process of development (which he refers to as "epigenesis") that: "…heredity is effected by the transmission of a nuclear preformation which in the course of development finds its expression in a process of cytoplasmic epigenesis" (Wilson, 1925; Peter and Davidson, 2013). Thus there had already coalesced the fundamental concept of a genomic program for development operative in all cells. For about the next half century there was no great conceptual advance in the state of this completely accurate idea per se. However, the fundamentally important point that all cells of the developing organism contain the same DNA genome was established in a particularly incontrovertible manner by John Gurdon's nuclear transplantation experiments of the 1960s (Gurdon and Uehlinger, 1966; Laskey and Gurdon, 1970). Evidence accumulated rapidly after the mid-1980s from what we have called the fundamental experiment in developmental gene regulation, i.e., gene transfer into eggs, demonstrating how spatially localized transcription is controlled by *cis*-regulatory interactions. Only then could come into focus what the regulatory system controlling development mechanistically entails, but its majestic dimensions emerged only with genomics.

7.2 Combinatorial transcriptional regulation of spatial gene expression

As we now know, *cis*-regulatory logic processing of multiple positive and negative inputs is the basic mechanism of spatial gene expression control, operating at all the nodes of developmental GRNs. But in terms of the overall history of our field, among the insights essential to our current concepts this was one of the last to coalesce. Until the late 1940s and early 1950s, what genes are in material terms remained very unclear, and the problem of how gene expression is regulated was inaccessible. This is not to say that no one earlier than this had figured out from first principles the outlines, if not the molecular biology, of how things actually do work: consider the remarkable prescience of the following quote, again from E.B. Wilson, but this from back in 1896: "If chromatin be the idioplasm (genetic material) in which inheres the sum total of hereditary forces, and if it be equally distributed at every cell division, how can its mode of action so vary in different cells as to cause diversity of structure, i.e., *differentiation*?… My own conception…is as follows. All the nuclei are equivalent, and all contain the same idioplasm… Through the influence of this idioplasm the cytoplasm of the egg or of the blastomeres derived from it undergoes specific progressive changes, each change then reacting upon the nucleus and then initiating a new change. These changes differ in different regions of the egg because of pre-existing differences…such as the distribution of different substances in the egg cytoplasm" (Wilson, 1896). If for "change" we substitute "new transcription factor", we have essentially how spatially differential gene expression does in fact get going in the early

embryo (see next chapter). But this was not a dominant view, particularly among those who actually worked on genes, i.e., geneticists. As T.H. Morgan complained of dominant opinion in the area of what we would call developmental genetics in 1934, "the implication in most genetic interpretations is that all the genes are acting all the time in the same way." He then proceeds to argue in the opposite direction: "…an alternative view would be that different batteries of genes come into action as development proceeds" (Morgan, 1934). This anticipates the correct view that genes are differentially expressed. Morgan went on in this passage to adduce exactly the same egg cytoplasmic mechanism as in Wilson's 1896 quote (without attribution). Thus, regulation of gene expression reentered the arena, this time to stay. By the early 1950s the theory of "variable gene activity" had been explicitly developed as a general explanation of differentiation (Brachet, 1949; Sonneborn, 1950; Stedman and Stedman, 1950; Mirsky, 1951, 1953). Overwhelming evidence accumulated in the next three decades from molecular genetics and molecular developmental biology that this theory is correct, and that during development regulatory control of gene activity is specifically what accounts for differential gene expression, which is in turn the driver of development.

The weakness of this body of mechanistic knowledge was that it was all about how individual genes behave with little direct illumination of the holistic processes of development. On the other hand, an entirely unrelated theoretical trajectory oriented toward these holistic aspects was based on evidence of inductive gradients. This gave rise to attempts to interpret the spatial patterns formed in development from a completely different starting point, as the global output of "reaction diffusion" gradient systems. The relevance of these ideas to developmental pattern formation was challenged by accumulating molecular evidence regarding the mechanisms for control of spatial gene expression. A particularly illuminating step forward was the identification of multiple, specific CRMs which respectively generate the stripes of pair rule gene expression in *Drosophila*, utilizing combinatorial inputs from activators and repressors which at the DNA level delimit transcriptional activity and hence form the sharp stripe boundaries (Stanojevic et al., 1991; Small et al., 1996). By now, insights have accumulated from thousands of structure/function experiments on *cis*-regulatory expression constructs inserted into eggs, which show exactly where the explanation for developmental spatial gene expression actually lies: in the encoded *cis*-regulatory DNA sequence (for review of spatial information processing in specific developmental CRMs from many different embryonic systems, see Davidson, 2006).

7.3 Hierarchical structure of GRNs

An early attempt at constructing a model for gene regulation in development was formulated in 1969 (Britten and Davidson, 1969). This was a model for gene regulation which proposed many large-scale structural and logical features that much later turned out to be intrinsic to the organization of actual developmental GRNs. These features include the following. First, this model assumed a very large-scale system of interacting regulatory genes. Second, it was a hierarchical model in which upper level regulatory genes control other regulatory genes at the top of the hierarchy in each developmental process. Third, it was a network model in which individual genes respond to multiple *trans*-regulatory inputs produced by other regulatory genes, while each such *trans*-regulator has multiple targets. Fourth, it treated inductive signals as activators of special response elements which cause transcription of upper level hierarchy regulatory genes. Fifth and most generally, its presumption was that development is indeed determined by a distributed genomic sequence code. Virtually nothing was then known of the molecular biology of transcriptional regulation of the gene in animals. This model assumed that RNA rather than protein is the active product of the regulatory genes. In retrospect it is interesting to see that this largely erroneous assumption had so very little effect on the structure of its underlying logic. The arguments of the 1969 model soon led to the explicit proposition that change in the genomic sequence code for development that results in change in the architecture of regulatory networks, would turn out to be the engine of evolutionary change in body plans (Britten and Davidson, 1971). The spectacular technical advance of gene cloning soon ensued, and for many years thereafter

the novel experimental opportunities to examine expression, function, and regulation of individual genes dominated research. As an indirect consequence, system-scale thought on the subject of developmental gene regulation was almost buried, and so were system-scale measurements, except for studies on a few specific large gene families, and some measurements of global parameters of genome organization and expression. However, by the end of the 1990s, in the light of all that had been learned about gene regulation, it became possible to perceive the shape of developmental gene regulatory networks the existence of which had been proposed by Britten and Davidson in 1969. The subject of developmental gene networks was rejuvenated first in principle (Arnone and Davidson, 1997), and then in respect to the initial experimental attempt at large-scale GRN analysis in the sea urchin embryo (Davidson et al., 2002). This has now led to the first relatively complete experimentally obtained GRN model for a large-scale developmental process for which the general explanatory value has been computationally demonstrated (Peter et al., 2012).

In the next three chapters we see all of the conceptual aspects of transcriptional regulatory networks that we have reviewed here in operation at different levels of the process of development. In Chapter 3, the GRNs considered control the formation of the embryo, in Chapter 4 they control the formation of adult body parts, and in Chapter 5 they control the specification of cell types from multipotential precursors. Throughout, our focus has been on developmental processes for which experimental analysis has revealed the genomic basis of the underlying regulatory logic. This leads inexorably to consideration of development through the lens of GRN structure and function.

REFERENCES

Ackers, G.K., Shea, M.A., Smith, F.R., 1983. Free energy coupling within macromolecules. The chemical work of ligand binding at the individual sites in co-operative systems. J. Mol. Biol. 170, 223–242.

Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., Schubert, L.A., Birditt, B., Shay, T., Goren, A., Zhang, X., Smith, Z., Deering, R., McDonald, R.C., Cabili, M., Bernstein, B.E., Rinn, J.L., Meissner, A., Root, D.E., Hacohen, N., Regev, A., 2009. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science 326, 257–263.

Arnone, M.I., Davidson, E.H., 1997. The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851–1864.

Arnosti, D.N., Kulkarni, M.M., 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. 94, 890–898.

Bell, A.C., West, A.G., Felsenfeld, G., 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98, 387–396.

Ben-Tabou de-Leon, S., Davidson, E.H., 2009. Modeling the dynamics of transcriptional gene regulatory networks for animal development. Dev. Biol. 325, 317–328.

Berger, M.F., Bulyk, M.L., 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. 4, 393–411.

Bolouri, H., Davidson, E.H., 2003. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. Proc. Natl. Acad. Sci. U.S.A. 100, 9371–9376.

Boveri, T., 1905. Über die Abhängigkeit der Kerngrösse und Zellenzahl bei Seeigellarven von der Chromosomenzahl der Ausgangszellen. Zellenstudien, Jena.

Boveri, T., 1907. Zellenstudien VI. Die Entwicklung dispermer Seeigeleier. Ein Beitrag zur Befruchtungslehre und zur Theorie des Kerns. Gustav Fischer, Jena.

Brachet, J., 1949. L'hypothese des plasmagenesdans le developpement et la differentiation. Colloq. Int. C.N.R.S 8.

Britten, R.J., Davidson, E.H., 1969. Gene regulation for higher cells: a theory. Science 165, 349–357.

Britten, R.J., Davidson, E.H., 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. 46, 111–138.

Busby, S., Bakken, A., 1979. A quantitative electron microscopic analysis of transcription in sea urchin embryos. Chromosoma 71, 249–262.

Calzone, F.J., Thézé, N., Thiebaud, P., Hill, R.L., Britten, R.J., Davidson, E.H., 1988. Developmental appearance of factors that bind specifically to *cis*-regulatory sequences of a gene expressed in the sea urchin embryo. Genes. Dev. 2, 1074–1088.

Christensen, R.G., Gupta, A., Zuo, Z., Schriefer, L.A., Wolfe, S.A., Stormo, G.D., 2011. A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic Acids Res. 39, e83.

Cotton, R.W., Manes, C., Hamkalo, B.A., 1980. Electron microscopic analysis of RNA transcription in preimplantation rabbit embryos. Chromosoma 79, 169–178.

Davidson, E.H., 1986. Gene Activity in Early Development, third ed. Academic Press/Elsevier, San Diego, CA.

Davidson, E.H., 2006. The Regulatory Genome. Gene Regulatory Networks in Development and Evolution. Academic Press/Elsevier, San Diego, CA.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Schilstra, M.J., Clarke, P.J., Rust, A.G., Pan, Z., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Dev. Biol. 246, 162–190.

Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., de Laat, W., 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes. Dev. 18, 2485–2490.

Elgar, G., Vavouri, T., 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet. 24, 344–352.

Emerson, B.M., Lewis, C.D., Felsenfeld, G., 1985. Interaction of specific nuclear factors with the nuclease-hypersensitive region of the chicken adult beta-globin gene: nature of the binding domain. Cell 41, 21–30.

Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., Wu, Q., 2012. CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. Proc. Natl. Acad. Sci. U.S.A. 109, 21081–21086.

Gurdon, J.B., Uehlinger, V., 1966. "Fertile" intestine nuclei. Nature 210, 1240–1241.

Hadchouel, J., Carvajal, J.J., Daubas, P., Bajard, L., Chang, T., Rocancourt, D., Cox, D., Summerbell, D., Tajbakhsh, S., Rigby, P.W., Buckingham, M., 2003. Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by a combination of elements dispersed throughout the locus. Development 130, 3415–3426.

Hare, E.E., Peterson, B.K., Eisen, M.B., 2008a. A careful look at binding site reorganization in the even-skipped enhancers of *Drosophila* and sepsids. PLoS Genet. 4, e1000268.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., Eisen, M.B., 2008b. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet. 4, e1000106.

Ho, M.C., Schiller, B.J., Akbari, O.S., Bae, E., Drewell, R.A., 2011. Disruption of the abdominal-B promoter tethering element results in a loss of long-range enhancer-directed Hox gene expression in *Drosophila*. PLoS One 6, e16283.

Istrail, S., Davidson, E.H., 2005. Logic functions of the genomic *cis*-regulatory code. Proc. Natl. Acad. Sci. U.S.A. 102, 4954–4959.

Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., Haussler, D., 2007. Human genome ultraconserved elements are ultraselected. Science 317, 915.

Kirchhamer, C.V., Davidson, E.H., 1996. Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the CyIIIa gene *cis*-regulatory system. Development 122, 333–348.

Krivega, I., Dean, A., 2012. Enhancer and promoter interactions-long distance calls. Curr. Opin. Genet. Dev. 22, 79–85.

Kulkarni, M.M., Arnosti, D.N., 2003. Information display by transcriptional enhancers. Development 130, 6569–6575.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., Cohen, B.A., 2012. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. Proc. Natl. Acad. Sci. U.S.A. 109, 19498–19503.

Laskey, R.A., Gurdon, J.B., 1970. Genetic content of adult somatic cells tested by nuclear transplantation from cultured cells. Nature 228, 1332–1334.

Laubichler, M.D., Davidson, E.H., 2008. Boveri's long experiment: sea urchin merogones and the establishment of the role of nuclear chromosomes in development. Dev. Biol. 314, 1–11.

Liberman, L.M., Stathopoulos, A., 2009. Design flexibility in *cis*-regulatory control of gene expression: synthetic and comparative evidence. Dev. Biol. 327, 578–589.

Lin, S., Riggs, A.D., 1975. The general affinity of lac repressor for *E. coli* DNA: implications for gene regulation in procaryotes and eucaryotes. Cell 4, 107–111.

Longabaugh, W.J., Davidson, E.H., Bolouri, H., 2005. Computational representation of developmental genetic regulatory networks. Dev. Biol. 283, 1–16.

Longabaugh, W.J., Davidson, E.H., Bolouri, H., 2009. Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. Biochim. Biophys. Acta 1789, 363–374.

Ma, J., 2011. Transcriptional activators and activation mechanisms. Protein Cell 2, 879–888.

Marsman, J., Horsfield, J.A., 2012. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. Biochim. Biophys. Acta 1819, 1217–1227.

Materna, S.C., Nam, J., Davidson, E.H., 2010. High accuracy, high-resolution prevalence measurement for the majority of locally expressed regulatory genes in early sea urchin development. Gene Expr. Patterns 10, 177–184.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., Kellis, M., Lander, E.S., Mikkelsen, T.S., 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. 30, 271–277.

Miller, O.L., Bakken, A.H., 1972. Morphological studies of transcription. Acta Endocrinol. Suppl. (Copenh) 168, 155–177.

Mirsky, A.E., 1951. Some chemical aspects of the cell nucleus. In: Dunn, L.C. (Ed.), Genetics of the 20th Century. Macmillan, New York.

Mirsky, A.E., 1953. The chemistry of heredity. Sci. Am. 188, 47–57.

Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., Nakao, M., 2009. Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J. 28, 1234–1245.

Morgan, T.H., 1934. Embryology and Genetics. Columbia Univ. Press, New York.

Nasiadka, A., Dietrich, B.H., Krause, H.M., 2002. Advances in Developmental Biology and Biochemistry. Elsevier.

Nolis, I.K., McKay, D.J., Mantouvalou, E., Lomvardas, S., Merika, M., Thanos, D., 2009. Transcription factors mediate long-range enhancer-promoter interactions. Proc. Natl. Acad. Sci. U.S.A. 106, 20222–20227.

Oliveri, P., Tu, Q., Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc. Natl. Acad. Sci. U.S.A. 105, 5955–5962.

Ong, C.T., Corces, V.G., 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. 12, 283–293.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M., Ahituv, N., Pennacchio, L.A., Shendure, J., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. 30, 265–270.

Peng, G.H., Chen, S., 2011. Active opsin loci adopt intrachromosomal loops that depend on the photoreceptor transcription factor network. Proc. Natl. Acad. Sci. U.S.A. 108, 17821–17826.

Peter, I.S., Davidson, E.H., 2011. A gene regulatory network controlling the embryonic specification of endoderm. Nature 474, 635–639.

Peter, I.S., Davidson, E.H., 2013. Transcriptional network logic: the systems biology of development. In: Walhout, A.J.M., Vidal, M., Dekker, J. (Eds.), Handbook of Systems Biology. Concepts and Insights. Academic Press/Elsevier, pp. 211–228.

Peter, I.S., Faure, E., Davidson, E.H., 2012. Feature Article: predictive computation of genomic logic processing functions in embryonic development. Proc. Natl. Acad. Sci. U.S.A. 109, 16434–16442.

Phillips, R., Kondev, J., Theriot, J., 2009. Physical Biology of the Cell. Garland Science, New York.

Poorey, K., Viswanathan, R., Carver, M.N., Karpova, T.S., Cirimotich, S.M., McNally, J.G., Bekiranov, S., Auble, D.T., 2013. Measuring chromatin interaction dynamics on the second time scale at single-copy genes. Science 342, 369–372.

Ren, X., Siegel, R., Kim, U., Roeder, R.G., 2011. Direct interactions of OCA-B and TFII-I regulate immunoglobulin heavy-chain gene transcription by facilitating enhancer-promoter communication. Mol. Cell. 42, 342–355.

Robasky, K., Bulyk, M.L., 2011. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 39, D124–D128.

Royo, J.L., Maeso, I., Irimia, M., Gao, F., Peter, I.S., Lopes, C.S., D'Aniello, S., Casares, F., Davidson, E.H., Garcia-Fernandez, J., Gomez-Skarmeta, J.L., 2011. Transphyletic conservation of developmental regulatory state in animal evolution. Proc. Natl. Acad. Sci. U.S.A. 108, 14186–14191.

Schwab, K.R., Patel, S.R., Dressler, G.R., 2011. Role of PTIP in class switch recombination and long-range chromatin interactions at the immunoglobulin heavy chain locus. Mol. Cell. Biol. 31, 1503–1511.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Small, S., Blair, A., Levine, M., 1996. Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. Dev. Biol. 175, 314–324.

Sonneborn, T.M., 1950. The cytoplasm in heredity. Hered. (Edinb) 4, 11–36.

Spitz, F., Furlong, E.E., 2012. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626.

Stanojevic, D., Small, S., Levine, M., 1991. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. Science 254, 1385–1387.

Stedman, E., Stedman, E., 1950. Cell specificity of histones. Nature 166, 780–781.

Stormo, G.D., Zhao, Y., 2010. Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. 11, 751–760.

Swanson, C.I., Evans, N.C., Barolo, S., 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev. Cell. 18, 359–370.

Thanos, D., Maniatis, T., 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83, 1091–1100.

Tu, Q., Cameron, R.A., Davidson, E.H., 2014. Quantitative developmental transcriptomes of the sea urchin *Strongylocentrotus purpuratus*. Dev. Biol. 385, 160–167.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., Blobel, G.A., 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol. Cell. 17, 453–462.

Von Hippel, P., Revzin, A., Wang, A., 1974. Non-specific DNA binding of genome regulating proteins as a biological control mechanism. Proc. Nat. Acad. Sci. U.S.A. 71, 4808–4812.

Wilson, E.B., 1896. The Cell in Development and Heredity, first ed. Macmillan, New York.

Wilson, E.B., 1925. The Cell in Development and Heredity, third ed. Macmillan, New York.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J., Cooke, J.E., Elgar, G., 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3, e7.

Yang, J., Corces, V.G., 2012. Insulators, long-range interactions, and genome function. Curr. Opin. Genet. Dev. 22, 86–92.

Yuh, C.H., Bolouri, H., Davidson, E.H., 1998. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896–1902.

Yuh, C.H., Bolouri, H., Davidson, E.H., 2001. *Cis*-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development 128, 617–629.

Yun, K., So, J.S., Jash, A., Im, S.H., 2009. Lymphoid enhancer binding factor 1 regulates transcription through gene looping. J. Immunol. 183, 5129–5137.

Zhao, Y., Stormo, G.D., 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. 29, 480–483.