Chapter 2

Components and Mechanisms of Regulation of Gene Expression

Alper Yilmaz and Erich Grotewold

Abstract

The control of gene expression is a biological process essential to all organisms. This is accomplished through the interaction of regulatory proteins with specific DNA motifs in the control regions of the genes that they regulate. Upon binding to DNA, and through specific protein–protein interactions, these regulatory proteins convey signals to the basal transcriptional machinery, containing the respective RNA polymerases, resulting in particular rates of gene expression. In eukaryotes, in addition and complementary to the binding of regulatory proteins to DNA, chromatin structure plays a role in modulating gene expression. Small RNAs are emerging as key components in this process. This chapter provides an introduction to some of the basic players participating in these processes, the transcription factors and co-regulators, the *cis*-regulatory elements that often function as transcription factor docking sites, and the emerging role of small RNAs in the regulation of gene expression.

Key words: Promoter, DNA-binding, operon, *cis*-regulatory element, microRNA, small interfering RNA.

1. Introduction

Cells can be considered as membrane-enclosed environments in which many different proteins undertake one or several specific functions. Thus, the proper development and the functional integration of cells within an organism depend on controlling the accumulation of these proteins within some defined concentration restrictions, which are space and time dependent. Consistent with the central dogma of biology, which states that the genetic information flow is, in general terms, from DNA to RNA and then to the proteins, the instructions on how much and when a

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_2,

protein needs to be made are encoded in the DNA. The process of transcription transfers the code responsible for making proteins, the cell workhorses, from the DNA to RNA and translation converts a messenger RNA (mRNA) sequence into a sequence of amino acids in a protein. Thus, protein levels can be controlled at multiple stages, including transcription, translation as well as mRNA and protein transport and stability. This chapter will primarily focus on the control mechanisms associated with transcription and responsible for how much mRNA is being made for each of the thousands (or tens of thousands) protein-encoding genes in a cell.

2. Description

2.1. Mechanisms of Transcription

In simple terms, the process of transcription involves the unwinding double stranded DNA and the chemical synthesis of RNA, using one of the two genomic DNA strands as the template for the RNA sequence. This is achieved by DNA-dependent RNA polymerases (RNAP). In prokaryotes, there is a single type of RNAP, which is responsible for the generation of various types of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). In eukaryotes, however, there are multiple RNAPs, each specialized in the production of particular types of RNA species. For example, RNAP I synthesizes rRNAs, RNAP II synthesizes mRNAs, and RNAP III synthesizes tRNAs. In addition, there are other RNAP with functions more restricted to particular kingdoms. For example, in plants, RNAP IV synthesizes small interfering RNA (siRNAs) (1, 2) and RNAP V transcribes intergenic and non-coding sequences, participating in the small interfering RNA(siRNA)-mediated transcriptional gene silencing (TGS) (3, 4).

To ensure proper gene expression levels, the activity of prokaryotic RNAP and eukaryotic RNAP II, in particular, are subjected to tight control. One of the best-studied mechanisms involved in regulating RNAP II activity is through the effect of transcription factors (TFs), which specify when and where RNAP II (and associated factors) is tethered to DNA, how RNAP II initiates (and re-initiates once a round of mRNA formation has been completed) transcription, and elongates nascent mRNAs. We define here TFs as proteins that bind DNA in a sequence-specific fashion to particular DNA sequences (*cis*-regulatory elements) located in the regulatory regions of the genes that they control. This definition excludes the large number of proteins that can affect gene expression without binding to specific DNA sequences. As these proteins often function by modulating the

action of specific DNA-binding TFs, there are few common characteristics that permit their easy identification.

TFs are usually classified into families, based on the presence of specific structures in their DNA-binding or protein–protein interaction domains. In vitro, TFs usually recognize DNA sequences 6–8-bp long, length that is clearly insufficient for the exquisite regulatory specificity that they display in vivo, suggesting that large number of TFs form the active regulatory complexes and providing the bases for the principle of combinatorial gene regulation (5).

In prokaryotes, binding of RNAP to specific regions is achieved by a particular protein factor, the sigma (σ) subunit. This prokaryotic TF increases the affinity of RNAP to certain promoter regions while decreasing its affinity to non-specific DNA. The σ factor responsible for the regulation of most "housekeeping" genes in *Escherichia coli* is σ^{70} and σ^{A} in *Bacillus subtilis*, which are responsible for initiating transcription from most promoters. Other σ factors are usually stress induced, to allow organisms to become virulent or adapt to any number of environmental changes such as hyperosmolarity, heat shock, oxidative stress, nutrient deprivation, and variations in pH (6, 7).

2.2. Organization of Gene Regulatory Sequences

2.2.1. Operons and Other Gene Clusters

One strategy by which prokaryotic organisms control the expression of genes that participate in a common process is to group the genes into operons, which are usually transcribed from a unique promoter resulting in a single (poly-cistronic) mRNA that is translated into multiple proteins, allowing the cell to streamline the control of transcription. Here, we describe the *lac* operon as an archetypical bacterial operon, as an example of how prokaryotes negotiate the control of gene expression (**Fig. 2.1**).

The *lac* operon encodes for three enzymes (*lacZ* encoding β-galactosidase, *lacY* encoding a lactose permease, and *lacA* encoding a trans-acetylase) necessary for the uptake and metabolism of lactose. Only when lactose but no glucose, a more favorable carbon source, is present in the environment, the *lac* operon is expressed. When grown in glucose, for example, regardless of whether lactose is present or not, the *lacZYA* genes are not expressed, a consequence of a repressor protein (*lac* repressor) recognizing the operator sequence of the operon regulatory region, preventing the recruitment of RNAP to the DNA. When lactose is present, this small molecule recognizes the *lac* repressor, preventing it from binding the operator sequence.

In eukaryotes, operon-like structures have been described, although they clearly differ from bacterial operons, since they

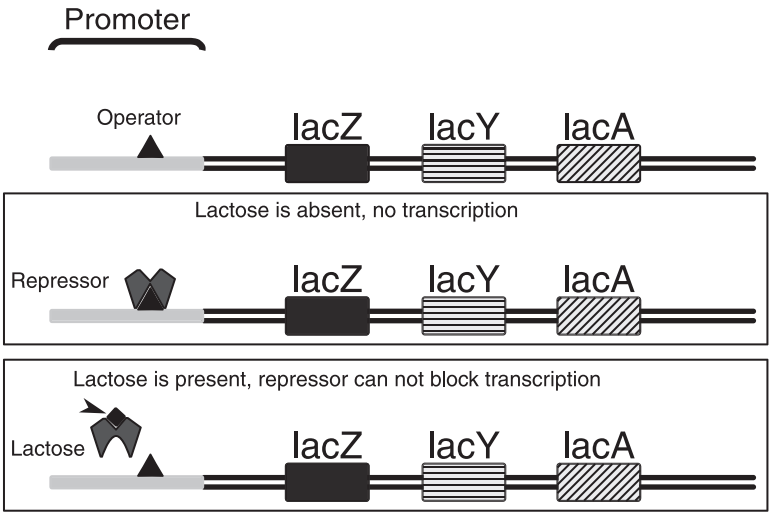

Fig. 2.1. Single RNAP transcribes multiple genes in an inducible lac operon. The repressor protein can bind to the operator region and hinder RNAP binding to the promoter region in the absence of lactose (lac). When lac is present, this small molecule binds to the repressor and dissociates it from operator, allowing RNAP to transcribe the *lacZYA* genes.

do not appear to produce poly-cistronic RNAs. Most of these gene clusters encode enzymes that participate in a common pathway. Plants have the best described examples. These gene clusters encode enzymes for multiple catalytic steps that synthesize compounds defending the host against pathogens (8–11). So far, the mechanisms involved in the coordinate regulation of these complex gene clusters have not been established.

2.2.2. The Organization of the Regulatory Regions of RNAP II-Transcribed Genes

The region of a gene, usually proximal to the transcription start site (TSS), to which RNAP II and associated factors are initially recruited, consists of the core or basal promoter. It assembles as a complex formed by the basal transcription factors (BTF). The precise boundaries of the core promoter must be empirically determined for each gene, but as a rule of thumb, it is considered to comprise ~50 bp to each site of the TSS. Note that the convention is to number the first nucleotide represented in the mRNA as +1, thus this interval can be represented as [−50; +50]. Core promoters contain a number of *cis*-regulatory elements, which include the TATA box and an Initiator (Inr) element (12–14). However, there is no *cis*-regulatory element that is universally present in all core promoters. Even the broadly distributed TATA motif involved in the recruitment of the TATA-binding protein (TBP), a central BTF involved in the assembly of the transcriptional pre-initiation complex (PIC), is present in just ~30% of all eukaryotic promoters (5). BTFs receive signals from other regulatory factors, the TFs, most likely mediated by mediator proteins (15). Textbooks indicate that the regulatory regions of genes are usually located upstream of the TSS. However, notable

recent evidence in large part provided by the Encyclopedia of DNA Elements (ENCODE) consortium suggest that regulatory sequences can be found in 5′- and 3′-untranslated regions (5′- and 3′-UTRs), introns, and even coding regions (16). Thus, it is clear that the definition of what the typical regulatory region of a gene includes needs to be broadened.

2.3. Transcription Factors as Key Regulators of Transcription

TFs are responsible for providing signals necessary for the correct assembly of the PIC and are therefore primarily responsible for controlling the time, amplitude, and duration of gene transcription. About 5–7% of the genome of an eukaryotic organism encodes for TFs (17), which can be grouped into 50–60 distinct groups of families. Some families have dramatically expanded while others might be absent altogether from particular organisms or kingdoms. For example, the MYB family, named after the avian *my*elo*b*lastocys virus from where the first protein harboring this domain was first identified (18, 19), is very large in plants (>180 members in *Arabidopsis*), while animal genomes contain just a handful of genes encoding proteins with this domain.

TFs can activate or repress transcription. If they function as transcriptional activators, they often harbor a transcriptional activation domain (TAD), responsible for interacting with mediator or other BTFs. The structure of TADs is significantly less conserved than the folds that characterize DNA-binding domains, and they are classified into various types (acidic, proline-rich, glutamine-rich, etc.) (20). The structure of the acidic TAD of the herpes simplex virus VP16 was determined and key residues identified for function (21).

2.3.1. De Novo Identification of TFs and Target Sites

Important questions that the biologist often encounters include (1) how to determine if a protein functions as TF or not and (2) what are the direct targets (defined as the genes directly regulated) of a TF.

2.3.1.1. De Novo Identification of TFs

For the identification of TFs from genome sequence or Expressed Sequence Tag (EST) information, specific signatures characteristic of TFs can be followed. As described earlier, TFs can be classified into families based on particular folds of the respective DNA-binding domains. These structures can often share little sequence identity, resulting in the need to investigate relatedness by using profiles that capture weak similarities or even information on neighbor amino acids. The PFAM database (http://pfam.sanger.ac.uk/) is a large collection of protein families, each represented by multiple sequence alignments and Hidden Markov Model (HMM) profiles (22). Within a protein family, multiple alignments reveal similarity in particular regions due to conserved amino acid sequences. These protein fragments correspond to one or more functional regions termed domains. PFAM

contains profiles of domains that carry DNA-binding protein–protein interaction functions and this information is used to predict if an unknown protein corresponds to a TF with a previously described DNA-binding domain or not.

2.3.1.2. Identification of Gene Directly Regulated by a TF

The second problem that the experimentalist often encounters is how to identify the genes that a TF directly regulates. In studying TF function, it is important to establish which DNA sequences they can bind to. This can be accomplished through in vitro protein–DNA interaction techniques that include electrophoretic mobility shift assays (EMSA) in combination with footprinting approaches or by the systematic evolution of ligands by exponential enrichment (SELEX). Using information derived from such experiments to predict TF targets in silico, however, is not trivial, as in vitro DNA-binding specificities established, for example, by SELEX are often not correlated with the sequences that a TF binds in vivo – a good example being provided by E2F factors (23). Thus, the alternative is to experimentally identify the in vivo targets of a TF. The participation of a TF in a given regulatory process can be inferred from mutant analyses or from gene expression profile clusters. However, determining the ultimate function of a TF depends on identifying which genes it can directly activate. Two main approaches are currently available to identify direct targets of TFs: (a) by expressing a fusion of the TF to the hormone-binding domain of the glucocorticoid receptor and identifying the mRNAs induced/repressed in the presence of the GR ligand (dexamethasone, DEX), in the presence of an inhibitor of translation (e.g., cycloheximide, CHX), or (b) by identifying the DNA sequences that a TF binds in vivo, using chromatin immunoprecipitation (ChIP) assays, which can be coupled with next generation sequencing methods (ChIP-Seq) (24) or by using the immunoprecipitated DNA to hybridize a tiling or promoter array representing all the genes in an organism (ChIP-chip) (25, 26). Information on TFs and their binding sequences for a number of species is available at several databases (**Table 2.1**).

2.4. Transcriptional Networks

TFs function in networks, in which a regulatory protein controls the expression of another, which in turn may modulate the expression of other regulatory proteins or control genes encoding structural proteins or enzymes. These hierarchical arrangements allow specific signals to be amplified, providing the information necessary for given sets of genes to be deployed with particular spatial and temporal patterns motifs (27). Gene regulatory networks (GRNs) are formed by motifs, and the dynamic properties of these motifs significantly contribute to the overall behavior of the network (28). MicroRNAs and other small RNAs (briefly described in the next section) are also emerging as key

Table 2.1
Online TF databases for various species. Online resources related to TFs are listed and marked for information provided on TF sequence (TFs), TF binding sequences (TF binding), promoter sequences, and TF binding locations in target gene promoters (Promoters) and regulatory networks. Circuitry of regulatory networks combines individual TF–target gene relationships into single comprehensive view. A list of plant *cis*-element resources and detailed discussion is available in (33)

Name	URL	TFs	TF binding	Promoters	Regulatory networks	Reference
AGRIS	arabidopsis.med.ohio-state.edu	√	√	√	√	(34)
DBD	www.transcriptionfactor.org	√				(35)
GRASSIUS	grassius.org	√	√	√	[a]	(17)
JASPAR	jaspar.cgb.ki.se		√			(36)
PAZAR	www.pazar.info	√	√	√		(37)
PLANTTFDB	planttfdb.cbi.pku.edu.cn	√				(38)
PLNTFDB	plntfdb.bio.uni-potsdam.de	√				(39)
TFCONES	tfcones.fugu-sg.org	√	√			(40)
TFdb	genome.gsc.riken.jp/TFdb	√				(41)
TRANSFAC[b]	www.gene-regulation.com	√	√	√	√	(42)

[a] Planned feature.
[b] Some features are available in commercial package.

components of GRNs [e.g., (29)], often participating in mixed network motifs (27).

2.5. Small RNAs and Gene Expression

One of the most significant discoveries of the past few years is the realization that most of the DNA that lies between genes is not really "junk," but that it participates in the formation and is the subject of regulation of a large number of non-coding RNAs, often groups under the term small RNA (to distinguish them from the longer mRNA, tRNA, or rRNA populations). Small RNAs have indeed been called the "Guardians of the Genome" (30), and one of their main functions appears to be to keep transposons (pieces of DNA that can move around the genome) at bay, preventing major genome damage. Small RNAs can be of different types and usually have lengths 20–30 nucleotides long. They appear to be broadly distributed in all eukaryotes, and even

prokaryotes express small RNAs with unique regulatory activities (31). One class of small RNAs, the microRNAs (miRNAs) participate in the post-transcriptional regulation of mRNA translation and stability. In contrast, small interfering RNAs (siRNAs) control gene expression by specifically targeting particular sequences for silencing in the process of TGS that involves histone modifications and DNA methylation (32).

Acknowledgments

Support in the Grotewold lab for projects involving regulation of gene expression is provided by NRI Grant 2007-35318-17805 from the USDA CSREES, DOE Grant DE-FG02-07ER15881, and NSF grant DBI-0701405. A.Y. is supported by NIH Ruth L. Kirschstein National Research Service Award 5 T32 CA106196-05 from NCI.

References

1. Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005) RNA polymerase IV directs silencing of endogenous DNA. *Science* 308, 118–120.
2. Pikaard, C.S., Haag, J.R., Ream, T., and Wierzbicki, A.T. (2008) Roles of RNA polymerase IV in gene silencing. *Trends Plant Sci* 13, 390–397.
3. Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* 135, 635–648.
4. Wierzbicki, A.T., Ream, T.S., Haag, J.R., and Pikaard, C.S. (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat Genet* 41, 630–634.
5. Grotewold, E., and Springer, N. (2009) Decoding the transcriptional hardwiring of the plant genome. In: *Plant systems biology* (Coruzzi, G., and R.A. Gutierrez, Eds.) pp. 196–228, Wiley-Blackwell, Chichester.
6. Gruber, T.M., and Gross, C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* 57, 441–466.
7. Kazmierczak, M.J., Wiedmann, M., and Boor, K.J. (2005) Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev* 69, 527–543.
8. Field, B., and Osbourn, A.E. (2008) Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science* 320, 543–547.
9. Jonczyk, R., Schmidt, H., Osterrieder, A., Fiesselmann, A., Schullehner, K., Haslbeck, M. et al. (2008) Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. *Plant Physiol* 146, 1053–1063.
10. Osbourn, A.E., Field, B. (2009) Operons. *Cell Mol Life Sci* 66, 3755–3775.
11. Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., and Osbourn, A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc Natl Acad Sci USA* 101, 8233–8238.
12. Gurley, W.B., O'Grady, K., Czarnecka-Verner, E., and Lawit, S.J. (2006) General transcription factors and the core promoter: ancient roots. In: *Regulation of transcription in plants* (Grasser, K., Eds.) pp. 1–27. Blackwell Pub, Oxford.
13. Smale, S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev* 15, 2503–2508.
14. Smale, S.T., and Kadonaga, J.T. (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* 72, 449–479.

15. Gustafsson, C.M., and Samuelsson, T. (2001) Mediator – a universal complex in transcriptional regulation. *Mol Microbiol* 41, 1–8.
16. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
17. Yilmaz, A., Nishiyama, M.Y., Jr., Fuentes, B.G., Souza, G.M., Janies, D., Gray, J. et al. (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol* 149, 171–180.
18. Klempnauer, K.H., Gonda, T.J., and Bishop, J.M. (1982) Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. *Cell* 31, 453–463.
19. Klempnauer, K.H., Ramsay, G., Bishop, J.M., Moscovici, M.G., Moscovici, C., McGrath, J.P. et al. (1983) The product of the retroviral transforming gene v-myb is a truncated version of the protein encoded by the cellular oncogene c-myb. *Cell* 33, 345–355.
20. Roberts, S.G. (2000) Mechanisms of action of transcription activation and repression domains. *Cell Mol Life Sci* 57, 1149–1160.
21. Uesugi, M., Nyanguile, O., Lu, H., Levine, A.J., and Verdine, G.L. (1997) Induced alpha helix in the VP16 activation domain upon binding to a human TAF. *Science* 277, 1310–1313.
22. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26, 320–322.
23. Rabinovich, A., Jin, V.X., Rabinovich, R., Xu, X., and Farnham, P.J. (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res* 18, 1763–1777.
24. Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat Methods* 4, 613–614.
25. Buck, M.J., and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 83, 349–360.
26. Herring, C.D., Raffaelle, M., Allen, T.E., Kanin, E.I., Landick, R., Ansari, A.Z. et al. (2005) Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *J Bacteriol* 187, 6166–6174.
27. Re, A., Cora, D., Taverna, D., and Caselle, M. (2009) Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. *Mol Biosyst* 5, 854–867.
28. Alon, U. (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* 8, 450–461.
29. Card, D.A., Hebbar, P.B., Li, L., Trotter, K.W., Komatsu, Y., Mishina, Y. et al. (2008) Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells. *Mol Cell Biol* 28, 6426–6438.
30. Malone, C.D., and Hannon, G.J. (2009) Small RNAs as guardians of the genome. *Cell* 136, 656–668.
31. Waters, L.S., and Storz, G. (2009) Regulatory RNAs in bacteria. *Cell* 136, 615–628.
32. Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A.J. (2007) Targets of RNA-directed DNA methylation. *Curr Opin Plant Biol* 10, 512–519.
33. Brady, S.M., and Provart, N.J. (2009) Web-queryable large-scale data sets for hypothesis generation in plant biology. *Plant Cell* 21, 1034–1051.
34. Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V., and Grotewold, E. (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* 140, 818–829.
35. Kummerfeld, S.K., and Teichmann, S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res* 34, D74–D81.
36. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32, D91–D94.
37. Portales-Casamar, E., Kirov, S., Lim, J., Lithwick, S., Swanson, M.I., Ticoll, A. et al. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol* 8, R207.
38. Guo, A.Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.H., Liu, X.C. et al. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res* 36, D966–D969.
39. Riano-Pachon, D.M., Ruzicic, S., Dreyer, I., and Mueller-Roeber, B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 8, 42.

40. Lee, A.P., Yang, Y., Brenner, S., and Venkatesh, B. (2007) TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. *BMC Genomics* 8, 441.
41. Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., and Suzuki, H. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun* 322, 787–793.
42. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24, 238–241.

http://www.springer.com/978-1-60761-853-9

Computational Biology of Transcription Factor Binding
Ladunga, I. (Ed.)
2010, XI, 454 p. 102 illus., 9 illus. in color., Hardcover
ISBN: 978-1-60761-853-9
A product of Humana Press