Transcriptional Regulation: a Genomic Overview

Author: José Luis Riechmann
Source: The Arabidopsis Book, 2002(1)
Published By: The American Society of Plant Biologists
URL: https://doi.org/10.1199/tab.0085
Transcriptional Regulation: a Genomic Overview

José Luis Riechmann

Mendel Biotechnology, 21375 Cabot Blvd., Hayward, CA 94545, USA
e-mail: jriechmann@mendelbio.com
and
California Institute of Technology, Division of Biology 156-29, Pasadena, CA 91125
e-mail: jriechma@caltech.edu

Abstract

The availability of the Arabidopsis thaliana genome sequence allows a comprehensive analysis of transcriptional regulation in plants using novel genomic approaches and methodologies. Such a genomic view of transcription first necessitates the compilation of lists of elements. Transcription factors are the most numerous of the different types of proteins involved in transcription in eukaryotes, and the Arabidopsis genome codes for more than 1,500 of them, or approximately 6% of its total number of genes. A genome-wide comparison of transcription factors across the three eukaryotic kingdoms reveals the evolutionary generation of diversity in the components of the regulatory machinery of transcription. However, as illustrated by Arabidopsis, transcription in plants follows similar basic principles and logic to those in animals and fungi. A global view and understanding of transcription at a cellular and organismal level requires the characterization of the Arabidopsis transcriptome and promoterome, as well as of the interactome, the localizome, and the phenome of the proteins involved in transcription.

Introduction.

Many of the biological processes in a plant are regulated at the level of transcription. Changes in gene expression have been shown to underlie the response to environmental cues and stresses (such as light, temperature, and nutrient availability), the defense response against pathogens, the regulation of metabolic pathways, the regulation of photosynthesis, or the establishment of symbiotic relationships, to name a few. In plants, as well as in animals, development is based on the cellular capacity for differential gene expression (reviewed in: Scott, 2000; Benfey and Weigel, 2001). Accordingly, many of the genes identified in screens for Arabidopsis mutants with altered, for example, flower or root development have been found to encode transcription factors. Alterations in gene expression are also emerging as a major source of the diversity and change that underlie the morphological evolution of eukaryotic organisms (Doebley and Lukens, 1998; Cubas et al., 1999b; Carroll, 2000; Tautz, 2000). In particular, morphological changes that occurred during plant domestication and crop improvement in agriculture have been associated with mutations in transcription factors (Peng et al., 1999), alterations in their expression (Doebley et al., 1997; Wang et al., 1999b), or changes in the expression of other types of regulatory proteins (Frary et al., 2000). Related transcription factors, such as the Arabidopsis MYB proteins WEREwOLF (WER) and GLABROUS1 (GL1), have been shown to be functionally equivalent, and owe their particular roles in plant development to differences in their expression patterns (Lee and Schiefelbein, 2001).

The availability of the Arabidopsis genome sequence (Lin et al., 1999; Mayer et al., 1999; Arabidopsis Genome Initiative, 2000; Salanoubat et al., 2000; Tabata et al., 2000; Theologis et al., 2000) allows a global, or genomic, analysis of transcriptional regulation in plants. Whereas the mechanisms of transcription are largely common across eukaryotes, their components vary among kingdoms. The complement of genes coding for transcriptional regulators in Arabidopsis has been described (Arabidopsis Genome Initiative, 2000; Riechmann et al., 2000). Their systematic functional characterization can be pursued with a variety of reverse genetic methods (Riechmann and Ratcliffe, 2000).
In addition, gene expression profiling technologies, such as DNA microarrays, allow monitoring transcription factor activity at a genome-wide level. These studies should eventually lead to an understanding of the interplay of the transcription factors with the genome whose expression they control.

This chapter intends to provide a genomic perspective on transcriptional regulation in Arabidopsis. The first section briefly reviews the different types of proteins directly involved in transcription in eukaryotes, and our current understanding on how they function. The following sections consist of a description of the Arabidopsis complement of genes and proteins involved in transcriptional control, in particular sequence-specific DNA-binding transcription factors and chromatin-related proteins. Transcriptional regulators often act in a combinatorial fashion, and this mode of action is reviewed in the context of Arabidopsis promoters and cis-regulatory sequences, and of protein-protein interactions. Finally, genome-wide functional analyses of transcription factors, the characterization of the Arabidopsis promorome, and of the transcriptome by gene expression profiling experiments, are considered. The availability of the genome sequence of different prokaryotic and eukaryotic organisms has provided for new ways of searching for unity and diversity among biological systems, and given birth to the field of comparative genomics. Although the subject of this book is Arabidopsis, reference is made in this chapter to other eukaryotic organisms, in order to situate the Arabidopsis genome information in a broader biological context.

2. Transcription machinery: concepts, components, and mechanisms

In eukaryotic organisms, regulation of gene expression proceeds through mechanisms that are fundamentally different from those in prokaryotes, which explains both the large number and diversity of proteins that are involved in the process, as well as how it can be tightly regulated to facilitate the diversification in expression patterns that is required for biological complexity (Struhl, 1999). In a prokaryote such as E. coli, the ground state for transcription is non-restrictive, that is, the RNA polymerase complex is not limited in its ability to gain access to the DNA and initiate RNA synthesis (Struhl, 1999). Negative regulation is rare, and exerted by sequence-specific repressors. Furthermore, it has been estimated that the global structure of the E. coli gene regulatory network possesses low complexity. On average, a transcription factor would regulate three genes, and an E. coli gene would be under the direct control of two transcription factors (Thieffry et al., 1998). There is a prominence of promoters controlled by a single regulator, and whereas many of the regulators regulate themselves (usually through auto-inhibitions), very rarely do they regulate other transcription factors (Thieffry et al., 1998).

In contrast, the ground state for transcription in eukaryotes is restrictive, as a result of the packing of the DNA into chromatin, which blocks the recognition of the core promoters by the basic transcription machinery (Kornberg, 1999; Struhl, 1999). The effects of chromatin structure on promoter accessibility makes chromatin modifying activities necessary for eukaryotic transcription, and has important implications for the way transcription factors act. In addition to the components of the basic transcription machinery and to scores of sequence-specific DNA-binding transcription factors, eukaryotic genomes contain a variety of genes that code for chromatin-related proteins. Furthermore, transcriptional regulators in eukaryotes operate following a combinatorial logic (an efficient way of increasing the number and diversity of gene regulatory activities), and the complexity of the regulatory networks can be great.

Prokaryotic sequence-specific DNA-binding transcription factors often recognize binding sites longer than 12 base-pairs (bp) (see RegulonDB, http://www.cifn.unam.mx/regulondb/ and DPInteract, http://arep.med.harvard.edu/dpinteract) (Robison et al., 1998; Salgado et al., 2001), whereas binding sites for eukaryotic transcription factors are usually shorter, 5 to 10 bp long. A combinatorial mechanism composed of factors that recognize short sequences is probably a more economical way (requires a reduced number of factors) of selectively regulating the expression of tens of thousands of genes, than a mechanism based upon factors that are each dedicated to control a small number of genes and operate through longer target sites. Thus, the DNA binding characteristics of eukaryotic transcription factors, and the mechanisms of transcription themselves, might be operationally and evolutionarily related to features of eukaryotic genomes such as the vast increase in genome size and in the number of genes to be regulated.

Briefly, the proteins involved in transcription in eukaryotes can be classified into four different functional groups: (1) the basic transcription apparatus and intrinsic associated factors (also known as general transcription factors, or GTFs); (2) large multi-subunit coactivators and other cofactors; (3) sequence-specific DNA-binding transcription factors; and (4) chromatin-related proteins. In contrast to the components of the basal transcription machinery, which in general are highly conserved, coregulators and transcription factors have diverged largely among eukaryotes. The roles that the proteins in these four classes play can be summarized as follows (after Lee and Young, 2000;

(1) The basic transcription apparatus, and intrinsic associated factors. In eukaryotic organisms, there are three different RNA polymerases, which are responsible for the synthesis of rRNA (Pol I), mRNA (Pol II), and tRNA, SS rRNA, and other small RNA molecules (Pol III). The focus of this chapter is the transcription of protein-encoding genes, which is carried out by Pol II exclusively. Pol II is a multi-subunit enzyme (Cramer et al., 2001) that requires accessory factors to recognize promoter sequences and accurately initiate transcription. These general transcription factors (GTFs) include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIH. GTFs carry out a variety of different functions, from positioning the polymerase on the promoter (TFIIB) to unwinding its DNA (TFIIF). TFIID is a multi-subunit complex that is generally responsible for promoter recognition. It contains the TATA-box binding protein (TBP) and several TBP-associated factors (TAFs) (reviewed in Green, 2000). The TAF subunits of TFIID are critical for the responsiveness of the basic apparatus to transcriptional activators. However, individual TAFs are not essential for transcription of all genes in a genome. TAFs contribute to the specificity and variety of transcriptional responses: distinct TAFs can be targeted by different classes of activators, and individual TAFs can function as promoter selectivity factors. Furthermore, some TAFs can form part of other multi-subunit regulatory complexes, in addition to TFIID, such as the histone acetylation SAGA complex; and whereas most of the TAFs are ubiquitously expressed, some are expressed in a tissue or cell-type specific manner, which can lead to the formation of different TAF-containing complexes (for a review on gene-selective roles of GTFs and TAFs, see Veenstra and Wolfe, 2001).

(2) Large multi-subunit coactivators, and cofactors that bind sequence specific transcription factors. This heterogeneous class of regulatory proteins includes cofactors that interact with sequence-specific transcription factors and modulate their DNA binding or interaction with the core machinery, as well as large multi-subunit coactivators such as the Mediator complex, initially identified in yeast. Multi-subunit coactivators interact with Pol II and/or with multiple types of activators, serving as a modular adapter to regulate transcription initiation (Hampsey and Reinberg, 1999). The Mediator (or Mediator-like) complex is found in organisms from yeast to humans, but its number of subunits vary, and the complex from one organism might contain subunits that have no orthologs in another (Malik and Roeder, 2000; Rachez and Freedman, 2001).

(3) Sequence-specific DNA-binding transcription factors (activators and repressors). These are transcription factors of the classic type: usually defined as proteins that show sequence-specific DNA binding and are capable of activating and/or repressing transcription. They are responsible for the selectivity in gene regulation, and are often themselves expressed in a tissue, cell-type, temporal, or stimulus-dependent specific manner. Transcription factors are modular proteins, with distinct and functionally separable domains, such as DNA-binding and activation domains. Most known transcription factors can be grouped into families according to their DNA binding domain (Luscombe et al., 2000). Transcription factors can interact directly with different components of the general machinery and with coactivators, affecting complex formation. They can also interact with chromatin remodeling complexes.

(4) Chromatin-related proteins. This group includes factors that covalently modify histones (such as histone acetylases and deacetylases), and remodeling complexes that hydrolize ATP for reorganizing chromatin structure (such as the SWI/SNF and ISWI complexes). Histone acetylation is generally a characteristic of transcribed chromatin, whereas deacetylation is associated with repression. Accordingly, histone acetyltransferase activities are found in coactivators, and deacetylase activities in corepressors. Chromatin proteins usually form part of multi-subunit complexes.

Using the regulation of the HO endonuclease gene in yeast (Saccharomyces cerevisiae) as a paradigm, the steps leading to transcriptional activation can be summarized as follows (Cosma et al., 1999; Cosma et al., 2001). Upstream sequences are recognized by a transcription (enhancer-binding) factor, with accessibility to its targets despite the packing of the DNA into chromatin fibers. This transcription factor recruits the SWI/SNF complex, which then recruits SAGA, and results in the remodeling of chromatin and localized histone acetylation, which facilitates the access of additional transcription factors to cis-regulatory sequences. The secondary activators direct gene transcription through multiple interactions with cofactors and the core machinery, recruiting the RNA polymerase complex to the transcription initiation site. The specific order in which the different chromatin-modifying complexes are recruited can vary among promoters and organisms, but the dual role of activators, first enlisting chromatin modifying activities and then inducing localization of the basal transcription apparatus, appears to be widespread in eukaryotes, including plants (see below, and: Agalioti et al., 2000; Brown et al., 2001; Merika and Thanos, 2001).

In many instances, the correct functioning of a gene requires the termination of the activation of its transcription to be as rapid or precise as its initial triggering. Termination
of activation can be accomplished by several mechanisms, among them the targeted destruction of transcription factors after their interaction with the basal transcription machinery. Phosphorylation of a transcription factor molecule by kinases that form part of the Pol II holoenzyme (such as Srb10 or TFIIH) would mark it for ubiquitin-mediated destruction, effectively preventing it from engaging into another Pol II initiation event, and freeing the promoter sequence to interact with another transcription factor molecule (reviewed in Tansey, 2001).

In addition to the mechanisms of transcriptional control that the classes of proteins described in this section mediate, there are at least two other possible levels of regulation of gene expression in eukaryotes: DNA methylation and nuclear organization. DNA methylation is associated with suppressed gene expression, and is reviewed in other chapters of this book (see also: Finnegan et al., 2000; Habu et al., 2001). Nuclear organization could provide for a higher level of regulation of gene expression, where different transcriptional functions might be segregated into distinct compartments (for models and reviews, see: Francastel et al., 2000; Lemon and Tjian, 2000; Cremer and Cremer, 2001; Misteli, 2001).

Of all the groups of proteins involved in transcription, the most numerous one is that of sequence-specific DNA-binding transcription factors. They are the principal factors upon which the mechanisms for selectivity of gene activation are built, and the basic (although not the only) protein components of the combinatorial logic of transcription.

3. Transcription factor gene content of the Arabidopsis genome

The analysis of the Arabidopsis genome sequence indicates that it codes for at least 1,572 transcription factors, which account for ~6% of its estimated ~26,000 genes (Arabidopsis Genome Initiative, 2000; Riechmann et al., 2000) (Table 1). This observation, however, represents an underestimate of the total number of transcription factors, given that, at present, approximately 40% of the proteins predicted from the genome sequence cannot be assigned to functional categories on the basis of sequence similarity to proteins of known biochemical function (Lin et al., 1999; Mayer et al., 1999; Arabidopsis Genome Initiative, 2000; Salanoubat et al., 2000; Tabata et al., 2000; Theologis et al., 2000). Some of those uncharacterized proteins are expected to be transcriptional regulators and, in fact, novel classes of transcription factors are still being discovered (for example: Boggon et al., 1999; Schauser et al., 1999; Kawaoka et al., 2000; Nagano et al., 2001; Windhövel et al., 2001). Therefore, the total number of transcription factor genes present in Arabidopsis (as well as, for the same reasons, in any other of the sequenced eukaryotic genomes) will be uncertain for some time.

A question pertaining to genome-wide surveys is whether all the proteins identified by sequence similarity searches do indeed belong to the functional groups into which they are being catalogued. In the case of transcription factors, the answer depends on the particular gene family that is considered. If the conserved DNA-binding domain that defines a gene family is poor in sequence information (for example, some zinc-coordinating motifs), the ratio of false positives in the searches can be relatively high (although it can often be reduced by additional sequence comparison strategies that are beyond the scope of this chapter, see: Riechmann et al., 2000). On the other hand, many families are defined by long DNA-binding domains (50 to 70 amino acids) with multiple residues being highly conserved (that is, the domains are rich in sequence information). The three-dimensional structure of these domains might have been solved, and revealed the contacts between some of the conserved residues and the DNA. In cases like these, such as for example the homeobox and the AP2/ERF (APETALA2/ethylene response factor) families, it is reasonable to expect all the members of the gene family to be transcription factors (activators or repressors) (the AP2/ERF family was initially referred to as AP2/EREBP for AP2/ethylene responsive element binding protein). However, there are cases in which a family of bona fide transcription factors might also contain members that have additional functions (for reviews on multifunctional transcription factors: Ladomery, 1997; Wilkinson and Shyu, 2001). For example, the Drosophila homeodomain protein Bicoid directs anterior embryo development both by regulating transcription and by interacting with Caudal mRNA and inhibiting its translation, thus restricting Caudal (which is another homeodomain protein) accumulation to the posterior part of the embryo through posttranscriptional control. Both DNA- and RNA-binding are specified by the Bicoid homeodomain, but by distinct subregions or residues in it (Niessing et al., 2000). The Arabidopsis MYB-related protein AtCDC5 is known to be homologous to the S. cerevisiae CEF1 and S. pombe Cdc5 proteins (Hirayama and Shinozaki, 1996; Ohi et al., 1998). Cdc5 proteins are essential for the G2/M progression, but their molecular functions are not completely understood, as they are required for pre-mRNA splicing and associate with core components of the splicing machinery, but also show sequence-specific binding to double stranded DNA and transactivation potential (Burns et al., 1999; Lei et al., 2000). It is thus possible that Cdc5 proteins are another example of factors with several distinct molecular
### Table 1. Arabidopsis transcriptional regulators

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Predicted number of genes</th>
<th>Genetically characterized Arabidopsis factors</th>
<th>Gene family functions</th>
<th>Family signature sequence (InterPro or GenBank accession number)</th>
<th>Gene family references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP2/ERF*</td>
<td>14</td>
<td>ANT, AP2</td>
<td>Flower and seed development; cell identity and proliferation</td>
<td>IPR001471 (Riechmann and Meyerowitz, 1999)</td>
<td></td>
</tr>
<tr>
<td>AP2 subfamily</td>
<td>14</td>
<td>ANT, AP2</td>
<td>Flower and seed development; cell identity and proliferation</td>
<td>IPR001471 (Riechmann and Meyerowitz, 1999)</td>
<td></td>
</tr>
<tr>
<td>ERF subfamily</td>
<td>124</td>
<td>ABH1, CBF1/DREB1B, CBF3/DREB1A, ERF1</td>
<td>Control of metabolic pathways; defense response; abiotic stress responses; hormone responses (ABA, ethylene); sugar response</td>
<td>IPR001092 (Kagaya et al., 1999)</td>
<td></td>
</tr>
<tr>
<td>RAV-like</td>
<td>6</td>
<td>GL3, HFR1/REPI/RSPI, PF3, rZuBP1, SPT, TT9</td>
<td>Control of secondary metabolism; cell-type specification (trichome and root development); phytochrome-mediated responses; carpel development; abiotic stress responses</td>
<td>IPR001092 (Kagaya et al., 1999)</td>
<td></td>
</tr>
<tr>
<td>bHLH</td>
<td>139</td>
<td>AS1, AMYB2, ATMYB4, ATR1, GL1, LAI1, PAPI, PAP2, TT2, WER</td>
<td>Stem cell function and leaf patterning; cell-type specification (trichome and root development); control of secondary metabolism; defense response; abiotic stress responses; hormone responses (ABA); G2/M-phase-specific transcription; phytochrome A signaling</td>
<td>IPR001005</td>
<td>(Martin and Paz-Ares, 1997; Stracke et al., 2001)</td>
</tr>
<tr>
<td>MYB superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYB-(R1)R2R3</td>
<td>136</td>
<td>AS1, AMYB2, ATMYB4, ATR1, GL1, LAI1, PAPI, PAP2, TT2, WER</td>
<td>Stem cell function and leaf patterning; cell-type specification (trichome and root development); control of secondary metabolism; defense response; abiotic stress responses; hormone responses (ABA); G2/M-phase-specific transcription; phytochrome A signaling</td>
<td>IPR001005</td>
<td>(Martin and Paz-Ares, 1997; Stracke et al., 2001)</td>
</tr>
<tr>
<td>MYB-related</td>
<td>54</td>
<td>CCA1, CPC, LHY</td>
<td>Circadian clock; cell-type specification (root development)</td>
<td>IPR000818</td>
<td>(Takatsuji, 1999)</td>
</tr>
<tr>
<td>C2H2 (Zn)</td>
<td>112</td>
<td>EMF2, FIS2, RHL41, SC, SUP, VRN2</td>
<td>Flower, ethylene, shoot and seed development; flowering time; root nodule development; abiotic stress responses; light acclimatization response</td>
<td>IPR000822</td>
<td>(Takatsuji, 1999)</td>
</tr>
<tr>
<td>NAC</td>
<td>109</td>
<td>CUC1, CUC2, NAC1, NAP</td>
<td>Meristem and organ development; organ separation; auxin response; virus resistance</td>
<td>BAB10725 (Xie et al., 2000)</td>
<td></td>
</tr>
<tr>
<td>HB</td>
<td>90</td>
<td>ANL2, ATIB-2, ATIB-8, ATIB-13, BEL1, GL2, KNAT1, PHB, PHV, REVIF1, STM, WUS</td>
<td>Development; radial patterning (adaxial identity), stem cell identity and maintenance, shoot apical meristem function, vascular meristem differentiation, and ovule, leaf, internode, and root development; shade avoidance response; anthocyanin distribution; auxin signalling; cell death</td>
<td>IPR001356 (Chua et al., 1995)</td>
<td></td>
</tr>
<tr>
<td>MADS</td>
<td>82</td>
<td>AG, ANR1, AP1, AP3, CAL, PLC, PUL, MALF1, PL, SEP1, SEP2, SEP3, SHP1, SHP2, SDC1, SVP</td>
<td>Meristem identity; flower, fruit, and root development; flower and fruit abscission and dehiscence; flowering time</td>
<td>IPR002100 (Riechmann and Meyerowitz, 1997b; Ng and Yanofsky, 2001)</td>
<td></td>
</tr>
<tr>
<td>bZIP</td>
<td>77</td>
<td>AB15, HYS, PAN</td>
<td>Semi-storage gene expression; flower and leaf development; photomorphogenic development; defense response; hormone responses (ABA, gibberellin); control of gibberellin biosynthesis</td>
<td>IPR001871 (Foster et al., 1994)</td>
<td></td>
</tr>
<tr>
<td>WRKY (Zn)</td>
<td>72</td>
<td>_</td>
<td>Defense response</td>
<td>S72443</td>
<td>(Eulgem et al., 2000; Yu et al., 2001)</td>
</tr>
<tr>
<td>GARP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2-like</td>
<td>44</td>
<td>KAn1, KAn2, PHR1</td>
<td>Photosynthetic development; phosphorus metabolism; abaxial cell identity</td>
<td>AAD55941 (Estoh et al., 2001; Rossi et al., 2001)</td>
<td></td>
</tr>
<tr>
<td>ARR-B class</td>
<td>12</td>
<td>ARR1, ARR2</td>
<td>Cytokinin signal transduction</td>
<td>BAA74528 (D'Agostino and Kasber, 1995; Hwang and Shean, 2001)</td>
<td></td>
</tr>
</tbody>
</table>

continued on following page
continued from previous page

<table>
<thead>
<tr>
<th>C2C2 (Zn)</th>
<th></th>
<th></th>
<th>Seed germination; endosperm-specific gene expression; carbon metabolism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dof</td>
<td>37</td>
<td>DAG1</td>
<td></td>
</tr>
<tr>
<td>CO-like</td>
<td>33</td>
<td>CO1</td>
<td>Flowering time; circadian rhythm</td>
</tr>
<tr>
<td>GATA</td>
<td>28</td>
<td>YABBY</td>
<td>Abaxial cell fate</td>
</tr>
<tr>
<td>CCAAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAP2 type</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAP3 type</td>
<td>11</td>
<td>LECI</td>
<td>Embryo development</td>
</tr>
<tr>
<td>HAP5 type</td>
<td>13</td>
<td>QRO</td>
<td></td>
</tr>
<tr>
<td>DrR</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRAS</td>
<td>32</td>
<td>GAI, PAT1, RGA1, SCR, SHR,</td>
<td>Root and shoot development (radial patterning); hormone responses (gibberellin; phytochrome-mediated responses</td>
</tr>
<tr>
<td>Trichix</td>
<td>28</td>
<td></td>
<td>Light-regulated gene expression</td>
</tr>
<tr>
<td>HSF</td>
<td>26</td>
<td>HSF1, HSF3</td>
<td>Heat shock response</td>
</tr>
<tr>
<td>TCP</td>
<td>25</td>
<td>PTF1</td>
<td>Flower development (asymmetry); cell proliferation; light-responsive chloroplast gene expression</td>
</tr>
<tr>
<td>ARF</td>
<td>23</td>
<td>ETT, MP, NPH4</td>
<td>Auxin responses: flower, embryo, and vascular tissue development</td>
</tr>
<tr>
<td>C3H-type 1 (Zn)</td>
<td>17</td>
<td>PEH1</td>
<td>Embryo development</td>
</tr>
<tr>
<td>GRF/ENBP (C3H-type 2: Zn)</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLP</td>
<td>16</td>
<td>SPL3</td>
<td>Flowering time, leaf organogenesis</td>
</tr>
<tr>
<td>Nin-like</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABI3-VP1</td>
<td>14</td>
<td>ABI3, FUS3, LEC2</td>
<td>Embryo and seed development and germination; hormone responses (ABA); plant differentiation</td>
</tr>
<tr>
<td>ZI-HE (Zn)</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TULP</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRS (Zn)</td>
<td>9</td>
<td>SHI</td>
<td>Hormone responses (gibberelin)</td>
</tr>
<tr>
<td>E2F/DP</td>
<td>8</td>
<td></td>
<td>Cell cycle</td>
</tr>
<tr>
<td>CPP (Zn)</td>
<td>8</td>
<td>TSO1</td>
<td>Cell division; leghemoglobin gene expression</td>
</tr>
<tr>
<td>Affin-like</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EII</td>
<td>6</td>
<td>EIN3, EIL1</td>
<td>Hormone responses (ethylene)</td>
</tr>
<tr>
<td>NtLIM1-like (Zn)</td>
<td>6</td>
<td></td>
<td>Lignin biosynthesis</td>
</tr>
<tr>
<td>STHA-like</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBF-2-like</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRT-like (Zn)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LFY</td>
<td>1</td>
<td>LFY</td>
<td>Meristem identity; flower development</td>
</tr>
<tr>
<td>NZZ</td>
<td>1</td>
<td>NZZ</td>
<td>Sporogenesis</td>
</tr>
<tr>
<td>SAP</td>
<td>1</td>
<td>SAP</td>
<td>Flower development</td>
</tr>
</tbody>
</table>

Other transcriptional regulators


continued on following page
functions. Two other members of the Arabidopsis MYB-related family, AtTRP1 and AtTBP1, have been identified as telomere-binding proteins (Chen et al., 2001a; Hwang et al., 2001), although a possible role in transcription cannot be ruled out because the 5' regions of some Arabidopsis genes contain two or more non-contiguous telomeric repeats (Regad et al., 1994). These examples illustrate the limitations of using sequence similarity to assign potential roles to proteins that are otherwise uncharacterized, and also how the determination of their molecular functions can be elusive. Similar cases might occur within some of the zinc-coordinating protein families, since the same or related motifs can be involved in DNA- and RNA-binding, and may be present in proteins with functions involving nucleic acid binding but not necessarily transcriptional regulation. For example, vertebrate Y-box proteins contain a zinc-coordinating cold-shock domain, and are often dual DNA- and RNA-binding proteins that can regulate transcription and/or translation (reviewed in: Matsumoto and Wolfe, 1998; Sommerville, 1999).

With these caveats in mind, the Arabidopsis complement of transcription factors has been the subject of an extensive genome-wide descriptive analysis, which also included a comparison with those of Drosophila melanogaster, Caenorhabditis elegans, and Saccharomyces cerevisiae (Riechmann et al., 2000). The main conclusions of that study are summarized here.

The 1,572 transcriptional regulator genes identified in the Arabidopsis genome are classified into more than 45 different gene families (Table 1; Figure 1), all of which are scattered throughout the genome. In addition, there are a few single-copy or "orphan" genes, such as LEAFY (LFY). Transcriptional regulators represent approximately 4.6, 3.5, and 3.5% of the genes in Drosophila, C. elegans, and yeast, respectively (Riechmann et al., 2000). Thus, the Arabidopsis content of transcription factors is 1.3 times that of Drosophila, and 1.7 times that of C. elegans and yeast (Riechmann et al., 2000). The large number and diversity of transcription factors in Drosophila were proposed to be related to its substantial regulatory complexity (Adams et al., 2000). Applying the same logic to Arabidopsis suggests that the regulation of transcription in plants is as complex as that in Drosophila. Furthermore, if the estimated total number of genes in humans, 30,000-40,000, and of transcription factor genes, 1,850-2,000, are correct (International Human Genome Sequencing Consortium, 2001; Tupler et al., 2001; Venter et al., 2001), then the transcription factor gene content of Arabidopsis and of H. sapiens are similar (~6% in Arabidopsis, versus 4.8-6.6% in humans). It should be noted, however, that there is a substantial degree of uncertainty about these estimates of gene numbers in humans (see, for example: Hogenesch et al., 2001; Wright et al., 2001).

Transcription factors, the networks that they form, and the genes that they regulate, have been proposed as a possible objective measurement (connectivity of gene-regulation networks) of the biological complexity of an organism (Szathmáry et al., 2001). From that point of view,
the large number of transcription factors in Arabidopsis was interpreted in the context of the complexity of secondary metabolism in plants (Szathmáry et al., 2001), but it might also be related to the complex interactions between plants and the environment (both biotic and abiotic) as well as to the degree of duplications in the genome (see below, and Arabidopsis Genome Initiative, 2000).

The extent to which the Arabidopsis complement of transcription factors represents that of other plants is still an open question. Since the evolutionary divergence between the monocot and dicot lineages is a relatively recent event, which perhaps occurred ~200 million years ago (Savard et al., 1994), it could be expected that the overall composition of monocot and dicot transcription factor complements would be similar. In fact, the largest transcription factor families in Arabidopsis also appear to

Figure 1. The Arabidopsis complement of transcription factors. Gene families are represented by circles, whose size is proportional to the number of members in the family. Domains that have been shuffled, and that therefore “connect” different groups of transcription factors are indicated with rectangles, whose size is proportional to the length of the domain. DNA binding domains are colored; other domains (usually protein-protein interaction domains) are shown with hatched patterns. Dashed lines indicate that a given domain is a characteristic of the family or subfamily to which it is connected. Gene names are written in italics. Whereas many of the indicated domain shuffling events are specific to plants, others likely predate the appearance of the three distinct eukaryotic kingdoms (for details, see Riechmann et al., 2000). This figure is an expanded and updated version of Figure 1 in Riechmann et al. (2000).
be the most prevalent ones in monocotyledonous plants. For example, the phylogenetic comparison of a subset of maize and Arabidopsis MYB-(R1)R2R3 sequences shows that the amplification of the gene family occurred prior to the separation of monocots and dicots (Rabinowicz et al., 1999). In addition, within phylogenetically well-studied families of transcription factors, such as the MADS-box family, many examples of orthology can be identified between Arabidopsis genes and those from rice or maize, and even from gymnosperms (reviewed in: Theissen et al., 2000; Ng and Yanofsky, 2001) (see also http://www.mpizo-koeln.mp.de/mads). Putative orthologous MADS-box genes have regularly maintained conserved functions, even after substantial sequence divergence (Theissen et al., 2000). However, it is also apparent that diversity in transcriptional regulators will be found within the plant kingdom, and between monocots and dicots. Many MADS-box gene duplication and diversification events occurred after separation of the moss and fern lineages from the lineage that originated the flowering plants (Münster et al., 1997; Hasebe et al., 1998; Krogan and Ashton, 2000; Svensson et al., 2000), and at least two clades of MADS-box genes appear to have been amplified in the phylogenetic lineage that led to grasses with respect to Arabidopsis (Theissen et al., 2000). Similarly, whereas most of the amplification of the MYB-(R1)R2R3 gene family occurred prior to the separation between monocots and dicots, several subgroups in maize appear to have originated recently or undergone duplication (some of these duplications are likely to be associated with the allohexaploid origin of the maize genome, but others do not reflect it: Rabinowicz et al., 1999). These recent expansions could have allowed a functional diversification that might not be present in Arabidopsis.

An issue that impinges on the question of the similarity of the Arabidopsis complement of transcription factors with that of other plants is the degree of completeness of the current characterization (i.e., sequence determination and analysis) of the Arabidopsis genome, in particular if that question is to be addressed on a gene-by-gene basis. TRM1 is a maize C2H2 zinc finger transcription factor involved in the repression of rbcS gene expression in mesophyll cells that is related to the mammalian transcription activator-repressor YY1 (Xu et al., 2001). A BLAST search of the higher plant DNA sequences available in GenBank (July 2001) identifies homologous genes in other monocots (Triticum aestivum) as well as in dicotyledonous plants (Nicotiana tabacum, Solanum tuberosum), but not in Arabidopsis. It is possible that an Arabidopsis TRM1 homolog resides in one of the still unsequenced segments of the genome (see http://www.arabidopsis.org). Similarly, there are a few Arabidopsis transcription factor genes represented by ESTs or BAC-end sequences that still cannot be identified in the genome sequence. The limitations of the current sequencing technologies make it impractical or impossible to determine the sequence of eukaryotic genomes to absolute completeness. Thus, a failure to identify a particular gene in the genome sequence of an organism should not be taken as a definitive proof of the absence of that gene. In addition, gene sequences might diverge more than expected, which might result in the identification of homologous genes requiring more sophisticated sequence analysis than a standard BLAST search. For example, a homolog of the mammalian tumor suppressor gene p53 can be identified in the sequence of the C. elegans genome, despite initial reports that no p53-like gene was present in that organism (Derry et al., 2001; Schumacher et al., 2001).

The genome-wide comparison of transcription factors among eukaryotic organisms (Arabidopsis, Drosophila, C. elegans, and S. cerevisiae, encompassing the plant, animal, and fungal kingdoms) reveals the evolutionary generation of diversity in the regulation of transcription (Riechmann et al., 2000). Each of these eukaryotic kingdoms has its own set of particular transcription factor families and genes. Members of kingdom-specific families represent 45% of the Arabidopsis complement of transcriptional regulators, whereas those of families that are present in all four organisms account for 53% (Figure 2). In each organism, a minority (2-5%) of its transcription factors belong to families that are present in two of the three kingdoms: in animals and yeast (SOX/TCF, Fork head, and RFX1-like transcription factors) or in plants and animals (TULP, CPP, and E2F/DP families) (Figure 2) (Riechmann et al., 2000). This distribution of genes and gene families reflects the evolutionary history of eukaryotes. According to molecular phylogenetic analyses, plants, animals and fungi all diverged from a common ancestor during a short period of time, ~1.5 billion years ago (Wang et al., 1999a; Philippe et al., 2000; Nei et al., 2001). Thus, most of the transcription factor families are either shared by the three kingdoms (those that were present in the common ancestor), or specific to each one (those families that arose independently following divergence).

Many of the Arabidopsis transcription factor gene families are large (Table 1). However, none has been so disproportionately amplified as the nuclear hormone receptors in C. elegans (~38% of its transcription factors), the C2H2 zinc finger proteins in Drosophila (~46%), or the C6 and C2H2 families in yeast (~25% each one) (Figure 3) (Riechmann et al., 2000). The three largest families of transcription factors in Arabidopsis, AP2/ERF, bHLH (basic-region helix-loop-helix), and MYB-(R1)R2R3, each represent only ~9% of the total, and there are several other families with comparable numbers of genes (Figure 3) (Riechmann et al., 2000). The two transcription factor
families that have been more substantially amplified in plants, as compared to animals and yeast, are the MYB and the MADS families. Another difference between the Arabidopsis complement of transcription factors and those of the other eukaryotes is that less than 25% of it consists of zinc coordinating proteins, whereas zinc coordinating transcription factors represent ~51% of the total in Drosophila, ~64% in C. elegans, and ~56% in yeast (Riechmann et al., 2000).

Figure 2. Distribution of transcriptional regulators in eukaryotic organisms (A. thaliana, D. melanogaster, C. elegans, and S. cerevisiae). Transcriptional regulators are kingdom-specific, common to plants, animals, and fungi, or present in only two of the three kingdoms. Members of kingdom-specific families represent only 14% of the total in Drosophila because of its extensive use of the C2H2 zinc finger proteins. The data represented in this figure are from Riechmann et al. (2000).
The Arabidopsis transcription factors that belong to families that are common to all eukaryotes do not share significant similarity with those from the other kingdoms, except in the conserved DNA binding domains that define the respective families (Riechmann et al., 2000). Furthermore, diversity in protein sequence and structure is increased by domain shuffling (Figure 1). Shuffling of some of the DNA-binding domains that are present in all eukaryotes has generated novel transcription factors with plant-specific combinations of modules, as for example in the homeodomain, MADS, and ARID protein families (Figures 1 and 4) (Riechmann et al., 2000).

The Arabidopsis genome contains many tandem gene duplications and large-scale duplications on different chromosomes (Arabidopsis Genome Initiative, 2000; Blanc et al., 2000; Vision et al., 2000). Whereas some of these duplications have been followed by rearrangements and divergent evolution, up to 40 to 60% of the Arabidopsis genes might comprise pairs of highly related sequences (the percentage depending on the parameters used in the analyses) (Arabidopsis Genome Initiative, 2000; Blanc et al., 2000). Transcription factor genes follow these general observations. A comparison of the transcription factor complement to itself (all-against-all) revealed that, on
average, closely related genes account for ~45% of the total number in the major families (a pair of proteins was considered highly similar if they showed >60% amino acid sequence identity along at least two-thirds of the length of one of them) (Riechmann et al., 2000). The pairs or groups of closely related genes most often correspond to duplications in different chromosomes (~65% on average), or to duplications in the same chromosome but at very

<table>
<thead>
<tr>
<th>Arabidopsis Homeobox family</th>
<th>Type</th>
<th>Number of genes</th>
<th>PHRA</th>
<th>Leu PHD zip finger</th>
<th>HAT7</th>
<th>Leu zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD-PHD finger</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-Zip classes I / II</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-Zip class III</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-Zip class IV</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WUS-like</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-BELL</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-KNOX</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.** The Arabidopsis Homebox (HB) and ZF-HB gene families. The Arabidopsis homeobox gene family can be subdivided into different groups according to the combinations of domains that the corresponding proteins contain, and to the phylogenetic analysis of the homeodomain. The number of members in each Arabidopsis homeobox gene subfamily is indicated (except for three genes, whose classification is unclear and that are not represented in this figure). Most of the combinations of a homeodomain with a domain of a different type (leucine zipper, PHD finger, START domain) are the result of domain shuffling events specific to the plant kingdom: those combinations are not found in Drosophila, C. elegans, or yeast homeodomain proteins. The only Arabidopsis homeodomain proteins that have an additional motif also found in animal homeodomain proteins are those of the KNOX class, which contain a MEINOX domain (Bürglin, 1998). Conversely, homeodomains in animals are associated with a large variety of motifs, such as the paired and POU-specific domains (which are themselves specific to animals), the LIM motif, or C2H2 zinc fingers, in combinations that are not present in Arabidopsis (for a similar depiction of the animal homeodomain proteins, see Gehring et al., 1994). The START domain is a lipid-binding domain that could provide regulation of HD-Zip class III protein function by sterols. It is found in a variety of eukaryotic proteins, but has been found associated with transcription factor domains only in this class of plant homeobox genes (Ponting and Aravind, 1999). Proteins of the ZF-HB family contain a homeodomain-related sequence that is more divergent from all the different groups of homeodomain sequences of the HB family than these are among themselves (Windhövel et al., 2001). In addition, ZF-HB proteins contain a plant-specific zinc coordinating motif (Windhövel et al., 2001).
large distances (~22%), than to tandem repeats (~13%) (Riechmann et al., 2000). In addition, clusters of three or more homologous transcription factor genes are very rare in the genome (Riechmann et al., 2000). This distribution indicates that it will be feasible to generate double or triple mutants for the majority of the pairs or groups of highly related genes that, because of their sequence similarity, might have overlapping or partially redundant functions (which might not be revealed by single mutant analyses; see below).

The analysis of ~120,000 Arabidopsis expressed sequence tags (ESTs) (sequences available in GenBank in January 2001) suggests that, in terms of overall expression and considered as a whole, transcription factor genes are not substantially different from the rest of the genes in the genome. Approximately half of the ~26,000 predicted genes are matched by an EST (Arabidopsis Genome Initiative, 2000; Theologis et al., 2000). Similarly, when the major Arabidopsis transcription families are considered, ~47% of the genes are represented by an EST (Table 2). This observation is in contrast to the sometimes common assumption that, because of their regulatory nature, genes of this class are generally expressed at low levels.


Chromatin structure is an important element of the mechanisms that determine gene expression patterns in eukaryotes, because nucleosome assembly eliminates the accessibility of promoter sequences for the basal transcription machinery. The unfolding of packed chromatin is necessary for gene expression and, conversely, repression requires the formation and maintenance of condensed chromatin structures. Gene silencing and epigenetic phenomena, in which chromatin structure and histone modifications play a role, are by themselves the subject of other chapters in this book.

As summarized above, one of the mechanisms of transcription factor action is the recruitment of chromatin remodeling complexes to target promoters. This mechanism has been deduced from research on transcription in yeast and mammalian cells, but studies on the regulation of the β-phaseolin (phas) gene in bean (Phaseolus vulgaris) suggest that it also operates in plants (reviewed in Li et al., 2001a). The phas gene is silenced in vegetative tissues as a consequence of the positioning of a nucleosome over the TATA boxes of the promoter, making them inaccessible to TBP, whereas nucleosome displacement allows the gene to be highly expressed during seed development (Li et al., 1998). Such modification in chromatin structure results from the presence of the seed-specific transcription factor PvALF, a member of the ABI3/VP1 family (Li et al., 1999). However, PvALF is not sufficient for phas transcriptional activation, which does not occur in the absence of abscisic acid (ABA) (Li et al., 1998). Thus, a plausible model is that PvALF mediates chromatin reconfiguration, then allowing the binding of ABA-responsive transcription factors and the recruitment and assembly of the basal transcription machinery on the phas promoter (Li et al., 2001a).

The remodeling or reconfiguration of chromatin involves different types of enzymes, such as members of the SWI2/SNF2 subfamily of the DEAD/H box superfamily of nucleic-acid stimulated ATPases, and proteins that covalently modify histones, such as acetyltransferases (HATs) and deacetylases (HDACs), kinases, and methyltransferases (for reviews: Kadonaga, 1998; Elgin and Workman, 2000; Fry and Peterson, 2001; Jenuwein, 2001; Urnov and Wolfe, 2001). All eukaryotes appear to contain several proteins belonging to each one of these types, and each type can be further divided into different structural subclasses. Such structural diversity allows different proteins with the same biochemical activity to be involved in specialized cellular functions. The chromatin proteins with enzymatic activity usually form part of multi-subunit complexes, which might be necessary for their specificity and functionality.

In general, histone acetylation is a characteristic of transcribed chromatin, whereas deacetylation is associated with repression. HATs acetylate the ε-amino groups of specific lysine residues in the amino-terminal tails of the histone proteins that form the octamer around which the DNA wraps in the nucleosomes. Histone deacetylase-containing complexes reverse this covalent modification (reviewed in Khochbin et al., 2001). The molecular mechanisms by which histone acetylation affects chromatin structure and influences transcription could involve the destabilization of interactions between the DNA and the histone octamer (by neutralizing positive charges), interference with the high-order packing of chromatin, or the modification of interactions between histones and other proteins (reviewed in: Marmorstein, 2001a; Marmorstein and Roth, 2001; Roth et al., 2001). Other types of post-translational modifications, such as phosphorylation and methylation, also occur on histones, and can regulate chromatin structure and transcriptional activation and repression (reviewed in Marmorstein, 2001a). Methylation of specific lysine residues in histone tails is a relatively stable modification, thus providing a stable epigenetic mark for transcriptional regulation and gene silencing via heterochromatin assembly (reviewed in: Jenuwein, 2001; Rice and Allis, 2001). Lysine methyltransferase activity has been demonstrated for several eukaryotic SET domain proteins (Table 3). In
addition, histone tails can also be methylated at arginine residues by a different class of enzymes, that act as coactivators of transcription (Chen et al., 1999).

A “histone code” hypothesis has been proposed, suggesting that distinct covalent histone Modifications might be used by the cell, sequentially or in combination, to generate a “code” that could be read by other proteins to produce different transcriptional outputs (Strahl and Allis, 2000). Reading the “histone code” would necessitate protein domains that recognize, in a receptor-ligand type of interaction, the different covalent modifications that can occur on histones (Strahl and Allis, 2000). Binding activities have been identified in several of the protein domains that are frequently found in chromatin-related proteins, such as the bromodomain and the chromodomian, which can recognize acetylated- and methylated-lysine residues of the histone tails, repectively (Table 3). A further level of complexity in regulatory mechanisms is inferred from the observation that the same covalent modifications that can be found on histones also occur on other proteins involved in transcriptional control. For example, histones are not the only targets for HATs, as HAT-catalyzed acetylation can also regulate the activity of transcription factors and co-factors (reviewed in: Sterner and Berger, 2000; Chen et al., 2001b). Lastly, another group of enzymes involved in chromatin remodeling is that of the DNA-dependent ATPases of the SWI2/SNF2 type. Yeast SWI2/SNF2 is the catalytic subunit of the multiprotein SWI/SNF remodeling complex, which can mediate the repositioning of nucleosomes by sliding histone octamers to other sites on the same DNA molecule, as well as by transferring them to other DNA molecules (reviewed in: Vignali et al., 2000; Flaus and Owen-Hughes, 2001).

A catalogue of known and putative chromatin proteins in Arabidopsis and maize has been compiled in ChromDB, a database that aims to present information on the entire complement of chromatin proteins in plants (http://chromdb.biosci.arizona.edu). ChromDB lists over 220 different Arabidopsis chromatin genes, including SWI2/SNF2 homologs (22 genes), HATs (12 genes; 10 are listed as HATs and 2 as TAFII250 homologs; Table 4), HDACs (17 genes; Table 4), and SET-domain-protein genes (29 genes), and also includes histones (50 genes) and homologs of subunits of global transcription factors.

A catalogue of known and putative chromatin proteins in Arabidopsis, or in any other euakaryotic organism, and in particular of the subset of those proteins that might be involved in transcriptional control, is complicated by several factors. First, chromatin remodeling is mediated by multiprotein complexes, some of which have already been purified and characterized from yeast and animal (mammalian, Drosophila) cells, but none from plants. Some of these complexes (for example, the yeast SAGA and human PCAF complexes) show a remarkable conservation in subunit composition, but there are also cases of proteins and complexes that are specific to one kingdom (Sterner and Berger, 2000, and see below). Thus, biochemical studies will be needed to obtain a complete description of the Arabidopsis complement of chromatin proteins. Another complication for the identification of bona fide chromatin proteins arises from their multi-domain architecture. Chromatin proteins frequently combine different domains or motifs of distinct molecular functions (Table 3). However, those domains are not necessarily unique to chromatin proteins, as they (or related sequences) can be present in other types of proteins. For example, an Arabidopsis protein that contains sequences related to the chromodomian is localized to the chloroplast and forms part of the chloroplast signal recognition particle pathway (Klimyuk et al., 1999). Lastly, the structure of chromatin influences not only transcription, but also other nuclear processes that are physically associated with the genome, such as replication, recombination, and DNA repair. Thus, that a protein is chromatin-related does not necessarily imply that it is involved in transcriptional control. For these different reasons, the identification and description by sequence similarity searches of the complement of chromatin proteins involved in transcriptional regulation, and of their biochemical and molecular functions, is more complicated than that of the sequence-specific DNA-binding transcription factors discussed above.

It is apparent from the content in known chromatin genes of the Arabidopsis genome that chromatin remodeling is important in plants for the control of gene
<table>
<thead>
<tr>
<th>Domain</th>
<th>Molecular function</th>
<th>Domain characteristics and protein types</th>
<th>InterPro accession number</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAT</td>
<td>Acetylation of lysine residues</td>
<td>A signature motif of the HAT domain is the acetyl-CoA substrate recognition and binding site. HAT proteins usually contain one or several additional domains, such as bromodomains, chromodomains, and zinc coordinating motifs of various types. HAT proteins can be classified into distinct subfamilies, with little sequence similarity among them (see Table 4).</td>
<td>IPR000182</td>
<td>(Marmorstein, 2001b; Marmorstein and Roth, 2001; Roth et al., 2001)</td>
</tr>
<tr>
<td>HDAC</td>
<td>Deacetylation of lysine residues</td>
<td>HDAC proteins can be classified into distinct subfamilies, with little sequence similarity among them (see Table 4).</td>
<td>IPR000286, IPR003084</td>
<td>(Koehnbin et al., 2001)</td>
</tr>
<tr>
<td>SW12/SNF2</td>
<td>DNA-dependent ATPase</td>
<td>The DNA-dependent ATPase domain characteristic of the SW12/SNF2 proteins belongs to the superfamilly of DEAD/H helicases. However, SW12/SNF2 proteins do not show helicase activity; rather, the ATPase motors are used for nucleosome displacement. The SW12/SNF2 domain contains seven distinct conserved motifs, but there are proteins, such as Arabidopsis MOM1 (see Table 5), whose SW12/SNF2 domain does not contain all of them.</td>
<td>IPR000330</td>
<td>(Vignali et al., 2000; Flaus and Owen-Hughes, 2001)</td>
</tr>
<tr>
<td>SET</td>
<td>Methylation of lysine residues</td>
<td>Several proteins that contain a SET domain flanked by cysteine-rich regions show histone methyltransferase activity, for which both the SET domain and the cys-rich regions are essential. Histones might not be the only substrate for SET domain proteins, and some SET domain proteins might be enzymatically inactive.</td>
<td>IPR001214</td>
<td>(Rea et al., 2000; Jenuwein, 2001)</td>
</tr>
</tbody>
</table>

Non-enzymatic domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Molecular function</th>
<th>Domain characteristics and protein types</th>
<th>InterPro accession number</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bromodomain</td>
<td>Interaction with acetylated lysine</td>
<td>Frequently found in transcriptional regulatory proteins, such as HATs, transcriptional coactivators, and corepressors, but not exclusive to them.</td>
<td>IPR001487</td>
<td>(Jeanmougin et al., 1997; Dhalluin et al., 1999; Jacobson et al., 2000)</td>
</tr>
<tr>
<td>Chromodomain</td>
<td>Binding to specific methylated lysine residues; RNA binding; protein-protein interaction?</td>
<td>Found in some HATs, in some SET-domain containing proteins, as well as in combination with the chromo-shadow domain. The chromodomain of the Drosophila HAT MOF is involved in RNA binding; it has been proposed that interaction with non-coding RNAs could target regulators to specific chromosomal locations.</td>
<td>IPR000953</td>
<td>(Akhatar et al., 2000; Jones et al., 2000; Bannister et al., 2001; Lachner et al., 2001)</td>
</tr>
<tr>
<td>Chromo-shadow domain</td>
<td>Protein-protein interaction</td>
<td>The chromo-shadow domain shows sequence similarity to the chromodomain. The chromo-shadow domain of Drosophila HP1 A can specifically recognize a consensus pentapeptide sequence, which is sufficient for specific interactions. The chromodomain does not bind to such consensus pentapeptide sequence.</td>
<td>(Brischer et al., 2000; Smothers and Henikoff, 2000)</td>
<td></td>
</tr>
<tr>
<td>PHD finger (Zn)</td>
<td>Protein-protein interaction</td>
<td>Initially identified in a subset of plant homeodomain proteins, the PHD finger is also present in a variety of proteins involved in chromatin-mediated transcriptional regulation, such as PKL (an Arabidopsis SW12/SNF2 protein), and members of the Trithorax (trx) and Polycomb (Pc) groups.</td>
<td>IPR001965</td>
<td>(Ausland et al., 1995; Halbach et al., 2000; Capili et al., 2001)</td>
</tr>
</tbody>
</table>

continued on following page
expression. That some of the molecular mechanisms for chromatin reconfiguration and transcriptional control are conserved among plants and the other eukaryotic kingdoms can be deduced from the presence of orthologous genes. Furthermore, similarities or functional equivalence at the molecular or physiological level has been demonstrated in some cases, as illustrated with the following examples.

An RPD3-type maize histone deacetylase has been shown to complement a S. cerevisiae null mutant in the homologous RPD3 gene (Rossi et al., 1998).

The Arabidopsis gene BUSHY (BSH), which codes for a protein with high sequence similarity to S. cerevisiae SNF5 (a component of the SWI/SNF remodeling complex), can partially complement a snf5 mutation in yeast (Brzeski et al., 1999).

Arabidopsis homologs of human CBP/p300 proteins recapitulate the binding specificity of p300 for the adenoviral oncoprotein E1A, in addition to being capable of activating transcription in mammalian cells (Bordoli et al., 2001).

The Arabidopsis protein that is orthologous to yeast GCN5 possesses HAT activity, and can interact with Arabidopsis ADA2 proteins, suggesting that a complex analogous to yeast SAGA (of which GCN5 and ADA2 form part) and human PCAF also exists in plants (Stockinger et al., 2001).

PICKLE (PKL; also initially referred to as GYMNOS) is an Arabidopsis protein of the SWI2/SNF2-type that appears to be involved in the repression of several important developmental regulators, such as LEC1 and meristematic genes (Eshed et al., 1999; Ogas et al., 1999). PKL is homologous to human Mi-2, a component of the NuRD complex. By virtue of its different subunits, the NuRD complex combines both ATP-dependent chromatin remodeling and HDAC activity. The homology between PKL and Mi-2 suggests that a NuRD-like complex might exist in plants; thus, a plausible mechanism for gene

<table>
<thead>
<tr>
<th>Domain</th>
<th>Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RING (Zn)</td>
<td>Protein-protein interaction</td>
<td>The RING finger can specifically interact with E2 ubiquitin conjugating enzymes, promoting ubiquitination of heterologous substrates or of the RING finger proteins themselves. The RING finger is found in a wide variety of otherwise unrelated proteins (including some chromatin-related proteins and transcription factors) that are involved in different biological processes. (Joazeiro and Weissman, 2000)</td>
</tr>
<tr>
<td>LIM (Zn)</td>
<td>Protein-protein interaction</td>
<td>The LIM finger is found in a wide variety of otherwise unrelated proteins (including some chromatin-related proteins and transcription factors) that are involved in different biological processes. (Sánchez-Garcia and Rabbits, 1994)</td>
</tr>
<tr>
<td>SAND</td>
<td>DNA binding (sequence specific)</td>
<td>The SAND domain is found in animals, but not in plants or in yeast. It is present in proteins that are involved in chromatin–mediated transcriptional regulation, in combination with other motifs, such as bromodomains and PHD fingers. Interacts with low complexity nucleotide sequences, although specifically. (Bottomley et al., 2001)</td>
</tr>
<tr>
<td>AT-Hook</td>
<td>DNA binding (AT rich sequences)</td>
<td>Defined as the DNA binding domain of the HMG-1 and HMG-Y proteins, the AT-hook domain can be found in a variety of nuclear proteins, including chromatin proteins and transcription factors. It can be found associated with a large variety of other domains involved in DNA-related processes. (Aravind and Landsman, 1998)</td>
</tr>
<tr>
<td>ARID</td>
<td>DNA binding (AT rich sequences)</td>
<td>Domain found in a variety of nuclear proteins, including chromatin proteins and transcription factors. In animal ARID proteins, the domain is combined with motifs such as PHD fingers or the jumonji domain. In the Arabidopsis ARID proteins, the ARID domain is associated with an HMG box. (Kortschak et al., 2000; Riechmann et al., 2000)</td>
</tr>
</tbody>
</table>

(a) This table presents domains that are frequently found in chromatin-related proteins, but not a comprehensive list of all types of domains and motifs that can be identified in such proteins. (b) (Zn) indicates a zinc coordinating motif.
<table>
<thead>
<tr>
<th>Type (Prototype member)</th>
<th>Number of genes in Arabidopsis</th>
<th>Arabidopsis proteins</th>
<th>Characteristics</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Histone Acetyltransferases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNAT (yeast GCN5 and HAT1)</td>
<td>4</td>
<td>AtGCN5/T14E10_180, At2g41450, At5g56740, At5g50520</td>
<td>Yeast GCN5 and human PCAF (which are the catalytic subunit of the SAGA and PCAF complexes, respectively) act as coactivators for transcription factors. GCN5 proteins contain HAT- and bromo-domains. In addition, metazoan GCN5 proteins contain an N-terminal extension that is absent in the homologs from yeast and Tetrahymena. AtGCN5 also contains an N-terminal extension, but unrelated in sequence to those of metazoans. It has been shown that AtGCN5 possesses HAT activity, and that it can interact with ADA2 proteins, implying the existence of a SAGA-PCAF-like complex in plants. Furthermore, an Arabidopsis transcription factor with an acidic activation domain (CBF1) was shown to interact with ADA2, suggesting that CBF1 might use interactions with chromatin-remodeling complex(es) to regulate transcription.</td>
<td>(Stecher and Berger, 2000; Lusser et al., 2001; Mann, 2001)</td>
</tr>
<tr>
<td>CBP (human CBP and p300)</td>
<td>5</td>
<td>PCAT1/At1g67210, PCAT2/At1g79000, PCAT3/At1g55970, PCAT4/At3g12980, At1g6710</td>
<td>Human CBP and p300 are transcriptional coactivators with HAT activity that interact, in distinct complexes, with many sequence-specific transcription factors. CBP/p300 homologs are found in animals and plants, but not in <em>S. cerevisiae</em>. Arabidopsis CBP/p300 homologs lack the bromodomain and the CRB-domain-binding region that are highly conserved in animal CBP/p300 proteins. It has been shown that Arabidopsis PCAT2 possesses HAT activity, and that it can activate transcription in an acetyltransferase-dependent manner in mammalian cells when fused to GAL4-DBD. PCAT1-4 are ubiquitously expressed.</td>
<td>(Goodman and Smolik, 2000; Bordoli et al., 2001; Vo and Goodman, 2001)</td>
</tr>
<tr>
<td><strong>MYST</strong> (yeast SAS2)</td>
<td>2</td>
<td>At5g64610, At5g09740</td>
<td>HATs of the MYST family participate in a variety of chromatin-related processes, including gene silencing (yeast SAS2 and SAS3) and transcriptional coactivation (yeast ESA1, as the catalytic subunit of the NuA4 complex).</td>
<td></td>
</tr>
<tr>
<td>TAF,250</td>
<td>2</td>
<td>At1g32750, At3g19040</td>
<td>TAF,250 is a TBP-associated factor, an is found in all eukaryotes. The modular structure of human TAF,250 includes an N-terminal kinase domain, a HAT domain, and two C-terminal bromodomains. The Arabidopsis TAF,250 homologs have only one bromodomain, also at the C-terminus, whereas the yeast homolog has none.</td>
<td>(Kocherin et al., 2001; Lusser et al., 2001)</td>
</tr>
<tr>
<td><strong>Histone Deacetylases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD2 (maize HD2a)</td>
<td>4</td>
<td>AtHD2A/HD2A/At3g44750, AtHD2B/HD2A/At5g22650, AtHD2C/HD2A/At5g03740, AtHD2D/HD2A/At2g27840</td>
<td>This class of HDACs is specific to the plant kingdom (no closely related proteins identified in animals or fungi). HD2 proteins do not show sequence similarity to RPD3- and HDAC1-type HDACs. They contain an N-terminal region of sequence similarity to peptide- and -polylypeptide isoforms (PIIases), a central acidic region (that has been shown to be phosphorylated in the case of HD2 from maize), and a carboxyl-terminal zinc finger structure that could be involved in protein-protein interactions (present in most, but not all of the plant proteins of this class). Maize HD2 appears to be located preferentially in the nucleolus, and could be involved in the regulation of rRNA genes. Arabidopsis AHHD2A was shown to be capable of repressing expression using a transient expression system.</td>
<td>(Lusser et al., 1997; Wu et al., 2000b; Dong et al., 2001; Lusser et al., 2001)</td>
</tr>
</tbody>
</table>

*continued on following page*
repression by PKL is via histone deacetylation mediated by NuRD (reviewed in Ahringer, 2000).

The Polycomb group (PcG) and the trithorax group (trxG) of proteins in Drosophila and mammals control the cellular inheritance of mitotically stable states of gene expression, homeotic genes in particular. PcG and trxG proteins (repressors and activators, respectively) are thought to regulate transcription by modulating the structure of chromatin (reviewed in: Brock and van Lohuizen, 2001; Francis and Kingston, 2001; Mahmoudi and Verrijzer, 2001). The proteins within each group (Pc or trx) can be unrelated in sequence; rather, their relationship to each other comes from the fact that they operate together in the form of multi-subunit complexes of a genetically defined function (Gould, 1997). Homologous or related proteins for some PcG and trxG factors have been identified in Arabidopsis, and in some cases functionally characterized. Three Arabidopsis proteins show homology to the Drosophila SET-domain PcG protein Enhancer of zeste (E(z)), CURLY LEAF (CLF), CURLY LEAF LIKE (CLK), and MEDEA (MEA) (Goodrich et al., 1997; Grossniklaus et al., 1998) (Table 5). CLF is a repressor of floral organ identity (i.e., homeotic) gene expression in vegetative tissues (Goodrich et al., 1997), whereas MEA is involved in the maternal control of embryogenesis (Grossniklaus et al., 1998). Another Arabidopsis PcG protein involved in seed development is FERTILIZATION-INDEPENDENT ENDOSPERM (FIE), which shows homology to PcG proteins with WD repeats, such as Drosophila extra sex combs (esc) (Ohad et al., 1999). Animal E(z) and esc proteins have been shown to interact and to co-localize in unique complexes. Similarly, Arabidopsis FIE and MEA also interact, which provides a molecular explanation for the similarities between the fie and mea mutant phenotypes (Spillane et al., 2000; Yadegari et al., 2000).

Other Arabidopsis proteins, such as EMBRYONIC FLOWER2 (EMF2), FERTILIZATION-INDEPENDENT SEED 2 (FIS2), and VERNALIZATION 2 (VRN2) are related to a different Drosophila PcG protein, Suppressor of zeste 12 (Su(z)12) (Luo et al., 1999; Birve et al., 2001; Gendall et al., 2001; Yoshida et al., 2001).

Despite these similarities, however, novel features in the chromatin-mediated regulation of gene expression have also evolved in plants. Plants contain what appears to be a kingdom-specific family of histone deacetylases, the HD2 class (Lusser et al., 1997; Aravind and Koonin, 1998; Dangl et al., 2001) (Table 4). Orthologs for some Arabidopsis chromatin proteins are not found in yeast or animals. This is the case, for example, of MOM1, a SWI2/SNF2-related protein that is involved in the maintenance of transcriptional gene silencing (Amedeo et al., 2000; Arabidopsis Genome Initiative, 2000). In addition, homologous chromatin proteins can show structural variation among the different eukaryotic kingdoms, and some of those variations appear to be specific to plants. In fact, eukaryotic chromatin proteins are a prominent example of evolutionary innovation by domain shuffling, deletion, and accretion (International Human Genome Sequencing Consortium, 2001). For example, Arabidopsis CBP/p300-like proteins lack the
### Table 5. Arabidopsis chromatin genes functionally characterized

<table>
<thead>
<tr>
<th>Gene</th>
<th>Functional class</th>
<th>Method of characterization</th>
<th>Expression pattern</th>
<th>Phenotype/Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>AtHD2A</em> /</td>
<td>HDAC HD2-type</td>
<td>Antisense RNA</td>
<td>Expressed primarily in flowers and young siliques</td>
<td>Silencing of <em>AtHD2A</em> results in aborted seed development and stunted siliques.</td>
<td>(Wu et al., 2000b)</td>
</tr>
<tr>
<td><em>AtHD1</em> /</td>
<td>HDAC RPD3-type</td>
<td>Antisense RNA</td>
<td>Ubiquitous</td>
<td>Reduction of <em>AtHD1</em> transcript levels caused pleiotropic developmental alterations, the most frequent of which was a delay in flowering time. These phenotypic alterations were accompanied by enhanced histone acetylation, and by the ectopic expression of tissue or cell-type specific genes. These results suggest that AtHD1 is a global regulator during development, and illustrate that, as in other eukaryotes, histone deacetylation is directly involved in suppression of gene expression in plants.</td>
<td>(Wu et al., 2000a; Tian and Chen, 2001)</td>
</tr>
<tr>
<td><em>AtHDA6</em> /</td>
<td>HDAC RPD3-type</td>
<td>Mutant</td>
<td>Ubiquitous</td>
<td>Mutations in <em>AtHDA6</em> affect transgene silencing, but the use of histone deacetylase inhibitors suggests that such effect might be independent of AtHDA6 deacetylase activity. The <em>hda6</em> mutants do not show obvious alterations in morphology and development. It is not yet clear the role that HDA6 might play in regulating the expression of Arabidopsis endogenous genes.</td>
<td>(Murfett et al., 2001)</td>
</tr>
<tr>
<td><em>BSH</em></td>
<td>SNF5 (subunit of SW1/SNF complex)</td>
<td>Antisense RNA</td>
<td>Ubiquitous</td>
<td>Reduction of <em>BSH</em> transcript levels caused developmental alterations: bushy growth and flowers that do not produce seeds.</td>
<td>(Brzeski et al., 1999)</td>
</tr>
<tr>
<td><em>MOM1</em></td>
<td>SW12/SNF2</td>
<td>Mutant, antisense RNA</td>
<td></td>
<td>Mutation of <em>MOM1</em> or reduction of transcript levels by antisense RNA cause the reactivation of transcription from methylated, silenced transgenic loci, implicating <em>MOM1</em> in the maintenance of transcriptional silencing. This role, however, might be very specific, because the <em>mom1</em> mutant shows normal growth.</td>
<td>(Amedeo et al., 2000)</td>
</tr>
<tr>
<td><em>DDM1</em></td>
<td>SW12/SNF2</td>
<td>Mutant</td>
<td></td>
<td>Loss of <em>DDM1</em> function results in a substantial reduction of genomic cytosine methylation, primarily in repeated sequences. Single copy sequences lose methylation over generations in a <em>ddm1</em> background, resulting in the plants accumulating developmental abnormalities. <em>DDM1</em> thus implicates chromatin remodeling in the maintenance of DNA methylation. In addition, <em>DDM1</em> activity is necessary to stabilize (reduce mobilization of) transposons.</td>
<td>(Jeddeloh et al., 1999; Miura et al., 2001; Singer et al., 2001)</td>
</tr>
<tr>
<td><em>PKL</em></td>
<td>SW12/SNF2</td>
<td>Mutant</td>
<td>Primarily throughout developing embryos and meristems and primordia</td>
<td>PKL is a SW12/SNF2 protein of the CHD3 subclass. CHD3 proteins have been identified as repressors of transcription, and PKL appears to be involved in the repression of several important developmental regulators, such as <em>LECI</em>, and meristematic genes. Consistently, <em>pk1</em> mutants show pleiotropic morphological phenotypes (that are dependent on the genetic background).</td>
<td>(Eshed et al., 1999; Ogas et al., 1999)</td>
</tr>
<tr>
<td><em>CLF</em></td>
<td>PeG protein of the E(z) type, with SET domain</td>
<td>Mutant</td>
<td>Primarily in meristems and primordia</td>
<td>CLF is a repressor of floral organ identity (i.e., homeotic) gene expression in vegetative tissues that encodes a SET-domain protein with homology to Drosophila Enhancer of zeste.</td>
<td>(Goodrich et al., 1997)</td>
</tr>
</tbody>
</table>

*continued on following page*
The Arabidopsis Book 20 of 46

bromodomain and the CREB-binding region that are highly conserved in animal CBP/p300 proteins (Bordoli et al., 2001) (Table 4; CBP/p300 proteins are not found in yeast). Instead, one of these Arabidopsis proteins (PCAT1) contains a repeated motif of unknown function that does not show sequence similarity to any other known amino acid motif (Bordoli et al., 2001).

Other Arabidopsis chromatin genes that have already been genetically or functionally characterized further show the importance of chromatin-mediated regulation of gene expression in multiple aspects of the plant life cycle (Table 5). Reduction of AtHD1 (an HDAC-coding gene, also referred to as AtRPD3A) transcript levels by using antisense RNA caused pleiotropic developmental alterations, suggesting a global role for AtHD1 in regulating gene expression during development (Wu et al., 2000a; Tian and Chen, 2001). Similarly, reduction of AtHD2A activity (which codes for an HDAC of the plant-specific HD2 class) resulted in aborted seed development (Wu et al., 2000b). In another study, mutants in the HDAC gene AtHDA6 were isolated, which were morphologically wild-type but showed deregulated expression of transgenes, suggesting that AtHDA6 might be specifically involved in (transgene) silencing processes (Murfett et al., 2001). In addition to MOM1 and PKL, mentioned above, another Arabidopsis gene coding for a SWI2/SNF2-type protein that has been functionally characterized is DECREASE IN DNA METHYLATION1 (DDM1). DDM1 is required to maintain normal cytosine methylation patterns and to stabilize transposon behavior (Jeddeloh et al., 1999; Miura et al., 2001; Singer et al., 2001).

In summary, genetic studies on a variety of biological processes in Arabidopsis, the determination of its genome sequence, and biochemical studies performed in maize, have all started to illuminate the different physiological functions that chromatin remodeling might play in plants. However, our understanding of chromatin remodeling at the molecular level, and on how it influences plant nuclear processes, is extremely limited, and mostly derived from comparisons with the better-studied systems of yeast, Drosophila, and mammalian cells. If chromatin research in these model organisms is to be viewed as an example, it is clear that biochemical studies will be essential to understand chromatin in plants.

5. The combinatorial nature of transcriptional regulation: promoters, cis-elements and trans-acting factors.

Whereas plants and animals (or, to be more precise, Arabidopsis and Drosophila, C. elegans, and humans) might have comparable contents of transcription factors (3.5-6.6% of the total number of genes; see above), the organization of the regulatory sequences on which these transcription factors act can be different in the two kingdoms. In animals, the regulatory sequences that determine the correct temporal and spatial expression of a gene can extend over tens of kilobases (kbs) of DNA (for a review, see Bonifer, 2000). In contrast, regulatory sequences of plant genes usually span much shorter DNA intervals, often less than 1 or 2 kbs. This is reflected in the compact organization of the Arabidopsis genome, in which gene density is high. Out of the sequenced 115.4 megabases (Mb) of the 125 Mb genome, 51.2 Mb (or 44% of the sequenced regions) correspond to predicted exons and introns (Arabidopsis Genome Initiative, 2000). On average, there is one gene per 4.5 kb of DNA: the gene length (exons plus introns) is approximately 2 kb, and ~2.5 kb correspond to intergenic regions. Considering the whole genome, transposons account for ~20% of the intergenic DNA, resulting in an average of 2 kb of DNA for the 5’ and 3’ regions of a particular gene (Arabidopsis Genome Initiative, 2000). Other plants, maize for example, have genomes that are much larger than that of
Arabidopsis, but with a similar organization of promoter sequences: in the maize genome, active genes are usually distributed in compact gene-rich islands, with much of the genomic DNA corresponding to repetitive sequences made up of retrotransposons (SanMiguel et al., 1996; Fu et al., 2001). As a result, regulatory sequences in Arabidopsis, and in plants in general, are easier to identify and delimit experimentally than in, for example, humans (for an introduction to the problem in mammals, see: Gumucio et al., 1993; Hardison et al., 1997; Bonifer, 2000; Fickett and Wasserman, 2000; Scherf et al., 2001). Compact Arabidopsis 5’ promoter sequences often recapitulate faithfully the expression of the native gene when assayed in transgenic plants by reporter gene fusions, that is, in a chromatin context. However, this is not always the case, because regulatory elements can also be localized downstream of the transcription start site: in introns, in the 5’ untranslated region, or in 3’ sequences (Larkin et al., 1993; Sieburth and Meyerowitz, 1997; Deyholos and Sieburth, 2000; Yu et al., 2001). For example, the large second intron of the MADS-box floral organ-identity gene AGAMOUS (AG) is essential for the correct expression of the gene, and contains binding sites for at least two AG regulators, LFY and WUSCHEL (WUS) (Sieburth and Meyerowitz, 1997; Bombiès et al., 1999; Busch et al., 1999; Deyholos and Sieburth, 2000; Lohmann et al., 2001).

In spite of the structural differences between animal and plant cis-regulatory and promoter regions, regulation of gene expression is often in both cases the result of multiple inputs, reflecting, or taking advantage of, the combinatorial nature of the mechanisms of eukaryotic transcription. Multiple stimuli can converge through different cis-acting elements on a promoter to coordinately regulate the expression of the corresponding gene (Arnone and Davidson, 1997; Yuh et al., 1998). The cis-acting elements are usually organized in a modular fashion: both in animals and plants, the regulatory region of a gene can be partitioned into discrete subelements, each one containing one or several binding sites for transcription factors and performing a certain regulatory function (Benfey and Chua, 1990; Arnone and Davidson, 1997). The modular nature of cis-regulatory systems is exemplified by the 2.3 kb promoter region of the sea urchin developmentally regulated Endo16 gene, one of the best characterized eukaryotic promoters (Yuh et al., 1998, 2001). It consists of six different regulatory modules, which provide different regulatory functions that are integrated through interrelations between the modules, and result in the spatial expression, and repression, of the gene, as well as on its variable rates of transcription (Yuh et al., 1998, 2001). This cis-regulatory system therefore acts like an information processing unit, and computational models for the modes of action of some of its modules have been established (Arnone and Davidson, 1997; Yuh et al., 1998, 2001). The view of cis-regulatory regions as information processing systems in which the output of developmental (or other) inputs is wired, is probably applicable to the eukaryotic genome as a whole (Arnone and Davidson, 1997; Davidson, 2001).

In plants, regulation of gene expression by systems of cis-acting modules, and the fact that these modules can interact synergistically (i.e., that combinations of modules direct gene expression in a manner not observed with the modules in isolation), was first described for the cauliflower mosaic virus (CaMV) 35S promoter (Benfey and Chua, 1990). The CaMV 35S promoter directs high levels of expression in most tissues and developmental stages when introduced as a transgene in plants, but can be dissected into subdomains that confer tissue-specific expression (Benfey and Chua, 1990; Benfey et al., 1990a, b). The combinatorial interaction of cis-elements has also been demonstrated, for example, for Arabidopsis light-regulated promoters. Several consensus cis-sequences that are necessary for high activity in the light have been identified in the promoters of photosynthesis-associated nuclear genes (such as the rbcS and cab genes). These consensus sequences are referred to as ‘light responsive elements (LREs)’. Minimal promoters, sufficient to confer light-dependent expression, contain several LREs, but no single LRE is found in all light-regulated promoters (in fact, some LREs are also present in promoters that are not regulated by light) (Argüello-Astorga and Herrera-Estrella, 1998). LREs function combinatorially: whereas they cannot confer proper light responsiveness in isolation, paired combinations of them are able (1) to respond to a wide spectrum of light through the phytochrome signal transduction pathways, (2) to respond to the chloroplast developmental state, and (3) to confer a photosynthetic-cell specific expression pattern, therefore satisfying the strict definition of light-inducible (Puente et al., 1996; Chattopadhyay et al., 1998b). Thus, it is the combination of LREs in the promoter what serves as the integration system for the coordination of different light and developmental inputs to regulate the expression of the photosynthesis-related genes (Puente et al., 1996; Chattopadhyay et al., 1998b). Similarly, the promoter of the meristem identity gene LEAFY serves as the convergence point for different signals that control flowering time in Arabidopsis, including both environmental cues (daylength pathway) and endogenous signals (gibberellins) (Blázquez and Weigel, 2000), in accordance with the concept of promoters acting as information processing systems.

The combinatorial and synergistic function of cis-elements in eukaryotic promoters is logically accompanied by the combinatorial mode of action of the trans-acting factors that bind to those sites, and allows for the generation of regulatory diversity by a limited number of factors and
binding sites. The requirement of several, often adjacent, cis-elements for the regulation of gene expression can be related to direct interactions between the proteins that bind to those elements. Direct interactions among transcription factors, however, is not the only molecular mechanism by which they can function combinatorially to regulate gene expression, since they can also interact with other components of the transcription machinery and with other classes of regulatory proteins. For example, LFY and WUS cooperatively participate in the regulation of AG expression, yet they bind independently to AG cis-regulatory sequences and a direct interaction between the two proteins has not yet been detected (Lohmann et al., 2001).

Several examples of direct interactions between different Arabidopsis transcription factors have been reported, although the number is still small. In addition to increasing the regulatory repertoire, direct interactions between transcription factors are one of the mechanisms by which proteins with very similar DNA binding domains might achieve regulatory specificity (see, for example, Grotewold et al., 2000). Direct interactions can occur between members of the same protein family, to form dimeric complexes that bind to palindromic DNA sequences (such as in the case of the MADS domain proteins: Huang et al., 1996; Riechmann et al., 1996a; Riechmann et al., 1996b), or between transcription factors of different families. Examples of the latter include Arabidopsis, maize, and petunia proteins of the MYB and bHLH families (Table 6), interactions between bZIP and AB13/VF1 proteins in rice and Arabidopsis (Hobo et al., 1999; Nakamura et al., 2001), between soybean C2H2 zinc finger and bZIP proteins (Kim et al., 2001), and between Dof and bZIP transcription factors in Arabidopsis and maize (Chen et al., 1996; Vicente-Carbajosa et al., 1997). The interaction between bZIP and AB13/VF1 proteins (TRAB1 and VF1, respectively, in the case of rice, and AB13 and AB15 in Arabidopsis) provides a mechanism for VF1-mediated, ABAd-inducible gene expression (Hobo et al., 1999; Nakamura et al., 2001). The interaction between bZIP and Dof proteins might mediate the endosperm-specific expression of seed-storage proteins (Vicente-Carbajosa et al., 1997).

Within the MADS domain family, interactions are not limited to the formation of protein dimers, but also include the formation of ternary complexes. APETALA1 (AP1), APETALA3 (AP3), PISTILLATA (PI), and AGAMOUS (AG) are MADS domain proteins that, together with AP2, control the development of floral organs in Arabidopsis (Bowman et al., 1991; Coen and Meyerowitz, 1991; Goto et al., 2001; Jack, 2001; Theiben, 2001). AP3 and PI bind to DNA forming a heterodimer, whereas AP1 and AG can both bind to DNA as homodimers or as heterodimers with other MADS domain proteins (Huang et al., 1996; Riechmann et al., 1996a; Riechmann et al., 1996b). The activity of AP1, AP3, PI, and AG, however, requires of floral cofactors, also MADS domain proteins, that are encoded by the SEPALLATA genes, SEP1, SEP2, and SEP3 (Pelaz et al., 2000; Honma and Goto, 2001; Pelaz et al., 2001a; Pelaz et al., 2001b). In yeast two-hybrid experiments, AP3 and PI together, but neither one of the two proteins individually, can physically interact with AP1 and with SEP3 (Honma and Goto, 2001). The ectopic expression of AP3, PI, and AP1, or of AP3, PI, and SEP3 converts vegetative leaves into petaloid organs (Honma and Goto, 2001), whereas AP3 and PI alone are not sufficient for such organ conversion (Krizek and Meyerowitz, 1996). These results indicate that the formation of ternary complexes might be necessary for the function of AP3 and PI. The role that ternary complex formation might play in AP3 and PI function could be several fold: from providing an activation domain that AP3 and PI appear to lack, but that AP1 and SEP3 have (Honma and Goto, 2001), to increasing the DNA-binding specificity/affinity of the complex versus that of the protein dimers, given that the organ identity activity of AP1, AP3, PI, and AG is independent of their individual DNA-recognition properties (Riechmann and Meyerowitz, 1997a). These results with the Arabidopsis floral organ identity proteins parallel and expand previously obtained data for the Antirrhinum majus MADS-domain proteins SQUAMOSA, DEFICIENS, and GLOBOSA (SQUA, DEF, and GLO, which are AP1, AP3, and PI orthologs, respectively). DEF and GLO were also found to form ternary complexes with SQUA, and the three proteins together to bind DNA with increased affinity versus SQUA or DEF/GLO alone (Egea-Cortines et al., 1999).

Whereas these isolated examples illustrate the importance of interactions between transcription factors for the regulation of transcription, and how the combinatorial logic can operate, they do not convey the scope of the regulatory interactions in which transcription factors could be involved. For this, the whole complement of proteins should be considered (see below).


The future of biological sciences in the “post-genome era” has been anticipated as an endeavor to generate a collection of comprehensive “functional maps” (corresponding to the “transcriptome”, the “phenome”, the “interactome”, the “localizome”, and so on), that would be compiled into a “biological atlas” which would represent the modular nature of biological processes in a holistic manner and allow the formulation of new hypothesis.
Transcriptional Regulation: a Genomic Overview 23 of 46

Greenbaum et al., 2001; Kim, 2001; Vidal, 2001). These maps could be visualized as two-dimensional matrices in which one axis represents all the genes or proteins that can be tested in an organism, and the other a comprehensive series of mutant backgrounds, conditions to which the organism can be exposed, etc. (Vidal, 2001). For instance, a yeast transcriptome map of this type is already being developed (Hughes et al., 2000b). The interactome would represent the map of physical interactions among all the proteins of a proteome (reviewed in Walhout and Vidal, 2001) for attempts to construct the yeast interactome, see: Uetz et al., 2000; Ito et al., 2001). The localizome map would describe in what cells and cellular compartments, and when, all the proteins of an organism’s proteome can be found; and to produce the phenome, a collection of mutants encompassing all the genes in a genome would be screened in a large series of phenotypic assays (for C. elegans and yeast, see: Ross-

<table>
<thead>
<tr>
<th>Organism</th>
<th>bHLH protein</th>
<th>MYB protein</th>
<th>Function</th>
<th>Characteristics</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>GL3</td>
<td>GLI</td>
<td>Trichome cell fate determination</td>
<td>GL3 (a homolog of maize R) interacts with itself, with GL1, and with the WD-40 repeat protein TTG1. Genetically, GL1, GL3, and TTG1 are all required for trichome formation. GL3 might also interact with additional MYB proteins, such as PAP1, and with other bHLH proteins, such as the one represented by EST I46D23T7. These other bHLH proteins might in turn have interaction properties similar to those of GL3 (that is, interact with the same MYB and WD-40 repeat proteins).</td>
<td>Payne et al., 2000)</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>R homolog</td>
<td>WER</td>
<td>Cell fate determination in root epidermis</td>
<td>WER and GL1 are highly related proteins. They specify different cell-types because of their distinct, non-overlapping expression patterns, but they are otherwise functionally equivalent. WER was shown to interact in the yeast two-hybrid system with the maize R protein. The Arabidopsis bHLH protein or proteins that interact with WER still remain to be identified.</td>
<td>Lee and Schiefelbein, 1999, 2001</td>
</tr>
<tr>
<td>Maize</td>
<td>R</td>
<td>C1</td>
<td>Anthocyanin biosynthetic pathway</td>
<td>C1 requires the interaction with the HLH protein R to activate transcription of the anthocyanin biosynthetic genes. The mechanism by which R controls transcription is still unknown (DNA-binding activity for this protein has not been described). P is a MYB protein, highly related to C1, that can activate some of the C1 targets independently of R.</td>
<td>Moi et al., 1998; Grotewold et al., 2000)</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>rd22BP1</td>
<td>ATMYB2</td>
<td>Drought- and ABA-regulated gene expression</td>
<td>rd22BP1 and ATMYB2 interact synergistically in the regulation of the dehydration- and ABA-responsive promoter from the rd22 gene, although a direct interaction between the two proteins still remains to be demonstrated. The rd22 promoter contains binding sites for rd22BP1 and ATMYB2 that are ~40 nucleotides apart. The two binding sites function in the dehydration-responsiveness of the promoter.</td>
<td>Abe et al., 1997)</td>
</tr>
</tbody>
</table>

(a) A similar set of interactions to the ones described in this table, among bHLH, MYB, and WD-40 repeat proteins, probably occurs among factors that control the anthocyanin pathway in Arabidopsis and petunia, although the protein-protein interactions have not been described in detail yet. In Arabidopsis, these proteins are TT2 (a MYB protein), TT8 (a bHLH factor), and TTG1 (a WD-40 repeat protein) (Nesi et al., 2000; Nei et al., 2001); the petunia proteins include AN2 (MYB), AN11 (WD-40), and AN1 (bHLH) (de Vetten et al., 1997; Quattrocchio et al., 1999; Speit et al., 2000).
Macdonald et al., 1999; Fraser et al., 2000; Gönczy et al., 2000; Maeda et al., 2001).

Such view is also appropriate when considering Arabidopsis transcriptional regulation at a global level in a cellular and organismal context, for whose understanding several of those functional maps would be required: the genome-wide transcriptome map, the interactome and the phenome of the transcriptional regulators, as well as other “-ome” maps not previously considered, such as the promotorome. Intrinsic to this view is the realization that none of these different “-ome” maps would lead, in isolation from the others, to a comprehensive or even logical understanding of transcriptional regulation and of its role as a major determinant of cellular and organismal functions and phenotypes.

To generate these functional maps, the systematic investigation of transcription factor function and transcriptional regulation in Arabidopsis can be pursued with a variety of tools for functional genomic analyses, including reverse genetics methods, gene expression profiling experiments, and protein-protein interaction screens (reviewed in Riechmann and Ratcliffe, 2000). Whereas the availability of the Arabidopsis genome sequence allows us to compile lists of proteins that are involved in the regulation of transcription, and of putative promoter and cis-acting sequences, a global understanding of this process is still in its infancy. However, somewhere along the way of generating these functional maps, and once a sufficient amount of data has been collected, it should be possible to start decoding the “language” of transcriptional control, and to eventually be able, for instance, to build synthetic promoters directing gene expression in novel, designed spatial and temporal patterns (for an example of an initial attempt to design an artificial expression cassette in plants simply by statistical analysis of nucleotide sequences, see Sawant et al., 2001).

6.1 The transcription factor phenome map.

The number of Arabidopsis transcription factors that have been functionally characterized is still small, approximately 10% of the total (an incipient phenome; Table 1). Most of these genes were characterized through the traditional genetic approach, whereby genes are first defined by a mutant phenotype and then isolated. For the majority of these transcriptional regulators, functional characterization is limited to the description of phenotypic differences between mutant and wild-type plants, and determination of their expression pattern, but there is very little knowledge on their modes of action, that is, on the genes that they regulate (the transcriptome map) and on the mechanisms that they use to achieve that regulation (the interactome and the promotorome). As a result, the dynamic relationship between the genome, the transcriptional regulators, and the transcriptome, remains largely uncharacterized.

Different reverse genetics strategies are or can be used in plants to generate and isolate mutants in known genes: T-DNA or transposon insertion mutagenesis (Krysan et al., 1999; Maes et al., 1999; Parinov and Sundaresan, 2000; Young et al., 2001), fast neutron deletion mutagenesis (Li et al., 2001b), targeted screening for induced local lesions (TILLING) (Colbert et al., 2001), and DNA/RNA oligonucleotide-mediated site-directed mutagenesis (Oh and May, 2001). In addition, gene function can be inhibited by RNA interference (RNAi) or by virus–induced gene silencing (Baulcombe, 1999; Chuang and Meyerowitz, 2000; Levin et al., 2000; Hammond et al., 2001; Wesley et al., 2001). All these methods have been extensively reviewed and will not be discussed here. They are being used in several large-scale reverse genetics efforts to characterize the function of Arabidopsis transcription factors and chromatin-related proteins (for example, Meissner et al., 1999) (see also http://Ag.Arizona.Edu/chromatin/chromatin.html).

Probably the two main difficulties for generating a comprehensive phenome of Arabidopsis transcriptional regulators are the finite number of assays in which the mutants can usually be screened, and the existence of functional redundancy or overlap among different genes (Riechmann and Ratcliffe, 2000). Many of the Arabidopsis knockout mutants thus far isolated through reverse genetics approaches, in transcription factor genes as well as in genes of other classes, do not exhibit obvious morphological phenotypic alterations (Meissner et al., 1999; Bouche and Bouchez, 2001). This finding parallels what has been observed in other eukaryotic organisms, such as C. elegans, Drosophila, and yeast, in both forward and reverse genetics screens (for an overview, see Thatcher et al., 1998). For instance, the systematic analysis by RNAi in C. elegans of 4,590 genes (contained in chromosomes 1 and 3) only revealed mutant phenotypes in ~14% of the cases (Fraser et al., 2000; Gönczy et al., 2000). However, it is likely that Arabidopsis mutants in “silent” or “nonessential” transcription factor genes (that is, they show no overt phenotype) might in fact reveal informative phenotypes when tested in comprehensive assays to characterize their physiology, metabolism, etc. (for an example of the use of metabolome data to reveal the phenotype of silent mutations in yeast, see Raamsdonk et al., 2001). For those genes that are involved in the plant’s response to the environment, either biotic or abiotic, mutant phenotypes might not be revealed unless specific environmental conditions are used in the experiments. However, the assumption that
if a gene is expressed or induced under a particular set of conditions then that gene is important for the organism’s growth or survival in those conditions, should be taken with some caution: in yeast, there appears to be little correlation between the two when large sets of genes are considered (Winzeler et al., 1999). Lastly, detection of slightly deleterious effects caused by mutations in “silent” genes might require multigenerational competition studies in which fitness can be assessed, as shown in Arabidopsis (actin genes) and in yeast (Gilliland et al., 1998; Thatcher et al., 1998; Winzeler et al., 1999). The analysis by deletion of 2,026 genes in yeast indicated that ~80% of them were nonessential for viability, but 40% of those silent deletants showed impaired growth in a simple competitive assay (Winzeler et al., 1999).

The extent of functional redundancy among related Arabidopsis transcription factors has been illustrated by several recent studies on factors from different groups, such as the MADS, GARP, YABBY, and GRAS gene families. MADS-box genes that act redundantly include: AP1, CAULIFLOWER (CAL), and FRUITFULL (FUL), in the control of floral meristem identity (Bowman et al., 1993; Kempen et al., 1995; Ferrándiz et al., 2000); the SHATTER-PROOF genes (SHP1 and SHP2), which are required for proper development of the fruit-valve margin (Liljegren et al., 2000); and the SEPALLATA genes (SEP1, SEP2, and SEP3), which are cofactors or interactors for the floral organ identity genes AP1, AP3, PI, and AG (see above, and: Pelaz et al., 2000; Honma and Goto, 2001; Pelaz et al., 2001a; Pelaz et al., 2001b). The redundancy among AP1, CAL, and FUL in specifying floral meristem identity is partial. ap1 plants show a mutant phenotype (a partial conversion of flowers into inflorescences and a disruption of sepal and petal development), whereas a mutation in CAL results in a mutant phenotype only when combined with an ap1 allele (Bowman et al., 1993; Kempen et al., 1995). ap1 cal mutant plants show a complete conversion of the floral meristems into inflorescence meristems (Bowman et al., 1993). In other words, AP1 can completely compensate for the loss of CAL function, but CAL can only compensate for part of AP1 activity. A mutation in FUL does not alter floral meristem identity in the presence of a functional copy of AP1 or CAL (Ferrándiz et al., 2000).

The SEP genes appear to have largely overlapping, although not identical, functions: the triple sep1 sep2 sep3 mutant shows a clear conversion of petals, stamens, and carpels, to sepals, whereas single or double sep mutants exhibit more subtle phenotypic alterations (Pelaz et al., 2000; Pelaz et al., 2001a) (for a review on the SEP genes and the ABC model of flower development: Jack, 2001). Similarly, only the shp1 shp2 double mutant, and not the single mutants, shows drastic phenotypic effects, in this case fruit that fails to dehisce (Liljegren et al., 2000).

Another example of related genes that act redundantly is provided by KANADI1 (KAN1) and KANADIZ2 (KAN2), which participate in the establishment of polarity in Arabidopsis lateral organs by determining abaxial cell fate (Eshed et al., 2001). KAN1 and KAN2 are members of the GARP family of plant-specific transcription factors, and they form part of a monophyletic group within the family (Eshed et al., 2001; Kerstetter et al., 2001). In fact, the genetic mechanism or network that controls lateral organ polarity in Arabidopsis appears to consist of multiple transcription factors from different gene families, with the corresponding genes within each group acting, at least in part, in a functionally overlapping manner (see below, and Eshed et al., 2001). Finally, GAI and RGA, which are highly related members of the GRAS gene family, have partially redundant functions as negative regulators of the gibberellin (GA) signaling pathway (Dill and Sun, 2001; King et al., 2001). In summary, situations of overlapping or partially redundant gene function among related genes are frequent within the different Arabidopsis transcription factor families (for general discussions on gene function after duplication, and on genetic redundancy and how it might be maintained by selection, see: Thomas, 1993; Cooke et al., 1997; Massingham et al., 2001).

Furthermore, in addition to redundancy resulting from the incomplete functional divergence between highly related (duplicated) members of a gene family, it can also arise from functional convergence of more distantly related genes (reviewed in: Pickett and Meeks-Wagner, 1995; Cooke et al., 1997). For instance, two divergent forkhead transcription factor genes from C. elegans, pes-1 and fkh-2, are partially redundant in embryonic development (Molin et al., 2000). Inactivation of pes-1 or fkh-2 alone caused no apparent phenotypic alteration during embryogenesis, whereas inactivation of both genes severely disrupted it. The functional association between pes-1 and fkh-2 was investigated because of the similarity in their expression patterns, but not because of sequence homology; pes-1 and fkh-2 belong to different clades within the forkhead gene family (Molin et al., 2000). The C. elegans genome contains 15 different forkhead genes (Riechmann et al., 2000), and the expression patterns had been determined for all of them. This example illustrates the limitations of sequence analysis as a tool to explore genetic redundancy. In addition, it suggests another reason why the complete functional characterization of the Arabidopsis complement of transcription factors will require the determination of the expression patterns of all of its members. Although the extent of this type of redundancy (genes that belong to the same family, but to different clades within it, and yet have the same or overlapping functions) in Arabidopsis is unknown, there is evidence that it exists. For example, AINTEGUMENTA (ANT) acts redundantly with APETALA2 (AP2) to repress AG expression in cells of...
the second whorl of developing Arabidopsis flowers (Krizek et al., 2000). Both ANT and AP2 are AP2/ERF proteins, but they belong to different clades within the AP2 subfamily.

Last, functional redundancy can also exist between genes of different classes or families (i.e., with distinct molecular functions), for instance if they form part of independent pathways controlling the same process. An example of this type is provided by members of the KANADI and YABBY gene families, involved in the determination of abaxial polarity in lateral organs (Eshed et al., 1999; Bowman et al., 2001; Eshed et al., 2001). For example, CRABS CLAW (CRC), the founding member of the YABBY family (Bowman and Smyth, 1999; Bowman, 2000), and KAN1 participate in the determination of abaxial polarity in the carpels (Eshed et al., 1999; Bowman et al., 2001). In contrast to redundancy between duplicated genes, it is not possible to predict functional redundancy between unrelated proteins by simply analyzing the genome sequence. Rather, these cases of functional overlap will be uncovered through classic mutagenesis screens for enhancers of a particular mutant phenotype (for example: Eshed et al., 1999; Bowman et al., 2001; Eshed et al., 2001), and as a result of genome-wide analyses of gene expression designed to identify the targets of different transcription factors (see below).

6.2 The transcriptome and promoterome maps.

Two of the different “-ome” maps are essential to understand transcripational regulation at a molecular, genome-wide level and, ultimately, to explain the basis for the effects that differential gene expression has on the functions and phenotypes of cells and organisms: the transcriptome and the promoterome maps. The transcriptome can be viewed as the collection of transcripts that are expressed from the genome at any particular temporal and physiological instance, considering both transcript identity and abundance (i.e., a description both qualitative and quantitative). The promoterome, as defined in this chapter, would consist of all the promoters and cis-acting elements in a genome, and of their interactions with the complement of transcriptional regulators.

6.2.1 DNA microarrays.

A comprehensive characterization of the Arabidopsis transcriptome in its multiple forms can be achieved using DNA microarray technologies, which allow the parallel monitoring of the expression of thousands of genes and, eventually, of the complete Arabidopsis genome (for reviews on DNA microarrays and gene expression: Lockhart and Winzeler, 2000; Richmond and Somerville, 2000; Young, 2000; Altman and Raychaudhuri, 2001; Schulze and Downward, 2001) (for information on microarray resources, see: http://www.arabidopsis.org/links/microarrays.html). Currently, the expression of up to ~8,500 different Arabidopsis genes, or approximately one third of the genome, has been monitored in DNA microarray experiments to generate catalogues of genes that are expressed in response to particular stresses or stimuli, or in certain tissues or developmental processes (Table 7). These early studies have included the response to different nutrient concentrations, to drought and cold stresses, to wounding and insect feeding, the disease response, and light-related processes, such as the circadian clock and phytochrome A signaling (Table 7). The most extensive dynamic reprogramming of the expression of the genome has been observed upon light stimulus or in light-related processes (Harmer et al., 2000; Schaffer et al., 2001; Tepperman et al., 2001) (Table 7). For instance, the analysis of circadian changes in the mRNA levels of more than 8,000 genes reveals how differential gene expression underlies many of the physiological changes that the plant undergoes in its daily life cycle. The expression of genes implicated in photosynthesis, in phenylpropanoid biosynthesis, in lipid modification, and in carbon, nitrogen, and sulfur pathways was found to be regulated by the circadian clock, and a physiological explanation can be reasoned for it: to prepare for light-harvesting, for protection against UV light, to increase chilling-resistance at night, and to coordinate the metabolism of the plant with its environment (Harmer et al., 2000).

Catalogues of expressed genes provide insights into a variety of biological processes. However, elucidating the relationship between the transcriptome, the promoterome, and the complement of transcription factors, to eventually understand the logic of transcription, requires additional types of genome-wide analyses and experiments, such as the following.

6.2.2 Comparative analysis of promoter sequences of...
<table>
<thead>
<tr>
<th>Table 7. Arabidopsis DNA microarray studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment type</strong></td>
</tr>
<tr>
<td>Technique development</td>
</tr>
<tr>
<td>Gene expression in different organs</td>
</tr>
<tr>
<td>Gene expression in developing seeds</td>
</tr>
<tr>
<td>Gene expression in different organs</td>
</tr>
<tr>
<td>Wounding and insect feeding response</td>
</tr>
<tr>
<td>Response to high or low nutrient (nitrate) concentration</td>
</tr>
<tr>
<td>Drought and cold stresses</td>
</tr>
<tr>
<td>Systemic acquired resistance</td>
</tr>
<tr>
<td>Defense responses</td>
</tr>
</tbody>
</table>

*continued on following page*
<table>
<thead>
<tr>
<th>Gene expression in mpk4 mutant</th>
<th>cDNA</th>
<th>~9,000</th>
<th>Inactivation of MAP kinase 4 results in constitutive systemic acquired resistance. Microarray analysis was used to confirm at the transcriptional level this observation, and to identify genes (16) that are downstream of NPK4 signaling. Putative WRKY binding sites were identified in the promoters of those genes, along with another conserved cis-element. (Petersen et al., 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circadian clock</td>
<td>Oligonucleotide (GeneChip)</td>
<td>~8,300</td>
<td>Circadian changes were observed for up to 6% (or &gt;450) of the genes represented on the microarray. Cluster analysis revealed groups of coregulated genes that form part of metabolic and light response pathways. Many of these regulatory events could be interpreted in the light of the daily physiological cycle of the plant (anticipation of dawn and dusk, coordination of carbon, nitrogen, and sulfur pathways, etc.). Comparison of promoter sequences identified a novel regulatory motif (the &quot;evening element&quot;) involved in circadian clock regulation. The activity of that element was demonstrated experimentally using reporter gene fusions. Incidentally, an &quot;evening element&quot;-like sequence is found in the promoter of TOC1, a gene that is regulated by the circadian clock and that is itself a component of its machinery. The cis-element is required for the correct TOC1 expression, and is recognized by the MYB-related proteins CCA1 and LHY, which are also components of the circadian clock and negatively regulate TOC1 (Alabadi et al., 2001). (Harmer et al., 2000)</td>
</tr>
<tr>
<td>Diurnal expression, circadian clock</td>
<td>cDNA</td>
<td>~7,800</td>
<td>Identification of diurnal and circadian regulated genes (11% and 2% of the genes represented on the microarray, respectively). The regulation of some of those genes was explained in the light of the daily physiological cycle of the plant. (Schaffer et al., 2001)</td>
</tr>
<tr>
<td>Phytochrome A signaling</td>
<td>Oligonucleotide (GeneChip)</td>
<td>~8,300</td>
<td>Characterization of the temporal transcriptional response to far-red light (using wild-type and phy:4 mutant seedlings). Up to 10% of the genes represented on the array showed differential expression. Transcription factors comprise a substantial fraction (~34%, or 21 genes) of the early genes (those that respond to the signal within 1 hour, 62 genes in total), whereas genes involved in photosynthesis and general metabolism comprise a majority (~35%; 266 genes) of the late-response genes (3-24 hours, 750 genes in total). (Tepperman et al., 2001)</td>
</tr>
<tr>
<td>Oxidative stress</td>
<td>cDNA</td>
<td>~7,800</td>
<td>Identification of genes that are differentially expressed in response to oxidative stress (a total of 175, 113 of which are induced and 62 repressed). (Desikan et al., 2001)</td>
</tr>
</tbody>
</table>

(a) As of October 2001, this list does not intend to be comprehensive, but rather to illustrate different types of studies performed using Arabidopsis microarrays, and summarize those of most relevant results. Microarray experiments have also been used, for example, to characterize the alterations in transcript levels in mutants in genes that do not code for transcriptional regulators themselves, such as BIG/DOCI (which codes for a callosin-like protein required for polar auxin transport), or ABH1 (which codes for an mRNA cap binding protein) (Gil et al., 2001; Hugouvieux et al., 2001).
genes with similar expression profiles.

The results of multiple DNA microarray experiments can be combined and analyzed together using clustering techniques, by which those genes that show similar expression patterns across the set of experiments are identified and grouped (reviewed in: Sherlock, 2000; Quackenbush, 2001; Raychaudhuri et al., 2001). The set of experiments to be compared might consist of different cell types or tissue samples, different physiological conditions, or might characterize over a time course the transcriptional response to a given stimulus. One assumption underlying the clustering analysis of time course experiments is that genes with highly related expression profiles might be regulated by the same mechanism. Thus, once groups of co-regulated genes are established, their promoter sequences can be compared to identify common cis-acting elements. The success of the approach is determined, at least in part, by the structural organization (i.e., size and complexity) of the regulatory regions, and it has proven particularly fruitful in yeast (Cho et al., 1998; Roth et al., 1998; Spellman et al., 1998; Tavazoie et al., 1999; Wolfsberg et al., 1999; Gasch et al., 2000; Hughes et al., 2000a; Lyons et al., 2000; Jakt et al., 2001). In animals, in which complete regulatory regions are difficult to delimit from sequence information alone, and cis-acting elements might be distributed over very long distances, the analyses might have to be restricted to the proximal promoter sequences (for example, Livesey et al., 2000). In contrast, as discussed above, plant regulatory regions are more similar to those of yeast, in that they are often completely encompassed within a few hundred base pairs upstream from the transcription start site. The comparison of the promoter sequences of a group of Arabidopsis genes that are co-regulated by the circadian clock, and that in the experiment showed the highest level of expression near the end of the subjective day, identified a novel motif that is conserved among those promoters, and that was then experimentally shown to mediate their regulation (Harmer et al., 2000) (Table 7). Similarly, a group of Arabidopsis genes that co-regulate with PR-1 over a series of systemic acquired resistance (SAR) inducing or repressing conditions was identified, and their promoters searched for the presence of known cis-elements for transcription factors. W boxes, the binding site for WRKY proteins, were the only known cis-element that was present in all the promoters, suggesting that WRKY transcription factor(s) participate in the control of the PR-1 regulon (Maleck et al., 2000). This latter example also illustrates that the identification of common elements in the upstream regions of co-regulated genes will usually not be sufficient, in the absence of other information, to pinpoint the identity of the specific regulators, because the majority of the Arabidopsis transcription factors form part of multigene families (Table 1), in which different members have related or similar DNA-binding specificities. The same limitation applies to genome-wide searches for transcription factor binding sites that are carried out without reference to expression data (for an example in Arabidopsis, Du and Chen, 2000). Such searches can lead to the identification of potential downstream genes, especially if the target sites for the DNA-binding protein(s) or complex(es) of interest are well characterized, but additional experimentation is usually required to establish the association between the identified elements and the transcription factor(s) under study (for an example in yeast, Zhong et al., 1999).

6.2.3 Phylogenetic footprinting.

A valuable approach to identify unknown cis-regulatory regions and elements, and that can be applied at a genome-wide scale, is phylogenetic footprinting: sequence comparisons across phylogenetically related species that reveal conserved cis-elements in the non-coding regions of homologous genes. Phylogenetic footprinting is based on the observation that regulatory regions are more conserved throughout evolution than regions that do not have a function that is dependent on their sequence (for reviews: Gumucio et al., 1996; Duret and Bucher, 1997; Hardison et al., 1997; Fickett and Wasserman, 2000). A critical factor for the success of the phylogenetic footprinting method is the choice of species for the comparative analysis: at a genome-wide scale, they should be similar enough so that most sequences can be aligned with the corresponding ortholog(s), but distant enough so that non-functional sequences have diverged by accumulating mutations at neutral positions. Logically, the frequency of detectable conserved elements in the non-coding regions of orthologous genes decreases as species separated by increasing evolutionary distances are compared (Duret and Bucher, 1997). Identifying the optimal species (or group of species) for phylogenetic footprinting might require surveying several species within the corresponding genus or family (for an example of such survey in Saccharomyces, see Cliften et al., 2001). In fact, it appears that the most comprehensive and meaningful results will be obtained in comparisons that include several species of different evolutionary distances (Duret and Bucher, 1997; Cliften et al., 2001). The use of groups of species might be particularly important for genome-wide phylogenetic footprinting, because the optimal evolutionary distance for comparison might vary across genes. The phylogenetic footprinting
method has been used to define regulatory elements by comparisons between human and mouse sequences and among mammals, and between C. elegans and C. briggsae, among other organisms (Gumucio et al., 1996; Thacker et al., 1999; Loots et al., 2000; Wasserman et al., 2000). Comparisons of the promoters of the CHALCONE SYNTHASE and AP3 genes across different cruciferous plant species has demonstrated the value of phylogenetic footprinting as a basis to functionally analyze Arabidopsis cis-regulatory regions (Koch et al., 2001), although the ideal species or group of species for that type of comparison with Arabidopsis at a genome-wide scale still needs to be identified.

Phylogenetic footprinting, however, will provide only a partial description of the promoterome, because only those elements that maintain similar functions across the species compared are likely to be conserved. Alterations in gene expression are an important mechanism of evolutionary change, and regulatory elements and functional features that arose after the divergence of the species used in the comparison might not be identified in the analysis (for variations of the phylogenetic footprinting method that, combined with experimental data, try to overcome these limitations, see Gumucio et al., 1996). Furthermore, there are instances in which a cis-region might maintain a given regulatory function despite considerable sequence variation.

6.2.4 Inducible activation of transcription factor activity.

The identification of the downstream genes of the many transcriptional regulators encoded by the genome is a necessary step to define the networks of gene activity that occur in a cell, tissue, or organism. The combination of DNA microarray technology with systems for the inducible activation of transcription factor function offers a way to dissect regulatory programs. The activity of transcription factors can be transcriptionally or posttranslationally regulated, using inducible gene expression systems or generating protein fusions to steroid-binding domains, such as the glucocorticoid receptor (GR) (Aoyama, 1999; Picard, 2000; Zuo and Chua, 2000). The advantage of using posttranslational regulation is that simple direct and indirect effects of transcription factor activity can be separated by using inhibitors of protein synthesis (for examples in Arabidopsis: Sablowski and Meyerowitz, 1998; Wagner et al., 1999; Samach et al., 2000; Sakai et al., 2001). Thus, an experiment to identify at a genome-wide scale the target genes of a particular transcription factor would consist of generating transgenic plants expressing a fusion of the factor to a steroid binding domain (ideally, in a mutant background in which the native transcription factor gene is inactivated), applying the hormone to the tissues under study (ideally, those in which the endogenous gene would be normally active), both in the presence and in the absence of an inhibitor of protein synthesis (cycloheximide), and following the effects on mRNA accumulation over time using DNA microarrays. Direct posttranslational regulation by fusion to GR has already been engineered for several plant transcription factors, including the maize bHLH R protein (Lloyd et al., 1994), the Arabidopsis homeodomain proteins ATHB-1, ATHB-2, and KNAT2 (Aoyama et al., 1995; Oghishi et al., 2001; Pautot et al., 2001), CON- STANS, a zinc finger transcriptional regulator (Simon et al., 1996; Samach et al., 2000), the MADS domain protein AP3 (Sablowski and Meyerowitz, 1998), LFY (Wagner et al., 1999), and ARR1, a GARP transcription factor of the ARR-B subclass (Sakai et al., 2001). However, the technique might not be universally applicable, since some transcription factor-GR fusion proteins might be inactive, or constitutively active in the absence of the hormone.

6.2.5 Genome-wide maps of in vivo DNA binding by transcription factors.

An alternative approach to identify transcription factor downstream targets has been recently developed in yeast, combining chromatin immunoprecipitation and DNA microarrays. The underlying assumption is that transcription factors bind to the promoters or regulatory regions of the genes whose expression they control. In this method, proteins are crosslinked to genomic DNA in living cells using formaldehyde. The DNA that is specifically crosslinked to the protein of interest is then enriched by immunoprecipitation, amplified by PCR, and labeled for its use as a probe in dual-color microarray experiments (the corresponding control consisting of a sample DNA that was not enriched). The microarray used for the hybridization contains all the intergenic regions of the yeast genome, and might also contain the corresponding ORFs (Ren et al., 2000; Iyer et al., 2001; Lieb et al., 2001). This approach has been used to identify the binding sites for several transcriptional regulators in the yeast genome (Ren et al., 2000; Iyer et al., 2001; Lieb et al., 2001), and to study the targeted recruitment of the yeast histone acetylase Esa1 (Reid et al., 2000). In a more global study, the binding by the nine yeast transcription factors that are known...
to regulate the cell cycle (Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5, and Ace2) was analyzed, showing that these factors form themselves a circular network of serial regulation (Simon et al., 2001). The results of all these experiments are also a testimony to the complexity of transcription, and to how much we still have to learn and to explain when considering the regulation of the expression of eukaryotic genomes as a whole. First, Gal4, SBF, MBF, and Rap1 were all found to bind preferentially to potential promoter regions, despite the fact that consensus binding site sequences for all of them are distributed all over the yeast genome (Ren et al., 2000; Iyer et al., 2001; Lieb et al., 2001). This biased recognition of binding sites suggests the existence of a superimposed level of regulation that might mark or distinguish regulatory regions from coding sequences, a component of which might be chromatin structure. The distribution of binding sites for a given transcription factor in the promoters of the genes that such factor regulates is not random either, suggesting the existence of, and constraints in, long-range interactions with other components of the transcription machinery. For example, Rap1 binding sequences were found to occur more often in tandem and at a certain upstream distance (250-450 bp), and to be located preferentially on the minus strand relative to the corresponding open reading frame (Lieb et al., 2001). In addition, not every promoter bound by, for example, SBF and MBF contains recognizable consensus sites (Iyer et al., 2001), indicating the existence of additional sources for specificity in transcription factor activity in vivo. Once again, the structural similarity between regulatory regions in yeast and plants suggests that the technique might in principle be applicable to Arabidopsis, provided that the corresponding experimental protocols (for crosslinking and immunoprecipitation) can be established, and with the caveat of the multicellular nature of plants (see below).

An alternative technique for the genome-wide identification of in vivo target loci has been developed in Drosophila, which makes use of E. coli DNA adenine methyltransferase (Dam). In this method, named DamID, a protein fusion between a chromatin protein of interest and Dam is expressed at low levels in Drosophila cells (in culture, or in the whole fly) (van Steensel and Henikoff, 2000; van Steensel et al., 2001). This leads to the methylation of the GATC sequences that might occur in the genome surrounding the binding sites of the target protein. Methylation events, at least in Drosophila, and it is not yet known if the approach will work for proteins that bind as single molecules (or as dimers) to specific, short cis-elements (as many transcription factors do), since all the experiments reported so far involved potential cooperative binding that could ‘coat’ a region of DNA (for more information and discussion on DamID, see: http://blocks.fhrc.org/DamID). If these technical hurdles can be overcome, chromatin profiling by DamID could represent a valuable alternative to immunoprecipitation-based methods.

The genome-wide maps of in vivo protein-DNA association will contribute to clarify a longstanding unanswered question in transcriptional regulation: the correlation between the DNA binding properties of the transcription factors (i.e., affinities, which are measured in vitro), and their effects on transcription in vivo (for a discussion on this topic, see Biggin and Tjian, 2001). Ultimately, quantitative studies and information will be needed to understand the transcriptional code, and to be able to model transcriptional regulation, both in vivo and in silico.

The identification of the target genes for the many transcriptional regulators encoded by the Arabidopsis genome, the compilation of lists of genes that are differentially regulated (activated or repressed) in particular biological processes, and, most importantly, the integration and combined analysis of these large genome-wide data sets, will eventually define the networks by which transcription factors act, and the pathways downstream of them. Questions that are difficult to address in gene-by-gene studies can now be considered: to what extent different environmental responses, or distinct developmental pathways, share effector mechanisms (that is, the same target genes)?; how many different patterns of expression are triggered by a particular stimulus, and how are those differences achieved molecularly (that is, the complexity of response pathways at the level of the regulation of effector-or “realizator”- genes)? Combined with the characterization and analysis of the promoterome, such studies would lead to understand the transcriptional code, and eventually to rationally manipulate transcriptional regulation. These genome-wide studies would also assess the degree of connectivity among regulatory networks (and thus start defining the “networkome”). Furthermore, the identification of (direct) target genes that are shared by different transcription factors might also provide clues about what regulators act together, both molecularly (interacting proteins, or proteins binding to the same promoters but not interacting directly), and at the genetic level (either related genes that are (partially) redundant, or genes that
form part of different pathways that control the same process). Such analysis would complement, and guide, other genetic studies to characterize overlapping functions among transcriptional regulators (for example, by identifying pairs or groups of genes to be analyzed in double, or multiple, mutant combinations; see above).

Last, it should be noted that the characterization of the Arabidopsis transcriptome, and its explanation in terms of transcription factor activity, faces one challenge not encountered in yeast, in which many of these types of studies have been pioneered: the multicellular nature of plants. The profiles of gene expression of different cell types and cells are logically different, and so can be their responses to the activity of particular transcription factors. Thus, in many instances the transcriptome that is characterized is in fact the average of those of the different cell types included in the study. It is still technically challenging, although not impossible, to achieve cellular resolution in genome-wide studies in multicellular organisms.

6.3 The transcription factor interactome map.

Genome-wide analyses of protein-protein interactions in eukaryotes have been pioneered for the proteomes of yeast and C. elegans using the yeast two-hybrid system. The general merits and problems of the approach in genome-wide screens have been discussed elsewhere (Riechmann and Ratcliffe, 2000; Hazbun and Fields, 2001; Legrain et al., 2001). For transcription factors, the two-hybrid system presents the added complication that their use as “baits” often requires the preparation of specialized constructs in which the sequences coding for activation domains have been removed (which is not a trivial hurdle if hundreds of proteins with presumed, but uncharacterized, activation domains have to be analyzed). This is because the read-out of the system consists on the transcriptional activation of reporter genes as a result of the interaction between the “bait” and “prey” fusion proteins. However, modifications of the two-hybrid system have already been devised (based on repression rather than on activation of transcription) that should be applicable for identifying interactions with transactivator proteins (for example: Hirst et al., 2001).

At present, there is very little data on the Arabidopsis transcription factor interactome, and only a small number of direct interactions between different plant transcription factors have been described (see above, and Singh, 1998). In addition to transcription factors directly interacting among themselves or with other components of the transcription machinery, they can also interact with other types of proteins, thus expanding the networks that would form the transcription factor interactome. Such interactions with proteins of other classes can be mechanistically important for the control of transcription, and they can also provide the link between transcription factor activity and signal transduction pathways, as for example in light- and disease-responses.

Arabidopsis perceives light using different types of light-absorbing photoreceptors, such as phytochromes (phyA through phyE), which absorb red and far-red light, and cryptochromes (cry1 and cry2), which absorb blue and UVA light (for review, Nagy and Schäfer, 2000). Phytochrome- and cryptochrome-mediated light responses involve differential regulation of gene expression. PIF3 is a transcription factor of the bHLH family that is involved in phytochrome signal transduction, in particular signaling by phyB (Ni et al., 1998; Halliday et al., 1999). PIF3 binds to a cis-element present in several light-regulated promoters, and phyB (which is translocated to the nucleus in a light dependent manner: Kircher et al., 1999; Yamanuchi et al., 1999) reversibly binds to DNA-bound PIF3 upon the light-triggered conversion to its biologically active form (Ni et al., 1998; Martinez-Garcia et al., 2000; Zhu et al., 2000). Thus, phytochromes might act as light-switchable components of transcription complexes, and their interaction with transcription factors might provide a short, direct pathway from light perception to photoresponsive nuclear gene expression (Martinez-Garcia et al., 2000).

Another Arabidopsis transcription factor involved in light-mediated responses is HY5, which controls the photomorphogenic development that is undertaken by seedlings grown in the light. HY5 is a bZIP protein that binds to a cis-element present in several light-responsive promoters (Oyama et al., 1997; Chattopadhyay et al., 1998a). The regulation of HY5 activity by light involves its interaction with COP1, a RING-finger protein with WD-40 repeats whose subcellular localization is light-dependent (nuclear in the dark and cytoplasmic in the light), and that might target HY5 for proteasome-mediated degradation in the nucleus (Osterlund et al., 2000). In this case, it is the COP1 protein that interacts with light-activated photoreceptors (cryptochromes), which repress COP1 activity, thus permitting HY5 accumulation and induction of gene expression (Wang et al., 2001).

The Arabidopsis ankyrin repeat-containing protein NPR1 has been shown to interact with some bZIP transcription factors of the TGA subfamily, which have been implicated in the activation of salicylic acid (SA)-responsive genes (Zhang et al., 1999). NPR1 is required for the induction of systemic acquired resistance (SAR) responses, such as the expression of pathogenesis-related (PR)
genes, and it has been shown to act downstream of SAR-inducing agents (SA and avirulent pathogens) (Cao et al., 1997). NPR1 enhances the DNA binding activity of the interacting TGA bZIP proteins, and the in vivo relevance of the protein–protein interaction is demonstrated by the observation that point mutations in NPR1 that abolish its function also disrupt the interaction with the TGA factors (Després et al., 2000; Zhou et al., 2000). NPR1 is localized in both the cytoplasm and the nucleus of unstimulated cells, but concentrates in the nucleus in response to SA and, in fact, nuclear localization of NPR1 is required for PR gene expression (Kinkema et al., 2000). Furthermore, using an in vivo protein fragment complementation assay, based on association of reconstituted murine dihydrofolate reductase (mDHFR) with a fluorescent probe to detect protein–protein interactions, it has been shown that the interaction between NPR1 and the bZIP factor TGA2 is itself induced by SA and localized predominantly in the nucleus (Subramaniam et al., 2001). Thus, the interaction of transcription factor(s) with NPR1 provides a link between an SAR-inducing agent, SA, and the changes in gene expression that are associated with SAR. The mechanism by which the SA signal is transduced to NPR1 still remains to be determined, but additional two-hybrid screens have identified novel proteins of still uncharacterized function that interact with NPR1 and also accumulate in the nucleus (Weigel et al., 2001).

Another example of how an external trigger for a defense response might modulate gene expression through the interaction between a transcription factor and a protein of a different type is provided by the tomato AP2/ERF protein PtI4. PtI4 was first identified in a two-hybrid screen by virtue of its interaction with Pto, a protein kinase that confers resistance to *Pseudomonas syringae* carrying the corresponding avirulence gene, *AvrPto* (Zhou et al., 1997). PtI4 is phosphorylated by Pto, which enhances the binding of PtI4 to the GCC-box present in the promoter of PR genes (Gu et al., 2000).

In summary, it is clear from these different examples and from the published literature that a comprehensive description of the transcription factor “interactome” will encompass, in addition to the more than 1,500 transcriptional regulators encoded by the Arabidopsis genome, many other proteins from a variety of functional classes. It is also apparent that such comprehensive description will not be attained using only two-hybrid experiments, which will reveal only a subset of the interactions that occur in a cell, and that additional, alternative techniques are needed. The mDHFR-based protein fragment complementation assay mentioned above permits direct visualization (through spectroscopy, fluorescence-activated cell sorting, or fluorescence microscopy) of protein–protein interactions in living cells, and can be used to detect interactions in a quantitative manner and to follow the temporal pattern of the interaction (Subramaniam et al., 2001). Although the assay has been used in isolated protoplasts, and still needs to be developed to study interactions in whole plant tissues or organs, it represents a valuable alternative to the two-hybrid system for studying and dissecting signaling cascades, and could be developed into a high-throughput screening system for pathway and network mapping or for the identification of molecules that modify protein–protein interactions (Subramaniam et al., 2001). Additional techniques that will be useful to characterize the transcription factor interactome include fluorescence resonance energy transfer (FRET) by generating protein fusions to spectral variants of the jellyfish green fluorescent protein (GFP) (Gadella et al., 1999; Shah et al., 2001).

**Conclusion**

Genomics research, and functional genomics in particular, has often been hailed as the provider of a new paradigm in biology research; it has also been sometimes reviled by describing such type of research as consisting of little else than “fishing” experiments. Both opposing views are based on one common premise: the contrast between a frequently hypothesis-free genomics and the hypothesis-driven research that has been so much favored in molecular biology over the past decades. But genomics might not fit either one of these two disparate views. In many aspects, genomics bears many similarities with classic biology disciplines, such as genetics (in which the whole genome is blindly mutagenized at random to “fish” for interesting, novel phenotypes), and taxonomy and phylogeny (in which lists of elements are compiled and the relationships between them have to be established). It is in many instances the methodic collection of unanticipated data what allows the formulation of new hypothesis. However, if not radically different in concepts, genomics certainly changes the scales of biology research, and provides new dimensions to it. The relative simplicity of the *Arabidopsis thaliana* genome, together with the availability of many modern genetic and genomic research tools in that species, indicates that Arabidopsis is a premier organism to elucidate the complex logic of transcription at a genome-wide level in multicellular eukaryotes.
Acknowledgments

I wish to acknowledge my colleagues at Mendel Biotechnology for their input and work in our transcription factor genomics research program, as well as for their discussions, insight, and comments. I also wish to acknowledge the work of all those who participated in the Arabidopsis Genome Initiative and sequenced the Arabidopsis genome.

References


Transcriptional Regulation: a Genomic Overview


