EVOLUTION OF LIGHT-REGULATED PLANT PROMOTERS

Gerardo Argüello-Astorga and Luis Herrera-Estrella
Departamento de Ingeniería Genética de Plantas, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 629, 36500 Irapuato, Guanajuato, México

KEY WORDS: gene evolution, plant evolution, plastid-nuclear interactions, photoreceptors

ABSTRACT
In this review, we address the phylogenetic and structural relationships between light-responsive promoter regions from a range of plant genes, that could explain both their common dependence on specific photoreceptor-associated transduction pathways and their functional versatility. The well-known multipartite light-responsive elements (LREs) of flowering plants share sequences very similar to motifs in the promoters of orthologous genes from conifers, ferns, and mosses, whose genes are expressed in absence of light. Therefore, composite LREs have apparently evolved from cis-regulatory units involved in other promoter functions, a notion with significant implications to our understanding of the structural and functional organization of angiosperm LREs.

CONTENTS
INTRODUCTION ......................................................... 526
LIGHT REGULATION OF GENE TRANSCRIPTION: A BRIEF OVERVIEW ........................................ 527
Control of Gene Expression by Photoreceptors .............................................................. 527
Light-Responsive Promoters ............................................................................................. 527
The Phylogenetic-Structural Sequence Analysis ............................................................... 528
EVOLUTION OF LREs IN Lhcb PROMOTERS ......................................................... 529
The Proximal LRE of Lhcb 1 Genes .................................................................................. 529
Evolution of the Proximal Region of Lhcb1 Promoters ...................................................... 531
LREs in the Central Region of Lhcb1 Promoters ............................................................... 532
Evolution of the Central LRE of Lhcb1 Promoters ............................................................. 532
Delimitation of a LRE in Lhcb2 Promoters ....................................................................... 533
Evolution of Lhcb2 Promoters ......................................................................................... 533
EVOLUTION OF LREs IN rbcS PROMOTERS .............................................................. 534
The Light-Responsive Box II-III Region ............................................................................ 534
Evolution of the Box II-III Homologous Regions ............................................................... 536

525

1040-2519/98/0601-0525$08.00
INTRODUCTION

Photosynthesis is an ancient biochemical process that probably evolved 3500 million years ago (111). The nuclear genome of higher plants encodes proteins homologous to cyanobacterial proteins of photosynthesis; because these bacteria are the living descendants of the original photosynthetic organisms, plant genes encoding such proteins are among the oldest genes of eukaryotes.

To date, molecular evolution studies have concentrated on the coding sequences of gene families. The evolution of regulatory sequences, which determine where, when, and the level at which genes are transcribed, has been largely neglected. In the case of the photosynthesis-associated nuclear genes (PhANGs) from higher plants, interesting evolutionary aspects of the molecular mechanisms by which transcription is activated by light receptors (e.g., phytochrome) could be addressed through the comparative analysis of promoter sequences. For instance, why does light profoundly affect transcription of PhANGs in monocotyledonous and dicotyledonous plants, while PhANG promoters in conifers, ferns, and mosses are either light insensitive or, at most, weakly photoresponsive (4, 71, 89, 99, 129). The systematic comparisons of angiosperm and nonflowering plant PhANGs promoter sequences provide a unique opportunity to explore how a new regulatory function, light responsiveness, was incorporated into the promoters of a wide range of genes whose expression is coordinately regulated.

Besides its obvious relevance for evolutionary studies, a comparative analysis of the structure of photoregulated promoters can be useful to address other important issues in plant gene expression. Comparative analysis of PhANG upstream sequences may contribute to reduce the apparent diversity of light-responsive elements (LREs) by revealing concealed phylogenetic and structural
relationships between dissimilar promoter regions with analogous functions. Other important issues concerning the composition and functional organization of LREs in plant genes could also be uncovered by their analysis from an evolutionary perspective.

The purpose of this review is to address the phylogenetic and structural relationships between light-responsive promoter regions from orthologous and paralogous plant genes. We review LREs from a wide range of genes, explore their common dependence on specific phototransduction pathways, and analyze correlations between the composition of multipartite LREs and their overall functional properties.

LIGHT REGULATION OF GENE TRANSCRIPTION:
A BRIEF OVERVIEW

Control of Gene Expression by Photoreceptors

The responses of plants to light are complex: seed germination, de-etiolation of seedlings, chloroplast development, stem growth, pigment biosynthesis, flowering, and senescence (67). Most of these responses require changes in both chloroplast and nuclear gene expression, which are mediated by three major classes of photoreceptors: phytochromes, blue/UV-A light receptors, and UV-B light receptor(s) (2, 61, 100). Light-regulated genes may respond to more than one photoreceptor, thus allowing a finely tuned control of their expression when stimulated by light (123).

The best characterized light receptor is phytochrome (PHY), which exists in two photochemically interconvertible forms, Pr and Pfr, and is encoded by a small family of genes in angiosperms (42, 99, 100). PHY controls the expression of diverse genes at the transcriptional, posttranscriptional, and translational levels (43, 112, 123). Gene expression is regulated by at least three different signal transduction pathways activated by PHY: one depends on cyclic GMP (cGMP), which regulates genes such as those involved in anthocyanin biosynthesis in some species; a second pathway depends on calcium/calmodulin, which activates a subset of chloroplast-associated nuclear genes (17, 85); and a third signal pathway, which requires both calcium and cGMP, activates a subset of genes necessary for chloroplast development (e.g. the gene encoding ferredoxin NADP+ oxidoreductase) (18), and represses transcription of phy-A and the gene encoding asparagine synthetase (91).

Light-Responsive Promoters

In flowering plants, light regulation of nuclear genes occurs mainly at the transcriptional level, as demonstrated by nuclear run-on transcription assays and
promoter-reporter gene fusions (112, 122, 126). The most extensively studied light-responsive genes are those encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rbcS) and the chlorophyll a/b binding proteins (Lhc, formerly called Cab), both of which are a paradigm for PhANGs controlled by the calcium/calmodulin phototransduction pathway (17). Chalcone synthase (chs) genes, on the other hand, have been the model for genes involved in the anthocyanin biosynthetic pathway in some dicots and, hence, for genes targeted by the phytochrome signal pathway dependent on cGMP (17, 85). More recently, PHY-A genes have become a model for the diverse genes whose transcription is down-regulated by light and which are controlled by the third phytochrome signaling pathway, dependent on calcium and cGMP (91).

A plethora of cis-acting elements and protein factors presumably involved in transcriptional light responses have been identified (13, 16, 122, 126); however, conclusive evidence for an essential role in light responsiveness has been obtained for only a few of them (5, 41, 66, 136). Several general conclusions are possible: 1. No single conserved sequence element is found in all light-responsive promoters. 2. The smallest native promoter sequences, sufficient to confer light inducibility on heterologous minimal promoters, are multipartite regulatory elements that contain different combinations of cis-acting sequences. 3. Even the smallest known photoresponsive promoter regions, when examined in a heterologous context, display fairly complex responses, often retaining the dependence on light wavelength, developmental stage of chloroplasts, and tissue specificity of the native promoter. 4. Numerous protein factors bind sequences within photoregulated promoters, although their actual contribution to light-activated transcription remains uncertain (13, 122, 126, 128).

The Phylogenetic-Structural Sequence Analysis

A commonly used approach to identify regulatory cis-elements is to search for conserved DNA motifs within the promoter region of orthologous genes. Such analyses have been successful in identifying cis-acting elements involved in the light responsiveness of PhANGs, such as the G-box and I-box elements from rbcS genes (48) and the GATA motifs of Lhcb1 genes (44). However, regulatory elements showing sequence degeneracy are not easily recognized in comparative analyses.

Sequence heterogeneity of regulatory elements may be functionally overcome if multiprotein regulatory complexes facilitate binding to imperfect target sites (86, 135). Because conventional computer programs for DNA sequence comparisons can fail to detect evolutionarily related but structurally variable promoter regions with analogous functions, alternative approaches have been developed, such as the “phylogenetic-structural method” of sequence analysis (8, 10). This method is based on the search of “homologous” (rather than
“similar”) DNA sequences of a functionally characterized promoter. Two sequences are homologous when they share common ancestry, regardless of the degree of similarity between them (37). In this sense, the G-box elements of \( rbcS, chs, \) and \( Lhcb1 \) genes are not homologous but only similar because they have different evolutionary origins, whereas the \( rbcS \) I-G unit of solanaceous species is probably homologous to the \( rbcS \) X-Y promoter region of \( Lemna \), an aquatic monocotyledonous plant, despite their rather low overall sequence similarities (10).

The individual elements found within a multipartite cis-regulatory region are termed phylogenetic footprints (PFs); they share high conservation over a segment of 6 contiguous base pairs in alignments of orthologous upstream sequences and represent potential binding sites for transcription factors (53). A cluster of PFs whose arrangement (combination, spacing, and relative orientation) is conserved through a phylogenetic series of homologous promoter regions is termed a conserved modular arrangement (CMA). If a given CMA consistently correlates with experimentally defined light-responsive promoter regions, then it is called an LRE-associated CMA. In this method, sequence comparisons are made in a phylogenetically ordered fashion, because the overall structure of multipartite regulatory units tends to diverge in evolution. Thus, homologous genes from species belonging to the same plant taxon should be compared first. Comparisons with other taxa follow. This procedure allows us to discern how an ancestral multipartite regulatory module has changed in an evolutionary context and can establish phylogenetic relationships between promoter regions that are apparently unrelated by structural criteria. An example of this analysis applied to \( Lhcb1 \) light-responsive regions is presented in Figure 1a.

EVOLUTION OF LREs IN \( Lhcb \) PROMOTERS

The Proximal LRE of \( Lhcb1 \) Genes

Three types of chlorophyll \( a/b \) binding proteins are found in the major light-harvesting complex of photosystem II, encoded by \( Lhcb1, Lhcb2, \) and \( Lhcb3 \) (52, 60). Most \( Lhcb \) promoters that have been functionally analyzed are from the \( Lhcb1 \) gene family and typically lack introns. In \( Lhcb1 \) genes from dicotyledons, an LRE is located in the proximal promoter region of these genes. This LRE is characterized by three conserved GATA motifs spaced by 2 and 6 bp, respectively, which are located between the CCAAT and TATA boxes (44, 87). In Arabidopsis \( Lhcb1^\ast \) (formerly \( Cab2 \)), the \(-111 \) to \(-33 \) region is sufficient to confer a pattern of expression dependent on both PHY and a circadian clock, to a heterologous minimal promoter (6). A nuclear factor that specifically interacts with the triple GATA repeats found in this LRE was identified and named
CGF-1 (6). If the three GATA motifs are altered by site-directed mutagenesis, the interaction of CGF-1 is disrupted and responsiveness to PHY and part of the light-induced circadian oscillation of Cab2 expression are comprised (5). Mutational analysis of the Arabidopsis Lhcb1*3 (Cab1) gene promoter identified additional cis-acting motifs, that could be involved in the functionality of this LRE. Mutations in a 27-bp region upstream the CCAAT box, which is bound by the CA-1 factor (75a), drastically reduced the overall promoter activity and also eliminated PHY responsiveness (68). More recently, a Myb-related transcription factor that specifically interacts with a conserved sequence motif [AA(C/A)AAA TCT] within the CA-1 region was cloned (131). Transgenic Arabidopsis expressing an antisense mRNA for this factor (called CCA-1) showed reduced PHY induction of the endogenous Lhcb1*3 gene, whereas expression of other PHY-regulated genes, such as rbcS, was not affected; thus CCA-1 has a specific role in Lhcb1 photoregulation (131).

Evolution of the Proximal Region of Lhcb1 Promoters

At least five PFs are found within the first 110–130 bp of the Lhcb1 promoter sequences from dicotyledons. All these PFs bind defined proteins. The most distal PF is the CUF-1 binding site, a G-box-related sequence which functions as a general activating element (6). The second PF is the CCA-1 binding site (131). The third is the CCAAT-box, which is part of the binding site of Tac, a protein factor proposed to be involved in the regulation of Arabidopsis cab2 by the circadian clock (22). The fourth and fifth PFs are located in the CGF-1 binding region, comprising the three GATA motifs located upstream to the TA TA box. Comparative analyses indicate that these GATA elements are distinct from both the phylogenetic and the structural point of view. 1. GATA I is part of a PF whose consensus is GATAAGR (the I-box motif) (48), whereas GATA II/III form a single PF with a GATANNGATA consensus in dicotyledons. 2. Apparently GATA I is a more ancient element than GATA II/III, inferred from the fact that the former, but not the latter, element is present in Lhcb promoters of gymnosperms (11, 71).

Figure 1  (a) Lhcb1 promoter regions that are homologous but dissimilar. The indicated CGF-1 binding site was defined by Teakle & Kay (121). The Z-DNA element of the Arabidopsis Lhcb1*3 promoter has been functionally characterized by Ha & An (54) and by Puente et al (98). The illustrated region corresponds to the cabCMA-3 of the genes indicated in the figure. (b) Hypothetical arrangement of the upstream LRE of Lhcb1 promoters in the common ancestor of Dicotyledons. Two putative LREs derived from the ancestor, which have been previously identified as cabCMA-3 and cabCMA-2 (10), are shown. GATA-containing sequences of maize genes that are inverted with respect to dicot homologous motifs are underlined. Notice that in some cases, a CCA-1 motif is the evolutionary counterpart of an I-box element.
Within dicots only the CCAAT and GATA motifs are invariably present in the TATA-proximal region. The CUF-1 and CCA-1 binding sites are found in some but not all \( Lhcb1 \) genes. Only one tomato gene \( (Lhcb1^+5) \) contains both CUF-1 and CCA-1 binding sites, four \( (Lhcb1^+1, ^+2, ^+3, \) and \( ^+4) \) contain the CUF-1 element but lack the CCA-1 site, and two \( (Lhcb1^+6 \) and \( ^+7) \) lack a discernible CUF-1 site but contain an inverted CCA-1 binding site. Although it is probable that these structural differences determine qualitative differences in the regulation of these paralogous promoters, transcript abundance in tomato leaves is not correlated with the composition of their proximal region (97).

Most of the known \( Lhcb1 \) genes of monocotyledons have promoters clearly distinct from those of dicots, with two exceptions: the \( \text{Lemna} \) \( Lhcb1^+1 \) \( (AB30) \) and the maize \( Lhcb1^+2 \) \( (cab-1) \) genes \( (69, 118) \), both containing a \( (\text{CUF-1 site})-(\text{CCAA T-box})-(\text{GA TA II})-(\text{I-box}) \) arrangement homologous to that of dicot promoters \( (10) \). This arrangement in \( Lhcb1 \) promoters may either predate the divergence of monocots and dicots or result from convergent evolution. Support for an ancient origin comes from an orthologous gene of a conifer \( (\text{Pinus contorta}) \) that has a promoter \( (\text{CCAA T-box})-(\text{I-box})-(\text{TATA-box}) \) arrangement that is similar to that found in angiosperms (see Figure 6). The pine gene lacks the GATA III, CCA-1, and CUF-1 elements, suggesting that these elements probably evolved after the divergence of the lineages leading to modern conifers and angiosperms.

**LREs in the Central Region of \( Lhcb1 \) Promoters**

\( Lhcb1 \) genes contain additional LREs, besides those found in the TATA-proximal region. Castresana et al \( (23) \) demonstrated by gain-of-function experiments the photoresponsiveness of the \( -396 \) to \( -186 \) region of the tobacco \( \text{Cab-E} \) gene. A light-responsive region has also been mapped to the \( -347 \) to \( -100 \) promoter region of the pea \( Lhcb1^+2 \) \( (AB80) \) gene. This 247-bp fragment was shown to function as a light-responsive, tissue-specific enhancer in gain-of-function experiments \( (113) \). Two nuclear factors binding this regulatory region were identified, one of which \( (\text{ABF-1}) \) is found only in green tissues \( (7) \). By systematic deletion the \( AB80 \) enhancer light-responsive core is located in the \( -200 \) to \( -100 \) region \( (\text{Argüello-Astorga} \ & \ \text{Herrera-Estrella, manuscript in preparation}) \). Evidence for LREs in the \( -240 \) to \( -100 \) region of the \( \text{Arabidopsis} \) \( \text{Cab2} \) promoter has been also reported \( (121) \).

**Evolution of the Central LRE of \( Lhcb1 \) Promoters**

The comparative analysis of the tobacco \( \text{CabE} \) and pea \( AB80 \) upstream LREs, with the corresponding promoter segments from other \( Lhcb1 \) genes from dicots, suggests a structure of this promoter region in the common ancestor of dicotyledons. An extensive ancestral CMA encompassing at least five PFs is
inferred (Figure 1b). This ancestral organization is still conserved in some genes of leguminous species [e.g. pea Cab-8 (3)]. The central PF of this CMA is a sequence similar, but not identical, to the G-box from rbcS genes. In tobacco CabE this G box–like element (G∗box) is associated with a Box II–like element corresponding to a PF with a GATA core motif (23). Homologous G box–like and 5′ associated GATA-motif elements (I∗box) exist in members of the Lhcb1 gene family in plants belonging to four orders of dicotyledons. A third PF (YCCACART) is found immediately upstream of the G-box element in several Lhcb1 genes of legumes and the tobacco Cab-E promoter but not in other dicot genes. Interestingly, this PF is also found in the homologous region of two maize Lhcb1 genes (Figure 1b), suggesting a very ancient origin of this PF. A fourth PF with the consensus CATGGCTA closely precedes a PF encompassing an I-box motif located 15–20 bp downstream of the G-box element. The arrangement of this Lhcb1 region (cabCMA2) is highly analogous to that of the G-, I-box region of dicot rbcS promoters, with a very similar PF associated 5′ to the I-box motif. The resemblance in some cases is so striking that a Lhcb1 promoter (e.g. pea AB80) can display more similarity in this region with the analogous rbcS promoter segment (e.g. pea rbcS 3A) than with certain homologous Lhcb1 regions (10). The analogies between the overall structural organization of the I∗-G∗-I promoter region of Lhcb1 genes and the I-G-I unit of rbcS promoters are intriguing and suggest either a common ancestry or convergent evolution of these regulatory promoter modules.

**Delimitation of a LRE in Lhcb2 Promoters**

Although comparatively few members of the Lhcb2 gene family have been studied, the promoter of one, the Lhcb2∗1 (formerly AB19) gene from Lemna gibba, has been characterized in great detail. Deletion analysis, linker scanning, and site-directed mutagenesis identified two 10-bp elements in the −134 to −105 region that are critical for light responsiveness. One of them contains the I box–related GATAGGG motif and the other a CCAAT motif. Mutation of the latter element led to high levels of expression in the dark, suggesting that it binds a repressor in the absence of light. Mutations in the I-box motif led to complete loss of red-light responsiveness, suggesting that this region is involved in PHY- mediated light activation of transcription (66). More detailed site-directed mutagenesis of those 10-bp regions allowed the identification of two shorter sequences (REα = AACCAA and REβ = CCGATA) that are critical for AB19 light regulation (31).

**Evolution of Lhcb2 Promoters**

Genomic clones of Lhcb2 genes have been isolated from the moss Physcomitrella patens (79), the fern Polystichum munitum (96), the conifer Pinus
thunbergii (71), the monocot Lemna gibba (64), and three dicotyledonous plants (pea, Petunia, and cotton) (39, 103, 115). These promoters contain a number of PFs, two of which correspond to the Lemna REα and REβ elements (Figure 2). Only a few of the identified PFs are present in all species. Among the PFs common to most if not all the known Lhcb2 promoters are the Lemna REα element, generally recognized as a conserved CCAAT box, and a G(G/A)AAATCT motif, which is similar to the CCA-1 binding site of the Lhcb1 genes.

An interesting observation is that the gene harboring sequences with the highest similarity to the Lemna REα and REβ light-responsive elements is that of P. thunbergii (see Figure 2), whose promoter directs light-independent gene expression in both its native context (71) and in transgenic angiosperms (70, 138). This paradoxical observation is discussed below. Another relevant observation is that Lhcb2 genes from dicotyledons lack a REβ element, instead displaying three lineage-specific PFs, two of which include inverted GA TA motifs (Figure 2). It would be interesting to determine whether those elements are functionally equivalent to REβ.

Upstream sequences of Lhcb2 and Lhcb1 genes are probably derived from a common ancestral promoter, because they display a similar overall structural organization, with a (G Box/CUF-1 site)-(CCA-1 element)-(CCAA T-box)-(I-box)-(TA TA box) basic arrangement. Spacing between these elements is, however, very different in the two gene families.

EVOLUTION OF LREs IN rbcS PROMOTERS

The Light-Responsive Box II-III Region

The pea rbcS-3A gene has been used as a paradigm for the study of light-regulated gene expression in dicotyledons. Analysis of the rbcS-3A promoter has uncovered three independent regions that contain an LRE (47). The LRE located in the −166 to −50 region has been characterized in most detail (45). This region includes two elements called Box II and Box III, both of which are binding sites for a nuclear factor named GT-1 (49). This factor binds in vitro to sequences related to the degenerate consensus (A/T) GTGPu (T/A) AA (T/A) (50). A synthetic tetramer of the pea Box II element (GTGTGGTTAA TAA TG) conferred light-responsive transcriptional activity to the −90 CaMV 35S promoter (75, 98) but was unable to enhance transcription when fused to either the −46 CaMV 35S or the −50 rbcS-3A minimal promoters (28). This element appears necessary, but not sufficient, for light-regulated transcriptional activation. Arabidopsis and tobacco cDNAs encoding GT-1, a Box II DNA binding proteins, have been cloned and partially characterized (46, 58, 95). Interestingly, GT-1 is closely related to the GATA-binding nuclear factors CGF-1 and IBF-2b and can bind to similar cognate DNA sequences (121). Two other
Figure 2  Sequence changes in *Lhcb2* gene promoters from land plants. DNA motifs represented by identical symbols but shaded differently indicate related but not identical sequences.
proteins called 3AF-3 and 3AF-5 probably act together with GT-1 to confer rbcS-3A light-regulated expression; they bind inverted GATA motifs located at each end of Box III. Mutations in these GATA sequences severely reduced promoter activity (104).

**Evolution of the Box II-III Homologous Regions**

Upstream sequences of many angiosperm rbcS genes are known. However, only two rbcS promoters are known from nonflowering plants, the conifer Larix laricina (59) and the homosporous fern Pteris vittata (56). Comparison of the Pteris and Larix proximal promoter regions (−130/−1) reveals only four PFs, two of which are inverted I box–related motifs: one overlapping the putative TATA-box of the Larix gene, and the second, with the sequence GTTATCC, found several bp upstream. In the Larix promoter, this PF is found as an imperfect direct repeat, flanked by two additional PFs, one of which encompasses a CCAAT motif. The relative position of this CMA in the conifer promoter (i.e. ~25 bp downstream to the I-box element) is practically identical to that of the Box II–3AF3 region in dicot rbcS promoters. Comparison of these CMAs uncovers several interesting characteristics:

1. The Box II element seems to be evolutionarily derived from two separate Pteris/Larix PFs (Figure 3). One of them is the most 5′ GTTATCC motif in Larix, which is homologous to the 3′ half of Box II. This relationship is especially clear in rbcS promoters of the Brassicaceae (see Figure 6). Therefore, Box II could be a composite element bound in vivo by two protein factors, one of them being a GATA-binding factor.

2. The second repeat of the GTTATCC motif of Larix is homologous to a conserved sequence immediately downstream to Box II, which in Arabidopsis and Brassica genes is almost identical to the so-called LAMP binding site (51), an inverted I-box element. In nonbrassicaceous dicots the sequences immediately downstream of Box II do not resemble the LAMP motif or the Pteris/Larix PF, but their structural relationship is easily recognizable in a phylogenetic series (not shown, but see 82).

3. The Pteris/Larix CCAAT-box is apparently homologous to the pea 3AF3 binding site (104). In the conifer promoter the LAMP-like motif and the CCAAT box are close, but in dicots the 3AF3 element is separated from the LAMP-related motif by 10–24 bp. Sequences in this intermediate DNA are functionally relevant, encompassing in pea rbcS-3A the Box III and the 3AF5 elements (104) (Figure 3).

The Box II–containing CMA is absent in all known rbcS genes of monocotyledons (~10), and in the orthologous genes of a dicot species, the common
Figure 3: Sequences of rbcS gene promoters in vascular plants.
ice plant (Aizoideae) (35). Our interpretation is that these groups have lost Box II, which we consider an ancestral feature of rbcS genes.

The Light-Responsive I-G Unit

rbcS promoters usually contain two closely associated elements, the I and G boxes. Mutations of either the G-box or the two flanking I-box elements of the Arabidopsis rbcS-IA promoter almost abolished its activity (36). A similar drastic drop of transcriptional activity was observed in the spinach rbcS-I promoter when the G-box element was mutated (80). In spite of its functional relevance in dicots, deletion of a G-box in the rbcSZm1 gene of maize had no significant effects on promoter activity. Deletion of the associated I-box motif reduced expression in light 2.5-fold (106).

Lemna rbcS promoters lack a typical G-box element but contain a canonical I-box motif within the so-called X-box element, which is part of the binding site of LRF-1, a light-regulated nuclear factor (21). This I-box motif is included in a 30-bp region of the Lemna rbcS SSU5B promoter necessary for PHY regulation (31).

Gain-of-function experiments showed that the region of the rbcS promoters encompassing the I- and G-box elements functions as a composite LRE, able to direct a tissue-specific pattern of expression almost identical to the native rbcS promoter, including the dependence on the developmental stage of plastids (9). Mutation of either the I-box or the G-box eliminated detectable transcription (9). The I-G region functions as a complex regulatory unit, similar to that of the light-responsive Unit 1 of the parsley chs gene (108).

Evolution of the I-G Region

The two I-box elements of the I-G-I CMA display a different sequence consensus and are flanked by different conserved sequence motifs. The more upstream I-box has the consensus GATAAGAT (A/T) and is adjacent 3′ to a PF with the consensus (A/T) ARGA TGA; the second I-box has the consensus ATGATAAGG and is 5′ flanked by a PF with the consensus TGGTGGCTA (Figure 3). The I-box-associated PFs are conserved in genes of plants belonging to five orders of dicotyledons. A homologous I-G-I arrangement is found in some maize rbcS genes but not in other genes from monocotyledons. However, CMAs that are probably derived from an ancestral I-G-I structure are present in all these promoters (Figure 3).

The finding that the I-G (-I) arrangement is also found in the homologous promoters of a conifer and a fern (Figure 3) is unexpected because these promoters are presumably light-insensitive, whereas the I-G unit is apparently involved in responses to light signals in angiosperms. Based on the persistence of the I-G-I arrangement since at least the divergence of ferns and seed plants, 395 mya (116), we propose that it has played, and probably still does, one or several
important functional roles, different to light-regulation, in the control of rbcS gene transcription.

EVOLUTION OF THE LIGHT-REPPRESSED PHYTOCHROME A PROMOTER

Functional Organization of phyA Promoters
Phytochrome (PHY) is encoded by small gene families in angiosperms; in Arabidopsis the PHY apoprotein is encoded by five genes (99). Each PHY is proposed to have a different physiological role (100). PHYs have been classified into two types. The type I, or “etiolated-tissue” PHY, is most abundant in dark-grown plants, and its Pfr form is rapidly degraded in light by an ubiquitin-mediated proteolytic process. Type II or “green-tissue” PHY are present in much lower levels, but their Pfr form is stable in light (99, 100). The only known type I PHY is that encoded by the phyA gene, whose mRNA abundance also decreases in light. This inhibition of phyA gene activity is autoregulatory (PHY-dependent) and operates at the transcriptional level (19, 65, 77).

The phyA promoters of two monocotyledons, oat and rice, have been functionally characterized. A combination of deletion analysis and linker-scan mutagenesis identified in oat three cis-regulatory elements designated PE1 (positive element-1), PE3 (positive element-3), and RE1 (repressor element-1) (19, 20). PE1 and PE3 act synergistically to support maximal expression under derepressed conditions (i.e., low Pfr levels); mutation of either element decreases expression to basal levels. In contrast, mutation of the RE1 element results in maximal transcription under all conditions, suggesting that Pfr represses phyA transcription through this negatively acting element (19).

The rice phyA promoter has no element similar to PE1. Instead, this promoter contains a triplet of GT-elements that have been shown, in transient expression assays, to be functionally equivalent to the oat PE1 element (32, 99). These elements, related in sequence to the GT-1 binding sites of rbcS promoters (65), are bound by a transcription factor, named GT-2 (32, 34).

In addition to phyA, several other genes are down-regulated by light (88, 92, 123), including the genes encoding asparagine synthetase (127). Recently a 17-bp element was identified that is both necessary and sufficient for the PHY-mediated repression of the pea asparagine synthetase gene. This sequence is very similar to the phyA RE1 element and is the target for a highly conserved PHY-generated repressor, whose activity is regulated by both calcium and cGMP (91).

Evolution of phyA Promoters
Phytochrome genes have been found in phototropic eukaryotic organisms ranging from algae to angiosperms (42, 99), and genomic clones of phy genes have been isolated from angiosperms and several lower plants. The latter include
species representative of lineages dated to the Silurian and Devonian, and which diverged more than 400 million years ago (27, 110, 116).

The cereal PE3-RE1 region is conserved in the phyA promoters of plant species belonging to three different orders of dicotyledons (1, 33, 105), but no obvious counterpart of the monocot GT-boxes region was detected. The conservation of the PE3-RE1 arrangement in dicots suggests a conserved regulatory activity of these elements (33). The phyA upstream sequences of bryophytes (the mosses Physcomitrella and Ceratodon; 107, 124), a lycopodiophyta (Selaginella; 57), and a psilotophyta (Psilotum; 107) contain a number of PFs. Two clusters of these PFs or CMAs are common to all reported lower plant phyA promoters (Figure 4). The CMA nearest the start codon of the mosses and lower vascular plant genes includes sequences similar to the PE3 core element and other flanking sequence elements. The central, most conserved DNA motif of the lower plant CMA is nearly identical to a sequence element in the Arabidopsis PE3–RE1 region, immediately upstream of the RE1 motif. No canonical RE1 element is found in lower plant phy promoters.

Interestingly, the second, more distal lower plant phyA-CMA seems to be the evolutionary counterpart of the region encompassing the GT boxes in monocot phyA promoters (Figure 4). Thus, the cis-regulatory elements of monocotyledon phyA genes seem to have evolved from cis-acting sequences already present in orthologous genes of the common ancestor of land plants. What the regulatory function was of such elements in the primitive land plants is an intriguing question that could be partially solved by the functional characterization of the identified phyACMAs in mosses and vascular cryptogams.

**Evolution of chs Promoters**

In some species, induction of chs gene expression by light is required for flavonoid accumulation, which provides a protective shield against potentially harmful UV irradiation (55).

The promoters of parsley and mustard chs genes have been functionally analyzed, and a 50-bp light-responsive conserved region, named light-regulatory unit I (LRU1), has been studied in great detail (41, 62, 63). LRU1 is sufficient to confer light-responsive heterologous minimal promoters (63, 102, 132) and consists of at least two distinct cis-acting elements, ACEchs and MREchs (formerly Box II and Box I, respectively) (108, 109). ACEchs contains a G-box element that interacts in vivo and in vitro with bZIP regulatory factors (40, 132). MREchs was originally identified as a 17-bp in vivo DNA footprint (109) with a conserved sequence motif called the H-box core (78), which is recognized by PcMYB1, a Myb-related transcription factor of parsley (41). Recently, it was shown that chs LRU1 activity is controlled by PHY through the cGMP-dependent transduction pathway (136). Comparative analysis of
Figure 4. Sequences of phyA promoters in land plants.
orthologous chs promoters showed that LRU1 homologous sequences are present only in genes of cereals and brassicaceous plants (10). Homologous genes of legumes, snapdragon, and carrot contain, in the same relative position as LRU1, a CMA comprising G-box and H-box elements as in LRU1, but spaced differently (6–8 bp). Several regulatory functions have been assigned to this G-H box CMA, including transcriptional activation by p-coumaric acid (78) and tissue-specific transcription (38). Although this CMA is found within light-responsive regions of some chs genes (e.g. Anthirrinum chs; 76), there is no direct evidence of a role in light regulation. Because of both their proximal position to the TATA box and their PF composition, LRU1 and the G-H CMA are clearly homologous cis-regulatory regions. Nonetheless, evolutionary divergence of chs genes harboring LRU1 and G-H CMA, respectively, date back to an era before the divergence of lines that gave rise to monocotyledons and dicotyledons, as inferred from the phylogenetical distribution of LRU1 (10, 26).

EVOLUTION OF LIGHT-REGULATED PARALOGOUS GENE PROMOTERS

Differential Activity of Paralogous Gene Promoters

Several light-regulated genes such as Lhcb, rbcS, and chs are found in multiple copies in most genomes. Members of these gene families frequently display quantitative and/or qualitative differences in expression reflecting in part transcriptional regulation (29a, 30, 82, 117, 133). Some of these functional differences correlate with differences in promoter architecture. For example, among the eight rbcS genes of Petunia, only the two most highly expressed (SSU301 and SS611) contain the I-G-box arrangement (29a). Insertion of an 89-bp fragment of the SSU301 promoter containing the I-G unit into the equivalent region of the weakly expressed SSU911 gene increased its expression 25-fold (29). Tomato has five rbcS genes (117). Three have promoters with the I-G unit. The mRNAs from all five tomato rbcS genes accumulate to similarly high levels in leaves and light-grown cotyledons; however, only the genes containing the I-G unit are coordinately expressed in dark-grown cotyledons, water-stressed leaves, and developing fruits (12). Interestingly, the spacer DNA sequence between the I-box and the G-box elements is highly divergent in tomato rbcS genes but conserved in orthologous genes from other plant species. A fruit-specific factor (FBF) specifically interacts with the I-G spacer DNA sequence of the tomato rbcS-3A promoter, which correlates with its reduced activity in developing tomato fruit (84). The tomato rbcS-2 and tobacco rbcS 8.0 display a very similar sequence to the pea rbcS-3A I-G spacer element, which is bound in vitro by GT-1 (50). It is conceivable that these “paralogous gene-specific
motifs” modulate environmental or tissue-specific effects on the structurally invariant I-G regulatory unit.

Do Paralogous Promoters Evolve by Nonrandom Processes?

Paralogous gene promoters often diverge in discrete segments, displaying nonrandom patterns of structural variation. For example, in the *rbcS* family of *Arabidopsis* the paralogous 1B, 2B, and 3B gene promoters clearly differ from that of *rbcS-1A* by a 60-bp internal deletion. This molecular event juxtaposed the photoresponsive Box II–LAMP element with downstream conserved motifs, creating a new combination of cis-regulatory elements. In the *rbcS-1B* promoter an additional deletion event removed a 44-bp region encompassing the I-G-I Unit, juxtaposing the two I boxes originally flanking the G-box sequence (72). This gene is the only member of the *Arabidopsis rbcS* gene family unable to respond to light pulses (30).

The evolution of discrete, short regulatory elements interspersed along the promoters of paralogous genes also seems to occur, and this is exemplified by the maize *rbcSZm1* and *rbcZm3* genes. Their promoter sequences are very similar but differ in the presence of small insertions, which are distributed in a discontinuous pattern (Figure 5a). Some of these paralogous gene-specific motifs are cis-regulatory elements (106).

![Figure 5](image-url)

**Figure 5** Paralogous gene promoters. (a) Schematic comparisons of two maize *rbcS* promoters mainly differing by small insertion/deletions events; some changes create cis-acting elements [based on data from Schäffner & Sheen (106)]. (b) Schematic comparison of two functionally distinct *Lhcb1* promoters of pea. Segments of high sequence divergence coinciding with relevant cis-regulatory elements are shown. Clear rectangles represent blocks of conserved sequences, whose similarity is indicated as the ratio between the number of identical nucleotides and the overall longitude (in bp) of the compared promoter segments.
A probable example of evolution of differential function by loss of specific cis-acting elements is available with the Lhcb1 genes of pea. Two of these genes (Cab-8 and AB96) showed significant transcript accumulation after a red light pulse, whereas the other three Lhcb1 genes (AB80, AB66, and Cab-9) require continuous red light for significant expression (133, 134). The pea Cab-8 gene encodes a mature protein 99.5% similar to that of the AB80 and AB66 genes (3, 125) and it displays significant similarity in the 350-bp proximal part of the promoter. Cab-8 has a consensus dicot Lhcb1 promoter, with a distal I*-G*-I structural unit and a proximal arrangement of elements (CUF-1)-(CCA-1)-(CCAA T-box)-(CGF-1 site). In the AB80 and AB66 promoters the original CCAA T box sequence seems to have been eliminated by an 8 nt internal deletion, the CUF-1 binding site is lost by multiple nucleotide substitutions, and the conserved GATA motif, upstream to the G-box, is mutated in the G residue and in a few additional upstream nucleotides (Figure 5b). Because the Cab-8 promoter has the organization of the hypothetical, ancestral dicot Lhcb1 promoter, the architecture of the AB80 and AB66 promoters could be considered as derived, by mutation and deletion of specific cis-regulatory elements, from a Cab-8-like promoter. The cause of the mutations and the selective forces that fixed them in the species remain unknown.

LRE-ASSOCIATED CMAs

**CMAs in Additional Plant Genes**

Photoresponsive regions have been defined in other genes by diverse approaches (13, 122). Some cis-regulatory elements different to those identified in rbcS, chs, and Lhcb promoters have been proposed for light-regulated transcription (14, 81, 93). Discrete arrays of sequence motifs in light-responsive regions of those genes are conserved between phylogenetically distant plant species (10). Some of these LRE-associated CMAs are present in orthologous genes from both monocots and dicots, indicating a very remote evolutionary origin. These ancient CMAs are found in genes encoding ferredoxin, and the pyruvate–orthophosphate dikinase, sedoheptulose-bisphosphatase, and the A subunit of chloroplast glyceraldehyde 3-phosphate dehydrogenase. LRE-associated CMAs were also identified in dicot genes encoding plastocyanin, subunits of the chloroplast ATPase, a 10-kDa protein of photosystem II, 4-coumarate:CoA ligase, and phenylammonia-lyase (10).

**Structural Analogies Between LRE-Associated CMAs**

Most of the ~30 identified CMAs (10) can be grouped in a small set of structural and phylogenetic types.
1. The (I-box)-(G-box)-(I-box) arrangement. Included are the I-G and G-I units of \( rbcS \) promoters and their evolutionary variants, including the (X-box)-(Y-box) region of \( \text{Lemna} \) genes (Figure 3); the \( \text{I}^*\text{-G}^*\text{-I} \) ancestral arrangement of \( \text{Lhcb1} \) genes found in legumes, and their evolutionary derivatives such as the (GATA-motif)-(Z-DNA) region of \( \text{Arabidopsis Lhcb1}^* \); the (Box II)-(G-box) arrangement of tobacco \( \text{CabE} \), and the (G-box)-(CCA-1 element) region of some spinach, tobacco, and \( \text{Arabidopsis} \) genes (Figure 1); also in this CMA group are included the \( \text{I}^*\text{-G}-(\text{I}) \) array of \( \text{Fed} \) promoters; and the inverted I-G-I unit of dicot and monocot \( \text{sbp} \) genes.

2. The (GT-1)-(LAMP-site)-(GT-1) arrangement, observed in both the Box II-3AF3 region of \( rbcS \) genes (\( rbcS \) CMA-3) and in the CMA-1 of \( \text{LS} \) genes.

3. The (CCAA T-motif)-(GA TA/I-box) combination, found in several light-responsive promoters, including the RE\( \alpha \)-RE\( \beta \) unit of \( \text{Lemna Lhcb2}^* \) (i.e. \( \text{cabCMA4} \)), the (CCAA T)-(GATA I-III) arrangement from \( \text{Lhcb1} \) promoters (i.e. \( \text{cabCMA1} \)), the (CCAA T box)-(motif 15) from solanaceous \( rbcS \) genes (i.e. \( rbcS \) CMA-2), and the CMA containing the PC2 region from plastocyanin gene promoters (i.e. \( \text{PcCMA-2} \)).

4. The (LAMP-site)-(TA TA-box) arrangement, characteristic of all of the solanaceous \( rbcS \) promoters (82) and found in a modified form in the \(-50/+15\) light-responsive region of pea \( rbcS-3A \) (73). This arrangement is also observed in the LRE-associated CMA of \( \text{atpC} \), in plastocyanin gene promoters (i.e. \( \text{PcCMA-1} \)), and in \( \text{gapA} \) CMA-1.

5. The (G-box)-(H-box) arrangement, found in the three identified CMAs of \( \text{chs} \) genes, including the photoresponsive units 1 and 2 of parsley \( \text{chs} \) gene (108, 109).

**Common Structural Features of Composite LREs**

Do all, or most, of the genes whose transcription is dependent on PHY harbor a common \( \text{cis} \)-regulatory element in their promoters? This appealing idea has not been confirmed by the sequence data from the dozens of PHY-dependent genes. The identification of LRE-associated CMAs provides a new opportunity to assess whether the regulatory regions of those genes share structural features that could explain their common dependence on such a photoreceptor. The comparative analysis of \(~30\) of these natural combinations of sequence motifs led to two important findings. 1. All of the LRE-associated CMAs present in PhANGs include at least one sequence identical, or related, to either the I-box core motif or its inverted version, the LAMP-site. 2. All of the CMAs found in genes encoding enzymes involved in the metabolism of phenylpropanoids.
(PhEMAGs) share a conserved module related to the chs H-box core motif, ACCTA(A/C) C (A/C) (10).

Because PHY regulates expression of its target genes by three different transduction pathways (85), the superfamily-specific conserved motifs could be binding sites for transcription factors targeted by specific PHY-signaling pathways (10). Based on the knowledge of the genes that are activated or repressed by these phototransduction pathways, it has been proposed that the I-box/GATA-binding factors are direct or indirect targets of the Ca²⁺/calmodulin-dependent transduction pathway, whereas transcription factors binding at the H-box motifs in PheMAG CMAs would be affected by the cGMP-dependent phototransduction pathway (10).

A general model of LRE function proposes that LREs are multipartite cis-regulatory elements with two general components: “light-specific” elements and “coupling elements.” The former are bound by transcription factors targeted by the light-signal transduction pathways (i.e. I-box/GATA-binding factors and HBFs), which confer photoresponsiveness. Coupling elements are bound by either cell-specific factors or regulatory proteins targeted by other signaling systems; consequently, the light stimulus to transcription is coupled to other endogenous and exogenous signals.

Using microinjection into single cells of the tomato aurea mutant, Wu et al (136) established that constructs containing either 11 copies of the rbcS Box II element or 4 copies of the chs Unit 1 are activated by different PHY signaling pathways. Box II is affected by the calcium-dependent pathway and the chs Unit 1 is activated by the cGMP pathway. Taking into account that GT-1, the factor that binds Box II, is highly related to the nuclear factors CGF-1 and IBF-2b, both of which bind I-box/GATA motifs (121) the work by Wu et al (136) supports the hypothesis that factors interacting with I-box-related sequences are targets for the Ca²⁺/calmodulin PHY activated pathway. Moreover, Feldbrügge et al (41) recently determined that the light-responsive core of parsley chs LRU1 is indeed the H-box, making it the most likely target of the cGMP pathway.

Evolution of LREs: The Chloroplast Connection
The finding that LREs from angiosperm gene promoters are very similar to putative regulatory units present in promoters from conifers and lower plants (Figure 6) is unexpected, because it is generally assumed that such promoters are either light insensitive or, at most, weakly photoresponsive. Such physiology is
I. REα-REβ region of the Lemna Lhcb2*1 (AB 19) promoter

Physcomitrella Lhcb 2*1

Lemma gibba AB19

Pinus thunbergii Cab 6

II. CCAAT-GATA region of Lhcb1 promoter

a) Petunia hybrida Cab 22R

Pinus thunbergii Cab 6

b) Lemma gibba AB30

Pinus contorta Lhcb1*1

Zea mays Cab-1

III. I-G unit of rbcS promoters

Pteris vititata rbcS-1

Larix laricina rbcS-1

Mesembryanthemum crystallinum rbcS-1

Nicotiana tabacum rbcS-1

Helianthus annus rbcS-1

IV. G-I unit of rbcS promoters

Larix laricina rbcS-1

Brassica napus rbcS-1

V. Box II element from rbcS promoters

Larix laricina rbcS1

Arabidopsis thaliana rbcS 3B

Brassica napus rbcS F1

PF-1

PF-2
well established in conifers (4, 71, 89, 94, 137), although to our knowledge, no
detailed molecular studies have been carried out in pteridophyta and bryophy-
tas. Conifers and other nonflowering plants (Gingko biloba being an exception;
24, 25) develop chloroplasts when grown in darkness (83, 129, 130), indicating
that their PhANG promoters are active in the dark. A central question emerges:
did light-responsiveness evolve by changes in cis-regulatory elements or trans-
acting factors, or by both.

Because I-box/LAMP elements, which are critical for PHY responsiveness in
angiosperms, are also components of ancestral, presumably non-photorespon-
sive, regulatory units of Lhcb and rbcS genes, it is probable that factors bind-
ing these motifs became direct or indirect targets of light-signaling pathways
in organisms preceding flowering plants. This notion is in agreement with the
available experimental evidence, including the recent demonstration that
synthetic pairwise combinations of I-box sequences with diverse conserved el-
ements function as complex LREs (98). However, the presence of an I-box
in combination with other conserved sequence motifs is not necessarily suf-
ficient for light regulation. The Pinus thunbergii cab-6 promoter containing
Reα and Reβ (an I-box motif) elements identical to those of the Lemna AB19
gene (Figure 6) directs a light-independent and tissue-specific expression of a
reporter gene in both dicots and monocots transgenic plants (70, 138). There-
fore, other cis-acting signals in addition to the I-box core seem to be necessary
for proper light control in flowering plants.

Our data suggest that composite LREs evolved from regulatory units that per-
formed functions other than light-regulation. What could these functions have
been? In the case of PhANGs, we hypothesize that these functions were related
to nuclear gene regulation by chloroplast-derived signals. This possibility is
supported by the finding that even the smallest light-responsive PhANG pro-
moter segments display a tissue-specific and chloroplast-dependent pattern of
expression similar to that of entire promoters; to date no evidence has been ob-
tained that these two functions can be separated (9, 14, 15, 74, 98, 114, 120).
Because coordination of gene expression between nuclear and chloroplast
genomes should have evolved a long time before terrestrial plants (101), it
is plausible that PhANG cis-acting promoter elements targeted by plastid sig-
nal transduction pathways evolved before LREs. Therefore, it is possible that
photoreceptor-mediated transcriptional regulation was produced during evo-
lution by targeting, either directly or through new regulators (i.e. via protein-
protein interactions), the same transcription factors and cis-regulatory elements
that mediated the influence of chloroplasts on PhANGs transcription. This possi-
bility is attractive because it suggests a simple mechanism by which different
gene families whose expression is coordinate could simultaneously acquire a
new, coordinated pattern of regulation.
CONCLUDING REMARKS

Much remains to be learned about the structure, mode of action, and evolution of LREs. It is clear that LREs are complex, composed of at least two cis-acting elements, that can be targeted by different photoreceptor-activated signal transduction pathways. The composite and variable structure of LREs could explain the specific properties of individual light-responsive promoters. Phylogenetic analysis of LREs clearly indicates that they have evolved from ancient regulatory elements, whose original, primary function was probably not light regulation. Transformation of mosses and other nonflowering plants in which the expression of photosynthesis-associated genes is not regulated by light could help in answering some questions concerning the evolution of LREs and the different signal transduction pathways that activate them. It will be of great interest to explore how a new mode of regulation, affecting many gene families involved in photosynthesis, arose during evolution.

ACKNOWLEDGMENTS

We are grateful to June Simpson for critical reading of this manuscript. This work was supported by a grant from the HHMI to LH-E (75 197-526902).

Literature Cited

ARGÜELLO-ASTORGA & HERRERA-ESTRELLA


34. Dehesh K, Hung H, Tepperman JM, Quail PH. 1992. GT2: a transcription factor with twin autonomous DNA-binding domains of closely related but different target sequence specificity. EMBO J. 11:4131–44
36. Donald RGK, Cashmore AR. 1990. Mutagenesis in either G box or I box sequences profoundly affects expression from the Arabidopsis thaliana rbcS-1A promoter. EMBO J. 9:1717–26
552 ARGUELLO-ASTORGA & HERRERA-ESTRELLA


74. Loake GJ, Faktor O, Lamb CJ, Dixon RA. 1992. Combination of H-box (CCA-TACC(N)7CT) and G-box (CACGTT) cis-elements is necessary for feed-forward stimulation of a chalcone synthase promoter by the phenylpropanoid-pathway intermediate p-coumaric acid.
PROMOTER EVOLUTION 553


90. Deleted in proof


ARGUELLO-ASTORGA & HERRERA-ESTRELLA

3AF5 and 3AF3, are required for rbcS-A expression. *Plant Cell* 4:473–83


119. Deleted in proof


